Comparative Analysis of Pre-trained Deep Learning Models for Facial Landmark Localization on Enhanced Dataset of Heavily Occluded Face Images

Zieb Rabie Alqahtani [1], Mohd Shahrizal Sunar [1,*], and Abdelmonim M. Artoli [2]

[1] Media and Game Innovation Centre of Excellence, Institute of Human Centered Engineering, University of Technology Malaysia, Johor, Malaysia
[2] Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
Email: zralqahtani@graduate.utm.my (Z.R.A.); shahrizal@utm.my (M.S.S.); aartoli@ksu.edu.sa (A.M.A.)
*Corresponding author

***Abstract*—The face is the main component in the human body to be considered in the physical world as it is read to know the feelings of someone, in the same way in computer vision its detection and its landmark localization are pivotal for applications spanning from facial recognition to emotion analysis and augmented reality. Existing datasets in this domain lack diversity especially, in terms of occluded faces, particularly those obscured by medical masks or niqabs. Moreover, a majority of images were captured in controlled environments with limited variations in pose and lighting. This paper addresses this gap by focusing on occluded face images and localizing five crucial landmarks or key points (eyes, nose, and mouth corners) of the face. The Niqab dataset was substantially enhanced with the addition of 11,000 images to the ENiqab-V1 dataset, predominantly featuring faces with 80 to 100% occlusions. Four deep learning models, three particularly belong to the same domain with high accuracy and one is a general object detection model, namely MediaPipe, face.evoLVE, TorchLM, and YOLOv5, were subjected to transfer learning over the ENiqab-V1 dataset. The goal is to perform a comparative analysis of the models and suggest future guidelines for potential accuracy improvement through fine-tuning. The models were evaluated based on accuracy and Mean Square Error (MSE), yielding accuracies of 48.56%, 59.62%, 52.8%, and 52.7%, and Mean Squared Errors (MSEs) of 0.78, 0.59, 1.2, and 0.85, respectively. The comparative analysis shows that face.evoLVE has the highest accuracy but for facial landmark localization over heavily occluded face images we suggest the general object detection model YOLOv5 due to its potential for optimization in terms of accuracy.**

***Keywords*—object detection, facial landmarks, heavily occluded face, deep learning**

I. Introduction

The objects (detection and classification or recognition) in human vision is a trivial task. A two-year-old child can instantly recognize tens of objects in an image effortlessly [1]. The task of object recognition is very hard to accomplish with computer vision; it requires the processing of highly dense data with expensive computation resources and through intelligent algorithms. Computer vision has a vital role and delivers remarkable outcomes in artificial intelligence. Computer vision enables to understand the digital images and videos [2] The development of smart methods based on deep learning models and computer vision, the computer can accurately detect and classify objects and respond accordingly to what they visualized [3]. The main aim of this study is on face detection and facial landmark localization of highly occluded faces, which is a sub-field of object detection and is considered as a single instance of object detection [4]. The computer vision-based face detection and facial landmarks localization starts from the low-level phase an image is acquired from a digital camera or video frame and delivered to the pre-processing stage. After pre-processing, the features are extracted either with hand-engineered design as in traditional machine learning techniques, or selected and extracted automatically and learned via deep-learning and convolutional neural networks [5]. The obtained results based on features are used for the detection and classification of objects. The classification area is used to classify different objects according to their class for segmentation or recognition while the detection of an object is further analyzed based on interest in face detection, face landmarks are detected and face alignment has been made for further elicitation of knowledge. The final output is in the form of extracted information such as face identification, emotion detection, or drowsiness detection based on some key points suitable or acceptable in that particular area of interest [6]. Further, Fig. 1 specifies the research focus area.

Manuscript received February 10, 2023; revised April 28, 2024; accepted May 28, 2024; published November 17, 2024.

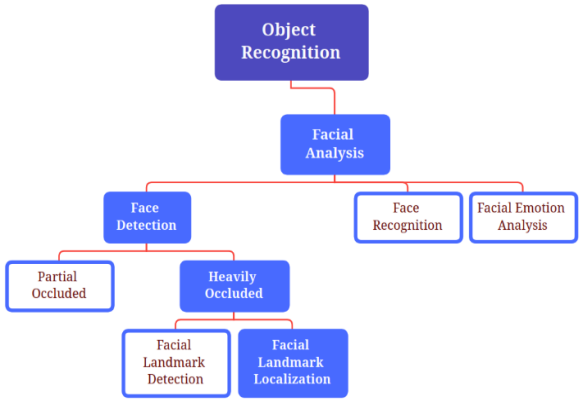

Fig. 1. Research area categorization.

The area of facial landmarks localization has shown considerable improvement in the last decade, however, the area has much more space in terms of accuracy for the situation in which the faces are covered or densely occluded due to masks (Medical health issues like recent compulsion in COVID-19) and Niqab (Religious or cultural norms of some traditions, civilizations, and religions). Face analysis-related research starts from its detection to localization of its landmarks, all having associated challenges due to the following factors:

Additional face components: Faces are covered with additional components like a beard, mustache, sunglasses or sight glasses, masks, and niqab. In addition, they are frequently changing from face-to-face and gender to gender [7]. Some of these components are:

1) Face posture: The face position or posture has variation, some face images are in frontal posture while some are half profile and in back position. All these variations make the task of face detection challenging [8].
2) Expression: Human expressions directly appear on the face or one may know the human intentions from his/her facial expressions which may be happiness, sadness, aggression, normal, and worry. These expressions change the face's appearance in various degrees and make it more complex for detection [9–11].
3) Size, illumination, and orientation: The size of faces varies according to age and gender, poor light conditions are also a hindrance while the orientation may also change from image to image or frame to frame. These factors always play a key role in the disturbance of accurate face detection [12–16].
4) Face with partial and full occlusion: The face is sometimes partially occluded with the face additional components like hair, glasses, etc., while fully occluded due to face masks and Niqab. In some cases, no key point or landmark is available [17–19].

The underlying study focuses mainly on addressing all these issues to accurately localize the face landmarks as shown in Fig. 2.

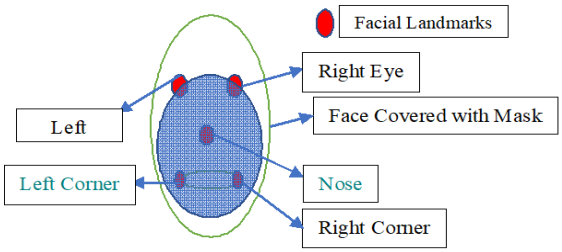

Fig. 2. Facial Landmarks highlighted in face covered with mask or niqab.

II. Literature Review

Facial landmark localization and recognition are not significant by themselves as an independent application but as a key step for many face application systems, such as gaze detection, drowsiness, mobile augmented reality, surveillance and security monitoring systems, face recognition, face reenactment, facial emotion recognition, and 3D face modeling [20]. All these facial applications are useless and could not work if the facial landmark localization system is not involved because it is the initial step for all these applications. Moreover, this study for the first time gives attention to highly occluded face images covered completely with niqab by considering the preprocessing techniques to enhance and make visible the facial landmarks. The niqabs in the majority of cases are worn by women due to cultural and religious norms. The piece of cloth that covers the face area is relatively thin as compared to other clothes in which the Landmarks are hidden but with less effort of preprocessing, they may be elaborated. The elaboration of at least one facial key point will enable us to calculate the rest of the Landmarks in heavily occluded face images. The evolution of face recognition related research is presented in Fig. 3.

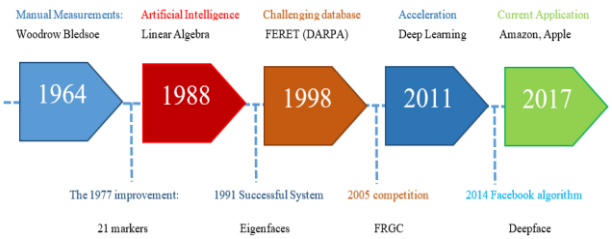

Fig. 3. Evolution of face recognition techniques.

Face recognition techniques relies on prominent face features referred as landmarks [21, 22]. These landmarks includes eyes, nose, lips and shape of face. Facial landmark localization is the initial step in research related to face analysis [23]. Face detection and facial landmark localization of human faces in digital images is a non-trivial task for computer vision. Facial landmark localization has received much focus from computer vision researchers and is still an ongoing challenge for solution or improvement [24]. The problem is mainly in an unconstrained environment in which the face appears in extreme variations in pose, lighting and heavy occlusions [25–28]. The main objectives of facial landmarks localization is to correctly localize the face attribute in its exact position. The face landmarks are

utilized in different face recognition-based applications. Emotion recognition application, utilize the face landmarks lips, eyes and eye brows to predict the emotional state of subject [29–30]. Biometric user verification through face recognition also utilizes face landmarks. Image enhancement applications and filters also need landmarks for putting application's features [31]. Fig. 4 depicts neural networks-based face landmark detection logical schema.

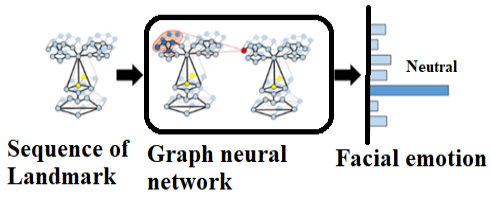

Fig. 4. Face landmarks and neural networks-based emotion recognition.

A comprehensive related work method review has been conducted by Song *et al.* [32] and Kowalski [33]. It was categorized into Parametric and Non-parametric Shape Model-based methods. In contrast, the research work reported in [33] has divided the previous approaches into conventional and deep learning approaches of machine learning to the same problem. These are discussed subsequently.

A. *Parametric Approaches*

This approach defines the data in the model to a specific distribution, it identifies each landmark by a distribution, such as the Gaussian method. In this subsection. The statistical distribution of facial feature points is constructed from 600 face images as in Fig. 5. It is further classified based on their appearance model into two classes: local part model-based [30, 34, 35] and holistic sketch-based methods, such as Active Appearance Models (AAM) [36, 37], where 600 shapes (black) normalized, red indicates the mean shape of all shapes adapted. Fig. 5 depicts statistical facial points.

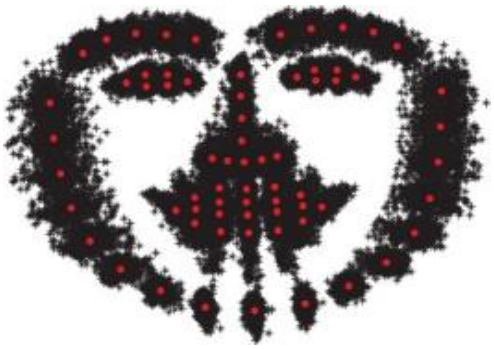

Fig. 5. Statistical distribution of facial feature points.

B. *Local Part or Constrained Local Model-based Methods*

The constrained local method is used in [38–40]. An input image is fit for a specific shape via an objective function which implies two parameters: shape prior R(P) [41, 42] and the sum of response maps [43] D(X; I) as presented in Eq. (1).

$$\min_{\mathrm{p}} \mathrm{R(p)} + \sum_{i=1}^{N} D(X;I) \tag{1}$$

C. *Holistic Model-Based Methods*

Holistic methods utilize holistic sketch information of face. It also exploits the global facial shape patterns for detecting facial landmarks. The Active Appearance Model (AAM) is a classic example of holistic modelling. It was proposed as a statistical model by Milborrow and Nicolls [44], Wimmer *et al.* [45]. The model analyzes facial key-point correlations and matches shape and texture simultaneously. This technique is actually matching a holistic model of that can fit the whole class of objects rather than tracking deformable objects [46]. In Ref. [47], some extensions by fitting more landmarks than needed and using a two-dimensional sketch of facial landmarks instead of one. Fig. 6 depicts the holistic shape model based landmark localization.

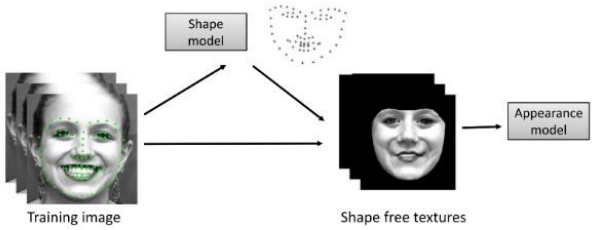

Fig. 6. Illustration of Holistic model approach.

The flexibility of parametric shape models is hard to achieve. Such as Principal Component Analysis (PCA), which is determined heuristically.

D. *Non-Parametric Model-Based Methods*

These approaches are distribution-free and do not rely on assumptions that the data are drawn from a given probability distribution [48]. The distinction between the two models is that a fixed number of parameters is defined in the parametric model. In contrast, in the non-parametric, the number of parameters grows gradually with the training data [28, 49].

E. *Cascaded Regression-Based Methods (CR)*

CR methods have recently become the most prominent methods for face alignment because of their reported accuracy and speed [23, 50–53]. This approach from image appearance learns a regression function to fit the aimed output. The regression estimation starts from initial shape s_0 and step by step refines the face shape s by estimating the shape increment Δs.

F. *Deep Learning CNN-Based Methods (DL-CNN)*

The Deep Learning CNN-Based Methods (DL-CNN) approach in the ImageNet competition [54] in 2012 [55], has been adopted, produced a high impact results on a variety of computer vision tasks and issues, as in image classification [56–58], object detection [59–61] and face detection [62–66]. The early work was carried out using a deep probabilistic model called Boltzmann machines

which were used to capture pose and expressions variations for face landmark detection [67, 68]. DL-CNN recently became the most popular model for all face-related applications, landmark localization, and detection. The study of Nan *et al.* [69] categorizes deep learning methods into two groups: those for nonlinear shape variations and those for nonlinear mapping from appearance to shape. Examples of handling nonlinear shapes are found by Ranjan *et al.* [70], previously mentioned, and in Zhou *et al.* [71], which introduced a hierarchical probabilistic model to address the challenges posed by variations in facial expressions and poses. Zhang *et al.* [72] categorized deep learning-CNN methods into two categories hybrid-based and pure-based learning methods. In pure learning methods, landmark locations are predicted directly by the CNN model [73]. Moreover, Ma *et al.* [73] used a cascade of four convolutional layers to predict five facial landmarks from a given bounding box of input image face. Each landmark point is then refined by an external network as illustrated in Fig. 7.

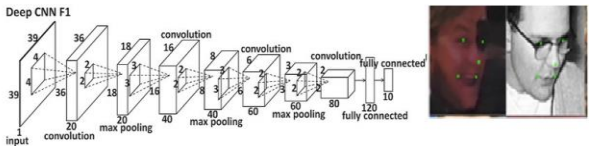

Fig. 7. CNN-Cascaded layers for five landmark point's prediction.

However, increasing landmark feature points would increase the computation time required for detecting all points [74]. Multi-task learning is based on task-based feature extraction and its adaptation for different tasks. The research work reported in Ranjan *et al.* [75], and Zhou *et al.* [76] proposed Tasks-Constrained Deep-CNN for gender, pose, and emotion estimation with facial landmark points. In Ref. [77], a similar CNN model was proposed to predict face detection, landmark localization, pose estimation, and girder detection in joint-related tasks. In this model, features from multiple layers were shared to utilize the low-level to high-level feature representation. In an improvement of the CNN cascaded framework, similar work was proposed by Nguyen *et al.* [78] to predict gradually 68 landmark facial points instead of five landmark points. An intuitive approach by Rothe *et al.* [79] and Hassaballah *et al.* [80] proposed a Multi-Task CNN (MTCNN) framework, consisting of three stages, as illustrated in Fig. 8, for predicting face detection and five landmark localizations. In the first stage, the CNN-layer called P-Net, inspired by Krizhevsky *et al.* [54], is a region proposal network that proposes regions with bounding boxes. The obtained regions are refined by the Non-Maximum Suppressions (NMS) technique to eliminate overlapped bounding boxes [81]. The output of the previous stage P-net is fed to the second stage network R-Net which performs more filtering on false positive candidates and applies NMS on bounding boxes for more calibration. The final stage O-Net takes the output of R-Net and outputs the face bounding box along with five facial landmark points. Since three tasks involved face/non-face classification, bounding box regression, and five landmark localization, three loss functions have been used during training. The 2-fold Cross-Entropy Loss was formulated as the learning objective so that the output belongs to either (0, 1).

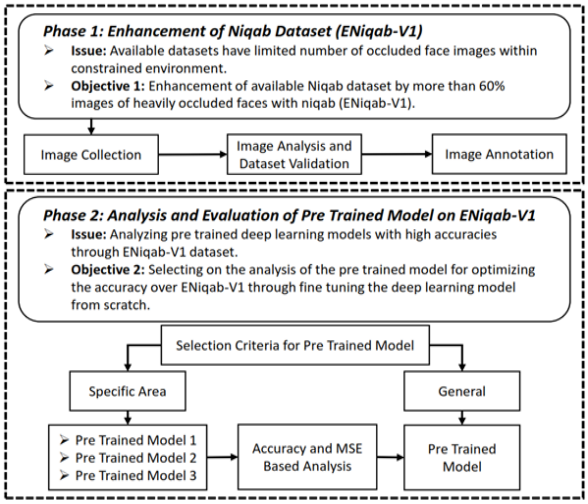

Fig. 8. Proposed methodology.

G. Face Landmark Localization in Heavily Occluded Face

Most of the facial landmark localization algorithm's attention is on images taken in a controlled environment with less variation. However, images of human faces would appear in huge varieties and variations in real-world situations due to lighting differences, pose occlusion, face occlusions, and other variations. Occlusion is one of the main reasons for the failure of the facial landmark localization algorithms [82], as the key points defined on a clear face may disappear in a face covered with masks and niqabs. Due to health face masks or fashion masks, the three landmarks, i.e., the left and right corners of the mouth and nose completely covered with masks, while in a niqab, all five landmarks disappear. Simple occlusions can happen due to extreme head pose or caused by overlaying of other objects such as glasses, eyes or hands on the mouth or nose. When handling occlusion, there are some challenges. Predicting which parts of the facial landmark are occluded is quite difficult. Another issue is that face landmarks could be occluded arbitrarily with arbitrary objects that can vary in size and shape [83]. Many of the models already developed give little attention to occluded faces or are ignored completely. These studies mainly trained their models on ideal or controlled images or partially uncontrolled images like hair on one eye or shadow. Most of the current works dealt with occlusion as a separate challenge, and therefore the designed individual models based on the assumption that some of the face parts are occluded and then concatenate these models for localization. Wu and Ji [84] split the face into a 3 × 3 grid. The total is nine regions assuming that one part is at least not occluded. Information about facial appearance from one part is utilized for predicting the localization of facial occlusion for all occluded parts by merging predictions of the nine parts probability. This approach helped in improving both occlusion detection and landmark prediction at once. However, the model will abruptly go

down in terms of accuracy when applying heavily occluded faces because, in some of these images, no key point is visible that estimation may be made to calculate the other landmarks. In Ref. [85], the face is described as a combination of parts, each part composed of local facial landmarks; training these separate models is based on predicting non-occluded parts. The drawback of these previously mentioned methods is occlusion-dependent models, which assume that some predefined parts are occluded; however, facial occlusion is arbitrary in unconstrained situations. These models might not be able to cover complex occlusion as in real-world scenarios [86]. For example, Muhi *et al.* [87] dealt with occlusion as an independent framework to avoid the mentioned challenge. A hierarchical deformable probabilistic model was designed, which encodes rules for occlusion modeling of facial landmarks and visibility of each landmark to predict joint landmark locations and landmark occlusion [21]. Bhatlawande *et al.* [88] proposed a cascaded regression model based on local appearance in order to predict probabilities of landmark locations and their visibility in an iterative manner, occlusion pattern is added as a constraint in the prediction. It depends more on appearance from visible facial landmarks rather than occluded landmarks. This framework can handle occlusions of object occlusion and extreme head poses [21]. Further in Table I, the summary of the previous studies deal with facial landmark localization is presented.

TABLE I. SUMMARY OF THE PREVIOUS STUDIES FOR FACIAL LANDMARK LOCALIZATION

Ref.	Methodology	Dataset	Result
[1]	Mobile net SSD architecture	Niqab dataset	99.6%
[16]	Recurrent CNN (R-CNN)	UBIRIS.v2 MICHE	0.02 NE
[17]	Adversarial Occlusion-aware Face Detector	MAFA, FDDB	97.88%
[21]	Recurrent Attentive-Refinement (RAR) LSTM	300-W, COFW, AFLW	16.3% error reduction
[23]	Generative Adversarial Network	Mask 300, Simulated Masked Face Recognition Dataset (SMFRD)	3.45% MSE
[30]	CNN (Face Attention Network)	Wider Face, MAFA	79%–89%
[48]	DCNN Face Detector	MTCNN	99.6%
[77]	Hierarchal Part Model for Occlusion Model	Multi PIE, IBUG HELEN, COFW	98.95%
[89]	CNN with grid loss and hinge loss	AFLW, FDDB PASCAL	87.1%
[90]	CNN	iBUG, LFPW, AFW,HELEN	3.37% NRMSE
[91]	Face Direct Vector and YOLO 3	Celeb A, Augmented dataset	0.02 Normalize error
[92]	Generative Adversarial Network	CelebA	78.7%
[93]	Occlusion Adaptive Deep Network	Occlusion AffectNet OcclusionFERPlus FED-RO	89.83%
[94]	GAN AIs (Generative Adversarial Network)	OCFW, COFW	0.06 NRMSE
[95]	Resnet and Unet	CelebAMask-HQ	98.3%
[96]	ADA Face	OCFR 2022	Rank 1

However, all the studies mentioned above used datasets, which have partially unconstrained images and when applied to highly occluded datasets, these approaches will show a drastic deviation in accuracy. Due to mentioned limitations in the literature, this paper sets the following points: to enhance the niqab dataset by 11,000 images of heavily occluded faces, implementing the existing high accuracies facial landmark localization techniques through transfer learning over the heavily occluded ENiqab-V1 dataset to analyze the accuracy and mean square error of each model. Based on the analysis of the per-trained models, give suggestion for adoption of one model for optimizing its accuracy through fine tuning from scratch and its evaluation on the ENiqab-V1 dataset.

III. MATERIALS AND METHODS

Face landmark localization in heavily occluded faces is simple in images with clear faces but become critical when the face images are fully covered with medical masks and niqabs because most of the key landmarks such as nose, eyes and lips are unavailable. The proposed research work is implemented in two phases as illustrated in Fig. 8.

Phase 1 of this research mainly focuses on the dataset enhancement to include heavily occluded face images, annotation of the images, and its evaluation while Phase 2 deals with analyzing the most prominent pre-trained deep learning models in the area on the ENiqab-V1 dataset. Finally, give recommendations for the selection of an appropriate model having the capacity for optimizing the accuracy on logical grounds.

A. Phase 1: Dataset Collection, Preparation and Annotations

In this phase the process starts with choosing the appropriate images that are aligned with the research scope, the face images with high occlusion (mostly from 80% to 100%) with high degree of variations. For the mentioned purpose the Niqab dataset is more than 60% enhanced to ENiqab-V1. The images were collected from various search engines and analyzed. More than 11,000 images were collected and added to the dataset which has already 10,000 images. The qualifying criteria of images to be included in the dataset are "human-in-the-loop". The search keywords are "femme niqab Maroc", "woman niqab France", "woman hijab Africa", "hijab + sunglasses", "niqab + sunglasses", and "woman + mask + sunglasses". The images are converted to .jpg format and renamed with "Bulkre_name utility". The dataset was carefully validated by two independent analysts to remove the outlier images while the validation was carried out with a statistical metric, Kohen's Kappa. The task of image analysis is performed manually. Two analysts work for 10 days on 11,000 images. The analysis task involves checking the "human in the loop process" with the help of Google, Bing, and Yahoo search engines with the help of niqab-related keyword queries. The collected images are renamed using the renaming library mentioned above. The image validation metric is Kohen's Kaffa, which finds the agreement probability between the two raters. The process of finding Kohen's Kappa is presented here to find the

agreement probability between the two raters. Initially total number of images collected was 15,000 according to the mentioned criteria while two sets of the same images were distributed between both the raters. Table II presents the summarized results of the images examined or rated by both raters.

TABLE II. PRESENTS RATING STATISTICS BY TWO RATERS

Rater	A	B	C	D
1 2	11000	47	50	2903

Note: A = Total number of Images that both the analyst agreed; B = The total number of Images analyst 1 considered correct while analyst 2 considered incorrect; C = The total number of Images analyst 2 considered correct while analyst 1 considered incorrect; D = The total number of Images analyst 1 and analyst 2 considered incorrect

The Cohen's Kappa is the probability of agreement between the two raters and calculated through Eq. (2).

$$K = \frac{Po-Pe}{1-Pe} \tag{2}$$

where, *Po* is Probability of agreement and *Pe* is random probability and both are calculated through Eqs. (3) and (4) respectively.

$$Po = \frac{Number\ in\ Agreement}{Total} \tag{3}$$

$$Pe = P(correct) + P(incorrect) \tag{4}$$

where, in Eq. (3), *P(correct)* is the probability of images declared correct by both the raters while *P(incorrect)* is the probability of images declared incorrect by both the raters. The calculations of both can be carried out with Eqs. (5) and (6) respectively.

$$P(correct) = \frac{(A+B)}{(A+B+C+D)} \times \frac{(A+C)}{(A+B+C+D)} \tag{5}$$

$$P(correct) = \frac{(C+D)}{(A+B+C+D)} \times \frac{(B+D)}{(A+B+C+D)} \tag{6}$$

According to Table II the calculations of Cohen's Kappa has been made and the results obtained are presented in graphical for in Fig. 9.

Fig. 9. The agreement variables versus the agreement or disagreement values on these variables for calculation of Cohen's Kappa.

The Cohen's Kappa calculated is 0.83, shows high agreement probability of both the raters on the dataset ENiqab-V1 images.

B. Phase 2: Pre Trained Deep Learning Models Evaluation on ENiqab-V1 Dataset

In this phase of the underlying research, first a criteria for the consideration of deep learning models is devised. In this back drop, this research considers four deep learning models on the criteria that three pre trained models were particularly taken from the area of facial landmarks localization with high prediction accuracies while one of the model was generally taken from the area of object detection with high accuracy and other robust features. The models considered under the mentioned criteria are: MediaPipe, face.evoLVe, Torchlm (Specifically trained and tested for facial landmarks localization) and YOLOv5 (General object detection model). The experimentation has been carried out on the ENiqab-V1 dataset applied to the pre trained models.

MediaPipe Model offers a pre-trained face landmark model that can be used for detecting facial landmarks on a face. The model is based on a deep neural network architecture and is trained to work in real-time on mobile devices and other platforms [97–99]. The pre trained face landmark model provided by MediaPipe is trained on the 300W-LP dataset, which is a large-scale benchmark dataset for facial landmark localization [100]. The dataset consists of over 60,000 images of faces, each with 68 annotated landmarks indicating the location of various facial features [101]. The 300W-LP dataset is widely used in the computer vision research community for training and evaluating facial landmark localization models. It includes a wide range of face images with variations in pose, expression, and lighting conditions, making it a challenging dataset for training robust models. MediaPipe also provides the option to train custom face landmark models using user-provided datasets. Overall, the use of the 300W-LP dataset for training the pre-trained face landmark model provided by MediaPipe ensures that the model has been trained on a large and diverse set of faces, which helps to improve its performance and robustness. The maximum achieved an average accuracy of the model is 95.7%.

face.evoLVe model is a deep learning model that can be used for facial landmarks localization. Specifically, it is a Convolutional Neural Network (CNN) architecture that uses a ResNet-50 backbone network with stacked hourglass modules for feature extraction and key point localization. The model takes an input image of a face and outputs a set of 68 landmarks corresponding to various facial landmarks, such as the corners of the eyes, nose, and mouth, as well as the outline of the face. These landmarks can be used for various applications, such as face recognition, emotion detection, and virtual makeup. The face. Evolve model was developed by the Insight Face research group and is part of the Insight Face toolkit, which is a collection of deep learning models and tools for face recognition and analysis. The model has been trained on several large-scale datasets [102]. 300W: A widely used dataset for facial landmark localization, which consists of 300 face images with 68 annotated landmarks per image. The dataset includes both indoor and outdoor images with varying lighting conditions, pose, and expression, AFLW:

The Annotated Facial Landmarks in the Wild dataset contains over 24,000 images of faces in the wild with annotated landmarks. The dataset includes a diverse set of images with varying pose, expression, and lighting conditions. COFW: The Calibration-Free Face Pose dataset contains 1,345 annotated face images with varying pose, expression, and lighting conditions. The dataset is designed to evaluate facial landmark localization in real-world scenarios. However, all these datasets have lack of occluded face images with masks or with niqab partially or fully. The model achieved the lowest state of the art Normalized Mean Error (NME) of 3.32% on 300W full test images while the achieved accuracy is 99.7%.

TorchLM is a PyTorch landmarks-only library with 100+ data augmentations, support training and inference. TorchLM aims only focus on any landmark localization, such as face landmarks, hand key points and body key points, etc. It provides 30+ native data augmentations and can bind with 80+ transforms from TorchVision and albumentations, no matter the input is an np.array or a torch Tensor, TorchLM will automatically be compatible with different data types and then wrap it back to the original type through a autotype wrapper. TorchLM ran experiments on three datasets: 300W, COFW68, and WFLW-68 and achieved an average accuracy of 96.45 %.

YOLOv5 is an open-source object detection framework developed by Ultralytics. YOLO stands for "You Only Look Once", which refers to the framework's single-shot object detection algorithm that can detect multiple objects in an image or video in real-time with high accuracy [103]. YOLOv5 is based on a deep Convolutional Neural Network (CNN) architecture that is designed to be faster and more accurate than its predecessors, such as YOLOv4 [103]. The YOLOv5 model uses a combination of anchor boxes, feature pyramid networks, and deep residual networks to identify objects in images or videos. YOLOv5 is lightweight, making it suitable for deployment on edge devices with limited computational resources. YOLOv5 model is trained on the COCO dataset, and includes simple functionality for Test Time Augmentation (TTA), model ensemble, hyper parameter evolution, and export to ONNX, CoreML and TFLite trained on the Objectron dataset. The Yolov5 achieved an average accuracy of 94.56 % trained on multi task facial dataset.

IV. Experimental Results and Discussion

The experimentation has been carried out on the ENiqab-V1 dataset applied to the pre-trained models. Accuracy and mean square error are considered as performance evaluation metrics. The results of each model are presented first for visual analysis and then its comparative analysis has been made for discussion.

A. Accuracy

Accuracy is one of the evaluation metrics to define how the model was accurate. The obtained accuracy is calculated by dividing the True Positive (TP)—Correct detection by the ground-truths of faces. Eq. (7) represents the accuracy computed for the mentioned deep learning model evaluation:

$$R = \frac{True\ Positive}{All\ Ground\ Truth} \tag{7}$$

If the accuracy is less than 75% indicates that the model is in conflict with the objectives of this paper.

B. Mean Squared Error

Error can be measured by computing the difference between the inferred values and the ground truth values [100]. The most commonly used error metrics are Mean Square Error (MSE) [101]. Standard evaluations in the face alignment literature are usually expressed as the point-to-point MSE error between each point of the predicted shape and the ground truth annotations. Eq. (8), presents the MSE computation.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{8}$$

where, Y is the ground-truth point, $\hat{Y}$ is the predicted point, and n are the number of samples.

C. Experimentation Results of MediaPipe

The experimentation with ENiqab-V1 dataset made on the MediaPipe model through transfer learning results a Mean Square Error (MSE) of 0.78 while an accuracy of 48.569% which is drastic fall while the same model in constrained and partially unconstrained approaches gives accuracy more than 90%. The MediaPipe framework is previously adapted for occluded face detection [102], face landmark localization, face pose detection and face mask detection. Three kinds of sample images were extracted:

Partial Detection Sample: The model did not detect heavy occluded faces to be able to detect the facial landmarks, while it detects un-occluded faces with mismatching facial landmarks.

No Detection Sample: The model did not detect the heavily occluded and non-occluded faces.

Key points Location Sample: The model can detect face well, but there is mismatch of facial landmarks key points

The visual results of the model are presented in Fig. 10, which clearly shows the missed and detected facial landmark key points.

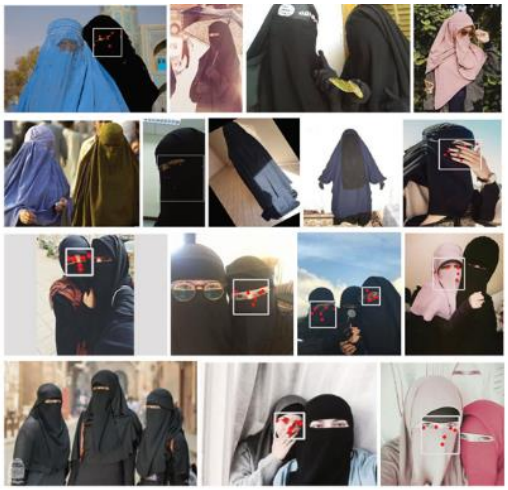

Fig. 10. Facial landmarks localized with MediaPipe pre trained model.

D. Experimentation Results of face.evoLVe

The experimentation with face.evoLVe model on the ENiqab-V1 dataset, in comparison to other models considered in this research improved the accuracy by almost 10 points, however the model needs heavy improvement. The highly and heavy occluded faces were extracted from the ENiqab-V1 dataset. The result of the model achieved 0.59 MSE and 59.4% accuracy. Three kinds of sample images were extracted:

Partial Detection Sample: This sample shows how the model is performing poor since the model was trained on selfie images and it did not recognize a semi-occluded face taken by a selfie camera.

No Detection Sample: The model did not detect the heavily occluded and non-occluded faces.

Key points Location Sample: The model can detect face well, but there is mismatch of facial landmarks key points

The visual results of the model are presented in Fig. 11, which clearly shows the missed and detected facial landmark key points.

Fig. 11. Facial landmarks localization with face, evoLVe pre trained model.

E. Experimentation Results of TorchLM

The highly and heavy occluded faces were extracted from the selected dataset. The result of the model achieved 1.01 MSE and 52.07% accuracy. There are three kinds of samples as mentioned:

Partial Detection Sample: As can be seen in the next Fig. 12, the model did not detect heavy occluded faces to detect the facial landmarks, while it detects un-occluded faces with mismatching facial landmarks.

No Detection Sample: The model did not detect the heavily occluded and non-occluded faces.

Key points Location Sample: The model can detect face well, but there is mismatch of facial landmarks key points

The visual results of the model are presented in Fig. 12, which clearly shows the missed and detected facial landmark key points.

Fig. 12. Facial landmarks detected with TorchLM pre trained model.

F. Experimentation Results of YOLOv5

The YOLOv5 model predicts the following results when applied on the ENiqab-V1 dataset through transfer learning. The reason behind the low accuracy is highly occluded faces in the dataset, however the model is pre trained for object detection and it detects the faces accurately but need fine tuning and from scratch training on the ENiqab-V1 dataset. The highly and heavy occluded faces were extracted from the selected dataset. The result of the model achieved 0.87 MSE and 52.6% accuracy. Three kinds of samples have been experimented as mentioned below:

Partial Detection Sample: As can be seen in the next Fig. 13, the model did not detected heavy occluded faces to detect the facial landmarks, while it detects un-occluded faces with mismatching facial landmarks.

No Detection Sample: The model did not detect the heavily occluded and non-occluded faces.

Key points Location Sample: The model can detect face well, but there is mismatch of facial landmarks key points.

Fig. 13. Facial landmark localization with YOLOv5 pre trained model.

G. Comparative Analysis of the Pre Trained Models

In this section, all the above pre trained models in detail were analyzed and one of the model was selected for

further experimentation on the basis of justified grounds that will be fine tune for optimal prediction in face detection and facial landmarks localization. As mentioned earlier that three out of four models are specifically applied in the area while the fourth model (YOLOv5) is used for object detection in real-time images and videos. The above mentioned models are used for experimentations in this research through transfer learning to analyze their MSE and accuracy. The results of all the four models are presented in Figs. 14 and 15, respectively.

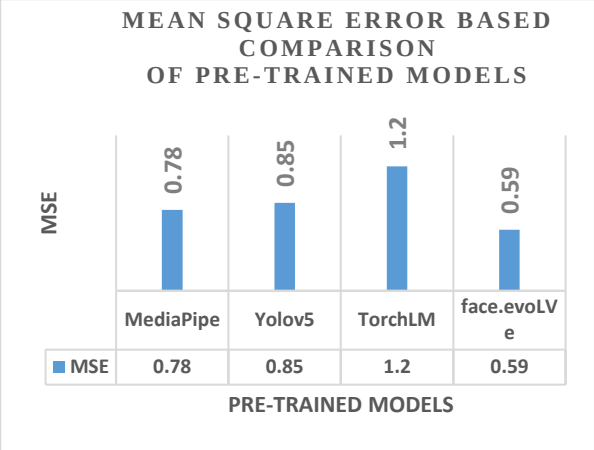

Fig. 14. Mean Square Error comparison of the four pre trained models tested over the ENiqab-V1.

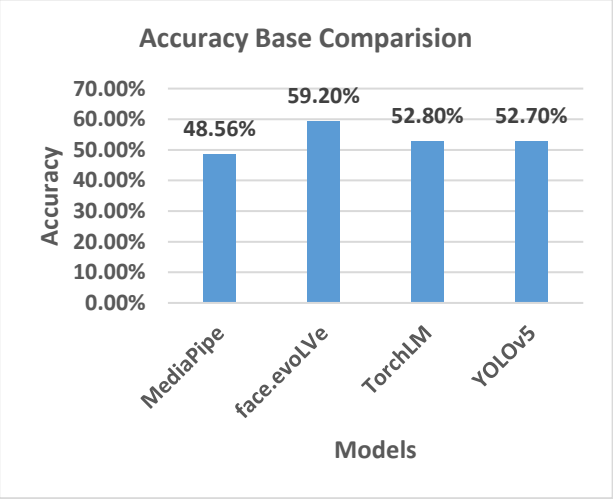

Fig. 15. Accuracy based comparison of the four pre trained models tested over the ENiqab-V1.

As all these models accuracies are below the criteria mentioned previously and their MSEs are also high than the mentioned criteria. In these four models the three (Mediapipe, TorchLM and face.evoLVe) are specifically trained for facial landmarks localization but their accuracies drastically degraded on ENiqab-V1 dataset (48.569, 52.8 and 59.2 respectively) while the MSEs are also high (0.72, 1.2 and 0.59 respectively). The fourth is general object detection model (YOLOv5) whose accuracy is 52.7 and MSE is 0.85. Now, if the models on the same basis are compared with their achieved accuracies on their own datasets shown abrupt fall in the accuracy as presented in Table III.

TABLE III. COMPARISON OF THE FOUR MODELS ON THE BASIS OF THEIR ACHIEVED ACCURACIES AND MSES WITH THE ACCURACIES AND MSE TESTED WITH ENIQAB-V1

Mode	Achieved Accuracy	Accuracy ENiqab-V1	Difference
MediaPipe	95.7%	48.56%	47.14%
face.evoLVe	99.7%	59.2%	40.50%
TorchLM	96.45%	52.8%	43.65%
YOLOv5	94.56%	52.7%	41.86%

The fall in accuracy by almost an average of 40% of all the models is due to the heavily occluded face images with 80 to 100% covered with niqab in ENiqab-V1 dataset. The potential drawback in the MediaPipe model, non-handling of large pose variations and requirement of significant computational resources are the main causes to drop the model in case of heavily occluded face images. In the same way, face.evLVe model requires high computational resources, limited applicability to the real world scenarios and potential bias make it out of the race. Further, TorchLM is designed specifically for facial landmarks localization, so it may not be optimized for this task to improve the accuracy. YOLOv5 is a general object detection model and has much more space for optimization to improve the accuracy over heavily occluded dataset. YOLOv5 will also be applicable for real world scenarios.

V. CONCLUSION

In the four pre-trained models, the three (Mediapipe, TorchLM and face.evoLVe) are specifically trained for facial landmarks localization but their accuracies drastically degraded on ENiqab-V1 dataset (48.569, 52.8 and 59.2 respectively) while the MSEs are also high (0.72, 1.2 and 0.59 respectively). The fourth is the general object detection model (YOLOv5) whose accuracy is 52.7 and MSE is 0.85. Now, this study suggest the selection of YOLOv5 model for enhancement to achieve high accuracy over any occluded face image dataset on the following grounds:

1) This research mainly involve two steps i.e., face detection and facial landmarks localization, while the YOLOv5 deep learning model is basically developed for real time object detection in images and videos, thus with fine tuning it has space for optimization of accuracy.
2) The model is specifically built for real-time object detection in images and videos while its use through transfer learning in this research shows considerable accuracy and shows more space for optimization as compare to other three models which are specifically build for facial landmarks key point localization.
3) YOLOv5 has achieved state-of-the-art accuracy on various object detection tasks, including detecting small objects and achieving high precision [103]. This may be advantageous for facial landmark key point localization tasks that require accurate and precise detection of facial features.
4) YOLOv5 is optimized for speed and can process images and videos in real-time on a CPU, making

it a suitable option for applications that require fast and efficient detection. This may be useful for facial landmark key point localization tasks that require real-time performance, such as in video analysis.

YOLOv5 is highly customizable, allowing for fine-tuning and adaptation to specific use cases and datasets. This flexibility may be useful for facial landmark key point localization tasks that require customization and adaptation to specific scenarios and applications [103, 104]. To summarize, harnessing the capabilities of the YOLOv5 model, along with the ability to customize and fine-tune its parameters, offers potential for improving accuracy in occluded face image datasets. This approach holds the promise of enhancing the performance of facial landmarks localization systems, making them more effective in real-world applications.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

ZA performed conceptualization, methodology, data collection, analysis, and writing original draft. MS performs supervision, reviewed, editing and approved the final version. AA did co-supervision, reviewed and editing. All the authors had approved the final version.

ACKNOWLEDGMENT

The authors would like to thank members of the Media and Game Innovation Centre of Excellence (MaGICX) and the Institute of Human Centered Engineering (iHumEn), Universiti Teknologi Malaysia, for their time, effort, and enthusiasm.

REFERENCES

[1] Martvel, I. Shimshoni, and A. Zamansky, "Automated detection of cat facial landmarks," *International Journal of Computer Vision*, pp. 1–6, 2024.

[2] Y. H. Chen, "Iterative refinement strategy for automated data labeling: Facial landmark diagnosis in medical imaging," arXiv preprint, arXiv:2404.05348, 2024.

[3] A. F. Jafargholkhanloo and M. Shamsi, "Quantitative analysis of facial soft tissue using weighted cascade regression model applicable for facial plastic surgery," *Signal Processing: Image Communication*, vol. 121, 117086, 2024.

[4] A. A. S. Alashbi, M. S. Sunar, and Z. Alqahtani, "Deep-learning-CNN for detecting covered faces with Niqab," *Journal of* Information *Technology Management*, vol. 14, pp. 114–123, 2022.

[5] N. O'Mahony, S. Campbell, A. Carvalho *et al.*, "Deep learning vs. traditional computer vision," in *Proc. Advances in Computer Vision: the 2019 Computer Vision Conference (CVC)*, 2020, pp. 128–144.

[6] W. AbdAlmageed, Y. Wu, S. Rawls *et al.*, "Face recognition using deep multi-pose representations," in *Proc. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

[7] H. Hoffmann, H. Kessler, T. Eppel *et al.*, "Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men," *Acta Psychologica*, vol. 135, no. 3, pp. 278–283, 2010.

[8] A. I. D. Paiva-Silva *et al.*, "How do we evaluate facial emotion recognition?" *Psychology & Neuroscience*, vol. 9, no. 2, 153, 2016.

[9] A. Bansal *et al.*, "Umdfaces: An annotated face dataset for training deep networks," in *Proc. 2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 464–473.

[10] Z. Lei, S. Liao, M. Pietikäinen, and S. Z. Li, "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 247–256, 2010.

[11] N. S. Vu and A. Caplier, "Face recognition with patterns of oriented edge magnitudes," in *Proc. the Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete*, Greece: Springer Berlin Heidelberg, 2010, pp. 313–326.

[12] F. M. Ramírez, "Orientation encoding and viewpoint invariance in face recognition: Inferring neural properties from large-scale signals," *The Neuroscientist*, vol. 24, no. 6, pp. 582–608, 2018.

[13] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint, arXiv:1803.01271, 2018.

[14] S. Liu, Y. Song, M. Zhang, J. Zhao, S. Yang, and K. Hou, "An Identity authentication method combining liveness detection and face recognition," *Sensors*, vol. 19, no. 21, 4733, 2019.

[15] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: A real-time face detector," *The Visual Computer*, vol. 37, pp. 805–813, 2021.

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

[17] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[18] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan, "Deep recurrent regression for facial landmark detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1144–1157, 2016.

[19] M. Hassaballah, E. Salem, A. M. M. Ali, and M. M. Mahmoud, "Deep recurrent regression with a heatmap coupling module for facial landmarks detection," *Cognitive Computation*, pp. 1–15, 2022.

[20] L. Zhou, H. Zhao, and J. Leng, "MTCNet: Multi-task collaboration network for rotation-invariance face detection," *Pattern Recognition*, vol. 124, 108425, 2022.

[21] Y. Chong *et al.*, "Automated anatomical landmark detection on 3D facial images using U-NET-based deep learning algorithm," *Quantitative Imaging in Medicine and Surgery*, vol. 14, no. 3, 2466, 2024.

[22] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3486–3496.

[23] X. Jin and X. Tan, "Face alignment in-the-wild: A survey," *Computer Vision and Image Understanding*, vol. 162, pp. 1–22, 2017.

[24] Y. Wu, C. Gou, and Q. Ji, "Simultaneous facial landmark detection, pose, and deformation estimation under facial occlusion," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3471–3480.

[25] Q. T. Ngoc, S. Lee, and B. C. Song, "Facial landmark-based emotion recognition via directed graph neural network," *Electronics*, vol. 9, no. 5, 764, 2020.

[26] A. Farkhod, A. B. Abdusalomov, M. Mukhiddinov, and Y. I. Cho, "Development of real-time landmark-based emotion recognition CNN for masked faces," *Sensors*, vol. 22, no. 22, 8704, 2022.

[27] R. D. Putra, T. W. Purboyo, and L. A. Prasasti, "A review of image enhancement methods," *International Journal of Applied Engineering Research*, vol. 12, no. 23, pp. 13596–13603, 2017.

[28] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, 2018.

[29] L. Song, C. Hong, T. Gao, and J. Yu, "Lightweight facial landmark detection network based on improved MobileViT," *Signal, Image and Video Processing*, pp. 1–9, 2024.

[30] M. Kowalski, "Localization and tracking of facial landmarks in images and video sequences," PhD thesis, Institute of Radioelectronics and Multimedia Technology, 2018.

[31] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "MOFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proc.*

the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1274–1283.
[32] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. the 5th European Conference on Computer Vision, Freiburg, Germany: Springer Berlin Heidelberg*, 1998, pp. 484–498.
[33] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *Proc. the 5th European Conference on Computer Vision, Freiburg*, Germany: Springer Berlin Heidelberg, 1998, pp. 581–595.
[34] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977–1981, 2005.
[35] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2617–2628, 2007.
[36] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. the 18th ACM International Conference on Multimodal Interaction*, ACM, 2016, pp. 433–436.
[37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.
[38] N. Ikezawa, T. Okamoto, Y. Yoshida, S. Kurihara, N. Takahashi, T. A. Nakada, and H. Haneishi, "Toward an application of automatic evaluation system for central facial palsy using two simple evaluation indices in emergency medicine," *Scientific Reports*, vol. 14, no. 1, 3429, 2024.
[39] L. Teijeiro-Mosquera and J. L. Alba-Castro, "Performance of active appearance model-based pose-robust face recognition," *IET Computer Vision*, vol. 5, no. 6, pp. 348–357, 2011.
[40] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. the IEEE International Conference on Computer Vision*, 2015, pp. 3676–3684.
[41] P. Kaur, K. Krishan, S. K. Sharma, and T. Kanchan, "Facial-recognition algorithms: A literature review," *Medicine, Science and the Law*, vol. 60, no. 2, pp. 131–139, 2020.
[42] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of Face Recognition*, 2011, pp. 487–519.
[43] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
[44] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proc. the 10th European Conference on Computer Vision, Marseille*, France: Springer Berlin Heidelberg, 2008, pp. 504–513.
[45] A. Wimmer, G. Soza, and J. Hornegger, "A generic probabilistic active shape model for organ segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, Springer Berlin Heidelberg, 2009, pp. 26–33.
[46] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao, "Integrating parametric and non-parametric models for scene labeling," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4249–4258.
[47] J. Zhang, Y. Deng, Z. Guo, and Y. Chen, "Face recognition using part-based dense sampling local features," *Neurocomputing*, vol. 184, pp. 176–187, 2016.
[48] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, pp. 177–190, 2014.
[49] B. Huang, Z. Wang, G. Wang, K. Jiang, K. Zeng, Z. Han, X. Tian, and Y. Yang, "When face recognition meets occlusion: A new benchmark," in *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4240–4244.
[50] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.
[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.
[52] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3642–3649.
[53] X. Mei *et al.*, "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 8, 963, 2019.
[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
[55] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
[56] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
[57] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. the 18th ACM International Conference on Multimodal Interaction*, ACM, 2016, pp. 433–436.
[58] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.
[59] J. Deng, J. Guo, and S. Zafeiriou, "Single-stage joint face detection and alignment," in *Proc. the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 1–4.
[60] F. Nan, W. Jing, F. Tian, J. Zhang, K. M. Chao, Z. Hong, and Q. Zheng, "Feature super-resolution based facial expression recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol. 236, 107678, 2022.
[61] O. Topsakal, J. Glinton, M. I. Akbas, and M. M. Celikoyar, "Open-source 3D morphing software for facial plastic surgery and facial landmark detection research and open access face data set based on deep learning (Artificial Intelligence) generated synthetic 3D models," *Facial Plastic Surgery & Aesthetic Medicine*, vol. 26, no. 2, pp. 152–159, 2024.
[62] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Detecting and aligning faces by image retrieval," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3460–3467.
[63] A. H. Hasan, A. A. Yasir, and M. J. Hayawi, "Driver drowsiness detection based on the DenseNet 201 model," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 13, pp. 3682–3692, 2021.
[64] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, 2018.
[65] Y. Wu, Z. Wang, and Q. Ji, "Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3452–3459.
[66] Y. Wu, Z. Wang, and Q. Ji, "A hierarchical probabilistic model for facial feature detection," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1781–1788.
[67] X. Zou, S. Zhong, L. Yan, X. Zhao, J. Zhou, and Y. Wu, "Learning robust facial landmark detection via hierarchical structured ensemble," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 141–150.
[68] F. Ma, B. Sun, and S. Li, "Robust facial expression recognition with convolutional visual transformers," arXiv preprint, arXiv:2103.16854, 2021.
[69] F. Nan, W. Jing, F. Tian, J. Zhang, K. M. Chao, Z. Hong, and Q. Zheng, "Feature super-resolution based facial expression recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol. 236, 107678, 2022.
[70] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.
[71] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386–391.
[72] J. Zhang, Y. Deng, Z. Guo, and Y. Chen, "Face recognition using part-based dense sampling local features," *Neurocomputing*, vol. 184, pp. 176–187, 2016.

[73] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q. V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, 103525, 2022.
[74] R. Rothe, M. Guillaumin, and L. V. Gool, "Non-maximum suppression for object detection by passing messages between windows," in *Proc. the 12th Asian Conference on Computer Vision*, Singapore: Springer International Publishing, 2015, 290–306.
[75] M. Hassaballah and K. M. Hosny, "Recent advances in computer vision," *Studies in Computational Intelligence*, vol. 804, pp. 1–84, 2019.
[76] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.
[77] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.
[78] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas, "Consensus of regression for occlusion-robust facial feature localization," in *Proc. the 13th European Conference, Zurich*, Switzerland: Springer International Publishing, 2014, pp. 105–118.
[79] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.
[80] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2385–2392.
[81] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proc. the IEEE International Conference on Computer Vision*, 2015, pp. 3658–3666.
[82] O. A. Muhi, M. Farhat, and M. Frikha, "Transfer learning for robust masked face recognition," in *Proc. 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2022, pp. 1–5.
[83] S. Bhatlawande, S. Shilaskar, T. Gadad, S. Ghulaxe, and R. Gaikwad, "Smart home security monitoring system based on face recognition and android application," in *Proc. 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2023, pp. 222–227.
[84] A. A. S. Alashbi and M. S. Sunar, "Occluded face detection, face in niqab dataset," in *Proc. Emerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing 4*, 2020, pp. 209–215.
[85] B. Thaman, T. Cao, and N. Caporusso, "Face mask detection using mediapipe facemesh," in *Proc. 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2022, pp. 378–382.
[86] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile GPUs," arXiv preprint, arXiv:1907.05047, 2019.
[87] N. Bayar, K. Güzel, and D. Kumlu, "A novel blazeface based pre-processing for MobileFaceNet in face verification," in *Proc. 2022 45th International Conference on Telecommunications and Signal Processing (TSP)*, 2022, pp. 179–182.
[88] Z. Pan, Y. Wang, and S. Zhang, "Joint face detection and facial landmark localization using graph match and pseudo label," *Signal Processing: Image Communication*, vol. 102, 116587, 2022.
[89] D. Miller, E. Brossard, S. Seitz, and I. Kemelmacher-Shlizerman, "MEGAface: A million faces for recognition at scale" arXiv preprint, arXiv:1505.02108, 2015.
[90] J. Wang, Y. Yuan, and G. Yu, "Face attention network: An effective face detector for the occluded faces," arXiv preprint, arXiv:1711.07246, 2017.
[91] Z. Yuan, "Face detection and recognition based on visual attention mechanism guidance model in unrestricted posture," *Scientific Programming*, vol. 2020, pp. 1–10, 2020.
[92] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: A real-time face detector," *The Visual Computer*, vol. 37, pp. 805–813, 2021.
[93] A. A Alashbi, M. S. Sunar, and Z. Alqahtani, "Deep-learning-CNN for detecting covered faces with Niqab," in *Proc. Journal of Information Technology Management, Special Issue: 5th International Conference of Reliable Information and Communication Technology (IRICT 2020))*, 2020, pp. 114–123.
[94] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu, and X. Wang, "Yolo-facev2: A scale and occlusion aware face detector," arXiv preprint, arXiv:2208.02019, 2022.
[95] S. Janahiram, A. Alsadoon, P. W. C. Prasad, A. M. S. Rahma, A. Elchouemi, and S. A. Senanayake, "Detecting occluded faces in unconstrained crowd digital pictures," in *Proc. 2016 First International Conference on Multimedia and Image Processing (ICMIP)*, 2016, pp. 5–9.
[96] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. L. Chang, M. G. Yong, J. Lee, and W. T. Chang, "Mediapipe: A framework for building perception pipelines," arXiv preprint, arXiv:1906.08172, 2019.
[97] M. Opitz *et al.*, "Grid loss: Detecting occluded faces," in *Proc. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam*, The Netherlands, 2016, vol. 14.
[98] Z. R. Alqahtani, M. S. Sunar, and A. A. Alashbi, "Landmark localization in occluded faces using deep learning approach," in *International Conference of Reliable Information and Communication Technology*, 2020, pp. 1023–1029.
[99] H. Guo, J. Liu, Z. Xiao, and L. Xiao, "Deep CNN-based hyperspectral image classification using discriminative multiple spatial-spectral feature fusion," *Remote Sensing Letters*, vol. 11, no. 9, pp. 827–836, 2020.
[100] M. Vajgl, P. Hurtik, and T. Nejezchleba, "Dist-YOLO: Fast object detection with distance estimation," *Applied Sciences*, vol. 12, no. 3, p. 1354, 2022.
[101] J. Cai, H. Han, J. Cui, J. Chen, L. Liu, and S. K. Zhou, "Semi-supervised natural face de-occlusion," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1044–1057, 2020.
[102] X. Ding, S. Zhang, L. Kang, and C. Liu, "Occlusion Adaptive Deep Network," *IEEE Transactions*, vol. 46, pp. 10–20, 2020.
[103] H. Liu, W. Zheng, C. Xu, T. Liu, and M. Zuo, "Facial landmark detection using generative adversarial network combined with autoencoder for occlusion," *Mathematical Problems in Engineering*, vol. 2020, pp. 1–8, 2020.
[104] Yin and L. Chen, "FaceOcc: A diverse, high-quality face occlusion dataset for human face extraction," arXiv preprint, arXiv:2201.08425, 2022.