Uncertainty-oriented Order Learning for Facial Beauty Prediction

Xuefeng Liang, Member, IEEE, Zhenyou Liu, Jian Lin, Xiaohui Yang and Takatsune Kumada

Abstract—Previous Facial Beauty Prediction (FBP) methods generally model FB feature of an image as a point on the latent space, and learn a mapping from the point to a precise score. Although existing regression methods perform well on a single dataset, they are inclined to be sensitive to test data and have weak generalization ability. We think they underestimate two inconsistencies existing in the FBP problem: 1. inconsistency of FB standards among multiple datasets, and 2. inconsistency of human cognition on FB of an image. To address these issues, we propose a new Uncertainty-oriented Order Learning (UOL), where the order learning addresses the inconsistency of FB standards by learning the FB order relations among face images rather than a mapping, and the uncertainty modeling represents the inconsistency in human cognition. The key contribution of UOL is a designed distribution comparison module, which enables conventional order learning to learn the order of uncertain data. Extensive experiments on five datasets show that UOL outperforms the state-of-the-art methods on both accuracy and generalization ability.

Index Terms—Pairwise Comparison, Uncertainty Modeling, Order Learning, Facial Beauty Prediction.

1 Introduction

 $oldsymbol{\mathbb{C}}$ OCIOLOGICAL and psychological studies [1] have shown That Facial Beauty (FB) has a great impact on career development, interpersonal relationships, social status and social acceptance. Thus, Facial Beauty Prediction (FBP), a challenging task in computer vision, has attracted much attention. In the last decade, several methods had been applied for FBP [2], [3], [4], automatic facial beautification [5], and makeup recommendation [6], [7], [8], [9]. The pioneer FBP methods focused on designing handcrafted features based on aesthetic knowledge, such as geometric features [10], [11], [12], [13], [14], holistic features [15], [16], [17], [18], [19], [20] or mixed features [2], [21], [22], [23], [24], and performing prediction by classifiers (such as KNN [10], [17], [21], SVM [14], [23], [25], [26], decision trees [14], Adaboost [17], etc.) or regression algorithms (such as linear regression [11], [22], [27], ridge regression [16], [19], Gaussian regression [26], etc.).

Later, deep learning was introduced into FBP due to its superiority at various vision tasks. Many studies tried to learn the mapping from FB features to FB scores (the mean of multiple ratings) [3], [4], [28], [29], [30]. Although existing regression methods perform well on a single dataset, they often show weak generalization when tested across datasets. We think they underestimate two inconsistencies existing in the FBP: (1) *Inconsistency of FB standards*. For different FB datasets, the volunteers who rate the facial images are often different populations with different cultural and educational backgrounds, and thus have different reference systems for their FB standards. Even FB scores across datasets are



(a) Inconsistent FB standards across datasets.



Fig. 1. Two inconsistencies in FBP problem. (a). Three images, coming from SCUT-FBP5500, Hot-Or-Not and MEBeauty datasets respectively, have similar normalized FB scores but different FB appearances. (b). Ratings of a face image from different people are commonly inconsistent. Many FBP methods take the mean of these ratings as the FB score.

normalized to the same scale, biases still exist, as shown in Fig. 1 (a). In our experiment, many volunteers ranked the right face in the MEBeauty dataset as the most beautiful one, and the left face in the SCUT-FBP5500 as the second one; (2) *Inconsistency of human cognition*. Studies [31], [32] pointed that the FB ratings made by different people were more likely to diverge. Figure 1 (b) illustrates an example.

The inconsistency of FB standards among datasets has not been addressed in this field yet, but could be regarded as a nonlinear label shift in domain adaptation problem, in which different datasets are the overlapping subsets of the universal domain. Existing domain adaptation methods aim to learn domain invariant representations from multiple datasets via minimizing domain shift measures [33], optimal distribution matching [34], [35] and domain adversarial training [36]. On the contrary, we expect to learn the invariant in FB from a single dataset, which can be easily applied to other datasets. Our observation shows the

X. Liang (corresponding author) and X. Yang are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China E-mail: xliang@xidian.edu.cn

J. Lin and Z. Liu are with the Guangzhou Institute of Technology, Xidian University, Guangzhou, 510555, China

T. Kumada is with the Graduate School of Informatics, Kyoto University, Kyoto, Japan.

[•] Jian Lin and Zhenyou Liu contributed equally to this work.

order of FB is highly consistent across different datasets and can be learned from one FBP dataset. A psychology study [37] has shown that human subconsciously cognize realworld scales by learning an order rather than measuring exact values. An order pattern is essentially an awareness of relative relation that is independent on the reference base. The research [38] also pointed out that relative relations can be measured much easier than estimating precise quantities in many cases. Lin et al. [28] introduced the relative ranking into the loss function of a regression model for FBP problem. Lim et al. [39] proposed an order learning based on relative relations, and applied it to estimate precise facial ages, which demonstrated a better performance. We then apply the idea of order learning to learn the FB order of instances in the dataset to address the problem of the inconsistency of FB standards.

For the inconsistency of human cognition, some studies [31], [32] applied the label distribution learning (LDL) as the objective of regression model. However, LDL essentially learns the mapping from a feature point on the latent space to a label distribution, and is still inclined to overfit the data. In psychophysics, Thurstone proposed *A Law of Comparative Judgment* [40] to address such inconsistency, which is also known as uncertainty problem. Thurstone argued that the discriminal processes generated by a stimulus were not always equal, therefore, modeled it as a Gaussian distribution on a psychological scale, known as **discriminal dispersion**. Inspired by this, we model the inconsistent cognition of FB as a multi-dimensional Gaussian distribution on a high-dimensional psychological scale space.

However, conventional order learning can only compare data with precise labels rather than uncertain data. In order to address both inconsistencies, we design a module to compare distributions based on the Monte Carlo sampling, which enables order learning to learn the order relations of uncertain data.

Moreover, to compare with competing methods, the order relations must be transferred to FB scores. To this end, order learning needs a reference set that must be balanced, continuous, and cover entire range. Unfortunately, FBP datasets are usually small, unbalanced (i.e. medium ratings are majority), even discontinuous. To relax this constraint, we introduce the Bradley-Terry model [41]. It applies the maximum likelihood to estimate overall distribution using partial comparison results.

We conduct extensive experiments on the FBP benchmark dataset SCUT-FBP5500 and related datasets, Color FERET, Hot-Or-Not, MEBeauty and MIFS. The results show that our method outperforms six competing methods.

The main contributions of our work are threefold:

- We propose Uncertainty-oriented Order Learning (UOL), which enables order learning to learn the relative relations of uncertain data by a distribution comparison module.
- To address the inconsistency of FB standards among datasets, we apply order learning to learn the relative relations between instances. To address the inconsistency of human cognition, we model FB features as multi-dimensional Gaussian distributions on a psychological scale space, which can learn more robust relative relations of FB features.

• Extensive experiments demonstrate that our UOL has a better performance and generalization over the competing methods on SCUT-FBP5500 and other FBP datasets.

2 RELATED WORK

2.1 Facial Beauty Prediction

The earliest methods of FBP generally used handcrafted features, such as geometric features [10], [11], [12], [13], [14], holistic features (e.g., color features [19], [22], eigenface [2], [16], [17], [21], LBP features [15], [18], [20], etc.) or mixed features (i.e., geometric and holistic features), then built classification [10], [14], [21], [23], [25] or regression [4], [11], [16], [19], [22], [26] models.

Recently, convolutional neural networks (CNN) have become a mainstream method for FBP. Some studies considered FBP as a classification task. Gan et al. [42] proposed the 2MBeautyNet to improve the accuracy. Zhai et al. developed three models consecutively, i.e. the BeautyNet model [43], the AFBS model [44] and the BLS method [45].

More studies treated FBP as a regression problem. Xu et al. [29] proposed a CRNet that jointly optimized the classification branch and regression branch through classification and regression losses. Later, they introduced an improved expectation loss into ComboLoss [30] to boost the performance. Lin et al. [28] proposed a R^3 CNN that used the relative ranking based loss. Shi et al. [4] used pixels to mark different parts of a face as meta information and applied the co-attention learning mechanism to characterize the importance of different regions and different facial components simultaneously. Lin et al. [3] proposed a facial attribute aware convolution neural network, AaNet, which used a parameter generator to adaptively adjust the filter of the main network. F. Bougourzi et al. [46] aimed to train a better regression model, so they proposed CNN-ER and applied dynamic loss parameters to minimize the effect of outliers.

As most FBP datasets are scored by multiple people, an alternative solution for FBP is label distribution learning (LDL). Fan et al. [31] used LDL to train CNN on the basis of residual neural network. Wang et al. [32] proposed the LDL method LDL-LDM to exploit global and local label correlations on label distribution manifolds.

All above methods model the FB features of a facial image as a point on the latent space, and learn a mapping between the point and the given FB score, ranking or label distribution. Few of them consider the human cognition bias and the FB standard bias across datasets. Such direct regression mappings may suffer from weak generalization.

2.2 Order Learning and Pairwise Comparison

In everyday life, people often cognize the world and make decisions through comparisons. In psychology, kinds of abstract attitude are commonly measured by pairwise comparison, in which the subject compares a series of objects in pairs and makes a choice between two objects based on a certain criterion. Some studies have introduced pairwise comparison into their tasks. Ko et al. [47] proposed a Pairwise Aesthetic Comparison Network (PACNet) to extract features for image aesthetic assessment (IAA). Lee et al.

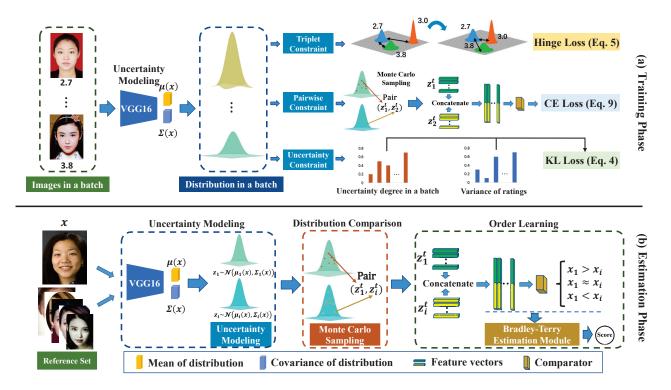


Fig. 2. (a) The training phase of UOL. The order of distributions is constrained by cross entropy loss and hinge loss, and the dispersion of the distributions is constrained by KL loss. (b) The estimation phase of UOL. In uncertainty modeling, the FB of a facial image is modeled by a multi-dimensional Gaussian distribution whose mean μ and diagonal covariance Σ are learned by VGG from the image. In distribution comparison, we sample from both the distributions of test image and reference image to form a pair and predict its order by a comparator in order learning. After having the order relations of T pairs between reference images and the test image, the Bradley-Terry model is applied to estimate the score of the test image.

[48] established the prototype of order learning, constructed a pairwise comparison matrix to predict image aesthetic scores. Hu et al. [49] proposed a learning framework based on pairwise comparison by focusing on the relative quality ranking of restored images. Lim et al. [39] formally proposed order learning and applied it to facial age estimation. It aims to discover the order of sequential patterns for a robust classification task. Both pairwise comparison methods and order learning can only work on data with certain label other than uncertain data, as they need to construct pairs for comparison.

2.3 Thurstone's Theory

Thurstone argued that a term is needed for the process by which the organism identifies, distinguishes, discriminates, or reacts to stimuli. It is known as the discriminal process and also called as attitude. Psychologically some of these attributes can be measured. FBP is a classical attitude measurement in the field of psychophysics, and the inconsistency of human cognition has been studied by psychologists for a long time. Thurstone's study showed that the discriminal processes generated by a stimulus were subject to noticeable fluctuation due to the different subjects who perceive the stimulus or the different environments. This fluctuation among the discriminal processes for a uniform repeated stimulus was designated the **discriminal dispersion**.

In his study, experiments showed that the discriminal dispersion which any given repeated stimulus produces on the psychological continuum is usually Gaussian. So Thurstone proposed the *Law of Comparative Judgment* [40] based on this theory and used the means and variances of discriminal dispersions to measure the attitudes on psychological scale. Later, Cohen [50] extended Thurstone's theory to multi-dimensional spaces.

2.4 Uncertainty Modeling

The uncertainty reflects the dispersion of a random variable. Since the FB ratings encode human cognition bias, FB is a kind of uncertain data. There have been a few preliminary works to model the uncertainty for other tasks. Gast et al. [51] proposed Probout to replace the intermediate activations with low-dimensional Gaussian distributions by adjusting the activation function, and obtained uncertainty in a lightweight manner instead of traditional Bayesian approaches. Liu et al. [52] considered the image quality as a distribution rather than a feature point, then modeled the uncertainty by a low-dimensional Gaussian distribution. However, low-dimensional Gaussian distribution naturally limited the feature expressiveness. In the field of face recognition, to reduce the uncertainty caused by image distortion, Shi et al. [53] applied probabilistic face embeddings to model the uncertainty of face features. Chang et al. [54] proposed data uncertainty learning (DUL) to achieve a joint learning of data embeddings and uncertainty. In the field of age estimation, Li et al. [55] proposed a probabilistic ordinal embedding (POEs) to treat each facial age data as a multivariate Gaussian distribution and applied a set of ordinal distribution to enforce ordinal property in the embedding space. Although DUL and POEs modeled facial

images as multi-dimensional Gaussian distributions, this uncertainty aimed to alleviate the effect of the inherent noise in the image, so only the "mean" of distribution is used for inference. Also, to avoid the distribution degenerating into deterministic embedding, the information bottleneck loss in POEs only ensures that the covariance matrix of distribution is close to the Identity matrix I. Instead, the uncertainty in our UOL denotes the inconsistency of human cognition, so we expect that the FB distribution to be close to the *discriminal dispersion* of human rating, and then take the variance of multiple ratings as the uncertainty to constrain the distribution. Thus, both the motivation and modeling approach of UOL have differences from above methods.

3 METHODS

Our proposed UOL consists of four modules: an order learning model in **section 3.1**; an uncertainty modeling module based on Thurstone's *discriminal dispersion* theory in **section 3.2**; a distribution comparison module in **section 3.3**, which enables order learning to learn the relative relations of uncertain data; and a FB score estimation module based on the Bradley-Terry model, which transforms the order relations to FB scores in **section 3.4**. Figure 2 shows the overall framework.

3.1 Order Learning

Order learning aims to learn the order relations between instances. As the order between FB is independent on the reference base, it can largely avoid the bias of FB standards introduced by different datasets. Following is the principle of order learning. Given two faces images, x_i and x_j , and their FB scores y_i and y_j respectively, the order between x_i and x_j can be defined and encoded by a one-hot label,

$$Y = \begin{cases} x_i \approx x_j : [1, 0, 0], & \text{if } |y_i - y_j| \le \theta, \\ x_i < x_j : [0, 1, 0], & \text{if } y_i - y_j < -\theta, \\ x_i > x_j : [0, 0, 1], & \text{if } y_i - y_j > \theta, \end{cases}$$
(1)

where $\theta=0.2$ is the threshold that represents the discrimination of FB.

The conventional order learning treats an instance as a fixed feature point on the latent space, which is learned by a network $g(\cdot)$, shown in the left of Fig.3. It then carries out pairwise feature comparisons between two instances. The comparator $f(\cdot,\cdot)$ in order learning consists of three fully connected layers and is formulated as

$$Y' = f(g(x_1), g(x_2)), \tag{2}$$

which learns order relation from precise labels of two samples in the pair.

3.2 Uncertainty Modeling

Previous methods of FBP usually use the mean of human ratings as the FB score for regression, which underestimate the human cognition biases. According to Thurstone's *discriminal dispersion* theory, the discriminal processes to the same stimulus are not always equal, but rather present a Gaussian distribution on the psychological scale. This

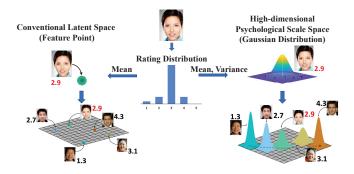


Fig. 3. In our UOL, FB features of facial images are modeled as the multi-dimensional Gaussian distributions on psychological scale space instead of fixed points on the conventional latent space.

theory is also applicable to FBP. We then design a high-dimensional psychological scale space to address the inconsistency of human cognition on FB. Specifically, we model the human ratings of an instance x as a multi-dimensional Gaussian distribution $z \sim \mathcal{N}(\mu(x), \Sigma(x))$ in the space, which is used as a feature for pairwise comparisons, as shown in the right of Fig. 3. A VGG16 is applied to encode mean vector $\mu(x)$ and covariance matrix $\Sigma(x)$ of the distribution. $\Sigma(x)$ represents the dispersion of the rating distribution. As $\Sigma(x)$ is a diagonal matrix, the degree of discriminal dispersion is the Frobenius norm of it,

$$\|\Sigma(x)\|_F = \sqrt{\sum_{i=j=1}^{D} |\sigma_{i,j}|}$$
 (3)

During training, the variance of the i-th instance's ratings is set as the ground truth, η_i , representing *discriminal dispersion* degree of the instance, and the network is optimized by minimizing the KL divergence between the predicted distribution of dispersion degrees $(\|\Sigma(x_1)\|_F, \|\Sigma(x_2)\|_F, \cdots, \|\Sigma(x_M)\|_F)$ and the ground truth distribution $(\eta_1, \eta_2, \cdots, \eta_M)$ of the M training instances. Unlike POEs [55], such operation can optimize the prediction to be as close as possible to the human ratings rather than a standard normal distribution,

$$\mathcal{L}_{Dis} = \sum_{m=1}^{M} \eta_m \cdot (\log(\eta_m) - \log(\|\Sigma(x_m)\|_F)), \quad (4)$$

where M is the total number of training instances.

It has been generally accepted in machine learning that data augmentation can improve the robustness of models. We think our uncertainty modeling can be considered as a specific form of data argumentation, because its process is similar to the feature-level data augmentation [56], [57]. Firstly, we build up a Gaussian distribution in the high-dimensional psychological scale space according to the human ratings. Then, we randomly sample from these Gaussian distributions for pairwise comparisons. This process can be considered as disturbing a single feature point on the latent space, which is the feature level augmentation. As the disturbed features usually belong to the same class of the original feature, such augmentation is often applied to classification tasks [56], [57]. Our order learning just is a triple classification problem.

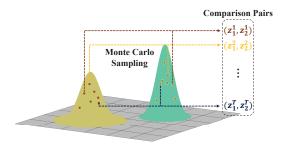


Fig. 4. Monte Carlo sampling of our distribution comparison module.

3.3 Comparison of Distribution

After modeling the FB uncertainty, an order should be established for these multi-dimensional Gaussian distributions on the phycological scale space. However, the conventional order learning cannot learn the order between distributions. Thurstone compares the observations of multiple subjects as the order of two stimuli on the psychological scale. To mimic this process, we design a uncertainty-oriented comparison module based on the Monte Carlo sampling.

To have a better order relation, we first constrain the Wasserstein distance between distributions on the psychological scale space. It allows that instances with similar scores have smaller distances between their distributions, while instances with significant different scores have larger distances. To this end, we apply a hinge loss to constrain the ordinal property of the psychological scale space and form a triplet for any three instances (x_l, x_m, x_n) from the dataset, who have ground truth (y_l, y_m, y_n) and corresponding feature distributions (z_l, z_m, z_n) ,

$$\mathcal{L}_{Ord} = \frac{1}{|S|} \sum_{(l,m,n) \in S} max(0, d(z_l, z_m) + \tau - d(z_l, z_n)),$$
 (5)

where $S=\{(l,m,n)\,|\,|y_l-y_m|<|y_l-y_n|\}$ and τ is the margin. $d(\cdot,\cdot)$ denotes the Wasserstein distance between two Gaussian distributions,

$$d(z_1, z_2)^2 = \sum_{j=1}^{D} (\mu_1^j - \mu_2^j)^2 + (\sigma_1^j - \sigma_2^j)^2,$$
 (6)

where μ_1^j , μ_2^j , σ_1^j and σ_2^j are the *j*-th dimension of μ_1 , μ_2 , $diag(\Sigma_1)$ and $diag(\Sigma_2)$ respectively. D is the dimensionality of the vector. The construction procedure of triplets can be found in Appendix A.

Afterwards, we apply T times Monte Carlo sampling on the distribution of instance x_i , which is analogous to the observations of multiple subjects on a stimulus. To make network be backpropagated, the random sampling and forward propagation must be separated. Thus, we apply the reparameterization sampling method [58] to get the t-th sampling $z_i^{(t)}$ from distribution z_i ,

$$z_i^{(t)} = \mu(x_i) + diag(\sqrt{\Sigma(x_i)}) \cdot \varepsilon^{(t)}, \ \varepsilon^{(t)} \sim \mathcal{N}(0, I), \quad (7)$$

where $\mathcal{N}(0,I)$ denotes the multi-dimensional Gaussian distribution with zero mean and identity covariance matrix I. The sampling process is shown in Fig. 4.

The comparator $f(\cdot,\cdot)$ in conventional order learning is applied to learn the order between two sampling feature

points. The relative relation Y' between two distributions of x_1 and x_2 is obtained by calculating the mean of T comparisons,

$$Y' = \frac{1}{T} \sum_{t=1}^{T} f(z_1^{(t)}, z_2^{(t)}).$$
 (8)

A cross-entropy loss \mathcal{L}_{Cls} for triple classification is used to optimize the comparator $f(\cdot, \cdot)$,

$$\mathcal{L}_{Cls} = -\log \frac{\exp(Y_c')}{\sum_{r=1}^{3} \exp(Y_r')},\tag{9}$$

where Y_c' and Y_r' denote the c-th and r-th dimensions of the output vector Y', c is the dimension where the ground truth is.

Thus, the entire loss of our UOL is

$$\mathcal{L} = \mathcal{L}_{Cls} + \alpha \mathcal{L}_{Ord} + \beta \mathcal{L}_{Dis}, \tag{10}$$

where α and β are weights to control the contribution of each loss function.

3.4 FB Score Estimation by Bradley-Terry Model

After establishing the order of samples, network cannot predict the FB score yet because the order is independent on the reference base. Conventional order learning [39] compares the input face image with a set of reference images whose labels cover the entire range of ages. These comparisons will find the most similar references to the input. Their precise labels will determine the label of input. This method is known as the maximum consistency (MC) rule [39], which requires the reference set must be balanced (the number of reference images must be the same for each interval) and continuous (no discontinue interval throughout the entire range).

However, most FBP datasets are unbalanced. Figure 5 (b) shows an example, the data distribution of SCUT-FBP5500. Thus, MC rule can only cover the range of 1.6~4.5, the blue box in Fig. 5 (c). To address this problem, we propose a score estimation method based on Bradley-Terry model. Specifically, an input with unknown score s_t is compared with a reference image with known score s_{r_i} . Bradley-Terry model tries to estimate the best s_t , and then models the possible order result Y and score difference $(s_t - s_{r_i})$ as the following probability distribution,

$$P(Y \le y_j, s_{r_i}; s_t) = \frac{e^{\delta_{y_j} + (ks_t - ks_{r_i})}}{1 + e^{\delta_{y_j} + (ks_t - ks_{r_i})}},$$

$$s_{r_i} \in S_{ref}, \quad y_j \in \{0, 1, 2\},$$
(11)

where 0, 1 and 2 represent the "<", " \approx " and ">" relations. $\delta_0 = -\delta_1 = -\delta$, $\delta_2 = +\infty$, δ is a positive parameter to control the probability of " \approx ". k is a scaling parameter to determine the change rate of probability. S_{ref} denotes the set of all scores in the reference set.

Suppose n images exist in the reference set and their ground truth scores are $\{s_1, s_2, \ldots, s_n\}, s_i \in S_{ref}$. We apply the optimized comparator $f(\cdot, \cdot)$ to predict the order between the input and each reference image, which results

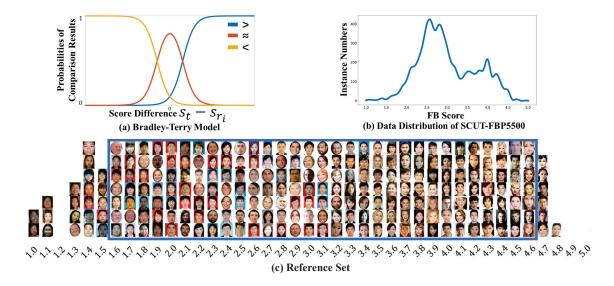


Fig. 5. Illustration of Bradley-Terry (BT) estimation module. (a) shows the probability distribution of BT model. (b) shows the data distribution of SCUT-FBP5500 dataset. (c) lists the reference set, all reference images with precise scores are going to be compared with the input image to estimate its FB score. The range marked by blue box is the reference set selected by MC rules [39], which cannot cover the entire range.

in $\{R_1,R_2,\ldots,R_n\}$, $R_i\in\{0,1,2\}$, then maximize the likelihood function,

$$L(s_t) = \prod_{i=1}^{n} P(Y = R_i, s_i; s_t).$$
 (12)

Finally, the FB score s_t of input image can be obtained.

The advantage of using maximum likelihood estimation is that FB scores can be estimated in the entire range by partial comparison results. Therefore, our UOL can work on unbalanced and discontinuous reference set. Obviously, the more complete the reference set is, the better performance our UOL achieves. Figure 5(c) shows the reference set for our UOL based on SCUT-FBP5500 dataset.

4 EXPERIMENT

4.1 Implementation Details

We use a pretrained VGG16 on ImageNet as the backbone of our method, and optimize it by Adam optimizer with a batch size of 32. The learning rate is 1e-4 at the beginning and a Cosine Annealing scheduler with the minimal learning rate 1e-6 is applied. We set the training epoch to 100 for all 5-fold cross validation. For data preprocessing, all the facial images (350×350) are resized to 256×256 firstly. Then a 224×224 center croping and a random horizontal flipping are performed, followed by per-pixel rescale to $0\sim1$ and mean value subtraction. The hyperparameters α and β in the loss function are 1e-4 and 1e-3, respectively.

We discretize the FB scores of training set at intervals of 0.1, and select $min(n_i, 10)$ images in each score interval from training set as the reference images to estimate the final score, where n_i denotes the total number of training data in the i-th score interval.

4.2 Datasets and Evaluation Metrics

4.2.1 Datasets

SCUT-FBP5500 [59] has 5500 frontal facial images, which was scored by 60 volunteers among the range of $1\sim5$. The

data consist of male/female and Asian/Caucasian faces, and have diverse annotation information (facial feature annotation, ratings and mean rating for each face by different volunteers). In this paper, we use the mean score and corresponding variance of different volunteers' ratings to train and test our model.

Hot-Or-Not [60] contains 2056 frontal female facial images aged 18-40 without constraint on race, lighting, pose or expression. The data was scored by 30 volunteers among the range of $-3\sim3$.

MEBeauty [61] includes 1300 females and 1250 males facial images. Each gender group is divided into six racial categories: Black, Asian, Caucasian, Hispanic, Indian and Middle Eastern. The data was scored by 300 volunteers among the range of $1\sim10$.

Color FERET [62] is a dataset for face recognition. It contains 11,338 color images of size 512×768 pixels captured in a semi-controlled environment with 13 different poses from 994 subjects. In this paper, 671 frontal images are selected to validate the generalization capability of our method.

MIFS (Makeup Induced Face Spoofing) [63] is a facial image dataset collected from YouTube videos of makeup impersonations, consisting of 107 makeup transformations. Each subject has two images with and without makeup. In real life, people usually believe faces who wear makeup are more attractive than those who do not. Therefore, we select facial image of each subject with and without makeup to evaluate the discrimination of UOL and competing methods.

In this work, all methods are only trained on the training set of SCUT-FBP5500, and evaluated without any fine-tuning on all five datasets, in which Hot-Or-Not, MEBeauty, Color FERET, and MIFS are unseen datasets for all methods during test. Such setting is for testing the generalization ability of UOL.

4.2.2 Evaluation Metrics

To test the effectiveness of UOL, we follow the evaluation metrics in [3], [4], [28], [29], [30], [46]: mean absolute error

TABLE 1
Performance comparison on SCUT-FBP5500. The results of competing methods are from their papers.

Methods	PC ↑	$MAE\downarrow$	$\text{RMSE}\downarrow$
R^{3} CNN [28]	0.9142	0.2120	0.2800
CRNet [29]	0.8869	0.2397	0.3186
AaNet [3]	0.9055	0.2236	0.2954
ComboLoss [30]	0.9199	0.2050	0.2704
Co-Attention [4]	0.9260	0.2020	0.2660
CNN-ER [46]	0.9250	0.2009	0.2650
UOL	0.9240	0.1975	0.2633

TABLE 2
Performance of different models on the Hot-Or-Not.

PC ↑	$MAE\downarrow$	RMSE ↓
0.3555	0.5741	0.7140
0.3250	0.5811	0.7294
0.2893	0.5923	0.7399
0.3329	0.6154	0.7677
0.2697	0.5613	0.7080
0.3513	0.5269	0.6653
0.4073	0.5410	0.6779
	0.3555 0.3250 0.2893 0.3329 0.2697 0.3513	0.3555 0.5741 0.3250 0.5811 0.2893 0.5923 0.3329 0.6154 0.2697 0.5613 0.3513 0.5269

(MAE) and root mean square error (RMSE). MAE measures the mean of absolute errors between the predictions and ground truth. RMSE measures the deviation between the predictions and ground truth. However, lower MAE and RMSE do not guarantee a better correlation of predictions and ground truth.

As SCUT-FBP5500, Hot-Or-Not and MEBeauty have varied ranges of FB scores rated by different people, MAE and RMSE are infeasible to evaluate the generalization ability of a method. The Pearson correlation coefficient (PC) [64] is a well-accepted metric for the evaluation of FB in psychological research. PC quantifies the degree of interdependence of prediction and ground truth. Thus, we employ PC as the metric to measure the generalization ability of a model trained on SCUT-FBP5500.

For MIFS dataset, we apply the accuracy rate (ACC) as the metric, which measures if the estimation of a face with makeup by a model is higher than that of the face without makeup.

4.3 Comparison with SOTA Methods

To verify the performance and generalization ability of UOL, we compare it with the state-of-the-art methods R^3 CNN [28], CRNet [29], Co-attention [4], AaNet [3], ComboLoss [30] and CNN-ER [46].

4.3.1 Performance Evaluation on SCUT-FBP5500

We firstly test all methods on the large-scale dataset SCUT-FBP5500, and report the result in Table 1. One can see that our UOL achieves the best on both MAE and RMSE, but slightly worse than Co-attention and CNN-ER on PC, which demonstrates that UOL has reached the state-of-the-art performance on SCUT-FBP5500.

4.3.2 Generalization Capability Evaluation

Our method aims to improve the generalization ability of the model by mimicking the human cognition. To this end,

TABLE 3
Performance of different models on the MEBeauty.

Methods	PC ↑	MAE ↓	RMSE ↓
R^{3} CNN [28]	0.5039	0.5329	0.6691
CRNet [29]	0.4380	0.5645	0.9019
AaNet [3]	0.3746	0.6102	0.7548
ComboLoss [30]	0.5078	0.5481	0.6888
Co-attention [4]	0.4976	0.5476	0.6907
CNN-ER [46]	0.4911	0.5753	0.6973
UOL	0.5532	0.5230	0.6489

TABLE 4
Comparison of the generalization ability of seven models on Color FERET and the accuracies on MIFS.

Methods	PC ↑ (Color FERET)	Acc(%) ↑ (MIFS)
R^{3} CNN [28]	0.8265	73.91
CRNet [29]	0.8564	96.15
AaNet [3]	0.7504	76.92
ComboLoss [30]	0.9146	92.31
Co-Attention [4]	0.7067	76.92
CNN-ER [46]	0.8490	96.15
UOL	0.9266	92.31

we train all methods on SCUT-FBP5500, and then test them on unseen datasets (Hot-Or-Not, MEBeauty, Color FERET, MIFS) which are not used to fine tune models. The competing methods are strictly implemented according to their open codes and papers.

(1) Experiments on Datasets with Human Ratings

We normalize scores of Hot-Or-Not and MEBeauty to the range $1{\sim}5$ and compute PC with the estimated scores by each method. Tables 2 and 3 show that the PC of UOL is considerably higher than those of other methods, which indicates that UOL has better generalization ability. The evaluations of MAE and RMSE show that UOL underperforms CNN-ER on Hot-Or-Not by $1\% \sim 2\%$, but outperform it on MEBeauty by 5%. One can see Co-attention performs worse than on SCUT-FBP5500, which shows it is sensitive to test data when FB standard shifts or image quality varies.

(2) Experiments on Datasets without Human Ratings

Color FERET and MIFS do not have human ratings. So we design two different experiments to evaluate the generalization ability of these methods.

For Color FERET, we select all 671 frontal face images as the test data, and apply UOL and six competing methods to give scores for each image for simulating human rating. After having all six ratings for each image, the lowest and highest ones are removed, the mean of the remaining ratings is considered as FB score. Afterwards, we do the same process on datasets with human ratings, and list the results in Table 4. UOL also achieves the highest PC.

For MIFS, facial images appear in pairs, in which one is with makeup and another one is not. We employ 7 volunteers to compare the image pairs, and clean the data according to the consistency of volunteers' results as follows:

Step 1: Manually select two frontal facial images of each face ID with and without makeup from MIFS, respectively. Group them as a pair.

Step 2: Show each pair to 7 volunteers. They vote the more beautiful image in the pair.

TABLE 5
The effectiveness evaluation of uncertainty modeling and order learning on SCUT-FBP5500 (SCUT), Hot-Or-Not (HON) and MEBeauty (MEB).

Methods	SCUT			HON	MEB
- Wichious	PC↑	MAE↓	$RMSE\downarrow$	PC↑	PC↑
VGG16(Regression)	0.9044	0.2248	0.2973	0.3675	0.5122
VGG16(Regression) + LDL	0.9076	0.2214	0.2909	0.3228	0.5338
VGG16(Regression) + Uncertainty Modeling	0.9080	0.2218	0.2920	0.3895	0.5367
VGG16(Order Learning)	0.9198	0.2025	0.2683	0.3958	0.5442
UOL	0.9240	0.1975	0.2633	0.4073	0.5532

TABLE 6
The effectiveness evaluation of three loss functions on Hot-Or-Not and MEBeauty.

CE Hinge K	KL	Hot-Or-Not				MEBeauty		
	KL	PC↑	$MAE\downarrow$	$\text{RMSE}\downarrow$	PC ↑	$MAE\downarrow$	RMSE ↓	
√			0.4007	0.5419	0.6844	0.5457	0.5252	0.6543
\checkmark	\checkmark		0.4036	0.5410	0.6793	0.5463	0.5231	0.6508
\checkmark		\checkmark	0.4036	0.5426	0.6792	0.5397	0.5290	0.6592
✓	✓	✓	0.4073	0.5410	0.6779	0.5532	0.5230	0.6489

Step 3: Calculate the consistency of the volunteers' votes of each face ID, select pairs with higher consistency (more than 5 volunteers give the same vote) as the test data. The ground truth is the majority voting of volunteers.

We apply each method on images in a pair. If the estimated score of image with makeup is higher than the one without makeup, the comparison is correct, otherwise incorrect. Table 4 shows UOL is the second best. After carefully examining the results, we find the two pairs misestimated by UOL are also misestimated by other four methods. It indicates that some unknown FB features have not been explored by these methods. CRNet just misestimates a pair, the best in this experiment, but performs worse than its upgrade version, ComboLoss, on other experiments.

All above results demonstrate the generalization ability of UOL outperforms the competing FBP algorithms.

4.4 Ablation Studies

4.4.1 Effectiveness of order learning and uncertainty modeling

To validate the effectiveness of order learning and uncertainty modeling respectively, we conduct ablation studies on three datasets with FB scores. It is worth noting that we do not separately evaluate Bradley-Terry module, because UOL cannot estimate scores covering whole range without it. All versions are trained on the SCUT-FBP5500. The backbone is VGG16. The results under different settings are listed in Table 5. One can see that order learning contributes a significant performance gain, because it is more consistent with human cognition to order patterns than regression approaches. Uncertainty modeling also boosts the performance of a regression model with a marginal gain. Their integration, UOL, can further boost the performance. These results demonstrate that order is a very valuable invariant in FB, and uncertainty modeling is more feasible for order learning (a classification model) than a regression model.

Label distribution learning (LDL) could be also considered an uncertainty modeling. We then apply LDL to the backbone and report the results in Table 5. It can be seen that LDL achieves a marginal gain on SCUT-FBP5500 and

MEBeauty, but performs worse on Hot-Or-Not. The possible reason is that SCUT-FBP5500 and MEBeauty have similar data distributions, but Hot-Or-Not has different distribution. LDL essentially learns the mapping from fixed points on the latent space to certain distributions and is inclined to overfit the label distribution. Please note that order learning has difficulty in using label distribution because its form is not comparable to learn order.

4.4.2 Effectiveness of three losses

UOL employs three loss functions that play different roles. CE Loss aims at training the comparator in Fig. 2 for estimating the relative order between instances. CE Loss is indispensable to our UOL. Hinge Loss constrains the distance between the modeled distributions of instances on the latent space, which can improve the representation of order relation and further boost the order estimation. KL Loss constrains the consistency between the modeled distribution of instances and the variance of human ratings for instances, which aims at a more accurate distance metric in Hinge Loss.

We also evaluate the effectiveness of them and report the results in Table 6. We can see that Hinge Loss helps CE Loss achieve better performance, KL Loss further boosts the performance when CE Loss and Hinge Loss work together. But CE + KL Losses degrade the performance because KL Loss cannot directly help CE Loss without Hinge Loss.

5 CONCLUSION

In this paper, we propose a novel Uncertainty-oriented Order Learning for facial beauty prediction. UOL enables order learning to learn the relative relations of uncertain data by a distribution comparison module, in which order learning addresses the inconsistency of FB standards between datasets, and uncertainty modeling tackles the inconsistency of human cognition of FB. In addition, we introduce the Bradley-Terry model into order learning to relax the restriction that the reference set must be continuous and balanced. Extensive experiments demonstrate that our method outperforms state-of-the-art methods on SCUT-FBP5500 in terms

of FB score prediction, and better generalization on other datasets. However, the improper use of FBP models might result in an unethical impact. Devising better data forensics approaches could be countermeasures. In the future work, we will explore the impact of face attributes on the UOL.

REFERENCES

- A. Laurentini and A. Bottino, "Computer analysis of face beauty: A survey," Computer Vision and Image Understanding, vol. 125, pp. 184–199, 2014.
- [2] H. Altwaijry and S. Belongie, "Relative ranking of facial attractiveness," in 2013 IEEE Workshop on Applications of Computer Vision (WACV). IEEE, 2013, pp. 117–124.
- [3] L. Lin, L. Liang, L. Jin, and W. Chen, "Attribute-aware convolutional neural networks for facial beauty prediction." in *IJCAI*, 2019, pp. 847–853.
- [4] S. Shi, F. Gao, X. Meng, X. Xu, and J. Zhu, "Improving facial attractiveness prediction via co-attention learning," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 4045–4049.
- Signal Processing (ICASSP). IEEE, 2019, pp. 4045–4049.

 [5] L. Liang, L. Jin, and X. Li, "Facial skin beautification using adaptive region-aware masks," IEEE transactions on cybernetics, vol. 44, no. 12, pp. 2600–2612, 2014.
- [6] T. Alashkar, S. Jiang, S. Wang, and Y. Fu, "Examples-rules guided deep neural network for makeup recommendation," in *Proceedings* of the AAAI conference on artificial intelligence, vol. 31, no. 1, 2017.
- [7] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, and S. Yan, "Wow! you are so beautiful today!" ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 1s, pp. 1– 22, 2014.
- [8] X. Ou, S. Liu, X. Cao, and H. Ling, "Beauty emakeup: A deep makeup transfer system," in *Proceedings of the 24th ACM interna*tional conference on Multimedia, 2016, pp. 701–702.
- [9] K. Scherbaum, T. Ritschel, M. Hullin, T. Thormählen, V. Blanz, and H.-P. Seidel, "Computer-suggested facial makeup," in *Computer Graphics Forum*, vol. 30, no. 2. Wiley Online Library, 2011, pp. 485–492.
- [10] P. Aarabi, D. Hughes, K. Mohajer, and M. Emami, "The automatic measurement of facial beauty," in 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), vol. 4. IEEE, 2001, pp. 2644–2647.
- [11] J. Fan, K. Chau, X. Wan, L. Zhai, and E. Lau, "Prediction of facial attractiveness from facial proportions," *Pattern Recognition*, vol. 45, no. 6, pp. 2326–2334, 2012.
- [12] L. G. Farkas and G. Cheung, "Facial asymmetry in healthy north american caucasians: an anthropometrical study," *The Angle Orthodontist*, vol. 51, no. 1, pp. 70–77, 1981.
- [13] H. Gunes and M. Piccardi, "Assessing facial beauty through proportion analysis by image processing and supervised learning," International journal of human-computer studies, vol. 64, no. 12, pp. 1184–1199, 2006.
- [14] H. Mao, L. Jin, and M. Du, "Automatic classification of chinese female facial beauty using support vector machine," in 2009 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2009, pp. 4842–4846.
- [15] J. Gan, L. Li, Y. Zhai, and Y. Liu, "Deep self-taught learning for facial beauty prediction," *Neurocomputing*, vol. 144, pp. 295–303, 2014.
- [16] Y. Mu, "Computational facial attractiveness prediction by aesthetics-aware features," *Neurocomputing*, vol. 99, pp. 59–64, 2013.
- [17] D. Sutić, I. Brešković, R. Huić, and I. Jukić, "Automatic evaluation of facial attractiveness," in *The 33rd International Convention MIPRO*. IEEE, 2010, pp. 1339–1342.
- [18] S. Wang, M. Shao, and Y. Fu, "Attractive or not? beauty prediction with attractiveness-aware encoders and robust late fusion," in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 805–808.
- [19] R. White, A. Eden, and M. Maire, "Automatic prediction of human attractiveness," UC Berkeley CS280A Project, vol. 1, no. 2, 2004.
- [20] H. Yan, "Cost-sensitive ordinal regression for fully automatic facial beauty assessment," *Neurocomputing*, vol. 129, pp. 334–342, 2014.

- [21] Y. Eisenthal, G. Dror, and E. Ruppin, "Facial attractiveness: Beauty and the machine," *Neural computation*, vol. 18, no. 1, pp. 119–142, 2006.
- [22] A. Kagian, G. Dror, T. Leyvand, D. Cohen-Or, and E. Ruppin, "A humanlike predictor of facial attractiveness," Advances in Neural Information Processing Systems, vol. 19, 2006.
- [23] J. Whitehill and J. R. Movellan, "Personalized facial attractiveness prediction," in 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2008, pp. 1–7.
- [24] D. Zhang, F. Chen, Y. Xu et al., Computer models for facial beauty analysis. Springer, 2016.
- [25] A. Bottino and A. Laurentini, "The intrinsic dimensionality of attractiveness: A study in face profiles," in *Iberoamerican Congress* on Pattern Recognition. Springer, 2012, pp. 59–66.
- [26] D. Xie, L. Liang, L. Jin, J. Xu, and M. Li, "Scut-fbp: A benchmark dataset for facial beauty perception," in 2015 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2015, pp. 1821– 1826.
- [27] 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.
- [28] L. Lin, L. Liang, and L. Jin, "Regression guided by relative ranking using convolutional neural network (r3cnn) for facial beauty prediction," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 122–134, 2022.
- [29] L. Xu, J. Xiang, and X. Yuan, "Crnet: Classification and regression neural network for facial beauty prediction," in *Pacific Rim Confer*ence on Multimedia. Springer, 2018, pp. 661–671.
- [30] L. Xu and J. Xiang, "Comboloss for facial attractiveness analysis with squeeze-and-excitation networks," *arXiv preprint arXiv*:2010.10721, 2020.
- [31] 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 Transactions on Multimedia*, vol. 20, no. 8, pp. 2196–2208, 2017.
- [32] J. Wang and X. Geng, "Label distribution learning by exploiting label distribution manifold," *IEEE Transactions on Neural Networks* and Learning Systems, pp. 1–14, 2021.
- [33] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.
- [34] R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 39, no. 9, pp. 1853–1865, 2017.
- [35] M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren, "Enhanced transport distance for unsupervised domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13936–13944.
- [36] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domainadversarial training of neural networks," *The journal of machine* learning research, vol. 17, no. 1, pp. 2096–2030, 2016.
- [37] H. A. Simon, "A behavioral model of rational choice," *The quarterly journal of economics*, vol. 69, no. 1, pp. 99–118, 1955.
- [38] T. L. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of mathematical psychology*, vol. 15, no. 3, pp. 234–281, 1977.
- [39] K. Lim, N.-H. Shin, Y.-Y. Lee, and C.-S. Kim, "Order learning and its application to age estimation," in *International Conference on Learning Representations*, 2019.
- [40] L. L. Thurstone, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.
- [41] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
- [42] J. Gan, L. Xiang, Y. Zhai, C. Mai, G. He, J. Zeng, Z. Bai, R. D. Labati, V. Piuri, and F. Scotti, "2m beautynet: Facial beauty prediction based on multi-task transfer learning," *IEEE Access*, vol. 8, pp. 20245–20256, 2020.
- [43] Y. Zhai, H. Cao, W. Deng, J. Gan, V. Piuri, and J. Zeng, "Beautynet: Joint multiscale cnn and transfer learning method for unconstrained facial beauty prediction," Computational intelligence and neuroscience, vol. 2019, 2019.
- [44] Y. Zhai, Y. Huang, Y. Xu, J. Gan, H. Cao, W. Deng, R. D. Labati, V. Piuri, and F. Scotti, "Asian female facial beauty prediction using

- deep neural networks via transfer learning and multi-channel feature fusion," *IEEE Access*, vol. 8, pp. 56892–56907, 2020.
- [45] Y. Zhai, C. Yu, C. Qin, W. Zhou, Q. Ke, J. Gan, R. D. Labati, V. Piuri, and F. Scotti, "Facial beauty prediction via local feature fusion and broad learning system," *IEEE Access*, vol. 8, pp. 218 444–218 457, 2020.
- [46] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed, "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression," *Knowledge-Based Systems*, vol. 242, p. 108246, 2022.
- [47] K. Ko, J.-T. Lee, and C.-S. Kim, "Pac-net: pairwise aesthetic comparison network for image aesthetic assessment," in 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018, pp. 2491–2495.
- [48] J.-T. Lee and C.-S. Kim, "Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1191–1200.
- [49] B. Hu, L. Li, H. Liu, W. Lin, and J. Qian, "Pairwise-comparison-based rank learning for benchmarking image restoration algorithms," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2042–2056, 2019.
- [50] H. S. Cohen, A MULTI-DIMENSIONAL ANALOGY TO THUR-STONE'S LAW OF COMPARATIVE JUDGMENT. University of Illinois at Urbana-Champaign, 1973.
- [51] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3369–3378.
- [52] Y. Liu, F. Wang, and A. W. K. Kong, "Probabilistic deep ordinal regression based on gaussian processes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5301–5309.
- [53] Y. Shi and A. K. Jain, "Probabilistic face embeddings," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6902–6911.
- [54] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5710–5719.
- [55] W. Li, X. Huang, J. Lu, J. Feng, and J. Zhou, "Learning probabilistic ordinal embeddings for uncertainty-aware regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13896–13905.
- [56] B. Li, F. Wu, S.-N. Lim, S. Belongie, and K. Q. Weinberger, "On feature normalization and data augmentation," in *Proceedings of* the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12383–12392.
- [57] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," Advances in Neural Information Processing Systems, vol. 32, 2019.
- [58] D. Liang, Ř. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 689–698.
- [59] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li, "Scut-fbp5500: a diverse benchmark dataset for multi-paradigm facial beauty prediction," in 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 1598–1603.
- [60] D. Gray, K. Yu, W. Xu, and Y. Gong, "Predicting facial beauty without landmarks," in European Conference on Computer Vision. Springer, 2010, pp. 434–447.
- [61] I. Lebedeva, Y. Guo, and F. Ying, "Mebeauty: a multi-ethnic facial beauty dataset in-the-wild," *Neural Computing and Applications*, pp. 1–15, 2021.
- [62] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.
- [63] C. Chen, A. Dantcheva, T. Swearingen, and A. Ross, "Spoofing faces using makeup: An investigative study," in 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). IEEE, 2017, pp. 1–8.
- [64] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

APPENDIX A

TRIPLET CONSTRUCTION FOR HINGE LOSS

To model the order on psychological scale space described in Sec. 3.3, we construct a triplet of instances (x_l,x_m,x_n) for the hinge loss, which ensures that their FB scores (y_l,y_m,y_n) meet $|y_l-y_m|<|y_l-y_n|$. Theoretically, selecting instances can be random, it makes training time consuming. Therefore, we define a hard triplet, whose difference between $|y_l-y_m|$ and $|y_l-y_n|$ is small, and construct M hard triplets from a batch of training data with M instances. We take the i-th instance as the anchor t, then set the instance with index t instance t from the remaining t instances as the third element, which satisfies:

$$\underset{n \neq l,m}{\arg \min} \left| \left| y_l - y_m \right| - \left| y_l - y_n \right| \right|,$$
where $\left| y_l - y_m \right| \neq \left| y_l - y_n \right|.$ (13)

Algorithm 1 describes the strategy of this hard triplets selection.

Algorithm 1: Hard Triplets Selection

```
Input: FB scores of M (batch size) instances
               y=\{y_1,y_2,\ldots,y_M\}.
   Output: Triplets of index, originized in batch form,
               (l = \{l_1, l_2, \dots, l_M\}, m = \{m_1, m_2, \dots, m_M\},\
               n = \{n_1, n_2, \dots, n_M\}).
1 for i ← 1 to M do
        l_i \leftarrow i;
        m_i \leftarrow (i+1) \mod M;
        dis_{12} \leftarrow |y_{l_i} - y_{m_i}|;
4
        sub = +\infty;
5
6
        for j \leftarrow 1 to M do
             if j \neq l_i and j \neq m_i then
                  dis_{13} \leftarrow |y_{l_i} - y_j|;
 8
                  tmp \leftarrow |\overrightarrow{dis}_{12} - \overrightarrow{dis}_{13}|;
 9
                  if tmp < sub and tmp \neq 0 then
10
                       sub \leftarrow tmp;
11
                       n_i \leftarrow j;
12
                  end
13
14
             end
        end
15
16 end
```

APPENDIX B

PAIR CONSTRUCTION FOR PAIRWISE COMPARISON

Assume FB scores are in the range of [1,H]. In order learning, we define a pair of two instances, who meet $|y_i-y_j|<\theta$, as a " \approx " pair, where $\theta\ll H$. If instances are selected randomly in each batch, the number of " \approx " pairs will be much smaller than ">" pairs and "<" pairs. Such data imbalance makes the trained comparator in **Sec. 3.3** overfit the ">" and "<" pairs. Meanwhile, random selection may let an instance frequently appear in a batch. It also makes networks overfit this instance. To tackle these problems, we design a balanced pairs selection strategy, which ensures the proportion of " \approx " pairs in a batch is close to $\frac{1}{3}$, and all instances appear almost equally in a batch. Algorithm 2 describes our balanced pairs selection strategy.

Algorithm 2: Balanced Pairs Selection

```
Input: FB scores of M (batch size) instances
            y = \{y_1, y_2, \dots, y_M\}; adjacency list flag = \{flag[1], flag[2], \dots, flag[M]\},
            where flag[i] is an empty list denotes which
            instances have been paired with the i-th
            instance; limitation of every instance can be
            selected N; threshold \theta.
   Output: Pairs of index p, and corresponding order
            relationships Y.
1 p \leftarrow \{ \};
2 Y \leftarrow \{\};
3 for i \leftarrow 1 to M do
      candidates \leftarrow \{1, 2, \dots, M\};
       // candidates is a list denotes which
       instances can be selected.
       DELETE(candidates, i);
       // DELETE denotes deleting an
       element from a list.
       for j \leftarrow 0 to LEN(flag[i]) do
          DELETE(candidates, flag[i][j]);
10
       end
       \mathbf{for}\ j \leftarrow 1\ \mathbf{to}\ M\ \mathbf{do}
11
           if flag[i] > N then
12
            DELETE(candidates, j);
13
14
15
       end
       while flag[i] < N and candidates \neq \{ \} do
16
           sim \leftarrow 0;
17
           for j \leftarrow 1 to LEN(candidates) do
18
19
               dif \leftarrow |y_i - candidates[j]|;
              if dif < \theta then
20
                  sim \leftarrow sim + 1;
21
22
                   // sim is the number of "pprox"
                  pairs in candidates.
23
              end
24
           end
           unsim \leftarrow LEN(candidates) - sim;
25
26
           prob = \{ \};
27
           // prob is the list denotes
           probability of every instance to
           be selected.
           for j \leftarrow 1 to LEN(candidates) do
28
               dif \leftarrow |y_i - candidates[j]|;
29
              if dif < \theta then
30
                  INSERT(prob, \frac{1}{3*sim});
31
                   // INSERT denotes appending
32
                  an element to a list.
               else
33
                INSERT(prob, \frac{2}{3*unsim});
34
35
              end
36
           end
           r \leftarrow \text{RAN}-
37
            DOM_CHOICE_BY_PROB(candidates, prob);
           // select an instance's index r
38
           from candidates by their
           probabilities prob, we implement
           this by using numpy.
           INSERT(flag[i], r);
INSERT(flag[r], i);
39
40
           INSERT(p, \{i, r\});
41
           order \leftarrow GENLABEL(|y_i - y_r|, \theta);
42
           // GENLABEL denotes generating the
43
           order label for a pair by their
           ground truth and threshold.
           INSERT(Y, order);
44
          DELETE(candidates, r);
45
46
      end
47 end
```