Fine-Tuning YOLOv8s for Unified Human and Face Detection in Crowded Environments

Oussama Lachihab
Department of Computer Science, Laboratory of Computer Science and Smart Systems, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco
o.lachihab.ced@uca.ac.ma (corresponding author)

Ahmed El Kiram
Department of Computer Science, Laboratory of Computer Science and Smart Systems, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco
kiram@uca.ac.ma

Latifa Errajy
Department of Computer Science, Laboratory of Computer Science and Smart Systems, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco
l.errajy@uca.ac.ma

Received: 4 October 2025 | Revised: 2 November 2025 | Accepted: 12 November 2025

 DOI: https://doi.org/10.48084/etasr.15307

ABSTRACT

Accurate detection of human bodies and faces in densely populated scenes remains challenging due to occlusions and overlapping instances. This paper presents a lightweight object detection solution built upon the You Only Look Once version 8 small (YOLOv8s) architecture, fine-tuned for challenging urban scenes where occlusion, density, and limited computing resources are common. Leveraging an enhanced dataset with detailed person and face annotations, our model achieves a good mean Average Precision at IoU threshold 0.5 (mAP@0.5) of 57.61%, with particularly robust performance on full-body human detection (Average Precision (AP) = 73.5%). Despite moderate face detection accuracy (AP = 42.1%), qualitative results demonstrate solid performance under real-world constraints. The model's compact size and high inference speed make it ideally suited for deployment on edge devices, such as mobile cameras and embedded Artificial Intelligence (AI) systems. A compelling use case is explored through the lens of crowd monitoring in Jamaa El-Fna square in Marrakech, a bustling and high-density public space that demands real-time situational awareness. This work offers a practical tool for urban analytics and public safety, and it lays the foundation for future improvements in face detection, post-processing, and real-time system integration.

Keywords-YOLOv8s; object detection; human detection; face detection; edge AI; real-time inference; crowd analysis; urban monitoring

I. INTRODUCTION

Computer vision, a rapidly evolving field within Artificial Intelligence (AI), aims to enable machines to see and interpret the world as humans do. Among its numerous applications, human detection [1] holds a particularly crucial role. The ability of machines to accurately identify and locate human beings within images and videos underpins a wide array of technologies that are increasingly integral to modern life.

Real-time human detection techniques have focused separately on either body detection [2] or face detection [3], depending on the application context. However, recent advancements in AI deployment across real-world environments reveal significant limitations in approaches that treat faces and bodies independently. To address the complexity of human-centered understanding, there is a growing need for systems capable of detecting both faces and bodies simultaneously [4].

In crowded [5], dynamic [6], and occluded scenes [7], the challenges associated with partial visibility, varying poses, diverse appearances, and environmental conditions require models that can reason holistically about human presence.

Research has shown that human perception itself relies on the integration of both facial and bodily cues to form accurate first impressions, highlighting the limitations of face-only or body-only approaches [8]. Face detection alone is often insufficient when faces are occluded or turned away, whereas

body detection alone may lack the specificity needed for identity verification, abnormal behavior analysis [9], or fine-grained interaction modeling. Running two separate detection models in parallel on the same image or video stream results in doubled inference time, increased memory consumption, and higher energy costs, especially problematic for real-time or resource-constrained applications like robotics, autonomous driving, or surveillance. Moreover, integration challenges emerge when combining the outputs: overlapping detections, conflicting bounding boxes, and inconsistent confidence scores can arise, complicating downstream tasks such as tracking and re-identification. A joint detection framework that considers both face and body cues offers improved robustness and richer semantic information, enabling more reliable and interpretable systems [10, 11].

The integration of face and body detection supports a range of critical applications. In public safety and surveillance, it allows for more accurate monitoring of individuals in complex scenes. In consumer technologies such as virtual reality, fitness applications, and online collaboration platforms, it enhances user experience by enabling the tracking of both facial expressions and body postures [12, 13]. In retail and healthcare analytics, the ability to simultaneously detect presence, attention, and emotional states leads to deeper insights and more personalized services. Furthermore, from an ethical standpoint, combining face and body detection can support the development of more privacy-conscious and less biased AI systems, moving away from exclusive reliance on facial recognition. Despite its potential, the simultaneous detection of faces and bodies presents unique challenges as well, including handling occlusions, optimizing real-time performance, and balancing accuracy across diverse demographics and environments. As a result, there is a critical need for advanced models and datasets specifically designed for joint face and body detection tasks. To address these limitations, we propose fine-tuning You Only Look Once version 8 small (YOLOv8s) for the simultaneous detection of faces and bodies. We leverage BFJDet annotations (an enhanced version of the CrowdHuman dataset with body and face labels) to train the model effectively for holistic human detection. Our fine-tuning methodology involved initializing from pretrained YOLOv8s weights, adapting the anchor and label structures for dual-class detection (face and body), and applying progressive learning rates and data augmentation techniques to maximize generalization while maintaining high inference speeds. Our main contributions are as follows:

- We fine-tune YOLOv8s to jointly detect human faces and bodies using a unified lightweight architecture.
- We demonstrate that our model achieves promising performance, with a mean Average Precision at IoU threshold 0.5 (mAP@0.5) of 0.5761 and mean Average Precision averaged over multiple IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05 (mAP@0.5:0.95) of 0.3638 on the validation set
- We validate the model's robustness under occlusion and scale variation scenarios.
- We present a real-time capable solution that significantly reduces computational overhead compared to separate detection pipelines.
- We highlight potential applications in surveillance, smart retail, robotics, and human-robot interaction where unified face-body detection is critical.

II. RELATED WORK

A. Object Detection

Object detection, a core task in computer vision, involves identifying and localizing objects within images or videos. Over the years, researchers have proposed a variety of frameworks to address the challenges of accuracy, speed, and scalability in object detection. These frameworks form the backbone of many modern applications such as autonomous driving, surveillance, robotics, and augmented reality [14]. Early object detection approaches, such as the Sliding Window + Classifier method (e.g., using Support Vector Machines (SVMs) with handcrafted features like Histogram of Oriented Gradients (HOG)), scanned images exhaustively at multiple scales. However, these approaches were computationally intensive and lacked real-time capability. Authors in [15] revolutionized computer vision by winning the prestigious ImageNet competition with AlexNet, a deep Convolutional Neural Network (CNN) that leveraged the Rectified Linear Unit (ReLU) activation function to address the gradient vanishing problem. Building upon this momentum, new architectures such as the Visual Geometry Group (VGG) networks [16], and Residual Networks (ResNet) [17] further advanced deep neural network design.

The emergence of deep learning revolutionized the field, giving rise to two main categories of object detection frameworks: two-stage detectors and one-stage detectors. Two-stage detectors, such as R-CNN [18], Fast R-CNN [19], and Faster R-CNN [20], pioneered the use of region proposals followed by classification. While effective, these methods often struggled with real-time performance requirements. Single-stage detectors emerged as more efficient alternatives, with You Only Look Once (YOLO) representing a paradigm shift in balancing speed and accuracy. The YOLO family has seen continuous improvements through successive iterations: YOLOv1 introduced the concept of predicting bounding boxes and class probabilities directly from full images in a single evaluation [21]. YOLOv2/YOLO9000 added batch normalization, anchor boxes, and multi-scale training [22]. YOLOv3 incorporated feature pyramid networks and residual connections [23]. YOLOv4 enhanced performance through architectural modifications and advanced training techniques [24]. YOLOv5 improved scalability with various model sizes, ranging from nano to extra-large [25]. YOLOv6 and YOLOv7 continued optimizing the architecture for edge devices [26, 27]. YOLOv8, which forms the basis of our work, introduced significant architectural changes including a Cross Stage Partial Fusion (C2f) module, enhanced loss functions, and improved anchor-free detection [28]. YOLOv9 brought Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN) to enhance training and feature aggregation [29]. YOLOv10 eliminated Non-Maximum

Suppression (NMS) for real-time efficiency and added design innovations like Partial Self-Attention (PSA) and large-kernel convolutions [30]. YOLOv11 introduced a transformer-based backbone, dynamic heads, and dual label assignment for better accuracy, especially in crowded scenes [31]. YOLOv12 focused on attention mechanisms like Area Attention (A2) and FlashAttention, offering high accuracy with low computational cost [32].

Other notable single-stage detectors include Single Shot MultiBox Detector (SSD) [33] and RetinaNet [34]. Our work builds upon YOLOv8s specifically due to its excellent balance between computational efficiency and detection accuracy.

B. Face Detection

The face detection process involves identifying and localizing human faces in digital images or video streams. It serves as a critical precursor to a wide range of downstream applications, including facial recognition, emotion analysis, and surveillance. Over the years, face detection has evolved dramatically, progressing from rule-based and machine learning methods to highly accurate deep learning-based solutions.

The earliest face detection systems relied on handcrafted features and statistical classifiers. A seminal work by authors in [35] introduced a real-time face detection framework using Haar-like features and an AdaBoost classifier within a cascade architecture.

Subsequent advancements introduced machine learning algorithms that utilized richer feature descriptors such as (HOG), Local Binary Patterns (LBP), and SVMs. These models offered improved accuracy and generalization, especially when combined with multi-scale sliding window strategies. Nevertheless, their reliance on manually engineered features limited their adaptability to highly variable face appearances in real-world scenarios.

The advent of deep CNNs has significantly advanced the field of face detection. One notable model, Multi-task Cascaded Convolutional Networks (MTCNN), employs a three-stage CNN cascade to iteratively refine face candidate regions and predict facial landmarks. MTCNN is particularly recognized for its high detection accuracy, especially in challenging scenarios involving rotated or partially occluded faces. However, its multi-stage architecture introduces computational overhead, making it less suitable for real-time applications where low latency is essential [36]. Consequently, MTCNN is more appropriate for tasks requiring precise facial feature localization, such as face alignment or emotion recognition.

On the other hand, SSD-based face detectors strike a compelling balance between detection speed and accuracy. By leveraging convolutional filters across multiple feature maps at different scales, SSD models effectively detect faces of varying sizes with relatively low computational cost [37].

In contrast, YOLO-based face detection approaches are optimized for real-time performance. By processing the entire image in a single forward pass, YOLO achieves exceptionally fast inference while maintaining competitive accuracy, making it particularly well-suited for dynamic environments and resource-constrained platforms [3, 38].

To assess the performance of different face detection algorithms, we present a comparative analysis of four widely used models: Viola–Jones, MTCNN, SSD, and YOLO. The comparison focuses on key criteria such as detection speed, accuracy, the ability to detect facial landmarks, and real-time processing capability, as shown in Table I.

TABLE I. COMPARATIVE ANALYSIS OF FACE DETECTION ALGORITHMS: VIOLA–JONES, MTCNN, SSD, AND YOLO, FOCUSING ON PERFORMANCE ASPECTS

Algorithm	Speed	Accuracy	Landmark detection	Real-time capability
Viola–Jones [35]	Moderate	Basic	No	Limited
MTCNN [36]	Low–medium	High	Yes	Moderate
SSD [33]	High	Moderate–high	No	Good
YOLO [28]	Very high	High	No	Excellent

YOLO stands out for its high speed and real-time performance, making it ideal for resource-constrained or dynamic environments. In contrast, MTCNN offers superior accuracy and facial landmark detection but at the cost of processing speed. This trade-off highlights the importance of selecting the appropriate algorithm based on specific application needs.

C. Body Detection

Human body detection has evolved through several methodological approaches. HOG + SVM classifiers were among the earliest successful techniques, using gradient-based features for pedestrian detection. Deformable Part Models (DPMs) improved robustness by modeling articulated body structures as compositions of parts, though they struggled with occlusion and real-time performance. These approaches laid the groundwork for human detection but were limited in handling complex real-world scenarios.

The advent of deep learning revolutionized detection pipelines, Faster R-CNN [20] and its variants integrated Region Proposal Networks (RPNs) with convolutional feature extractors, significantly improving localization accuracy. However, their computational cost remained high for real-time applications.

Subsequent architectures like YOLO and SSD prioritized speed, making them suitable for video analysis, but sometimes at the expense of precision in crowded scenes [39-41]. Other approaches that are not strictly detection frameworks, such as human pose estimation and person segmentation, have also contributed to human body understanding. For example, OpenPose [42] pioneered real-time multi-person 2D pose estimation using part affinity fields, enabling robust tracking in crowded environments, HRNet [43] preserved high-resolution feature representations throughout the network, improving keypoint localization accuracy.

Mask R-CNN [44] and YOLO-Pose [45] leverage multi-scale feature fusion to detect partially visible humans, which is

critical for crowded scenes. Hybrid models, such as AlphaPose [46], first detect persons and then estimate poses, demonstrating how detection pipelines can guide finer-grained tasks.

D. Multitask Detection

Simultaneous detection of multiple object types, particularly faces and bodies, represents a significant challenge in computer vision that has received increasing attention. This approach offers several advantages, including reduced computational overhead, simplified system architecture, and the potential for improved performance through shared feature representations.

Earlier multitask detection systems often employed cascaded approaches, where one detection task would feed into another. JointDet [47] explored joint detection of pedestrians and faces in surveillance contexts by first detecting pedestrians and then searching for faces within those regions, achieving improved efficiency over separate detectors. BFJDet [11], which our work builds upon, introduced a comprehensive benchmark for body–face detection with a novel dataset containing detailed annotations for both faces and bodies across various scenarios. It proposed a unified detection framework that explicitly modeled the relationship between co-occurring faces and bodies. PairDETR presented a novel Transformer-based approach for joint detection and association of human bodies and faces. Unlike traditional methods that perform detection first and association later, PairDETR directly predicts body–face pairs as a unified entity.

III. METHODOLOGY

In our work, we explore multitask learning for joint object detection by fine-tuning a lightweight YOLOv8s architecture using the BFJDet-annotated CrowdHuman dataset to simultaneously detect faces and full human bodies within crowded environments. By leveraging a shared representation across semantically related detection tasks, our model achieves a balance between accuracy and speed, making it suitable for real-time deployment scenarios such as surveillance and human–robot interaction.

As shown in Figure 1, we conducted inference on a sample image with dimensions 480 × 640 pixels. A total of 36 faces and persons were identified, with bounding boxes enclosing full human bodies, with 28.6 ms preprocessing time and 1007.0 ms inference time, demonstrating the real-time viability of our model.

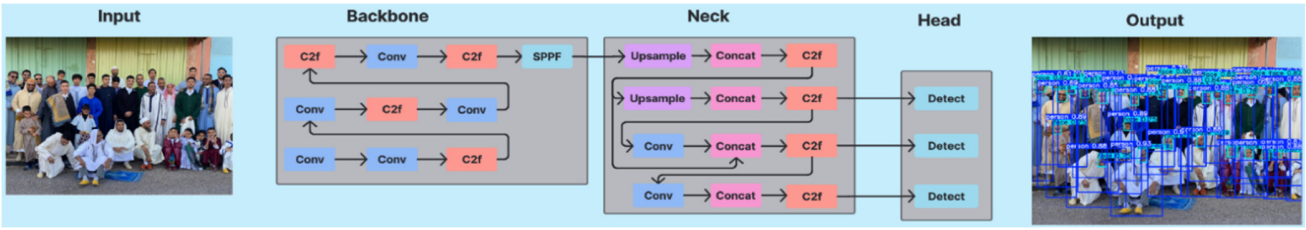

Fig. 1. Process of joint face and body detection using the YOLOv8s architecture.

A. Preprocessing and Training Settings

1) Dataset Source

We employed the CrowdHuman dataset [48] augmented by the BFJDet project [11], which provides additional annotations for full human bodies and faces.

2) Cleaning and Preprocessing

The crowdHuman dataset contains 15,000 images, of which over 5,000 corrupted or incomplete image files (broken JPGs or missing labels) were automatically detected and removed. Custom scripts were written to validate annotation consistency, class integrity, and coordinate format. Bounding boxes with zero area or misaligned coordinates were filtered out. The cleaned dataset was split into training (≈8,000 images), validation (≈1,000 images), and test (≈1,000 images) subsets, with test images remaining unseen during training.

Only two classes were retained: person and face, simplifying the task into a binary multi-object detection scenario.

3) Annotation Format Conversion

Original Common Objects in Context (COCO)–style annotations were converted to YOLOv8 format, ensuring compatibility with Ultralytics' pipeline. Each object label in the dataset was normalized to the format (class_id, x_center, y_center, width, height), which aligns with the YOLO annotation convention.

4) Training Configuration

Training was conducted on an NVIDIA Tesla T4 GPU (16 GB VRAM) using the Kaggle cloud platform. The model was trained for 100 epochs with an input resolution of 640 × 640 pixels and a batch size of 16. All training, validation, and logging were managed using the Ultralytics YOLOv8 training interface, and the best model checkpoint (based on the highest mAP@0.5) was saved for subsequent inference and evaluation.

All training and evaluation scripts, along with configuration files, will be available in our public Kaggle repository: https://www.kaggle.com/code/oussamalach/yolo-facebody1training.

B. Model Architecture: YOLOv8s

The YOLOv8 architecture consists of three main components: a backbone, a neck, and detection head (Figure 2).

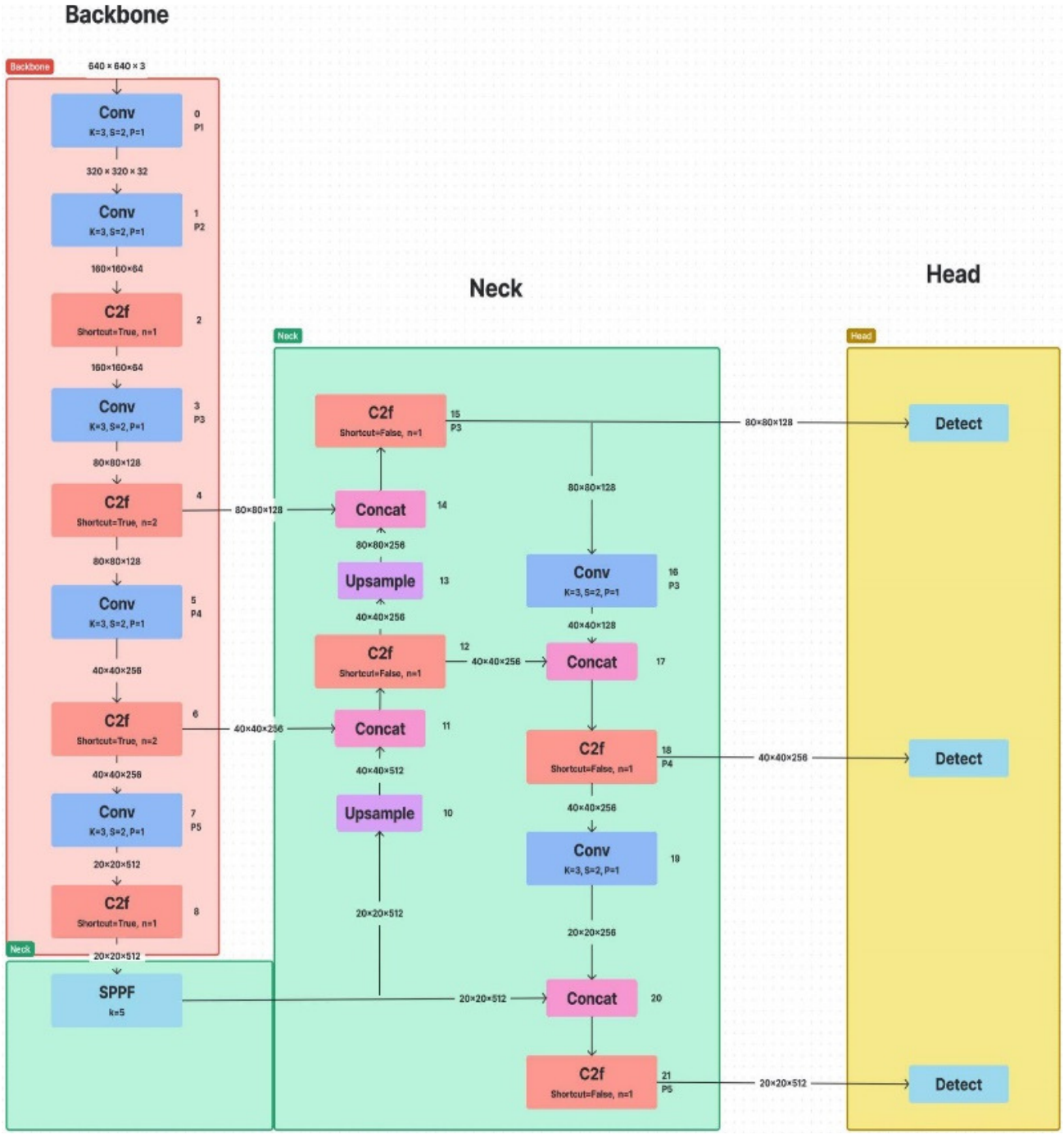

Fig. 2. YOLOv8s architecture.

The backbone performs initial feature extraction through convolutional layers and C2f blocks that enhance feature reuse via residual connections. YOLOv8s applies a depth multiplier (d = 0.33) and width multiplier (w = 0.50) to balance accuracy and speed, resulting in a lightweight model that is ideal for real-time use.

The neck aggregates multi-scale features from stages P3–P5 using upsampling, concatenation, and convolution operations, whereas the head produces predictions at three scales for small, medium, and large objects. A final Spatial Pyramid Pooling Fast (SPPF) layer increases the receptive field without extra computational cost.

Compared to larger YOLOv8 variants (m, l, x), the YOLOv8s version maintains high detection accuracy with reduced model size and real-time inference capability, making it well-suited for edge deployment in human and face detection tasks.

IV. RESULTS AND EVALUATION

A. Evaluation Protocol

To evaluate the performance of the proposed model, we adopted the COCO evaluation protocol, which includes standard object detection metrics such as precision, recall, mAP@0.5, and mAP@0.5:0.95. These metrics are well-established for measuring both classification accuracy and localization precision in object detection tasks.

B. Model Performance

The YOLOv8s model was fine-tuned on the BFJDet-enhanced CrowdHuman dataset for simultaneous detection of human bodies and faces in crowded environments. After 100 epochs of training, the model achieved a precision of 0.8609, recall of 0.4690, mAP@0.5 of 0.5761, and mAP@0.5:0.95 of 0.3638 on the validation set. The model processes images at an average inference time of 13.1 ms per image, corresponding to approximately 76 FPS, confirming its real-time capability. The model size is 21.9 MB, maintaining a lightweight profile suitable for deployment on edge devices.

The precision value suggests the model is effective at minimizing false positives, which is crucial in dense scenes such as shopping centers or public gatherings. However, the recall score indicates that while many true objects are detected, some remain missed. This is a known limitation of lightweight models like YOLOv8s when dealing with heavily occluded or small targets.

In contrast to prior multi-branch or transformer-based approaches such as PairDETR and BFJDet, our method employs a unified YOLOv8s detection head trained jointly on full-body and face annotations. This design simplifies deployment by avoiding specialized subnetworks for each target type, preserving the model's lightweight structure and enabling real-time inference.

C. Interpretation of Results

The results demonstrate that the fine-tuned YOLOv8s model effectively balances accuracy and efficiency for simultaneous human and face detection in crowded scenes. The high precision (0.8609) indicates that the model produces few false positives, which is crucial in dense environments. The moderate recall (0.4690) suggests that while most objects are detected, some occluded or small targets may still be missed, which is a common limitation for lightweight detection models.

The mAP (mAP@0.5 = 0.5761; mAP@0.5:0.95 = 0.3638) reflects the model's performance across multiple scales, demonstrating reliable detection of both full-body and facial features. Combined with an inference speed of 13.1 ms per image (≈76 FPS) and a model size of 21.9 MB, YOLOv8s proves to be a lightweight, real-time solution suitable for deployment on edge devices.

D. Visual Results and Qualitative Analysis

Figures 3 and 4 illustrate the precision–recall curve and qualitative detection results. As shown, the precision–recall curves indicate that the model performs significantly better on the person class (Average Precision (AP) = 73.5%) than on the face class (AP = 42.1%). The person curve remains close to the top-right corner, suggesting high precision and recall, even at lower confidence thresholds. In contrast, the face detection curve drops rapidly, indicating challenges in identifying smaller or occluded facial regions. The overall mAP@0.5 of 57.61% demonstrates reasonably solid performance for object detection in crowded environments. These qualitative observations indicate that the model performs well in moderately crowded scenes but may miss highly occluded individuals or very small faces. These limitations suggest that future improvements could include multi-scale training, enhanced data augmentation, or integration with association modules to better handle overlapping targets.

The unified detection approach successfully demonstrates the feasibility of real-time human and face detection in dense environments, balancing computational efficiency with practical accuracy.

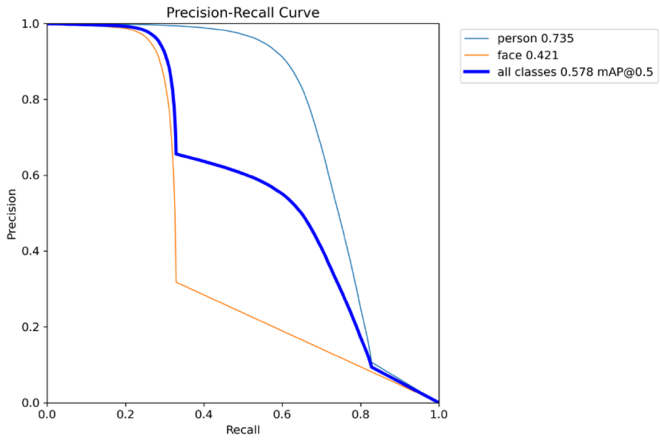

Fig. 3. Precision–recall curve of the fine-tuned YOLOv8s model.

V. USE CASE IMPLICATIONS

As shown in Figure 5, two example scenes illustrate the model's performance before and after inference. The raw frames depict crowded environments with many people, varying occlusions, and different lighting conditions, reflecting real challenges found in public spaces such as shopping centers or metro stations. After inference, the model produces bounding boxes for both humans and faces. The full-body detections are generally accurate and consistent, showing the model can handle dense crowds effectively.

Facial detections are less consistent, particularly for small, partially visible, or occluded faces. This indicates that additional training or refinement may be needed to improve performance for faces in complex scenes.

The unified detection approach reduces the need for separate models for body and face recognition, which simplifies deployment and system maintenance. Its lightweight design also makes it suitable for devices with limited computing power, such as edge devices or mobile platforms. Overall, the observations suggest the model is useful for real-world applications that require awareness of both people and faces, while highlighting areas for further improvement.

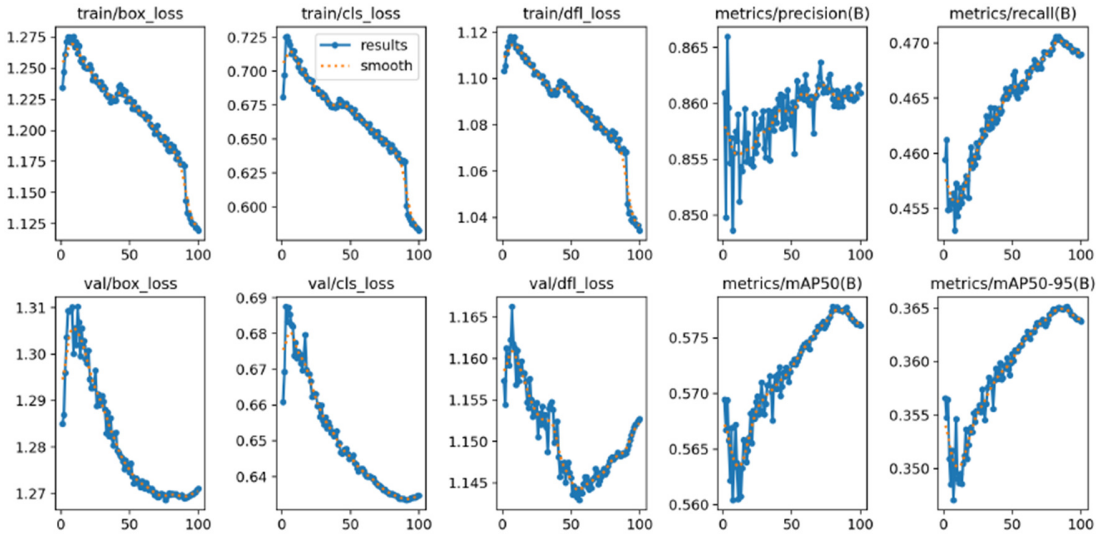

Fig. 4. Qualitative detection results on the validation set.

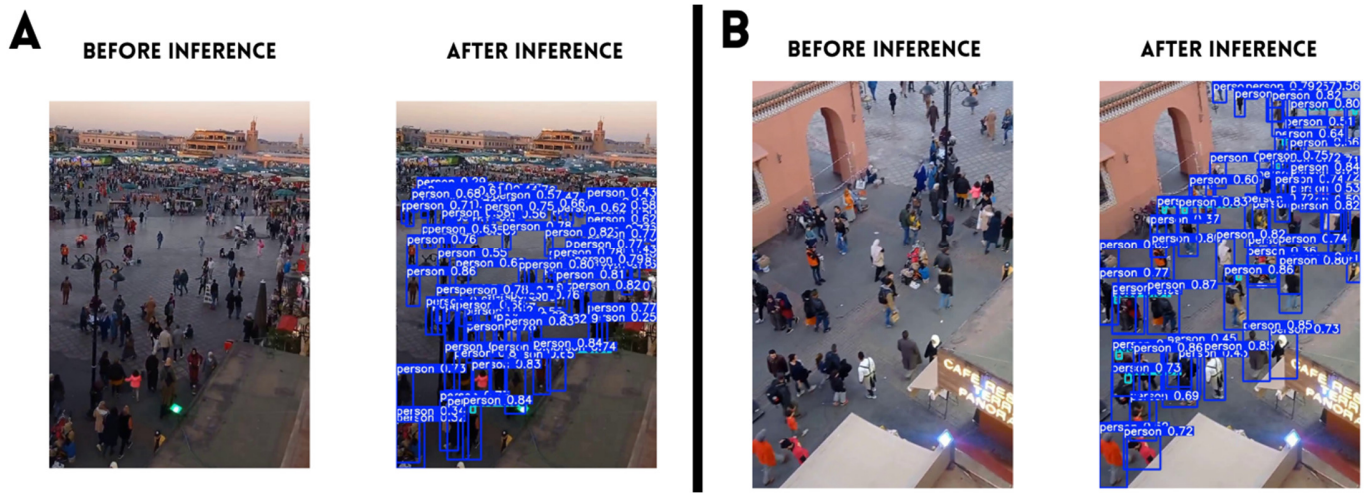

Fig. 5. Before-and-after inference results for two scenes captured from an elevated viewpoint: (a) densely crowded and more distant, and (b) moderately crowded and closer. Bounding boxes highlight detected persons and faces.

VI. CONCLUSION

This study presented a lightweight and efficient object detection system based on the You Only Look Once version 8 small (YOLOv8s) architecture, tailored for simultaneous human and face detection in crowded environments. Through fine-tuning on an enriched dataset with annotated person and face regions, the model achieved a precision of 86.1%, a recall of 46.9%, and a mean Average Precision at IoU threshold 0.5 (mAP@0.5) of 57.61%, indicating strong detection capabilities for full-body human instances. However, the face detection performance (Average Precision (AP) = 42.1%) revealed limitations, particularly in scenarios involving occlusion, small facial regions, or overhead viewpoints. Despite these challenges, the compact YOLOv8s variant demonstrated an excellent trade-off between accuracy and computational efficiency, making it highly suitable for deployment in resource-constrained environments such as embedded systems or mobile surveillance units.

Qualitative and visual analyses further highlighted the model's reliability in detecting people in dense and dynamic scenes, including real-world scenarios such as crowd monitoring in Jamaa El-Fna square in Marrakech. The results underline the model's potential as a foundation for real-time analytics in public safety, smart city infrastructure, and crowd behavior analysis.

Future work will focus on improving face detection accuracy, enhancing post-processing methods, and integrating the system into a complete real-time video processing pipeline. These developments aim to ensure the solution's robustness, scalability, and effectiveness in complex, real-world environments.

REFERENCES

[1] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognition*, vol. 51, pp. 148–175, Mar. 2016, https://doi.org/10.1016/j.patcog.2015.08.027.

[2] H. Mokayed, T. Z. Quan, L. Alkhaled, and V. Sivakumar, "Real-Time Human Detection and Counting System Using Deep Learning Computer Vision Techniques," *Artificial Intelligence and Applications*, vol. 1, no. 4, pp. 205–213, Oct. 2023, https://doi.org/10.47852/bonviewAIA2202391.

[3] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: a real-time face detector," *The Visual Computer*, vol. 37, no. 4, pp. 805–813, Apr. 2021, https://doi.org/10.1007/s00371-020-01831-7.

[4] M.-A. Fiedler, P. Werner, A. Khalifa, and A. Al-Hamadi, "SFPD: Simultaneous Face and Person Detection in Real-Time for Human–Robot Interaction," *Sensors*, vol. 21, no. 17, Sept. 2021, Art. no. 5918, https://doi.org/10.3390/s21175918.

[5] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-End People Detection in Crowded Scenes," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2325–2333, https://doi.org/10.1109/CVPR.2016.255.

[6] L. Stearns and A. Thieme, "Automated Person Detection in Dynamic Scenes to Assist People with Vision Impairments: An Initial Investigation," in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, Galway, Ireland, 2018, pp. 391–394, https://doi.org/10.1145/3234695.3241017.

[7] L. Van Ma, T. T. D. Nguyen, C. Shim, D. Y. Kim, N. Ha, and M. Jeon, "Visual multi-object tracking with re-identification and occlusion handling using labeled random finite sets," *Pattern Recognition*, vol. 156, Dec. 2024, Art. no. 110785, https://doi.org/10.1016/j.patcog.2024.110785.

[8] Y. Hu and A. J. O'Toole, "First impressions: Integrating faces and bodies in personality trait perception," *Cognition*, vol. 231, Feb. 2023, Art. no. 105309, https://doi.org/10.1016/j.cognition.2022.105309.

[9] L. M. Wastupranata, S. G. Kong, and L. Wang, "Deep Learning for Abnormal Human Behavior Detection in Surveillance Videos—A Survey," *Electronics*, vol. 13, no. 13, June 2024, Art. no. 2579, https://doi.org/10.3390/electronics13132579.

[10] A. Ali, G. Gaikov, D. Rybalchenko, A. Chigorin, I. Laptev, and S. Zagoruyko, "PairDETR : Joint Detection and Association of Human Bodies and Faces," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, pp. 423–432, https://doi.org/10.1109/CVPR52733.2024.00048.

[11] J. Wan, J. Deng, X. Qiu, and F. Zhou, "Body-Face Joint Detection via Embedding and Head Hook," in *2021 IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, 2021, pp. 2939–2948, https://doi.org/10.1109/ICCV48922.2021.00295.

[12] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4011–4018, Oct. 2019, https://doi.org/10.1109/LRA.2019.2930434.

[13] T. Zhou, S. Gao, Y. Mei, and L. Wang, "Facial Expressions and Body Postures Emotion Recognition based on Convolutional Attention Network," in *2021 International Conference on Computer, Information and Telecommunication Systems*, Istanbul, Turkey, 2021, pp. 1–5, https://doi.org/10.1109/CITS52676.2021.9618520.

[14] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object Detection With Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019, https://doi.org/10.1109/TNNLS.2018.2876865.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017, https://doi.org/10.1145/3065386.

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015, https://doi.org/10.48550/arXiv.1409.1556.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580–587, https://doi.org/10.1109/CVPR.2014.81.

[19] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1440–1448, https://doi.org/10.1109/ICCV.2015.169.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017, https://doi.org/10.1109/TPAMI.2016.2577031.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 6517–6525, https://doi.org/10.1109/CVPR.2017.690.

[23] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement." arXiv, Apr. 08, 2018, https://doi.org/10.48550/arXiv.1804.02767.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv, Apr. 23, 2020, https://doi.org/10.48550/arXiv.2004.10934.

[25] G. Jocher *et al.*, *ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support.* (2021), Zenodo. https://doi.org/10.5281/zenodo.5563715.

[26] C. Li *et al.*, "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications." arXiv, Sept. 07, 2022, https://doi.org/10.48550/arXiv.2209.02976.

[27] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, 2023, pp. 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721.

[28] G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLOv8*. (2023), Github. Accessed: Jan. 08, 2026. [Online]. Available: https://github.com/ultralytics/ultralytics.

[29] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," in *18th European Conference on Computer Vision*, Milan, Italy, 2024, pp. 1–21, https://doi.org/10.1007/978-3-031-72751-1_1.

[30] A. Wang *et al.*, "YOLOv10: Real-Time End-to-End Object Detection," in *38th Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024, pp. 107984–108011, https://doi.org/10.52202/079017-3429.

[31] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements." arXiv, Oct. 23, 2024, https://doi.org/10.48550/arXiv.2410.17725.

[32] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors." arXiv, Feb. 18, 2025, https://doi.org/10.48550/arXiv.2502.12524.

[33] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *14th European Conference on Computer Vision*, Amsterdam, Netherlands, 2016, pp. 21–37, https://doi.org/10.1007/978-3-319-46448-0_2.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2999–3007, https://doi.org/10.1109/ICCV.2017.324.

[35] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, 2001, p. I–I, https://doi.org/10.1109/CVPR.2001.990517.

[36] N. Zhang, J. Luo, and W. Gao, "Research on Face Detection Technology Based on MTCNN," in *2020 International Conference on Computer Network, Electronic and Automation*, Xi'an, China, 2020, pp. 154–158, https://doi.org/10.1109/ICCNEA50255.2020.00040.

[37] B. Ye, Y. Shi, H. Li, L. Li, and S. Tong, "Face SSD: A Real-time Face Detector based on SSD," in *2021 40th Chinese Control Conference*, Shanghai, China, 2021, pp. 8445–8450, https://doi.org/10.23919/CCC52363.2021.9550294.

[38] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu, and X. Wang, "YOLO-FaceV2: A scale and occlusion aware face detector," *Pattern Recognition*, vol. 155, Nov. 2024, Art. no. 110714, https://doi.org/10.1016/j.patcog.2024.110714.

[39] M. Ş. Gündüz and G. Işık, "A new YOLO-based method for real-time crowd detection from video and performance analysis of YOLO models," *Journal of Real-Time Image Processing*, vol. 20, no. 1, Jan. 2023, Art. no. 5, https://doi.org/10.1007/s11554-023-01276-w.

[40] S. Ennaama, H. Silkan, A. Bentajer, and A. Tahiri, "Enhanced Real-Time Object Detection using YOLOv7 and MobileNetv3," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19181–19187, Feb. 2025, https://doi.org/10.48084/etasr.8777.

[41] H. H. Nguyen, T. N. Ta, N. C. Nguyen, V. T. Bui, H. M. Pham, and D. M. Nguyen, "YOLO Based Real-Time Human Detection for Smart Video Surveillance at the Edge," in *2020 IEEE Eighth International Conference on Communications and Electronics*, Phu Quoc Island, Vietnam, 2021, pp. 439–444, https://doi.org/10.1109/ICCE48956.2021.9352144.

[42] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021, https://doi.org/10.1109/TPAMI.2019.2929257.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 5686–5696, https://doi.org/10.1109/CVPR.2019.00584.

[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 2980–2988, https://doi.org/10.1109/ICCV.2017.322.

[45] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, New Orleans, LA, USA, 2022, pp. 2636–2645, https://doi.org/10.1109/CVPRW56347.2022.00297.

[46] H.-S. Fang *et al.*, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, June 2023, https://doi.org/10.1109/TPAMI.2022.3222784.

[47] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational Learning for Joint Head and Human Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 10647–10654, Apr. 2020, https://doi.org/10.1609/aaai.v34i07.6691.

[48] S. Shao *et al.*, "CrowdHuman: A Benchmark for Detecting Human in a Crowd." arXiv, Apr. 30, 2018, https://doi.org/10.48550/arXiv.1805.00123.