MDPI

Article

Joint Face Detection and Multi-Task Estimation of Age, Gender, and Ethnicity Using Deep Neural Networks

Abdessamad Hnioua[1], Imane Lakouira[1] and Abdessamad El Boushaki[2]

[1] Master IAII, Department of Computer Science, Faculty of Sciences and Techniques (FST), Cadi Ayyad University, Marrakech, Morocco
[2] Department of Computer Science, Faculty of Sciences and Techniques (FST), Cadi Ayyad University, Marrakech, Morocco
* Correspondence: a.hnioua5280@uca.ac.ma ,i.lakouira4247@uca.ac.ma, a.elboushaki@uca.ac.ma

Abstract

Facial analysis is a crucial task in computer vision with applications in security, human-computer interaction, social media analytics, and demographic studies. This work presents a robust and scalable framework for joint face detection and multi-task facial attribute analysis, including age estimation, gender classification, and ethnicity recognition. The proposed system integrates a YOLOv8-based face detection module with a Multi-Task ResNet50 architecture, enabling shared feature extraction while simultaneously predicting multiple attributes. The YOLOv8 detector is trained on a dedicated face dataset of approximately 16,700 images annotated with precise bounding boxes, ensuring robust face localization under varying poses, lighting, and occlusions. The Multi-Task ResNet50 model leverages pre-trained ImageNet weights and task-specific heads to predict age, gender, and ethnicity from cropped facial regions. Experimental results on both the custom face detection dataset and the UTKFace dataset demonstrate high detection accuracy (mAP@0.5 = 0.884) and strong multi-task performance, with a gender classification accuracy of 93.03%, ethnicity F1-scores up to 0.89, and age MAE ranging from 1.42 to 8.95 across age groups. Qualitative results confirm reliable predictions under challenging real-world conditions. Overall, this framework provides an effective and efficient solution for comprehensive facial analysis in unconstrained environments.

Keywords: Facial analysis; Deep learning; Face detection; YOLO; Multi-task learning; ResNet50; Age estimation; Gender classification; Ethnicity recognition

1. Introduction

Automatic facial analysis has become a central topic in computer vision, due to its wide range of practical applications including security surveillance, access control, demographic analytics, and human-computer interaction. Key tasks in facial analysis include face detection, age estimation, gender classification, and ethnicity recognition. Traditional approaches often treat these tasks independently, training separate models for each attribute. While effective, this strategy is computationally expensive and fails to exploit the correlations between facial attributes.

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved the performance of facial analysis tasks. Multi-task learning (MTL) frameworks have emerged as an effective solution, enabling a shared backbone for feature extraction and task-specific heads for individual predictions. This approach reduces redundancy, improves generalization, and allows simultaneous prediction

Received:
Accepted:
Published:

of multiple facial attributes. Meanwhile, real-time object detection architectures such as YOLO have set new benchmarks for face detection, providing accurate localization even in complex and unconstrained scenes.

In this work, we propose a unified, modular framework that combines YOLOv8-based face detection with a Multi-Task ResNet50 model for facial attribute analysis. The system operates in a two-stage pipeline: first, the YOLOv8 module detects and localizes faces in the input image; second, the cropped facial regions are processed by the Multi-Task ResNet50 model to estimate age, predict gender, and classify ethnicity. This design ensures high robustness to occlusions, varying illumination, pose variations, and diverse demographic profiles.

1.1. Datasets and Evaluation

The framework is trained and evaluated on two datasets:

- A custom face detection dataset with 16,700 images annotated with YOLO-formatted bounding boxes, providing diverse real-world conditions for robust detection.
- The UTKFace dataset, containing over 23,000 aligned and cropped faces labeled with age, gender, and ethnicity, used to train and validate the Multi-Task ResNet50 model.

Evaluation includes both quantitative metrics and qualitative visual analysis. Face detection performance is assessed using mAP, precision, and recall. Multi-task attribute prediction is evaluated using Mean Absolute Error (MAE) for age, classification accuracy and F1-score for gender and ethnicity, and confusion matrices to highlight class-wise performance.

1.2. Contributions

The main contributions of this work are as follows:

- A cascaded deep learning framework integrating YOLOv8 for precise face detection and Multi-Task ResNet50 for simultaneous age, gender, and ethnicity prediction.
- Detailed analysis of both detection and multi-task performance, including metrics for age, gender, ethnicity, and visual qualitative results.
- Demonstration of robust performance under challenging conditions, including varying poses, illumination, partial occlusions, and diverse demographic profiles.
- Modular and scalable architecture that allows future extension to additional facial attributes or biometric tasks.

2. State of the Art

2.1. Multi-Task Facial Attribute Analysis

Automatic facial attribute analysis has gained significant attention in computer vision, particularly for estimating age, gender, and race from facial images. Recent advances in deep learning, especially Convolutional Neural Networks (CNNs), have led to substantial improvements in performance across these tasks.

Early approaches treated age estimation, gender classification, and race recognition as independent problems, relying on separate models for each task. Although effective, these single-task approaches are computationally expensive and fail to exploit the strong correlations between facial attributes. To address these limitations, *Multi-Task Learning* (MTL) frameworks have been proposed, enabling a shared feature extraction backbone with task-specific output heads.

Several state-of-the-art multi-task CNN models have demonstrated excellent performance on benchmark datasets such as UTKFace, Adience, and MORPH II. Notably, the

work by Butera *et al.*[1] reports a Mean Absolute Error (MAE) of approximately **2.95 years** for age estimation, gender classification accuracy reaching **98.3%**, and race classification accuracy up to **93.1%**. These results clearly outperform many traditional single-task approaches by jointly learning discriminative representations.

Despite these strong performances, existing MTL models still face notable challenges. Performance degradation is observed under unconstrained conditions such as occlusions, low-resolution images, extreme head poses, and illumination variations. Moreover, prediction errors increase significantly for extreme age ranges (children and elderly subjects), and dataset bias remains a critical limitation affecting generalization and fairness.

2.2. Human and Face Detection

Human and face detection has experienced remarkable progress with the rise of CNN-based detectors. Traditional methods based on hand-crafted features and sliding window strategies, such as ACF, LDCF, and Checkerboard models, relied on Integral Channel Features (ICF) and achieved reasonable performance on early benchmarks.

Modern CNN-based detectors have significantly improved detection accuracy by learning hierarchical feature representations. These approaches commonly exploit multi-scale feature maps to detect objects of different sizes and incorporate part-based models to better handle occlusion. Multi-task detection frameworks and specialized loss functions have also been introduced to enhance performance in crowded scenes.

However, widely used datasets such as Caltech-USA and KITTI suffer from low crowd density, with an average of fewer than **1 person per image**. CityPersons partially alleviates this issue, with approximately **6 persons per image**, but still exhibits limited instance overlap. To address these limitations, the **CrowdHuman** dataset was introduced by Shao *et al.*[2] providing an average of **22.6 persons per image** and a significantly higher overlap between instances (IoU > 0.5), enabling more realistic evaluation of detection algorithms in crowded environments.

2.3. Unified Face Detection and Landmark-Based Models

Beyond pure detection, unified models have been proposed to jointly perform face detection, head pose estimation, and facial landmark localization. A notable example is the work of Zhu and Ramanan[3], which models facial landmarks as shared parts within a tree-structured framework, allowing efficient global optimization through dynamic programming.

Despite being trained on relatively small datasets (often only a few hundred annotated faces), such unified models achieve state-of-the-art performance for face detection in unconstrained environments. Their results are competitive with commercial systems, such as Google Picasa and Face.com, demonstrating strong robustness to pose variation and partial occlusion.

.

3. Proposed Approach

This work aims to develop an intelligent framework for comprehensive facial analysis from images by jointly addressing face detection and the extraction of multiple biomet-

[1] K. B. J. Bosco and Y. Chi, "Enhanced Age, Gender, Race Estimation Using Multi-task CNN," *International Journal of Latest Trends in Engineering and Management Sciences*, 2025. DOI: https://doi.org/10.51583/IJLTEMAS.2025.1409000046

[2] S. Shao *et al.*, "CrowdHuman: A Benchmark for Detecting Human in a Crowd," Megvii Inc. Dataset available at: https://sshao0516.github.io/CrowdHuman

[3] X. Zhu and D. Ramanan, "Face Detection, Pose Estimation, and Landmark Localization in the Wild," University of California, Irvine.

ric attributes. To achieve this objective, a two-stage modular architecture based on deep learning is proposed, enabling robust face localization followed by detailed facial attribute analysis.

3.1. Overall Framework

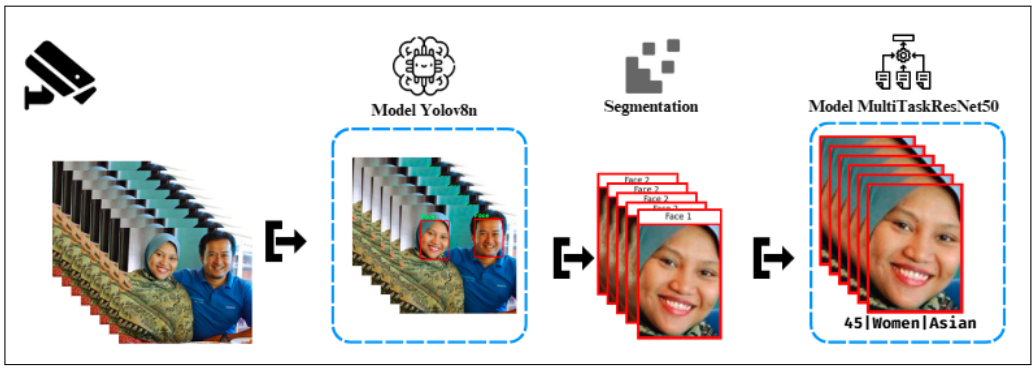

Figure 1. Overall pipeline of the proposed facial analysis framework.

The proposed framework adopts a cascaded processing pipeline composed of two complementary deep learning models: a face detection model dedicated to localizing facial regions in an input image, and a multi-task facial analysis model built upon a *ResNet50* backbone. This separation allows each model to focus on a specific task, ensuring accurate face localization in unconstrained environments and reliable attribute prediction from the extracted facial regions.

3.2. Face Detection Module

The first stage of the pipeline focuses on detecting and localizing faces within the input image. A YOLO-based face detection model is employed and trained on a dedicated dataset comprising approximately **16,700 images** annotated with facial bounding boxes. The annotations are provided in both pixel-coordinate format and normalized YOLO format. The dataset was collected from Google Open Images using the *OIDv4* toolkit and exclusively targets face detection scenarios. This enables the model to robustly detect faces under varying lighting conditions, poses, and backgrounds, while producing accurate bounding box coordinates required for face cropping.

3.2.1. YOLOv8 Architecture and Hyperparameters for Face Detection

The YOLOv8 model used for face detection follows a modern architecture composed of a **backbone**, a **neck**, and a **detection head**. The backbone extracts multi-scale features from the images, the neck fuses these features for better multi-scale detection, and the head predicts bounding box coordinates and face probabilities.

The model keeps **feature maps** at three scales (80Œ80, 40Œ40, 20Œ20), which allows detecting small, medium, and large faces. Unlike traditional architectures, YOLOv8 does not use fully connected layers for detection; instead, 1Œ1 convolutions are applied on each feature map.

This configuration enables the model to detect faces robustly under various lighting conditions, poses, and backgrounds.

Part	Module / Block	Main Function
Backbone	Convolutions + C2f + SPPF	Extract multi-scale features from input images
Neck	Upsample + Concat + C2f	Fuse features at different scales to improve detection accuracy
Head	Detect Layer	Predict bounding boxes and face probabilities
Feature Maps Flatten	Flatten	Convert final feature maps (e.g., 20Œ20Œ256) into a 1D vector (e.g., 102,400)
Fully Connected	Linear layers (e.g., 102,400 512)	Produce embedding vector (512-d) for face recognition or final classification

Table 1. Extended YOLOv8 architecture for face detection and recognition (with embedding)

Hyperparameter	Value
Epochs	50
Image size	640
Batch size	16
Initial learning rate	0.01
IoU threshold	0.7
Optimizer	auto (SGD/Adam)
Pretrained weights	True
Device	GPU

Table 2. Main hyperparameters used for training

3.3. Multi-Task ResNet50 Architecture

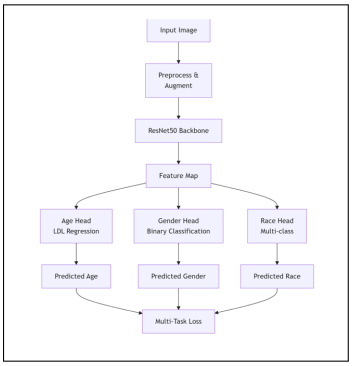

Figure 2. Shows the Details of the model architecture, Multi-Task Deep CNN Model for Age, Gender, and Race Estimation .

The multi-task model predicts age, gender, and race simultaneously from a facial image. It uses a **ResNet50 backbone pre-trained on ImageNet** and task-specific heads. Shared feature representations improve generalization and reduce computational redundancy.

3.3.1. Model Architecture Table

Step	Module / Block	Output / Features	Key Parameters
1	Input	224 Œ 224 Œ 3 image	-
2	Shared Backbone	ResNet50	Feature map 2048, includes Conv layers, BatchNorm, ReLU activations, and Global Average Pooling (GAP)
3	Age Head	Fully Connected (FC)	1 output (regression), Dropout 0.5 FC Dropout 0.3 FC
4	Gender Head	Fully Connected + Sigmoid	1 output (binary classification), Dropout 0.5 FC Dropout 0.3 FC
5	Race Head	Fully Connected + Softmax	5 outputs (multi-class classification), Dropout 0.5 FC Dropout 0.3 FC
6	Loss Functions	Multi-task	MAE (Age) + Binary Cross-Entropy (Gender) + CrossEntropy (Race) with class weights [0.47, 1.05, 1.38, 1.19, 2.80]
7	Optimizer	Adam + StepLR	Initial learning rate 0.0001, weight decay 0.0001 (L2 regularization), learning rate decay factor $\gamma = 0.5$ every 10 epochs

Table 3. Compact architecture of the Multi-task ResNet50 model with adjustable width using `tabularx`

3.3.2. Training Hyperparameters

Hyperparameter	Value / Description
Epochs	50
Batch size	32
Optimizer	Adam
Initial learning rate	0.0001
Scheduler	StepLR, decay $\gamma = 0.5$ every 10 epochs
Weight decay	0.0001 (L2 regularization)
Early stopping	patience=25
Dropout (Heads)	0.5 and 0.3
Loss (Age)	MAE (L1Loss)
Loss (Gender)	BCE
Loss (Race)	CrossEntropy with class weights

Table 4. Training hyperparameters for the Multi-task ResNet50 model

3.4. Advantages of the Proposed Method

The proposed approach offers several key advantages: enhanced robustness through task-specific model specialization, reduced background interference via prior face localization, efficient feature sharing enabled by multi-task learning, and a modular and scalable architecture suitable for extension to additional facial attributes. Overall, the integration of a YOLO-based face detection module with a multi-task facial analysis network pro-

vides an effective and scalable solution for automatic facial analysis in real-world, unconstrained image settings.

4. Datasets

4.1. Face Detection Dataset

This dataset is used for training and evaluating the face detection module of the proposed framework. It was collected from Google Open Images using the *OIDv4* toolkit and is exclusively dedicated to face detection tasks. The dataset contains approximately **16,700 high-quality images**, each annotated with bounding boxes corresponding to facial regions.

4.1.1. Dataset Organization

The dataset is structured following the standard YOLO directory format and is divided into training and validation subsets:

- **Training set**: approximately 13,400 images,
- **Validation set**: 3,347 images.

Each image is associated with a corresponding annotation file stored in text format (`.txt`), sharing the same filename as the image.

Table 5. Dataset split for face detection

Subset	Number of Images	Purpose
Training	13,400	Model learning
Validation	3,347	Model evaluation

4.1.2. Annotation Format (YOLO)

Each annotation file contains one or more lines, where each line corresponds to a detected face in the image. The annotations follow the YOLO format:

$$(c, x, y, w, h)$$

where:

- c denotes the class index (face = 0),
- x and y represent the normalized coordinates of the bounding box center,
- w and h denote the normalized width and height of the bounding box.

All coordinates are normalized with respect to the image width and height, ensuring scale invariance across different image resolutions.

4.1.3. Mathematical Interpretation of Annotation Files

An example annotation file, `00006c07d2b033d1.txt`, contains the following entries:

```
0 0.3943359375 0.325390125 0.139063 0.192188
0 0.6656245    0.315625    0.142188 0.189062
```

Each line mathematically represents a bounding box defined as:

$$B_i = (x_i W,\ y_i H,\ w_i W,\ h_i H)$$

where W and H denote the image width and height, respectively. The presence of two lines indicates that the image contains two distinct faces.

4.1.4. Annotation File Structure

For each image file:

- one corresponding `.txt` file exists,
- each line represents a single face instance,
- multiple lines indicate multiple faces in the same image.

This lightweight text-based annotation structure allows efficient parsing and seamless integration with YOLO-based detection architectures.

4.1.5. Example of Annotated Image

Figure 3. Example image with face bounding boxes annotated in YOLO format.

5. UTKFace Dataset

The UTKFace dataset is a large-scale face dataset containing over **23,000 aligned and cropped images** with annotations for age, gender, and ethnicity. This processed version includes a `labels.csv` file, which simplifies integration into deep learning pipelines.

Training Set

The training set contains **16,593 images** and is used for model learning.

Validation Set

The validation set contains **3,556 images** and is used to tune hyperparameters and evaluate the model during training.

Test Set

The test set contains **3,556 images** and is reserved for the final evaluation of the trained model.

Table 6. Dataset split for UTKFace

Subset	Number of Images	Purpose
Training	16,593	Model learning
Validation	3,556	Model evaluation
Test	3,556	Final evaluation

5.1. Annotation Format

5.1.1. CSV Structure

Annotations are stored in `labels.csv` with the following fields:

- **Age**: integer representing the age of the person,

- **Gender**: 0 for Male, 1 for Female,
- **Ethnicity / Race**:
  - 0: White
  - 1: Black
  - 2: Asian
  - 3: Indian
  - 4: Others

5.1.2. Gender Distribution

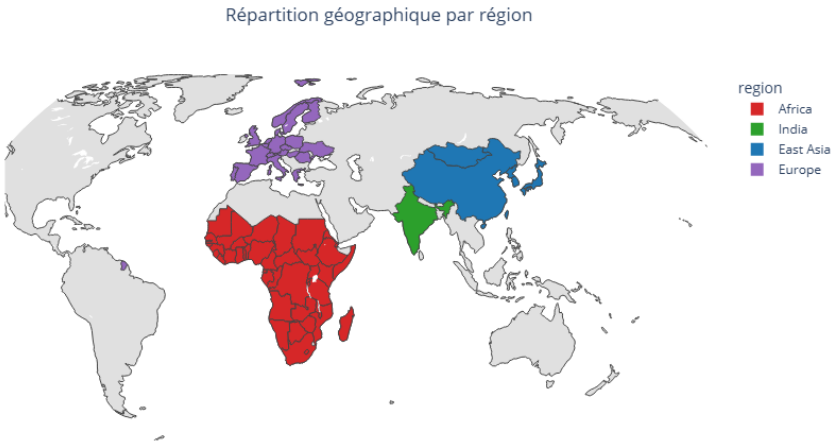

Figure 4. Geographical distribution by region.

5.2. Visual Exploration of the Dataset

5.2.1. Gender Distribution

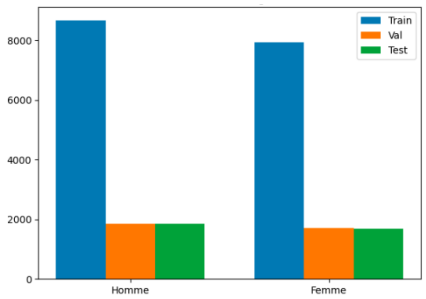

Figure 5. Distribution of images by gender across training, validation, and test sets.

5.2.2. Ethnicity Distribution

The UTKFace dataset provides a rich, annotated collection suitable for age estimation, gender classification, and ethnicity recognition. Its aligned and cropped faces ensure consistency, and the CSV annotations facilitate seamless integration into deep learning workflows. This structure makes the dataset ideal for multi-task learning and demographic analysis.

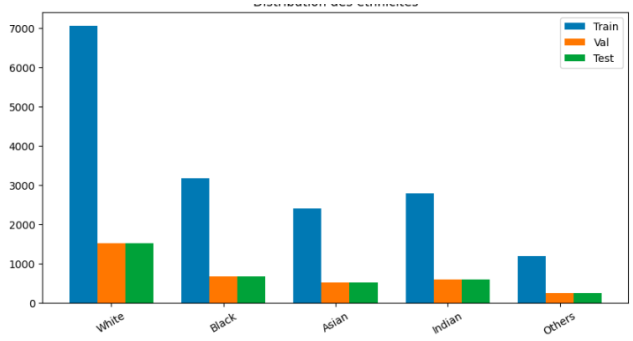

Figure 6. Distribution of images by ethnicity across training, validation, and test sets.

6. Results and Evaluation Metrics

6.1. Face Detection YOLOv8

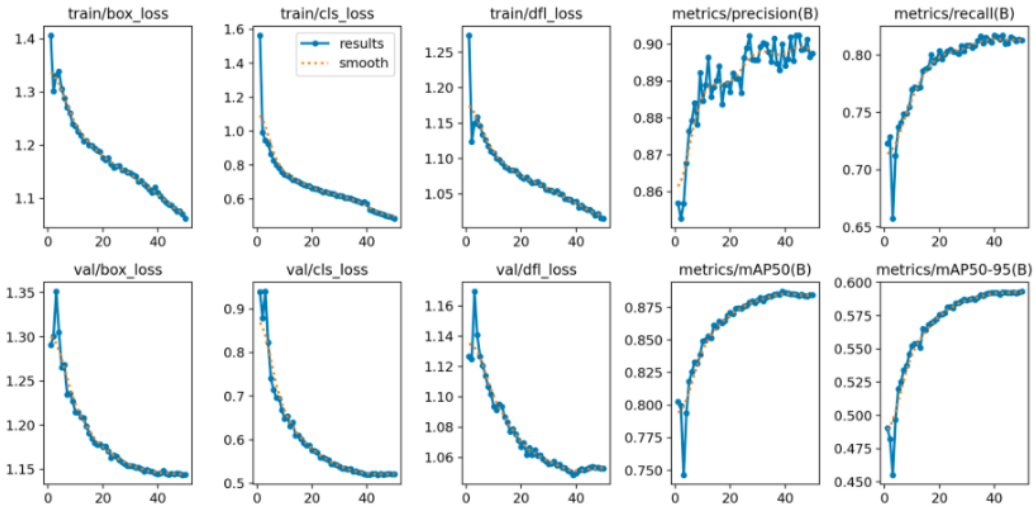

Figure 7. Training and validation curves for YOLOv8 face detection over 40 epochs. The curves show box loss, classification loss, distribution focal loss, precision, recall, and mAP metrics.

6.1.1. Analysis of Training Graphs

Training Losses

The training losses decrease steadily over the 40 epochs. The `train/box_loss` drops from approximately 1.4 to 1.1, indicating continuous improvement in bounding box accuracy. The `train/cls_loss` falls sharply from 1.6 to 0.6, demonstrating rapid learning in classification. The `train/dfl_loss` decreases steadily from 1.25 to 1.05.

Performance Metrics (Precision and Recall)

Precision and recall are defined mathematically as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

where TP = True Positives, FP = False Positives, FN = False Negatives.

Precision (`metrics/precision(B)`) fluctuates between 0.85 and 0.89 with a slight upward trend. Recall (`metrics/recall(B)`) improves significantly, rising from around 0.65 to 0.80, showing better detection of positive objects.

Validation Losses

The validation curves closely follow the training curves, with no signs of overfitting. The `val/box_loss` decreases from 1.35 to 1.15, `val/cls_loss` from 0.9 to 0.6, and `val/dfl_loss` from 1.16 to 1.06.

mAP Metrics (Mean Average Precision)

The mAP metrics are calculated as:

$$\text{AP} = \int_0^1 P(R)\, dR \tag{3}$$

$$\text{mAP} = \frac{1}{N_c} \sum_{i=1}^{N_c} AP_i \tag{4}$$

where $P(R)$ is the precision as a function of recall, and N_c is the number of classes.

The `metrics/mAP50(B)` progresses from 0.75 to 0.875, reflecting strong performance at an IoU threshold of 0.5. The stricter `metrics/mAP50-95(B)` rises from 0.45 to 0.575, indicating room for improvement in bounding box precision at higher overlap thresholds.

Overall Conclusion

The training is successful: losses decrease, metrics increase, and validation curves remain aligned with training curves. The model generalizes well, achieving solid performance (mAP50 0.875). Minor fluctuations in precision and recall suggest potential for further optimization (e.g., learning rate adjustments or additional data augmentations).

6.1.2. Final Training Results

The final performance of the YOLO model for face detection is summarized in Table 7.

Table 7. Performance metrics of the final YOLO model

Metric	Value
Box Loss	1.061
Classification Loss	0.4843
DFL Loss	1.015
Precision (P)	0.898
Recall (R)	0.813
mAP@0.5	**0.884**
mAP@0.5:0.95	0.593

Figure 8. Example of face detection on a test sample

Commentary:

The obtained results demonstrate the high effectiveness of the proposed model, achieving a **mAP@0.5 of 88.4%**, which indicates strong reliability in face detection. As illustrated in Figure 8, the model accurately localizes faces with high confidence scores (up to 0.88), even under partial occlusions such as sunglasses or varying facial expressions. Furthermore, the low *Classification Loss* value (0.4843) confirms the networks ability to correctly distinguish faces from background regions.

6.2. Evaluation of the Multi-Task ResNet50 Model

This section provides a comprehensive evaluation of the multi-task ResNet50 architecture trained to jointly perform age estimation, gender classification, and ethnicity recognition. The analysis combines quantitative metrics with qualitative visual validation to assess the robustness and effectiveness of the proposed model.

6.2.1. Age Prediction Analysis

The age estimation task is evaluated using residual error distribution, correlation analysis, and age-group-based regression metrics.

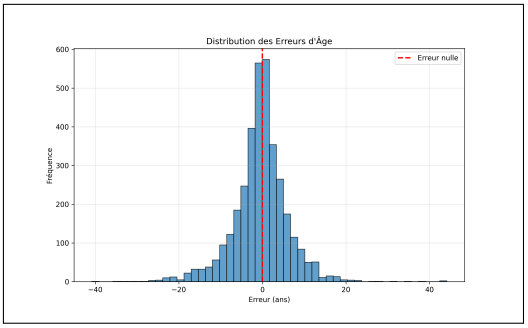

Figure 9. Distribution of age prediction residuals

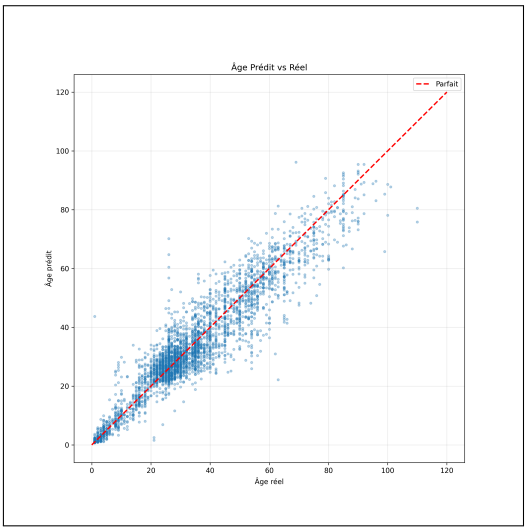

Figure 10. Predicted age vs. actual age correlation

Interpretation:

Figure 9 shows that the residual errors are symmetrically distributed around zero, indicating the absence of systematic bias in the age predictions. The scatter plot in Figure 10 demonstrates a strong linear correlation between predicted and ground-truth ages, confirming the regression capability of the model.

However, prediction accuracy decreases with increasing age. As reported in Table 9(b), the Mean Absolute Error (MAE) is lowest for the 0–10 age group (1.42) and increases progressively, reaching 8.95 for the 81+ group. This behavior reflects the higher facial variability and aging patterns present in older populations.

6.2.2. Gender and Ethnicity Classification Performance

The classification performance for gender and ethnicity is evaluated using confusion matrices and standard classification metrics.

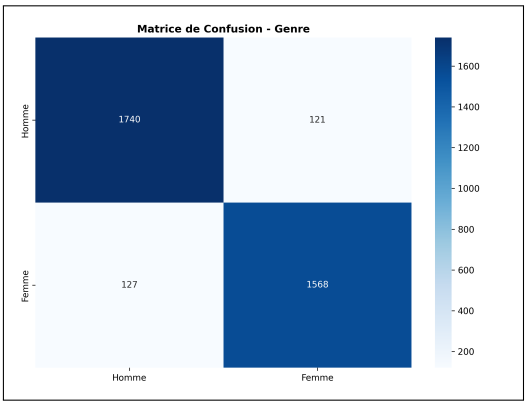

Figure 11. Confusion matrix for gender classification

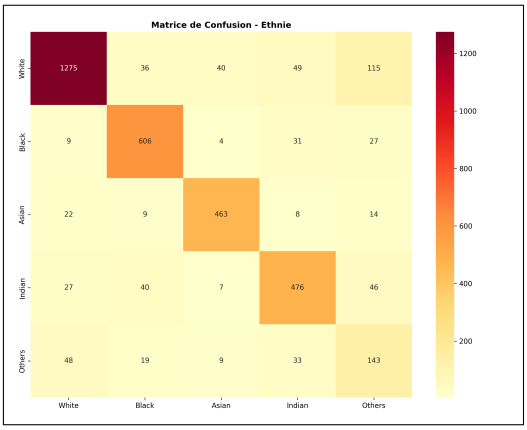

Figure 12. Confusion matrix for ethnicity classification

Interpretation:

Gender classification achieves a high overall accuracy of **93.03%**, with balanced precision and recall for both Male and Female classes (Table 8). This indicates that the shared feature representations learned by the multi-task architecture are highly discriminative for gender prediction.

Ethnicity classification results, summarized in Table 9(a), show strong performance for the *White*, *Black*, and *Asian* classes, with F1-scores above 0.87. In contrast, the *Others* category exhibits a lower F1-score (0.4791), mainly due to its high intra-class diversity and potential class imbalance, which make consistent feature learning more challenging.

6.2.3. Classification Reports – Multi-Task ResNet50

Table 8. Gender classification results of the multi-task ResNet50 model

Metric	Value
Precision (Male)	0.9320
Precision (Female)	0.9284
Recall (Male)	0.9350
Recall (Female)	0.9251
F1-score (Male)	0.9335
F1-score (Female)	0.9267
Accuracy	0.9303
Macro Avg F1-score	0.9301
Weighted Avg F1-score	0.9303
Support (Total)	3556

Table 9. Ethnicity classification and age prediction results (Multi-Task ResNet50)

(a) Ethnicity Classification Results

Metric	Value
Precision (White)	0.9232
Precision (Black)	0.8535
Precision (Asian)	0.8853
Precision (Indian)	0.7973
Precision (Others)	0.4145
Recall (White)	0.8416
Recall (Black)	0.8951
Recall (Asian)	0.8973
Recall (Indian)	0.7987
Recall (Others)	0.5675
F1-score (White)	0.8805
F1-score (Black)	0.8738
F1-score (Asian)	0.8912
F1-score (Indian)	0.7980
F1-score (Others)	0.4791
Accuracy	0.8332
Macro Avg F1-score	0.7845
Weighted Avg F1-score	0.8385
Support (Total)	3556

(b) Age Prediction Results by Age Group

Age Group / Metric	Value
0–10 MAE	1.42
0–10 RMSE	3.44
11–20 MAE	3.83
11–20 RMSE	5.29
21–30 MAE	3.65
21–30 RMSE	5.36
31–40 MAE	4.72
31–40 RMSE	6.02
41–50 MAE	6.51
41–50 RMSE	7.93
51–60 MAE	6.49
51–60 RMSE	8.59
61–70 MAE	6.68
61–70 RMSE	9.30
71–80 MAE	7.44
71–80 RMSE	9.27
81+ MAE	8.95
81+ RMSE	11.51

6.2.4. Qualitative Validation and Final Synthesis

Figure 13. Examples of unified multi-task predictions for age, gender, and ethnicity

Interpretation:

The qualitative results in Figure 13 confirm that the ResNet50 backbone effectively learns shared representations suitable for all three tasks. The model maintains high confidence predictions across different facial poses, illumination conditions, and demographic profiles. The proposed multi-task framework offers a computationally efficient solution while preserving strong predictive performance across regression and classification tasks.

7. Conclusion and Future Work

In this project, we developed a comprehensive framework for automatic facial analysis by integrating a YOLO-based face detection module with a multi-task ResNet50 model for age, gender, and ethnicity estimation. The system demonstrates high effectiveness in unconstrained environments, achieving strong detection precision and reliable attribute prediction across diverse demographic groups.

The results show that combining precise face localization with multi-task learning significantly improves both computational efficiency and predictive accuracy. Quantitative evaluations (mAP, MAE, F1-scores) and qualitative visualizations confirm the robustness of the proposed approach, even under challenging conditions such as partial occlusions, varying lighting, and different facial poses.

Future Research Directions:

In the context of this project, future work could explore the development of a **unified framework capable of combining multiple modalities**, including image description and temporal analysis for video streams. Specifically:

- Implementing architectures that generate **textual descriptions of images** using RNNs or LSTMs to enhance semantic understanding of facial scenes.
- Extending the framework for **real-time video processing**, enabling continuous face recognition and attribute analysis.
- Applying the system for **security and crowd management**, including monitoring of public spaces, detection of suspicious activities, and management of high-density events.
- Investigating the integration of **multi-task learning with spatio-temporal models** to handle both individual facial analysis and crowd-level dynamics in real-world scenarios.

Overall, this project lays the foundation for a scalable and modular facial analysis framework, with the potential to expand toward intelligent surveillance systems and advanced crowd monitoring applications.

1. UTKFace Dataset - Face Aligned and Labeled. Kaggle. Available at: https://www.kaggle.com/datasets/alifshahariar/utkface-dataset-face-aligned-and-labeled.
2. Face Detection Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/fareselmenshawii/face-detection-dataset.
3. Open Images Dataset V4 (OIDv4) Toolkit for dataset scraping. Available at: https://github.com/thepanacealab/OIDv4_ToolKit.
4. YOLOv8 Documentation, Ultralytics. Available at: https://docs.ultralytics.com/.
5. Jocher, G., et al. "YOLOv8: You Only Look Once Real-Time Object Detection." Ultralytics, 2023. Available at: https://arxiv.org/abs/2305.16250.
6. He, K., Zhang, X., Ren, S., Sun, J. "Deep Residual Learning for Image Recognition." In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. DOI: https://doi.org/10.1109/CVPR.2016.90.
7. ResNet Documentation. PyTorch. Available at: https://pytorch.org/vision/stable/models.html#resnet.
8. Butera, K. B. J., Bosco, Y. Chi. "Enhanced Age, Gender, Race Estimation Using Multi-task CNN." *International Journal of Latest Trends in Engineering and Management Sciences*, 2025. DOI: https://doi.org/10.51583/IJLTEMAS.2025.1409000046.
9. Zhu, X., Ramanan, D. "Face Detection, Pose Estimation, and Landmark Localization in the Wild." University of California, Irvine. Available at: https://www.ics.uci.edu/~xzhu/face/.
10. Shao, S., et al. "CrowdHuman: A Benchmark for Detecting Human in a Crowd." Megvii Inc. Dataset. Available at: https://sshao0516.github.io/CrowdHuman.
11. Goodfellow, I., Bengio, Y., Courville, A. *Deep Learning*. MIT Press, 2016. Available at: https://www.deeplearningbook.org/.