Received: April 25, 2022. Revised: June 21, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Multi-Output Convolutional Neural Network for Automatic Human Head Attributes Classification

Loai Alamro[1]* Yuhanis Yusof[1] Nooraini Yusoff[2]

[1]*School of Computing, Universiti Utara Malaysia, Malaysia*
[2]*Department of Data Science, Universiti Malaysia Kelantan, City Campus, Malaysia*
* Corresponding author's Email: loaialamro@gmail.com

Abstract: Human head attribute classification (HHAC) is a fundamental and substantial research domain in pattern recognition and computer vision. However, recent HHAC networks do not consider the correlation of common characteristics among the attributes over different regions of the human head. To address the above problem, this study proposes a multi-output convolutional neural network to jointly learn the features of human head attributes with common characteristics. The proposed network contains two convolutional blocks and five output layers, where each output layer learns to predict a specific group of human head attributes. In order to properly learn the correlation among the human head attributes, this study divides these attributes into five groups: hair, face, style, accessories, and appearance. Extensive experiments showed that the proposed network obtained an average classification accuracy of 95.29% and 97.93% on the challenging CelebA and LFWA datasets, respectively. Thus, the proposed network is approximately 2% and 10% superior to the closest competitor (i.e., PS-MCNN) on both datasets. In addition, the proposed network achieved higher classification accuracy compared to the existing networks almost in all human head attributes. That findings demonstrate the effectiveness of the proposed network and the attributes grouping method in learning the correlations among human head attributes correctly.

Keywords: Deep learning, Convolutional neural network, Multi-label learning, Human head attribute classification, Attribute correlation.

1. Introduction

Human Head Attribute Classification (HHAC) is a fundamental and substantial research domain in pattern recognition and computer vision. The main goal of HHAC is to predict the human head attributes of a given image, including gender, age group, smiling, attraction, etc. During recent years, HHAC has attracted significant attention due to its widespread applications, including object recognition [1, 2], face recognition [3, 4], face verification [5, 6], face retrieval [7], image retrieval [8], image search [9]and recommendation systems [10]. However, HHAC remains a challenging problem in practice because of the large variability of human head appearances in illumination, pose, expression, etc.

Recently, due to the outstanding performance of Deep Learning (DL) networks, especially the Convolutional Neural Network (CNN), most HHAC networks mainly focus on using DL to predict human head attributes. Generally speaking, the HHAC networks can be divided into two categories: single-label learning based networks [11, 12] and multi-label learning based networks [13-19]. The single-label learning-based networks firstly employ the CNN to extract human head features then predict the head attributes using the Support Vector Machine (SVM) [20]. In this manner, Zhang, Paluri, Ranzato, Darrell and Bourdev [11] proposed the Pose Aligned Networks for Deep Attribute modelling (PANDA) by combining the deep representations extracted from every pose region of a human head with deep representations of the entire human head to train the SVM classifier for HHAC. Liu, Luo, Wang and Tang [12] employ Localization Networks (LNets) for face localization and Attribute Network (ANet) for feature

extraction, then employs a single SVM classifier for each face attribute. However, these networks consider the classification of every human head attribute as an individual and independent problem, thus ignoring the correlations between attributes. In addition, these networks use DL for feature extraction and require training an external classifier to classify human head attributes.

Alternatively, multi-label learning-based networks, predict multiple head attributes simultaneously in an end-to-end DL network. These networks use the lower layers of CNN to extract the shared features of the head attributes and learn the head attributes on the upper layers of CNN. Rudd, Günther and Boult [13] proposed a novel Mixed Objective Optimization Network (MOON) with a loss function that mixes multiple-task to address the different distribution of attribute labels. He, Wang, Fu, Feng, Jiang and Xue [14] proposed the Adaptively Weighted Multi-task CNN (AW-CNN) to jointly learn multi-attributes with the validation loss trend algorithm that updates the weights of the weighted loss layer automatically. Guo, Fan and Wang [15] proposed the Class Activation Map (CAM) network to highlight the relevant image regions of each head attribute. Mahbub, Sarkar and Chellappa [16] proposed the Normalized Score Aggregation (NSA) which utilize keypoints to directly segment faces into several image patches that fed into two-step CNN for feature extraction and learning prediction. Xu, Chen, Li, Shen, Lv, Zhou and Ji [17] combined the Bio-inspired Facial Aesthetic Ontology (Bio-FAO) and CNN to predict the human head attribute. Huang, Li, Cheng, Zhang and Hauptmann [19] propose a Greedy Neural Architecture Search (GNAS) for automatically discovering the optimal tree-like architecture to predict multi-attributes. Zhuang, Yan, Chen and Wang [18] proposed a Multi-task Framework of Cascaded CNN termed (MCFA), which comprised of three cascaded sub-networks to jointly learn multiple tasks (i.e., face detection, face landmark & localization, and face attribute classification). However, these networks treated the head attributes equally during the training phase, ignoring the various learning complexities of these attributes. Therefore, the performance of these networks may not be optimal since the correlations between the head attributes are not effectively exploited.

Moreover, several multi-label learning-based networks [21-24] proposed to divide head attributes into several groups. For instance, Hand and Chellappa [21] proposed to divide the head attributes into 9-groups according to their locations on a human head and learn the relationships between the attributes in these locations. Accordingly, they proposed the Multi-task CNN (MCNN) combined with an Auxiliary Network (AUX), which benefit from the attribute relationships and an improved classification. Han, Jain, Wang, Shan and Chen [22] proposed to divide the head attributes according to the heterogeneity (i.e., ordinal vs. nominal and holistic vs. local) in terms of data type and semantic meaning. Accordingly, they later proposed the Deep Multi-Task Learning (DMTL) which consist of 4 structurally identical sub-networks for each group of attributes with separate loss functions for each one. On the other hand, Cao, Li and Zhang [23] proposed to divide the head attributes into 4 groups (i.e., upper, middle, lower, and whole image) according to their locations on the human head image. They proposed Partially Shared Multi-task CNN (PS-MCNN), which is composed of 4 sub-networks where each one corresponds to a specific attributes group and a single shared sub-network for HHAC. Mao, Yan, Xue and Wang [24] which proposed to divide the 40 categories into two groups: objective categories and subjective categories. They proposed Deep Multi-task Multi-label CNN (DMM-CNN), which involved two networks for feature extraction and a novel dynamic weighting scheme to automatically specify the loss weight for each category. All the stated studies presented several criteria to group the human head attributes but have not considered the common characteristics among the attributes. Therefore, this study aims to address these challenges and propose: 1) a novel grouping method that classify human head attributes into five groups: hair, face, style, accessories and appearance; and 2) a deep Multi-Output Convolutional Neural Network (MOCNN) to learn the joint features among the attributes based on the proposed groups. The proposed network contains two convolutional blocks and five output layers, where each output layer learns to predict a specific group of human head attributes.

The rest of the article is organized as follows. Section 2 presents a brief Background on convolutional neural network. Section 3 shows the mechanism of datasets understanding and attribute grouping. The proposed network with the parameters used in building the layers of the networks is detailed in Section 4. In Section 5, experimental results of the proposed network are reported. Finally, Section 6 presents the conclusion.

2. Background

This section includes brief explanation on Convolutional Neural Network (CNN) which is the fundamental architecture in the study. CNN was

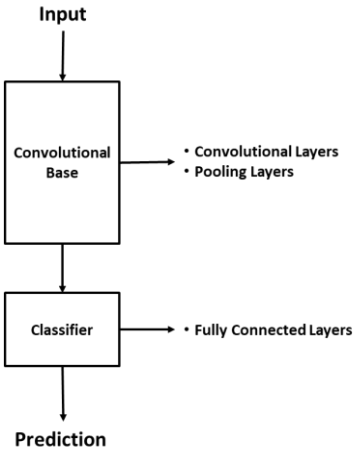


Figure. 1 Architecture of CNN (adapted from [36, 37])

proposed in 1980 by Fukushima [25], and it was later revised by LeCun [26]. Recently, CNNs achieved outstanding performance over many fields such as object recognition [1], face recognition [27], scene recognition [28] and natural language processing [29]. Due to the availability of large-scale labeled datasets such as ImageNet [30] and the high-performance Graphics Processing Units (GPUs), training complex models and huge dataset becomes possible. In this context, several networks have been proposed, such as AlexNet [31], VGGNet [32], GoogleNet [33], Residual Networks (ResNet) [34], and DenseNet [35]. A typical CNN architecture consists of two main parts, namely, the convolutional base and classifier, as shown in Fig. 1.

The convolutional base is composed of a convolutional stack and pooling layers. Convolution is a mathematical procedure for integrating two sets of input, namely, an image matrix with a dimension of $h \times w \times d$ and a filter (kernel) with a dimension of $f_h \times f_w \times d$, to produce a feature map with a volume dimension of $(h - f_h + 1) \times (w - f_w + 1) \times 1$. This procedure is formulated by Eq. (1).

$$X_j^l = f\left[\sum_{i \in M_j}(X_i^{l-1} * K_{ij}^l + b_j^l\right] \quad (1)$$

Where:

X_j^l signifies the j^{th} feature-map of the l layer, $f[\cdot]$ signifies the activation function, M_j signifies the input images, $*$ signifies the convolution operation, X_j^{l-1} the i^{th} feature-map of the $l-1$ layer, K_{ij}^l signifies the convolutional filter linking the j^{th} feature-map of the l layer and the i^{th} feature-map of the $l-1$ layer, b_j^l signifies the bias.

There are several activation functions that are commonly used such as sigmoid, tanh, and ReLU, which are formulated as follows:

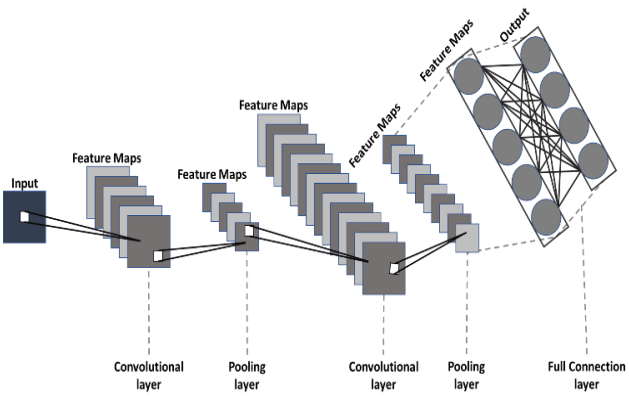


Figure. 2 CNNs layers

$$\text{Sigmoid} \quad f(x) = \frac{1}{1+e^{-z}} \quad (2)$$

$$\text{Tanh} \quad f(x) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad (3)$$

$$\text{ReLU} \quad f(x) = \max(0, x) \quad (4)$$

The pooling layers refer to a sample-based discretization process usually performed after the convolutional layers to reduce dimensionality. The goal of pooling layers is to downsample the representation of input (i.e., input matrix, hidden layer, and output matrix) through decreasing the total number of parameters, which then reduces the training time and avoids overfitting. Pooling layers can perform various functions, such as maxpooling or average-pooling. The former selects the maximum value in a certain filter area, whereas the latter selects the average value in a filter area. The pooling layers is formulated by Eq. (5). A CNN classifier is composed of fully connected layers, which learn how to use the features produced by earlier convolutional layers to achieve the explicit expression of classification. Fig. 2 illustrates the basic layers of CNNs architecture.

$$X_j^l = f\left[\beta_j^l \, down(X_i^{l-1} + b_j^l)\right] \quad (5)$$

Where:

X_j^l signifies the j^{th} feature-map of the l layer, $f[\cdot]$ signifies the activation function, β signifies the subsampling coefficient, $down(\cdot)$ signifies the subsampling function, X_j^{l-1} the i^{th} feature-map of the $l-1$ layer, b_j^l signifies the bias.

3. Understanding datasets and attribute grouping

The training process of a dataset is considered critical for building, testing, and subsequently

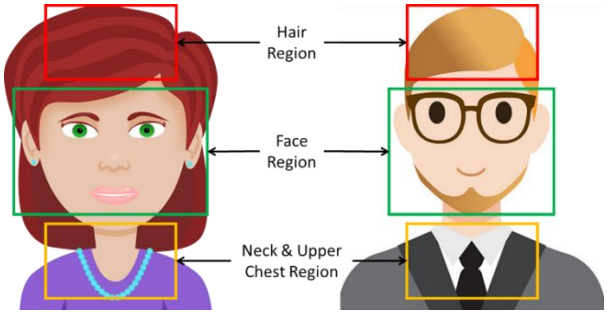


Figure. 3 Regions of human head

creating a successful DL network. Therefore, understanding content of the dataset is essential. This study investigates two widely used datasets (i.e., Large-scale CelebFaces Attributes (CelebA) [38] and Labeled Faces in the Wild-a (LFWA) [39]) of Human Head Attribute Classification (HHAC). Both datasets classify a human head into a set of categories (i.e., labels), where each category corresponds to an attribute of the human head. The CelebA dataset has 40 categories, while the LFWA dataset has 73 categories in which each one of them differs in its training, for example, training a network to classify the "Wearing-Eyeglasses," may be easier than training it to classify "Pointy Nose". In this study, the work focuses on analyzing 40 categories in both datasets and determine type of correlation between the categories in accordance with their locations on the human head. The categories in the datasets are distributed over three different regions of the human head; hair, face and neck & upper chest region, as shown in Fig. 3.

The hair region includes three sets of categories: 1) Hair colour (Black Hair, Blond Hair, Brown Hair, Gray Hair, Bald), 2) Hair style (Straight Hair, Wavy Hair, Receding Hairline, Bangs) and 3) Accessories (Wearing Hat). On the other hand, the face region details eight sets of categories: 1) Eyebrows (Arched Eyebrows, Bushy Eyebrows), 2) Eyes (Narrow Eyes, Bags under Eyes), 3) Nose (Big Nose, Pointy Nose), 4) Mouth (Big Lips, Mouth Slightly Open), 5) Bones (Double Chin, High Cheekbones, Oval Face), 6) Beard style (5 O'clock Shadow, Goatee, Mustache, No Beard, Sideburns), 7) Makeup (Heavy Makeup, Wearing Lipstick, Rosy Cheeks, Pale Skin), and 8) Accessories (Eyeglasses, Wearing Earrings). The final region which is the neck & upper chest region only include Accessories (Wearing Necklace, Wearing Necktie). We can observe that each set involves several categories with similar criteria in which each category belongs to a specific attribute of the human head. Moreover, there are several categories (i.e., Attractive, Blurry, Chubby, Male, Smiling and Young) which does not belong to any region of the human head regions nor to any set of

Figure. 4 Distribution of categories on human face

categories that have mentioned. These categories are considered global categories where each one represents a specific theme of human appearance. Fig. 4 illustrates distribution of the categories based on human head regions.

After studying the categories, it has been found that they are varied, concerning to the essential ingredients of the human head such as hair, eyes, nose, etc., regardless the artificial categories which are human-made accessories such as "Wearing Hat", "Eyeglasses", "Wearing Lipstick", and the like. Moreover, many of these categories have common characteristics as some of them belong to the same kind, location, or they are sometimes having both characteristics concerning to the kind and locations. For example, the categories of hair color and hairstyle describe the characteristics of the human head hair, as well these categories are located in the same region of the human head. On the other hand, although the categories of eyebrows, eyes, nose, mouth, bones, beard style and makeup are located on the face region, they have different characteristics. While the categories of eyebrows, eyes, nose, mouth, and bones describe the characteristics of the whole human face, the beard style and makeup describe the style characteristics of the human face. In addition, although categories such as Wearing Hat, Eyeglasses, Wearing Earrings, Wearing Necklace, and Wearing Necktie are distributed in all regions of the human head, these categories have common characteristics where they describe the accessories. Furthermore, the global categories such as gender and age group also have common characteristics where they describe the general appearances of the human.

In the final analysis, this study found that there is a disparity ratio in the learning of categories because each one has its own complexity. Hence, the classification accuracy among the categories is also different as some categories get lower classification accuracy than others. In contrast, there are many

categories that may be strongly correlated according to their common characteristics or locations on the human head, and they are sometimes having both correlations. Therefore, exploiting intrinsic correlations among categories could lead to extracting the optimal features and boosting classification accuracy. Subsequently, based on the assumption that many categories are strongly correlated as they have common characteristics the study splits all the 40 categories into five groups as they are explained below:

1- Hair Categories (Black Hair, Blond Hair, Brown Hair, Gray Hair, Bald, Bangs, Straight Hair, Wavy Hair, Receding Hairline).

2- Face Categories (Arched Eyebrows, Bushy Eyebrow, Narrow Eyes, Bags Under Eyes, Big Nose, Pointy Nose, Big Lips, Mouth Slightly Open, Double Chin, High Cheekbones, Oval Face).

3- Style Categories (5 o Clock Shadow, Goatee, Mustache, No Beard, Sideburns, Heavy Makeup, Wearing Lipstick, Rosy Cheeks, Pale Skin).

4- Accessories Categories (Eyeglasses, Wearing Earrings, Wearing Hat, Wearing Necklace, Wearing Necktie).

5- Appearance Categories (Attractive, Blurry, Chubby, Male, Smiling, Young).

4. Proposed approach

The proposed network consists of sixteen layers with one input layer, two convolutional layers, two pooling layers, one flattens layer, five fully connected layers, and five output layers. All of the convolutional layers of the proposed network involve ReLUs units to implement the linear transformation and the nonlinear mapping. The central concept in constructing this network is the deployment of filters with different sizes, where a larger filter was used in the first convolutional layer to extract shallow features, and a smaller filter was used in the second convolutional layer to extract deeper features.

The input layer of the proposed network is fed by images of size $(178 \times 218 \times 3$ or $256 \times 256 \times 3)$ pixels, where each image was preprocessed by the procedure described in section 5.1. In addition, a 5×5 convolutional filter was designed in the first convolutional layer to implement the convolution operations and extract the shallow features. Next, a 2×2 top pooling layer has been used to reduce the dimensionality of the convolutional layer outputs. Then, the outputs of the Maxpooling layer fed into the second convolutional layer to implement its convolution operations. In the second convolutional layer, a 3×3 convolutional filter has been designed to implement the convolution operations and produce

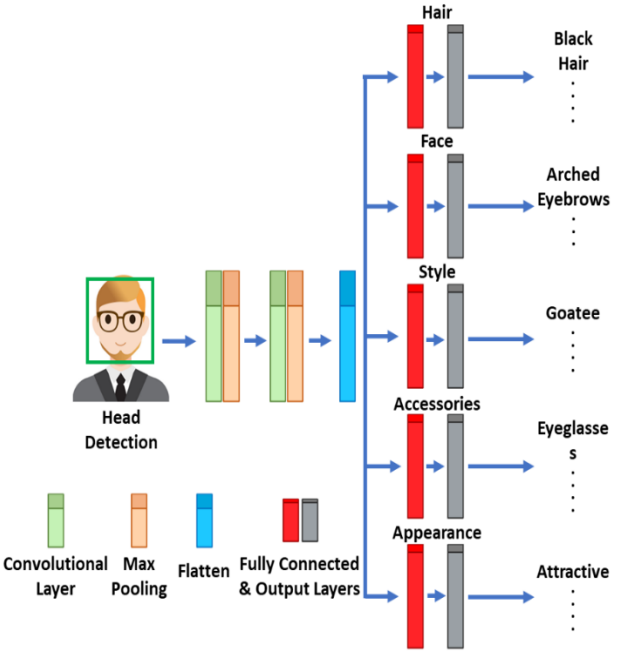


Figure. 5 Structure of the proposed network

the feature maps. A Maxpooling layer of size 2×2 was deployed following the second convolutional layer to reduce the dimensionality of the produced feature maps. The samples of the learned convolutional filters to implement the convolution operations of convolutional layers and the samples of the produced features are illustrated in Fig. 5 and 6 sequentially. Moreover, a flatten layer has been used to flatten outputs of the Maxpooling layer into a 1-dimensional vector. Then, the extracted features through the convolutional and pooling layers are passed into the fully connected layers. Since this study has proposed to divide the categories of datasets into five groups (Hair Categories, Face Categories, Style Categories, Accessories Categories and Appearance Categories), the proposed network included five fully connected layers, each one connected to an output layer. In addition, each output layer employed a Sigmoid classifier to implement the multi-category classification task. The exact structure of the proposed network is illustrated in Fig. 5.

5. Experiment results and discussions

This section covers the experiments performed to evaluate the effectiveness of the proposed HHAC networks. First, in Subsection 5.1, the datasets and parameter settings used for the evaluation are described. In subsection 5.2, the experiment settings are presented. Finally, in subsection 5.3, the performance of the proposed network is compared against several state-of-the-art networks.

5.1 Datasets and parameter settings

The experiments have been conducted on two challenging, large-scale Human Head Attribute datasets, namely CelebA and LFWA. CelebA dataset [18] includes 202,599 images of 10,177 celebrities' identities; each celebrity has approximately an average of 20 images. In addition, each head image of the CelebA dataset was provided with 40 binary attribute annotations. The CelebA dataset is split into three portions training, validation, and testing. The training set includes 162,770 images, validation set with 19,867 images, and the testing set includes 9,962 images. On the other hand, the LFWA dataset is built upon 13,233 images of 5,749 identities, where 1,680 identities have two or more images. Besides, each image of the LFWA dataset is provided with 73 binary attribute annotations. However, the LFWA dataset does not contain any validation images. Furthermore, the number of images it quite small and not superabundant to train an accurate DL network. Thus, this study proposes to: 1) deploy augmentation techniques to increase the number of images in LFWA dataset. Augmentation techniques include, rotation, scaling, flipping, shifting and zoom, as shown in Fig. 6; and 2) deploy transfer learning technique that will fine-tune the network trained on CelebA dataset to be used on LFWA dataset. After performing the data augmentation techniques, the LFWA dataset has over 25,000 images for training, 6,880 images for testing and 0.2% of the training images were used for validation. This study uses the same 40 categories of CelebA and LFWA in the experiments.

5.2 Experiment settings

The proposed network is implemented based on the open-source DL platform TensorFlow 2.0, where Intel Core i7-6700 CPU, 64 Gigabyte RAM and NVIDIA GEFORCE RTX 3090 GPU is used to train the networks for ten epochs with the batch size of 32. The learning rate for both networks is set to 0.0001 with a decay rate of 1×10^{-6}. The parameters of the input layer for each network are changed to fit the size of the images of each dataset, $178 \times 218 \times 3$ for CelebA dataset and $256 \times 256 \times 3$ for LWFA dataset. This study employed a binary cross-entropy loss for each categories group to implement the training.

5.3 Performance results on CelebA and LFWA dataset

In this subsection, we evaluate the effectiveness of the proposed network by comparing its performance against several state-of-the-art HHAC networks on CelebA and LFWA datasets sequentially. The performance of the proposed network has been compared with eleven networks on CelebA datasets including the PANDA [11], LNets+ANet [12], MOON [13], NSA [16], MCNN-AUX [21], MCFA [18], GNAS [19], AW-CNN [34], PS-MCNN [23], DMM-CNN [24] and DMTL [22]. In addition, the proposed network has been also compared with nine networks on LFWA datasets includes PANDA [11], LNets+ANet [12], NSA [16], MCNN-AUX [21], MCFA [18], GNAS [19], PS-MCNN [23], DMM-CNN [24] and DMTL [22].

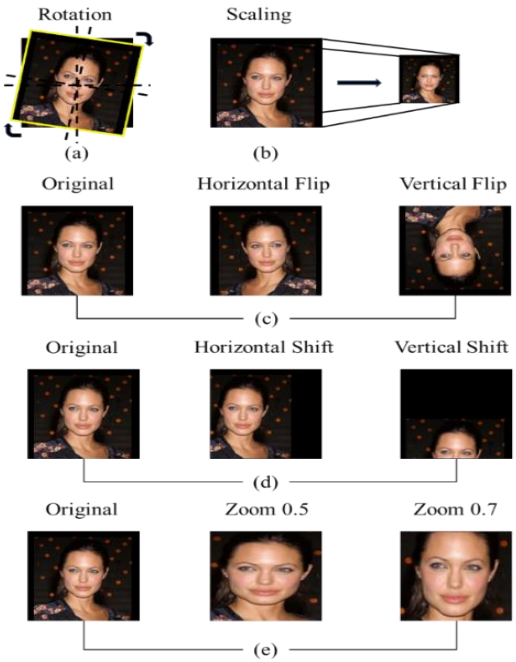


Figure. 6 Five techniques for data augmentation

The classification accuracy of each category on the CelebA and LFWA datasets is reported in Table 1 and 2, respectively. Fig. 7 and 8 show that the proposed Network outperforms the contend networks and achieves the average accuracy of 95.29% and 97.93% on CelebA and LFWA datasets, respectively. Compared with PANDA and LNets+ANet that used a single SVM classifier for each category, Network achieves superior performance by exploiting the multi-label learning with improvements of 9.86%, 7.96%, 16.9% and 14.3% on CelabA and LFWA datasets, respectively. The proposed Network also achieves better performance than NSA, GNAS and MCFA on both datasets; it achieves a gain of 4.68%, 3.66%, and 4.06% on CelebA dataset, respectively, while outperforming NSA by 12.11%, GNAS by 11.56%, and MCFA by 14.3% on LFWA dataset, respectively. In addition, The MOON and AW-CNN networks did not provide results on the LFWA dataset; therefore, the comparison is conducted only

on the CelebA dataset for these networks. As a result, the proposed Network achieves better performance than MOON and AW-CNN by a difference of 4.35% and 3.49%, respectively.

The proposed network leverages from dividing 40 datasets categories into five groups (i.e., hair, face, style, accessories, and appearance). It jointly learns features of each group of attributes independently according to their species, location or both together.

Table 1. The classification accuracy (in %) achieved by all the contend networks on the CelebA dataset. The accuracy of each category achieved by the proposed network is highlighted in bold. '-' indicates that the network does not provide the corresponding result of the category

CelebA	PANDA [11]	LNets+ANet [12]	MOON [13]	NSA [16]	MCNN-AUX [21]	MCFA [18]	GNAS [19]	AW-CNN [34]	PS-MCNN [23]	DMM-CNN [24]	DMTL [22]	**Proposed Network**
5 o'clock Shadow	88	91	94.03	93.1	94.51	94	94.76	-	96.6	94.84	95	**97.34**
Arched Eyebrows	78	79	82.26	82.6	83.42	83	84.25	-	85.77	84.57	86	**91.41**
Attractive	81	81	81.67	82.8	83.06	83	83.06	-	84.39	83.37	85	**89.06**
Bags Under Eyes	79	79	84.92	84.9	84.92	85	85.87	-	87.29	85.81	85	**91.73**
Bald	96	98	98.77	98	98.9	99	98.96	-	99.41	99.03	99	**99.63**
Bangs	92	95	95.8	95.7	96.05	96	96.2	-	98	96.22	99	**98.6**
Big Lips	67	68	71.48	69.3	71.47	72	71.79	-	73.13	72.93	96	**86.91**
Big Nose	75	78	84	83.8	84.53	84	85.1	-	86.4	84.78	85	**91.44**
Black Hair	85	88	89.4	89	89.78	89	90.24	-	91.66	90.5	91	**94.75**
Blond Hair	93	95	95.86	95.8	96.01	96	96.11	-	97.93	96.13	96	**98.28**
Blurry	86	84	95.67	96	96.17	96	96.42	-	98	96.4	96	**98.04**
Brown Hair	77	80	89.38	88.3	89.15	88	89.75	-	91.03	89.46	88	**92.03**
Bushy Eyebrows	86	90	92.62	92.7	92.84	92	92.99	-	94.51	93.01	92	**95.93**
Chubby	86	91	95.44	94.9	95.67	96	95.93	-	97.66	95.86	96	**98.26**
Double Chin	88	92	96.32	95.8	96.32	96	96.48	-	98.29	96.39	97	**98.68**
Eyeglasses	98	99	99.47	99.5	99.63	100	99.69	-	99.85	99.69	99	**99.9**
Goatee	93	95	97.04	96.7	97.24	97	97.59	-	97.74	97.63	99	**98.91**
Gray Hair	94	97	98.1	97.5	98.2	98	98.37	-	98.66	98.27	98	**99.4**
Heavy Makeup	90	90	90.99	91.6	91.55	92	91.82	-	93.31	91.85	92	**96.17**
High Cheekbones	86	88	87.01	87.6	87.58	87	88.05	-	89.5	87.73	88	**92.85**
Male	97	98	98.1	98	98.17	98	98.5	-	98.81	98.29	98	**99.36**
Mouth Open	93	92	93.54	93.8	93.74	93	94.16	-	95.99	94.16	94	**97.36**
Mustache	93	95	96.82	95.9	96.88	97	97.03	-	98.56	97.03	97	**99.07**
Narrow Eyes	84	81	86.52	86.9	87.23	87	87.66	-	89.07	87.73	90	**94.78**
No Beard	93	95	95.58	96.2	96.05	96	96.3	-	98.03	96.41	97	**98.34**
Oval Face	65	66	75.73	74.9	75.84	75	75.57	-	77.43	75.89	78	**84.06**
Pale Skin	91	91	97	97	97.05	97	97.24	-	98.84	97	97	**98.95**
Pointy Nose	71	72	76.46	76.5	77.47	77	78.24	-	79.32	77.19	78	**85.22**
Receding Hairline	85	89	93.56	92.3	93.81	94	93.94	-	95.85	94.12	94	**97.04**
Rosy Cheeks	87	90	94.82	94.8	95.16	95	95.01	-	96.92	95.32	96	**98.16**
Sideburns	93	96	97.59	97.2	97.85	98	97.96	-	98.22	97.91	98	**99.26**
Smiling	92	92	92.6	92.7	92.73	93	93.24	-	94.85	93.22	94	**96.86**
Straight Hair	69	73	82.26	80.4	83.58	85	84.77	-	85.96	84.72	85	**88.17**
Wavy Hair	77	80	82.47	81.7	83.91	85	84.52	-	86.39	86.01	87	**89.00**
Wearing Earrings	78	82	89.6	89.4	90.43	90	90.98	-	92.66	90.78	91	**93.83**
Wearing Hat	96	99	98.95	98.7	99.05	99	99.12	-	99.43	99.12	99	**99.76**
Wearing Lipstick	93	93	93.93	93.2	94.11	94	94.41	-	95.7	94.49	93	**97.14**
Wearing Necklace	67	71	87.04	85.6	86.63	88	87.61	-	88.98	88.03	89	**93.05**
Wearing Necktie	91	93	96.63	96.1	96.51	97	96.76	-	98.52	97.15	97	**98.76**
Young	84	87	88.08	88	88.48	88	88.89	-	90.54	88.98	90	**94.21**
Accuracy (average)	85.43	87.33	90.94	90.6	91.29	91.23	91.63	91.8	92.98	91.7	92.6	**95.29**

Table 2. The classification accuracy (in %) achieved by all the contend networks on the LFWA dataset. The accuracy of each category achieved by the proposed network is highlighted in bold. '-' indicates that the network does not provide the corresponding result of the category

LFWA	PANDA [11]	LNets+ANet [12]	NSA [16]	MCNN-AUX [21]	MCFA [18]	GNAS [19]	PS-MCNN [23]	DMM-CNN [24]	DMTL [22]	**Proposed Network**
5 o'clock Shadow	84	84	77.6	77.06	75	-	78.17	79.18	80	**95.54**
Arched Eyebrows	79	82	81.7	81.78	79	-	83.53	82.7	86	**96.76**
Attractive	81	83	80.2	80.31	77	-	81.84	81.1	82	**96.23**
Bags Under Eyes	80	83	82.6	83.48	79	-	86.74	82.7	84	**96.85**
Bald	84	88	91.9	91.94	91	-	92.6	91.96	92	**99.45**
Bangs	84	88	90.7	90.08	89	-	91.45	91.3	93	**99.52**
Big Lips	73	75	79	79.24	75	-	82.7	79.82	77	**96.08**
Big Nose	79	81	83.1	84.98	81	-	86.48	83.67	83	**97.76**
Black Hair	87	90	92.5	92.63	91	-	92.96	91.55	92	**99.63**
Blond Hair	94	97	97.5	97.41	97	-	98.51	97.17	97	**99.88**
Blurry	74	74	86.4	85.23	86	-	87.2	87.58	89	**98.45**
Brown Hair	74	77	80.9	80.85	77	-	81.87	81.56	81	**97.33**
Bushy Eyebrows	79	82	84.3	84.97	76	-	85.72	85.33	80	**96.8**
Chubby	69	73	76.1	76.86	74	-	78.11	77.66	75	**95.88**
Double Chin	75	78	80.5	81.52	77	-	86.7	80.98	78	**96.39**
Eyeglasses	89	95	91.5	91.3	91	-	92.78	92.83	92	**99.42**
Goatee	75	78	83	82.97	80	-	84.11	82.82	86	**96.85**
Gray Hair	81	84	88.5	88.93	88	-	91.04	89.38	88	**99.37**
Heavy Makeup	93	95	95.4	95.85	94	-	96.6	95.68	95	**99.48**
High Cheekbones	86	88	88.3	88.38	85	-	88.77	88.13	89	**98.18**
Male	92	94	92.6	94.02	93	-	95.18	94.14	93	**99.22**
Mouth Open	78	82	82.5	83.51	78	-	84.6	84.45	86	**95.44**
Mustache	87	92	93	93.43	91	-	94.47	94.46	95	**99.38**
Narrow Eyes	73	81	82.8	82.86	78	-	83.51	83.67	82	**96.65**
No Beard	75	79	80.8	82.15	79	-	82.01	82.48	81	**97.18**
Oval Face	72	74	76.8	77.39	74	-	77.9	76.94	75	**95.74**
Pale Skin	84	84	91	93.32	82	-	94.97	91.86	91	**99.64**
Pointy Nose	76	80	84.2	84.14	80	-	87.52	84.51	84	**97.86**
Receding Hairline	84	85	84.9	86.25	85	-	87.5	86.3	85	**98.88**
Rosy Cheeks	73	78	87.1	87.92	85	-	88.81	86.44	86	**98.97**
Sideburns	76	77	81.8	83.13	78	-	84.42	82.99	80	**96.88**
Smiling	89	91	90.8	91.83	88	-	92.7	92.24	92	**98.95**
Straight Hair	73	76	78.9	78.53	77	-	79.65	79.2	79	**96.07**
Wavy Hair	75	76	78.3	81.61	79	-	83.35	79.87	80	**97.12**
Wearing Earrings	92	94	94.8	94.95	93	-	95.54	94.14	94	**99.38**
Wearing Hat	82	88	90.2	90.07	91	-	91.21	90.84	92	**99.52**
Wearing Lipstick	93	95	94.1	95.04	94	-	95.7	95.11	93	**99.59**
Wearing Necklace	86	88	89.6	89.94	89	-	90.92	89.47	91	**99.02**
Wearing Necktie	79	79	81.4	80.66	82	-	82.18	81.28	81	**98.02**
Young	82	86	85.7	85.84	87	-	86.88	88.94	87	**98.08**
Accuracy (average)	81.03	83.85	85.8	86.31	83.63	86.37	87.67	86.56	86.15	**97.93**

In contrast, the NSA, GNAS, MCFA, MOON, and AW-CNN learn the standard features among the 40 categories. Compared to MCNN-AUX, PS-MCNN-LC, DMTL, and DMM-CNN adopt different criteria (i.e., location, relationship or heterogeneity) to divide the datasets categories. The proposed network outperformed the benchmark networks on all categories, which validates the effectiveness of the grouping method used in this study. For the CelebA dataset, the proposed network outperformed the

Received: April 25, 2022. Revised: June 21, 2022.

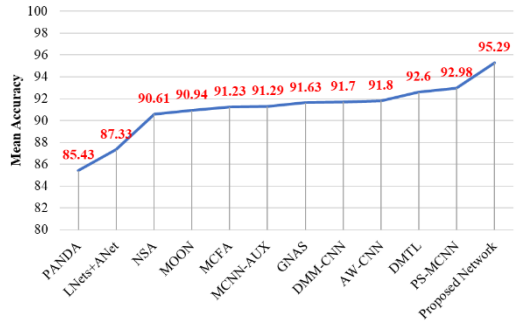


Figure. 7 Performance comparison between contend networks on the CelebA dataset

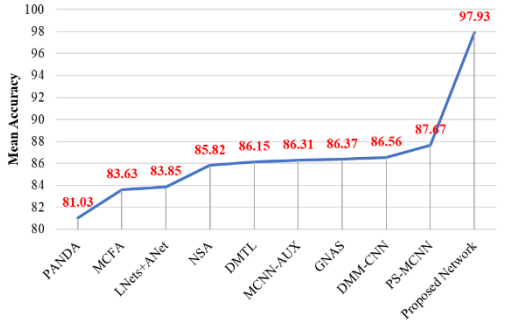
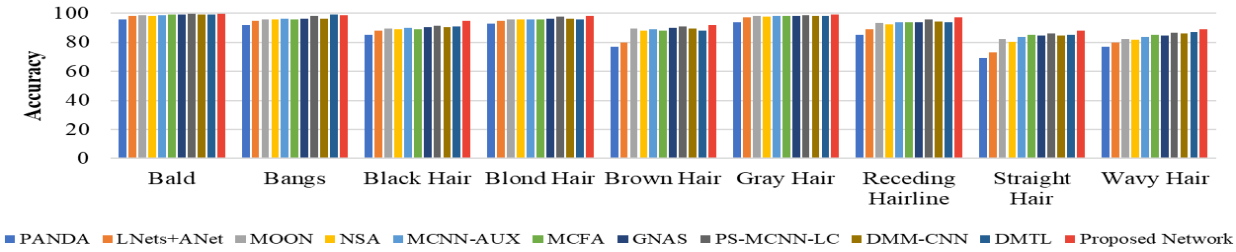
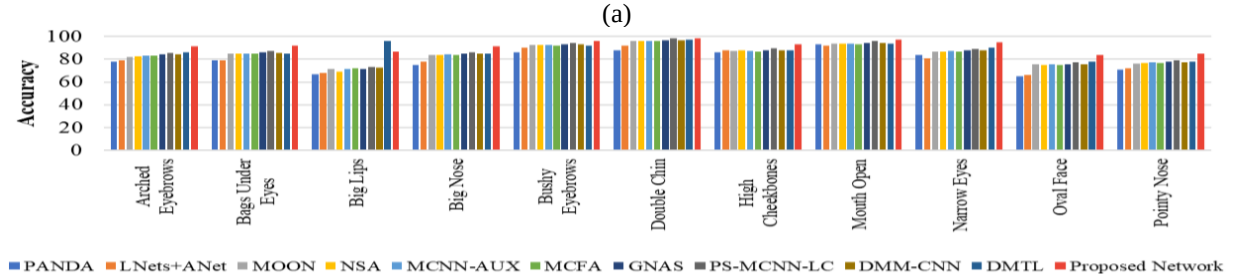


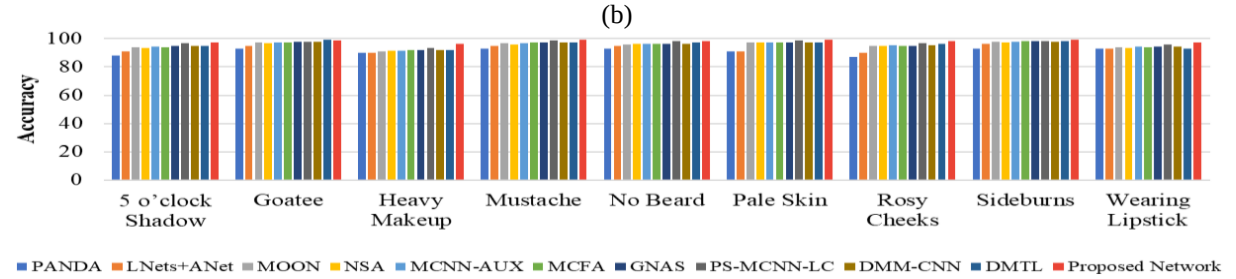
Figure. 8 Performance comparison between contend networks on the LFWA dataset



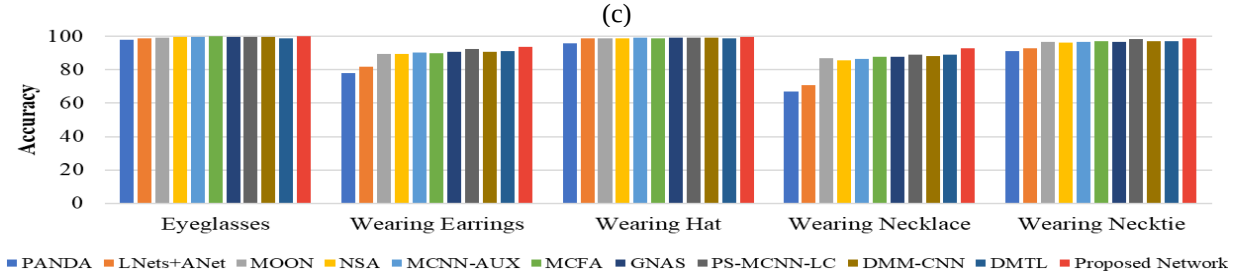
(a)



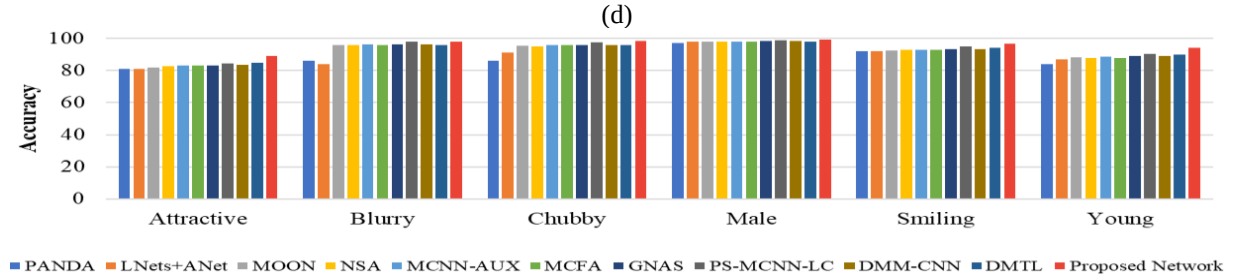
(b)



(c)

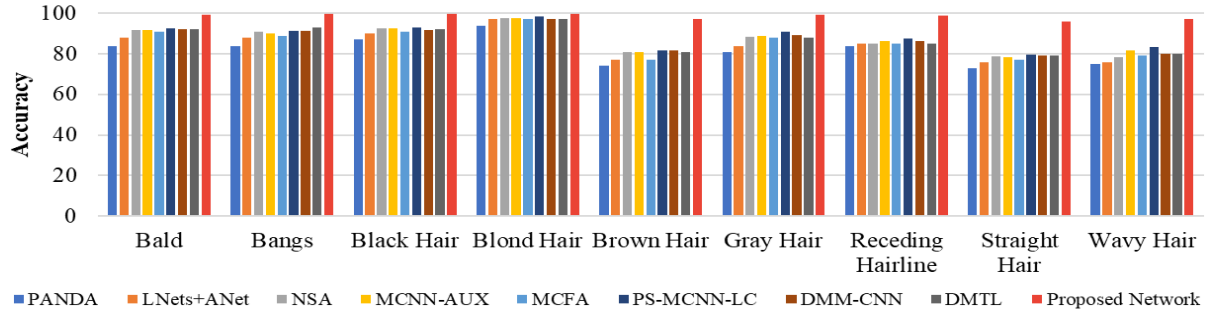


(d)

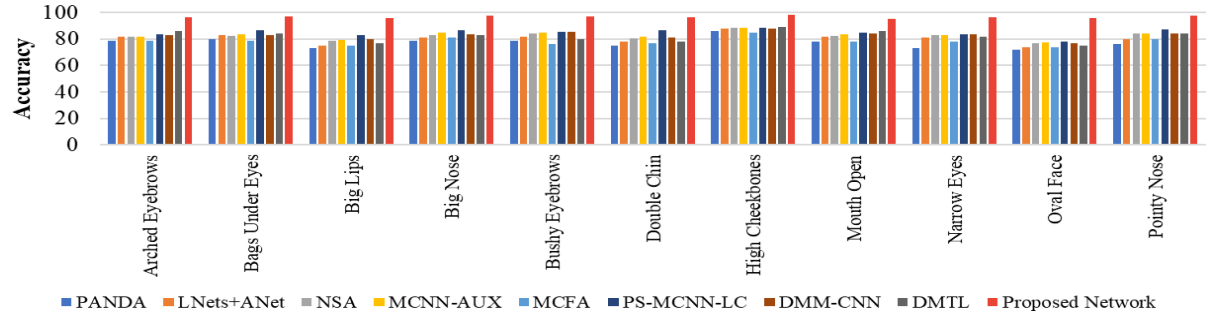


(e)

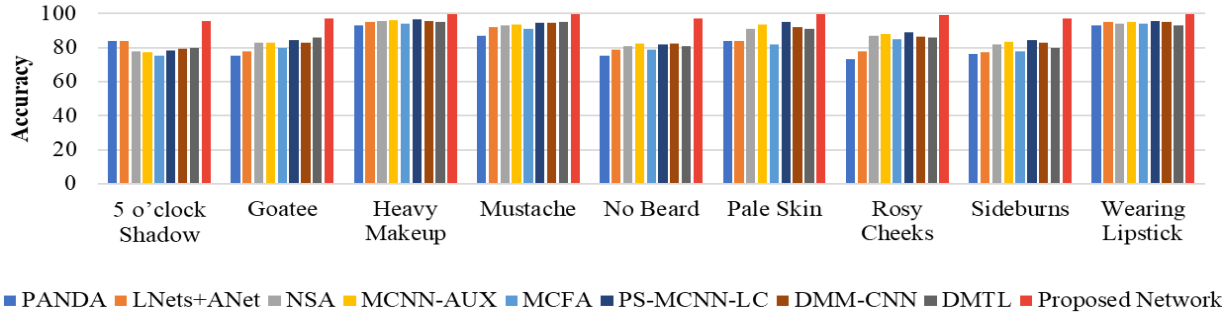
Figure. 9 Comparison of CelebA dataset accuracy based on groups: (a) hair, (b) face, (c) style, (d) accessories, and (e) appearance



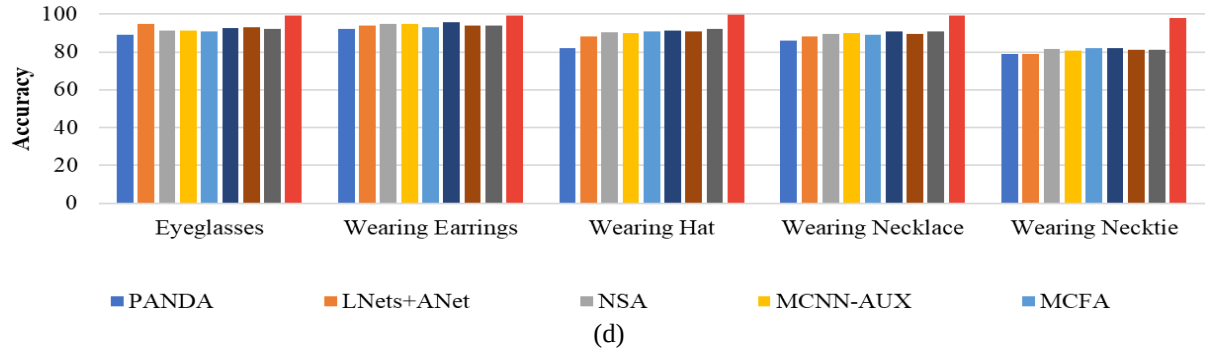
(a)



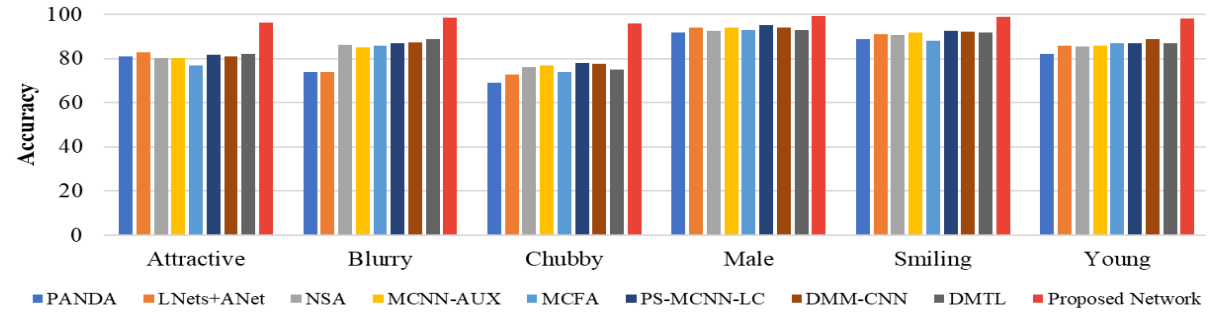
(b)



(c)



(d)



(e)

Figure. 10 Comparison of LFWA dataset accuracy based on groups: (a) hair, (b) face, (c) style, (d) accessories, and (e) appearance

MCNN-AUX by 4%, PS-MCNN-LC by 2.31%, DMTL by 2.69% and DMM-CNN by 3.59%, and the corresponding accuracy improvements on LFWA dataset are 11.62%, 10.26%, 11.78% and 11.37%, respectively.

To further demonstrate the effectiveness of the proposed network, the study compared the performance of the network on the categories of each group with the performance of contending networks. Fig. 9 and 10 presents the performance comparison on the categories of each group for all the contending networks on CelebA and LFWA datasets, respectively. On the CelebA dataset, the proposed Network outperforms the contending networks in all categories groups, except the DMTL network, which obtained a higher accuracy for two categories only (i.e., bangs and big lips). In contrast, the proposed Network significantly outperforms the contending networks in all categories groups without exception on the LFWA dataset. The significant performance of the network on LFWA is due to the techniques (i.e., data augmentation and transfer learning) used in training the proposed Network. Furthermore, the proposed Network shows strength in appearance categories such as attractive, chubby and young, and fine-grained categories such as pointy nose, five o clock shadow, and Mustache.

6. Conclusions

This paper presents a new multi-output deep learning network for HHAC. The proposed network can effectively improve the performance of existing HHAC by learning the joint features among the attributes according to their common characteristics. Based on the proposed attributes groups (i.e., hair, face, style, accessories, and appearance), the proposed network is built upon two convolutional blocks of feature learning and five output layers to predict the attributes of each group. Extensive experiments on the CelebA and LFWA benchmark datasets showed that the proposed network achieved superior performance over recent HHAC networks. The proposed network achieved an average accuracy of 95.29% on and 97.93% on the CelebA and LFWA datasets, superior by 2% and 10% to the competition's HHAC networks. Moreover, the proposed network outperforms the competition's networks in the classification accuracy of almost all human head attributes. Finally, it is noteworthy to note that the outstanding performance of the proposed network is due to the attributes grouping method and the structure of the proposed network. The attributes grouping allows learning the explicit correlations between the attributes with similar components and extracting their common features, which helps improve classification accuracy. Besides, the structure of the proposed network includes a single output layer for each group of attributes, allowing feature learning and attributes prediction for each group independently. In the future, this study can extend to develop assistive visual recognition applications such as visually impaired systems and robotic vision systems.

Conflicts of Interest

The authors declare no conflict interest.

Author Contributions

The article conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing original draft preparation, writing review, editing, and visualization, have been done by 1st author. The supervision and study administration have been done by the 2nd and 3rd authors.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”, In: *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection”, In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[3] X. Wu, R. He, Z. Sun, and T. Tan, “A Light CNN for Deep Face Representation With Noisy Labels”, *IEEE Transactions on Information Forensics and Security*, Vol. 13, No. 11, pp. 2884-2896, 2018.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition”, In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690-4699, 2019.

[5] G. B. Huang, H. Lee, and E. L. Miller, “Learning hierarchical representations for face verification with convolutional deep belief networks”, In: *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2518-2525, 2012.

[6] Z. Fan and Y. P. Guan, “A deep learning framework for face verification without alignment”, *Journal of Real-Time Image Processing*, Vol. 18, No. 4, pp. 999-1009, 2021.

[7] M. Zhang, X. Zhe, S. Chen, and H. Yan, "Deep center-based dual-constrained hashing for discriminative face image retrieval", *Pattern Recognition*, Vol. 117, p. 107976, 2021.

[8] R. R. Saritha, V. Paul, and P. G. Kumar, "Content based image retrieval using deep learning process", *Cluster Computing*, Vol. 22, No. 2, pp. 4187-4200, 2019.

[9] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep Image Retrieval: Learning Global Representations for Image Search", *Computer Vision – ECCV 2016*, pp. 241-257.

[10] A. M. Elkahky, Y. Song, and X. He, "A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems", In: *Proc. of the 24th International Conference on World Wide Web*, pp. 278-288, 2015.

[11] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose Aligned Networks for Deep Attribute Modeling", In: *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637-1644, 2014.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild", In: *Proc. of the IEEE International Conference on Computer Vision*, pp. 3730-3738, 2015.

[13] E. M. Rudd, M. Günther, and T. E. Boult, "MOON: A Mixed Objective Optimization Network for the Recognition of Facial Attributes", *Computer Vision – ECCV 2016.* pp. 19-35.

[14] K. He, Z. Wang, Y. Fu, R. Feng, Y. G. Jiang, and X. Xue, "Adaptively Weighted Multi-task Deep Network for Person Attribute Classification", In: *Proc. of the 25th ACM International Conference on Multimedia*, pp. 1636-1644, 2017.

[15] H. Guo, X. Fan, and S. Wang, "Human attribute recognition by refining attention heat map", *Pattern Recognition Letters*, Vol. 94, pp. 38-45, 2017.

[16] U. Mahbub, S. Sarkar, and R. Chellappa, "Segment-Based Methods for Facial Attribute Detection from Partial Faces", *IEEE Transactions on Affective Computing*, Vol. 11, No. 4, pp. 601-613, 2018.

[17] M. Xu, F. Chen, L. Li, C. Shen, P. Lv, B. Zhou, and R. Ji, "Bio-Inspired Deep Attribute Learning Towards Facial Aesthetic Prediction", *IEEE Transactions on Affective Computing*, Vol. 12, No. 1, pp. 227-238, 2018.

[18] N. Zhuang, Y. Yan, S. Chen, and H. Wang, "Multi-task Learning of Cascaded CNN for Facial Attribute Classification", In: *Proc. of 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2069-2074, 2018.

[19] S. Huang, X. Li, Z. Q. Cheng, Z. Zhang, and A. Hauptmann, "GNAS: A Greedy Neural Architecture Search Method for Multi-Attribute Learning", In: *Proc. of the 26th ACM International Conference on Multimedia*, pp. 2049-2057, 2018.

[20] A. M. Andrew, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods by Nello Christianini and John Shawe-Taylor", *Robotica*, Vol. 18, No. 6, pp. 687-689, 2000.

[21] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network for attribute classification", *arXiv Preprint arXiv:1604.07360*, pp. 1-16, 2016.

[22] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 11, pp. 2597-2609, 2018.

[23] J. Cao, Y. Li, and Z. Zhang, "Partially Shared Multi-task Convolutional Neural Network with Local Constraint for Face Attribute Learning", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4290-4299, 2018.

[24] L. Mao, Y. Yan, J. Xue, and H. Wang, "Deep Multi-task Multi-label CNN for Effective Facial Attribute Classification", *IEEE Transactions on Affective Computing*, pp. 1-1, 2020.

[25] K. Fukushima, "A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", *Biol. Cybern.*, Vol. 36, pp. 193-202, 1980.

[26] Y. Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural Computation*, Vol. 1, No. 4, pp. 541-551, 1989.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition", *British Machine Vision Association*, Vol. 1, No. 3, pp. 1-12, 2015.

[28] Y. Yuan, L. Mou, X. J. I. T. O. N. N. Lu, and L. systems, "Scene recognition by manifold regularized deep learning architecture", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 10, pp. 2222-2233, 2015.

[29] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search", In: *Proc. of the 23rd International Conference on World Wide Web*, pp. 373-374, 2014.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge", *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252, 2015.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, pp. 1097-1105, 2012.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", In: *Proc. of International Conference on Learning Representations*, San Diego, CA, USA, Vol. 6, pp. 1-14, 2014.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[35] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", In: *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708, 2017.

[36] M. D. Zeiler, "Hierarchical convolutional deep learning in computer vision", *New York University*, 2013.

[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning: MIT press*, 2016.

[38] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification", *Advances in neural information processing systems*, pp. 1988-1996, 2014.

[39] G. B. Huang, M. Mattar, T. Berg, and E. L. Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments", In: *Proc. of Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Amherst, pp. 07-49, 2008.