Facial Beauty Prediction Based on Vision Transformer

Djamel Eddine Boukhari[1,2,*], Ali Chemsa[1], and Riadh Ajgou[1]
[1] LGEERE Laboratory Department of Electrical Engineering, University of El Oued, 39000 El-Oued, Algeria
[2] Scientific and Technical Research Centre for Arid Areas (CRSTRA),07000 Biskra, Algeria
Email: boukhari-djameleddine@univ-eloued.dz (D.E.B.), d-technologie@univ-eloued.dz (A.C.), riadh-ajgou@univ-eloued.dz (R.A.)

***Abstract*—Facial beauty analysis is a crucial subject in human culture among researchers through different applications. Recent studies used multidisciplinary approaches to examine the relation between facial traits, age, emotions, and other factors. However, facial beauty prediction is a significant visual recognition challenge for the evaluation of facial attractiveness for human perception, which requires a considerable effort due to the field's novelty and lack of resources with a small database for facial beauty prediction. In this vein, a deep learning method has recently demonstrated remarkable qualities in facial beauty prediction. Additionally, vision Transformers have recently been introduced as novel Deep Learning approaches and have presented a strong performance in a number of applications. The key issue is that vision transformer performs significantly worse than ResNet when trained on a small ImageNet database. In this paper, we propose to tackle the difficulties of facial beauty prediction, using vision transformers as opposed to feature extraction based on Convolutional Neural Networks commonly used in traditional methods. Moreover, we define and optimize a set of hyper-parameters according to the SCUT-FBP5500 benchmark dataset the model obtains 0.9534 Pearson Coefficient. Experimental results indicated that using this proposed network leads to better predicting of facial beauty closer to human evaluation than conventional technology that provides facial beauty assessment.**

***Index Terms*—Facial beauty prediction, vision transformer, deep learning, convolutional neural networks, performance evaluation**

I. Introduction

The human face has a prominent role in our social interactions and the pursuit of beauty since facial beauty is a feature of human nature. In recent years, there has been a significant increase in the demand for aesthetic surgery, which underscores the importance of a nuanced understanding of beauty in medical settings [1]. Nevertheless, the investigation of physical beauty in humans has a history of more than 4,000 years, which shows the persistent relevance of this topic [2].

The importance of physical beauty in the face has been studied for years by ancient civilizations, including the sculptors of Egypt, Greece, and Rome, which defined a series of ratio rules for the human body to create attractive works since it affects social decisions such as partner choices and hiring decisions [1]. The perception of facial attractiveness is considered a highly desirable physical trait, with philosophers, artists, and scientists all attempting to understand the secrets of beauty [3, 4].

Facial beauty prediction is a promising topic with great attention among researchers and users, particularly in the field of facial recognition and understanding [5]. Yet, beauty is a form of information in computer-based face analysis related to attractiveness perception. Additionally, several theories in psychology have been conducted on how people observe facial attractiveness. Furthermore, studying face attractiveness using computers is a relatively recent research study, with few resources and articles published on the subject. Several works have focused on analyzing the irregular features of face attractiveness [6, 7].

The analysis of facial attractiveness presents two main challenges. Firstly, the complexity of human perception and the wide variety of facial traits make it difficult to build robust and effective models for evaluating beauty. Secondly, many face reference databases are primarily configured for face recognition problems, and they are unsuitable for attractiveness prediction [8]. Therefore, most facial beauty studies focus on designing facial beauty descriptors [9, 10]. In addition, the perceived attractiveness of a face is influenced by its symmetry and, to a lesser extent, by sex characteristics [11, 12].

In recent years, most facial beauty prediction research studies have relied on deep learning methods [13, 14]. Furthermore, the development of deep learning architecture has been driven by the strength and adaptability of these algorithms, particularly convolutional neural networks (CNNs) [15, 16]. However, these algorithms provide a novel perspective on the facial beauty prediction problem, with promising results for several computer vision applications such as face recognition, object identification, semantic segmentation, image classification, biomedical analysis, captioning, and biometrics [17].

Vision Transformer, or ViT [18], is the new state-of-the-art method for image classification. ViT was posted to the archive in October 2020 and officially published in 2021 on all the public data sets [19]. ViT beats the best RestNet by a small margin, provided that ViT has been pre-trained on a sufficiently large data set. The goal of this paper is to provide an overview of deep learning

Manuscript received September 13, 2023; revised November 9, 2023; accepted November 13, 2023.
*Corresponding author

techniques, particularly vision transformers in the field of facial beauty prediction [20].

The contributions of this work are presented as follows:

- We propose ViT-FBP: vision transformer architecture for facial beauty prediction.
- Firstly, we tackle the difficulties of facial beauty prediction on a small dataset using a vision transformer as opposed to feature extraction based on CNN commonly used for facial beauty prediction methods.
- Our ViT-FBP model consistently outperforms all previously published approaches on different facial beauty prediction tasks.
- We present state-of-the-art results on SCUT-FBP5500 dataset for facial beauty prediction by using our face alignment system making our scripts and the method of preprocessed faces accessible to the public at (https://github.com/DjameleddineBoukhari/ViT-FBP)

The structure of this paper is as follows: Some related research issues on face attractiveness prediction are included in Section II. Section III explains the selection process for the used architectures. Section IV shows the experimental results and the assessment performance based on the SCUT-FBP5500 data set.

II. Related Works

Deep learning is a subset of machine learning. It proved an efficient tool compared to the traditional methods, such as geometric and textural features. In FBP, deep features from an input image are extracted using deep convolutional neural networks. However, the comparison between machine learning and deep learning is how each technique and data are used. A short brief on facial beauty prediction methods based on deep learning, supervised learning, or semi-supervised learning

A. Supervised Learning

Geometric prior GPNet: Peng *et al*. in 2023 [21] propose using a dual-branch structure and geometric regularization with a hybrid network named GPNet. The two branches that identify the model structure are a local CNN branch and a global Swin Transformer branch, both of which are multi-scale feature fusion modules. An ensemble DCNNs: Saeed *et al*. (2023) [22] propose an ensemble DCNN-based regression model with an architecture of two fine-tuned, well-known pre-trained CNNs, namely AlexNet and VGG16, plus one network built entirely from scratch. The CNN-ER: Bougourzi *et al*. in 2022 [23] proposed an ensemble CNN with two branches, ResneXt-50 and Inception-v3, for face beauty estimation. This ensemble is trained with four loss functions (dynamic ParamSmoothL1, dynamic Huber, dynamic Tukey, and MSE). Thus, most work uses transfer learning for facial beauty perdition. In addition, several pre-trained CNN models show their accuracy for the beauty evaluation task. CNNs usually use an end-to-end model to complete a classification or regression task due to their fully connected layers; however, we can extract deep features by removing the fully connected layers and keeping only the convolutional ones [24, 25]. A CNN typically consists of convolutional layers, pooling layers, and fully connected layers. Convolutional layers are the core building blocks of a CNN [26]. CNN – SCA: Cao *et al*. [27] used residual-in-residual (RIR) groups to build a deeper network, where a combined spatial-wise and channel-wise attention mechanism is introduced for better feature comprehension. Authors presented their face beauty database SCUT-FBP5500 [28], with two evaluation protocols (5-fold cross validation 80%–20 % and 60%–40% split). They tested three CNN architectures Alexnet [29], Resnet-18 [30] and ResneXt-50 [30]. Consequently the results show better feature comprehension. R3CNN: Lin *et al*. [31] proposed R3CNN architecture to integrate relative ranking into the regression to improve the performance of FBP that could be flexibly implemented by using existing CNNs as backbone network. This architecture provides better results than SCUT-FBP [11] and SCUT-FBP5500 [28] dataset.

FSCLDE: Dornaika *et al*. in 2020 [32] proposed efficient deep discriminant embedding , using a cascaded feature extraction and selection architecture that can turn noisy and weak descriptors into strong ones, the structure enables the transformation of any linear approach into a deep variation. The rate classification (%) of face beauty achieved by a 1-NN classifier over different embedding spaces on three face beauty datasets SCUT-FBP5500 SCUT-FBP and M2B with evaluation protocol 5-fold cross validation .

B. Semi-Supervised Learning

MSMFME : Dornaika 2023 [33] proposed a multi-view semi-supervised technique that fuses various graphs to create a unifying flexible manifold embedding model, which has been trained and tested using fivefold cross-validation, where tests are conducted on the SCUT FBP-5500 dataset. NFME: F. Dornaika *et al*. in 2020 [34] propose a graph-based semi-supervised facial beauty prediction. The proposed method, NFME, is based on texture and handles the scenario of real score propagation as they modify and kernelize an existing linear flexible manifold embedding technique. Performance of face beauty is achieved by NFME on three face beauty datasets, SCUT-FBP5500, SCUT-FBP, and M2B, with an evaluation protocol of 5-fold cross-validation.

III. Methods

In this section, we initially provide an outline of the proposed architecture. A typical transformer makes use of the attention mechanism in neural networks. The attention mechanism was first proposed for the language translation problem [35]. Our proposed ViT-FBP architecture uses the core network following the ViT standard, with 8 layers of transformer blocks for fundamental feature extraction. If we add two fully connected layers, performance could be improved. As a result, the network's capacity has increased [36, 37].

A. Overview of FBP Vision Transformer

We apply data augmentation to images and create a layer that splits the image into patches. Then, encode the patches into the vector that stores image patches. The core network is consistent with or adopts the original ViT to create multiple layers of the transformer block, a data normalization layer, and a multi-head attention layer to skip connections of encoded patches, a data normalization layer, a multi-layer perceptron, and skip connections. Then, the regression token is used with an MLP head to extract features, which are then calculated by two fully connected layers (FC). Two linear layers in the MLP have a GELU activation function, as shown in Fig. 1.

The input image is transformed using a set of parameters called d, such as rotation angle and crop coordinates. Saving it directly becomes memory-inefficient since d varies for every image in every cycle. To solve this issue, in order to encode d, we just use one argument, $d_0 = E(d)$, where E(·) is the encoder shown in Fig. 1.

The multi-head attention consists of several self-attention blocks in order to capture as many complicated interactions as possible between the various items in the sequence. Essentially, we repeatedly use the cycle through the attention process [38]. With d_{model}-dimensional keys, values, and queries, many attention functions should be used instead of just one, and the attention function is carried out in parallel with each of the projected versions of queries, keys, and values. Each matrix is multiplied by a different weight matrix to create the mapping. The attention mechanism has the ability to concentrate on any object on the input image compared to Convolutional Neural Networks (CNNs), which use variable-size convolution kernels to scan across the different levels of the architecture; besides it functions inside a single network layer. Moreover, Tokenization occurs at the pixel level, as indicated in Fig. 1 below, meaning that each pixel in the grid cares about each other pixel. In order to fix this, the input image is divided into equal-sized square blocks, or image patches. Then, for later use and retrieval, each image patch is unrolled into a one-dimensional sequence (n×1) and given a positional embedding to a table.

The dense layer inside a fully connected layer consists of 2048 nodes, and the second dense layer inside a fully connected layer block consists of 1024 nodes.

For regression models, PC improvements do not only attach to the model structural design but also to the loss functions used. During training batches, the loss function computes the total error and uses back propagation to change the weights. To deal with different domains, several loss functions have been developed, some of which are derivations of already existing loss functions. The imbalances in the dataset are also taken into consideration by these loss functions. In the case of regression model of FBP, the default and the most frequently option is MSE.

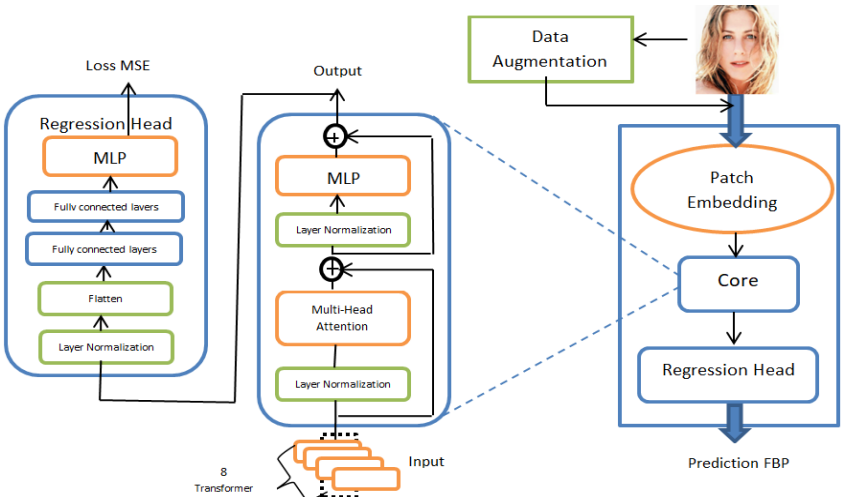


Fig. 1. Algorithm architecture, the core block consists transformer of an MHA, an MLP, skip connections, and layer normalizations.

IV. Experiments

The SCUT-FBP5500 dataset [28] is used for network training. Our network is trained for 600 iteration with batch size of b = 256. The Adam optimizer updates the parameters. The selected loss function was MSE.

A. The SCUT-FBP5500 Dataset

The standard SCUT-FBP5500 dataset [28] is introduced in this study and it comprises 5500 frontal face images at 350×350 resolutions with various attributes, including race (Asian/Caucasian), gender (female/male), and age (15 to 60 years old).

As shown in Fig. 2, Female Asian samples and the corresponding scores are from right to left: (1.56; 2.45; 3.51; 4.28), Male Asian samples and the corresponding scores are from right to left: (1.53; 2.46; 3.53; 4.23), Female Caucasian samples and the corresponding scores are from right to left: (1.93; 2.45; 3.66; 4.45) and Male Caucasian samples and the corresponding scores are from right to left: (1.53; 2.66; 3.45; 4.2), the ground truth rating for each face in the dataset is the average of all evaluations given on a scale from 1 to 5 by the 60 ratters. This enables the use of various computational models with various facial attractiveness prediction paradigms. The 2000 Asian females (AF), 2000 Asian men (AM),

750 Caucasian females (CF), and 750 Caucasian males (CM) are the four subsets of the SCUT-FBP5500 Dataset that may be separated according to race and gender [39, 40].

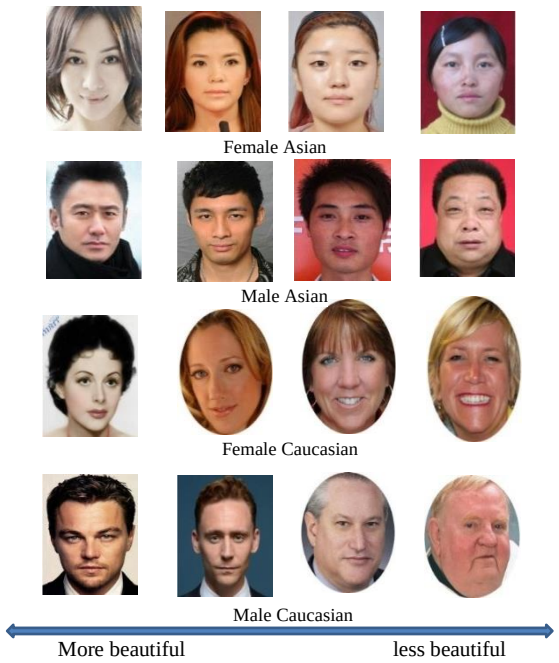


Fig. 2. Images of various facial features and beauty ratings from the SCUT-FBP5500 benchmark dataset.

B. Performance Evaluation

In this study, beauty prediction approach is assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Pearson Correlation (PC) [41, 42]. The following is a definition of the evaluation metrics:

Mean Absolute Error (MAE) is defined by:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

Root Mean Squared Error (RMSE) is defined by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2} \tag{2}$$

Pearson Correlation (PC) is defined by:

$$\mathrm{PC} = \frac{\sum_{i=1}^{N}(y_i - \bar{y_i})\left(\hat{y}_i - \bar{\hat{y}_i}\right)}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y_i})^2}\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}_i}\right)^2}} \tag{3}$$

where y_i and $\hat{y}_i$ represent the ground truth label and prediction score of the ith image, and $\bar{y}$ and $\bar{\hat{y}}_i$ represent the average of all ground truth labels and prediction scores respectively. Higher PC and lower MAE and RMSE indicate better performance achieved by the FBP system [42, 43].

C. Compared with State-of-the-Art Methods

We conducted comparisons utilizing a range of techniques, including geometric feature-based and deep learning-based techniques, such as LR, GR, SVR, AlexNet, ResNet-18 and ResNeXt-50, etc. MAE, RMSE and PC are chosen as the metrics.

Five-fold cross-validation of facial beauty prediction is used to testify the network capacity via comparison, which holds 80% to 20% splitting for each fold in Table I and Fig. 3 show the comparison.

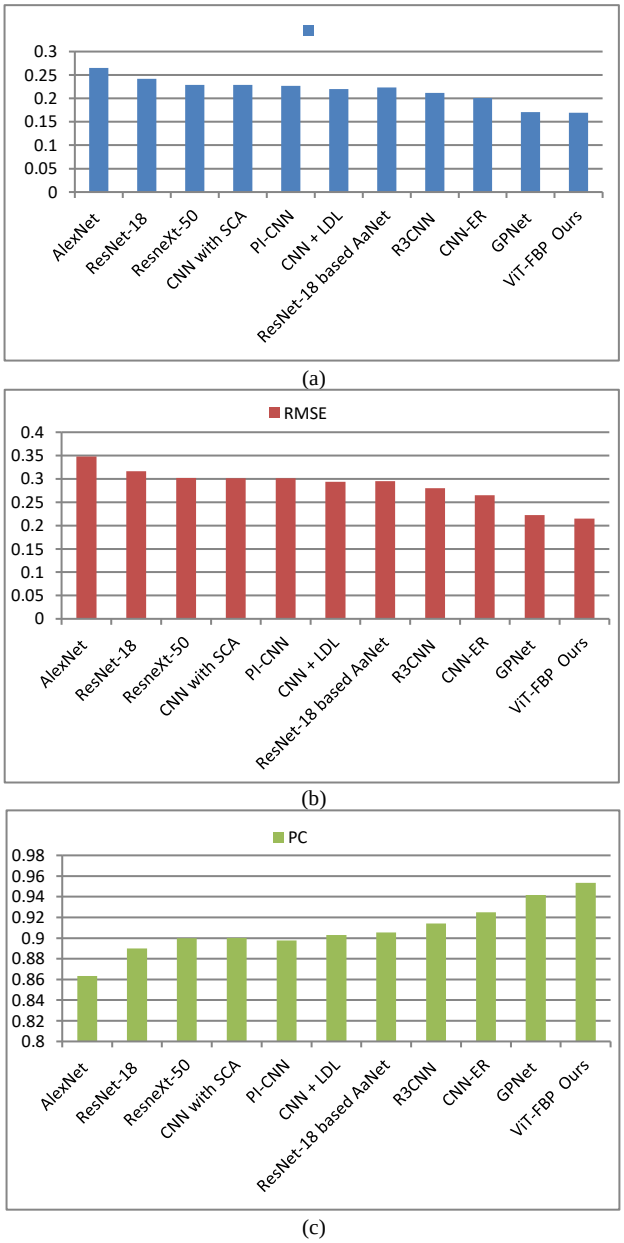


Fig. 3. Performance comparison of the five-fold cross validation, (a) Mean Absolute Error (MAE), (b) Root Mean Squared Error (RMSE) and (c) Pearson Correlation (PC).

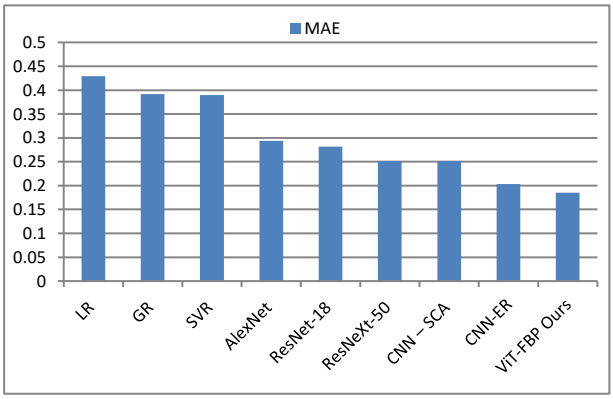
TABLE I: PERFORMANCE COMPARISON OF THE FIVE-FOLD CROSS VALIDATION

Methods	MAE	RMSE	PC
AlexNet [26]	0.2651	0.3481	0.8634
ResNet-18 [27]	0.2419	0.3166	0.89
ResneXt-50 [27]	0.2291	0.3017	0.8997
CNN with SCA [24]	0.2287	0.3014	0.9003
PI-CNN [44]	0.2267	0.3016	0.8978
CNN + LDL [20]	0.2201	0.294	0.9031
ResNet-18 based AaNet [45]	0.2236	0.2954	0.9055
R3CNN [28]	0.212	0.28	0.9142
CNN-ER [20]	0.2009	0.265	0.925
GPNet [18]	0.1706	0.2225	0.9415
ViT-FBP Ours	0.1691	0.2149	0.9534

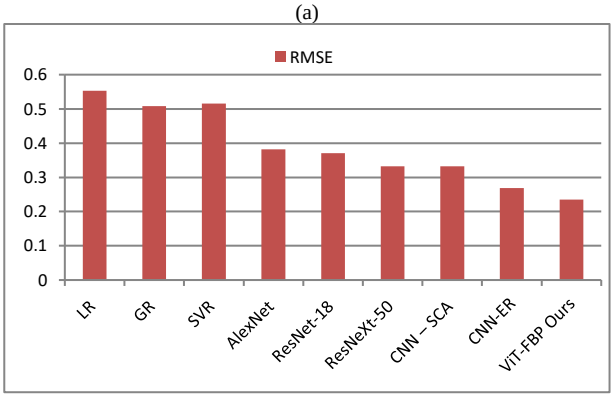
For 0.6 of the dataset is used for training, while 0.4 is used for testing. This means, 40% of the data set's instances are randomly chosen for testing, while the remaining 60% are randomly picked for training in Table II and Fig. 4 show the comparison.

TABLE II: PERFORMANCE COMPARISON OF DIFFERENT METHODS BY 60–40% SPLITTING

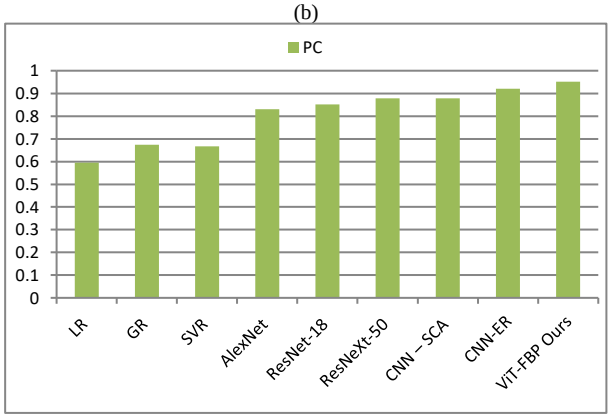
Methods	MAE	RMSE	PC
LR [24]	0.4289	0.5531	0.5948
GR [24]	0.3914	0.5085	0.6738
SVR [24]	0.3898	0.5152	0.6668
AlexNet [26]	0.2938	0.3819	0.8298
ResNet-18[27]	0.2818	0.3703	0.8513
ResNeXt-50 [27]	0.2518	0.3325	0.8777
CNN – SCA [24]	0.2517	0.332	0.878
CNN-ER [20]	0.2032	0.2683	0.9207
ViT-FBP Ours	0.1854	0.2347	0.9519



(a)



(b)



(c)

Fig. 4. Performance comparison of different methods by 60–40% splitting, (a) Mean Absolute Error (MAE), (b) Root Mean Squared Error (RMSE) and (c) Pearson Correlation (PC).

D. Discussion

According to our study, most researchers prefer supervised pre-trained models over those semi-supervised models, or create ones from the beginning. This is due to different reasons. First, pre-trained models are typically quicker to train since they simply need the hyper-parameters fine-tuned. Additionally, pre-trained models need the output layer to be modified depending on the task to provide the desired outputs. Typically, the quantity of parameters presents a limitation on performance improvement. The proposed model performs better than other models (AlexNet, ResNet-18, ResNeXt-50, CNN-SCA, and R3CNN).

Our network VIT-FBP holds 82.62 M parameters. CNN – SCA has 6.75 M parameters. ResNeXt-50 has 25.03 M parameters. AlexNet has 62.38 M parameters. The comparison reveals that our network is better than the cited works. The well-designed ViT operation can prevent the plentiful number of parameters. The comparisons of the two scenarios use the five folder of cross-validation 80%–20% split demonstrate the efficiency of our proposed approach. This tends to confirm that both the proposed ViT FBP Network played a crucial role in outperforming the State-of-the-Art methods.

Furthermore, some recent research in supervised learning applied an ensemble of deep convolutional neural network architectures, a pre-trained model, where the CNN-ER model uses two branches, Inception-v3 and ResneXt-50, with a dynamic loss function, or three various DCNNs, including pre-trained VGG16, AlexNet, and simple CNNs. In addition, geometric regulation The PGNet model, a hybrid network local CNN and a global Swin-Transformer structure, is used by regression-based face landmarks, which provides superior performance than earlier research using typically gender recognition and race classification as the auxiliary tasks. Thus, the comparisons of the two scenarios using the five folders of cross-validation and a 60%–40% split demonstrate the efficiency of our proposed approach. This tends to confirm that both of the proposed ViT-FBP networks played a crucial role in outperforming the state-of-the-art methods. However, since our goal is to predict the scores of predicting facial beauty, this model estimates its parameters using data from scores of facial beauty falling within a certain range. It can be deduced that the ground truth correlates with the most values of prediction.

VIT-FBP has the potential to be used in a variety of applications, such as:

- Facial beauty assessment: VIT-FBP could be used to develop more objective and reliable methods for assessing facial beauty. This could be useful for applications such as matchmaking, beauty pageants, and product design.
- Facial editing: VIT-FBP could be used to develop more realistic and effective facial editing tools. This could be used to help people improve their appearance or to create more believable digital avatars.

Although facial beauty estimation is adaptable and could be approached as a regression, classification, or

hybrid issue, most research approaches rely on a regression issue to obtain an accurate face beauty forecast.

E. Future Research

For more than a decade, convolutional neural networks have dominated computer vision research. To improve the efficiency and performance of tasks like image segmentation, object identification, and classification, recent developments in innovative topology, such as vision transformers, have shown a lot of promise. These models enable the modeling of long-range dependency by combining the Transformers' attention mechanism with convolutional neural networks' effectiveness. According to recent studies, the use of vision transformers for facial beauty prediction is a promising area for future research because the results are on par with or even better than those obtained using state-of-the-art deep convolutional neural network techniques. Furthermore, investigating the physiological reasons for facial preference, such as using fMRI to monitor brain activity during the perception of facial attractiveness or examining the impact of hormone levels on facial preference, can provide valuable insights into the underlying mechanisms of facial beauty perception.

Another promising area for future research is 3D facial beauty prediction. While deep convolutional neural networks have been effective in predicting facial beauty for 2D images, there is a lack of research on 3D images. Exploring the use of 3D datasets for facial beauty prediction in people of various ethnic backgrounds, as well as for applications such as 3D face plastic surgery, could lead to valuable contributions to the field. However, the use of facial beauty technology in medical and industrial settings still faces challenges. Therefore, understanding the notion of beauty is becoming increasingly crucial in medical settings due to the sharp rise in demand for cosmetic surgery in recent years.

V. Conclusion

Convolutional neural networks, a deep learning technique, have demonstrated promising outcomes in image processing, particularly for facial beauty prediction. However, they continue to face several difficulties, such as the overfitting issue and the need for huge datasets and powerful computing power.

In this paper, regardless of whether deep networks proved their effectiveness for facial beauty prediction tasks, we propose ViT-FBP as a vision transformer framework for facial beauty prediction, which has been successively applied to various deep learning techniques. The experimental findings show that our network can perform better than previous CNN baseline approaches. Experimental results showed that the proposed network achieved better performance compared to several works available in the open literature (AlexNet, ResNet-18, ResNeXt-50, CNN-SCA, and R3CNN). It improves the congruence assessment with human judgment.

Overall, VIT-FBP is a promising new approach to facial beauty prediction. It is more accurate and interpretable than previous methods, and it has the potential to be used in a variety of applications.

Future research directions for facial beauty prediction using a hybrid of Visual Transformers have achieved significant progress and show encouraging results that, on a number of benchmarks, are on par with or even outperform the current state-of-the art deep convolutional neural network approaches. The use of facial beauty technology in medical settings and industry is still fraught with difficulties. Knowledge of beauty is becoming crucial for medical settings because of the sharp rise in demand for cosmetic surgery over the past several years.

Besides, it is important to consider the variety of raters, including race, gender, age, and ethnicity. This study analyzed the current state of the art in the field of facial beauty prediction and offered a clear vision for future research.

Acknowledgments

This work was partially funded by Directorate General for Scientific Research and Technological Development (DGRSDT), Ministry of Higher Education and Scientific Research – Algeria.

Data Availability Statement

The dataset SCUT-FBP5500 analyzed during the current study is available in the Github repository, https://github.com/HCIILAB/SCUT-FBP5500-Database-Release

Conflict of Interests

The authors declare no conflict of interest.

Author Contributions

Djamel Eddine Boukhari; contribution was the proposition of the present method, design, and writing a draft of the manuscript, Ali Chemsa; formal analysis, interpretation, revision and proofreading, Riadh Ajgou; worked on Concepts, data analysis, and discussed the results.

References

[1] D. Zhang, F. Chen, and Y. Xu, *Computer Models for Facial Beauty Analysis*, Switzerland: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-32598-9

[2] B. N. Deekshith, Annapoornima, K. V. Pooja *et al*., "Facial expression recognition using property of symmetry," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 3, no. 3, pp. 61–67, 2014.

[3] H. Knight and O. Keith, "Ranking facial attractiveness," *The European Journal of Orthodontics*, vol. 27, no. 4 pp. 340–348, 2005.

[4] N. D. Rao, S. Thaherbasha, P. Balakrishna *et al*., "Face recognition by PHAse congruency modular kernel principal component analysis," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 6, no. 2, pp. 30–36, 2017.

[5] B. Gowthami, C. Maheswari, and K. Neelima, "Face recognition based on SLTP method under different emotions," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 6, no. 2, pp. 59–66, 2017.

[6] T. Leyvand, D. Cohen-Or, G. Dror and D. Lischinski, "Data-driven enhancement of facial attractiveness," *ACM Trans. on Graphics*, vol. 27, no. 3, pp. 1–9, 2008.

[7] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*, vol. 1, New York: Springer, 2011.
[8] J. Saeed and A. M. Abdulazeez, "Facial beauty prediction and analysis based on deep convolutional neural network: A review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 1–12, 2021.
[9] D. Gray, K. Yu, W. Xu *et al*., "Predicting facial beauty without landmarks," *Lecture Notes in Computer Science*, vol. 6316, pp. 434–447, 2010.
[10] T.-T.-K. Nga, P.-V. Tuan, I. Koo *et al.*, "Enhancing the classification accuracy of rice varieties by using convolutional neural networks," *International Journal of Electrical and Electronic Engineering & Telecommunications*, vol. 12, no. 2, pp. 150–160, 2023.
[11] D. Xie, L. Liang, L. Jin *et al.*, "SCUT-FBP: A benchmark dataset for facial beauty perception," *IEEE Int. Conf. on Systems, Man, and Cybernetics*, Hong Kong, China, 2015, pp. 1821–1826.
[12] A. Kagian, G. Dror, T. Leyvand *et al.*, A machine learning predictor of facial attractiveness revealing human-like psychophysical biases," *Vision Research*, vol. 48, no. 2, pp. 235–243, 2008.
[13] J. Gan, L. Xiang, Y. Zhai *et al*., 2M BeautyNet: Facial beauty prediction based on multi-task transfer learning," *IEEE Access*, vol. 8, pp. 20245–20256, 2020.
[14] L. Lin, L. Liang and L. Jin, R2-ResNeXt: A ResNeXt-based regression model with relative ranking for facial beauty prediction, in *Proc. 24th Int. Conf. on Pattern Recognition (ICPR)*, Beijing, China, 2018, pp. 85–90.
[15] D. E. Boukhari, A. Chemsa, R. Ajgou *et al.*, "An ensemble of deep convolutional neural networks models for facial beauty prediction," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27 no. 5. 2023.
[16] O. Guehairia, F. Dornaika, A. Ouamane *et al.*, "Facial age estimation using tensor based subspace learning and deep random forests," *Information Sciences*, vol. 609, pp. 1309–1317, 2022.
[17] O. Guehairia, A. Ouamane, F. Dornaika *et al.*, "Deep random forest for facial age estimation based on face images," in *Proc. 1st Int. Conf. on Communications, Control Systems and Signal Processing (CCSSP)*, El Oued, Algeria, 2020, pp. 305–309.
[18] K. Han, Y. Wang, H. Chen *et al*., "A survey on vision transformer," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *preprint arXiv*: 2010.11929, 2020.
[20] S. Khan, M. Naseer, M. Hayat *et al.*, "Transformers in vision: A survey," *ACM Computing Surveys*, vol. 54, no. 200, pp. 1–41, 2022.
[21] T. Peng, M. Li, F. Chen *et al.*, "Geometric prior guided hybrid deep neural network for facial beauty analysis," *CAAI Trans. on Intelligence Technology*, 2023. https://doi.org/10.1049/cit2.12197
[22] J. N. Saeed, A. M. Abdulazeez, and D. A. Ibrahim, "An ensemble dcnns-based regression model for automatic facial beauty prediction and analyzation," *Traitement du Signal*, vol. 40, no.1, pp. 55–63, 2023.
[23] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed, "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression," *Knowledge-Based Systems*, vol. 242, #108246. 2022.
[24] A. Chouchane, A. Ouamane, Y. Himeur *et al.*, "Improving CNN-based person re-identification using score normalization," in *Proc. of IEEE Int. Conf. on Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 2023, pp. 2890–2894.
[25] T. Lin, X. Chen, X. Tang *et al*., "Deep learning based classification of radar spectral maps," *International Journal of Electrical and Electronic Engineering & Telecommunications*, vol. 10, no. 2, pp. 99–104, 2021.
[26] M. Khammari, A. Chouchane, A. Ouamane *et al.*, "High-order knowledge-based Discriminant features for kinship verification," *Pattern Recognition Letters*, vol. 175, pp. 30–37, Nov. 2023.
[27] K. Cao, K. Choi, H. Jung *et al.*, "Deep learning for facial beauty prediction," *Information*, vol. 11, no. 8, #391, 2020.
[28] L. Liang, L. Lin, L. Jin *et al.*, "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," in *Proc. 24th Int. Conf. on Pattern Recognition (ICPR)*, Beijing, China, 2018, pp. 1598–1603.
[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[30] K. He, X. Zhang, S. Ren *et al*., "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.
[31] L. Lin, L. Liang and L. Jin, "Regression guided by relative ranking using convolutional neural network (R3CNN) for facial beauty prediction," *IEEE Trans. on Affective Computing*, vol. 13, no. 1, pp. 122–134, 2022.
[32] F. Dornaika, A. Moujahid, K. Wang *et al.*, "Efficient deep discriminant embedding: Application to face beauty prediction and classification," *Engineering Applications of Artificial Intelligence*, vol. 95, #103831, 2020.
[33] F. Dornaika, "Multi-similarity semi-supervised manifold embedding for facial attractiveness scoring," *Soft Computing*, vol. 27, pp. 5099–5108, Mar. 2023.
[34] F. Dornaika, K. Wang, I. Arganda-Carreras *et al.*, "Toward graph-based semi-supervised face beauty prediction," *Expert Systems with Applications*, vol. 142, #112990, 2020.
[35] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. on Neural Information Processing Systems*, 2017, pp. 6000–6010.
[36] K. Peng, A. Roitberg, K. Yang *et al.*, "TransDARC: Transformer-based driver activity recognition with latent space feature calibration," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 278–285.
[37] P. Mishra, R. Verk, D. Fornasier *et al.*, "VT-ADL: A vision transformer network for image anomaly detection and localization," in *Proc. IEEE 30th International Symposium on Industrial Electronics (ISIE)*, Kyoto, Japan, 2021, pp. 1–6.
[38] X. Wang, S. Zhang, Z. Qing *et al.*, "OadTR: Online action detection with transformers," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*, Montreal, QC, Canada, 2021, pp. 7545–7555.
[39] J. Gan, X. Xie, Y. Zhai, *et al.*, "Facial beauty prediction fusing transfer learning and broad learning system," *Soft Computing*, vol. 27, pp. 13391–13404, Nov. 2023.
[40] S. Shi, F. Gao, X. Meng *et al.*, "Improving facial attractiveness prediction via co-attention learning," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 4045–4049.
[41] I. Lebedeva, Y. Guo and F. Ying, "Transfer learning adaptive facial attractiveness assessment," *Journal of Physics: Conference Series*. vol. 1922, no. 1, #012004, 2021.
[42] I. Lebedeva, F. Ying, and Y. Guo, "Personalized facial beauty assessment: a meta-learning approach," *The Visual Computer: International Journal of Computer Graphics*, vol. 39, no. 3, pp. 1095–1107, 2023
[43] F. Chen and D. Zhang, "A benchmark for geometric facial beauty study," *Lecture Notes in Computer Science*, vol. 6165, pp. 21–32, 2010.
[44] J. Xu, L. Jin, L. Liang *et al.*, "Facial attractiveness prediction using psychologically inspired convolutional neural network (PI-CNN)," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 1657–1661.
[45] L. Lin, L. Liang, L. Jin *et al.*, "Attribute-aware convolutional neural networks for facial beauty prediction," in *Proc. the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp.847–853.

Djamel Eddine Boukhari is a research engineer in a scientific research center called CRSTRA. He received the M.S. degree from Biskra University in 2011. He is a PhD student in the Department of Electrical Engineering at University of El Oued. His research interest deep learning, computer vision, image compression and signal processing.

Ali Chamsa is a research professor in the Department of Electrical Engineering at the University of Eloued, Algeria. The PhD was received in automatic engineering from University of Biskra, Algeria in 2016. His research interest telecommunications, signal processing, estimation and detection theory.

Riadh Ajgou is a research professor in the Department of Electrical Engineering at the University of Eloued (Algeria). He received the doctorate and Master's and Engineer's degrees in 2016, 2010 and 2004 respectively, at the University of Biskra (Algeria). His research interest signal processing, speech processing and telecommunications.