PAPER

A Ranking Information Based Network for Facial Beauty Prediction

Haochen LYU[†], Jianjun LI[†,††a)], Yin YE[†††], *Nonmembers*, *and* Chin-Chen CHANG[††††], *Member*

SUMMARY The purpose of Facial Beauty Prediction (FBP) is to automatically assess facial attractiveness based on human aesthetics. Most neural network-based prediction methods do not consider the ranking information in the task. For scoring tasks like facial beauty prediction, there is abundant ranking information both between images and within images. Reasonable utilization of these information during training can greatly improve the performance of the model. In this paper, we propose a novel end-to-end Convolutional Neural Network (CNN) model based on ranking information of images, incorporating a Rank Module and an Adaptive Weight Module. We also design pairwise ranking loss functions to fully leverage the ranking information of images. Considering training efficiency and model inference capability, we choose ResNet-50 as the backbone network. We conduct experiments on the SCUT-FBP5500 dataset and the results show that our model achieves a new state-of-the-art performance. Furthermore, ablation experiments show that our approach greatly contributes to improving the model performance. Finally, the Rank Module with the corresponding ranking loss is plug-and-play and can be extended to any CNN model and any task with ranking information. Code is available at https://github.com/nehcoah/Rank-Info-Net.

key words: *ranking information, facial beauty prediction, computer vision, deep learning*

1. Introduction

Facial attractiveness plays a significant role in our daily lives and has attracted much attention from scholars who have conducted extensive research on the subject [1], [2]. In recent years, people have become increasingly concerned about their facial beauty. It is worth noting that having a more beautiful appearance can give individuals an advantage in various aspects of life, such as public speaking, presentations, job interviews, and career opportunities. With the advancement of computer technology, there has been a shift towards utilizing computers to assess people's appearance, giving rise to Facial Beauty Prediction (FBP). FBP aims to automatically evaluate facial attractiveness based on human aesthetics standards. It can also be applied in conjunction with practical tasks such as face beautification [3], [4], automatic face make-up [5], and plastic surgery [6].

In recent years, there have been significant breakthroughs in FBP tasks. These methods can be broadly categorized into two types: handcrafted feature-based and deep learning-based approaches. In the earlier studies, researchers evaluated facial attractiveness based on heuristics rules, such as facial landmarks, facial texture representations, symmetry and golden ratio proportions [7]. However, these methods had limitations and lacked fine-grained extraction and application of facial features. With the flourishing of deep learning, it has demonstrated unique capabilities in various domains, including facial beauty prediction. Various types of Convolutional Neural Networks (CNN), such as VGG [8], ResNet [9], MobileNet [10]–[12], EfficientNet [13], etc., have been applied to FBP tasks. The powerful feature extraction ability of these networks allows for a more comprehensive measurement of facial beauty. Specifically, facial beauty prediction methods can roughly be divided into regression methods and classification methods. Regardless of whether it is classification or regression, most researchers are dedicated to optimizing the feature extraction methods of CNN networks or exploring better ways to utilize the extracted facial features to achieve higher accuracy. Additionally, most facial beauty prediction datasets, such as the SCUT-FBP5500 dataset [14], are obtained by collecting ratings from a large number of volunteers, which serve as the corresponding ground truth. Some researchers utilize the label distribution from all volunteers' ratings for each image and employ Label Distribution Learning (LDL) to optimize the performance of the model [15], [16].

However, for tasks such as facial beuaty prediction and most scoring tasks, both classification and regression methods fail to effectively or even utilize the potential ranking information of images. We have also noticed that some researchers have applied pairwise ranking methods to facial beuaty prediction tasks [17], [18], but these methods require two backbone networks in the training process in order to introduce ranking loss, which increases the model parameters and extends the training time. Therefore, we propose a new end-to-end model based on ranking information, which only requires one backbone network in both training and testing stages. The overall framework is shown in Fig. 1. We add a Rank Module and an Adaptive Weight Module to the traditional convolutional neural networks and design a methods to extract ranking information and the corresponding pairwise ranking loss functions to fully utilize the underlying ranking information in the images. The outputs of the classifier

Manuscript received October 2, 2023.
Manuscript revised November 15, 2023.
Manuscript publicized January 26, 2024.

[†] The authors are with School of Computer Science and Technology, Hangzhou Dianzi University, China.

[††] The author is with School of Information Science and Technology, Hangzhou Normal University, China.

[†††] The author is with CEC Huada Electronic Design Co., Ltd., China.

[††††] The author is with Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan.

a) E-mail: lijjcan@gmail.com (Corresponding author)

DOI: 10.1587/transinf.2023EDP7203

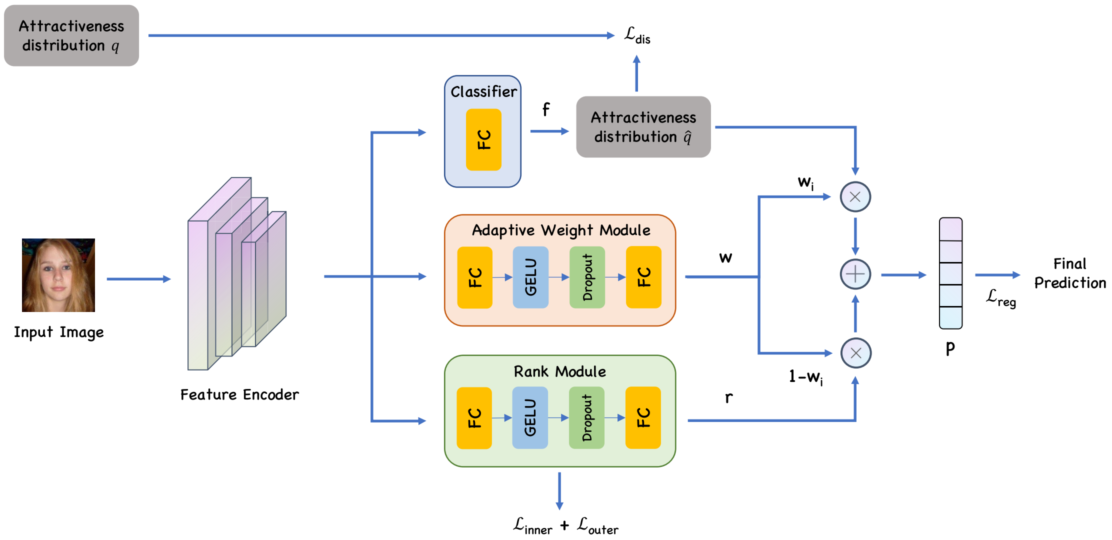

Fig. 1 The architecture of our proposed model (some graphic templates come from the GitHub repository ml-vision [19]), where an adaptive weight module and a rank module are added after the feature encoder. The classifier is a fully connected layer. The adaptive weight module and the rank module consist of fully connected layers, a GELU activation layer and a dropout layer. The outputs of the classifier, adaptive weight module, and rank module are denoted as f, w, and r respectively. The attractiveness distribution generated from the original data and from the classifier are respectively labeled as q and $\hat{q}$. The final output of the network is denoted as p.

and the Rank Module are integrated through the Adaptive Weight Module to obtain predicted scores. Considering the inference capability and training efficiency required for extracting ranking information, we finally choose ResNet-50 as the backbone network for our experiments in this paper. The Rank Module is responsible for learning the ranking information of the images, while the classifier learns the label distribution, and the regression loss is used to constrain the overall predicted scores. We test our method on the SCUT-FBP5500 dataset [14] and obtain exciting results. The experimental results show that our model achieves a new state-of-the-art performance. Ablation experiments also demonstrate the crucial effect of the proposed Rank Module and the pairwise ranking loss functions on improving accuracy of the model. Furthermore, our proposed Rank Module is plug-and-play and can be extended to any convolutional neural networks, while the methods of extracting ranking information and the pairwise ranking loss functions can be applied to most tasks with ranking information, including various scoring tasks and age estimation tasks, etc.

The main contributions of this paper can be summarized into three points:

- We propose a novel end-to-end network based on ranking information. This model extends the traditional Convolutional Neural Networks (CNN) by incorporating a Rank Module and an Adaptive Weight Module, along with corresponding pairwise ranking loss functions. The model demonstrates transferability, as the Rank Module and Adaptive Weight Module can be seamlessly integrated into any CNN-based network model. Moreover, the ranking information-based approach can be applied to any task that involves ranking information.
- Compared to other methods based on ranking information, our proposed model only requires one backbone network. Under the same backbone network conditions, our approach addresses the issue of excessive model parameters during the training phase, saving training time while also achieving better performance.
- We conducts various experiments on the SCUT-FBP5500 dataset and the final results achieves a new state-of-the-art performance. Additionally, the results from ablation experiments demonstrate that our method significantly improves the performance of the network.

This paper is structured in a manner that includes an overview of the related work in Sect. 2, followed by a detailed description of our proposed method in Sect. 3. The experimental results are presented in Sect. 4 and Sect. 5 provides a conclusion of our study. Finally, some development plans, funds and institutions are acknowledged at the end.

2. Related Work

2.1 Pairwise Method of Ranking

Learning to Rank is widely applied for document ranking, such as recommendation algorithms, and can be roughly divided into three methods: Pointwise, Pairwise, and Listwise. Many researchers have summarized these methods [20], [21]. In the Pairwise approach, numerous applications have emerged, including Ranking SVM [22], RankBoost [23], RankNet [24], LambdaRank [25], and so on.

With the development of deep learning, some researchers have applied these ranking methods, especially the pairwise methods, to various fields. The introduction of Siamese networks [26] has paved the way for the application of pairwise methods in computer vision. Its dual-branch network structure provides conditions for obtaining pair information. Therefore, most researchers choose similar network structures when migrating pairwise methods to other tasks to facilitate pair information extraction. Gattupalli et al. [27] carefully selected images from the AVA dataset and constructed a new dataset of image pairs with relative labels. They also proposed a neural network-based method to train the ranking information. Lin et al. [18] also proposed a general convolutional neural network architecture based on the Siamese network's framework and applied the pairwise method to the facial beauty prediction task.

2.2 Facial Beauty Prediction

Before the popularity of deep learning, researchers primarily used traditional methods for facial beauty prediction, including studying facial symmetry, texture features of the face and the golden ratio proportion. However, these methods had significant limitations as they lacked fine-grained feature extraction of facial characteristics and lacked a relatively systematic feature extraction approach, leading to poor performance. With the development of deep learning, a series of evaluation methods based on Convolutional Neural Network (CNN) have emerged.

Gray et al. [28] initially proposed a feature extraction method similar to CNN, eliminating the need for manual annotation of facial features for prediction. After the introduction of the VGG network [8], Xu et al. [29] applied it to facial beauty prediction tasks. Inspired by psychology, Xu et al. [30] introduced a hierarchical model called PI-CNN, using a cascaded fine-tuning approach to optimize predictors. Liang et al. [31] combined deep convolutional networks based on scattering transform with facial texture and shape features, proposing the RegionScarNet model. To address the issue of fixed-parameter convolutional kernels failing to fully utilize facial attributes, Lin et al. [32] introduced AaCNN, which can adaptively adjust the kernel size of the network. Xu et al. [33] proposed CRNet, which can simultaneously perform classification and regression tasks. Xu et al. [34] introduced a hierarchical multi-task network capable of simultaneously identifying the gender, race, and facial attractiveness of face images. Similarly, Xu [35] developed the multi-task model, which can automatically recognize facial attractiveness scores and gender. Lin et al. [17] presented R2-ResNeXt, which shared the weights of two ResNeXt networks [36] and used ranking loss to optimize network performance during training. Subsequently, they proposed a general CNN architecture called R^{3}CNN [18], which considers facial beauty prediction as a ranking-guided regression problem, using two CNNs to simultaneously perform ranking and regression tasks. Fan et al. [15] reshaped facial attractiveness as a label distribution learning problem and proposed an end-to-end framework, incorporating low-level geometric features for feature-level fusion. Later, Liu et al. [16] introduced a lightweight end-to-end FBP method, which achieved promising results by training with an improved label distribution learning approach based on [15]. Wei et al. [37] proposed a method that utilizes facial landmarks to compute facial features with a low computational cost. Saeed et al. [38] proposed FIAC-Net, which is a light deep convolutional neural network for facial images attractiveness assessment. Later, they [39] integrated three regression-loss functions to capitalize on the unique traits of each loss function in the facial beauty prediction. Bougourzi et al. [40] proposed an architecture with two backbones (2B-IncRex) and introduced a parabolic dynamic law to control the behavior of the robust loss parameters during training. Yang et al. [41] aimed to train a model for assessing facial beauty using transfer learning while also using the fine-grained image model to separate similar images by first learning features.

3. Method

In this section, we will provide a detailed description of our proposed method. To make full use of the ranking information of images, this paper extends the pairwise method to the Facial Beauty Prediction (FBP) task. In tasks where there is a sequential ranking relationship between samples, it is evident that the reasonable utilization of these ranking information during the training process will improve the performance of the model.

3.1 Network Architecture

To better integrate the pairwise method into FBP tasks, we optimize the traditional neural network architecture. As shown in Fig. 1, we take ResNet50 as backbone network and add a Rank Module and an Adaptive Weight Module. Specifically, we utilize the Adaptive Weight Module to adjust the contributions of the classifier and the Rank Module, and combine them with weighted summation to obtain the final result. The classifier consists of a single fully connected layer, while the Adaptive Weight Module and the Rank Module are composed of fully connected layers, a GELU activation layer, and a Dropout layer. We apply the pairwise method to the Rank Module. Additionally, we draw inspiration from [16] and apply the label distribution learning to the classifier. Finally, we constrain the integrated output of the Adaptive Weight Module using regression methods.

3.2 Label Distribution Learning

In most datasets for facial beauty prediction, multiple individuals rate the same images and the ground truth score is determined by taking the average of their ratings. In the SCUT-FBP5500 dataset, all images were rated by 60 volunteers on a scale of 1 to 5. This dataset is also used in the experimental part of this paper and detailed information about the dataset will be provided in Sect. 4.1. To make these ratings informative for facial beauty prediction, we follow [16] and apply label distribution learning. At the image level, we can calculate the mean μ and variance σ of all volunteer ratings for each image. Using these statistics as reference, we model the label distribution corresponding to each image using a Gaussian distribution. Considering network inference capability and training efficiency, we choose a sampling interval of $\Delta l = 0.05$. This means we divide the range $[1, 5]$ into 80 equal intervals and perform 80 sampling iterations within the range. Suppose the current sampling interval is $I_j = [s_j, s_j + \Delta l]$; then, the corresponding probability q_j for the current interval can be calculated using the probability distribution function $F(x|\mu,\sigma)$ of the Gaussian distribution.

$$q_j = F(s_j + \Delta l|\mu,\sigma) - F(s_j|\mu,\sigma) \tag{1}$$

We combine all the sampled values to obtain the label distribution, denoted as q, and normalize q using L1 normalization. Additionally, we apply a softmax operation to the output results of the classifier and denote the resulting distribution as $\hat{q}$. The label distribution loss $\mathcal{L}_{dis}$ can be calculated using the Euclidean distance.

$$\mathcal{L}_{dis} = \frac{1}{n}\sum_{i=1}^{n} \|\hat{q}^{(i)} - q^{(i)}\|_2 \tag{2}$$

where n represents the number of samples in a batch.

3.3 Ranking Loss within Images

We denote the output of the Rank Module as $r \in \{r_1, r_2, \cdots, r_c\}$, where c represents the total number of categories. For each image, ideally, the predicted probability distribution of model should resemble a Gaussian distribution, with the highest probability value located at the ground truth label and decreasing towards both sides, forming an ordered sequence. Empirically, when the probability distribution obtained by the network exhibits this ideal pattern, more accurate results can be achieved. Therefore, we utilize the ranking information to optimize the predicted probability distribution into the ideal state.

We sample the feature distribution and use the ranking method to optimize the network prediction. Specifically, we start sampling from the ground truth position in r, moving towards both sides. The sampled data is denoted as (r_i, m_i), where i refers to the i-th class, and m_i is the given sequential tag during the sampling process. For example, we denote the sampled data set as $\mathcal{R} \in \{(r_\ell, 0), (r_{\ell-1}, 1), (r_{\ell+1}, 1), (r_{\ell-2}, 2), \cdots\}$, where ℓ corresponds to the ground truth label of the current image. In the sequential tags, smaller values indicate positions closer to the beginning in the sequence. During the sampling process, we specify that the sequential tags of the sampled data on both sides of the ground truth label ℓ are the same if they have the same interval, which means they hold the same positions in the overall sequence.

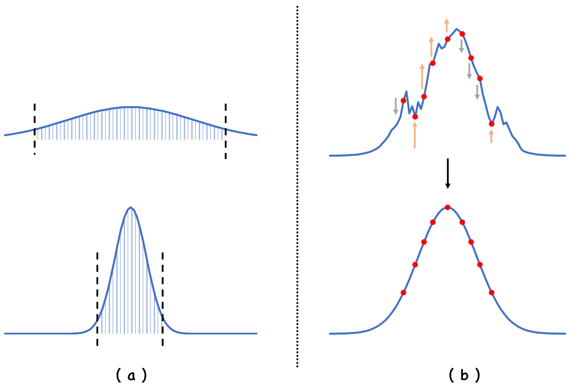

Fig. 2 The sampling range and working mechanism of $\mathcal{L}_{inner}$. Gaussian distribution is used as an example. Figure (a) shows the sampling range of different feature distributions under the influence of thresholds. Figure (b) demonstrates the working mechanism of $\mathcal{L}_{inner}$.

Algorithm 1 The Sampling Process

Input: Rank Module output $r \in \{r_1, r_2, \cdots, r_c\}$, thresholds t, label ℓ, number of classes c

Output: The set of samples $\mathcal{R}$

```
Let $\hat{r}$ = softmax($r$), cur = $\hat{r}_l$, j = 1
add sample ($r_\ell$, 0) to $\mathcal{R}$
while cur ≤ t do
    if $\ell$ - j ≥ 0 then
        cur += $\hat{r}_{\ell-j}$
        add sample ($r_{\ell-j}$, j) to $\mathcal{R}$
    end if
    if $\ell$ + j <c then
        cur += $\hat{r}_{\ell+j}$
        add sample ($r_{\ell+j}$, j) to $\mathcal{R}$
    end if
    j += 1
end while
return The set of samples $\mathcal{R}$
```

Additionally, due to the different feature distributions presented by different images, we set a threshold in sampling to flexibly adjust the sampling range. The sampling threshold is denoted as t, and the pseudocode for the sampling algorithm is shown in Algorithm 1, which aims to generate reasonable samples for subsequent loss calculation. As different images have different feature distributions after feature extraction, thanks to the existence of the sampling threshold, the sampling range of the above process will be located in a reasonable range of each image, as the shaded area shown in Fig. 2 (a), ensuring the utilization of as complete feature information as possible for different images. Note that we will perform sampling on all the images.

After obtaining the all the samples, we will select two samples with different sequential tags to form a sample pair for further processing each time. We donate these two samples as $a = (r_a, m_a)$ and $b = (r_b, m_b)$, and define the score $S_{a,b}$ of the sample pair as

$$S_{a,b} = \frac{\exp(r_a - r_b)}{1 + \exp(r_a - r_b)} \tag{3}$$

The label for the pair of samples, denoted as $y_{a,b}$, is defined as follows: if $m_a < m_b$, then $y_{a,b} = 1$; otherwise,

$y_{a,b} = 0$. Finally, the loss function of the sample pair is referred to as

$$\begin{aligned}\mathcal{L}_{inner}(a,b,y_{a,b}) = &- y_{a,b}\log(S_{a,b}) \\ &- (1-y_{a,b})\log(1-S_{a,b})\end{aligned} \tag{4}$$

We will consider all samples with different sequential tags when calculating the loss. Figure 2 (b) illustrates the working mechanism of $\mathcal{L}_{inner}$. Ideally, $\mathcal{L}_{inner}$ can arrange the sampled samples in the specified order.

3.4 Ranking Loss between Images

For different images in a batch, their ground truth labels are not completely identical. There is also a ranking order relationship between different images based on ground truth label. In order to calculate the ranking loss between images more conveniently, we simplify the feature information of images in the same batch to the expectation $\mathbb{E} \in \{\mathbb{E}_1, \mathbb{E}_2, \cdots, \mathbb{E}_n\}$, where n is the total number of images in the current batch. The expectation $\mathbb{E}_k$ for each image in the current batch is defined as

$$\mathbb{E}_k = \sum_{i=1}^{c} \hat{r}_i * i, \quad k = 1, 2, \cdots, n \tag{5}$$

where $\hat{r}$ represents the result of applying softmax to the output of the Rank Module r, and c represents the number of classes. Similar to the approach mentioned in Sect. 3.3, we can denote the expectations and labels of all the images in a batch as a set like $\{(\mathbb{E}_1, \ell_1), (\mathbb{E}_2, \ell_2), \cdots, (\mathbb{E}_n, \ell_n)\}$, where $(\mathbb{E}_j, \ell_j)$ is the j-th image in the batch, with the expectation $\mathbb{E}_j$ and the ground truth label ℓ_j. Here, we define that images with bigger ground truth scores indicates positions closer to the beginning in the sequence. Similarly, we select two samples with different labels for processing each time. We donate these two samples as $u = (\mathbb{E}_u, \ell_u)$ and $v = (\mathbb{E}_v, \ell_v)$, and the score between the two samples, named $S_{u,v}$, is calculated as

$$S_{u,v} = \frac{\exp(\mathbb{E}_u - \mathbb{E}_v)}{1 + \exp(\mathbb{E}_u - \mathbb{E}_v)} \tag{6}$$

The label for the pair of samples, denoted as $y_{u,v}$, is defined as follows: if $\ell_u > \ell_v$, then $y_{u,v} = 1$; otherwise, $y_{u,v} = 0$. Finally, the loss function between these two samples is referred to as

$$\begin{aligned}\mathcal{L}_{outer}(u,v,y_{u,v}) = &- y_{u,v}\log(S_{u,v}) \\ &- (1-y_{u,v})\log(1-S_{u,v})\end{aligned} \tag{7}$$

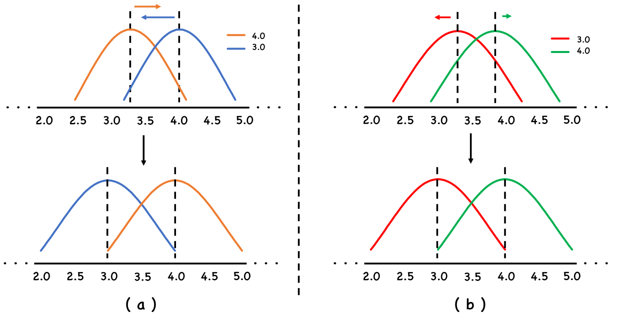

Fig. 3 The working mechanisms of $\mathcal{L}_{outer}$ in different scenarios. Gaussian distribution is used as an example. Figure (a) shows an incorrect relative position between the two distributions, while (b) depicts the correct relative position but with a too close distance.

Here we will select samples with different labels for processing. When there is a deviation in the relative positioning of expectations between images, ideally, $\mathcal{L}_{outer}$ can move the overall probability distribution to the correct position. As shown in Fig. 3, we divide the relative positional deviations of probability distribution into two categories. First, there is a relative positional error, as shown in Fig. 3 (a), where the blue curve and orange curve correspond to the probability distributions for images with labels 3.0 and 4.0, respectively. The upper graph of Fig. 3 (a) shows that their relative positions are incorrect; theoretically, the blue curve should be on the left side of the orange curve. In this case, $\mathcal{L}_{outer}$ will pull the blue curve to the left and the orange curve to the right, placing them in the correct positions. The second category is the relative positions are correct but the intervals are too close, as shown in Fig. 3 (b), where the red curve and green curve represent the probability distributions for images with labels 3.0 and 4.0, respectively. From the upper graph of Fig. 3 (b), we can see that although their relative positions are correct, they are too close to each other. In this case, $\mathcal{L}_{outer}$ will pull both curves apart, to some extent, making them move away from each other.

3.5 Adaptive Weight Module

Furthermore, we introduce an Adaptive Weight Module to integrate the outputs of the classifier and the Rank Module. We denote the output of the classifier as $f \in \{f_1, f_2, \cdots, f_c\}$, the output of the Adaptive Weight Module as $w \in \{w_1, w_2, \cdots, w_c\}$, and the output of the Rank Module as $r \in \{r_1, r_2, \cdots, r_c\}$. The Adaptive Weight Module combines the outputs of the classifier and the Rank Module to generate a new output, denoted as $p \in \{p_1, p_2, \ldots, p_c\}$, which serves as the final output of the network. Specifically, each p_i can be represented as follows:

$$p_i = w_i * f_i + (1 - w_i) * r_i, \quad i = 1, 2, \cdots, c \tag{8}$$

where c represents the number of classes. We define the predicted facial beauty score x as

$$x = \sum_{i=1}^{c} \hat{p}_i * i \tag{9}$$

where $\hat{p}$ is the result of applying softmax to p. For the final prediction of the network, we choose the Smooth L1 loss to minimize the discrepancy between the ground truth and the predicted scores by the network,

$$\mathcal{L}_{reg}(x,y) = \begin{cases} 0.5(x-y)^2, & if \quad |x-y| < 1 \\ |x-y| - 0.5, & otherwise \end{cases} \tag{10}$$

Table 1 Comparison with state-of-the-art on SCUT-FBP5500 dataset with five-folds cross validation. Best results are marked in bold.

Method	Backbone	MAE ↓	RMSE ↓	PC ↑
AaNet [32]	ResNet-18	0.2236	0.2954	0.9055
Co-attention learning [42]	MobileNetV2x2	0.2020	0.2660	0.9260
MT-ResNet [35]	ResNet-50	0.2459	0.3208	0.8905
R^{3}CNN [18]	ResNeXt-50	0.2120	0.2800	0.9142
Dual Label Distribution [16]	MobileNetV2	0.1964	0.2585	0.9276
FIAC-Net + Loss Ensembles [39]	FIAC-Net	0.2028	0.2614	**0.9305**
Dynamic ER-CNN [40]	ResNeXt-50 + Inception-v3	0.1998	0.2633	0.9262
Ours	ResNet-50	**0.1913**	**0.2551**	0.9288

where y is the ground truth label and $|\cdot|$ represents absolute value. The final loss function can be expressed as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{inner} + \lambda_2\mathcal{L}_{outer} + \lambda_3\mathcal{L}_{reg} + \lambda_4\mathcal{L}_{dis} \qquad (11)$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters that balance the four losses.

4. Experiments

In this section, we design various experiments to validate our proposed method. We provide a detailed analysis of the experimental setup and compare the results with state-of-the-art works. We also conducted ablation experiments to demonstrate the effectiveness of the Rank Module and the pairwise ranking loss functions.

4.1 Dataset and Evaluation Metrics

The SCUT-FBP5500 dataset consists of 5500 facial images. A total of 60 volunteers were asked to rate each photo on a scale ranging from 1 to 5. The ground truth label for each image is obtained by averaging the ratings from these 60 volunteers. In addition to the ground truth label for each image, the dataset also provides detailed rating scores from the 60 volunteers for each image.

In terms of evaluation criteria, we use Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson Correlation coefficient (PC) to measure the performance of the model. A more outstanding model will exhibit lower MAE and RMSE values while having a higher PC value. The specific formulas for calculation are shown in Eq. (12).

$$\begin{aligned}
MAE &= \frac{1}{N}\sum_{i=1}^{N}|x^{(i)} - y^{(i)}| \\
RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - y^{(i)})^2} \qquad (12) \\
PC &= \frac{\sum_{i=1}^{N}(y^{(i)} - \bar{y})(x^{(i)} - \bar{x})}{\sqrt{\sum_{i=1}^{N}(y^{(i)} - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(x^{(i)} - \bar{x})^2}}
\end{aligned}$$

where N represents the number of images in the test set, x is the predicted score by the network, y is the ground truth label, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$, and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}$.

Table 2 Details of five-folds cross validation on SCUT-FBP5500 dataset.

Split	MAE ↓	RMSE ↓	PC ↑
1	0.1919	0.2569	0.9266
2	0.1948	0.2626	0.9238
3	0.1925	0.2593	0.9278
4	0.1885	0.2475	0.9338
5	0.1887	0.2491	0.9321
Avg	0.1913	0.2551	0.9288

4.2 Implementation Details

In this paper, we do not perform extensive operations on the input images. For each input image, we first resize it to 256×256. During the training phase, the images are randomly cropped to 224×224 and subjected to random horizontal flipping with a probability of 0.5. During the testing phase, the images are center-cropped to 224×224. We use ResNet-50 as the backbone network, initialized with ImageNet-pretrained weights, and modify the output layer's channel to 80. We use the SGD optimizer with a momentum of 0.9 and weight decay of 0.0005. The initial learning rate is set to 0.001 and decreased by a factor of 0.3 every 15 epochs. Each model is trained for 90 epochs with a batch size of 64. Regarding the threshold t mentioned in Sect. 3.3, it is set to 0.95 for the first 15 epochs and then increased to 0.98 for the remaining epochs. Additionally, we set the hyperparameters $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ in Eq. (11). All experiments are conducted on an NVIDIA Titan GPU. We perform five-fold cross-validation and the average results are reported.

4.3 Comparison with the State of the Art

We compared our approach with recent state-of-the-art works, as shown in Table 1. Additionally, detailed information for each split of the five-fold cross-validation is provided in Table 2. The data in Table 1 demonstrates that our method performs significantly better than other approaches.

On the SCUT-FBP5500 dataset, our method achieves state-of-the-art performance on MAE and RMSE. Specifically, compared to previous methods that utilize pairwise ranking method, such as R^{3}CNN [18], our model shows significant improvements. Furthermore, compared to methods using the same backbone, our results demonstrate a substantial advancements. In comparison to the recently proposed method Dual Label Distribution [16] and Dynamic ER-CNN [40], our model still outperforms it with superior

Table 3 Ablation study on different combinations of methods. Best results are marked in bold.

Method	MAE ↓	RMSE ↓	PC ↑
B + LD	0.1940	0.2602	0.9258
B + LD + RI	0.1932	0.2591	0.9262
B + LD + RO	0.1924	0.2576	0.9275
B + LD + RI + RO	**0.1913**	**0.2551**	**0.9288**

Table 4 Ablation study on different setting of hyperparameters. Best results are marked in bold.

Hyperparameters	MAE ↓	RMSE ↓	PC ↑
$\lambda_1 = 1, \lambda_2 = 0$	0.1932	0.2591	0.9262
$\lambda_1 = 0, \lambda_2 = 1$	0.1924	0.2576	0.9275
$\lambda_1 = 2, \lambda_2 = 1$	0.1921	0.2580	0.9272
$\lambda_1 = 1, \lambda_2 = 2$	0.1919	0.2569	0.9278
$\lambda_1 = 1, \lambda_2 = 1$	**0.1913**	**0.2551**	**0.9288**

performance. Our proposed method slightly lags behind FIAC-Net with loss ensembles [39] in terms of PC metrics. It is worth noting that FIAC-Net is a network specifically designed for assessing facial attractiveness. However, overall, the results achieved by our method are inspiring and reach the state-of-the-art level.

4.4 Ablation Study

4.4.1 Different Combinations of Methods

In order to clarify the improvement of each method on the network performance, we conduct experiments by combining different methods with each other, and the results are shown in Table 3. We use B to represent training the network using $\mathcal{L}_{reg}$, LD to represent the label distribution, which is $\mathcal{L}_{dis}$, RI to represent $\mathcal{L}_{inner}$, and RO to represent $\mathcal{L}_{outer}$. The hyperparameters λ_i in each ablation experiment are set to 1.

From Table 3, we can observe that even without using ranking information during the training phase, we have already obtained a remarkable model. This is because, in order to comprehensively utilize ranking information, we change the number of fully connected channels of the network to 80, enabling the network to output more refined prediction results. With the introduction of the Ranking Module and the pairwise ranking loss, the accuracy of the network has been further improved. Both $\mathcal{L}_{inner}$ and $\mathcal{L}_{outer}$ have a certain enhancing effect on the network, and the best result is obtained by combining the two methods. Therefore, we can conclude that the Ranking Module with pairwise ranking loss plays a significant role in facial beauty prediction tasks. It can effectively extract ranking information from image features and optimize them accordingly.

4.4.2 Different Hyperparameters

In order to clarify the contributions of $\mathcal{L}_{inner}$ and $\mathcal{L}_{outer}$ to the improvement of model performance respectively, we conduct multiple experiments by setting different values for the hyperparameters λ_1 and λ_2 in Eq. (11). Meanwhile, the hyperparameters λ_3 and λ_4 for $\mathcal{L}_{reg}$ and $\mathcal{L}_{dis}$ are set to 1 in this section of experiments. The results are shown in Table 4.

From the experimental results, it can be observed that the different value of hyperparameters also has a significant impact. A larger λ_1 will make the network more focused on adjusting the predicted probability distribution of each image towards an ideal state, which is discussed in Sect. 3.3, while a larger λ_2 will prioritize the ranking relationships between images. For the SCUT-FBP5500 dataset, the most outstanding results are achieved when $\lambda_1 = \lambda_2 = 1$. However, for different tasks, determining the values of λ_1 and λ_2 based on the characteristics of the specific task is crucial in achieving optimal performance.

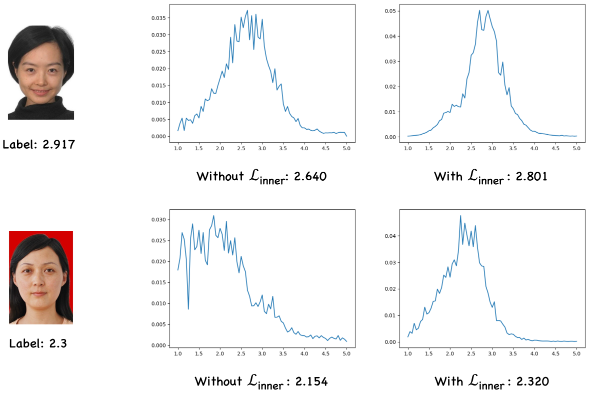

Fig. 4 The visualization of probability distributions of models with and without $\mathcal{L}_{inner}$ during training stage.

4.5 Visualization

To better illustrate the effect of $\mathcal{L}_{inner}$ during training stage, we select some samples and visualized their probability distributions as shown in Fig. 4.

In the example shown above in Fig. 4, although the probability distribution predicted by the model with $\mathcal{L}_{inner}$ is not very smooth itself, the jagged regions are significantly reduced compared to the distribution without $\mathcal{L}_{inner}$. In the example shown below in Fig. 4, the probability distribution predicted by the model without $\mathcal{L}_{inner}$ does not even approximate the ideal shape mentioned in Sect. 3.3. However, when $\mathcal{L}_{inner}$ is added, the distribution exhibits a rudimentary form of the ideal state, and the jagged regions of the distribution are also greatly reduced.

Based on the above, we can come to the conclusion that $\mathcal{L}_{inner}$ is able to utilize ranking information effectively during the training phase, optimizing the probability distribution of the images to a relatively ideal state and improving the accuracy of network predictions.

5. Conclusions

In this paper, we propose a novel end-to-end network architecture based on ranking information. We introduce a Rank Module and an Adaptive Weight Module to the Convolutional Neural Network (CNN) model, along with pairwise ranking loss functions. Unlike most methods that utilize ranking information, our approach only requires a single backbone network during training stage instead of sharing parameters between two backbone networks. This significantly reduces the training time of the network and even achieves better performance. Our experimental results on the SCUT-FBP5500 dataset reach a new state-of-the-art performance, and ablation experiments demonstrate that our method greatly assists in improving the performance of the model. Moreover, the Rank Module and Adaptive Weight Module we designed can be easily transferred to almost all CNN models. Additionally, the corresponding ranking information-based methods can be applied to any dataset that involves ranking information, such as most rating tasks and age estimation tasks.

Acknowledgments

This work was supported in part by the Key Research and Development Plan of Zhejiang: No.2021C03131; National Science Fund of China no.61871170; CEC Cooperation Funds: Intelligent Recognition Collaboration Project and TinyML Recognition Algorithms and Applications; Opening Fund of key Laboratory of Data Link Technology: CLDL-20202207.

References

[1] R. Thornhill and S.W. Gangestad, "Facial attractiveness," Trends in cognitive sciences, vol.3, no.12, pp.452–460, 1999.

[2] M. Bashour, "History and current concepts in the analysis of facial attractiveness," Plastic and reconstructive surgery, vol.118, no.3, pp.741–756, 2006.

[3] J. Li, C. Xiong, L. Liu, X. Shu, and S. Yan, "Deep face beautification," Proc. 23rd ACM international conference on Multimedia, pp.793–794, 2015.

[4] L. Liang, L. Jin, and D. Liu, "Edge-aware label propagation for mobile facial enhancement on the cloud," IEEE Trans. Circuits Syst. Video Technol., vol.27, no.1, pp.125–138, 2017.

[5] X. Ou, S. Liu, X. Cao, and H. Ling, "Beauty emakeup: A deep makeup transfer system," Proc. 24th ACM international conference on Multimedia, pp.701–702, 2016.

[6] A. Bottino, M. De Simone, A. Laurentini, and C. Sforza, "A new 3-d tool for planning plastic surgery," IEEE Trans. Biomed. Eng., vol.59, no.12, pp.3439–3449, 2012.

[7] K. Schmid, D. Marx, and A. Samal, "Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios," Pattern Recognition, vol.41, no.8, pp.2710–2717, 2008.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.770–778, 2016.

[10] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.4510–4520, 2018.

[12] A. Howard, M. Sandler, G. Chu, L.C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., "Searching for mobilenetv3," Proc. IEEE/CVF International Conference on Computer Vision, pp.1314–1324, 2019.

[13] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," International Conference on Machine Learning, pp.6105–6114, PMLR, 2019.

[14] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," 2018 24th International conference on pattern recognition (ICPR), pp.1598–1603, IEEE, 2018.

[15] Y.-Y. Fan, S. Liu, B. Li, Z. Guo, A. Samal, J. Wan, and S.Z. Li, "Label distribution-based facial attractiveness computation by deep residual learning," IEEE Trans. Multimedia, vol.20, no.8, pp.2196–2208, 2017.

[16] S. Liu, E. Huang, Y. Xu, K. Wang, X. Kui, T. Lei, and H. Meng, "Lightweight facial attractiveness prediction using dual label distribution," arXiv preprint arXiv:2212.01742, 2022.

[17] L. Lin, L. Liang, and L. Jin, "R^2-resnext: A resnext-based regression model with relative ranking for facial beauty prediction," 2018 24th International Conference on Pattern Recognition (ICPR), pp.85–90, IEEE, 2018.

[18] L. Lin, L. Liang, and L. Jin, "Regression guided by relative ranking using convolutional neural network (r^3cnn) for facial beauty prediction," IEEE Transactions on Affective Computing, vol.13, no.1, pp.122–134, 2022.

[19] E. Saravia, "ML Visuals," https://github.com/dair-ai/ml-visuals, 2021.

[20] T.-Y. Liu, "Learning to rank for information retrieval," Foundations and Trends® in Information Retrieval, vol.3, no.3, pp.225–331, 2009.

[21] C.J. Burges, "From ranknet to lambdarank to lambdamart: An overview," Learning, vol.11, no.23-581, p.81, 2010.

[22] T. Joachims, "Optimizing search engines using clickthrough data," Proc. eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.133–142, 2002.

[23] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," Journal of Machine Learning Research, vol.4, no.Nov, pp.933–969, 2003.

[24] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," Proc. 22nd international conference on Machine learning, pp.89–96, 2005.

[25] C.J.C. Burges, R. Ragno, and Q.V. Le, "Learning to rank with nonsmooth cost functions," Advances in neural information processing systems, vol.19, 2006.

[26] J. Bromley, J.W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," Advances in neural information processing systems, vol.6, 1993.

[27] V. Gattupalli, P.S. Chandakkar, and B. Li, "A computational approach to relative aesthetics," 2016 23rd International Conference on Pattern Recognition (ICPR), pp.2446–2451, IEEE, 2016.

[28] D. Gray, K. Yu, W. Xu, and Y. Gong, "Predicting facial beauty without landmarks," Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, Sept. 5-11, 2010, Proceedings, Part VI 11, pp.434–447, Springer, 2010.

[29] L. Xu, J. Xiang, and X. Yuan, "Transferring rich deep features for facial beauty prediction," arXiv preprint arXiv:1803.07253, 2018.

[30] J. Xu, L. Jin, L. Liang, Z. Feng, D. Xie, and H. Mao, "Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn)," 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp.1657–1661, IEEE, 2017.

[31] L. Liang, D. Xie, L. Jin, J. Xu, M. Li, and L. Lin, "Region-aware scattering convolution networks for facial beauty prediction," 2017 IEEE International Conference on Image Processing (ICIP), pp.2861–2865, IEEE, 2017.

[32] L. Lin, L. Liang, L. Jin, and W. Chen, "Attribute-aware convolutional neural networks for facial beauty prediction.," IJCAI, pp.847–853,

2019.

[33] L. Xu, J. Xiang, and X. Yuan, "Crnet: classification and regression neural network for facial beauty prediction," Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, Sept. 21-22, 2018, Proceedings, Part III, pp.661–671, Springer, 2018.

[34] L. Xu, H. Fan, and J. Xiang, "Hierarchical multi-task network for race, gender and facial attractiveness recognition," 2019 IEEE International conference on image processing (ICIP), pp.3861–3865, IEEE, 2019.

[35] J. Xu, "Mt-resnet: a multi-task deep network for facial attractiveness prediction," 2021 2nd International Conference on Computing and Data Science (CDS), pp.44–48, IEEE, 2021.

[36] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1492–1500, 2017.

[37] W. Wei, E.S.L. Ho, K.D. McCay, R. Damaševičius, R. Maskeliūnas, and A. Esposito, "Assessing facial symmetry and attractiveness using augmented reality," Pattern Analysis and Applications, vol.25, pp.635–651, 2021.

[38] J.N. Saeed, A.M. Abdulazeez, and D.A. Ibrahim, "Fiac-net: Facial image attractiveness classification based on light deep convolutional neural network," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), pp.1–6, IEEE, 2022.

[39] J.N. Saeed, A.M. Abdulazeez, and D.A. Ibrahim, "Automatic facial aesthetic prediction based on deep learning with loss ensembles," Applied Sciences, vol.13, no.17, p.9728, 2023.

[40] F. Bougourzi, F. Dornaika, N. Barrena, C. Distante, and A. Taleb-Ahmed, "Cnn based facial aesthetics analysis through dynamic robust losses and ensemble regression," Applied Intelligence, vol.53, pp.10825–10842, 2023.

[41] C.-T. Yang, Y.-C. Wang, L.-J. Lo, W.-C. Chiang, S.-K. Kuang, and H.-H. Lin, "Implementation of an attention mechanism model for facial beauty assessment using transfer learning," Diagnostics, vol.13, no.7, p.1291, 2023.

[42] S. Shi, F. Gao, X. Meng, X. Xu, and J. Zhu, "Improving facial attractiveness prediction via co-attention learning," ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4045–4049, IEEE, 2019.

Haochen Lyu is a graduate student at Hangzhou Dianzi University. He received the B.Sc. degree in computer science from Hangzhou Dianzi University in 2022. His research interests include image processing and video understanding algorithm.

Jianjun Li received the B.Sc. degree in information engineering from XiDian University, Xi'An, China, and the M.Sc. and Ph.D degrees in electrical and computer from The University of Western Ontario and University of Windsor, Canada separately. He is currently working at HangZhou Normal University as a chair professor and a joint professor at Hangzhou Dianzi University. His research interests include computer vision algorithms, micro-electronics, audio, video and image processing algorithms and implementation.

Yin Ye received the B.Sc. and M.Sc. degree in information engineering from XiDian University, Xi' an, China. She is currently working at Huada Electronic Design Corp., Ltd. as the Chief Engineer. Her work areas include information security technology, tinyML system design and IC design.

Chin-Chen Chang received the Ph.D. degree in computer engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1987. On numerous occasions, he was invited to serve as a visiting professor, the chair professor, an honorary professor, an honorary director, an honorary Chairperson, a distinguished alumnus, a distinguished researcher, and a research fellow by universities and research institutes. He has been the Chair Professor with the Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan, since February 2005. His current research interests include database design, computer cryptography, image compression, and data structures.