Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06535-0

S.I. : EMERGING TRENDS IN AI & ML

MEBeauty: a multi-ethnic facial beauty dataset in-the-wild

Irina Lebedeva[1] · Yi Guo[1] · Fangli Ying[1]

Received: 31 March 2021 / Accepted: 9 September 2021

Abstract

Although facial beauty prediction (FBP) has achieved high accuracy on images captured in a constrained environment, it is still a challenging task on face images in-the-wild. Moreover, there is no FBP benchmark dataset that includes images with various ethnic, age, gender properties and unrestricted in terms of face expression and pose. In this work, the issue of FBP in real-life scenario is addressed and a multi-ethnic facial beauty dataset, namely MEBeauty, is introduced. All face images are captured in an unconstrained environment and rated by volunteers with various ethnicity, age and gender in order to avoid any cultural and social biases in beauty perception. Different well-known CNNs with layer-wise transfer learning are performed on the dataset. Moreover, the evaluation of knowledge learning from the face recognition task across FBP is conducted. The expected high number of aberrant and outlier faces is considered and the use of various robust loss functions in order to learn deep regression networks for facial beauty prediction is evaluated. Several FBP frameworks are performed on the proposed dataset and widely-used SCUT-FBP 5500 in order to compare their effectiveness on face images in constrained and unconstrained environments.

Keywords Facial beauty prediction · Deep learning · Robust loss functions · Knowledge transfer · Regression

1 Introduction

A primary task in Computer Vision is to reproduce the human ability to recognize various visual concepts. One such concept is facial beauty, and to let a machine predict attractiveness is a challenging task. Facial beauty plays an important role in the face-to-face communication between humans, and influences one person's attraction to another. Therefore, it impacts many aspects of life.

Automatic facial beauty assessment has application potential in makeup recommendation [27], aesthetic surgery [51], face beautification [25], content-based image retrieval [29] and social media recommendation. A range of facial beauty prediction methods have already been proposed, including shallow predictors [1, 22] and deep-learning based approaches [18, 42]. Although the latter achieved relatively high accuracy, the performance still remains unsatisfied on face images captured in-the-wild conditions. Moreover, there is no benchmark on facial beauty datasets with images of various gender, ethnicities, age and expressions that can allow to create and evaluate FBP in real-life conditions taking into consideration specific ethnic attributes.

All existing datasets have different limitations. The main bottleneck of the earliest ones is the small number of face images [1, 22]. It only allows the evaluation of an FBP method from a very narrow perspective and, additionally, can not be effective in the combination with a data-driven method. The later datasets contain hundreds or thousands of face images. However, some of these datasets include faces of only one or two ethnicities (mostly Asian and Caucasian) [26], or one gender [19, 30]. Others contain images captured in a constrained environment, in other words, have restrictions on pose, expression, lightning and background. All of these bottlenecks make facial beauty prediction effective only on a limited number of images.

Since the perception of facial beauty is highly subjective, image quantity and quality are not the only important characteristics of an FBP dataset. The number of raters and their demographic information, e.g., age, gender and ethnicity, are also critical. The beauty evaluation limited to

✉ Irina Lebedeva
irina@mail.ecust.edu.cn

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Published online: 04 October 2021

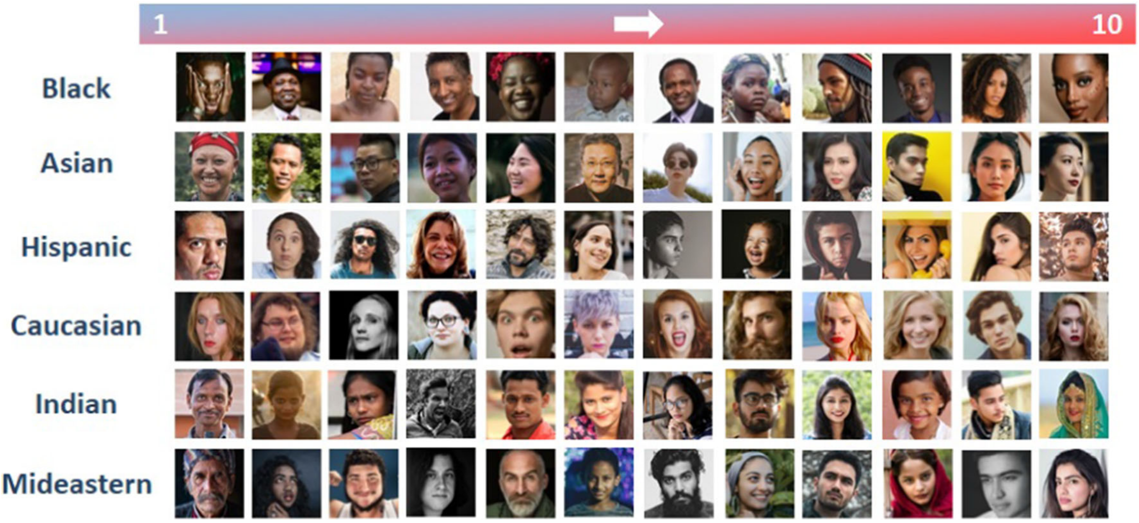


Fig. 1 Image samples of the MEBeauty dataset. Face photos of Caucasian, Asian, Black, Indian, Hispanic and Mideastern ethnic groups are presented in the dataset. The images are rated from 1 to 10, where 1 means not attractive and 10-very attractive face

one or few ethnic, age or gender groups can result in biased attractiveness assessment [2]. Unfortunately, face images in existing beauty datasets are usually rated by one ethnic group [16, 26] (Asian), one age group [28, 35] (20's). Moreover, the use of raters' demographic information in prediction modeling potentially improves its accuracy. Existing datasets include only scores given to an image, but there is no information about the raters' age, gender and ethnicity. In addition, there is another problem that restricts the research in the direction of facial beauty prediction—most FBP datasets are not publicly available for various reasons, including the privacy issue [35].

In this work, a multi-ethnic dataset, namely MEBeauty, that overcomes all issues above, is introduced. The dataset includes Black, Indian, Asian, Hispanic, Mideastern and Caucasian faces with a rich diversity in age, gender, face expression and pose. All images are captured in-the-wild and rated by several hundred representatives of different ethnic, gender and age groups. Figure 1 demonstrates MEBeauty sample images grouped by ethnicity and sorted by beauty score. Moreover, the MEBeauty dataset contains all scores given by all raters and the demographic information of a part of raters including their age, gender and ethnicity. Thus, the dataset can be useful to study the personalized aspect of facial beauty or personal beauty preferences prediction. The MEBeauty dataset can be found by following this link.[1]

Different convolutional neural networks (CNN) with layer-wise transfer learning are evaluated on the proposed dataset in order to find the most sufficient approach. Moreover, knowledge transfer from another face related task is evaluated for facial beauty prediction. Several FBP frameworks are performed on MEBeauty and the most widely-used facial beauty dataset, namely SCUT-FBP 5500, in order to compare their effectiveness on face images in constrained and unconstrained conditions.

In addition, since the dataset is highly diverse, the number of aberrant and outlier faces is expected to be high. Thus, the use of different robust loss functions to learn deep regression networks for facial beauty prediction is evaluated.

The main contributions of our work are the following:

1. A multi-ethnic facial beauty dataset, namely MEBeauty, that includes Black, Indian, Asian, Hispanic, Mideastern and Caucasian faces with a rich diversity in age, gender, face expression and pose, is collected.
2. The face images are rated by several hundred volunteers with different social and cultural backgrounds. The comprehensive statistical analysis of MEBeauty and its comparison to the other FBP datasets in many aspects are presented.
3. Different well-known CNN architectures with layer-wise transfer learning are evaluated on MEBeauty. Knowledge transfer from the face recognition task across FBP is also performed.
4. Several FBP frameworks are evaluated on the proposed dataset and widely-used SCUT-FBP 5500 in order to compare their effectiveness in both constrained and unconstrained environments.
5. Since the proposed dataset has a rich diversity, the influence of aberrant and outlier faces is considered and the use of robust loss functions in deep regression networks for facial beauty prediction is explored.

[1] https://github.com/fbplab/ME-beautydatabase.

6. The performance evaluation of the proposed FBP method is conducted separately by various ethnic and gender groups.

2 Related works

In this section, the classification and description of existing facial beauty prediction approaches are first presented. Then, the comprehensive analysis and comparison of facial beauty datasets are provided.

2.1 Facial beauty prediction methods

Since dozens of approaches have been proposed in order to automatically assess facial attractiveness, the illustrated classification of beauty evaluation methods can help to better understand the current situation in this area. Figure 2 demonstrates that two main machine learning concepts have been applied so far: traditional machine learning and deep learning. In most studies, facial beauty prediction is considered as a fully supervised task. Only a few works with a semi-supervised approach have been presented recently. This subsection outlines the classification of FBP methods in further details.

The earliest attempts to predict facial beauty are based on the combination of shallow predictors and geometric features, such as facial ratios and landmark distances [7, 16, 22, 34]. Textural features, like Gabor [8] and SIFT descriptor [48], are also exploited to automatically assess attractiveness. However, most of these works require manually labeled or adjusted features that makes the preprocessing of these approaches time-consuming. Another shortcoming of this line of FBP methods is the low accuracy produced by traditional machine learning in combination with the features mentioned above. Additionally, these approaches are not able to demonstrate their effectiveness on face images in-the-wild.

Significant results in facial beauty prediction have been achieved due to the advancement in deep learning [18, 42]. New CNN architectures are specifically designed for automatic attractiveness assessment [44, 45]. Deep facial features extracted by a CNN are utilized to predict personal beauty preferences in [23]. A three-level residual-in-residual network structure for better information transmission is also introduced for automatic beauty assessment [4]. Recently transfer learning has been widely used for facial beauty prediction [24, 36, 47]. A range of pre-trained CNN models showed their effectiveness for the attractiveness evaluation task, e.g., VGG16 [32], ResNeXt-50 [26]. A CNN for attractiveness assessment on 3D faces with both geometry and texture information is introduced in [41]. Although deep learning based FBP methods demonstrate high prediction accuracy, there are a range of restrictions in their performances. First of all, the approaches are able to show their effectiveness only on face images with a limited number of properties. They might assess attractiveness well on female faces, while demonstrating much less reliable results on male images. This

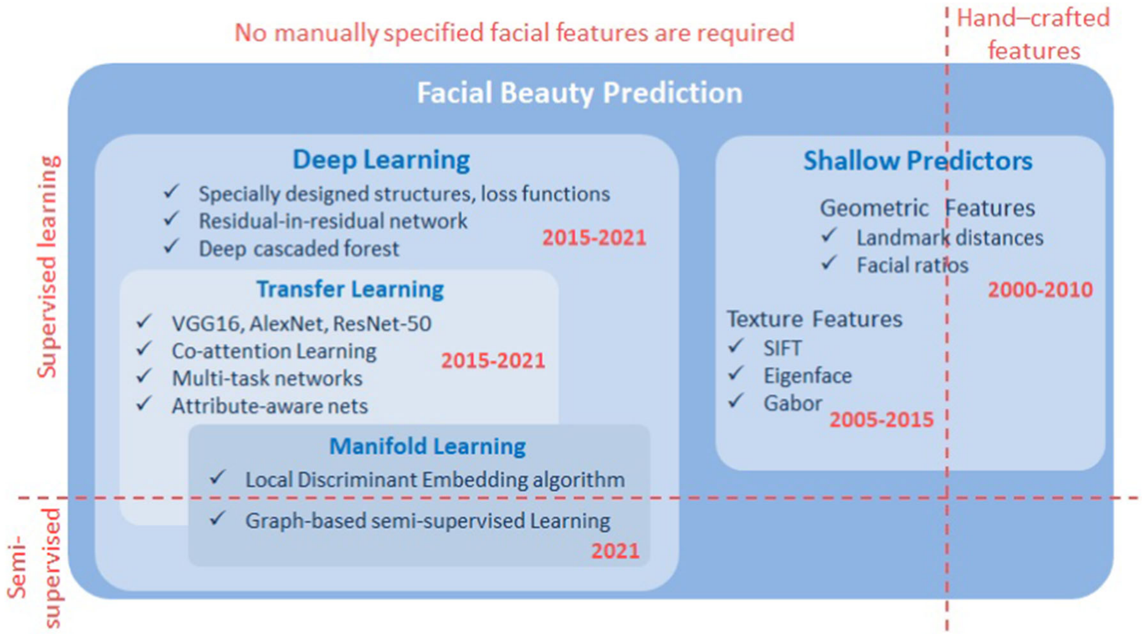


Fig. 2 The classification of facial beauty prediction methods. In most works, FBP is considered as a fully supervised task with no manual annotation required

issue is particularly critical for the ethnic diversity in face images. Secondly, existing deep learning based methods are sensitive to pose, expression, background varieties. In other words, these approaches are not able to demonstrate high performances on images in-the-wild.

Multi-task networks that simultaneously recognize gender, ethnicity and facial beauty from a face image is presented in [37, 43, 46]. Lately, manifold learning theory has been used for the facial beauty prediction task. A deep manifold-learning method with the supervised Local Discriminant Embedding algorithm is proposed in [14]. A range of manifold learning based FBP methods with the use of semi-supervised learning is also introduced [12, 13, 15]. An augmented reality tool for facial attractiveness evaluation is presented in [38]. The dynamic attractiveness prediction in short videos is studied in [39]. Since multi-task, manifold based and other methods mentioned in this paragraph are data-driven, a dataset used for training and its properties play a significant role in the performances of these approaches. As a result, these methods have shortcomings similar to deep learning based methods, including low performances on face images in unconstrained environments with ethic, age, gender diversities.

2.2 Facial beauty datasets

The analysis and comparison of existing facial beauty datasets are presented in Table 1, where the datasets are sorted by their publication year and the sign ”-” means that the related work has no information about the particular characteristic.

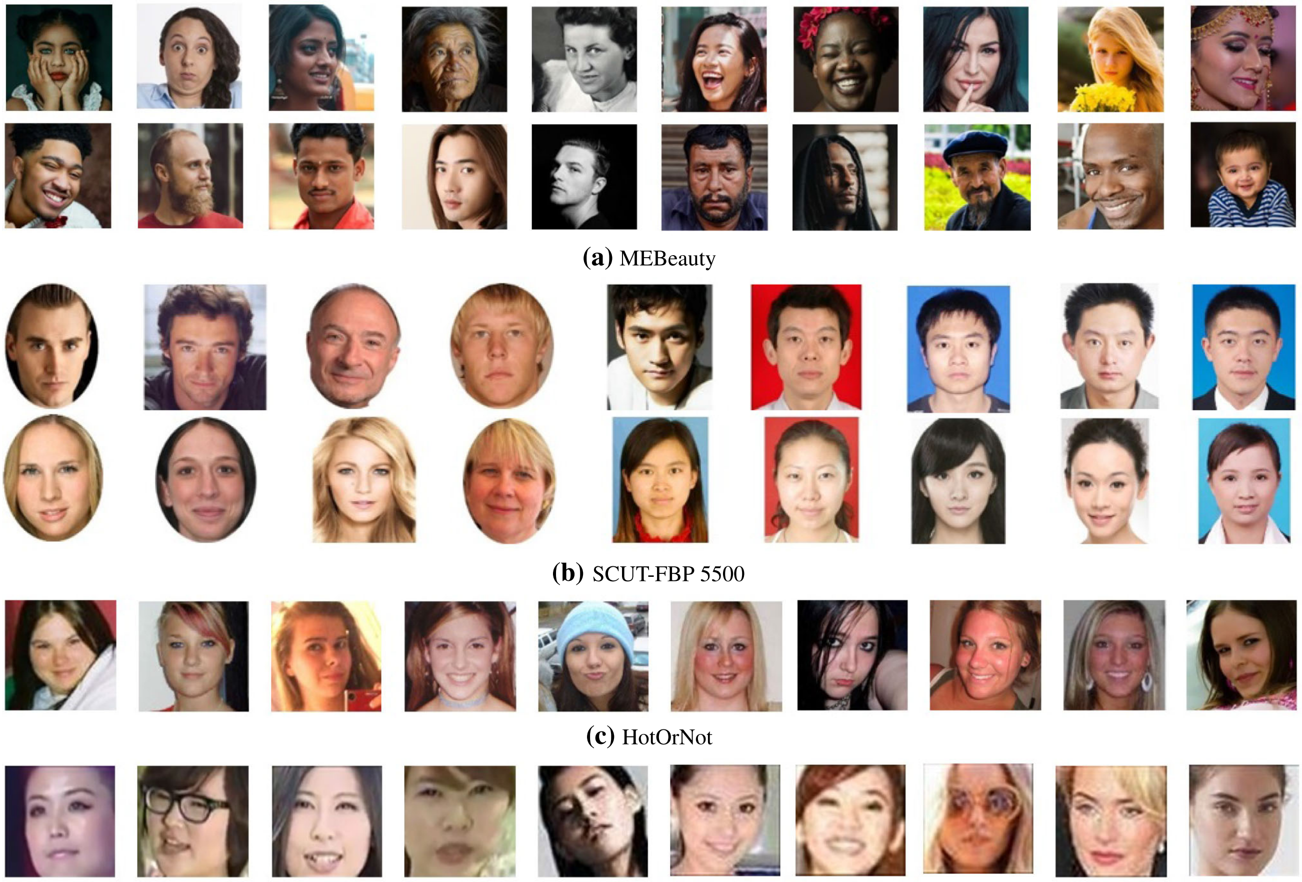
The earliest datasets collected for facial attractiveness assessment contain 200 or less face images [1, 16, 20, 22]. This results in low performance and strong bias, especially in the deep learning era. In [17], a dataset with 432 computer-generated female faces with the same expression, hairstyle, skin tone is introduced. The later datasets consist of 10 times more face images, but they have different restrictions. Some datasets are restricted in terms of face pose and contain only frontal faces [7, 28], while others are limited to only one gender and include only female images [19, 30].

Almost all FBP datasets are ethnicity-limited. Zhai [50] collected a dataset with more than 20 000 faces, but all of them are Asian looking. The SCUT-FBP 500 dataset [42] also consists of Asian faces. Its extended version SCUT-FBP 5500 contains Asian, Caucasian female and male faces with neutral expressions captured in a constrained environment. The image samples of publicly available

Table 1 Evaluation of facial beauty datasets and their key characteristics

Dataset	Publication year	Beauty category	Image number	Rater number	Unconstrained environment	Rater info	Publicly available
Aarabi et al. [1]	2001	1–4	80	12	×	×	×
Gunes et al. [20]	2004	1–10	215	48	✔	×	×
Eisenthal et al. [16]	2006	1–7	184	18/28	×	×	×
Kagian et al. [22]	2007	1–7	91	28	×	×	×
Whitehill et al. [40]	2008	−1–2	1000	8	×	×	×
Chang et al. [6]	2009	1–7	160	57	×	×	×
Mao et al. [28]	2009	1–4	510	–	×	×	×
Chen et al. [7]	2010	binary	875	–	×	×	×
Gray et al. [19]	2010	−3–3	2056	30	✔	×	✔
M2B [30]	2012	1–10	1240	40	✔	×	✔
Fan et al. [17]	2012	1–9	432g	30	×	×	×
De Vries et al. [10]	2015	binary	9364	–	✔	×	×
SCUT-FBP 500 [42]	2015	1–5	500	75	×	×	✔
LSFBD [50]	2016	0–4	20000	200	×	×	×
SCUT-FBP 5500 [26]	2018	1–5	5500	60	×	×	✔
Tong et al. [35]	2020	Binary	4512	20	–	×	×
LSAFBD [49]	2020		20 000	20	–	×	×
MEBeauty	2021	1–10	2550	300	✔	✔	✔

The datasets are sorted by their publication year and the sign ”-” means that the related work has no information about the particular characteristic



(a) MEBeauty

(b) SCUT-FBP 5500

(c) HotOrNot

(d) M2B

Fig. 3 Image samples from publicly available facial beauty datasets. SCUT-FBP5500 **b** contains female and male frontal faces, while Gray dataset **d** and M2B **e** contain female face captured in an unconstrained environment. **b**, **d** consist of Asian and Caucasian faces, while **c** only Caucasian faces. The MEBeauty dataset **a** combines the advantages of all the presented datasets

beauty datasets are shown in Fig. 3. Face and rater properties of FBP datasets are demonstrated in Table 2. Face images in some presented datasets are rated by only Asian volunteers [16, 26], while only raters in their 20's evaluated face images in [28, 35].

Besides, there are a study addressed to similar but not exactly the same task [31]. The dataset proposed in this work is not included to the table, but should be mentioned. The aim of this study is to estimate the photographic quality of the representation of the face, independent from the beauty of this face.

3 MEBeauty dataset

3.1 Images collection

Popular free photo sharing services have been used to collect human images.[2] All contents on the chosen services are copyright free and released under Creative Commons licenses[3] that makes the proposed dataset clean from the privacy and copyrights point of view. We select photographs manually in order to make the dataset as diverse as possible and avoid any bias in ethnicity, age, face expression, pose and background. More than 3000 photograhs of Black, Asian, Indian, Hispanic, Mideastern, Caucasian females and males of all ages are collected. Then faces are detected on the photographs and cropped with some background around them. This is done to concentrate a rater's attention on a face and avoid the body and clothes influence on a rater' beauty perception. The final dataset consists of 2550 face images of size 500 × 500. This relatively high resolution is used for further possible processing. The detailed ethnic, gender and age composition of the proposed multi-ethnic dataset, namely MEBeauty, and its statistical analysis are presented in Sect. 3.3. The dataset also contains all the links where each of the images can be found.

[2] https://pixabay.com/https://unsplash.com/. https://pexels.com/.

[3] https://creativecommons.org/.

Table 2 Face and rater properties of facial beauty datasets

Dataset	Face				Rater		
	Expression	Gender	Age	Ethnicity	Gender	Age	Ethnicity
Aarabi et al. [1]	–	f	–	–	–	–	–
Gunes et al. [20]	–	f	All	All	f m	All	All
Eisenthal et al. [16]	Neutral	f	Young	Caucasian	f m	–	Asian
Kagian et al. [22]	Neutral	f	Young	–	f m	–	–
Whitehill et al. [40]	All	f m	–	–	f m	20–35	Caucasian
Chang et al. [6]	Neutral	f m	18–30	Asian, Caucasian Indian, Black	f m	–	Asian
Mao et al. [28]	–	f	–	Asian	–	20–30	–
Chen et al. [7]	Neutral	f m	–	Asian	–	20–45	–
Gray et al. [19]	All	f	18–40	Caucasian	–	–	–
M2B et al. [30]	All	f	–	Caucasian Asian	–	–	Caucasian Asian
Fan et al. [17]	Neutral	f	–	Computer-generated	f m	20–25	Asian
De Vries et al. [10]	–	f m	–	–	f m	–	–
SCUT-fbp 500 [42]	Neutral	f	–	Asian	–	–	Asian
LSFBD [50]	All	f m	All	Asian	f m	20–35	–
SCUT-fbp 5500 [26]	Neutral	f m	–	Caucasian Asian	f m	18–27	Asian
Tong et al. [35]	–	f m	–	–	f m	20–30	–
LSAFBD [49]	–	f	–	Asian	f m	20–35	–
MEBeauty	All	f, m	All	Asian, Caucasian Indian, Black Hispanic, Mideastern	f m	All	Asian, Caucasian, Indian, Black, Hispanic, Mideastern

The attractiveness of Asian and Caucasian young female faces rated by the same ethnic groups are the most studied in the literature

3.2 Beauty score collection

For more objective beauty assessment of the faces, more than 300 human raters are involved in the beauty scoring. Moreover, in order to avoid any cultural and social biases, the images are rated by representatives of different ethnic, gender and age groups. In other words, volunteers of various ethnic, gender and age groups expressed their opinions about the beauty of each face in the proposed dataset to reduce cultural influence on the ground-truth score of a face image. Amazon Sagemaker[4] is exploited to collect scores from raters on scale 1 to 10, where 1 means not attractive, 10 is very attractive. This scale is used in order to give more options to raters to express their opinion about the beauty of a face than in previous works.

Since there is a chance that a rater did not pay enough attention to the tasks or accidentally made a mistake in scoring, the label control or data cleaning is necessary. Firstly, there are several dozens of faces that are duplicated in the dataset. Some of them are the same images, while other faces are pictured with different face poses or expressions. If the score difference in > 5 of these pairs given by a rater is > 3, all scores provided by this rater are removed from the dataset. Secondly, if an image has less than 10 scores or a volunteer rated less than 30 images, the image and the volunteer, respectively, are removed from the dataset.

Besides, two different objectives are assigned to raters. The first task is to evaluate the generic beauty of a face. That means, no matter how old the person is in an image, whatever ethnicity and gender he or she is—how good looking or attractive this person is according to the rater's opinion. The second task for a rater is to evaluate a face from the rater's interest in this person as a romantic partner. While the result of the first task might be applied to the universal concept of facial beauty, one of the potential applications of the results of the second task is social media recommendation including online dating. The final dataset with both types of scores before and after the data cleaning described above, as well as their average results is available by the link presented in Sect. 1.

[4] https://aws.amazon.com/sagemaker/.

3.3 MEBeauty composition and analysis

In this section, the detailed structure of the proposed dataset and its statistical analyses are presented. First of all, gender, age and ethnic composition are described, then distribution and standard deviation of beauty scores are demonstrated. The following must be mentioned: all conclusions, visual representations and analysis are based on the data in MEBeauty and are not specific to any gender, ethnic or age in the world as a whole.

3.3.1 Dataset composition

The total number of images in the dataset is 2550, including 1300 female and 1250 male faces. Every gender group is divided into 6 ethnic categories: Black, Asian, Caucasian, Hispanic, Indian, and Mideastern, that in turn include faces of various ages. Figure 4 demonstrates the proportion of a particular ethnic and age group among female and male images in the dataset. Caucasian faces have the highest percentage among all ethnicities, 35% and 40% of male (a) and female (c) images, respectively. Equal shares are occupied by Asian, Indian, Black, Hispanic, and Mideastern ethnic groups in the male part of the dataset (a), while 15% of female faces are Asian. Regarding the age composition, 47% males (b) and 42% (d) females belong to the age category of 20-35 years old, while about 25% of them are under the age of 14-20 and 35-55. The lowest percentage of both female and male groups occupied by faces under 14 or older than 55 years old.

3.3.2 Beauty score distribution

Figure 5 demonstrates the distribution of beauty scores among all female and male face images in the dataset (a) as well as inside every ethnic group. On the graphs, the red

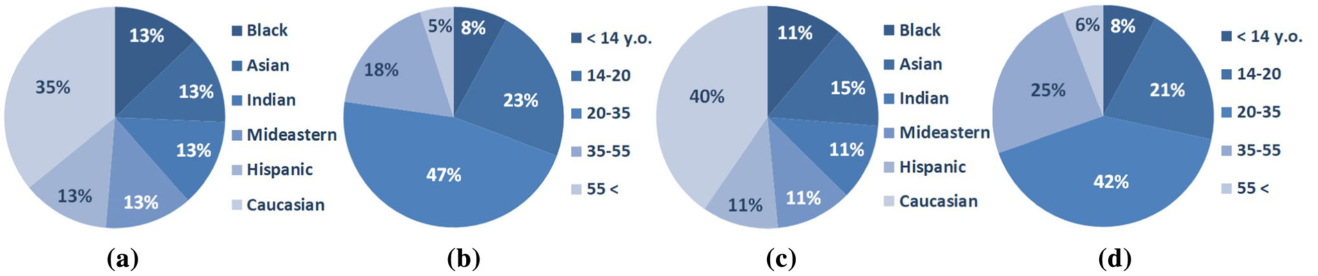
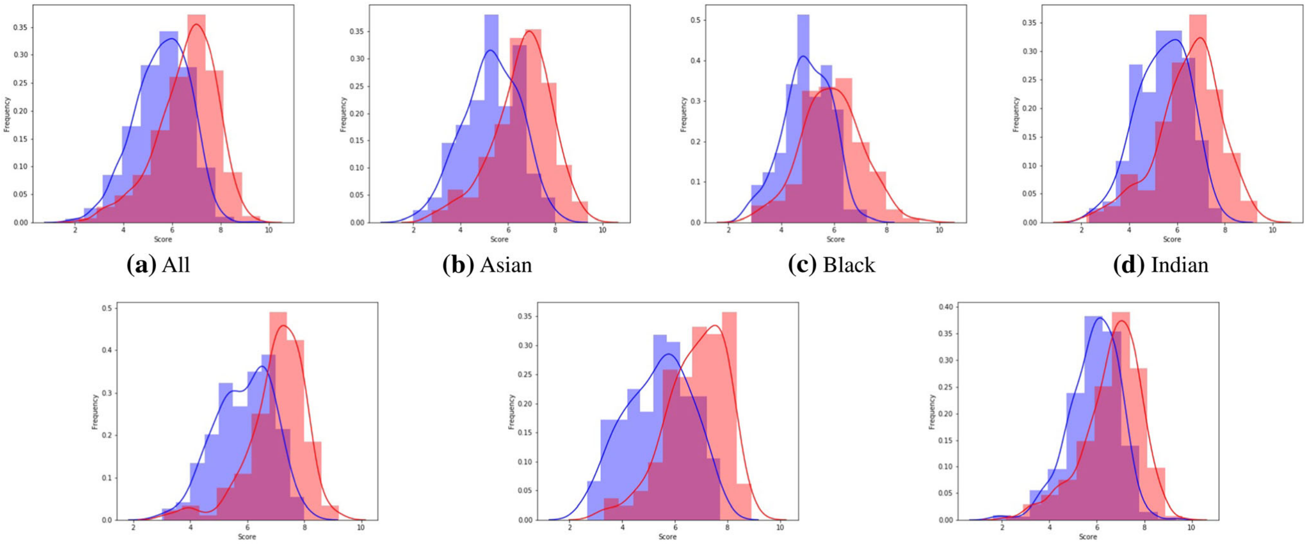


Fig. 4 The MEBeauty dataset ethnic and age composition. 40% female faces have Caucasian ethnicity (**a**) and 42% of them are aged from 20 to 35 (**b**), while Black, Asian, Hispanic, Indian, Mideastern male faces have equal parts (**c**) and 31% of them are under 20 years old



(a) All **(b)** Asian **(c)** Black **(d)** Indian

(e) Hispanic **(f)** Mideastern **(g)** Caucasian

Fig. 5 Score Distribution in MEBeauty. The following must be mentioned: all conclusions, visual representations and analysis are based on the data in MEBeauty and are not specific to any gender, ethnic or age in the world as a whole. Female faces (red curve) have higher beauty scores than male faces (blue curve) in all ethnic groups presented in the dataset. This difference is the most pronounced in the Mideastern group **f**. Scores of black female faces **c** are the most distributed, while Hispanic **e** and Caucasian **g** are the least spread out

curve is assigned with scores of female faces, while the blue curve—for male faces. Since the number of images belong to the whole datasets and its ethnic groups are different for female and male images, relative frequency is used for better clarity. Figure (a) shows that the attractiveness of female images trends to be higher than male images in the dataset. Mideastern (f) and Hispanic (e) female faces have the highest beauty scores among other faces presented in the ME-beauty dataset. The most distributed scores are assigned to Black female faces (b).

3.3.3 Standard deviation of beauty scores

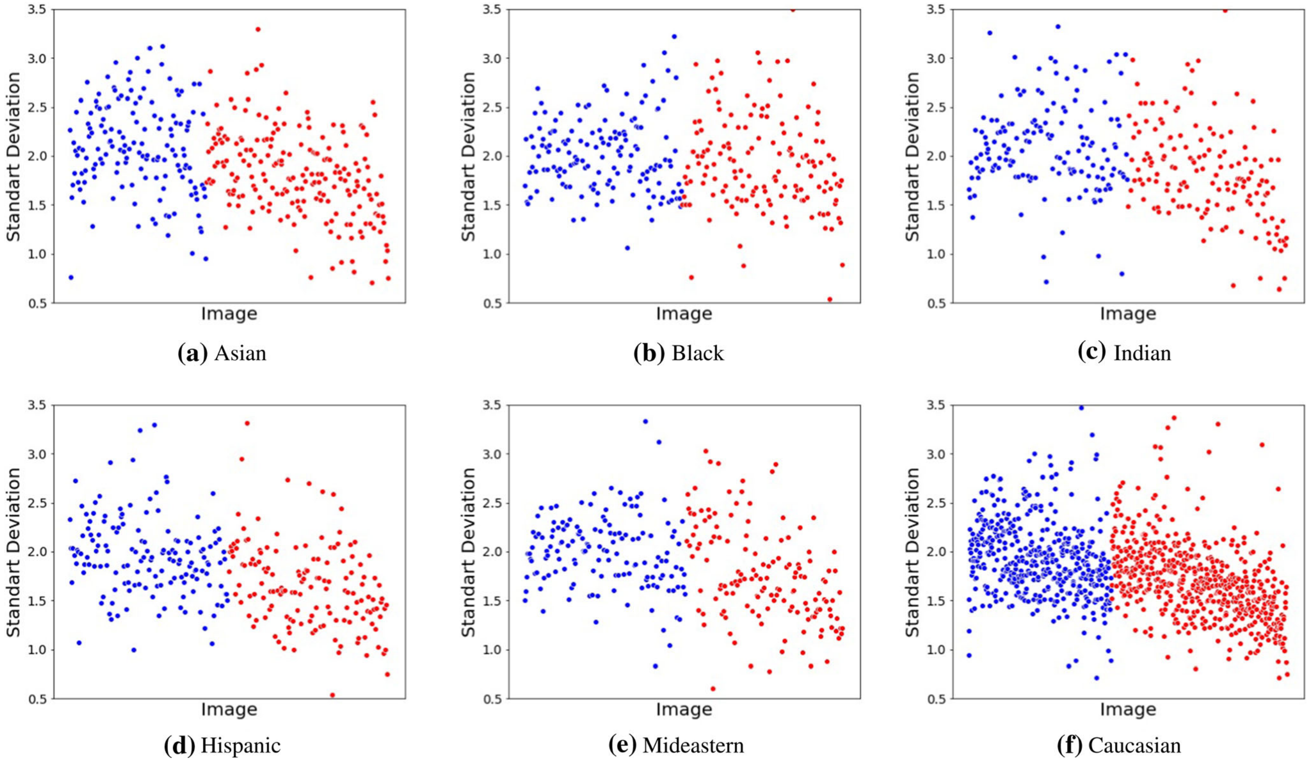
The standard deviation of all scores given to an image to its ground-truth is calculated. The point in Fig. 6 means the standard deviation of the scores of a particular image gathered from different raters to the ground-truth of this image. The red-color point is the standard deviation of the scores of a female face image, while the blue one is associated to a male image. Since the number of Caucasian face images is the largest among other ethnic groups, its chart (f) has the highest density. The scores of female images have lower standard deviation than scores of male images in all ethnic groups, except the Black group (b). Caucasian and Hispanic face images have less variability in scores than images in other groups. The scores of Asian (a) and Indian (c) male images are the most spread out in the dataset.

4 Methodology

Figure 7 illustrates the pipeline of the facial beauty prediction framework proposed in this study. Face cropping, alignment and resizing are first performed on a full color image. Then, pre-processed images are passed through a deep CNN to compute beauty patterns. Finally, a deep regression network with a robust loss function performs on these patterns to predict beauty scores. The details of these steps are presented in the following sections. In addition, Table 3 summarizes all the variables, sets and notations used in the system modeling.

4.1 Image preprocessing

In order to enhance deep learning methods applied in this work, the image preprocessing is exploited. Multi-task Cascaded Convolutional Networks [52] are exploited for face detection and alignment. Image augmentation using



(a) Asian **(b)** Black **(c)** Indian **(d)** Hispanic **(e)** Mideastern **(f)** Caucasian

Fig. 6 Standard deviation of beauty scores. Red points are the std of the scores assigned to female images, blue points—male images. The scores of Caucasian **f** and Hispanic **d** faces have the lowest std, while the scores given by different raters to Asian male **a** and Black **b** faces have the highest variability

rotation, flipping and shifting is also conducted on training data.

4.2 Deep beauty pattern learning

Once the face is cropped and aligned, it is then passed through a CNN to learn beauty patterns required for further training and prediction. Well-known CNN architectures are exploited for the task. First of all, VGG16 that is formed of 13 convolutional layers, and 3 fully connected layers [33] is used. Another CNN that is applied to learn beauty patterns is Xception [9]. This network contains depth-wise separable convolution layers that consist in performing convolution separately on each channel of the input. DenseNet [21], where each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers, is also exploited. Xception and DenseNet pre-trained on ImageNet [11] are used to learn beauty patterns. However, this is not optimal as face datasets are available and the facial beauty prediction task can benefit from pre-training with another face related task [3]. Thus, knowledge transfer from face recognition for FBP is performed. VGG16 pre-trained on the VGG Face 2 dataset [5] for face recognition is studied for the task.

4.3 Beauty prediction

The CNNs exploited in this work are originally created to provide discrete classification on the ImageNet dataset and have an output softmax layer, where the number of channels equals the number of classes in the dataset (1000). However, in most FBP studies and this work is no exception, facial beauty prediction is considered as a regression task. In other words, beauty score is a continuous value rather than a set of discrete classes. Thus, the last softmax layer is replaced by a fully connected layer with linear activation function. Since the training data have a relatively small number of images compared to the high dimension of the features, a dropout layer is added before the last layer in order to overcome overfitting. The regression for predicting beauty scores is learned by optimizing a robust loss function. In contrast to previous deep regression FBP frameworks, the proposed method is not limited to the classic L2 loss, different robust functions, such as Huber, and L1 loss, are used to train CNNs with a various number of unfreezing layers. The training process adopts the Adam optimizer in order to minimize the loss function over the training data.

Let x_i denotes a face image in the dataset $\mathcal{D}$, where n is the total number of samples. Then, y_i is the ground-truth

Table 3 Notations

Variable	Definition
$\mathcal{D}$	FBP dataset
x_i	i face image (sample) in the dataset $\mathcal{D}$
y_i	Ground-truth score of x_i image
$y_i{}^p$	Predicted score associated with x_i image
$\bar{y}$	The mean of the ground-truth scores
$\bar{y^p}$	The mean of the predicted scores
n	Number of face images in the dataset
$L1$	L1 loss
$L2$	L2 loss
δ	Hyperparameter to define the range for L1 and L2
L_δ	Huber loss
PC	Pearson correlation coefficient
MAE	Mean absolute error
MSE	Root mean squared error
$f(\bullet)$	Learning algorithm

The variables, sets, notations used in the approach modeling and its evaluation

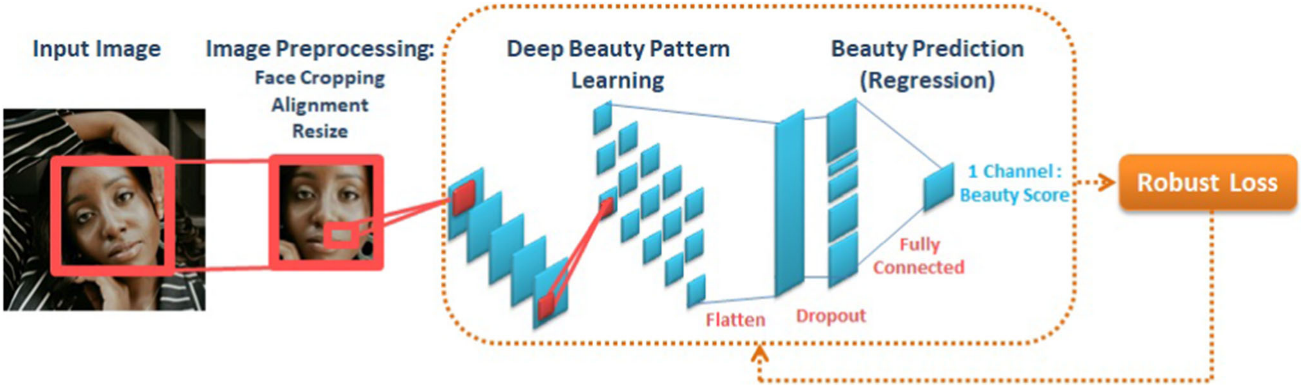


Fig. 7 The flowchart of the proposed approach. Each face is cropped aligned and resized. Then, pre-processed images are passed through a deep CNN to compute beauty patterns. Finally, a deep regression network with a robust loss function is trained on these patterns in order to predict beauty scores

score of x_i face image and $y_i{}^p$ is the predicted score associated with x_i face image.

4.4 Robust loss

In this section, the widely-used L2 loss as well as two robust loss functions used in the CNN training are briefly described.

4.4.1 L2 loss

The L2 loss measures the average of squared distances between ground-truth and predicted scores. For n predictions, the L2 loss is given by:

$$L2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i{}^p)^2 \tag{1}$$

where y_i is the ground-truth beauty score associated with the i image, and y_i^p is the predicted score that can also be explained as the output of the deep regression network.

4.4.2 L1 loss

The L1 loss function is the average of absolute residual error over the training set. For n training images, their predicted scores y_i^p and corresponding targets y_i, the L1 loss is defined by:

$$L1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i{}^p| \tag{2}$$

4.4.3 Huber loss

Huber loss is the combination of two functions presented above. It is L2 loss when the error is small, else L1 loss. Here δ is the hyperparameter to define the range for L1 and L2. The Huber loss is given by:

$$L_\delta = \begin{cases} \frac{1}{2}(y_i - y_i^p)^2 & \text{for} \quad |y_i - y_i^p| \leqslant \delta \\ \delta|y_i - y_i^p| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \tag{3}$$

where n, y_i^p, y_i mean number of training images, predicted score and its ground-truth. Huber loss approaches L2 when $\delta \sim 0$ and L2 when $\delta \sim \infty$.

5 Experimental results

First of all, the settings and the performance metrics applied in the work are described. Secondly, CNNs pre-trained for face recognition and object classification are performed on the proposed dataset with different numbers of unfreezing layers. Thirdly, regression loss functions defined in the previous section are exploited to re-train the CNNs. Fourth, the results are compared to the other FBP frameworks on the proposed dataset and widely-used SCUT-FBP 5500. Finally, the performance on separated ethnic and gender groups of MEBeauty is compared.

5.1 Experimental settings

The MEBeauty dataset is randomly divided into 80% for training and 20% for testing. In the training phase, 90% of the data is used for learning and 10% is used for validation. The Adam optimizer with its default configuration is experimentally chosen for the training of all CNNs. The training rate of 0.001 is also found as the most sufficient for the task. The batch size is fixed to 32 samples. The rate of 0.5 is chosen for the dropout layer shown in Fig. 7. All experiments are run on Amazon Elastic Compute Cloud.[5] A deep learning virtual machine with Ubuntu 18.04 is selected for the experiments.

5.2 Evaluation metrics

The facial beauty prediction methods presented in this work are evaluated in terms of Pearson Correlation (PC), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

$$\mathrm{PC} = \frac{\sum_{i=1}^{n}(f(x_i) - f(\bar{x}))(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(f(x_i) - f(\bar{x}))^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(f(x_i) - y_i)^2} \tag{6}$$

where n denotes the number of image samples in the set, x_i—the face image i, $f(\bullet)$—the learning algorithm, y_i—the ground-truth beauty score of an image x_i, $\bar{y}$ denotes the mean of the ground-truth scores, $f(\bar{x})$—the mean of the predicted scores.

PC is a measure of the linear correlation between $f(x_i)$ and y_i. It has a value between 1 and -1, where 1 means total positive linear correlation, 0—no linear correlation, and -1 means total negative linear correlation. MAE and RMSE measure the quality of a machine learning model, the values close to zero mean better performance.

[5] https://aws.amazon.com/ec2/.

5.3 Datasets

In order to compare the effectiveness of FBP frameworks on face images with different properties and restrictions, these methods are performed not only on the proposed multi-ethnic FBP dataset in-the-wild, but also on most known facial beauty prediction dataset SCUT-FBP 5500 that includes 5500 Asian, Caucasian female and male frontal faces with neutral expression.

5.4 CNN performance on MEBeauty

Different CNN architectures are exploited as a backbone of the proposed approach. Moreover, layer-wise transfer learning is applied to save time and computational power for training. Since the number of images in the dataset is still relatively low, the use of transfer learning potentially enhances the performance of the CNNs. Moreover, facial beauty prediction can benefit from pre-training with other face related task. Thus, knowledge transfer from the face recognition task is performed. Table 4 demonstrates the results of the approach with: VGG16 pre-trained on VGGFace2 dataset, Xception and DenseNet pre-trained on ImageNet.

Since the VGG16 network has the lowest number of layers among all the presented CNNs, it achieves the worst results with many unfreezing layers. On the other hand, VGG16 in this experiment is pre-trained for the face recognition task, thus, it presents good performances with a small number of unfreezing layers. Xception and DenseNet are both pre-trained on ImageNet, and demonstrate similar performances. However, due to its higher number of layers, DenseNet outperforms the Xception network. Both networks achieved the best prediction accuracy with 50% unfreezing layers.

The graph of the loss function vs. the number of epochs during the training of DenseNet with 50% of unfreezing layers and L2 loss is shown in Fig. 8. The blue curve is assigned to the loss on the training dataset, while the red curve is the loss on the validation data. At the beginning of the training process, the loss is high and unstable on both training and validation data. The optimal result from time- and power-consuming point of view is achieved with 30 training epochs. The continuation of the training process results in overfitting on the training data and less accuracy and stability on the validation data.

5.5 Loss functions evaluation

Since the configuration with 50% of unfreezing layers demonstrates the best performance on two CNN architectures, it is selected to evaluate various loss functions for facial beauty prediction on the proposed dataset. The comparison of L2 loss that is the classic for deep regression networks in age estimation and facial beauty prediction; L2 loss that is also called MAE loss in the literature; and robust Huber loss; is presented in Table 5. The L1 and Huber loss functions improve the performance of VGG16 on the dataset, while these loss functions demonstrate lower accuracy on DenseNet. The L2 loss function shows slight improvements in terms of PC, but achieves worse results in terms of MAE on the Xception network.

5.6 Comparison to the state-of-the-art

Table 6 shows the performance of FBP methods conducted under different conditions: on face images in-the-wild (MEB eauty) and images captured in a constrained environment (SCUT-FBP 5500). The images are rated on the scale of 1-10 in MEBeauty, while the scale of 1-5 is used in the SCUT-FBP 5500 dataset. The results of the proposed approach on both FBP datasets are also compared to the other methods.

First of all, a shallow predictor trained on Gabor features [7] is exploited and it achieves the lowest results on both types of face images. The CNN specially created for the facial beauty prediction task [42] is also evaluated on MEBeauty and SCUT-fbp5500. The network outperforms the traditional machine learning methods for FBP, but shows poor performances on images in-the-wild. Transfer

Table 4 Performance of CNNs with layer-wise transfer learning

Unfreezing %	VGG16			DenseNet			Xception		
	PC	MAE	RMSE	PC	MAE	RMSE	PC	MAE	RMSE
0%	0.641	0.875	1.078	0.663	0.830	1.026	0.571	0.893	1.126
25%	**0.709**	**0.789**	**0.967**	0.731	0.704	0.909	0.721	0.729	0.936
50%	0.671	0.785	0.978	**0.748**	**0.674**	**0.877**	**0.730**	**0.723**	**0.933**
75%	0.558	0.927	1.102	0.718	0.780	1.017	0.702	0.810	1.037
100%	0.263	1.073	1.323	0.484	0.985	1.209	0.441	1.012	1.236

VGG16 is pre-trained for face recognition, while DenseNet and Xception are pre-trained for general object classification. Bold indicates the best results

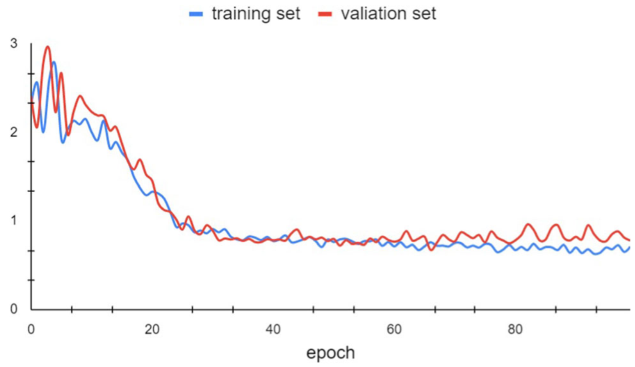


Fig. 8 The graph of the loss function vs. the number of epochs. The blue curve is associated with the loss on the training data, while the red curve shows the loss on the validation data. The DenseNet-backed approach with 50% unfreezing layers and L2 loss is chosen to demonstrate the training phase with 100 epochs on the MEbeauty dataset

learning based frameworks are also performed on both datasets. The ResNeXt50 and AlexNet networks pre-trained on ImageNet presented in [26] show good performances on face images with restrictions, while achieving lower accuracy on images in-the-wild. In contrast, another transfer learning based approach, namely TransFBP [47], improves the results on ME-Beauty better than SCUT-FBP 5500. This two-stage method includes the feature extraction from VGG-16 pre-trained for the face verification and Bayesian ridge regression. The FBP framework proposed in this work outperforms the state-of-the-art on the ME-Beauty dataset with face images in-the-wild, while also achieving very competitive performance on SCUT-FBP 5500.

5.7 Beauty prediction in terms of ethnicity and gender

The CNN with the configuration that demonstrates the best performances in Sects. 5.4 and 5.5 is evaluated on separated ethnic and gender groups of MEBeauty. Table 7 shows that the accuracy in terms of PC is higher on female images than male images, while the performance in terms of MAE and RMSE is better on male images. This difference is the most obvious for Asian and Black ethnic groups. It could be attributed to the lower distribution of the scores of male faces in the dataset. The best performances are achieved on Mideastern faces, Caucasian and Asian female faces, while the scores of Indian male faces are the least predictable in the dataset. Some examples of the prediction results of the proposed method are presented in Fig. 9.

6 Conclusion and future works

The facial beauty prediction issue on images in-the-wild is addressed. A multi-ethnic dataset, namely MEBeauty, with a rich diversity in age, gender, face pose and expression is collected. The performance of different CNNs with layer-wise transfer learning is done to study the optimal architecture and number of layers to fine-tune on FBP task.

Table 5 Loss function evaluation

CNN	L2			L1			Huber		
	PC	MAE	RMSE	PC	MAE	RMSE	PC	MAE	RMSE
VGG16	0.671	0.785	0.978	**0.694**	**0.728**	**0.918**	0.687	0.742	0.945
DenseNet	**0.748**	**0.674**	**0.877**	0.734	0.730	0.938	0.724	0.719	0.928
Xception	0.730	**0.723**	**0.933**	**0.736**	0.756	0.969	0.718	0.774	0.993

The configuration with 50% of unfreezing layers is selected to evaluate the loss functions

Table 6 FBP state-of-the-art performance on publicly available facial beauty datasets

Method	MEBeauty			SCUT-FBP 5500		
	PC	MAE	RMSE	PC	MAE	RMSE
Gabor + shallow predictor [7]	0.371	1.020	1.208	0.669	0.389	0.506
CNN based [42]	0.467	0.978	1.186	0.818	0.315	0.391
AlexNet [26]	0.493	0.933	1.167	0.843	0.272	0.359
ResNeXt50 [26]	0.523	0.918	1.140	0.877	0.245	0.328
TransFBP [47]	0.583	0.836	1.03	0.857	0.259	0.339
The proposed framework	**0.748**	**0.674**	**0.877**	**0.885**	**0.243**	**0.326**

SCUT-FBP 5500 contains Asian and Caucasian faces captured in constrained environment, while ME-Beauty is a multi-ethnic dataset with face images in-the-wild

Table 7 Evaluation on MEBeauty in terms of ethnicity and gender

	Black		Asian		Caucasian		Hispanic		Indian		Mideastern		All	
	f	m	f	m	f	m	f	m	f	m	f	m	f	m
PC	0.642	0.577	**0.753**	0.622	0.717	0.648	0.449	0.421	0.702	0.404	**0.749**	**0.693**	0.724	0.629
MAE	0.758	**0.662**	0.781	0.718	0.735	0.703	0.829	0.780	0.837	0.863	0.769	0.802	0.710	0.744
RMSE	0.818	0.594	0.912	0.807	0.831	0.779	1.004	0.960	1.122	1.018	1.012	0.940	0.923	0.821

The proposed method demonstrates better performances on female images than male images. The best results are shown on Mideastern faces, while the lowest accuracy is achieved on Indian male images

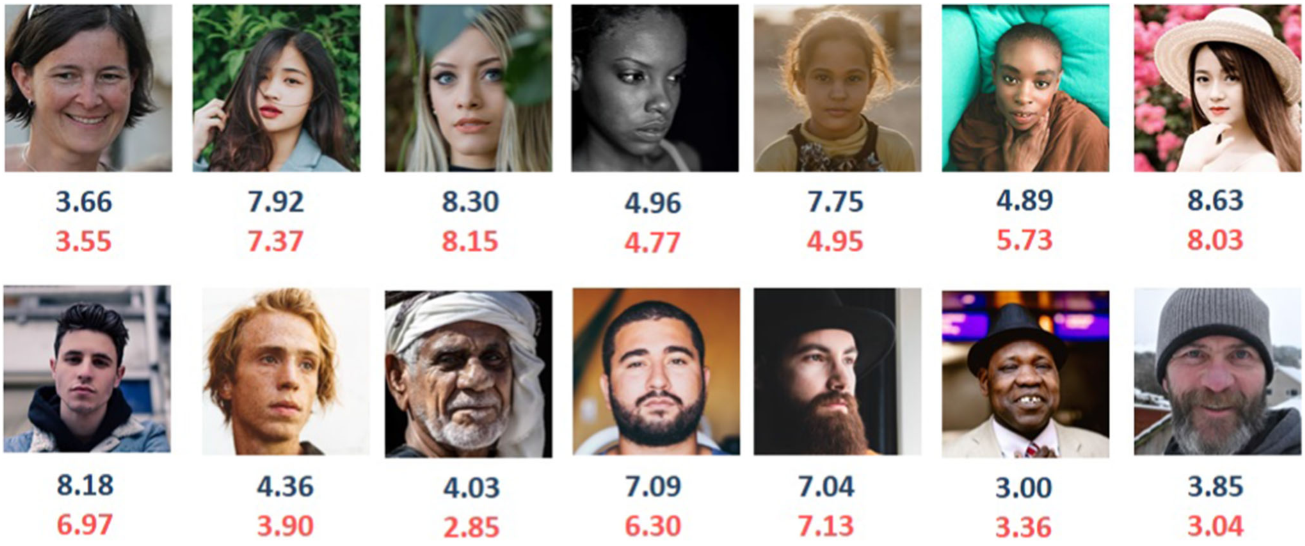


Fig. 9 Accurate and inaccurate prediction results on female and male images. The blue score represents the ground-truth of an image, while the red score is its predicted result

Knowledge transfer evaluation from face recognition task across FBP demonstrates that facial beauty prediction can benefit more from general tasks when an applied CNN architecture is deeper and the number of re-trained layers is relatively high. The use of robust loss functions to learn deep regression networks for beauty prediction is studied. The performance evaluation of several FBP frameworks on the proposed MEBeauty dataset and widely-used SCUT-FBP 5500 is conducted in order to compare their effectiveness on images captured in-the-wild and in an unconstrained environment. The evaluation of the proposed method on separated ethnic and gender groups demonstrates that the FBP achieved the best on Mideastern faces, Asian and Caucasian female faces, while the lowest results are shown on Indian male images.

Since the dataset contains scores given by a relatively high number of raters, the personalized aspect of facial beauty prediction will be studied in the future works. The use of generative adversarial networks (GAN) for beauty enhancement in multi-ethnic scenario and according personal preferences is another direction of the work. The dataset extension from the quantity and consistency point of view is also planned to be done.

Declarations

Conflicts of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Aarabi P, Hughes D, Mohajer K, Emami M (2001) The automatic measurement of facial beauty. In: Proceedings of the 2001 IEEE international conference on systems, man and cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), vol. 4, pp 2644–2647. IEEE
2. Agthe M, Strobel M, Spörrle M, Pfundmair M, Maner JK (2016) On the borders of harmful and helpful beauty biases: the biasing effects of physical attractiveness depend on sex and ethnicity. Evol Psychol 14(2):1474704916653968

3. Antipov G, Baccouche M, Berrani SA, Dugelay JL (2017) Effective training of convolutional neural networks for face-based gender and age prediction. Pattern Recogn 72:15–26
4. Cao K, Choi Kn, Jung H, Duan L (2020) Deep learning for facial beauty prediction. Information 11(8):391
5. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: a dataset for recognising faces across pose and age. In: Proceedings of the 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018), pp 67–74. IEEE
6. Chang F, Chou CH (2009) A bi-prototype theory of facial attractiveness. Neural Comput 21(3):890–910
7. Chen F, Zhang D (2010) A benchmark for geometric facial beauty study. International conference on medical biometrics. Springer, New York, pp 21–32
8. Chen Y, Mao H, Jin L (2010) A novel method for evaluating facial attractiveness. In: Proceedings of the 2010 international conference on audio, language and image processing, pp 1382–1386. IEEE
9. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
10. De Vries H, Yosinski J (2015) Can deep learning help you find the perfect match? Deep Learning Workshop at ICML 2015
11. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. IEEE
12. Dornaika F, Elorza A, Wang K, Arganda-Carreras I (2019) Nonlinear, flexible, semisupervised learning scheme for face beauty scoring. J Electron Imaging 28(4):043013
13. Dornaika F, Elorza A, Wang K, Arganda-Carreras I (2020) Image-based face beauty analysis via graph-based semi-supervised learning. Multimedia Tools Appl 79(3):3005–3030
14. Dornaika F, Moujahid A, Wang K, Feng X (2020) Efficient deep discriminant embedding: application to face beauty prediction and classification. Eng Appl Artif Intell 95:103831
15. Dornaika F, Wang K, Arganda-Carreras I, Elorza A, Moujahid A (2020) Toward graph-based semi-supervised face beauty prediction. Expert Syst Appl 142:112990
16. Eisenthal Y, Dror G, Ruppin E (2006) Facial attractiveness: beauty and the machine. Neural Comput 18(1):119–142
17. Fan J, Chau K, Wan X, Zhai L, Lau E (2012) Prediction of facial attractiveness from facial proportions. Pattern Recogn 45(6):2326–2334
18. Gan J, Li L, Zhai Y, Liu Y (2014) Deep self-taught learning for facial beauty prediction. Neurocomputing 144:295–303
19. Gray D, Yu K, Xu W, Gong Y (2010) Predicting facial beauty without landmarks. European conference on computer vision. Springer, New York, pp 434–447
20. Gunes H, Piccardi M, Jan T (2004) Comparative beauty classification for pre-surgery planning. In: Proceedings of the 2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583), vol. 3, pp 2168–2174. IEEE
21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
22. Kagian A, Dror G, Leyvand T, Cohen-Or D, Ruppin E (2007) A humanlike predictor of facial attractiveness. In: Advances in neural information processing systems, pp 649–656
23. Lebedeva I, Guo Y, Ying F (2021) Deep facial features for personalized attractiveness prediction. In: Thirteenth international conference on digital image processing (ICDIP 2021), vol. 11878, p 118780A. International society for optics and photonics
24. Lebedeva I, Guo Y, Ying F (2021) Transfer learning adaptive facial attractiveness assessment. In: Journal of Physics: Conference Series, vol. 1922, p 012004. IOP Publishing
25. Li J, Xiong C, Liu L, Shu X, Yan S (2015) Deep face beautification. In: Proceedings of the 23rd ACM international conference on Multimedia, pp 793–794. ACM
26. Liang L, Lin L, Jin L, Xie D, Li M (2018) Scut-fbp5500: a diverse benchmark dataset for multi-paradigm facial beauty prediction. In: Proceedings of the 2018 24th international conference on pattern recognition (ICPR), pp 1598–1603. IEEE
27. Liu L, Xing J, Liu S, Xu H, Zhou X, Yan S (2014) Wow! you are so beautiful today! ACM Trans Multimedia Comput Commun Appl (TOMM) 11(1s):20
28. Mao H, Jin L, Du M (2009) Automatic classification of Chinese female facial beauty using support vector machine. In: Proceedings of the 2009 IEEE international conference on systems, man and cybernetics, pp 4842–4846. IEEE
29. Murray N, Marchesotti L, Perronnin F (2012) Ava: a large-scale database for aesthetic visual analysis. In: Proceedings of the 2012 IEEE conference on computer vision and pattern recognition (CVPR), pp 2408–2415. IEEE
30. Nguyen TV, Liu S, Ni B, Tan J, Rui Y, Yan S (2012) Sense beauty via face, dressing, and/or voice. In: Proceedings of the 20th ACM international conference on multimedia, pp 239–248
31. Redi M, Rasiwasia N, Aggarwal G, Jaimes A (2015) The beauty of capturing faces: rating the quality of digital portraits. In: Proceedings of the 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol. 1, pp 1–8. IEEE
32. Rothe R, Timofte R, Van Gool L (2016) Some like it hot-visual guidance for preference prediction. In: Proceedings CVPR 2016, pp 1–9
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. ICLR
34. Sutić D, Brešković I, Huić R, Jukić I (2010) Automatic evaluation of facial attractiveness. In: MIPRO, 2010 proceedings of the 33rd international convention, pp 1339–1342. IEEE
35. Tong S, Liang X, Kumada T, Iwaki S (2020) Putative ratios of facial attractiveness in a deep neural network. Vision Res 178:86–99
36. Vahdati E, Suen CY (2019) Female facial beauty analysis using transfer learning and stacking ensemble model. International conference on image analysis and recognition. Springer, New York, pp 255–268
37. Vahdati E, Suen CY (2020) Facial beauty prediction using transfer and multi-task learning techniques. International conference on pattern recognition and artificial intelligence. Springer, New York, pp 441–452
38. Wei W, Ho ES, McCay KD, Damaševičius R, Maskeliūnas R, Esposito A (2021) Assessing facial symmetry and attractiveness using augmented reality. Pattern Anal Appl 1–17
39. Weng N, Wang J, Li A, Wang Y (2021) Two-stream temporal convolutional network for dynamic facial attractiveness prediction. In: Proceedings of the 2020 25th international conference on pattern recognition (ICPR), pp 10026–10033. IEEE
40. Whitehill J, Movellan JR (2008) Personalized facial attractiveness prediction. In: Proceedings of the 8th IEEE international conference on automatic face and gesture recognition, 2008. FG'08, pp 1–7. IEEE
41. Xiao Q, Wu Y, Wang D, Yang YL, Jin X (2021) Beauty 3D facenet: deep geometry and texture fusion for 3D facial attractiveness prediction. Comput Graph 98:11–18
42. Xie D, Liang L, Jin L, Xu J, Li M (2015) Scut-fbp: a benchmark dataset for facial beauty perception. In: Proceedings of the 2015 IEEE international conference on systems, man, and cybernetics (SMC), pp 1821–1826. IEEE

43. Xu J (2021) Mt-resnet: a multi-task deep network for facial attractiveness prediction. In: Proceedings of the 2021 2nd international conference on computing and data science (CDS), pp 44–48. IEEE
44. Xu J, Jin L, Liang L, Feng Z, Xie D (2015) A new humanlike facial attractiveness predictor with cascaded fine-tuning deep learning model. arXiv preprint arXiv:1511.02465
45. Xu J, Jin L, Liang L, Feng Z, Xie D, Mao H (2017) Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn). In: Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1657–1661. IEEE
46. Xu L, Fan H, Xiang J (2019) Hierarchical multi-task network for race, gender and facial attractiveness recognition. In: Proceedings of the 2019 IEEE international conference on image processing (ICIP), pp 3861–3865. IEEE
47. Xu L, Xiang J, Yuan X (2018) Transferring rich deep features for facial beauty prediction. arXiv preprint arXiv:1803.07253
48. Yan H (2014) Cost-sensitive ordinal regression for fully automatic facial beauty assessment. Neurocomputing 129:334–342
49. Zhai Y, Huang Y, Xu Y, Gan J, Cao H, Deng W, Labati RD, Piuri V, Scotti F (2020) Asian female facial beauty prediction using deep neural networks via transfer learning and multi-channel feature fusion. IEEE Access 8:56892–56907
50. Zhai Y, Huang Y, Xu Y, Zeng J, Yu F, Gan J (2016) Benchmark of a large scale database for facial beauty prediction. In: Proceedings of the 2016 international conference on intelligent information processing, pp 1–5
51. Zhang D, Chen F, Xu Y et al (2016) Computer models for facial beauty analysis. Springer, New York
52. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503