Applied Intelligence (2023) 53:26110–26125
https://doi.org/10.1007/s10489-023-04931-8

TransBLS: transformer combined with broad learning system for facial beauty prediction

Junying Gan[1] · **Xiaoshan Xie**[1] · **Guohui He**[1] · **Heng Luo**[1]

Accepted: 29 July 2023 / Published online: 18 August 2023

Abstract
Facial beauty prediction (FBP) is a frontier topic in the fields of machine learning and computer vision, focusing on how to enable computers to judge facial beauty like humans. The existing FBP methods are mainly based on deep neural networks (DNNs). However, DNNs lack global characteristics and only build local dependencies, so FBP still suffers from insufficient supervision information, low accuracy and overfitting. A transformer is a self-attention-based architecture that possesses better global characteristics than DNNs and can build long-term dependencies. Transformers have been widely used to solve some computer vision problems in recent years and have produced better results. In this paper, we propose an adaptive transformer with global and local multihead self-attention for FBP, called GLAFormer. However, GLAFormer does not converge and is prone to overfitting when the training samples are insufficient. The broad learning system (BLS) can accelerate the model convergence process and reduce overfitting. Therefore, we further combine GLAFormer with the BLS to form TransBLS, in which a GLAFormer block is designed as a feature extractor, the features extracted by it are transferred to the BLS for further refining and fitting, and the results are output. Experimental results indicate that TransBLS achieves state-of-the-art FBP performance on several datasets with different scales, better solving the low accuracy and overfitting problems encountered in FBP. It can also be widely applied in pattern recognition and object detection tasks.

Keywords Adaptive transformer · Broad learning system · Facial beauty prediction · Global self-attention · Local self-attention

1 Introduction

Plato once said that "beauty is a natural advantage". There is no doubt that beauty affects all aspects of our lives all the time. Since Plato presented his theory of aesthetics, research has been performed in the fields of philosophy, psychology, and medicine to study the nature of beauty and the criteria for evaluating it; however, no scientific definition has been formed [1]. Facial beauty prediction (FBP) is a cutting-edge topic in artificial intelligence that concerns the nature and laws of human cognition; it focuses on how to enable computers to judge facial beauty like humans. FBP research can promote the development of related industries [2], such as beauty recommendations, aesthetic surgery planning, facial beautification and cartoon character design. Chen et al. [3] indicated that facial beauty can be learned by data-driven methods. FBP is usually formulated as a supervised learning problem involving level classification [4–6] or score regression [7–9]. Regardless of the chosen formulation, a prediction model and a means of facial representation are two key factors of FBP. In recent years, with the emergence of massive data and the development of deep learning techniques, data-driven and deep neural networks (DNNs)-based methods have been widely studied, and the accuracy and Pearson coefficient of FBP have been significantly improved [10–13]. However, due to the lack of global properties and long-term dependencies, DNNs tend to fall into local optima, resulting in poor generalizability. Additionally, FBP suffers from insufficient supervised information, low accuracy and overfitting.

Junying Gan and Xiaoshan Xie are both contributed equally.

✉ Junying Gan
junyinggan@163.com

Xiaoshan Xie
xiaoshanxie.xsx@gmail.com

Guohui He
ghhe126@126.com

Heng Luo
bigboloo@163.com

[1] Department of Intelligent Manufacturing, Wuyi University, Canghou, Jiangmen 529020, Guangdong, China

A transformer is a self-attention-based architecture [14] constructed by multihead self-attention (MSA) and multilayer perceptrons (MLPs). Compared with DNNs, transformers have better global properties and can build long-term dependencies. Therefore, for the past few years, many scholars have tried to use transformers, such as the vision transformer (ViT) [16], Swin transformer [17] and data-efficient image transformers (DeiT) [18], to address some issues in computer vision and have achieved good results [15], Furthermore, transformers have been widely used in face recognition and detection [19–22], and these transformer-based methods have produced significant results. However, the superior performance of transformer-based methods relies on large-scale training data to avoid the negative effects of a lack of local characteristics. Unfortunately, facial beauty databases are small-scale databases, so transformers do not converge, and overfitting occurs. The conventional method labels more facial beauty data, but it usually requires intensive feature annotation, which consumes considerable labour and time. Therefore, applications of transformers for FBP have not yet been reported in the literature.

The broad learning system (BLS) is a fast incremental learning system without a deep architecture that is composed of feature and enhancement nodes, a weight layer and an output layer [23]. The BLS maps the features of the input image directly into its feature nodes and enhancement nodes, and then these features are multiplied by a randomly generated weight matrix to obtain output weights and results. The BLS is forging ahead towards building faster and more efficient machine learning methods. In recent years, the BLS and its variants have already been applied in face recognition and detection studies [24, 25]. Naturally, the BLS has also been widely used in FBP [26, 27], and the results show that it can effectively speed up the model training and feature fitting processes. However, the overly simple structure of the BLS limits its feature extraction ability, resulting in low accuracy and poor generalizability. The performance of the BLS alone is far inferior to that of DNNs and transformers.

To address the issues described above, we propose a new FBP framework consisting of a transformer and the BLS, called TransBLS. In particular, we propose an adaptive transformer with global and local MSA, called GLAFormer, to better learn the global and local features of face images. We also improve a valid adaptive multilayer perceptron (AdaptMLP), effectively adapting the pretrained ViT to FBP to reduce the need for training data. However, GLAFormer does not converge, and overfitting occurs when the input training samples are insufficient. Therefore, we further fuse GLAFormer with the BLS. In our TransBLS, GLAFormer is used to extract facial beauty features, and then these features are migrated to the BLS for refining and fitting, effectively speeding up the convergence process and reducing overfitting. Facial beauty level classification experiments are performed on several databases, including SCUT-FBP [28], SCUT-FBP5500 [29], the Large Scale Asian Facial Beauty Dataset (LSAFBD) [30] and CelebA [31], where our TransBLS achieves state-of-the-art FBP performance. Extensive results indicate the effectiveness of TransBLS for FBP in comparison with common DNNs, including ResNet50 [32], InceptionV3 [33] and EfficientNetB7 [34], and transformer-based methods, including the ViT and Swin transformer. They also demonstrate that TransBLS is superior to DNN-based approaches and transformer-based methods in terms of FBP.

The contributions of this paper are summarized as follows.

- We are the first to merge a transformer and the BLS to solve the FBP problem. We propose a TransBLS framework for FBP that combines the global characteristics of a transformer with the fast incremental learning characteristics of the BLS while avoiding their disadvantages.
- We present GLAFormer with global and local MSA to better learn the global and local features of face images. We also improve an AdaptMLP to adapt a pretrained ViT to FBP to reduce the need for training data.
- We perform extensive experiments on several facial beauty databases, and the results show that our method achieves state-of-the-art performance, demonstrating its effectiveness, superiority and generalizability.

The rest of this paper is organized as follows. The related works are briefly reviewed in Section 2. In Section 3, we build the TransBLS architecture and elaborate on its algorithmic implementation. In Section 4, extensive experiments are conducted on several databases, and the results are analysed. Section 5 introduces the conclusion and future work ideas.

2 Related works

2.1 Facial beauty prediction

Early FBP studies adopted classic pattern recognition methods, which achieved some success by combining artificial features with shallow predictors [35]. However, these artificial features are low-level features, making it difficult to obtain effective facial representations. Fortunately, with the emergence of large-scale training data and the development of deep learning techniques, some works [7–13] combining FBP with DNNs have achieved good results in recent years. FBP methods based on DNNs have been shown to surpass traditional methods due to the nonlinear transformation of their hierarchical structures [36]. However, due to the lack of global properties and long-term dependen-

cies, DNNs tend to fall into local optima, resulting in poor generalizability and causing FBP to suffer from insufficient supervised information, low accuracy and overfitting. Furthermore, the variability of face images and the complexity of human perception make it difficult to quickly construct robust and effective FBP models. In this paper, we design a simple but effective self-attention-based architecture, namely, TransBLS, to fuse more supervised information into the level classification task of FBP to achieve improved classification performance.

2.2 Visual transformer

Transformers are self-attention-based architectures [14] that have been widely used in recent years to solve some problems in computer vision because of their excellent global properties and ability to establish long-term dependencies [15]. Dosovitskiy et al. [16] proposed a ViT architecture for image recognition, in which an image is split into fixed-size patches, and each patch is linearly embedded, with position embedding added. Then, these patches are fit to a standard transformer encoder, successfully extending the ViT to computer vision. Soon after, Liu et al. [17] designed a Swin transformer architecture, a hierarchical ViT using shifted windows, achieving advanced image classification, object detection and semantic segmentation performance. Chen et al. [37] further optimized the ViT and proposed an efficient transformer adaptation method, namely, AdaptFormer, which can adapt a pretrained ViT to many different image and video tasks. Vaswani et al. [38] designed a local self-attention scaling method for constructing parameter-efficient visual backbones.

Furthermore, transformers have been widely applied in face recognition and detection tasks. A new visual grammar and coding approach was designed with an unsupervised transformer [19]; this approach constructs visual sequences from a given cluster and uses the cosine distance coding method to obtain good face recognition results. Moreover, it utilizes a full temporal convolutional network as a feature extractor, transfers the extracted facial features to the transformer and recognizes forged faces in videos [20], demonstrating the effectiveness of transformers in face detection. A novel end-to-end network was designed for 3D face reconstruction based on a transformer and a conditional generative adversarial network [21]. There is no doubt that transformers have shown great potential for face detection and recognition. However, transformers need to be trained on large-scale data to avoid negative effects due to the lack of local features and deviations. Facial beauty databases are small-scale datasets, so transformers do not converge, and overfitting occurs. Thus, in this paper, we design a fusion module with global and local multihead self-attention, namely, GLAFormer, to better learn the global and local features of face images. We also improve an AdaptMLP to adapt a pretrained ViT to FBP to reduce the need for training data.

2.3 Broad learning system

The BLS and its variants have been widely applied in various vision tasks due to their simple and effective network architectures [39], which have proven to be valid for face feature extraction. The BLS has been applied in face recognition and understanding tasks [24, 25], such as face recognition, face expression recognition and face micro expression recognition, in which facial features (in terms of facial geometry, colour and texture) were directly characterized, and good results were achieved.

Furthermore, the BLS can accelerate model convergence and reduce overfitting, and it is widely used in FBP. Zhai et al. [26] presented an FBP method based on local feature fusion and the BLS, in which two-dimensional principal component analysis was used to reduce the dimensionality of the local fusion features of face images, and then these features were input into the BLS to train an efficient FBP model, achieving an accuracy of 58.97% within 13.33 s on the LSAFBD. Gan et al. [27] proposed a fusion model via transfer learning and a broad learning system for FBP, which regards DNNs as feature extractors via transfer learning. Additionally, a connection layer was designed to fuse these extractors with the BLS, achieving an accuracy of 62.13% within 2286.54 s on the LSAFBD and an accuracy of 74.69% within 1291.83 s on SCUT-FBP5500, with better balancing accuracy and time efficiency. Some works have focused on the combination of the BLS and DNNs [40, 41], in which convolutional and pooling layers are used to efficiently extract image features and reduce the required number of parameters. The BLS is used to fit the features. Combining the advantages of a convolutional neural network (CNN) and BLS, it has been proven that CNN-BLS is better than the BLS in terms of facial feature extraction. Recently, Su et al. proposed a multi-attention BLS architecture by fusing multihead self-attention with the BLS [42]. From that above, it is clear that the BLS achieves low FBP accuracy. Therefore, in this paper, we propose a GLAFormer block to extract facial beauty features, which are transferred to the BLS for refining and fitting.

3 Method

3.1 Overall architecture

The overall architecture of TransBLS is shown in Fig. 1(a), and it consists of a linear projection mechanism, position embeddings, a GLAFormer block and a BLS module. An input image is split into fixed-size patches. Each patch is

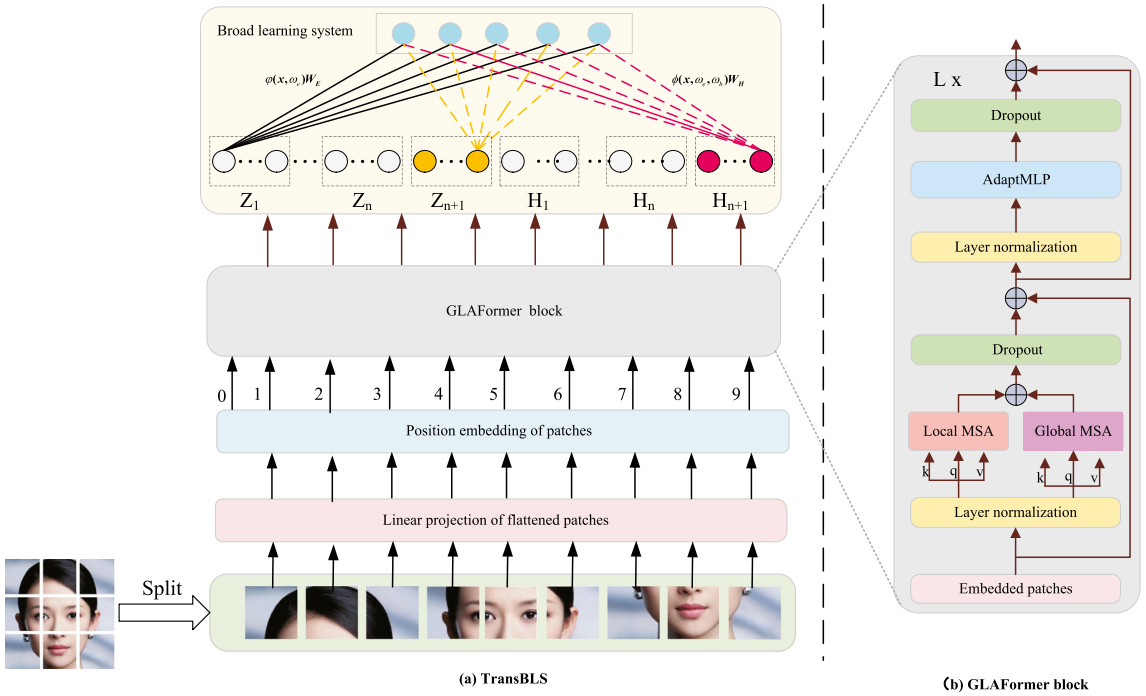


Fig. 1 (a) The overall architecture of TransBLS; (b) The architecture of GLAFormer Block

treated as a token after performing linear embedding and position embedding. Several GLAFormer blocks and an AdaptMLP are applied on these patch tokens to extract features, these features are then transferred to the BLS for refining and fitting, and the FBP results are output. More details are shown in Algorithm 1.

3.2 GLAFormer block

Each GLAFormer block is constructed by a global and local MSA fusion module and an AdaptMLP module. As illustrated in Fig. 1(b), a GLAFormer block contains a fusion module-based global and local MSA, followed by an AdaptMLP with a GeLU function in between them. Layer normalization (LN) is applied before each fusion module and each AdaptMLP. Dropout and residual connection are used after each fusion module and each AdaptMLP.

3.3 Global and local multihead self-attention fusion module

The standard ViT architecture for image recognition conducts global MSA, where the relationships between a token and all other tokens are computed, leading to quadratic complexity [16]. Therefore, the Swin transformer architecture was designed by computing self-attention within local windows. Vaswani et al. [38] designed a scaling local two-dimensional self-attention mechanism. To ensure that our model can better learn the global and local features of face images, we propose a global and local MSA fusion module,

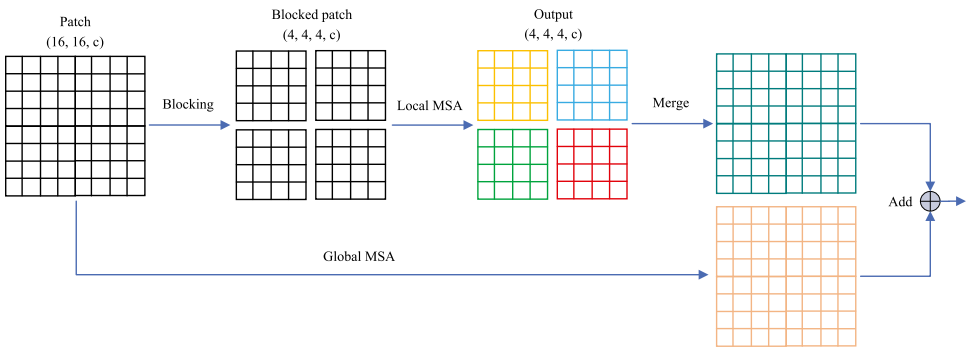


Fig. 2 The global and local multi-head self-attention fusion module

as shown in Fig. 2. Assuming that the size of each patch is $m \times m$ and that the size of the input face image is $h \times w$, the computational complexities of the fusion module, global MSA and local MSA are, respectively:

$$\Phi(\text{GLMSA}) = \Phi(\text{GMSA}) + \Phi(\text{LMSA}) \tag{1}$$

$$\Phi(\text{GMSA}) = 4hwC^2 + 2(hw)^2C \tag{2}$$

$$\Phi(\text{LMSA}) = 4hwC^2 + 2(s \cdot m)^2 hwC \tag{3}$$

where C is the number of hidden layer channels and s is a scaling factor used to adjust the patch size. In TransBLS, $s = 0.25$. The computational complexity does not increase.

3.4 Adaptive multi-layer perceptron

In AdaptFormer [37], the AdaptMLP architecture, a plug-and-play bottleneck module, was proposed to adapt the pretrained ViT to many tasks, and the module consists of two subbranches. It downsamples the features, then upsamples the features, and combines them with an MLP layer. To adapt the pretrained ViT to FBP and reduce the need for training data, we improve an AdaptMLP, as shown in Fig. 3; our version consists of an MLP layer and an adaptive branch. The adaptive branch includes downsampling, followed by upsampling with the GeLU function in between them. Then, multiplication is performed with a scale factor a. In TransBLS, $a = 10$. It is connected to the MLP layer through a residual connection. The AdaptMLP is computed as

$$x_a' = \text{MLP}(x_a) + a \cdot \text{GeLU}(x_a \cdot \omega_u) \cdot \omega_d + x_a \tag{4}$$

where x_a is the input feature of the AdaptMLP and x_a' is the output feature. ω_d is the downsampling parameter and ω_u is the upsampling parameter.

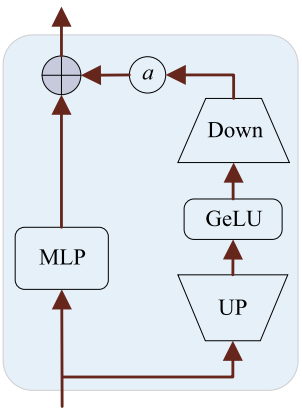


Fig. 3 The AdaptMLP module

With the fusion module and AdaptMLP module, the GLAFormer block is computed as follows:

$$Z_l' = \text{GLMSA}\left(\text{LN}\left(Z_{l-1}\right)\right) + Z_{l-1}, l = 1, \cdots, L \tag{5}$$

$$Z_l = \text{AdaptMLP}\left(\text{LN}\left(Z'\right)_l\right) + Z_l', l = 1, \cdots, L \tag{6}$$

$$x' = \text{LN}(Z_L^0) \tag{7}$$

where x' are output features of the GLAFormer block.

Algorithm 1 Facial beauty prediction via TransBLS

Input: training sample set x
Output: output matrix set y
1: Split x into fixed-size patches;
2: Perform linear embedding on the flattened patches;
3: Add the position embeddings of the flattened patches;
4: Input the patches into GLAFormer blocks, and extract the facial beauty features x' by (1),(2),(3),(4),(5),(6),(7);
5: **for** $i = 0, i \leqslant n$ **do**
6: Randomly initialize ω_e, ω_h;
7: Calculate $A = [\Phi(x, \omega_e)|\Omega(x, \omega_e, \omega_h)]$;
8: Calculate W_E, W_H by (9);
9: Calculate y by (8);
10: **end for**
11: **while** t **do**he loss threshold is not satisfied do
12: Add feature and enhancement nodes or new data;
13: Repeat steps 5 to 10;
14: Output the results of TransBLS;
15: **end while**

3.5 Broad learning system block

The features output from a GLAFormer block are migrated to the BLS block for further refining and fitting. In this part, we use φ to indicate the mapping function of the feature nodes and ϕ to represent the composite mapping function of the feature and enhancement nodes. The structure of the BLS can be expressed as

$$y = \varphi(x, \omega_e)W_E + \phi(x, \omega_e, \omega_h)W_H \tag{8}$$

where x denotes the input features of the BLS and y represents the output results of the BLS. ω_e and ω_h are random weights, and W_E and W_H are the output weights of the feature and enhancement nodes, respectively, calculated by

$$[W_E|W_H] = A^{-1}y = \lim_{\lambda \to 0} (\lambda I + A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y \tag{9}$$

where $A = [\Phi(x, \omega_e)|\Omega(x, \omega_e, \omega_h)]$. A^{-1} and A^{T} are the inverse matrix and the transpose matrix of A , respectively. I is the identity matrix.

Table 1 Specifications of the TransBLS architectures

Specification	TransBLS-T	TransBLS-S	TransBLS-B	TransBLS-L	TransBLS-H
Number of layers	12	12	12	24	32
Hidden sizes	192	384	768	1024	1280
MLP sizes	768	1536	3072	4096	5120
Number of Heads	3	6	12	16	16
Number of feature nodes	784	960	992	2278	2482
Number of enhancement nodes	1130	1888	1932	2936	3566

4 Experimental results and analysis

4.1 Experimental models

We build our base model, called TransBLS-B, in which the model size and computational complexity are similar to those of ViT-B. To further clarify the characteristics of each component of TransBLS, TransBLS-T, TransBLS-S, TransBLS-L and TransBLS-H are also built. The hyperparameters of these model variants are listed in Table 1, which mainly include the number of layers, hidden size, multilayer perceptron size, multihead attention size, number of feature nodes, and number of enhancement nodes. While the number of layers, hidden size, multilayer perceptron size, and multihead attention size are consistent with those of ViT, Hyperopt [43] is used to optimize the numbers of feature nodes and enhancement nodes in this paper.

4.2 Experimental objects

The LSAFBD [30] is a facial beauty prediction database constructed by our group, which contains 100,000 frontal images that cover different backgrounds, poses and ages. A total of 20000 images are labelled, and 80000 are unlabelled with the resolutions of 144 × 144. However, we only use 10000 labelled female images to study the properties of the proposed method. The images are divided into five categories, represented by "0", "1", "2", "3" and "4", corresponding to "extremely unattractive", "unattractive", "average", "attractive" and "extremely attractive", respectively; these are the

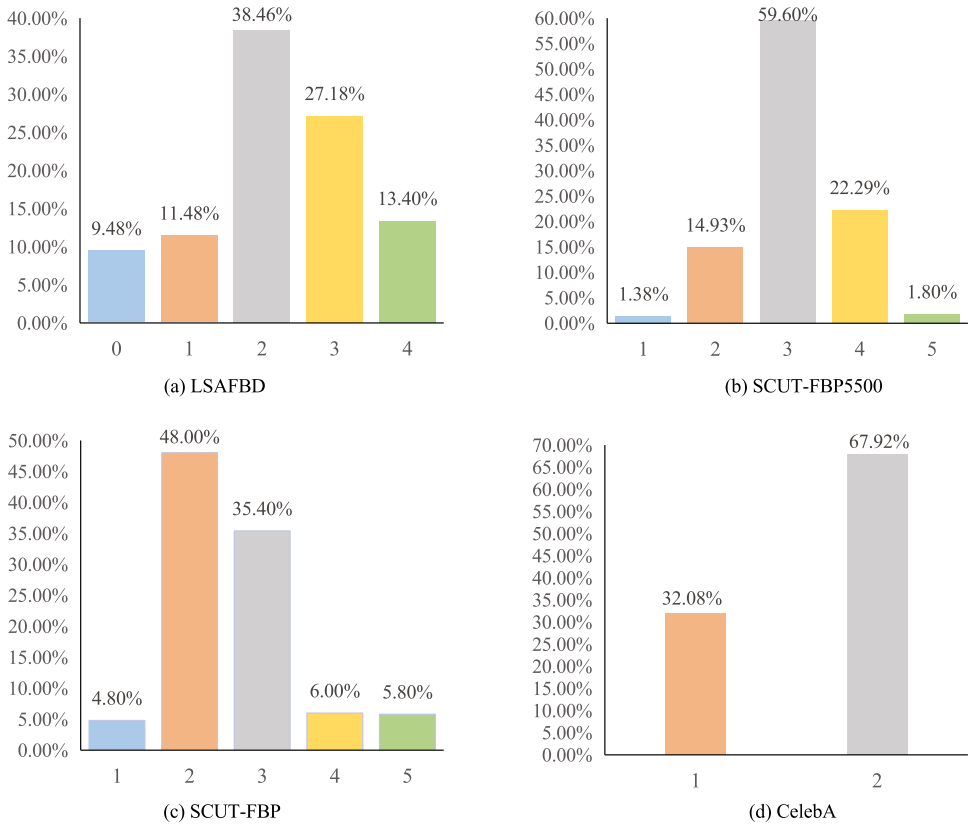


Fig. 4 Fig. 4 The beauty level distributions of different experimental databases. (a) The beauty level distribution of LSAFBD; (b) the beauty level distribution of SCUT-FBP5500; (c) the beauty level distribution of SCUT-FBP; (d) the beauty level distribution of CelebA

average results scored by 75 volunteers. There are 948 images in level "0", 1148 images in level "1", 3846 images in level "2", 2718 images in level "3" and 1333 images in level "4". Fig. 4(a) shows the distribution of the beauty labels.

SCUT-FBP5500 [29] is a facial beauty prediction database constructed by South China University of Technology; it contains 5,500 unobstructed frontal images with the resolutions of 350 × 350,covering different ages, genders and races. The beauty scores of all the images range from 1 to 5, with higher scores representing more beauty, as rated by 60 volunteers. Each image has 60 corresponding scores. We use the mode as the criterion, and all the images are divided into five grades, represented by "1", "2", "3", "4" and "5", which correspond to "extremely unattractive", "unattractive", "average", "attractive" and "extremely attractive", respectively. There are 76 images in level "1", 821 images in level "2", 3278 images in level "3", 1226 images in level "4", and 99 images in level "5". Fig. 4 (b) presents the distribution of the beauty labels.

SCUT-FBP [28] is another facial beauty prediction database built by South China University of Technology, which contains 500 frontal images of Asian females. The beauty scores of all the images range from 1 to 5, as rated by 75 volunteers. Each image has 60 corresponding scores. Utilizing the mode as the criterion, all the images are classified into five grades, represented by "1", "2", "3", "4" and "5", which correspond to "extremely unattractive", "unattractive", "average", "attractive" and "extremely attractive", respectively. There are 24 images in level "1", 240 images in level "2", 177 images in level "3", 30 images in level "4", and 29 images in level "5". Fig. 4(c) shows the distribution of the beauty labels.

CelebA [31] is a celebrity face database built by the Chinese University of Hong Kong, which contains 202,599 images of 10,177 celebrities. Each image is marked with 40 attribute features, such as "willow eyebrows", "attractive" and "pointy nose". There are 118,165 female images and 84,434 male images. According to the "attractive" attribute, we divide all the female images into two categories, "unattractive" and "attractive", which are denoted by "1" and "2", respectively. There are 37,911 images in level "1" and 80,254 images in level "2". Fig. 4(d) presents the distribution of the beauty labels.

4.3 Experiment environment and evaluation indices

In our experiment, the SCUT-FBP, SCUT-FBP5500, LSAFBD and CelebA databases are randomly divided into a training set and a testing set at a ratio of 8:2. ResNet50, InceptionV3, EfficientNetB7, the ViT, the Swin transformer and MLP-Mixer [44] based on transfer learning are trained with the same hyperparameters. The initial learning rate is 0.0001. When the training accuracy does not improve for more than 3 epochs, the multiplicative learning rate decay factor is 0.5. The batch size and number of epochs are 16 and 50, respectively. To ensure the reliability of the experiment, the categorical cross-entropy loss is used for cross-validation in this paper. The initial weights of these models are derived from ImageNet. All the experiments are implemented on a Python platform with a desktop computer possessing a NVIDIA GeForce RTX 3090 Ti, and all the algorithms are realized under the TensorFlow and Keras frameworks.

In this paper, FBP is formulated as a supervised classification learning problem. Therefore, the accuracy, precision, recall, F1 score and area under the curve (AUC) metrics [45] are used as indices for evaluating the performance of all the models. Accuracy is the ratio of the number of correctly classified samples to the total number of samples. Precision is the ratio of predicted positive samples. Recall is the number of correctly classified results divided by the number of results that should have been returned as positive. The F1 score is defined as the summed average of precision and recall. The AUC indicates the probability that a positive prediction example ranks ahead of a negative example, which is an evaluation metric that measures the merit of the tested classification model. These metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{12}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{13}$$

where TP is the number of true facial beauty results. TN is the number of true facial nonbeauty results. FP is the number of false facial beauty results. FN is the number of false facial nonbeauty results.

4.4 Experiments conducted on a small-scale database

We conduct classification experiments concerning FBP on a small-scale database, i.e., SCUT-FBP, to initially explore the properties of TransBLS and evaluate its effectiveness. Table 2 lists comparisons with the other models, including transformer-based methods, MLP-Mixers and common DNNs.

As seen from Table 2, when the image size is 224×224, TransBLS-B achieves 66.67% accuracy, 66.53% precision, 64.58% recall, a 65.54% F1 score and an 80.06% AUC, which are the best results among all the models. In particular, TransBLS-B provides a significant improvement over

Table 2 Results obtained by different methods on SCUT-FBP

Methods	Image size	Accuracy	Precision	Recall	F1 Score	AUC
InceptionV3 [33]	224×224	57.29	56.58	44.79	50.00	78.74
EfficientNetB7 [34]	224×224	58.33	56.63	48.96	52.52	76.63
ResNet50 [32]	224×224	60.42	60.67	56.25	58.38	78.35
Mixer-B/16 [44]	224×224	56.25	57.30	53.12	55.13	78.65
Mixer-L/16 [44]	224×224	55.21	55.95	48.96	55.58	77.68
ViT-B/16 [16]	224×224	53.62	51.51	42.92	46.82	65.75
Swin-B [17]	224×224	63.54	64.21	63.54	63.87	78.18
TransBLS-B (ours)	224×224	66.67	66.53	64.58	65.54	80.06
ViT-B/16 [16]	384×384	54.79	53.85	41.88	47.12	65.82
Swin-B [17]	384×384	64.58	63.44	61.46	62.43	78.73
TransBLS-B (ours)	384×384	67.63	67.70	65.63	66.08	80.61

the vision transformer and Swin transformer, where the accuracy of TransBLS-B is 13.05% and 3.13% higher than those of ViT-B/16 and Swin-B, the precision of TransBLS-B is 15.02% and 2.32% higher than those of these two models, the recall of TransBLS-B is 21.66% and 1.04% higher than those of these two models, the F1 score of TransBLS-B is 18.72% and 1.67% higher than those of these two models, and the AUC of TransBLS-B is 14.31% and 1.88% higher than those of these two models, respectively. When the image size is 384×384, the performance of TransBLS-B is improved. It is important to note that TransBLS-B also achieves better performance than ViT-B/16 and Swin-B.

Compared with the common DNNs and MLP-Mixers, the proposed TransBLS-B methods also outperforms InceptionV3, EfficientNetB7, ResNet50, Mixer-B16 and Mixer-L16 on these evaluation metrics. There is no doubt that TransBLS achieves improved FBP performance.

4.5 Experiments conducted on medium-scale databases

We perform classification experiments concerning FBP on medium-scale databases, i.e., SCUT-FBP5500 and the LSAFBD, and the results are listed in Tables 3 and 4, respectively.

We can see from Table 3 that TransBLS-B displays the best performance on SCUT-FBP5500 in terms of all defined metrics in comparison with the other models. Specifically, when the image size is 224×224, our TransBLS-B obtains an accuracy of 77.32% with significant increases of 15.21% and 1.36% over those of the vision transformer and Swin transformer, a precision of 77.19% with significant increases of 11.00% and 1.26% over those of these two models, a recall of 77.05% with significant increases of 22.13% and 1.18% over those of the two models, an F1 score of 77.12% with significant increases of 17.09% and 1.22% over those of these

Table 3 Results obtained by different methods on SCUT-FBP5500

Methods	Image size	Accuracy	Precision	Recall	F1 Score	AUC
InceptionV3 [33]	224×224	73.13	73.60	69.58	71.53	83.54
EfficientNetB7 [34]	224×224	71.40	72.74	69.03	70.84	83.62
ResNet50 [32]	224×224	71.31	71.76	70.13	70.94	83.34
Mixer-B/16 [44]	224×224	65.03	65.73	64.12	64.92	83.71
Mixer-L/16 [44]	224×224	64.39	66.15	62.30	64.17	82.90
ViT-B/16 [16]	224×224	62.11	66.19	54.92	60.03	81.69
Swin-B [17]	224×224	75.96	75.93	75.87	75.90	85.20
TransBLS-B (ours)	224×224	77.32	77.19	77.05	77.12	86.93
ViT-B/16 [16]	384×384	61.84	64.06	57.47	60.59	84.29
Swin-B [17]	384×384	74.95	74.93	74.86	74.89	85.31
TransBLS-B (ours)	384×384	77.69	77.85	77.59	77.72	87.07

Table 4 Results obtained by different methods on the LSAFBD

Methods	Image size	Accuracy	Precision	Recall	F1 Score	AUC
InceptionV3 [33]	224×224	57.67	65.49	44.79	53.20	85.93
EfficientNetB7 [34]	224×224	56.76	65.52	41.03	50.46	85.83
ResNet50 [32]	224×224	61.62	65.45	54.86	59.69	86.57
Mixer-B/16 [44]	224×224	58.72	60.02	56.71	58.32	85.14
Mixer-L/16 [44]	224×224	59.02	60.24	57.16	58.66	84.80
ViT-B/16 [16]	224×224	58.12	60.47	53.96	57.03	85.52
Swin-B [17]	224×224	62.90	62.89	62.75	62.82	86.74
TransBLS-B (ours)	224×224	63.83	63.72	63.08	63.40	86.46
ViT-B/16 [16]	384×384	58.32	61.13	48.15	53.87	85.46
Swin-B [17]	384×384	63.35	63.57	62.95	63.26	87.07
TransBLS-B (ours)	384×384	63.97	64.14	63.52	63.83	87.17

two models, and an AUC of 89.93% with significant increases of 7.24% and 1.73% over those of these two models. When the image size is 384×384, the performance of TransBLS-B is improved and outperforms ViT-B/16 and Swin-B on these evaluation metrics. Simultaneously, compared with InceptionV3, EfficientNetB7, ResNet50, Mixer-B16 and Mixer-L16, TransBLS-B has better performance, demonstrating its validity and generalisation

Similarly, we can see from Table 4 that when the image size is 224×224, TransBLS-B obtains 63.83% accuracy, 63.72% precision, 63.08% recall, a 63.40% F1 score and an 86.46% AUC on the LSAFBD. Furthermore, when the image size is 384×384, TransBLS-B achieves the highest facial beauty classification performance on the LSAFBD, reaching 63.97% accuracy, 64.14% precision, 63.52% recall, a 63.83% F1 score and an 87.17% AUC. It is clear that TransBLS-B has better FBP performance than with the other models, demonstrating the validity of the proposed method.

4.6 Experiments conducted on a large-scale database

We perform FBP level classification experiments on a large-scale database, i.e., CelebA, and the results are listed in Table 5. On CelebA, little difference is observed among the classification accuracies of all the models for FBP, but TransBLS-B still has state-of-the-art performance; its accuracy, precision, recall, F1 score and AUC reach 81.91%, 81.26%, 81.65%, 81.45% and 86.02%, respectively.

We study the characteristics of TransBLS on four databases. The results show the following.

(1) On SCUT-FBP, the performance of the transformers is far inferior to that of the common DNNs. This is because a transformer needs to be trained on numerous data to avoid the negative effects induced by a lack of local characteristics and deviations. However, TransBLS-B achieves good results, effectively solving this problem.

Table 5 Results obtained by different methods on CelebA

Methods	Image size	Accuracy	Precision	Recall	F1 Score	AUC
InceptionV3 [33]	224×224	78.88	78.90	78.81	78.85	83.93
EfficientNetB7 [34]	224×224	79.72	79.65	79.70	79.67	85.17
ResNet50 [32]	224×224	79.42	79.37	79.30	79.33	84.95
Mixer-B/16 [44]	224×224	74.99	74.68	74.77	74.72	81.95
Mixer-L/16 [44]	224×224	76.57	76.41	76.34	76.37	82.65
ViT-B/16 [16]	224×224	75.41	75.26	75.44	75.35	80.20
Swin-B [17]	224×224	81.06	81.10	81.02	81.06	84.12
TransBLS-B (ours)	224×224	81.85	81.13	81.51	81.32	85.90
ViT-B/16 [16]	384×384	75.65	75.35	75.47	75.41	82.64
Swin-B [17]	384×384	81.17	81.15	81.10	81.12	84.18
TransBLS-B (ours)	384×384	81.91	81.26	81.65	81.45	86.02

Table 6 Results obtained by models with different parameter scales on SCUT-FBP

Model	Parameter Scale	Accuracy	Precision	Recall	F1 Score	AUC
ViT [16]	ViT-T/16 [15]	50.71	53.85	38.54	44.93	67.90
	ViT-S/16 [15]	51.79	53.66	41.67	46.91	72.35
	ViT-B/16 [15]	53.62	51.51	42.92	46.82	69.81
	ViT-L/16 [15]	54.50	52.33	46.88	49.46	70.64
Swin [17]	Swin-T [16]	61.46	63.44	61.46	62.43	73.82
	Swin-S [16]	58.33	59.14	57.29	58.20	73.73
	Swin-B [16]	63.54	64.21	63.54	63.87	78.18
	Swin-L [16]	62.50	65.26	64.58	64.92	77.20
TransBLS (Ours)	TransBLS-T	64.58	64.86	62.83	63.83	77.12
	TransBLS-S	65.63	66.59	64.04	65.29	78.10
	TransBLS-B	66.67	67.70	64.61	66.12	80.06
	TransBLS-L	68.75	69.24	67.04	68.12	81.67
	TransBLS-H	68.91	69.48	67.18	68.31	81.85

(2) On SCUT-FBP5500 and the LSAFBD, the performance of the transformers is improved, but they are still not as good as the common DNNs overall. Even on CelebA, the transformers still fail to outperform the common DNNs, further demonstrating that a transformer relies heavily on a large amount of training data. TransBLS-B achieves state-of-the-art performance, indicating that our method is a good solution to this problem.

(3) TransBLS cleverly combines the advantages of transformers and the BLS while overcoming their disadvantages. On all the databases, TransBLS achieves state-of-the-art performance and outperforms the previously developed methods, which demonstrates its superiority and validity for FBP.

4.7 Ablation study

Different parameter scales. The ablation results obtained by TransBLS variants with different parameter scales on SCUT-FBP and CelebA are listed in Tables 6 and 7, respectively, and they are compared with the ViT and Swin transformer.

The FBP performance of the ViT and Swin transformer models does not always improve when the parameter scale increases, but these models consume more time and computing resources. For example, on CelebA, ViT-T/16 has the smallest parameter scale, but it outperforms ViT-S/16, ViT-B/16 and ViT-L/16 in terms of all the evaluation metrics. Similarly, on SCUT-FBP, Swin-T has a smaller parameter

Table 7 Results obtained by models with different parameter scales on CelebA

Model	Parameter Scale	Accuracy	Precision	Recall	F1 Score	AUC
ViT [16]	ViT-T/16 [16]	76.48	76.38	76.31	76.34	84.41
	ViT-S/16 [15]	75.96	75.88	75.67	75.77	84.20
	ViT-B/16 [15]	75.41	75.26	75.44	75.35	80.20
	ViT-L/16 [15]	75.62	75.57	75.52	75.54	79.40
Swin [17]	Swin-T [16]	81.08	81.03	81.08	81.05	84.46
	Swin-S [16]	80.98	80.90	80.84	80.87	83.96
	Swin-B [16]	81.06	81.10	81.02	81.06	84.12
	Swin-L [16]	81.31	81.19	81.13	81.16	84.57
TransBLS (Ours)	TransBLS-T	81.41	81.08	81.14	81.11	85.63
	TransBLS-S	81.52	81.10	81.30	81.20	85.76
	TransBLS-B	81.85	81.13	81.51	81.32	85.90
	TransBLS-L	81.87	81.34	81.71	81.52	86.23
	TransBLS-H	81.95	81.47	81.82	81.64	86.71

Table 8 Results obtained by TransBLS variants with different patch resolutions on SCUT-FBP5500

Patch Resolution	Model	Accuracy	Precision	Recall	F1 Score	AUC
8×8	TransBLS-T	74.03	71.80	73.84	72.81	80.22
	TransBLS-S	74.75	72.86	74.21	73.53	80.44
	TransBLS-B	75.21	73.35	75.03	74.18	81.83
	TransBLS-L	75.67	74.96	75.39	75.17	82.27
	TransBLS-H	76.67	75.83	76.12	75.97	82.57
16×16	TransBLS-T	75.23	73.03	75.05	74.03	81.58
	TransBLS-S	76.05	73.85	75.59	74.71	83.88
	TransBLS-B	77.32	77.19	77.05	77.12	86.93
	TransBLS-L	78.32	77.44	78.14	77.79	87.72
	TransBLS-H	78.46	77.53	78.32	77.92	88.01
32×32	TransBLS-T	74.21	72.67	74.03	73.34	80.74
	TransBLS-S	75.22	73.41	74.86	74.13	83.62
	TransBLS-B	76.04	74.94	75.87	75.40	85.50
	TransBLS-L	76.86	75.51	76.68	76.09	86.60
	TransBLS-H	77.58	76.10	77.30	76.70	87.09

scale and better performance than Swin-S in terms of all the evaluation metrics.

The FBP performance of TransBLS is significantly improved as the parameter scale increases. For instance, on SCUT-FBP, the accuracy improves from 64.58% for TransBLS-T to 68.91% for TransBLS-H, the precision increases from 64.86% for TransBLS-T to 69.48% for TransBLS-H, the recall improves from 62.83% to 67.18%, the F1 score increases from 63.83% to 68.31%, and the AUC improves from 77.12% to 81.85%. The same pattern is found on CelebA. Although the improvement is small, the performance of TransBLS is improved in general. Moreover, TransBLS always outperforms the ViT and the Swin transformer, so it can also be widely applied in pattern recognition and object detection tasks.

Different patch resolutions. The ablation results obtained by TransBLS variants with different patch resolutions on the LSAFBD and SCUT-FBP5500 are listed in Tables 8 and 9, respectively.

As we can see from Table 8, on SCUT-FBP5500, when the patch resolution increases from 8×8 to 16×16, the performance of TransBLS exhibits an increasing trend. For example, the accuracy of TransBLS-B rises from 75.21% to 78.32%, and the precision of TransBLS-B rises from 73.35% to 77.19%. However, when the patch resolution increases from 16×16 to 32×32, a decreasing trend is observed. For

Table 9 Results obtained by TransBLS variants with different patch resolutions on the LSAFBD

Patch Resolution	Model	Accuracy	Precision	Recall	F1 Score	AUC
8×8	TransBLS-T	60.45	60.07	60.19	60.13	80.99
	TransBLS-S	61.60	61.80	61.25	61.51	82.47
	TransBLS-B	62.50	62.03	62.35	62.19	82.96
	TransBLS-L	63.25	63.18	63.07	63.12	84.21
	TransBLS-H	64.50	63.96	64.15	64.05	85.36
16×16	TransBLS-T	61.37	61.29	61.17	61.23	83.39
	TransBLS-S	62.98	62.63	62.58	62.60	86.18
	TransBLS-B	63.83	63.08	63.40	63.24	86.46
	TransBLS-L	66.34	65.90	65.83	65.86	87.21
	TransBLS-H	66.49	66.21	66.03	66.12	87.39
32×32	TransBLS-T	61.34	60.47	60.29	60.38	81.87
	TransBLS-S	62.63	62.71	62.48	62.59	83.87
	TransBLS-B	63.18	62.78	63.08	62.93	85.82
	TransBLS-L	63.68	63.72	63.43	63.57	85.22
	TransBLS-H	64.25	64.02	64.79	64.40	85.57

Table 10 Results obtained by TransBLS-L on SCUT-FBP5500 with enhancement nodes added

Feature nodes	Enhancement nodes	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
2278	2000	77.32	75.41	74.85	75.13	86.57	1129.21
2278	2000→2500	77.78	75.93	74.92	75.42	87.69	1142.59
2278	2500→3000	77.69	75.56	74.76	75.16	87.18	1147.06
2278	3000→3500	78.14	76.35	75.73	76.04	87.72	1149.91
2278	3500→4000	77.96	75.76	75.06	75.41	87.21	1154.81
2278	4000→4500	78.32	76.57	75.89	76.23	87.77	1159.32

instance, the accuracy of TransBLS-B drops from 78.32% to 76.04%, and the precision of TransBLS-S drops from 78.32% to 76.04%. Similarly, the same pattern is produced on the LSAFBD.

As we can see from Table 9, for three patch resolutions, with the increase in the parameter scale, the performance of TransBLS is also improved. On the LSAFBD, when the patch resolution is 16×16, from TransBLS-T to TransBLS-H, the recall increases from 61.17% to 66.03%, the F1 score rises from 61.23% to 66.12%, and the AUC is improved from 83.39% to 87.39%. When the patch resolutions are 16×16 and 32×32, from TransBLS-T to TransBLS-H, similar patterns are observed. Similarly, on SCUT-FBP5500, the performance of TransBLS is enhanced when the parameter scale increases.

Incremental learning. TransBLS is a dynamic network, and we can improve its performance with incremental learning algorithms. In this part, experiments are implemented on SCUT-FBP5500, the LSAFBD and CelebA to verify their effectiveness and superiority of the proposed approach.

Firstly, we dynamically expand the enhancement nodes. In the experiments, the numbers of feature nodes and enhancement nodes for TransBLS-L are initialized to 2278 and 2000, respectively. In addition, 500 enhancement nodes are added in each update. The results are listed in Tables 10 and 11. As we can see, the overall performance of both TransBLS-L and the BLS are improved with the addition of more enhancement nodes. Although it takes 1159.32 s to update and retrain TransBLS-L, which is higher than the 696.35 s required by the BLS, the performance of TransBLS-L is substantially better than that of the BLS in terms of the accuracy, precision, recall, F1 score and AUC metrics.

Secondly, we simultaneously expand the feature nodes and enhancement nodes. Analogously, the numbers of feature nodes and enhancement nodes of TransBLS are initialized to 1938 and 2000, respectively. In addition, 170 feature nodes and 500 enhancement nodes are added in each update. The results are listed in Tables 12 and 13. As we can see, the overall performances of both TransBLS-L and the BLS are improved with the addition of more feature nodes and enhancement nodes. It only costs 2078.91 s to update and retrain TransBLS-L, which is better than the 2296.68 s required by the BLS. Simultaneously, the performance of TransBLS-L is substantially better than that of the BLS in terms of all the evaluation metrics.

Finally, incremental algorithms with new added inputs are tested and compared with the BLS and ViT-B/16. On CelebA, the training samples include 94,533 images, and the testing samples include 23,632 images. The testing samples are not changed. On the one hand, suppose that the initial network is trained on the first 44,533 training samples of CelebA. Then, an incremental algorithm is applied to dynamically add 10,000 input patterns each time until all 94,533 training samples are fed. The nodes of TransBLS-B are set as 2278 feature nodes and 2936 enhancement nodes, and the hyperparameters and weights of the transformer encoder remain unchanged. The results obtained after each update are listed in Tables 14 and 15. On the other hand, experiments involving ViT-B/16 with different numbers of training samples are conducted, and the results are listed in Table 16.

At the beginning, the performance of TransBLS-B and the BLS experience a catastrophic drop, and little change is observed in the performance of ViT-B/16. Fortunately, with the addition of input data, the performance is improved, and

Table 11 Results obtained by BLS on SCUT-FBP5500 with enhancement nodes added

Feature nodes	Enhancement nodes	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
2278	2000	66.68	58.21	61.50	59.81	70.10	525.24
2278	2000→2500	66.76	59.30	61.27	60.27	70.34	578.22
2278	2500→3000	66.85	58.22	62.19	60.14	70.46	580.76
2278	3000→3500	66.94	61.03	60.77	60.90	71.07	586.15
2278	3500→4000	66.58	58.82	61.37	60.07	70.29	594.09
2278	4000→4500	66.85	59.25	61.49	60.35	70.49	696.35

Table 12 Results obtained by TransBLS-L on the LSAFBD with feature and enhancement nodes added

Feature nodes	Enhancement nodes	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
1938	2000	65.08	65.00	63.77	64.38	86.24	2026.01
1938→2108	2000→2500	65.18	65.11	63.76	64.43	86.39	2048.60
2108→2278	2500→3000	65.13	65.14	64.01	64.57	86.26	2057.98
2278→2488	3000→3500	65.38	65.29	64.10	64.69	86.35	2066.17
2488→2618	3500→4000	65.63	65.55	64.26	64.90	86.95	2075.20
2618→2788	4000→4500	66.15	66.08	64.85	65.42	87.14	2078.91

Table 13 Results obtained by BLS on the LSAFBD with feature and enhancement nodes added

Feature nodes	Enhancement nodes	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
1938	2000	51.15	52.02	48.37	50.13	64.03	1499.28
1938→2108	2000→2500	52.51	52.37	48.87	50.56	64.54	1562.88
2108→2278	2500→3000	52.66	52.41	48.97	50.63	64.63	1630.97
2278→2488	3000→3500	52.91	52.69	49.55	51.07	65.52	1727.02
2488→2618	3500→4000	52.56	52.54	48.98	50.70	64.76	1763.95
2618→2788	4000→4500	52.76	52.57	49.24	50.85	65.12	2296.68

Table 14 Results obtained by TransBLS-B on CelebA with input patterns added

Input Patterns	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
44533	72.56	77.35	69.83	73.40	77.25	15522.03
44533→54533	72.65	77.85	69.86	73.64	77.26	15607.80
54533→64533	78.05	81.00	76.47	78.67	79.18	15632.92
64533→74533	80.73	81.43	80.52	80.97	79.42	15643.64
74533→84533	81.44	81.28	81.42	81.35	78.31	15667.12
84533→94533	81.70	81.33	80.73	81.43	77.17	15675.20

Table 15 Results obtained by BLS on CelebA with input patterns added

Input Patterns	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
44533	60.11	74.91	50.67	60.45	55.16	7982.08
44533→54533	70.88	75.58	52.89	62.23	67.92	12067.36
54533→64533	72.12	75.71	66.35	70.72	71.69	14265.92
64533→74533	73.66	76.08	72.30	74.14	71.99	15922.56
74533→84533	75.52	76.31	74.37	75.33	71.17	23380.32
84533→94533	76.02	76.94	73.40	75.13	69.67	25318.56

Table 16 Results obtained by ViT-B/16 with different numbers of training samples on CelebA

Training Samples	Accuracy	Precision	Recall	F1 Score	AUC	Time (s)
44533	74.30	74.22	74.39	74.30	78.29	70120.88
54533	74.98	74.61	75.03	74.82	78.85	76767.34
64533	75.19	74.92	75.21	75.06	79.41	97534.82
74533	75.42	75.28	75.50	75.39	81.91	114537.29
84533	75.74	75.59	75.91	75.75	82.33	128934.69
94533	75.41	75.26	75.44	75.35	80.20	137405.62

Table 17 Accuracy comparison with the state-of-the-art methods

Methods	LSAFBD	SCUT-FBP5500	CelebA
BLS+2DPCA [26]	58.98	66.85	74.92
Gan et al. [46]	60.80	75.50	-
ER-BLS [17]	62.13	74.69	79.05
TransBLS-T (ours)	61.27	75.23	81.41
TransBLS-S (ours)	63.03	76.05	81.52
LDCNN [47]	63.50	-	-
TransBLS-B (ours)	63.53	77.32	81.85
Cross Network [48]	64.76	-	-
BeautyNet [49]	64.84	-	-
Gan et al. [50]	65.40	74.80	79.50
2M BeautyNet [51]	65.59	-	-
TransBLS-L (ours)	66.34	78.32	81.87
TransBLS-H(ours)	66.49	78.46	81.95

TransBLS-B finally reaches 81.70% accuracy, 81.33% precision, 81.42% recall, a 81.43% F1 Score and an 79.42% AUC, which are better than those of the BLS and ViT-B/16. Furthermore, as we can see, it takes 25318.56 s to update and retrain the BLS and 137405.62 s to retrain the ViT. Our TransBLS only requires 15675.20 s, and its training efficiency is 1.62 and 8.77 times higher than those of the BLS and ViT-B/16, respectively.

4.8 Comparison with state-of-the-art methods

To further validate the effectiveness of our method, we also compare TransBLS with the other state-of-the-art FBP approaches on the LSAFBD, SCUT-FBP5500 and CelebA. The results are listed in Table 17. The performance of TransBLS-L and TransBLS-H surpasses that of the other approaches on all three databases, which can be combined to achieve a further performance boost. The best results of TransBLS further indicate that transformers exhibit superiority and effectiveness in FBP. They not only solve the problems of insufficiently supervised information and overfitting for FBP, the overfitting and low training efficiency of transformers, and the low accuracy of the BLS, but also effectively achieve improved FBP performance. Thus, transformers can be widely applied in image classification, pattern recognition and object detection tasks.

5 Conclusion

In this paper, we formulate FBP as a special level classification problem guided by supervised information. We propose an adaptive transformer with global and local self-attention, called GLAFormer, to achieve improved FBP performance. Because GLAFormer does not converge and overfitting arises on small-scale data, we fuse GLAFormer with the BLS to form TransBLS, which uses the BLS to speed up the convergence of GLAFormer and reduce overfitting. Numerous results show its superiority over previously developed DNN-based FBP methods and transformer-based FBP methods on several databases and demonstrate its effectiveness and robustness.

In the future, we will continuously study the characteristics of TransBLS for multitask prediction. How to design an adaptive multitask learning-based TransBLS architecture, how to combine local information and other factors that influence facial beauty, and how to use the effective information between multiple related tasks to continue to achieve improved FBP performance will be deeply studied.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61771347, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515010716 and in part by the Basic Research and Applied Basic Research Key Project in General Colleges and Universities of Guangdong Province under Grant 2018KZDXM073.

Declarations

Competing interest All the the authors declare that they have not known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Perrett DI, Lee KJ, Penton-Voak I, Rowland D, Yoshikawa S, Burt DM, Henzi SP, Castles DL, Akamatsu S (1998) Effects of sexual dimorphism on facial attractiveness. Nature 394(6696):884–887
2. Fan YY, Liu S, Li B, Guo Z, Samal A, Wan J, Li SZ (2017) Label distribution-based facial attractiveness computation by deep residual learning. IEEE Trans Multimed 20(8):2196–2208
3. Chen F, Xiao X, Zhang D (2016) Data-driven facial beauty analysis: prediction, retrieval and manipulation. IEEE Trans Affect Comput 9(2):205–216
4. Gan J, Jiang K, Tan H, He G (2020) Facial beauty prediction based on lighted deep convolution neural network with feature extraction strengthened. Chin J Electron 29(2):312–321
5. 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
6. Xu L, Xiang J, Yuan X (2018) Crnet: Classification and regression neural network for facial beauty prediction. In Pacific Rim Conf Multimed 661–671
7. Lin L, Liang L, Jin L, Chen W (2019) Attribute-Aware Convolutional Neural Networks for Facial Beauty Prediction. The 28th International Joint Conference on Artificial Intelligence (IJCAI) 847–853
8. Lebedeva I, Guo Y, Ying F (2021) MEBeauty: a multi-ethnic facial beauty dataset in-the-wild. Neural Comput Appl 1–15

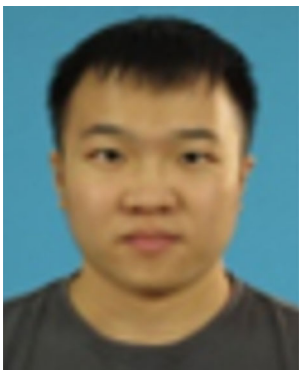
9. Bougourzi F, Dornaika F, Barrena N, Distante C, Taleb-Ahmed A (2022) CNN based facial aesthetics analysis through dynamic robust losses and ensemble regression. Appl Intell 1–18
10. Lin L, Liang L, Jin L (2019) Regression guided by relative ranking using convolutional neural network (R3CNN) for facial beauty prediction. IEEE Trans Affect Comput 1–14
11. 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
12. Cao K, Choi KN, Jung H, Duan L (2020) Deep learning for facial beauty prediction. Information 11(8):391
13. Bougourzi F, Dornaika F, Taleb-Ahmed A (2022) Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. Knowl-Based Syst 242:108246
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 5998–6008
15. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu Ch, Xu Y, Yang Z, Zhang Y, Tao D (2023) A survey on vision transformer. IEEE Trans Pattern Analy Mach Intell 45(1):87–110
16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. The 9th International Conference on Learning Representations (ICLR) 1–22
17. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision 10012–10022
18. Touvron H., Cord M, Douze M, Massa F, Sablayrolles A, Jgou H (2021) Training data-efficient image transformers & distillation through attention. In: International conference on machine learning 10347–10357
19. Heo YJ, Yeo WH, Kim BG (2022) Deepfake detection algorithm based on improved vision transformer. Appl Intell 1–16
20. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H (2022) Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 1–53
21. Al-Refai R, Nandakumar K (2023) A Unified Model for Face Matching and Presentation Attack Detection Using an Ensemble of Vision Transformer Features. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 662–671
22. Shao Z, Li F, Zhou Y, Chen H, Zhu H, Yao R (2023) Identity-invariant representation and transformer-style relation for micro-expression recognition. Appl Intell 1–12
23. Chen CP, Liu Z (2017) Broad learning system: An effective and efficient incremental learning system without the need for deep architecture. IEEE Trans Neural Netw Learn Syst 29(1):10–24
24. Li Y, Zhang T, Chen CP (2021) Enhanced Broad Siamese Network for Facial Emotion Recognition in Human?Robot Interaction. IEEE Trans Artif Intell 2(5):413–423
25. Li P, Sheng B, Chen CP (2021) Face sketch synthesis using regularized broad learning system. IEEE Trans Neural Netw Learn Syst 5346–5360
26. Zhai Y, Yu C, Qin C, Zhou W, Ke Q, Gan J, Labati RD, Piuri V, Scotti F (2020) Facial beauty prediction via local feature fusion and broad learning system. IEEE Access 8:218444–218457
27. Gan J, Xie X, Zhai Y, He G, Mai C, Luo H (2023) Facial beauty prediction fusing transfer learning and broad learning system. Soft Comput 27:13391–13404
28. Xie D, Liang L, Jin L, Xu J, Li M (2015) SCUT-FBP: A benchmark dataset for facial beauty perception. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics 1821–1826
29. Gan J, Zhai Y, Wang B (2017) Unconstrained Facial Beauty Prediction Based on Multi-scale K-Means. Chin J Electron 26(3):548–556
30. Liang L, Lin L, Jin L, Xie D, Li M (2018) SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In: 2018 24th International conference on pattern recognition (ICPR) 1598–1603
31. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision 3730–3738
32. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition 770–778
33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2818–2826
34. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, 6105–6114
35. Zhang L, Zhang D, Sun MM, Chen FM (2017) Facial beauty analysis based on geometric feature: Toward attractiveness assessment application. Expert Syst Appl 82:252–265
36. Yu Z, Qin Y, Li X, Zhao C, Lei Z, Zhao G (2023) Deep learning for face anti-spoofing: A survey. IEEE Trans Pattern Anal Mach Intell 45(5):5609–5631
37. Chen S, Chongjian GE, Tong Z, Wang J, Song Y, Wang J, Luo P (2022) AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. In: Advances in Neural Information Processing Systems, 1-21
38. Vaswani A, Ramachandran P, Srinivas A, Parmar N, Hechtman B, Shlens J (2021) Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 12894-12904
39. Gong X, Zhang T, Chen CP, Liu Z (2021) Research review for broad learning system: Algorithms, theory, and applications. IEEE Transactions on Cybernetics 8922–8950
40. Yang, F. (2018). A CNN-based broad learning system. In: 2018 IEEE 4th International Conference on Computer and Communications (ICCC) 2105-2109
41. Huang H, Liu Z, Chen CL, Zhang Y (2022) Hyperspectral image classification via active learning and broad learning system. Appl Intell 1–12
42. Su L, Xiong L, Yang J (2023) Multi-Attn BLS: Multi-head attention mechanism with broad learning system for chaotic time series prediction. Appl Soft Comput 132:109831
43. Deng L, Xiao M (2023) A New Automatic Hyperparameter Recommendation Approach Under Low-Rank Tensor Completion Framework. IEEE Trans Pattern Anal Mach Intell 45(4):4038–4050
44. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic MX, Dosovitskiy A (2021) Mlp-mixer: An all-mlp architecture for vision. Adv Neural Inf Process Syst 34:24261–24272
45. Tharwat A (2020) Classification assessment methods. Appl Comput Inf 17(1):168–192
46. Gan J, Wu B, Zhai Y, He G, Mai C, Bai Z (2022) Self-corrected noise labeling for face beauty prediction. Chinese Journal of Graph Graph 27(08):2487–2495
47. Gan J, Jiang K, Tan H, He G (2020) Facial beauty prediction based on lighted deep convolution neural network with feature extraction strengthened. Chin J Electron 29(2):312–321
48. Gan J, Wu B, Zou Q, Zheng Z, Mai C, Zhai Y, He G, Bai Z (2022) Application Research for Fusion Model of Pseudolabel and Cross Network. Comput Intell Neurosci 99866:1–10
49. Zhai Y, Cao H, Deng W, Gan J, Piuri V, Zeng J (2019) BeautyNet: Joint multiscale CNN and transfer learning method for unconstrained facial beauty prediction. Comput Intell Neurosci 1910624:1–14

50. Gan J, Wu B, Zou Q, Zheng Z, Mai C, Zhai Y, He G (2022) Two-input dual-task attention network incorporating noisy label correction mechanism for face beauty prediction. Signal Process 38(10):2124–2133
51. Gan J, Xiang L, Zhai Y, Mai C, He G, Zeng J, Bai Z, Labati RD, Piuri V, Scotti F (2020) 2M BeautyNet: Facial beauty prediction based on multi-task transfer learning. IEEE Access 8:20245–20256

Junying Gan was born in Ji'an, Jiangxi, China, she received the B.S., M.S., and Ph.D. degrees in electrical information engineering from BUAA, in 1987, 1992, and 2003, respectively. In 1992, she joined Wuyi University, Guangdong, China, as a Full Professor. She has published more than 50 journal articles. Her research interests include biometric recognition and image processing.

Xiaoshan Xie is a M.D. student at the department of intelligent manufacturing, Wuyi University. He obtained his bachelor's degree in engineering from the faculty of intelligent manufacturing at the Wuyi University. His research interests are artificial intelligence in deep learning, computer vision, biometric recognition, and medical image analysis.

Guohui He was born in Pingxiang, Jiangxi, China. He received the B.S. degree in computer engineering from the Shenyang University of Aeronautics and Astronautics, in 1983, and the M.S. degree from BUAA, in 1991. He joined Wuyi University, as a Full Professor in 1994. His research interest includes multimedia information systems.



Heng Luo was born in Shaoyang, Hunan, China. He is a M.D. student at the department of intelligent manufacturing, Wuyi University. His research interest includes biometric identification.