IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 January 2024, accepted 9 February 2024, date of publication 14 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3366451

RESEARCH ARTICLE

Deep Learning Based Multi Pose Human Face Matching System

MUHAMMAD SOHAIL[1], IJAZ ALI SHOUKAT[1], ABD ULLAH KHAN[2,3], (Member, IEEE), HARAM FATIMA[1], MOHSIN RAZA JAFRI[2], MUHAMMAD AZFAR YAQUB[4,5], (Member, IEEE), AND ANTONIO LIOTTA[4], (Senior Member, IEEE)

[1]Department of Computing, Riphah International University, Faisalabad Campus, Faisalabad 38000, Pakistan
[2]Department of Computer Sciences, National University of Sciences and Technology, Balochistan Campus, Quetta 87000, Pakistan
[3]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Suwon, Gyeonggi-do 17104, South Korea
[4]Faculty of Engineering, Free University of Bozen-Bolzano, 39100 Bolzano, Italy
[5]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Islamabad 44000, Pakistan

Corresponding author: Muhammad Azfar Yaqub (myaqub@unibz.it)

This work was supported by the Open Access Publishing Fund of the Free University of Bozen-Bolzano.

ABSTRACT Current techniques for multi-pose human face matching yield suboptimal outcomes because of the intricate nature of pose equalization and face rotation. Deep learning models, such as YOLO-V5, etc., that have been proposed to tackle these complexities, suffer from slow frame matching speeds and therefore exhibit low face recognition accuracy. Concerning this, certain literature investigated multi-pose human face detection systems; however, those studies are of elementary level and do not adequately analyze the utility of those systems. To fill this research gap, we propose a real-time face matching algorithm based on YOLO-V5. Our algorithm utilizes multi-pose human patterns and considers various face orientations, including organizational faces and left, right, top, and bottom alignments, to recognize multiple aspects of people. Using face poses, the algorithm identifies face positions in a dataset of images obtained from mixed pattern live streams, and compares faces with a specific piece of the face that has a relatively similar spectrum for matching with a given dataset. Once a match is found, the algorithm displays the face on Google Colab, collected during the learning phase with the Robo-flow key, and tracks it using the YOLO-V5 face monitor. Alignment variations are broken up into different positions, where each type of face is uniquely learned to have its own study demonstrated. This method offers several benefits for identifying and monitoring humans using their labeling tag as a pattern name, including high face-matching accuracy and minimum speed of owing face-to-pose variations. Furthermore, the algorithm addresses the face rotation issue by introducing a mixture of error functions for execution time, accuracy loss, frame-wise failure, and identity loss, attempting to guide the authenticity of the produced image frame. Experimental results confirm effectiveness of the algorithm in terms of improved accuracy and reduced delay in the face-matching paradigm.

INDEX TERMS Deep learning, face recognition, pattern matching, YOLO-V5.

I. INTRODUCTION

Over the last three decades, facial recognition has garnered significant attention due to its perceived ease of use as an image analysis and pattern recognition application. Two of the most important reasons to understand the trend are: firstly, the diverse range of commercial and legal requirements, and secondly, the ubiquity of relevant technologies such as smartphones, digital cameras, and GPUs [1]. Despite the advancements made in machine learning and recognition systems, their performance remains constrained by real-world conditions. For instance, accurately identifying facial images in unconstrained environments characterized by lighting variations, diverse postures, facial expressions, partial occlusion, disguises, or camera movement continues to pose daunting challenges. In other words, existing technologies are still lagging behind the visual capabilities of the human mind [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

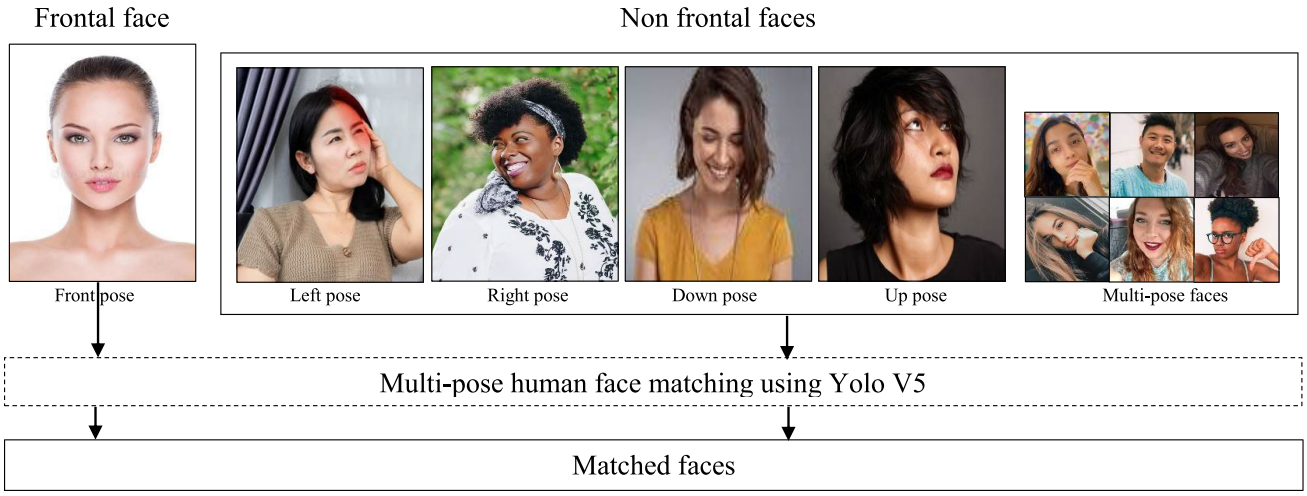

FIGURE 1. Structure of multi-pose human face recognition.

Recently, machine learning approaches have gained significant success in computer vision applications, specifically face recognition [3]. Two major applications of face recognition are face identification and face verification. In face identification, facial images of a person can be utilized for his identity, whereas in face verification, a face image and identity estimation is given to system to verify whether this image belongs to a specific person or not [4].

A major problem associated with face detection is detection accuracy. Different face scale for the same image varies dramatically for detector [5]. Deep learning-based approaches, i.e., Region-based Convolutional Neural Network (RCNN) [6] for face recognition has improved performance significantly as compared to traditional algorithms, i.e., AdBoost and Deep Pyramid Deformable (DPM) [7], [8]. Upcoming machine learning algorithms i.e. Spatial Pyramid Pooling (SPP-net) [9], Fast Region-based Convolutional Neural Network (Fast-RCNN) [10], Faster Region-based Convolutional Neural Network (Faster-RCNN) [11], and Region-based Fully Convolutional Network (R-FCNN) [12] improved their accuracy and speed with passage of time and thus the authors have achieved significant progress in face recognition domain but still there are many challenges need to be addressed, especially face recognition duration. Already available methods discuss feature selection based on duel systems [13] and single platform [14], which are comparable to strict scientific edge image retrieval approaches utilizing analytic infrastructure and local factors-centered approach tools. The limitations associated with [13] and [14] are blurry descriptions and missed facial expressions in foregrounded faces, as well as shading problems in small areas. To improve the limitations associated with already available methods discussed in section II, the authors therein chose one of the advanced algorithms, i.e., You Only Look Once (YOLO), and presented their work. But they lacked significant improvement in speed in multiple human face matching and higher precision.

To overcome the limitations associated with already available methods, a modified version of YOLO-V5 [15] is presented to improve face matching from different poses and reduce face recognition time. For this purpose detecting face benchmarks or recovering the depth appearance of a face from an image is utilized, which is relatively straightforward compared to the inherent difficulty in resolving position fluctuations. The YOLO-V5 method is utilized for multi-pose human face matching, enabling the evaluation or detection of individuals using their face photos captured in various postures. This approach has prompted the development of current rotation estimation techniques such as Insight position, recursive convolution photograph, detailed points, facial expression, and smoke, discussed in [16]. Figure 1 presents a framework for the detection of multi-pose challenge, including classifying the multi-pose face to a (left, right, front, top, bottom) appearance while maintaining identification order. The proposed method uses a multi-pose base YOLO-V5 to train a dataset based on face orientation and alignments between frontal and rotated faces in multiple poses.

The major contributions of this study are summarized below.

1) We proposed YOLO-V5-based simple and robust algorithm for real-time multi-pose human face-matching and recognition. The algorithm learns modeling across multi-pose faces and forehead appearances in input images to identify all types of rotation images, such as front, left, right, and side views.
2) By introducing a mixture of error functions for execution time, accuracy loss, frame-wise failure, and identity loss, the face rotation challenge is handled. Additionally, a fundamental weight matrix is used in the training process to increase accuracy.

Consequently, a higher accuracy of up to 99% and a faster face-matching speed of about 34s is achieved. Furthermore, actual experiments are carried out to show that the YOLO-V5 formalization, matching, and identification created excellent results in real-world circumstances.

A. PROBLEM MOTIVATION

Current multi-pose human face matching techniques face challenges in handling pose equalization and face rotation, resulting in suboptimal outcomes. Deep learning models, such as YOLO-V5, proposed to address these complexities, suffer from slow frame matching speeds, leading to lower face recognition accuracy. Existing literature on multi-pose human face detection lacks comprehensive analysis, leaving a research gap for practical and effective face matching algorithms. To bridge this gap, we present a real-time face matching algorithm based on YOLO-V5. Our method leverages multi-pose human patterns and various face orientations to improve recognition accuracy. By addressing face rotation with error functions, our algorithm offers high accuracy and reduced delay, making it valuable for identifying and monitoring humans in real-world applications.

II. RESEARCH BACKGROUND

The research background is discussed in this section in detail.

A. DEPTH POSE ALIGNMENT OF IMAGE

Depth pose alignment plays a crucial role in face recognition systems, especially when dealing with images captured under different poses. Conventional face recognition algorithms often encounter difficulties when faced with images where the subject's face is not directly facing the camera, leading to compromised performance. The primary objective of depth pose alignment is to address these challenges by transforming the pose of the face in the image into a standardized frontal view. This standardization greatly facilitates the recognition algorithm in accurately matching and identifying faces. A comprehensive analysis of this concept can be found in the literature, as presented below.

The existing computational intelligence learning models are unable to accurately distinguish faces in images with varying perspectives [17]. This is because changes in surface and pattern caused by shifting perspectives often outweigh the differences between individuals, as highlighted in [18] and [19]. The more recently developed face recognition techniques for face recognition can be broadly categorized into two groups [20]. The first group includes single-stage techniques such as edge detection and feature extraction, which represent bottom-up style as deep features. These techniques have been successfully applied to in-depth pose face recognition from images. The second group is multi-pose area approaches, which integrate the internal capabilities of face angles into a shared latent region that allows for the concept of multi-face recognition. Both of these approaches are elaborated in [21] and [22].

The authors in [23] proposed a pose estimation method based on bin classification. The proposed method is designed to accurately estimate head poses using a deep learning approach. They utilized predicted probabilistic labels to regress with a discrete Gaussian distribution, which models the diverse range of true head poses. This Gaussian distribution is used to supervise the deep neural network by employing a maximum mean discrepancy loss. Additionally, the authors also introduced a spatial channel-aware residual attention structure to enhance the intrinsic pose features, further improving the prediction accuracy and speeding up the training convergence process.

Accessible position in face recognition pertains to the system's ability to process face images captured in diverse real-world situations. This encompasses scenarios where the face may be partially obscured or only partly visible. Additionally, accessible position handling includes situations where the face is not directly facing the camera, but rather tilted or captured at an angle. Literature given below discusses the accessible position in detail.

Sign language recognition depends on three main channels of information i.e. hand gesture, body pose, and facial expression. The authors in [24] utilized SMPL-X, a modern parametric model that allows the extraction of 3D body shape, face, and hand information from a single image. By using this comprehensive 3D reconstruction, the authors conducted SLR and found that it resulted in greater accuracy compared to recognizing information from raw RGB images or 2D skeletons. Additionally, the authors highlighted the significance of combining information from all three channels to achieve the best recognition outcomes.

The authors in [25] proposed a method to improve the speed and accuracy of face recognition system. he system utilized a combination of mixed methods and strategies, incorporating deep learning and machine learning techniques. The project consisted of four primary stages. Initially, the Histogram of Oriented Gradients (HOG) was applied to swiftly identify faces in digital images. Following successful face detection, a customized facial landmark estimation process was employed to delineate five distinct facial regions. Subsequently, the segmented face was passed through a pre-trained facial model for recognition purposes. The study's results indicate that face recognition algorithms designed to operate in real-time using modern deep learning techniques can be efficiently deployed on inexpensive computing devices.

B. BLAZE POSE

The authors in [26] proposed a novel technique that can estimate 3D face shapes and animatable details that are unique to an individual but vary with expressions. Their proposed approach, named Detailed Expression Capture and Animation (DECA), is trained to produce a UV displacement map from a low-dimensional latent representation that contains both person-specific and generic expression parameters.

The authors developed a regressor that can predict detail, shape, albedo, expression, pose, and illumination parameters using only one image. The proposed model is trained on images captured in-the-wild, without any paired 3D supervision. Consequently, the model achieved significant improvement in shape reconstruction accuracy over two benchmarks.

In [27], a new network for 3D face reconstruction, called CED-Net, is presented. The network incorporates contextual information at both the shape and feature level. The loss function is constrained by considering the shape context relationship, where the Euclidean norm and vector angle similarity are computed for each contextual vector. To incorporate contextual information at the feature level, the network uses a local feature correlation modulator in its center section. This allows the network to capture the relationship between facial features from a spatial perspective.

A face tracking method for Human Robot Interaction (HRI) is discussed in [28]. The proposed method is presented for face detection using the Viola-Jones algorithm, while face tracking is achieved using the Kanade-Lucas-Tomasi (KLT) algorithm with different pose conditions. The camera motion is controlled based on the displacement between the frames, which is obtained from the tracking result of the previous stage. Real-time experiments show that the proposed system can successfully track human faces even when the subjects are wearing glasses, hats, or in lateral face postures.

The authors in [29] presented a method to overcome the challenges associated with face recognition. In face recognition applications, a major obstacle is the significant differences between profile and frontal faces. Existing techniques address this issue by either synthesizing frontal faces or by learning pose invariance. The authors propose a new approach using Lie algebra theory to investigate how rotating a face in 3D space affects the process of generating deep features with CNNs. The paper demonstrates that face rotation in the image space is equivalent to an additional residual component in the CNN feature space, which is determined solely by the rotation. Based on this finding, the paper proposes a Lie Algebraic Residual Network (LARNet) to address the issue of pose robust face recognition. The LARNet consists of a residual subnet for decoding rotation information from input face images, and a gating subnet to learn the rotation magnitude and control the strength of the residual component involved in the feature learning process.

C. SINGLE AND MULTIPLE MODELS FOR FACE RECOGNITION SYSTEMS

In face recognition systems, two main approaches are commonly used: the single model approach and the multiple models approach.

The single model approach involves training a single model on a dataset containing images of various individuals. While simple and efficient, this approach may struggle to handle diverse face variations and can result in reduced accuracy in challenging scenarios.

On the other hand, the multiple models approach utilizes a collection of specialized models, each trained to excel in recognizing specific subsets of faces or handling particular challenges. This approach enhances the system's performance and robustness, especially in dealing with variations in face appearance, pose, and lighting conditions. However, it introduces increased complexity and computational overhead in managing and maintaining multiple models.

The authors in [30] summarize the CNN-based methods used for face identification selecting face models from a larger population. The authors provide an outline of the latest developments in this field and examine the current state-of-the-art CNN-based face recognition and verification systems.

A method based on YOLO for facial recognition is presented in [31]. The authors therein proposed a system for recognizing facial expressions in a smart classroom setting. To achieve improved results, the authors employed YOLO to extract face images from high-resolution videos of multiple students. After pre-processing the images, a self attention based model called Vision Transformer (ViT) is utilized to recognize facial expressions. The authors then utilize the classified facial expressions to help teachers analyze the learning status of their students and provide suggestions for improving teaching effectiveness.

The authors in [32] proposed an online platform for face recognition. They prove that the proposed platform provides features such as user and criminal information management and real-time facial recognition for identifying criminals through a live stream camera feed. The system is designed for use by two types of users: police employees and administrators who have higher-level access and database maintenance responsibilities. The Haar Cascade algorithm is used and extended for efficient real-time recognition. The website is developed following the MVC pattern and includes a live feed section with video filters to optimize recognition results. The development process involved extensive research on face recognition algorithms and related platforms, requirements definition, persona and scenario development, communication with stakeholders, heuristic evaluation, and feedback collection via a questionnaire. The approach was successful in achieving its goals, as evidenced by the results of the feedback analysis.

To overcome the limitations associated with face recognition by using video surveillance cameras, a dataset is presented in [33], wherein it is shown that, though deep learning models render impressive performance in facial recognition, they perform poorly in surveillance scenarios. It is further shown that the accuracy of face recognition depends not only on the structure of the model but also on the quality and diversity of the training samples. It is demonstrated that the existing multi-pose face datasets do not include complete top-view face samples, which limits the accuracy of the models trained on them.

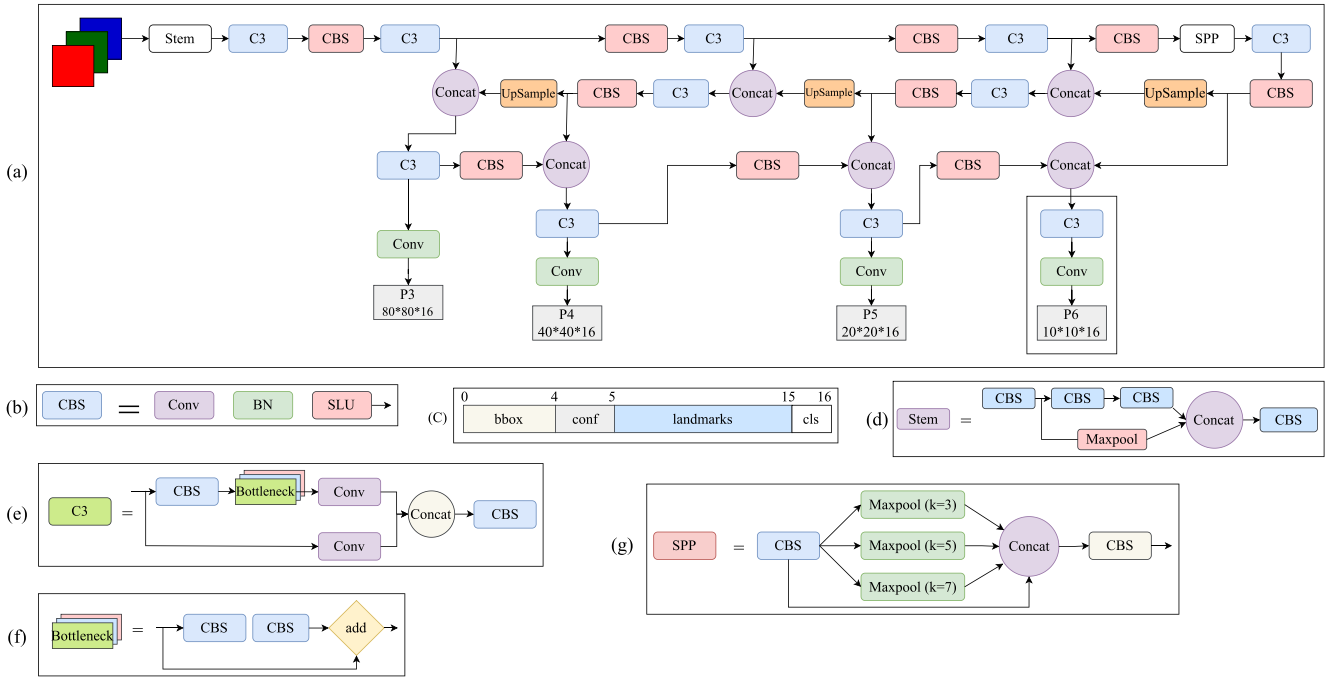

FIGURE 2. Changing the background Tier of the YOLO-V5 model [38].

A YOLO-V4 based scheme for head detection and people counting is presented in [34]. It is shown that counting accuracy is affected by pedestrians blocking each other and occluding their heads, especially in crowded areas.

The authors in [35] discussed a method for public identification. Increasing attention has been given to face recognition due to its importance in public identity verification and security, as well as in information management and digital entertainment. Existing face recognition systems encounter various challenges such as pose variation, illumination variation, and occlusion issues. The authors also propose a face recognition system based on the VGG16 deep learning model to address these problems. To achieve robust pose and view variant face recognition, the system utilizes MultiTask Convolutional Neural Network (MT-CNN) for face detection and VGGNet for face recognition. A real-time database of facial images of 50 subjects was used for evaluation, and cross-validation accuracy was used to assess the system's performance. The proposed method achieved an improved accuracy of 95.80%, 77.50%, and 98.20% under extreme uncontrolled conditions for ORL, FERET, and the real-time face database, respectively.

Similarly, the authors in [36] highlighted the recent progression in 2D face recognition and pointed out that the existing literature had limitations with respect to lighting conditions, poses, and face spoofing. 3D face recognition provides a solution to these limitations. However, constructing a suitable database for 3D face recognition is a major challenge. To overcome this challenge, the authors present a new database called CAS-AIR-3D Face, which contains 24713 videos from 3093 individuals captured by Intel RealSense SR305. This database includes three modalities: color, depth, and near infrared and contains variations in pose, expression, occlusion, and distance. We preprocess the data using a face alignment method, and a Point Cloud Spherical Cropping Method (SCM) is applied to remove background noise in the depth images. We also design an evaluation protocol for fair comparison and perform extensive experiments with different backbone networks to provide different baselines on this database. To our knowledge, CAS-AIR-3D Face is the largest low-quality 3D face database in terms of the number of individuals and the sample variations.

III. PROPOSED YOLO-V5 BASED MODEL

In our proposed model, the architecture of YOLO-V5 is reconfigured by incorporating a CSP darknet slim layer and adding a p6 shell at the neck level for optimal results. During the paired learning phase, the framework's source obtained from Robo-flow can consist of one or more multi-pose expressions. The representation of multiple human faces is achieved by collecting the frontal view and other aligned faces from the provided dataset [37].

A. MODEL STRUCTURE

A detailed detection mechanism based on YOLO-V5 is shown in Figure 2. Architectural application for the proposed model is presented in 2(a). In 2(c), a direction bundle termed CBS is demonstrated, which further emerges as a fully-connected stratum, packet similarity unit, and a fractal dimension storage facility visual stimuli activity. This is used in a variety of contexts. The ending summary for the upstairs,

TABLE 1. Ablations reading output on the custom or real and fake dataset with new changing of this YOLO-V5 model.

Changing	Model	Easy Bounding Box Accuracy	Medium Bounding Box Accuracy	Hard Bounding Box Accuracy	Time at GPU	Time at CPU
Arise hinder	Attraction +Conv	95.56%	97%	92%	30min	34s
Arise hinder	Stem block	96%	97%	92%	30min	34s
Kernel with SPP	(12,8,7)	97%	96%	90%	31min	25s
Add p6	Yes	99%	96%	94%	20min	20s
At augmentation of data	Mosaic	88%	85%	83%	15min	19s
At augmentation of data	Upper/lower flipping	87%	84%	82%	10min	17s
At augmentation of data	Reject small faces/ignore	30%	25%	23%	5min	10s
At augmentation of data	Crop randomly	98%	96%	94%	2min	5s

which includes structure framing, bravery, classifying, etc. is detailed in Figure 2(c). The observations in this investigation lead to the further development of YOLO-V5 to generate a head classifier that provides better results. If there is an apparent lack of different point tags, the terminal version sixteen must be six, and the number of individuals must be scaled accordingly. The stem infrastructure employed to recover the true awareness level in 2(d) is detailed as YOLO-V5. The exploration of the spine surface using YOLO-V5 for feature extraction is an example of this experiment's acknowledgment. Restricted fulfillment concern updates are explored in 2(e). Additionally, instead of integrating clear knowledge and outcomes from trained neural levels, the signal is decomposed into multiple proportions. A midpoint travels through a CBS place, and the number of superior throat sections, subsequent Cns layer, as shown in Figure 2(f). The back half flows below the surrounding matrix; after that, all pair are merged and delivered through a different broadcast region. In 2(g), the 3 technique categories are ready (7, 5, 3) inside this profile identity (13, 9, 5 with YOLO-V5).

B. OVERVIEW OF SIGNIFICANT CHANGES

The latest advancements in the YOLO-V5 framework have significantly enhanced its capabilities. One crucial modification is the inclusion of a preview retrieval portion. However, this cutting-edge feature incurs a significant learning rate penalty. The measurement system used in the blocks has also been improved, making it more impactful and useful in various functions. The locations of objects are now more accurately determined, which improves the identification's robustness. The diacritical marks level has also been transformed, resulting in increased predictive accuracy and reduced workload while maintaining efficiency. Additionally, the overall computer has been revamped, and the SPP domain has been altered, resulting in considerably higher resolution and making the YOLO-V5 more suitable for face identification. Furthermore, a pooling tier 6 trade part with a distance of 48 has been integrated to enable the processing of larger photos. The design includes multiple data amplification strategies for universal feature extraction, such as webcam movement and sketching, which have been found to be ineffective for biometric technology. The platform's efficiency has been boosted by extending the display, and the technology is exceptionally adept at detecting patterns in pitch-black environments. The CSP connectivity in this system is distinct from other varieties, making it a highly suitable computer vision feature for integrated or wearable devices despite its small size. Table 1 provides further details on the advancements and accomplishments in the feature selection design.

C. WORKING OF YOLO-V5

YOLO [39] is a popular object detection algorithm, largely utilized in machine learning and computer vision applications. The complete framework of the YOLO algorithm is explained in 3.

There are three major components of YOLO algorithm explained as:

1) Backbone: The initial stage of the YOLO algorithm is referred to as the backbone, which is primarily accountable for feature extraction from the input image. A CNN is usually employed as the backbone, pre-trained on a massive dataset. Its primary objective is to recognize high-level characteristics in the image, including edges, corners, and textures. These features are subsequently forwarded to the neck, the subsequent component in the process.
2) Neck: The neck constitutes the second phase of the YOLO algorithm, with the primary task of merging the features obtained from the backbone to develop an array of feature maps. It commonly comprises a set of convolutional layers that employ spatial filters to condense the size and intricacy of the feature maps. The ultimate goal of the neck is to create a condensed depiction of the input image that can be handled more efficiently by the subsequent phase, i.e., the head.
3) Head: The last component of the YOLO algorithm is known as the head, which is accountable for identifying the objects in the input image. The head comprises multiple convolutional layers that implement object detection algorithms to the feature maps generated by the neck. In this phase, the head anticipates a sequence of bounding boxes encircling each object present in the image, together with their respective class labels. Moreover, the head executes non-maximum

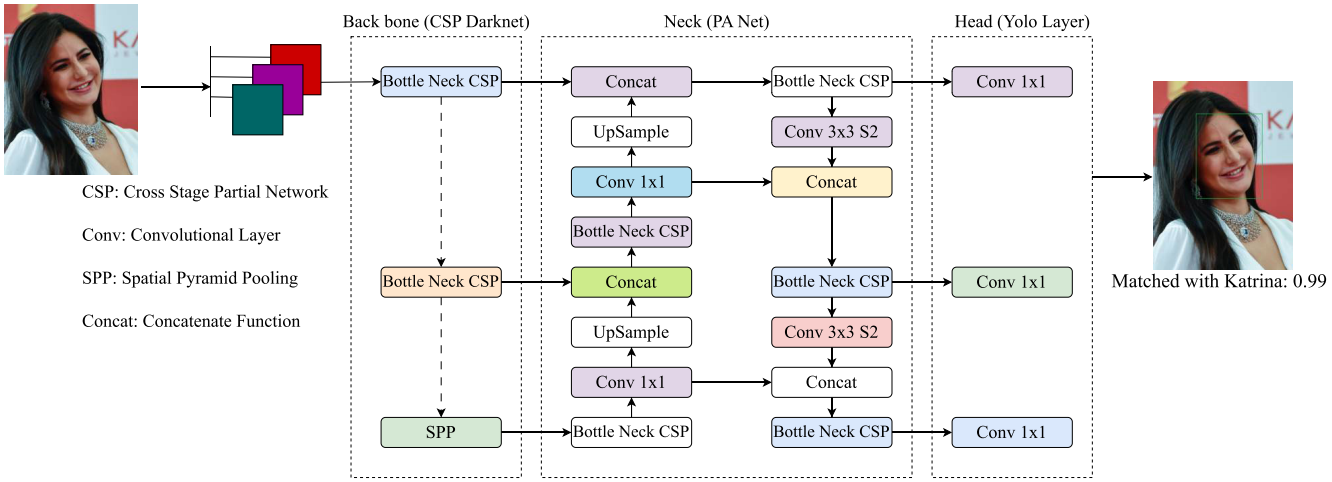

FIGURE 3. Complete framework of YOLO-V5.

suppression (NMS) to get rid of overlapping bounding boxes, enhancing the accuracy of the detection process.

1) MATHEMATICS BEHIND YOLO ALGORITHM

The optimization of the loss function is a crucial aspect of the YOLO algorithm during training. This function is entirely reliant on the sum-squared error, as expressed in [40]:

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(x_i - \hat{x})^2 + (y_i + \hat{y}_i)^2]$$
$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(\sqrt{\omega} - \sqrt{\hat{\omega}} \right)^2 + \left(\sqrt{i} - \sqrt{\hat{i}} \right)^2 \right]$$
$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(C_i - \hat{C}_i)^2]$$
$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{noobj} [(C_i - \hat{C}_i)^2]$$
$$+ \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2$$

The initial part of the equation calculates the loss based on the predicted position of the bounding box and the actual position of the bounding box. $\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(x_i - \hat{x})^2 + (y_i + \hat{y}_i)^2]$ by using (x_{center}, y_{center}) coordinates [39]. 1_i^{obj} represents whether the object exists in a specific cell of the grid made over detection i and 1_{ij}^{obj} represents that j^{th} bounding box predictor is in that specific cell i. In the second part, the loss function $\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(\sqrt{\omega} - \sqrt{\hat{\omega}})^2 + (\sqrt{i} - \sqrt{\hat{i}})^2]$ calculates the error in bounding box prediction.

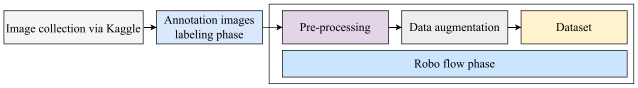

FIGURE 4. Dataset creation flowchart.

Confidence score, which determines the presence of an object within the bounding box, is computed and incorporated into the loss function to account for errors in object detection. 1_{ij}^{obj} will have value of '1' if the object is present in the bounding box and otherwise it will be '0'. The last part of the function $\sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2$ is responsible for the class probability loss. Whenever there is no object, YOLO does not care about the classification error [39]. The loss function is optimized during the training process and is responsible for minimizing the classification error of an object present in a particular grid cell and reducing the error in the coordinates of the bounding box.

IV. EXPERIMENTATION

Experimental work carried out for the proposed model is shown in Figure 8.

A. DATASET CREATION

For the implementation of the proposed model Private and RFFD-based database is utilized and accessed through Kaggle. The initial step is the level creating samples earlier learning, with the 1680 phases as shown in Figure 4.

B. COACHING DATA SOURCE

The subsequent phase entails employing a YOLO-V5 algorithm for training the model, and the weight file obtained from this process is utilized for testing, as depicted in Figure 5.

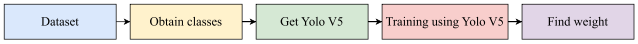

FIGURE 5. Dataset training flowchart.

FIGURE 6. Image classes are described as real and fake faces.

A distinctive YOLO-V5 recognition method is generated at Google Collab from the sample, which is then utilized to train the sample. Upon completion of the retraining process, the material can be evaluated using images. This process began with the dataset collection in 2019, as outlined in [41]. The collection comprised around 150,000 captioned features across a frequency range and profile trait filtration, along with 3,892 augmented raw snapshots used to establish fake and genuine face detection. Figure 7 shows the outcomes of this phase.

C. CLASSIFICATION

The third phase involves the identification and classification of multi-pose human faces using a pre-trained model. During the training of the YOLO-V5 framework, the process commences with 100-150 iterations, also known as epochs. The results are shown in categories of real face and fake face as shown in Figure 6, calculate based on which consists of 416 width size, 16 blocks, and 100 epochs.

1) PERFORMANCE IMPROVEMENT

Last 50% of the dataset is utilized to achieve improvement by dividing the data into blocks that are equivalent in terms of accuracy. The pixel count is primarily 415 on the seventh line and is further categorized into averages, top speeds, and estimation squares. Ultimately, the results are presented using confidence scores, frame categories, and certainty numbers, as depicted in Figure 8.

D. PREDICTION

In the next phase, results are predicted based on the proposed model as shown in Figure 10. The outcomes are represented

FIGURE 7. After Changing the Dataset from Raw to Labeled Face with Bounding Box looking like this.

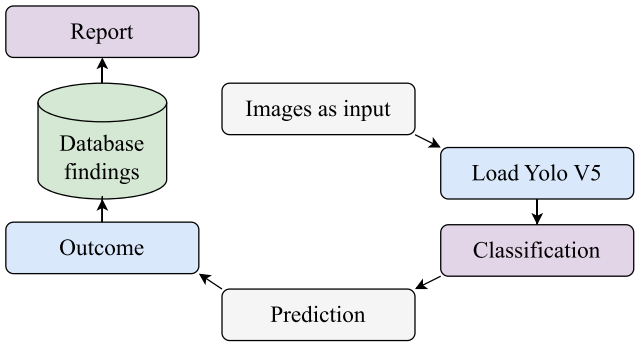

FIGURE 8. Validation Flowchart.

FIGURE 9. Output of testing images with their accuracy.

by inference squares, precision levels, and image classes. The recognition results of an entity type are then deposited in the summary dataset, and subsequently, the entries in the document registry are exhibited on the page. The report segment is designed to present the face-matching accuracy in numerical terms, along with the user's name, date, and time of participation. All these transformations are illustrated in Figure 9, which depicts the data's appearance at the GPU-based endpoint, following the completion of the entire process.

E. OUTCOME

The final stage of the proposed model is the outcome. At this phase, the final results are resented and a report is prepared.

V. EVALUATION MATRICES OF PROPOSED MODEL

In 1 C represents the total number of specific problem domains. If the participant's center does not fall within this element, the image will not be retrieved. Equation 1 is used to determine the exact position of the customer's estimation shot.

$$C_j^i = P_{i,j} \times IOU_{prep}^{truth} \tag{1}$$

The equation (C,J,I) represents the strength proportion of the j-th shape identification frame of the i-th template in terms of the bandwidth strokes. The parameters i and j correspond to the point of the i-th and j-th elements in the equation. The quadratic formula is used to determine whether there is a body or target, where j equals 1 if there is a Figure in the j-th panel and 0 otherwise. The output anticipate is a frequently used parameter that bridges the gap between the selected and fundamental constitutional images. The accuracy of the identified image increases with a higher threshold ratio. The loss value of the model is defined in Equation 2.

$$\begin{aligned} Loss\ Value = Loss\ Box + Loss\ Clssification \\ + Loss\ Object \end{aligned} \tag{2}$$

L-bx, l-clas, and l-obj serve as threshold markers for filtering out false positives, measuring classification accuracy, and identifying loss attributes respectively, throughout the preceding computation. The class label used for structuring the element for the object is demonstrated in the above formula. The pattern grid cell is utilized for the categorization of both the illustration and shaping of the basic, as shown in Equation 3.

$$b'_{bx} = 1 - IoU + \frac{p^2(b, b^{\Theta})}{c^2} P_{i,j} \times IOU_{prep}^{truth} \tag{3}$$

The assessment of image quality is imperative to ensure the development of an accurate image detection model that can identify individuals in the model effectively. When comparing the F1 rate value and the rate, the former exhibits higher values, thus representing the key measure of the efficiency of our model, as illustrated in Equation 4 [42].

$$\begin{aligned} Precision &= \frac{TP}{TP \pm FN} \times Recall \\ &= \frac{TP}{TP \pm FN} \times F1 \\ &= \frac{2 \times P \times R}{P \pm R} \end{aligned} \tag{4}$$

The recall ratio of 100, the performance rate of 0.985, and F1 score of 0.998 were obtained from the equation presented. This methodology was effectively applied to the generated dataset as described in equation 4. Furthermore, based on the fundamental principles of score analysis, it is evident that the proposed approach delivers enhanced outcomes.

VI. DISCUSSIONS

Ultimately, the primary objective of this study is to identify facial landmarks using a vast database containing various poses and subcategories of genuine and counterfeit appearances. This is commonly used to differentiate between a genuine image and one with numerous distinguishing features as well as bogus heads. The dataset employed in this research yields superior outcomes compared to prior datasets, as demonstrated in the comparison presented in Table 2. To establish the data source, it is necessary to gather pictures in a JPG format, followed by cataloging or marking each image with a box. The marker then generates a file name as an output. Lastly, the file scripts are integrated with the information for image preparation [43].

TABLE 2. Comparison of the dataset with previous datasets.

Dataset	Images	Human Faces	Accuracy
AFW [44]	12880	400	97%
FDDB [45]	13000	500	96%
Pascal face [37]	201609	650	95%
IJB-A [46]	22320	497	92%
MALF [47]	51009	960	91%
RFFD [48]	393703	1680-1750	99.64%

Table 2 compares the dataset design and evaluation to ensure the validity of this research. The dataset comprises meticulously enhanced and altered facial landmarks that are composites of several individuals differentiated by glasses, facial features, neck, or natural appearance. Furthermore, it examines each pattern in every round. This approach is consistent with previous studies that employed both supervised and unsupervised techniques to construct the YOLO-V5 framework, which was utilized in this study, resulting in an accuracy of 99%. However, in contrast to previous investigations, the raw data used here is applied to recognize facial images from real-time streams.

VII. EXPERIMENTAL RESULTS

Experimental results performed for the proposed research work are discussed here.

A. GRAPHICAL INTERFACE OF DATASET AND MATRICES USED IN YOLO-V5 WHICH BASED AT GPU

The point at which the graph starts to slope downwards (in the multi-pose method) after 100 iterations at mA@0.5, Precision, and Memory represents the threshold for reduced output, in accordance with the desired learning objectives. The results of this can be seen in Figure 10.

The graphical representation of the RFFD dataset training and implementation with the YOLO-V5 model for multi-pose human face recognition is displayed in Figure 10, depicting the bounding box accuracy and the result precision of the images. The second block demonstrates the image classification accuracy, while the third block highlights the precision accuracy of the dataset. The final block illustrates the recall accuracy of the model, and the results are based on mean average precision and time with various parameters' accuracy [43]. The outcomes are a consequence of the learning process, as demonstrated by the graph in Figure 10, where the maximum accuracy priority of 0.99% is achieved

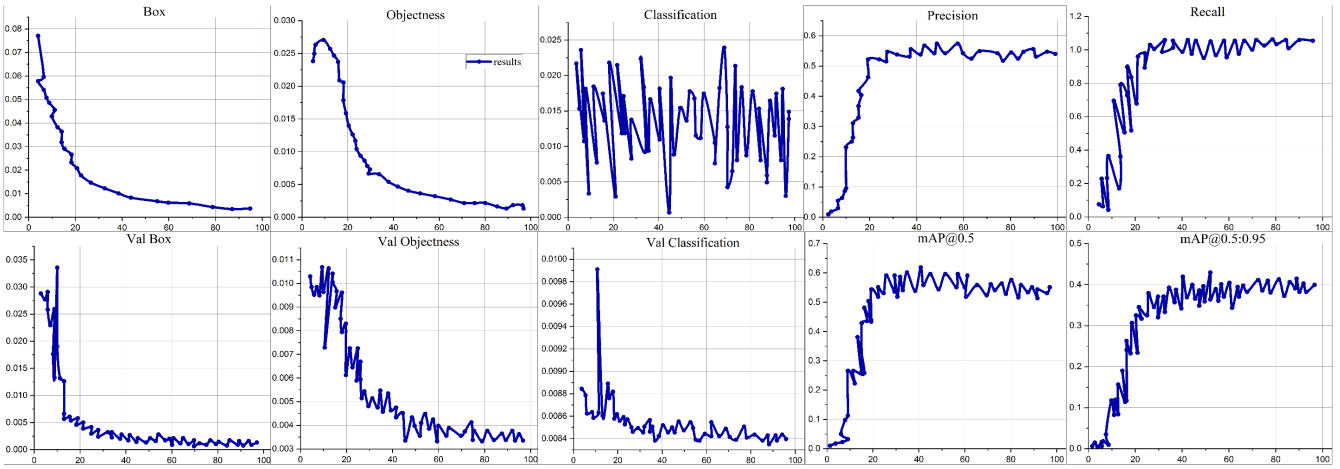

FIGURE 10. Graphical Representation of Dataset Training with YOLO-V5 Matrices.

at 100 epochs. Consequently, increasing the number of iterations above 100 yields greater rewards.

The accuracy of face recognition is assessed separately on both GPU and CPU levels for specific reasons. This approach allows researchers and developers to compare the system's performance under different computational configurations. By doing so, they can understand the hardware's impact, evaluate scalability, and make informed decisions about resource allocation based on speed and accuracy. The distinction between GPU and CPU accuracy provides valuable insights into how each configuration affects the face recognition system, ensuring a well-informed deployment strategy.

B. FACE RECOGNITION ACCURACY RESULTS OF PROPOSED MODEL AT GPU

Previous studies on human cross profile detection models utilizing YOLO-V5 and multiple databases have shown significant performance. However, many experiments have struggled to accurately identify natural human appearances, and improving their accuracy has posed challenges. Prior research has compared the proposed strategy presented in Table 3. In 2021, an enrollment individual recognition system was developed to enhance education and employment processes. Individuals trained their YOLO-V5 framework for template matching using 1280 facial landmark images. The quality of the study was based on retrieval but rather implies reliability precision to the facial expression of sentient pixels occurring in various components termed (simple, mild, tough) structures [49]. Although it takes a long time to discover individuals, as it requires 5000 cycles, they also struggle to find many facial landmarks, as shown in Table 3 below.

Table 3 presents a comprehensive overview of various face detection studies utilizing YOLO-V5 as the base framework. Among the methods, YOLO-V5 base Attendance system [43] exhibits robust accuracy at the basic frame level (97.5%), but its performance falters in more challenging scenarios (76%). Similarly, YOLO-V5 based Live human detection [50] method achieves commendable accuracy at the basic level (95.6%) but faces notable difficulties in handling tough frames (68%). While the Structure of multi-patterns with YOLO-V5 [51] demonstrates promising results at the basic level (93.1%), it reveals limitations in more complex scenarios (80%). On the other hand, YOLO5Face algorithm [38] showcases remarkable accuracy at both the basic (96%) and middle (95%) frame levels but encounters a decline in performance in challenging frames (86%). YOLO-V5 base live streaming human face detection algorithm [52] also performs well at the basic level (91.9%), but its accuracy decreases in tough frames (74.9%). The Angular measurement method [53] exhibits commendable accuracy at the basic (94%) and middle (93%) levels, but it faces challenges in tough scenarios (80%). Despite Human multi-pose recognizer with YOLO-V5 [54] achieving high accuracy at all frame levels (95.5%, 94.5%, and 88% at basic, middle, and tough levels, respectively), it still shows potential limitations in more difficult scenarios.

The proposed method, a modified version of YOLO-V5, emerges as a groundbreaking solution to the shortcomings of existing methods in face detection. Based on the results presented in Table 3, the proposed model showcases better performance, surpassing all other available methods across all frame levels. With a perfect 100% accuracy at the basic frame level and an impressive 99.5% accuracy at the middle frame level, the proposed model demonstrates its ability to accurately detect faces in both simple and moderately complex scenarios. Most notably, the proposed method excels in tackling tough frame-level situations, achieving a remarkable 99% accuracy, showcasing its robustness and adaptability to challenging real-world conditions. The utilization of the Private and RFFD database, consisting of 1680 faces, further enhances the model's ability to generalize and detect faces with high precision. The proposed method's superior performance can be attributed to its effective modifications

TABLE 3. Accuracy comparison of YOLO-V5 algorithms with this proposed YOLO-V5 model (online GPU).

Previous Models	Databases	Frames Quantity	Accuracy at basic frame level	Accuracy at Middle frame level	Accuracy at tough frame level	Iterations
YOLO-V5 base Attendance system [43]	Multi dataset	1430 faces	97.5%	87%	76%	1000
Live human detection YOLO-V5 [50]	Google, COCO database	2300 images	95.6%	90%	68%	100
Structure of multi-patterns with YOLO-V5 [51]	Scut head dataset	845 faces	93.1%	78%	80%	100
YOLO5Face [38]	Wider Face dataset	1056 images	96%	95%	86%	100
YOLO-V5 base live streaming human face detection [52]	Sample with wider face	960 samples	91.9%	81%	74.9%	100
Angular measurement [53]	Collection of COCO, Custom	300 faces	94%	93%	80%	100
Human multi-pose recognizer with YOLO-V5 [54]	Databases with wider face	880 images	95.5%	94.5%	88%	100
Short frame detection depends YOLO-V5 [55]	Informative IMFD database	900 samples	82%	77%	75%	300
The proposed model	Private and RFFD database	1680 faces	100%	99.5%	99%	100

and optimizations to the YOLO-V5 architecture, making it a pioneering advancement in the field of face detection and a promising foundation for future research and real-world applications.

C. CPU BASED INTERFACE FOR MULTI-POSE FACE RECOGNITION

The Python interface is utilized to convey the facial recognition results to the CPU, where Indian celebrity images are procured from online search engines to assess the efficacy of our YOLO-V5 model interface. The interface's impact on retention is gauged by several factors, including accuracy, the number of participants, and the identification of users along with their respective names. Furthermore, the interface architecture is agnostic, meaning that it can be accessed without any language constraints [56]. The multi-pose recognition model for human faces is acquired using Python, and the recognition time for each individual takes approximately 2 seconds, as demonstrated in Table 4.

The research project involving Classifier edition is noteworthy for its use of rigorous tasks, including 460 visuals sourced from Web search, examination of a limited dataset of only 15 image data to derive conclusions for mouth matching, resulting in a reliability of 87% based on 14 images. However, a previous report highlights the unreliability of sample sizes, with only 90 individuals identified and misleading results generated, resulting in a completely false value of 15%. Profile pairing is also a time-consuming process, from ideation to obtaining results using supplied information on individuals. Nevertheless, YOLO-V5 utilized in this experiment yields superior benefits when compared to other versions, such as YOLO-V3 which used 40 screenshots to evaluate their model with multiple categories. Despite the challenge of head identification with only 36 screenshots, a precise pairing of mouth and six different illustrations resulted in useful output. However, the accuracy of prediction is inferior compared to the conceptual framework presented in this study. The database comprises sixty photos of facial

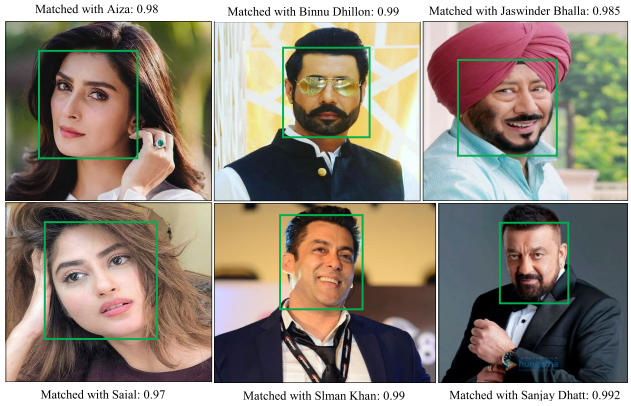

FIGURE 11. Multi-pose face recognition system results in images.

expressions with seven categories, and the (RFFD) sets are used to create YOLO-V4 patterns for each individual. While the personal YOLO-V4 shows advancements over past versions, it requires less data and takes longer than the YOLO-V5 employed in this work.

Their approach yields an 8% error rate within a lengthy timeframe of seventy seconds, requiring significant effort, and offering minimal return with a 2% false positive rate. Thus, in comparison to other models, our system proves to be more accurate. The experimentation was conducted through digital means, specifically photographs, on a runtime NVidia P100-PC-64GB with 52280 MB of storage, using a computer vision learning algorithm. The system executes image face detection, processing 40 photos containing multiple faces. The face-matching stage for the 40 photographs was completed within 5 seconds, including the pre-treatment phase. The average time for each portrait's face recognition is 0.02 seconds. Refer to Figure 11 for a visual representation of some of the matching results.

Based on the analysis of the first iteration of the YOLO model, the authors used 160 images from the internet, out of which only 15-20 were analyzed, and proved to be

TABLE 4. Average recognition measurement of performance with other earlier utilised YOLO editions for multi-posse appearance identification (Desktop-based).

Previous Methods	Database quantity	Different human Patterns	Complete Testing accuracy	Wrong detection False rate	Interface Execution time in seconds
YOLO-V1 [56]	450 images from Google	4 pattern of each person	86%	14%	68 seconds
YOLO-V2 [57]	20 images from COCO	3 pattern of every person	90%	10%	7.36 seconds
YOLO [31]	50 images from wider face	5 pattern of each person	85%	5%	55 seconds
YOLO-V4 [58]	100 images from private dataset	6 patterns of each person	98.3%	6%	23.83 seconds
YOLO-V4 [59]	61 images from FDDB	4 patterns of every one	93%	7%	46 seconds
YOLO-V5 [54]	340 images from wider face	6 patterns with each person	95%	5%	207.7 milli seconds
YOLO-V5 [52]	15 images from COCO	5 patterns with each person	95%	7%	75 seconds
Proposed Model	180 images from Google/ private	5 patterns of each person	99%	1%	25 seconds

successful in identifying profiles with an accuracy of 86% for 14 images [57]. However, six photos were not recognized, leading to false positives and limiting the combined outcome to 71 people, resulting in a false incidence rate of 15%. Profile identification also requires a significant amount of time, from contributing ideas to receiving conclusions using provided information about an individual. However, the YOLO-V5 model employed in this investigation offers superior outcomes compared to other editions. In contrast to the YOLO-V3 model, which used 41 images to evaluate their model across four categories, our model uses only 36 images for profile identification, resulting in more accurate classification performance. In addition, the YOLO-V4 model employs RFFD sources of data involving sixty photos of human facial expressions with seven categories, resulting in advancements over older designs but requiring less information and running slower than the YOLO-V5 model used in this report.

On the other hand, the YOLOv5 model used in 2022 mostly employed seventeen images to evaluate its prediction performance, requiring a lot of effort and producing a 9% false cost in seventy milliseconds. However, our system corresponds accurately in a shorter amount of time and yields a one-part incorrect ratio, making it more reliable than other models.

In the face detection post, the records used include 1680 images, consisting of 1420 input samples for the training phase, 220 images for validation accuracy, and 40 samples for testing data. The distribution pattern is 92.9% for the training set, 4.5% for priority basis, and 2.6% for system testing. The matrices evaluated with their results during training and testing time give good results based on precision, recall, and mean precision average. The output of the tensor board, including loading the model results and running it, is shown in Figure 12.

In the context of human detection, overall accuracy serves as an impartial estimate that is widely accepted for edge detection, such as characterization and indexing. Throughout this experiment, this metric is employed to select significant features, determine the head structure, and identify individuals, all while achieving an impressive 100% accuracy and a 0.97 reliability score for the exact textual location of selected features, such as a face pattern with a profile name and serial number. This is noteworthy as it

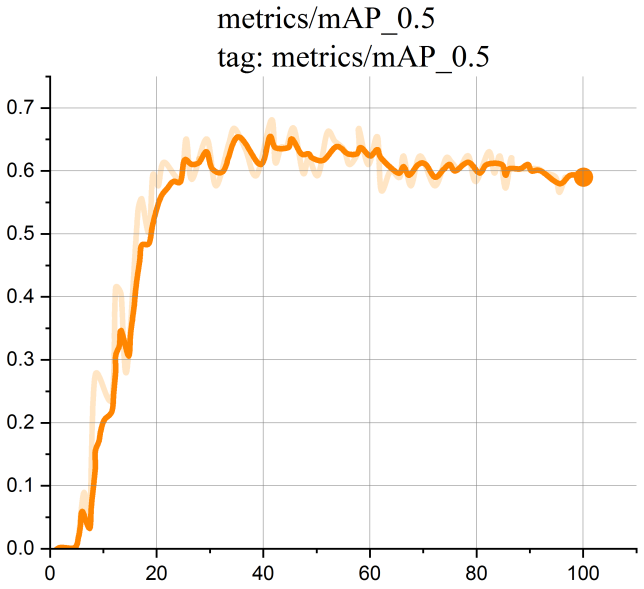

FIGURE 12. The result of mean average precision as Length (y) provides the IoU threshold numbers, Area (x) describes the exact efficiency of the system.

enables the retrieval of critical information. The potential for superior composite classification performance is estimated to be 96%. In a previous assessment, the vectorized platform was evaluated using an IoU benchmark of 0.5. In this survey, the benchmark was raised to 0.6, resulting in a production of 0.95 accuracy and reliability in diagnosis, as depicted in Figure 11. The precision outcomes of the multi-pose human faces recognition model are represented in the form of a graph, as illustrated in Figure 13.

The parameter of interest is the number of true positives divided by the sum of true negatives and false negatives in the quality matrix. As previously demonstrated, the value at a bounding box target of 0.5 is 100%, which is represented by the formula $1/(1+1)=0.5$. This precision rate is excellent for both the RFFD dataset and COCO dataset. In the case of multi-pose face detection using YOLO-V5, the recall value is 100%, represented as a range from 0.5 to 1, as shown in Figure 14.

The outcome of the parameter is calculated as the ratio of true positive results to the sum of true positive and false

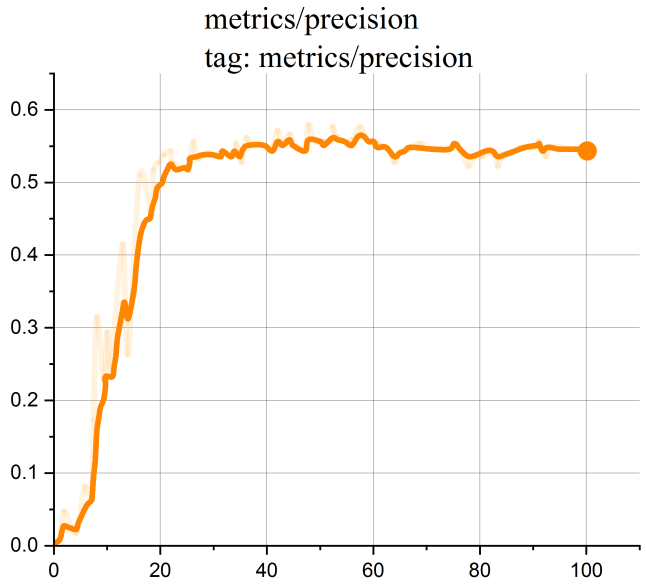

FIGURE 13. Result of precision as length (y) displays the IoU points, Area of width (x) shows the correct accuracy of the system.

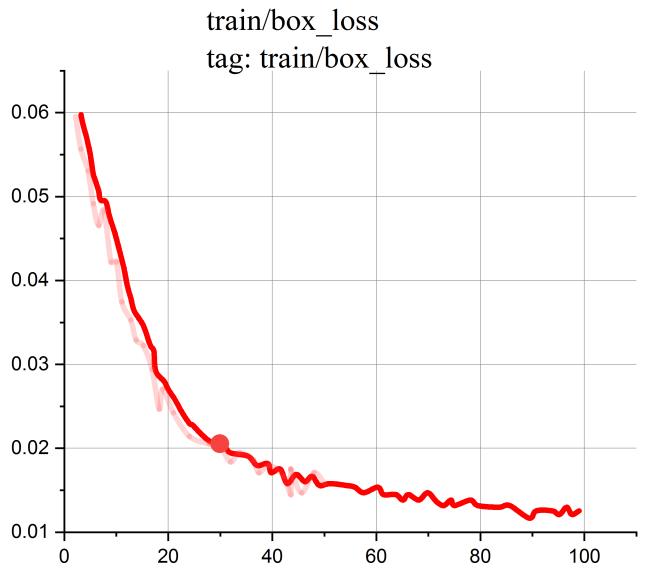

FIGURE 15. The results of package damage throughout the samples training period like (y) point shows the range of sample (x) displays the iterations correct accuracy fail in decreasing level.

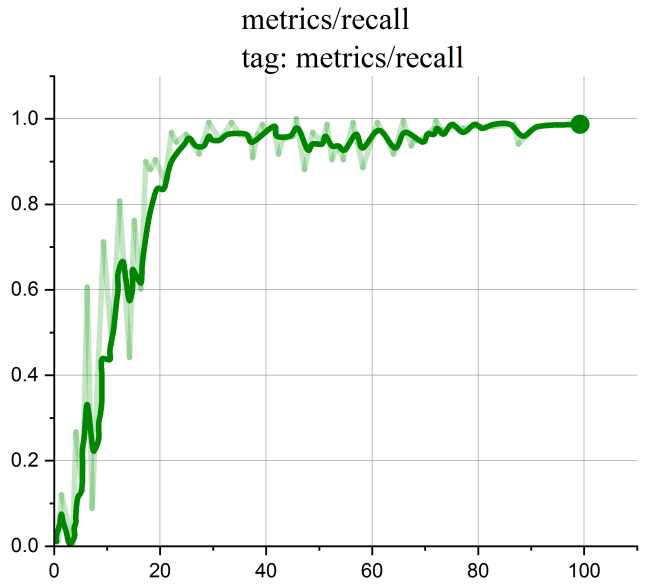

FIGURE 14. The outcome of recall shows as (y) defines the IoU graphic series, (x) shows the correctness of the design.

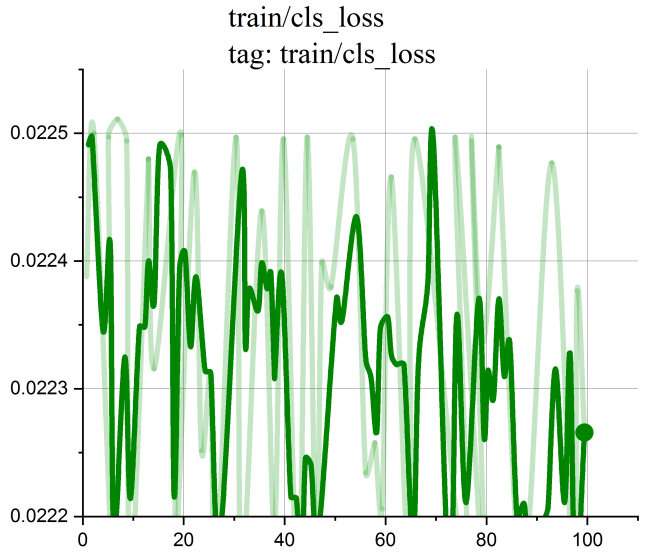

FIGURE 16. The results of transmission line losses throughout the pre-processing step of the sample are as follows: (y) Shown that is symbolized either by axis (x) The period reliability damage is given by pivot in decreasing order.

positive outcomes, within the recall vector. As previously mentioned, this ratio is 100% when a single threshold value is used, which can be represented by the equation 1/1+0=1. While previous studies have reported some losses in the training dataset, this study proposes the use of three different loss functions during training, namely boxing loss, classification loss, and face recognition loss. The severity of the loss due to inadequate training is depicted in Figures 14, 15, and 16, ranging from 95 to 0 and extending beyond infinity. The peak of each graph represents the magnitude of the loss, while the spread of the error function represents the precision of the model. Figure 15 illustrates the boxing loss in the dataset.

During the training phase of the dataset, all images are considered as objects due to the classification process. The loss of images is minimal at 0.5%. Following the boxing training of the given dataset, the classification loss of the dataset is depicted in Figure 16.

Therefore, each set of photographs requires a unique identifier during the training phase to be recognized and identified later. While there may be some minor losses in a categorized training dataset, there are no losses at 225.

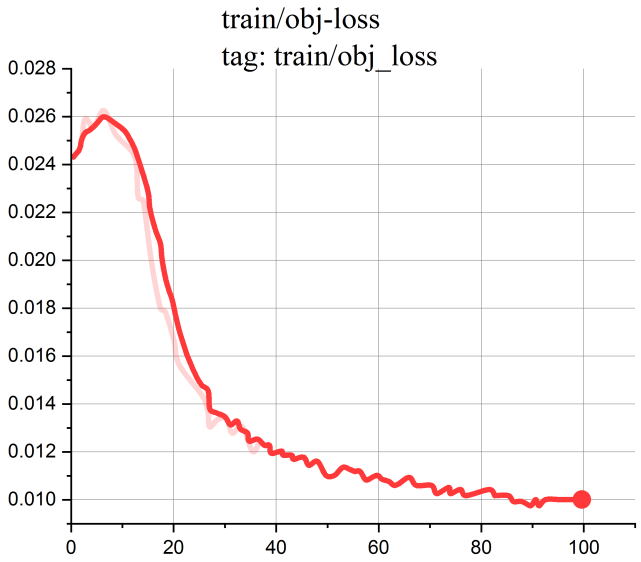

FIGURE 17. The outcome of face matching loss during the training time of dataset looks like as (y) Each facing the quantity of data (x) The era stability loss is described by a vector in lowest to highest.

Figure 17 displays the face-matching loss of the dataset after the boxing and classification training.

At the outset of the photo training process, there was a 2% incidence of item loss. Upon declaration of 35 instances, the algorithm selectively identifies only those which had been previously activated for this purpose, resulting in occasional shortfalls.

VIII. CONCLUSION AND DISCUSSION

This work primarily focuses on utilizing the YOLO-V5 framework for image processing to classify individual interactions and body language from assertive postures, left and right sides, head orientation, and angular orientation of shapes via webcam. This approach proves to be efficient in addressing problems that require quickness and effectiveness. Our research findings indicate that a highly certified dataset, along with an active GPU and CPU, can efficiently handle such scenarios. Cross-facial expression identification in responsive sight is a crucial research subject closely related to our daily routine. Processor-based actual template matching techniques facilitate human-computer interactions, reducing instances and criminals. In-hospital treatment and consultation benefit from an interpretive approach, significantly reducing the need for awareness efforts and resource expenditure, thereby improving our quality of life. Advanced analytics has emerged as a comprehensive education requirement as the world's methodological capabilities continue to evolve. This approach involved gathering and analyzing blockchain materials to extract the suggested YOLO-V5 techniques and head improvements, resulting in a significant increase in the accuracy and efficiency of human identification. The directional percentage, moving up shapes, stylistic layering, x and y-directional actions, and angles and height were consistently picked for this method. The predictor was trained using the sample, resulting in precise individual identification. Mouth detection and face recognition in low-light environments have also improved, addressing difficulties regarding actual posture evaluation when fitted with mirrored surfaces or protective suits. We aim to conduct experiments from various angles, assuming the social world to be free of distortions. Our observations demonstrate the possibility of separating this group of actions. However, different databases, including the multifractal element of three-dimensional graphics, present harsh and complex drawings that require significant research and resources to understand multimedia webcams before handling large amounts of data. Path length assessment is the primary focus for further development, given its poorer efficiency and unclear shots when recognizing a user and the database from a distant location. Our objective is to assert super-intelligent categorization and determine the most suitable procedure to maintain the accuracy and reliability of human activity recognition. We chose the RFFD dataset and a personalized sample using the YOLO-V5 model for greater concentrations, despite its high response rate, being acceptable for our investigation.

REFERENCES

[1] Y. Kortli, M. Jridi, A. A. Falou, and M. Atri, "A novel face detection approach using local binary pattern histogram and support vector machine," in *Proc. Int. Conf. Adv. Syst. Electric Technol. (IC_ASET)*, Mar. 2018, pp. 28–33.

[2] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020.

[3] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Jul. 2018.

[4] S. Setiowati, E. L. Franita, and I. Ardiyanto, "A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods," in *Proc. 9th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, Oct. 2017, pp. 1–6.

[5] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: A real-time face detector," *Vis. Comput.*, vol. 37, no. 4, pp. 805–813, 2021.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 1–12.

[8] S. Ioffe and D. A. Forsyth, "Probabilistic methods for finding people," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 45–68, 2001.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[10] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–12.

[12] J. Dai, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[13] M. Bizjak, P. Peer, and Ž. Emeršic, "Mask R-CNN for ear detection," in *Proc. 42nd Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2019, pp. 1624–1628.

[14] D. Li, Z. Li, R. Luo, J. Deng, and S. Sun, "Multi-pose facial expression recognition based on generative adversarial network," *IEEE Access*, vol. 7, pp. 143980–143989, 2019.

[15] F. Zhang, J. Gao, H. Zhou, J. Zhang, K. Zou, and T. Yuan, "Three-dimensional pose detection method based on keypoints detection network for tomato bunch," *Comput. Electron. Agricult.*, vol. 195, Apr. 2022, Art. no. 106824.
[16] M. D. Putro, Wahyono, and K.-H. Jo, "Multiple layered deep learning based real-time face detection," in *Proc. 5th Int. Conf. Sci. Technol. (ICST)*, vol. 1, Jul. 2019, pp. 1–5.
[17] Z. Ren and X. Xue, "Research on multi pose facial feature recognition based on deep learning," in *Proc. 5th Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Dec. 2020, pp. 1427–1433.
[18] S. Ruan, C. Tang, X. Zhou, Z. Jin, S. Chen, H. Wen, H. Liu, and D. Tang, "Multi-pose face recognition based on deep learning in unconstrained scene," *Appl. Sci.*, vol. 10, no. 13, p. 4669, Jul. 2020.
[19] S. B. Ahmed, S. F. Ali, J. Ahmad, M. Adnan, and M. M. Fraz, "On the frontiers of pose invariant face recognition: A review," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2571–2634, Apr. 2020.
[20] M. Ben Gamra and M. A. Akhloufi, "A review of deep learning techniques for 2D and 3D human pose estimation," *Image Vis. Comput.*, vol. 114, Oct. 2021, Art. no. 104282.
[21] M. Toshpulatov, W. Lee, S. Lee, and A. H. Roudsari, "Human pose, hand and mesh estimation using deep learning: A survey," *J. Supercomput.*, vol. 78, no. 6, pp. 7616–7654, Apr. 2022.
[22] P. Gao, K. Lu, J. Xue, L. Shao, and J. Lyu, "A coarse-to-fine facial landmark detection method based on self-attention mechanism," *IEEE Trans. Multimedia*, vol. 23, pp. 926–938, 2021.
[23] Y. Zhang, K. Fu, J. Wang, and P. Cheng, "Learning from discrete Gaussian label distribution and spatial channel-aware residual attention for head pose estimation," *Neurocomputing*, vol. 407, pp. 259–269, Sep. 2020.
[24] A. Kratimenos, "3D hands, face and body extraction for sign language recognition," in *Proc. Sign Lang. Recognit., Transl. Prod. (SLRTP) Workshop-Extended Abstr.*, vol. 4, 2020, pp. 1–4.
[25] Y. S. Ismael, "Deep learning based real-time face recognition system," *NeuroQuantology*, vol. 20, no. 6, pp. 7355–7366, 2022.
[26] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, Aug. 2021.
[27] L. Zhu, S. Wang, Z. Zhao, X. Xu, and Q. Liu, "CED-net: Contextual encoder–decoder network for 3D face reconstruction," *Multimedia Syst.*, vol. 28, no. 5, pp. 1713–1722, Oct. 2022.
[28] M. D. Putro and K.-H. Jo, "Real-time face tracking for human–robot interaction," in *Proc. Int. Conf. Inf. Commun. Technol. Robot. (ICT-ROBOT)*, Sep. 2018, pp. 1–4.
[29] X. Yang, X. Jia, D. Gong, D.-M. Yan, Z. Li, and W. Liu, "LARNet: Lie algebra residual network for face recognition," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11738–11750.
[30] A. Bansal, R. Ranjan, C. D. Castillo, and R. Chellappa, "Deep CNN face recognition: Looking at the past and the future," in *Deep Learning-Based Face Analytics*. Springer, 2021, pp. 1–20. [Online]. Available: https://doi.org/10.1007/978-3-030-74697-1_1
[31] X. Ling, J. Liang, D. Wang, and J. Yang, "A facial expression recognition system for smart learning based on YOLO and vision transformer," in *Proc. 7th Int. Conf. Comput. Artif. Intell.*, 2021, pp. 178–182.
[32] E. Michos, "Development of an online platform for real-time facial recognition," Postgraduate thesis, Masters HCI, Joint Program ECE CEID, Univ. Patras, Greece, 2021.
[33] N. Wang, Z. Wang, Z. He, B. Huang, L. Zhou, and Z. Han, "A tilt-angle face dataset and its validation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 894–898.
[34] Z. Zhang, S. Xia, Y. Cai, C. Yang, and S. Zeng, "A soft-YoloV4 for high-performance head detection and counting," *Mathematics*, vol. 9, no. 23, p. 3096, Nov. 2021.
[35] K. Bhangale, P. Ingle, R. Kanase, and D. Desale, "Multi-view multi-pose robust face recognition based on VGGNet," in *Proc. 2nd Int. Conf. Image Process. Capsule Netw. (ICIPCN)*. Cham, Switzerland: Springer, Feb. 2022, pp. 414–421.
[36] Q. Li, X. Dong, W. Wang, and C. Shan, "CAS-AIR-3D face: A low-quality, multi-modal and multi-pose 3D face database," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.
[37] T. Xie, Z. Chen, M. Cao, P. Hu, Y. Zeng, and Z. Pan, "Face detection in VR games," in *Proc. 3rd Int. Conf. Control Comput. Vis.*, Aug. 2020, pp. 7–10.
[38] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," in *Proc. Comput. Vis.–ECCV Workshops*, Tel Aviv, Israel. Cham, Switzerland: Springer, 2023, pp. 228–244.
[39] D. Thuan, "Evolution of YOLO algorithm and YOLOv5: The state-of-the-art object detention algorithm," Bachelor's thesis, DIN16SP, Inf. Technol., Oulu Univ. Appl. Sci., Finland, 2021.
[40] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[41] A. A. Una, E. Haque, N. S. Ritu, Z. T. Haque, and R. S. Opal, "Classification technique for face-spoof detection in artificial neural networks using concepts of machine learning," Bachelor's thesis, Dept. Comput. Sci. Eng., Brac Univ., Bangladesh, 2021.
[42] V. Shinde, N. Jagtap, and H. Shukla, "Deep learning based face-mask and shield detection," in *Proc. Int. Conf. Comput. Intell. Comput. Appl. (ICCICA)*, Nov. 2021, pp. 1–4.
[43] Mardiana, M. A. Muhammad, and Y. Mulyani, "Library attendance system using YOLOv5 faces recognition," in *Proc. Int. Conf. Converging Technol. Electr. Inf. Eng. (ICCTEIE)*, Oct. 2021, pp. 68–72.
[44] H. Wu, D. Ma, Z. Mao, and J. Sun, "SSRFD: Single shot real-time face detector," *Appl. Intell.*, vol. 52, no. 10, pp. 11916–11927, 2022.
[45] F. A. M. Ali and M. S. Al-Tamimi, "Face mask detection methods and techniques: A review," *Int. J. Nonlinear Anal. Appl.*, vol. 13, no. 1, pp. 3811–3823, 2022.
[46] Q. Xu, Z. Zhu, H. Ge, Z. Zhang, and X. Zang, "Effective face detector based on YOLOv5 and superresolution reconstruction," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–9, Nov. 2021.
[47] S. Dooley, G. Z. Wei, T. Goldstein, and J. P. Dickerson, "Are commercial face detection models as biased as academic models?" 2022, *arXiv:2201.10047*.
[48] A. Douklias, L. Karagiannidis, F. Misichroni, and A. Amditis, "Design and implementation of a UAV-based airborne computing platform for computer vision and machine learning applications," *Sensors*, vol. 22, no. 5, p. 2049, Mar. 2022.
[49] A. Ali-Gombe, E. Elyan, C. F. Moreno-García, and J. Zwiegelaar, "Face detection with YOLO on edge," in *Proc. 22nd Eng. Appl. Neural Netw. Conf. (EANN)*. Cham, Switzerland: Springer, 2021, pp. 284–292.
[50] G. Castellano, B. De Carolis, N. Marvulli, M. Sciancalepore, and G. Vessio, "Real-time age estimation from facial images using YOLO and efficientnet," in *Proc. Comput. Anal. Images Patterns, 19th Int. Conf. (CAIP)*. Cham, Switzerland: Springer, Sep. 2021, pp. 275–284.
[51] A. Wang, X. Cao, L. Lu, X. Zhou, and X. Sun, "Design of efficient human head statistics system in the large-angle overlooking scene," *Electronics*, vol. 10, no. 15, p. 1851, Jul. 2021.
[52] A. Ghimire, N. Werghi, S. Javed, and J. Dias, "Real-time face recognition system," 2022, *arXiv:2204.08978*.
[53] I. H. Al Amin and F. H. Arby, "Implementation of YOLO-v5 for a real time social distancing detection," *J. Appl. Informat. Comput.*, vol. 6, no. 1, pp. 1–6, Jul. 2022.
[54] N. Kim, J.-H. Kim, and C. S. Won, "FAFD: Fast and accurate face detector," *Electronics*, vol. 11, no. 6, p. 875, Mar. 2022.
[55] R. Chatterjee, A. Chatterjee, S. H. Islam, and M. K. Khan, "An object detection-based few-shot learning approach for multimedia quality assessment," *Multimedia Syst.*, vol. 29, no. 5, pp. 1–14, Oct. 2023.
[56] N. Darapaneni, A. K. Evoori, V. B. Vemuri, T. Arichandrapandian, G. Karthikeyan, A. R. Paduri, D. Babu, and J. Madhavan, "Automatic face detection and recognition for attendance maintenance," in *Proc. IEEE 15th Int. Conf. Ind. Inf. Syst. (ICIIS)*, 2020, pp. 236–241.
[57] H. Deshpande, A. Singh, and H. Herunde, "Comparative analysis on YOLO object detection with OpenCV," *Int. J. Res. Ind. Eng.*, vol. 9, no. 1, pp. 46–64, 2020.
[58] J. Yu and W. Zhang, "Face mask wearing detection algorithm based on improved YOLO-v4," *Sensors*, vol. 21, no. 9, p. 3263, May 2021.
[59] Y. Liu, R. Liu, S. Wang, D. Yan, B. Peng, and T. Zhang, "Video face detection based on improved SSD model and target tracking algorithm," *J. Web Eng.*, vol. 21, no. 2, pp. 545–568, 2022.

MUHAMMAD SOHAIL received the bachelor's degree in computer engineering from the University of Engineering and Technology (UET), Taxila, and the master's degree in computer science from the Riphah College of Computing, Riphah International University, Faisalabad, in 2023. His research interests include smart cities, smart transportation, advanced driving assistance systems, the Internet of Vehicles, smart education, and machine learning.

IJAZ ALI SHOUKAT received the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM). The Ph.D. thesis is on information security and applied cryptography. Currently, he is an Associate Professor with the Computing Department, Riphah Internaitonal University, Faisalabad Campus. He retains extensive academic, industry, and research experiences. His academic brilliance reflects the talent reward through the Outstanding Talent Support Scheme by the Punjab Information Technology Board, Government of Punjab, Pakistan.

ABD ULLAH KHAN (Member, IEEE) received the Ph.D. degree in computer science from the Ghulam Ishaq Khan Institute of Engineering Science and Technology, Pakistan, in 2021. He is currently an Assistant Professor with the National University of Science and Technology, Pakistan. He is also the Founding Member of the Highly Secure and Spectrum-Efficient Network (HiSEN) Research Group. He has over 20 journal publications in diverse reputed venues, such as IEEE Internet of Things Journal, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Network and Service Management, IEEE Wireless Communications Letters, *Future Generation Computer Systems* (Elsevier), and *Journal of Network and Computer Applications* (Elsevier). His research interests include resource allocation and management in wireless networks and network security. Besides, he is an Active Reviewer of IEEE Network, IEEE Internet of Things Journal, IEEE System Journal, IEEE Access, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Vehicular Technology, IEEE Transactions on Intelligent Transportation Systems, and *Computer Communications* (Elsevier).

HARAM FATIMA received the B.S. degree in computer science from Government College University Faisalabad, in 2020, and the M.S. degree in computer science from Riphah International University, Faisalabad Campus, Pakistan. Her research interests include machine learning and multi-pose face recognition.

MOHSIN RAZA JAFRI received the Ph.D. degree in computer science from Universitã Ca' Foscari Venezia, in 2019. He has been an Assistant Professor of computer science with the National University of Sciences and Technology (NUST), Pakistan, since June 2019. His research interests include communication system design, wireless sensor networks, and underwater sensor networks. He has contributed to developing network simulators and energy-efficient algorithms for wireless communication. Moreover, he has also developed stochastic models for the performance analysis of wireless sensor networks.

MUHAMMAD AZFAR YAQUB (Member, IEEE) received the bachelor's degree from COMSATS University Islamabad (CUI), Pakistan, in 2007, the master's degree from Lancaster University, U.K., in 2010, and the Ph.D. degree from the School of Computer Science and Engineering (SCSE), Kyungpook National University (KNU), Republic of Korea, in 2019. He is currently an RTDA Research Assistant with the Faculty of Engineering, Free University of Bozen-Bolzano, Italy. Previously, he was a Lecturer with the Department of Electrical and Computer Engineering, CUI, from 2008 to 2021, where he was an Assistant Professor, from 2021 to 2023. His research interests include future internet architectures, information-centric networks, CCN/NDN, wireless ad-hoc networks, sensor networks, connected vehicles, and video streaming. He is an ACM Member and serves as a TPC/reviewer for several conferences and journals.

ANTONIO LIOTTA (Senior Member, IEEE) is currently a Full Professor with the Faculty of Computer Science, Free University of Bozen-Bolzano, Italy, where he teaches data science and machine learning. Previously, he was the Founding Director of the Data Science Research Centre, University of Derby, U.K. He is credited with over 350 publications involving, overall, more than 150 coauthors. His research interests include artificial intelligence theories and applications, particularly artificial vision, e-health, intelligent networks, and intelligent systems. He is the Editor-in-Chief of the *Internet of Things* (Springer) book series (springer.com/series/11636) and an associate editor of several prestigious journals.

Open Access funding provided by 'Libera Università di Bolzano' within the CRUI CARE Agreement