Decoding Face Attributes: A Modified AlexNet Model with Emphasis on Correlation-Heterogeneity Relationship Between Facial Attributes

Abdelaali Benaiss[1], Otman Maarouf[2], Rachid El Ayachi[3], Mohamed Biniz[4], Mustapha Oujaoura[5]
Department of Computer Science-Faculty of Sciences and Technics-Laboratory TIAD,
Sultan Moulay Slimane University, BP: 592, Beni Mellal, Morocco[1, 3]
Department of Computer Science-Faculty of Sciences, Agadir Ibn Zohr University, BP 8106, Agadir, Morocco[2]
Department of Computer Science-Polydisciplinary Faculty-Laboratory LIMATI,
Sultan Moulay Slimane University, BP 523, Beni Mellal, Morocco[4]
Department of Computer Science, Network & Telecom (GIRT)-Laboratory of Informatics Mathematics & Communication Systems (MISCOM)-National School of Applied Sciences (ENSA), Cadi Ayyad University, BP 575, Safi, Morocco[5]

***Abstract*—Face attribute estimation has several applications in computer vision, biometric systems, face verification /identification and image retrieval. The performance of face attribute estimation has been improved by using machine learning algorithms. In recent years, most algorithms have addressed this problem in multiple binary problem. Specifically, CNN-based approaches, which we can divide them into two classes; shared features and parts-based approaches. In shared features approach, the model uses two types of CNNs: one for feature extraction succeed by another one, for attribute classification. In the parts-based approaches, the approaches split the face image into multiple parts according to the geometric position of each attribute and train a CNN model for each part of the face. However, the shared features approach can handle attributes correlation but ignored attribute heterogeneity and gain in training time. On the other hand, the parts-based approaches can handle attributes heterogeneity but ignore attributes correlation and need more time in the training set compared with a shared feature approach. In this work, we propose a face attribute estimation method, which combined shared features and a parts-based approach into one model. Our model splits the input face image into five parts: whole image part, face part, face upper part, lower part, and nose part. In the same manner, the face attributes are subdivided into five groups according to the geometric position in the face image. We train shared feature model for each part, and we proposed an algorithm for feature selection task followed by AdaBoost algorithm to handle attribute classification task. Through a set of experiments using the LFWA and IIITM Face Emotion datasets, we demonstrate that our approach shows higher efficiency of face attribute estimation compared with the state-of-the art methods.**

***Keywords*—Face attribute estimation; biometrics; Convolutional Neural Network (CNN); face verification; computer vision**

I. INTRODUCTION

The face recognition systems are the most attractive topics in the biometrics systems because of their convenience, hygiene, and low cost. Specifically, the face as the objective signal can be acquired in a contactless manner without requiring any special equipment. Moreover, due to the multiple advantages provided by the face recognition system, these kinds of systems are the most used in industry, combined with fingerprint-based systems as personal authentication for smartphones, security gates, payment services, computer human interaction, etc. In addition, the explosive development of Convolutional Neural Networks (CNNs) models, Graphic Units Process (GPU) material and opensource frameworks (e.g, Keras, PyTorch, Caffe, Dlib, etc), the Convolutional Neural Networks (CNNs) have replaced most traditional methods for face recognition problems. To further illustrate the versatility and application of convolutional neural networks (CNNs) in various domains, it is noteworthy to mention recent advancements in related fields. For instance, the work on automatic translation from English to Amazigh using transformer learning [1] demonstrates the adaptability of deep learning models in language processing tasks. Similarly, the recognition of Tifinagh characters using optimized convolutional neural networks [2] highlights the effectiveness of CNNs in character recognition tasks. These studies underscore the broad applicability and potential of CNNs, reinforcing their relevance in face recognition systems as well. To address the great demand for face recognition on face retrieval systems [3], video surveillance environments [4], and criminal investigation protocols [5]. There are multiples studies aimed at improving the performance of face recognition by improving the face attribute estimation.

For improving the face attributes task, the research community provides a number of face databases with their attributes like: CelebA [6] and LFWA [7] datasets, at each dataset, we have a number of face attributes (40/73 attributes) that describe the biological characteristics of the face (e,g ; color, shape and texture) or give information about a subject, weather it is wearing ornaments such as glasses or earrings or not. In addition, the works in study [8] and [9] make an assumption about relationship between face attributes based on the co-occurrence probabilities of two attributes in some databases. To illustrate, the attribute ‘Male’ has a high probability of attributes such as ‘Goatee’, while ‘Female’ has high probability of attributes such as ‘Heavy Makeup’ and ‘Wearing Lipstick’. On the contrary, the lower probability of occurrence for some attributes, the more likely those attributes are to be heterogenous. In the same manner, the works [10], [11], [12] and

[13] show that the relationship between attributes is based on the face parts. For example, 'Black Hair' and 'Blond Hair' are attributes related to 'Hair', which has a position in the upper face part and 'Big Nose' with 'Pointy Nose' are attributes related to 'Nose', which has a position in middle face part, where 'Double Chin' and 'Goatee' are attributes related to the low face part.

In general, face attribute estimation approaches consist of three processes: face detection, feature extraction and attribute classification. Among these processes, feature extraction and attributes classification are the most important, since they have the greatest impact on the estimation accuracy. To deal with the challenging problem of feature extraction and attribute classification, multiple works have been published about this object. In all published works, specifically CNN-based methods, they have a significant impact on feature extraction in terms of accuracy. Some works like in [14] and [10], use the same features for estimating multiple attributes without considering the attribute heterogeneity, and some others are limited to estimating a single attribute [15], or training a separate model for each face attribute without considering the attribute correlation [3]. Although, all previous works have a challenging problem when dealing with non-frontal face images, low image quality, occlusion, and pose variations. The region of interest (ROI) is often suitable and it may cover only a small part of the image, while the face image is dominated by the effects of position pose and viewpoint. Since then, Part-based methods in [5], [10], [11] and [16] have recently become the most popular approaches to dealing with pose variation, occlusion and low image quality.

In this paper, we present a novel method which combined three principal approaches; multi-task, part-based and attributes relationship to achieves better results on face attributes estimation. However, the contributions of this research aim to present;

1) Image denoising and online data augmentation with a specific technic, which get the experimental condition close to real-world scenarios.

2) Data balanced process to handle the challenge of minority class in databases.

3) Face split process combined with attributes subgrouping to handle respectively; head pose and attributes heterogeneity in same task.

4) An algorithm to handle feature selection step, followed by Adaboost classifier to addressed the attributes correlation and achieved the final estimation of each attribute.

The remainder of this paper is organized as follows; Section II gives a brief overview of the related work on face attributes estimation. The proposed approach with the baseline networks is outlined in Section III. Section IV displays the different experiments that have been conducted to achieve better results than competing methods. A discussion of the results is addressed in Section V. Finaly, we conclude our work in Section VI.

II. Related Work

A. Multi-Tasks Learning Approaches

Multi-task learning (MTL) in facial attribute estimation consists of training a model to achieve multiple attribute prediction using, in some cases, shared representation approaches. For instance, the work in study [6], proposes two CNNs; LNet and ANet; the first one is pre-trained for face localization and the second one is pre-trained for attribute prediction. Those two CNNs are succeeded by an SVM classifier for attribute classification. Also, the approach in study [16] proposed Deep Multi-task Learning (DMTL) network consists of learning a modified AlexNet with a batch normalization (BN) layer inserted after each Conv Layer for the shared part of model, followed by, a category-specification block for attributes estimation. This method can handle heterogeneous information about attributes. In addition, this shows superior performance compared to state-of-the-art methods on public benchmarks. Moreover, another architecture is designed in study [9], where a ResNet50 Network is adapted as the backbone architecture, they take the first 46 layer as Shared Layers succeed by Task-specific Layers consist of two branches res5C1 and res5C1 corresponding to smile and gender prediction (res5C1and res5C2 are two Residual Blocks of ResNet50). Those two branches are passed by attention block (coined as Att_C) to represent the dependency between smile and gender attributes. Furthermore, two novel approaches have been proposed in study [8], the first one called HyperFace uses AlexNet as backbone of model, while the second one called HyperFace-ResNet is based on ResNet-101. Those two architectures perform well in the face detection, landmarks localization, pose estimation and gender recognition on various public available unconstrained datasets, those approaches show a novel hypothesis about fusing the intermediate layers of the backbone structure. In other words, the introduction of multi-tasks learning approaches in CNN-based models shows better results compared with single-task learning approaches in the terms of attribute correlation.

B. Parts-Based Approaches

In parts-based approaches, the object image and face image are split into small parts and each part is taken as input of the feature extraction process. For example, the work in study [10], proposed Pose Aligned Networks for Deep Attribute Modeling (PANDA), which divide person image into small parts coined as Poselets and each one is passed by a trained CNN. The top-level activations of all CNNs are concatenated to obtain a pose-normalized deep representation and then Linear SVM classifier is trained for attribute classification. In some recent works, the facial attributes are clustered in many groups according to their location in facial parts and relationship in terms of correlation and heterogeneity, before the process of feature extraction and attribute classification. However, in study [5] they split the 40 face attributes into six or nine groups, and they proposed a multi-task deep CNN (MCNN) combined with an auxiliary network (AUX) to achieve the final estimation for each attribute group. Based on the MCNN-AUX model, the study [11] proposed Partially Shared Multi-task CNN (PS-CNN), since they split all the 40 attributes into four attribute groups including Upper, Middle, Lower, and Whole Image. Then the attributes classification of each group can be considered as individual attribute learning task. This method shows the promise results on two databases CelebA [6] and LFWA [7]. In short, parts-based approaches still the better approaches to addressing heterogenous attributes and face parts occlusion.

C. Attributes Grouping

Inspired by the fact that information contained in a face image is geometrically distributed on face parts like; eyes, nose, mouth, etc. For this reason, many recent works are based on this assumption. They have been adding attribute grouping approach as pre-processing step before feature extraction in attribute estimation process. For example, the work in study [8] split 40 attributes of LFWA and CelebA dataset into six subgroups based on attribute color and texture. This approach shows better results compared with other work like PANDA [10] with more than 10% of improvement in term of accuracy detection. Besides, in [5] the attributes have been separated into nine groups; Gender group, Nose group, Eyes group, Face group, AroundHead group, FacialHair group, Cheecks group and Fat group, each one contains a number of attributes according to their facial parts. Furthermore, the Faceness-Net approach proposed in study [12] separate attributes of CelebA dataset into five groups, similarly to [5] and [8]. This work proposes: Hair group, Eye group, Nose group, Mouth group and a Beard group. Each group is represented by branch in the model, those branches was summed into a face label map, which clearly suggests face's location in the image. Moreover, The Faceness-net approach shows better results compared to state-of-the-art methods in term of face detection problems (near 98.05 % AP = average precision on AFW dataset and 92.11 % AP on PASCAL dataset). Therefore, in [6] the subject face in dataset has been split into 14 parts and in the same manner, the 40 face attributes have been separated into 14 subsets. Each part of the face has a subset of attributes based on the visibility value of this attributes in this part of face (visibility value > τ = visibility threshold). In short, the attribute grouping step handles the attribute correlation and reduces model training time and hyperparameters, but still handless the attribute heterogeneity. On the other hand, in study [17] the group of attributes is subdivided into four groups based on attribute categories (nominal, ordinal, holistic and local). The approach trains four sub-network types according to each subgroup of attributes. This approach can handle attribute heterogeneities compared with the previous works in studies [5], [6], [8], [12], [16], and shown better results on average accuracy reach 93% on CelebA dataset and 86.3% on LFWA dataset. In conclusion, all the previous works prove that the attributes grouping step as pre-processing step can guide the model to achieve expected results, in terms of attribute relationship.

On the whole, the goal of this work is to combine an attribute grouping approach, parts-based approach and multi-tasks learning in order to make a novel model that can handle attribute correlation and heterogeneity at the same time and reach better results compared with state-of-the-art methods.

III. Proposed Method

In the majority approach's, the first step consists of taking the crop face from image and normalized it based on facial landmarks or splitting it to small parts before attacking CNN structure. In two cases, the crop face has less information about some attributes like; Wear Hat, Bald and 5 O 'Clock Shadow. Since, we propose an approach consists of five CNNs models, one of them, specifically design to extracted the information about Hair, Face background and Neck, at the same time, the other fours CNNs are specialized in facial details. More details about the proposed approaches are giving in following sections.

A. Face Split and Attributes Grouping

To split face image into three parts, we have used the position of facial keypoints return by MTCNN method presented in [17] (as shown in Fig. 1). The segmentation of face image and face attributes are pre-processing step of input image before it has been processed by CNNs models, those three parts are coined in this work as UP for upper-part, LP for lower-part and NP for nose-part indicated in Fig. 1, respectively, by (A), (B) and (C). (E) represent the face part.

In this approach, we have five CNNs coined as UP-Net (for upper part convolutional neural network), LP-Net (for lower part convolutional neural network), NP-Net (for nose part convolutional neural network), FP-Net (for face bounding box convolutional neural network) and WI-Net (for whole face image convolutional neural network). Each CNN take a specific input image to predicts a subset of specific attributes. Therefore, UP-Net take upper part of face as input and predict their corresponding attributes (attributes with index number #2, #4, #6, #13, #16 and #24), LP-Net take lower part of face as input to predict the attributes with index number (#7, #17, #22, #23, #25, #32 and #37). Where, NP-Net predict the attributes with index number (#8 and #28) according to nose part of face and FP-Net predict the attributes with index number (#5, #19, #20, #21, #26, #27, #30 and #40) which contain the information about face region. Finally, WI-Net receive whole face image as input and given the prediction of attributes with index number (#1, #3, #10, #11, #12, #14, #15, #18, #29, #31, #33, #34, #35, #36, #38 and #39). The label of each index number is described in Table I.

Indeed, certain segments are more effective at predicting a subset of attributes than others. For example, we can expect that upper-part (UP) would contain information about the person being bald, wearing hat or having certain types and color of hair. Therefore, this part of face can still predict attributes related to eyes, eyebrows and hair, at the same time, the lower-part (LP) can predict attributes related to mouth, goatee and moustache. Where, nose-part give information about nose. In short, we join each part with their corresponding attributes.

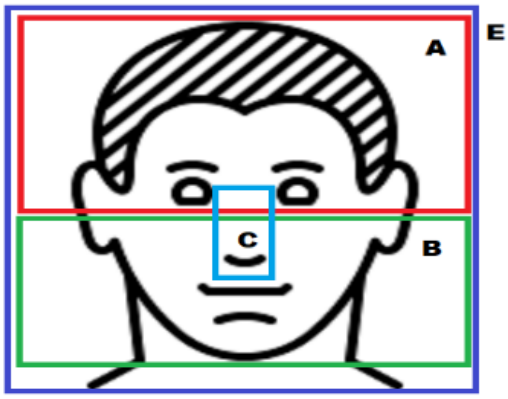

Fig. 1. The face image is divided into four parts; (E) face part, (A) upper part, (B) lower part, (C) nose part. The segmentation of face has been made based on keypoints return by MTCNN method in [17] (Best viewed in color).

TABLE I. FACE ATTRIBUTE LABELS DEFINED IN LFWA [7] DATASET

Index	*Attribute*	*Index*	*Attribute*
#1	5 O 'Clock Shadow	#21	Male
#2	Arched Eyebrows	#22	Mouth Slightly Open
#3	Attractive	#23	Mustache
#4	Bags Under Eyes	#24	Narrow Eyes
#5	Bald	#25	No Beard
#6	Bangs	#26	Oval Face
#7	Big Lips	#27	Pale Skin
#8	Big Nose	#28	Pointy Nose
#9	Black Hair	#29	Receding Hairline
#10	Blond Hair	#30	Rosy Cheeks
#11	Blurry	#31	Sideburns
#12	Brown Hair	#32	Smiling
#13	Bushy Eyebrows	#33	Straight Hair
#14	Chubby	#34	Wavy Hair
#15	Double Chin	#35	Wearing earrings
#16	Eyeglasses	#36	Wearing Hat
#17	Goatee	#37	Wearing Lipstick
#18	Gray Hair	#38	Wearing Necklace
#19	Heavy Makeup	#39	Wearing Necktie
#20	High Cheekbones	#40	Young

B. Attributes Correlation and Heterogeneity

The face verification and image search fields are the first methods has been introduced image attributes as descriptor. They used a subset of 40 binary attributes to describe each face in dataset (see Table I). They later extended the number of attributes with addition ones to achieve 73 binary attributes. In the recent years, more approach's shown attributes dependency and non-dependency [5] to improve accuracy of attributes detection. Further, as shown in the Fig. 2, the attribute with clear color square shown strong positive correlation between their two corresponding attributes respectively, in x-axis and y-axis. On the other hand, attribute with dark color square shown low correlation (heterogeneity). To draw the image in Fig. 2, we have been calculated co-occurrence matrix of each attribute index in LFWA dataset and each square in co-occurrence matrix present the probability to have a specific two attributes in same image. For example, the attribute No_beard (#25) has a strong correlation with Heavy_Makeup (#19), Wearing_Earrings (#35) and Wearing_Lipstick (#37) which has a probability more than 90% (shown by clear color square in Fig. 2) even though the attribute No_beard (#25) has weak correlation with Mustache (#23) which has a probability near 0% (shown by dark color square). In short, the proposed approach is motivated by previous assumptions and it has been introduced in the task specification process which well be detailed in following sections.

C. Network Architecture

Different parts of the face may have different signals for each attribute and sometimes signals coming from one part cannot infer certain attributes accurately. For example, the information about nose like Big_Nose (#8) (in the Nose Part of face) cannot give information about Wearing_hat (#36) (in the top of head, Whole image part). Therefore, based on these assumptions, we have been explored the advantages of part-based approach to handle the non-dependency of attributes in the shared feature process of our network and we have been explored paire-wise co-occurrence matrix of attributes to handle dependency attributes in task specification process.

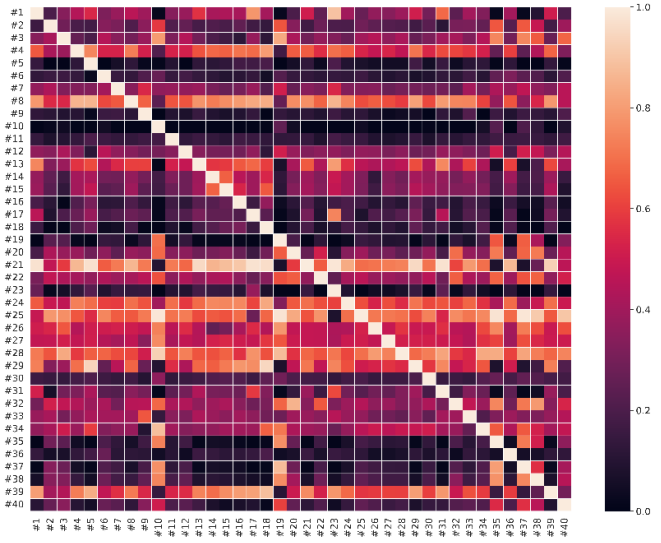

Fig. 2. Pair-wise co-occurrence matrix of the 40 face attributes (see Table I) provided with the LFWA [7] database (Best viewed in color).

Inspired from the works in [8] and [16], we have chosen to work with AlexNet as base structure of our model because it has a good result in the term of accuracy on challenging database and faster than GoogleNet and has a small structure compared with VGG and ResNet models. However, the AlexNet consists of five convolutional layers, three max pooling layer and the three Full connected layers. Based on work in study [18], we have been inserted a batch normalization layer after each convolution layer to avoiding overfitting problem. The original and modified AlexNet are shown in the Fig. 3.

Inserting Batch Normalization Layer to AlexNet model. Regularization stands out as an effective technique to combat overfitting issues. Incorporating Batch Normalization (BN) [14] between convolution layers can enhance model stability and regularization. The BN layer normalizes the output from the preceding activation layer by subtracting the mini-batch mean and dividing by its standard deviation. Essentially, this normalization process adjusts the means and variances of layer inputs by introducing two trainable parameters per layer. Additionally, BN reduces the network's sensitivity to the initialization of individual layers, enabling the use of higher learning rates. In our approach, we set a fixed learning rate of 0.001. The Fig. 3 shown more details about original and modified AlexNet model used in this work. The batch normalization layers add into AlexNet model are motioned in Fig. 3 by blue square. C shown the number of attributes returned by AlexNet block according to input facial part. The numbers denote the kernel size, cardinality and features maps for given layer.

Overall structure. The proposed approach consists of denoising process followed by image splitting process to get five parts (deemed as WI, FP, UP, LP and NP). Those five parts are resized into 224x224 size image whose have been taken as input of five subnets (coined as WI-Net, FP-Net, UP-Net, LP-Net and NP-Net). Each subnet of them has a modified AlexNet described in Fig. 3 as base structure.

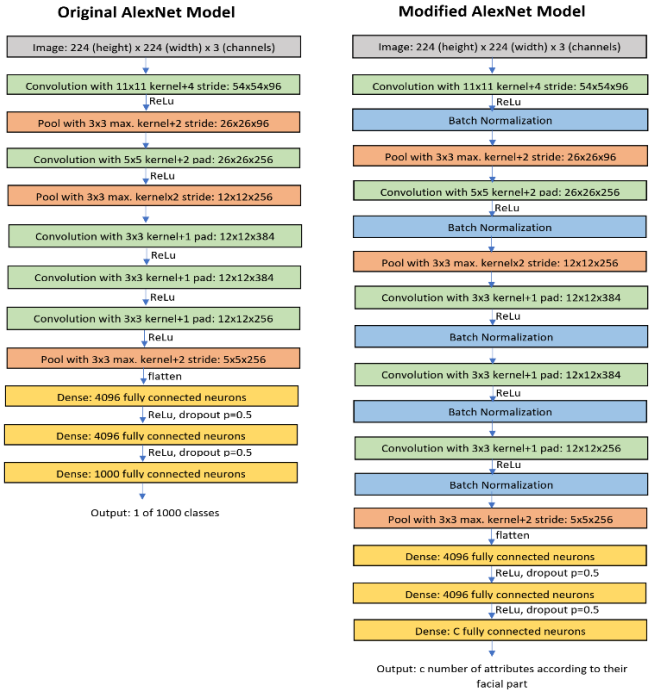

Fig. 3. The original and modified AlexNet used in backbone of our structure (Best viewed in color).

Since, the batch normalization (BN) layers add into AlexNet backbone adjusted the means and variations between the Conv. layers and make the main structure more stable in the learning process. Moreover, at the output of each subnet, we have a specific number of face attributes see Subsection A for more details (17 face attributes for WI, 8 for FP, 6 for UP, 7 for LP and 2 NP). To show the correlation between attributes, we have been proposed a coding algorithm repose on occurrence matrix to handle face attributes correlation see subsection E for more details. The feature selection has been followed by Adaboost classifier to achieve the final estimation for each attribute, see Fig. 4 to have an overview of our proposed method.

D. Simulation of Real-World Scenarios by Data Augmentation

In real-world, there are no universal patterns for facial attributes across all individuals and all datasets present in the literature are limited and don't accurately mirror real-world scenarios based on this assumption, a more realistic attributes estimation system should be trained on dataset prepared with data augmentation process to achieve real-world conditions.

Data augmentation, as discussed in study [19] and study [20], serves as a regularization technique by introducing synthetic images to the neural network, simulating more realistic conditions and viewpoints. This approach helps mitigate overfitting issues stemming from limited datasets. Various transformations can be applied to create additional modified images, including translation, zooming, brightness adjustments, and more. These subtle variations enable the model to generalize better to unseen data and enhance its robustness when exposed to slightly altered images. In our study, we adopted online augmentation, applying transformations to images as batches are processed during training. This approach accelerates the training process and eliminates the need to store augmented data alongside the original dataset in memory, as required in offline augmentation as per the protocol in study [19]. Specifically, our data augmentation involved shifting image appearance by 0.2 of the image height, adjusting brightness within a range of values between 0.1 and 0.2, and applying zooming. Fig. 5 illustrates the results of the data augmentation process on an image from the LFWA [7] database. The image in the left side of Fig. 5 is the original image after denoising process which loaded from LFWA [7] dataset and the group of images in right side are the data augmentation images after different transformation functions.

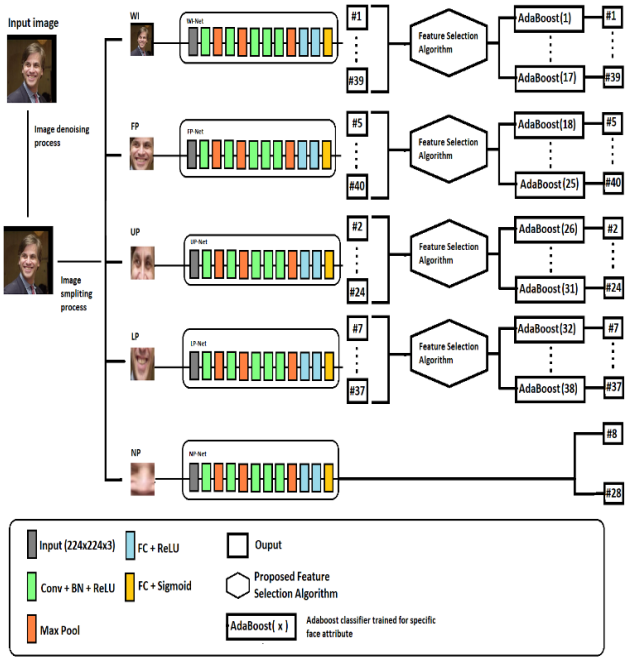

Fig. 4. Overview of the proposed architecture. (Best viewed in color).

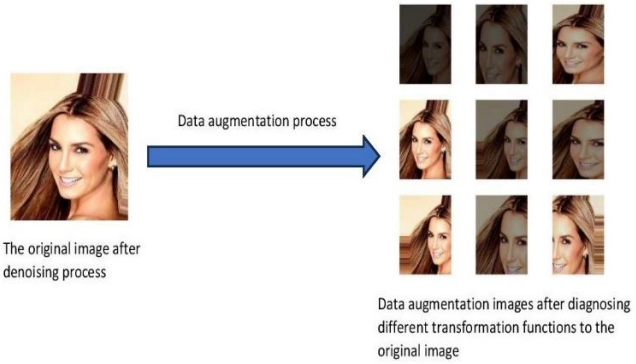

Fig. 5. Data augmentation process of each image in dataset (Best viewed in color).

In addressing this aspect of our methodology, we utilized the Keras ImageDataGenerator class, which offers a convenient and efficient method for image augmentation. This class provides a range of augmentation technics, including standardization, rotation, shifts, flips, brightness adjustments, and more. However, the primary advantage of employing the Keras ImageDataGenerator class is its capability for real-time data augmentation. This means it generates augmented images on-the-fly during the training phase of your model. In short, by utilizing this class, images are loaded in batches, which means saving more memory in training process.

E. Task-Specification Process

To achieve task-specification process, many approaches has been shown in the literature. For example, in study [6] an automatic attributes grouping method has been proposed which take columns of weights matrix returned fully-connected layer ANet as a decision hyperplane to partition the negative and positive samples of attribute. By sample applying k-means to these vectors, the clusters show clear grouping patterns. These are used as system features whish passed by SVM to achieve final attribute detection. Furthermore, the work in study [21] used Multi-Kernels Maximum Mean Discrepancies (MK-MMD) proposed in study [22] to show the correlation between features returned by MNet and TNet. Another approach shown with PANDA [10] which take a signal returned by each poselet as pose normalization of each image part and used a linear classifier (Logistic Regression in this case) to achieve a Task-specification process. In addition, FaceTracer [15] used SVM algorithm which deemed by Attribute-Tuned Global SVM to achieve final attribute detection. In short, the task-specification process consists of two steps feature selection combined with an algorithm of classification.

In contrast with the previous approaches, we have been proposed an encoding algorithm for features selection step and AdaBoost algorithm to achieve final prediction of each attribute. The proposed algorithm has been presented in the following section and the based co-occurrence matrices of FP-Net, UP-Net and LP-Net have been presented respectively, by (a), (b) and (c) in Fig. 6. The proposed Algorithm used those matrices to achieve feature selection step.

Algorithm 1: LPBT← (find list of attributes with probability greater than specific threshold)

```
Initialize
I  ← Specific attribute
M ← Matrix of occurrence
T  ← Threshold
   N  ← Number of rows in M
Compute
|  For i ← 0 to i ← N-1 do
|  Update
|        |     Update and analyze
|        |  |     If M[I][i] >= T then
|        |  |        L[C] ← i
|        |  |     C ← C + 1
|  End   |  |     End
|  End
```

Algorithm 2: CAGT← Calculate the margin of error between list attributes return by subnet and ground truth

```
Initialize
I  ← Specific attribute
L  ← List returned by Algorithm 1 (LPBT function)
N  ← lent of L list
T  ← Matrix values return by subnet (WI-Net, FP-Net, UP-Net or LP-Net)
P  ← Matrix of ground truth values of attributes group according to each subnet.
R  ← Number of rows in T matrix
C  ← Number of columns in T matrix
Compute
|  For i ← 0 to i ← N-1 do
|  Update
|        |     Update and analyze
|        |  |     If M[I][i] >= T then
|        |  |        L[C] ← i
|        |  |     C ← C + 1
|  End   |  |     End
|  End
|
|  For i ← 0 to i ← N-1 do
|  Update
|           Update and analyze
|              TV[i] ← T[:][L[i]]
|  End
|  For i ← 0 to i ← R-1 do
|  Update
|        |     Update and analyze
|        |        For j ← 0 to j ← C-1 do
|        |        Update
|        |              |     Update and analyze
|        |              |     If TV[i][j] == 0 then
|        |              |        TV[i][j] ← -1
|        |        End   |     End
|        |        End
|  End
|  For i ← 0 to i ← C-1 do
|  Update
|           Cpt ← 0
|        |     Update and analyze
|        |        For j ← 0 to j ← R-1 do
|        |        Update
|        |              |     Update and analyze
|        |              |     Cpt ← Cpt + TV[j][i]
|        |        End   |     End
|        |        CV[i] ← Sigmoid(Cpt)
|  End   |     End
```

Algorithm 3: we have in output of this algorithm the list shown the relationship between attributes.

```
Initialize
M ← Co-occurrence matrix in the output of each subnet (WI-Net, FP-Net, UP-Net or LP-Net)
T ← Matrix values return by each subnet
P ← Ground truth matrix for each subnet
    R ← Number of rows in T matrix
Compute
    For i ← 0 to i ← R-1 do
    Update
        Update and analyze
            Accu ← 0
            For j ← 0 to j ← 9 do
            Update
                    Update and analyze
                    S ← j/10
                    CO ← LPBT(i, M, S) /* Algorithm 1*/
                        If Accu < CAGT(i, CO, T, P) then
                            Accu ← CAGT(i, CO, T, P)
                            Find ← CO /*Algorithm 2*/
            End         End
            F[i] ← Find
    End
```

Effectively, in the training set, we have been taken the output of each subnet and we have been applied this algorithm to show the correlation between face attributes. See Table II for more details.

TABLE II. ATTRIBUTES IN FORT CORRELATION WITH EACH ATTRIBUTE RETURNED BY PROPOSED ALGORITHM FOR FEATURE SELECTION STEP IN OUR MODEL

Attribute index	***Attributes in correlation***	***Attribute index***	***Attributes in correlation***
#1	#1, #29, #31, #33, #39	#21	#21, #26, #27
#2	#2, #4, #6, #13, #16, #24	#22	#22, #25, #32
#3	#3, #33, #34, #35, #38	#23	#7, #17, #23
#4	#4, #13, #24	#24	#4, #13, #24
#5	#5, #21, #26, #27	#25	#22, #25, #32
#6	#2, #6, #24	#26	#19, #21, #26, #27
#7	#7, #22, #25	#27	#19, #21, #26, #27, #40
#8	#8	#28	#28
#9	#3, #9, #33, #35, #38, #39	#29	#1, #14, #15, #18, #29, #31, #33, #34, #39
#10	#1, #3, #9, #10, #11, #12, #14, #15, #18, #29, #31, #33, #34, #35, #36, #38, #39	#30	#5, #19, #20, #21, #26, #27, #30, #40
#11	#11, #12, #29, #33, #34, #35	#31	#1, #31, #39
#12	#3, #12, #33, #35, #38	#32	#22, #25, #32
#13	#4, #13, #24	#33	#1, #9, #14, #15, #18, #29, #33, #38, #39
#14	#14, #15, #33	#34	#3, #34, #38
#15	#14, #15, #29, #33, #39	#35	#3, #35, #38
#16	#13, #16, #24	#36	#14, #33, #36, #39
#17	#7, #17, #23	#37	#7, #22, #25, #32, #37
#18	#18, #29, #38	#38	#3, #35, #38
#19	#19, #26, #27, #40	#39	#1, #15, #29, #33, #39
#20	#19, #27	#40	#19, #21, #26, #27, #40

Finally, the results returned by Algorithm for each attribute has been passed in the followed step by Adaboost classifier to achieve the final estimation.

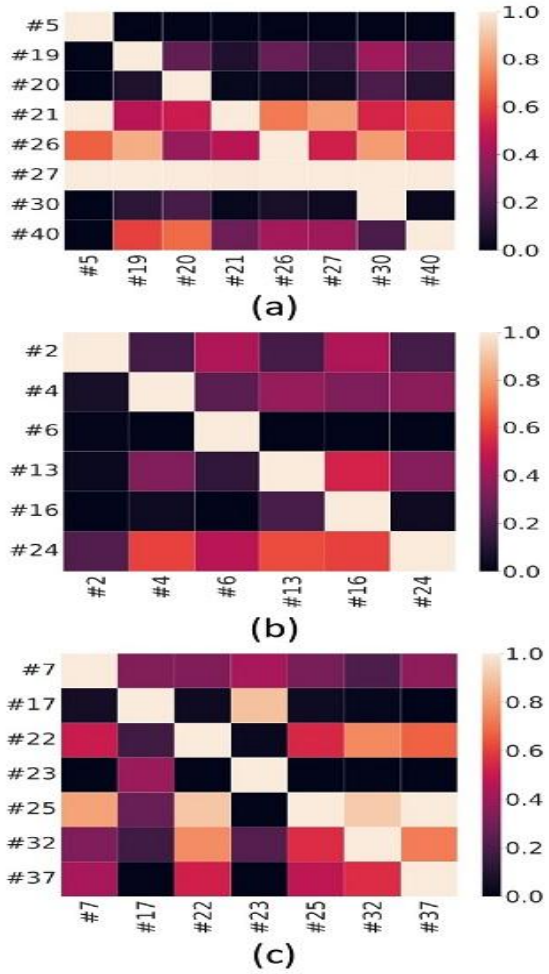

Fig. 6. The matrix of co-occurrence for each subnet in LFWA dataset. (Best viewed in color).

IV. EXPERIMENTAL RESULTS

In the experimental section, we have been shown all the experimental steps provide in this work. In first time, we have

been presented Evaluation Metrics used in this work (subsection A). In the followed subsection B, we have been described the databases used in this work to made training, validation and testing steps. The data pre-processing has been presented in subsection C, which contains image denoising, splitting and database balanced. Further, we have been shown the process to determine the values of some specific parameters of our networks in the subsection D. Finally, the performance of our approach for 40 face attributes, gender recognition and smile estimation have been shown in the subsection E.

A. Evaluation Metrics

The most common metrics for attributes estimation is Accuracy, Precision, Recall and Fi-score. Those metrics can be more representative than others metrics in du literature to evaluate the attributes estimation system, since it can be shown the difference between the estimated value and their ground truth in the statistics manner. Those metrics can be mathematically defined as following:

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FP} \quad (3)$$

$$F1-score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Where:

- True positive (TP): An instance for which both estimated and ground truth values are positive.
- True negative (TN): An instance for which both estimated and ground truth values are negative.
- False Positive (FP): An instance for which estimated value is positive but ground truth value is negative.
- False Negative (FN): An instance for which estimated value is negative but ground truth value is positive.

B. Datasets

LFWA dataset [7] is a large-scale face attribute database with 13143 images of 5.749 subjects in unconstrained environment. Each image is annotated with 40/73 attributes (see Table I more description). The images in this dataset are in color space and contain large variations in pose, expression, race, background, etc., making it challenging for face attribute estimation. Moreover, the split protocol suggested by dataset is 6263 for training and 6880 for testing. Some images from the dataset have been shown in Fig. 7.

Fig. 7. Samples from LFWA dataset (Best viewed in color).

More descriptions about LFWA dataset have been illustrated in Fig. 8.

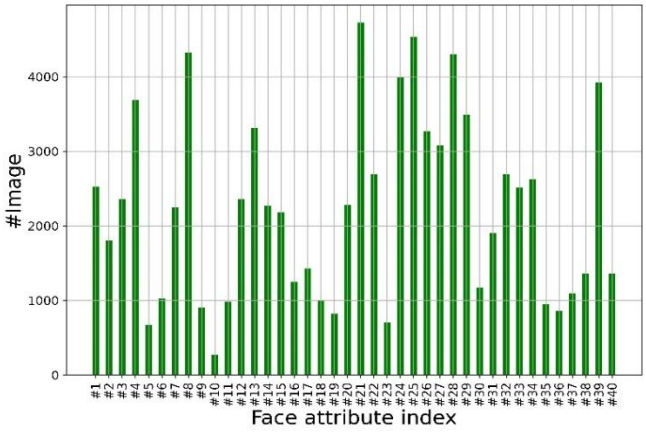

Fig. 8. Details about data distribution in LFWA dataset (Best viewed in color).

Strating from the Fig. 8, the LFWA dataset suffer from large imbalanced in data distribution. For example; number of man subjects in data reach 4727 when female subjects equal to 2153, which make gender estimation task harder for female subjects in compared with man subjects. In the same manner, smiling people in train part of dataset equal to 2687 when no smiling people reach 4193, which make this task hard in the training process. To deal with this problem, we have adopted SMOTE [23] algorithm to balance a training data for each subnet (WI-Net, FP-Net, UP-Net, LP-Net and NP-Net). More details have been described in the followed Subsection C.

The IIITM Face Emotion dataset [24] originates from the IIITM Face Data and comprises 1,928 images from 107 participants, including 87 males and 20 females. These images have been captured in three distinct vertical orientations (Front, Up, and Down) and has been featured six different facial expressions: Smile, Surprise, Surprise with Mouth Open, Neutral, Sad, and Yawning. The original IIITM Face dataset includes additional attributes such as gender, presence of facial hair like mustaches and beards, eyeglasses, clothing, and hair density. For this study, the IIITM Face dataset was adapted to focus on facial expressions across different orientations. The IIITM Face Emotion dataset features only the facial region segmented for each subject, with all images resized to fixed dimensions of 800 x 1000 pixels, maintaining an aspect ratio of 4:5. This resizing approach ensures consistent scaling across various facial positions for each subject. Some images from the dataset have been shown in Fig. 9.

Fig. 9. Samples from IIITM Face Emotion dataset (Best viewed in color).

C. Data Pre-processing

Despite of the approaches in the literatures, we have been intruding denoise process as a data pre-processing step. The method has been used to denoising face image is the method called Non-Local Means proposed in study [25] which based on a simple principle: replacing the color of a pixel with an average of the colors of similar pixels. But the most similar pixels to a

given pixel have no reason to be close at all. It is therefore licit to scan a vast portion of the image in search of all the pixels that really resemble the pixel one wants to denoise. Thus, we have split each image into five parts and we resize them into 224x224 resolution to be compatible with our modified AlexNet input layer. To resize images, we have used Cubic Interpolation algorithm. In short, for denoising and rescaling images, we have been used the implementation of Cubic Interpolation and Non-Local Mean algorithms available in OpenCV library. The face bounding box detection has been achieved by Haar-like detector present in study [26].

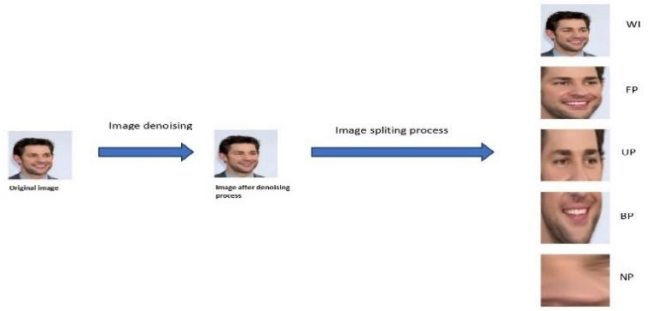

Fig. 10. Example of pre-processing steps for each image in LFWA datasets (Best viewed in color).

Furthermore, we have been processing to cross-validation approaches to predict the skill of our subnets.

In general, cross-validation is a statistical technique employed to assess the performance of machine learning models. When you have both a machine learning model and data at hand, you aim to determine its capability to fit the data. One common approach is to divide the data into training and test sets, training the model on the former and evaluating its performance on the latter. However, a single evaluation may not be sufficient to ascertain whether a favorable outcome is due to genuine model efficacy or mere chance. By conducting multiple evaluations through cross-validation, you can gain greater confidence in the robustness and design of your model.

The method involves a parameter known as 'k,' which denotes the number of subsets the data sample will be divided into. Hence, this technique is commonly referred to as 'k-fold cross-validation.' When a particular value is selected for 'k,' it can replace 'k' in the model reference; for example, setting k=10 would be termed '10-fold cross-validation.

Unfortunately, k-fold cross-validation may not be suitable for assessing imbalanced classifiers. This is because the data is divided into k-folds based on a uniform probability distribution.

While this approach may be effective for datasets with a balanced class distribution, it can falter when faced with severely skewed distributions. In such cases, one or more folds may contain minimal or no instances of the minority class. As a result, many model evaluations could be misleading, since the model could achieve high accuracy by simply predicting the majority class.

Balanced dataset steps. Machine learning algorithm performance is often assessed using public datasets like LFWA (see Fig. 10), but this approach can be problematic for imbalanced data. For instance, consider the task of gender classification in face attributes. A typical face dataset might have a distribution of 98% male and 2% female samples. Simply guessing the majority class would result in a predictive accuracy of 98%. However, the application demands high accuracy for detecting the minority class (female) while allowing for some errors in the majority class (male) to achieve this precision. Relying solely on straightforward predictive accuracy is not suitable in such scenarios. This realization underscores the necessity of balancing the dataset to obtain more accurate and meaningful results.

In this study, we adopted the method proposed in study [25] to balance the dataset for each subnet within our proposed pipeline. We opted for this algorithm due to its approach of over-sampling the minority class by generating "synthetic" examples rather than simply duplicating existing ones. This method creates additional training data by applying specific operations to real data, such as selecting k nearest neighbors from the minority class. However, it should be noted that this approach generates synthetic examples in a more generalized manner, operating in "feature space" rather than directly in "data space". Conversely, the majority class is addressed by under-sampling, wherein samples are randomly removed until the minority class comprises a specified percentage of the majority class.

As motioned in the sections above, our model combined five subnets, each one has a specific subset of attributes. Therefore, we have been applied SMOTE algorithm for each part of subnets parts. More details have been shown in figures; Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig.16, Fig. 17 and Fig. 18.

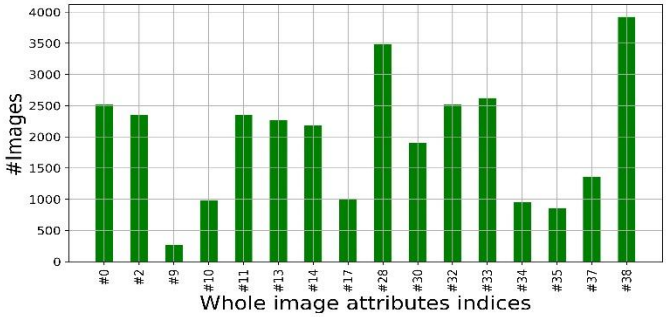

Fig. 11. Data distribution (LFWA dataset) before applied SMOTE algorithm to 16 face attributes according to WI-Net (Best viewed in color) (I).

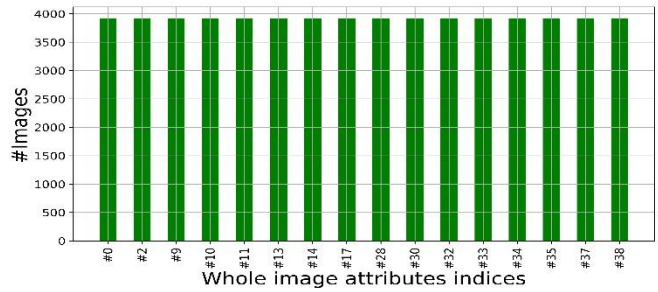

Fig. 12. Data distribution (LFWA dataset) after applied SMOTE algorithm to 16 face attributes according to WI-Net (Best viewed in color) (II).

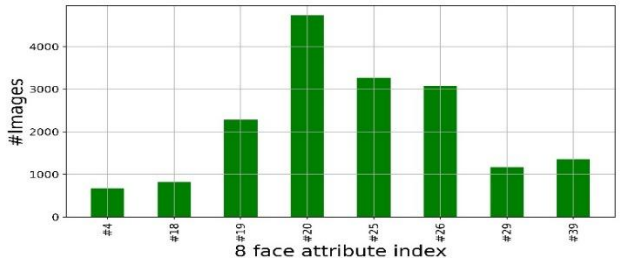

Fig. 13. Data distribution (LFWA dataset) before applied SMOTE algorithm to 8 face attributes according to FP-Net (Best viewed in color) (I).

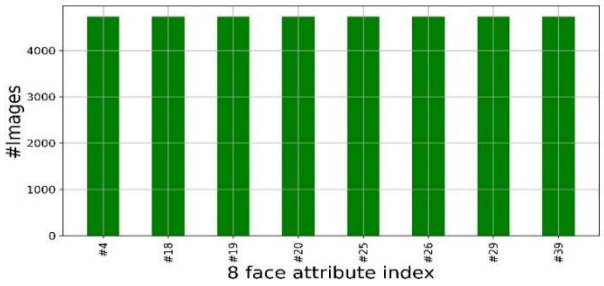

Fig. 14. Data distribution (LFWA dataset) after applied SMOTE algorithm to 8 face attributes according to FP-Net (Best viewed in color) (II).

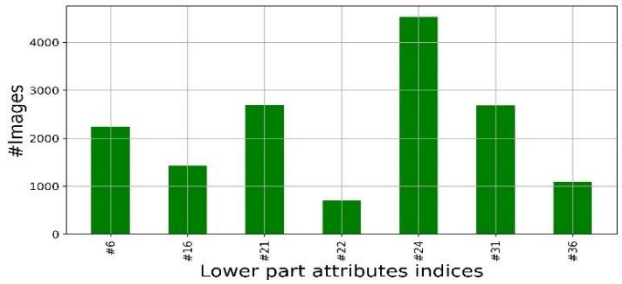

Fig. 15. Data distribution (LFWA dataset) before applied SMOTE algorithm to 7 face attributes according to LP-Net (Best viewed in color) (I).

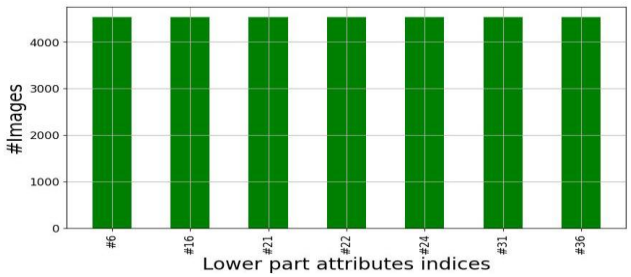

Fig. 16. Data distribution (LFWA dataset) after applied SMOTE algorithm to 7 face attributes according to LP-Net (Best viewed in color) (II).

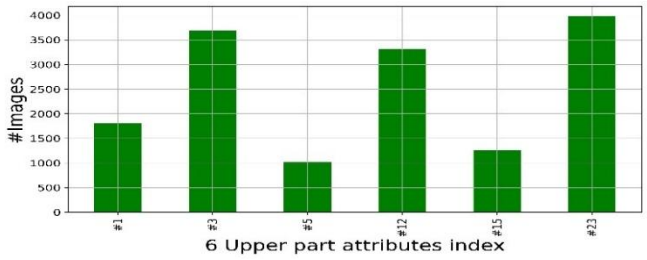

Fig. 17. Data distribution (LFWA dataset) before applied SMOTE algorithm to 6 face attributes according to UP-Net (Best viewed in color) (I).

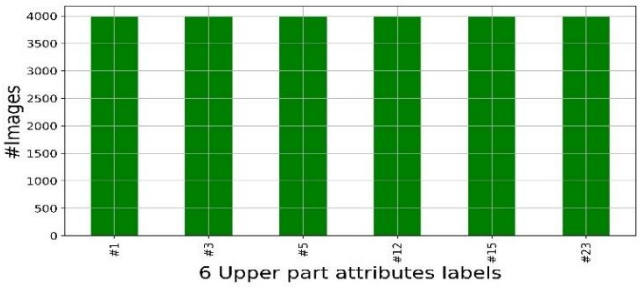

Fig. 18. D Data distribution (LFWA dataset) after applied SMOTE algorithm to 6 face attributes according to UP-Net (Best viewed in color) (II).

For nose part of our model, we have approximation the same number of classes; 4302 for Pointy Nose compared with 4321 for Big Nose. In this case there is no need to applied SMOTE algorithm.

D. Network Parameters

In the previous subsections, we have been described evaluation metrics, data augmentation and data pre-processing steps and this subsection, we will give more details about some parameters of our method. Therefore, we have been used the grid search algorithm [27], implemented in the kears framework to determine the values of some parameters; optimizer, Mini-batch size and Initial learning rate. See Table III for more details about search space of those parameters.

TABLE III. DETAILS OF SEARCH SPACE FOR EACH PARAMETER IN GRID SEARCH ALGORITHM [27]. WE HAVE BEEN INSPIRED FROM WORK IN [28] TO CHOOSE THE INTERVAL OF VALUES SPECIFIED FOR MINI-BATCH SIZE

Parameters	***Values***
Optimizer[a]	SGD; RMSprop; Adam; AdamW; Adadelta; Adagrad; Adamax; Adafactor; Nadam and Ftrl.
Mini-batch size	16; 28; 32; 64; 128; 256
Initial learning rate	0.1; 0.01; 0.001

[a]. All optimizer function names are reported from there implementation in Keras framework.

Thought a set test, we have concluded that the best values of the previous parameters are; SGD (the Stochastic Gradient Descent) as optimizer, mini-batch size equal to 28 and Initial learning rate equal to 0.001. All test has been made on FP-Net subnet for gender estimation attribute (#21) with a number of epochs equal to 100. In the next step, we have been increased the number of epochs from 0 to 600 in the training experiments, in order to set the epoch number which, get the better results in term of accuracy. This experiment shown that the achieves 100% and 0%, respectively, for Accuracy and Loss at the number of epoch equal to 500 (See Fig. 19 for more details). Therefore, we have been based on this conclusion to applied the parameters for all five subnets listed above (WI-Net, FP-net, UP-Net, LP-Net and NP-Net).

E. Experimental Results

This subsection summarizes the results that were obtained from the experiments for both datasets LFWA and IIITM face emotion. The most methods in the literature use LFWA to make those evaluation. In addition, we have been choosing to use IIITM face emotion, which has Asian people as subject, since the LFWA dataset has Asian people as minority class compared

with others ethnicity (White, Africans). Another reason to use IIITM face emotion dataset is all images has been taken in constrained condition for head pose, emotions and light when LFWA is unconstrained dataset. In short, we have been used the LFWA dataset to shown the performance of our approach compared with state-of-the-art methods and we have been used IIITM dataset to shown behavior of our method in constrained conditions on Asian people and to show the general ability of proposed model.

Through recent research, we have Gender and Smile are the most interested among all face attributes. To evaluate our proposed method on those two attributes, we have been used the FP-Net subnet to evaluate Gender estimation and LP-Net subnet to evaluate Smile estimation. The experimental set has been made on LFWA [7] and IIITM face emotion [24] since those two datasets provide gender and smile information; therefore, LFWA dataset specified the smile by attribute number 32 (#32) and gender by attribute number 21 (#21) when the IIITM face emotion dataset given that two information on image name (see [24] for attribute description).

The remaining subsections of this part are organized as follows. The results obtained on 40 face attributes compared with several state-of-the-art methods is described in subsection a). The subsection b) shown the performance of our model in Gender estimation. Finaly, we have been shown the behaviors of our model on Smile estimation task in subsection c).

1) 40 face attribute estimations: We have been training the five subnets on LFWA dataset trough 500 epochs after having applied pre-processing and data augmentation steps. All subnet parameters have been fixed like described in previous Network parameters subsection. However, the results of competing methods are reported from those originals paper whose respect the same protocol provided by dataset owners. Despite, in our method we have been applied SMOTE algorithm on dataset before training process. The classification results of proposed method and the competing methods on LFWA dataset have been presented in Table IV.

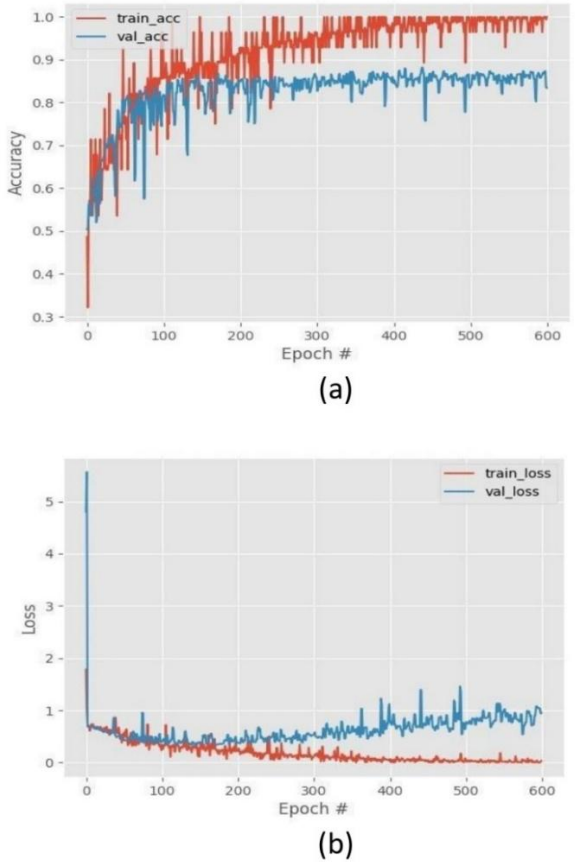

Fig. 19. Training and validation Accuracy/Loss for FP-Net on LFWA dataset (a and b) (Best viewed in color).

TABLE IV. ATTRIBUTE ESTIMATION ACCURACIES (IN %) FOR THE 40 BINARY ATTRIBUTES ON THE LFWA DATABASE BY THE PROPOSED APPROACH AND STATE-OF-THE ART METHODS [5], [6], [10], [14], [16], [15], [29]. THE AVERAGE ACCURACIES OF [5], [6], [10], [14], [16], [15], [29], AND THE PROPOSED APPROACH ARE 86.0%, 83.8%, 81.0%, 86.3%, 86.1%, 73.9%, 84.7% AND 86.8% RESPECTIVELY. SEE TABLE I FOR MORE DESCRIPTION ABOUT ATTRIBUTES LABELS

Attribute index	State-of-the Art Methods							Proposed Approach
	FaceTracker[15]	*PANDA[10]*	*LNets+Anet[6]*	*CTS-CNN[29]*	*MCNN-AUX[5]*	*DMTL[16]*	*MM-CNN[14]*	
1	70	84	84	77	77	80	78	**85.00**
2	67	79	82	83	82	86	81	**83.00**
3	67	79	82	83	85	82	81	**85.00**
4	71	81	83	79	80	84	83	**85.50**
5	65	80	83	83	83	92	93	**97.00**
6	77	84	88	91	92	93	92	**94.00**
7	72	84	88	91	90	77	79	**82.00**
8	76	87	90	90	93	83	84	**94.00**
9	88	94	97	97	97	92	92	**97.00**
10	62	74	77	76	81	97	97	**97.00**
11	78	81	84	87	89	89	85	83.00
12	68	73	75	78	79	81	82	**82.00**
13	73	79	81	83	85	80	85	**86.00**

14	73	74	74	88	85	75	76	79.00
15	67	69	73	75	77	78	82	**83.00**
16	70	75	78	80	82	92	92	**95.00**
17	90	89	95	91	91	86	84	90.00
18	69	75	78	83	83	88	89	**97.00**
19	88	93	95	95	96	95	95	86.00
20	77	86	88	88	88	89	87	81.00
21	84	92	94	94	94	93	94	**95.00**
22	77	78	82	81	84	86	82	**88.00**
23	83	87	92	94	93	95	94	**97.00**
24	73	73	81	81	83	82	82	**86.00**
25	69	75	79	80	82	81	81	**86.00**
26	66	72	74	75	77	75	79	**80.00**
27	70	84	84	73	93	91	91	68.00
28	74	76	80	83	84	84	85	**87.00**
29	63	84	85	86	86	85	87	**90.00**
30	70	73	78	82	88	86	87	**89.00**
31	71	76	77	82	83	80	84	**88.00**
32	78	89	91	90	92	92	91	**92.00**
33	67	73	76	77	79	79	79	**80.00**
34	62	75	76	77	82	80	82	**82.00**
35	88	92	94	94	95	94	94	90.00
36	75	82	88	90	90	92	91	90.00
37	87	93	95	95	95	93	95	86.00
38	81	86	88	90	90	91	90	87.00
39	71	79	79	81	81	81	83	**85.00**
40	80	82	86	86	86	87	85	78.00
Average	**73.9**	**81.0**	**83.8**	**84.7**	**86.3**	**86.1**	**86.3**	**86.83**

2) Gender estimation: In this subsection, we have been shown the results of our proposed approach for gender estimation task. As mentioned in section above, we have been pre-processed each image by denoising and splitting to remove the noise and adjusted them to input blocks.

We have been compared our approach with FaceTracer [15], PANDA[10], LNets+ANets [6] and all models (R-CNN_Gender, Multitask_Face, HyperFace and HF-ResNet) proposed in [8]. The gender estimation performance of different methods is reported in Table V. On the LFWA dataset, our method outperforms all competing methods listed above. Unlike all these methods in our approach, we have been processed to balanced dataset before the training step.

The imbalanced distribution in datasets make behavior of each model change from dataset to another. In addition, we have been evaluated the generalization ability of our FP-Net approach with cross-database testing on the IIITM face emotion and LFWA. See Table VI for more details.

3) Smile estimation: Smile detectors find applications across various sectors, including the media industry. Here, they play a crucial role in enabling companies to gauge public sentiment towards their products and services. For this reason, in this part of our paper, we have been interested to Smile estimation task. However, we have been presented the Smile estimation performance on LFWA and IIITM face emotion datasets since these datasets come with Smile information. We have been compared our LP-Net with MCFA [30], PANDA [10], FMTNet [21] and LNets+ANets [6]. The Smile estimation performance of different method is reported in Table VII. Furthermore, we have been evaluated the generalization ability of our LP-Net approach with cross-database testing on the IIITM face emotion and LFWA. See Table VIII for more details.

TABLE V. PERFORMANCE COMPARISON (IN %) OF GENDER ON LFWA DATASET

Method	***LFWA dataset***
FaceTracer[15]	84.00
PANDA[10]	92.00
LNets+ANets[6]	94.00
R-CNN_Gender[8]	91.00
Multitask_Face[8]	93.00
HyperFace[8]	94.00
HF-ResNet[8]	94.00
Proposed FP-Net	95.00

TABLE VI. CROSS-DATABASE TESTING ACCURACIES (IN %) OF FP-NET USING LFWA AND IIITM FACE EMOTION DATABASES FOR GENDER CLASSIFICATION

Database		***Metrics***			
Training	Testing	Accuracy	Precision	Recall	F1-score
IIITM	IIITM	99%	100%	99%	99%
IIITM	LFWA	50%	50%	100%	67%
LFWA	IIITM	79%	99%	80%	88%

TABLE VII. PERFORMANCE COMPARISON (IN %) OF SMILE ON LFWA DATASET

Method	***LFWA dataset***
MCFA[30]	88.00
PANDA[10]	89.00
FMTNet[21]	89.49
LNets+ANets[6]	91.00
Proposed LP-Net	92.00

TABLE VIII. CROSS-DATABASE TESTING ACCURACIES (IN %) OF LP-NET USING LFWA AND IIITM FACE EMOTION DATABASES FOR SMILE CLASSIFICATION

Database		***Metrics***			
Training	Testing	Accuracy	Precision	Recall	F1-score
IIITM	IIITM	91%	94%	96%	95%
IIITM	LFWA	54%	51%	92%	65%
LFWA	IIITM	54%	99%	48%	65%

V. DISCUSSION

In this section, summarized the discussion of the results achieved about 40 faces attributes, gender recognition, smile estimation and generalization ability of the proposed system. In the first subsection, we have been presented the effectiveness in the most face attributes (30/40) of the proposed approaches and some lower results (10/40). Additional results analysis about gender recognition has been shown in the second subsection. the third subsection contains behavior of our proposed approach in smile estimation task, and finally we a subsection on generalization ability of our system evaluated on cross-database testing.

A. 40 Face Attributes Estimation

The results on LFWA dataset by proposed approach and the state-of-the-art are reported in Table IV. The proposed approach outperforms [5], [6], [10], [14], [16], [15] and [29] for the most of the 40 attributes. Specifically, the proposed approach out performs competing methods by 30 face attributes. The remaining 10 attributes belongs to WI-Net, FP-Net and LP-Net. Those attributes can be subdivided into three groups; {#14, #35, #36 and #38} from Whole Image group, {#19, #20, #27 and #40} from Face Part group and {#17, #37} from Lower Part group. For WI-Net, we have four attributes has lower results from 17 attributes according to this subnet. All those attributes have a number of attributes in correlation, less than 5 attributes (see Table II) and belongs to image segments which need Meged-CNN structure like [14], very deep structure like in [29] or more the 4 attributes in correlation like the work in [16]. Further, the subnet FP-Net show 4 attributes with poor results from 8 attributes according to this subnet, while the competing methods in [16] and [6] show better results than our FP-Net. Despite, they use similar network structure to ours (AlexNet) but they use more attributes in correlation task (attributes inter-correlation). In addition, the attributes #17 from LP-Net subnet show similar results to [15] which use pixel of image combined with AdaBoost algorithm to achieve classification process but this approach show a limitation for profile face image even though, our subnet show lower results compared to [6], [5] and [29]. Since, the work in [5] use network structure similar to ours, and they add auxiliary block (coined AUX) which use fully connection approach between all attributes to handle attribute #17 estimation, when our LP-Net use just 3 attributes (see Table II for more details). In the same manner, the work in [6] use more than 3 attributes to make #17 estimation. However, the good results provided in [27] has been achieved by a deeper structure (similar to VGG) than ours. On the other hand, our proposed subnet shows better results about #17 than [10], [14] and [16] who's adopted AlexNet as backbone like ours, but LP-Net outperforms the estimation in task specification step (feature selection and AdaBoost classifier). Finaly, for attribute #37 our LP-Net shows lower results compared to all competing methods listed in Table IV. Thus, the works [8], [12] and [13] use similar backbone model like ours when [3], [4] use a deeper structure than AlexNet when the number of attributes used in classification step for all those methods is bigger than 5 attributes (see Table II) used in our proposed method.

Through all previous analysis, we have been concluded for WI-Net and LP-Net necessity to increase a number of attributes used in classification process to a number bigger than 5 and we have been concluded about FP-Net necessity to change AlexNet model by another structure like VGG or ResNet.

B. Gender Recognition

We present the gender recognition performance on LFWA and IIITM Face Emotion datasets since these datasets come with gender information. Our approach shows better results compared with PANDA [10] (see Table V). Based on results reported from there paper, this approach shows good performance in gender estimation on LFWA (92%) even when images are tightly cropped and variation in pose is reduced, but our FP-Net reach 95% with a wide variation in facial pose (our approach gain more from parts-based assumptions in term of head pose challenge). In the same manner, our FP-Net model outperformed FacTracer [15] methods by 11% in term of improvement. The FaceTracer [15] approach spite the input face to 10 parts while our FP-Net split the face into 4 parts only which get further advantage in terms of pre-processing time and has limitation with non-frontal face. In addition, our FP-Net outperforms LNets+ANets[6] by 1% with 5 attributes groups for our and six groups for LNets+ANets (more advantage in terms of group number). However, our FP-Net use one Modified AlexNet structure when LNets+ANets use a cascade of two AlexNet model which make advantage, on parameter number, train/test time complexity and memory consumption. Although, HyperFace models proposed in [8] has the same backbone network like our FP-Net (Modified AlexNet) even the FP-Net

shows better results more than R-CNN_Gender and Multitask_Face. Respectively, by 4% and 2% in the term of improvement. R-CNN_Gender predicts gender task only when our FP-Net can estimate 8 more attributes in addition of gender task (multi labels estimation against one task estimation). Since, Multitask_Face predicts gender in multitask approach which make training process hard than ours (multi labels against multitask process in the training set). Furthermore, the HF-ResNet approach use ResNet-101 model as a backbone network combined with AlexNet model which make it a very deep structure compared with our FP-Net model and slower than our train/test set. However, our FP-Net approach improved HF-ResNet by 1% in the term of accuracy. On the whole, the proposed FP-Net model for joint estimation of gender attribute demonstrates their effectiveness compared with existing approaches at many specific levels like; parameters number, memory consumption, number of attributes subgroups, accuracy and time complexity.

C. Smile Detection

Our proposed LP-Net model outperformed the state-of-the-art methods (see Table VII) in Smile detection on LFWA dataset. All accuracies value presented in Table VII has been reported from their original papers. To illustrate, our LP-Net outperformed the work in [30] by 4% of accuracy and the cited work used a cascade of three VGG-16 models (SNet, MNet and LNet) which made the detection process more complex and harder to train. Even though, our LP-Net use just one Modified AlexNet with seven attributes in the output combined with Adaboost classifier which made it less complex in train and test sets. Further, FMTNet presented in [21] take 40 attributes and split them to many subgroups, while each attribute has weight depending on the number of groups and the number of attributes in each group. Which means this approach investigate 40 attributes to estimate smile attribute when our LP-Net use just 3 attributes selected by cited algorithm from 7 attributes in relation with lower part of face. Therefore, LP-Net proposed in this work use less attributes to show a attributes correlation and gained 3% in the term of accuracy compared with FMTNet. On the other hand, the work in [21] investigate three model use VGG-16 as backbone (FNet, MNet and TNet) when our LP-Net use just one AlexNet model, to handle the smile task. The proposed LP-Net shows improvement retch 1% and 3%, compared the results provide, respectively, by LNets+ANets and PANDA methods. In short, we find that LP-Net proposed in this work performs better for Smile detection task compared to competing methods. In the other hand, our approach shows a discriminating capability for multis and individual tasks.

D. Generalization Ability

We believe that the real scenario is different from the laboratory scenario which mean the generalization ability provided in [16] for the first time (for the best of our knowing) can give more information about our proposed approach. Hence, we evaluate the generalization ability of the proposed approach with cross-database testing on LFWA and IIITM Face emotion databases.

Specifically, cross-database testing of gender and smile estimation between LFWA and IIITM Face emotion databases is performed by training our approach on LFWA and testing it on IIITM Face emotion, and vice versa. The estimation results with cross-database are shown in Table VI and Table VIII. The results provided by cross-database testing is lower than intra-database testing. Image conditions (constrained in IIITM Face emotion and unconstrained in LFWA) and the number of images (1,928 images in IIITM face emotion and 13,143 images in LFWA) are responsible for the drop in performance. This experiment suggests that varying image sources can introduce additional hurdles for accurately estimating facial attributes. Nevertheless, we maintain confidence that our proposed approach yields commendable results even within this demanding context.

VI. Conclusions

The paper proposes a method to decode face attributes using a multi-task, part-based approach and attribute relationships. In contrast of exciting works, it introduces two preprocessing steps: image denoising with the Non-Local Means algorithm and dataset balancing using the SOME algorithm. The feature selection has been done by splitting images into five parts (WI, FP, UP, LP, NP) and each one has been processed by a corresponding subnet (modified AlexNet as backbone). The correlation-heterogeneity relationship between attributes has been achieved by a novel feature selection Algorithm (proposed in this work) combined with AdaBoost algorithms.

To evaluated the proposed approach, we have been used LFWA and IIITM Face Emotion datasets. The first one has images in unconstrained conditions and large scale of illumination, head pose, … when the second one has images with constrained conditions of head pose, illumination and expression. This strategy helps to studies the performance of proposed approach in different conditions and ethnicity.

Trought a set of experiments has been made in this work, our approach performs well the state-of-the-arts methods specifically, on gender and smile attributes. Nevertheless, the results presented in this work shown that our subnets; WI-Net, UP-Net, LP-Net and NP-Net outperformances the competing methods on specific attributes groups, according to those parts of face, when the subnet FP-Net shown some lower results for attributes {#19, #20, #27 and #40}. One possible solution to this issue could be replaced AlexNet with deeper structure similar to VGG or ResNet. While, the lower results present by this work for attributes number {#14, #35, #36 and #38} and {#17, #37} returned respectively, by WI-Net and LP-Net, could be handle by increasing a number of attributes used to shown the relationship in classification process.

Finally, we have been studied the generalization ability of the proposed approach under cross-database testing scenarios on LFWA an IIITM Face Emotion datasets. Through a results analysis, the cross-database testing highlights the importance of training database in real-world face attributes estimation systems.

For future work, we will try to use a deeper structure for attributes with lower results in FP-Net subnet and we will investigate more time in feature selection to get better results for attributes number {#14, #35, #36 and #38} and {#17, #37}. On the other hand, we will adapt the proposed approach to estimate

age task and face emotion. The age task will be studied in regression manner.

REFERENCES

[1] O. Maarouf, A. Maarouf, R. El Ayachi, et M. Biniz, « Automatic translation from English to Amazigh using transformer learning », Indones. J. Electr. Eng. Comput. Sci., vol. 34, no 3, p. 1924, 2024, doi: 10.11591/ijeecs.v34.i3.pp1924-1934.

[2] M. Biniz et R. El Ayachi, « Recognition of Tifinagh Characters Using Optimized Convolutional Neural Network », Sens. Imaging, vol. 22, no 1, p. 28, 2021, doi: 10.1007/s11220-021-00347-1.

[3] U. D. Dixit, M.S. Shirdhonkar, « Face-based Document Image Retrieval System », Procedia Computer Science, Volume 132, 2018, Pages 659-668, ISSN 1877-0509, doi :10.1016/j.procs.2018.05.065.

[4] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, « A review of video surveillance systems », Journal of Visual Communication and Image Representation, Volume 77, 2021, doi:10.1016/j.jvcir.2021.103116..

[5] E. Hand et R. Chellappa, « Attributes for Improved Attributes: A Multi-Task Network Utilizing Implicit and Explicit Relationships for Facial Attribute Classification », Proc. AAAI Conf. Artif. Intell., vol. 31, no 1, 2017, doi: 10.1609/aaai.v31i1.11229.

[6] Z. Liu, P. Luo, X. Wang, et X. Tang, « Deep Learning Face Attributes in the Wild », in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile: IEEE, 2015, p. 3730-3738. doi: 10.1109/ICCV.2015.425.

[7] B. H. Gary, R. Manu, B. Tamara, et L.-M. Erik, « Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments », p. 07-49, 2007.

[8] R. Ranjan, V. M. Patel, et R. Chellappa, « HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition », IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no 1, p. 121-135, janv. 2019, doi: 10.1109/TPAMI.2017.2781233.

[9] D. Fan, H. Kim, J. Kim, Y. Liu, et Q. Huang, « Multi-Task Learning Using Task Dependencies for Face Attributes Prediction », Appl. Sci., vol. 9, no 12, p. 2535, 2019, doi: 10.3390/app9122535.

[10] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, et L. Bourdev, « PANDA: Pose Aligned Networks for Deep Attribute Modeling », in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA: IEEE, 2014, p. 1637-1644. doi: 10.1109/CVPR.2014.212.

[11] J. Cao, Y. Li, et Z. Zhang, « Partially Shared Multi-task Convolutional Neural Network with Local Constraint for Face Attribute Learning », in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT: IEEE, 2018, p. 4290-4299. doi: 10.1109/CVPR.2018.00451.

[12] S. Yang, P. Luo, C. C. Loy, et X. Tang, « Faceness-Net: Face Detection through Deep Facial Part Responses », IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no 8, p. 1845-1859, 2018, doi: 10.1109/TPAMI.2017.2738644.

[13] U. Mahbub, S. Sarkar, et R. Chellappa, « Segment-Based Methods for Facial Attribute Detection from Partial Faces », IEEE Trans. Affect. Comput., vol. 11, no 4, p. 601-613, 2020, doi: 10.1109/TAFFC.2018.2820048.

[14] H. Kawai, K. Ito, et T. Aoki, « Face Attribute Estimation Using Multi-Task Convolutional Neural Network », J. Imaging, vol. 8, no 4, p. 105, avr. 2022, doi: 10.3390/jimaging8040105.

[15] N. Kumar, P. Belhumeur, et S. Nayar, « FaceTracer: A Search Engine for Large Collections of Images with Faces », in Computer Vision – ECCV 2008, vol. 5305, D. Forsyth, P. Torr, et A. Zisserman, Éd., in Lecture Notes in Computer Science, vol. 5305. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, p. 340-353. doi: 10.1007/978-3-540-88693-8_25.

[16] H. Han, A. K. Jain, F. Wang, S. Shan, et X. Chen, « Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach », IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no 11, p. 2597-2609, 2018, doi: 10.1109/TPAMI.2017.2738004.

[17] J. Xiang et G. Zhu, « Joint Face Detection and Facial Expression Recognition with MTCNN », in 2017 4th International Conference on Information Science and Control Engineering (ICISCE), Changsha: IEEE, 2017, p. 424-427. doi: 10.1109/ICISCE.2017.95.

[18] S. Ioffe et C. Szegedy, « Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift », 2015, doi: 10.48550/ARXIV.1502.03167.

[19] C. Shorten et T. M. Khoshgoftaar, « A survey on Image Data Augmentation for Deep Learning », J. Big Data, vol. 6, no 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.

[20] A. Shannaq et L. Elrefaei, « AGE ESTIMATION USING SPECIFIC DOMAIN TRANSFER LEARNING », Jordanian J. Comput. Inf. Technol., no 0, p. 1, 2020, doi: 10.5455/jjcit.71-1571410322.

[21] N. Zhuang, Y. Yan, S. Chen, H. Wang, et C. Shen, « Multi-label learning based deep transfer neural network for facial attribute classification », Pattern Recognit., vol. 80, p. 225-240, 2018, doi: 10.1016/j.patcog.2018.03.018.

[22] H. Song et H. Chen, « A Fast and Effective Large-Scale Two-Sample Test Based on Kernels », 2021, doi: 10.48550/ARXIV.2110.03118.

[23] F. Charte, A. J. Rivera, M. J. Del Jesus, et F. Herrera, « MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation », Knowl.-Based Syst., vol. 89, p. 385-397, 2015, doi: 10.1016/j.knosys.2015.07.019.

[24] Rishi Raj Sharma , K V Arya, April 3, 2023, "IIITM Face Emotion (An Indian Face Image Data)", IEEE Dataport, doi: 10.21227/rers-ck04.

[25] A. Buades, B. Coll, et J.-M. Morel, « Non-Local Means Denoising », Image Process. Line, vol. 1, p. 208-212, 2011, doi: 10.5201/ipol.2011.bcm_nlm.

[26] P. Viola et M. Jones, « Rapid object detection using a boosted cascade of simple features », in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA: IEEE Comput. Soc, 2001, p. I-511-I-518. doi: 10.1109/CVPR.2001.990517.

[27] P. Liashchynskyi et P. Liashchynskyi, « Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS », 2019, doi: 10.48550/ARXIV.1912.06059.

[28] I. Kandel et M. Castelli, « The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset », ICT Express, vol. 6, no 4, p. 312-315, 2020, doi: 10.1016/j.icte.2020.04.010.

[29] Y. Zhong, J. Sullivan, et H. Li, « Face Attribute Prediction Using Off-the-Shelf CNN Features », 2016, doi: 10.48550/ARXIV.1602.03935.

[30] N. Zhuang, Y. Yan, S. Chen, et H. Wang, « Multi-task Learning of Cascaded CNN for Facial Attribute Classification », 2018, doi: 10.48550/ARXIV.1805.01290.