UNCERTAINTY-ORIENTED ORDER LEARNING FOR FACIAL BEAUTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

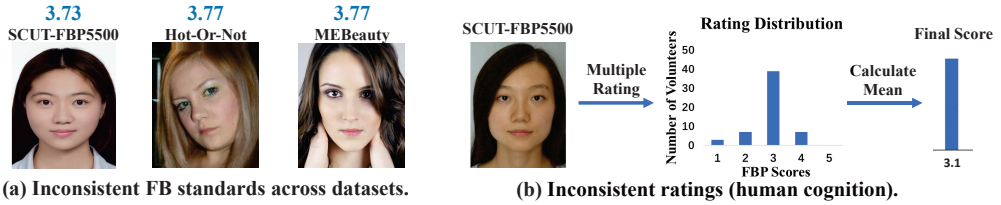
Previous Facial Beauty Prediction (FBP) methods generally model FB feature of an image as a point on the latent space, and learn a mapping from the point to a precise score. Although existing regression methods perform well on a single dataset, they are inclined to be sensitive to test data and have 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 show that UOL outperforms the state-of-the-art methods on both accuracy and generalization ability.

1 INTRODUCTION

Sociological and psychological studies (Laurentini & Bottino, 2014) 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 (Altwaijry & Belongie, 2013; Lin et al., 2019; Shi et al., 2019), automatic facial beautification (Liang et al., 2014), and makeup recommendation (Alashkar et al., 2017; Liu et al., 2014; Ou et al., 2016; Scherbaum et al., 2011).

The pioneer FBP methods focused on designing handcrafted features based on aesthetic knowledge, such as geometric features (Aarabi et al., 2001; Fan et al., 2012; Farkas & Cheung, 1981; Gunes & Piccardi, 2006; Mao et al., 2009), holistic features (Gan et al., 2014; Mu, 2013; Sutić et al., 2010; Wang et al., 2014; White et al., 2004; Yan, 2014) or mixed features (Altwaijry & Belongie, 2013; Eisenthal et al., 2006; Kagian et al., 2006; Whitehill & Movellan, 2008; Zhang et al., 2016), and performing prediction by classifiers (such as KNN (Aarabi et al., 2001; Sutić et al., 2010; Eisenthal et al., 2006), SVM (Mao et al., 2009; Whitehill & Movellan, 2008; Bottino & Laurentini, 2012; Xie et al., 2015), decision trees (Mao et al., 2009), Adaboost (Sutić et al., 2010), etc.) or regression algorithms (such as linear regression (Fan et al., 2012; Kagian et al., 2006; Schmid et al., 2008), ridge regression Mu (2013); White et al. (2004), Gaussian regression (Xie et al., 2015), 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) (Lin et al., 2019; Shi et al., 2019; Lin et al., 2022; Xu et al., 2018; Xu & Xiang, 2020). 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 among datasets due to the nonalignment of FB reference bases between datasets. Even FB scores across different datasets are normalized to the same scale, obvious biases still exist, as shown in Fig. 1 (a); (2) inconsistency of human cognition. Studies (Fan et al., 2017; Wang & Geng, 2021) pointed that the FB ratings made by different people were more likely to diverge. Figure 1 (b) illustrates an example.



(a) Inconsistent FB standards across datasets. (b) Inconsistent ratings (human cognition).

Figure 1: Two inconsistencies in FBP problem. (a). Three images, coming from SCUT-FBP5500, Hot-Or-Not and MEBeauty datasets respectively, have similar 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.

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 (Sun et al., 2016), optimal distribution matching (Flamary et al., 2017; Li et al., 2020) and domain adversarial training (Ganin et al., 2016). 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 order of FB is highly consistent across different datasets and can be learned from one FBP dataset. A psychology study (Simon, 1955) has shown that human subconsciously cognize real-world 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 (Saaty, 1977) also pointed out that relative relations can be measured much easier than estimating precise quantities in many cases. Lin et al. (2022) introduced the relative ranking into the loss function of a regression model for FBP problem. Lim et al. (2019) 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 (Fan et al., 2017; Wang & Geng, 2021) 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, thus still suffers from weak generalization. In psychophysics, Thurstone (1927c) proposed *A Law of Comparative Judgment* 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, called **discriminal dispersion**[1]. Inspired by this, we introduce an uncertain modeling method, which models 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.

To compare with competing methods, we compare the input face image with a set of reference images from the dataset to transfer the relative relations to FB score, whose labels cover the range of FB scores. The label of the most similar references will be the score of the input.

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 source code will be available soon.

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.

[1]Please refer to Appendix A for more details about *Law of Comparative Judgment*.

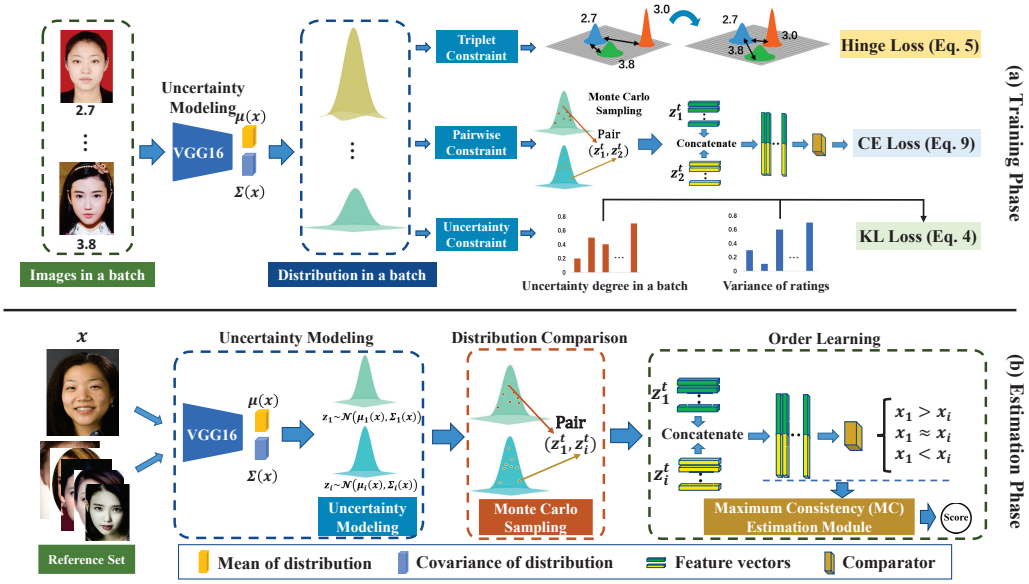


Figure 2: (a) The training phase of UOL. The order of distributions is constrained by cross entropy loss and hinge loss, and the dispersion of the distributions is constrained by KL loss. (b) The estimation phase of UOL. In uncertainty modeling, the FB of a facial image is modeled by a multi-dimensional 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 Maxinum Consistency (MC) rule (Lim et al., 2019) is applied to estimate the score of the test image.

• 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. The code will be available soon.

2 METHODS

Our proposed UOL consists of four modules: an order learning model in **section 3.1**; an uncertainty modeling module in **section 3.2**; a distribution comparison module in **section 3.3**; a FB score estimation module in **section 3.4**. Figure 2 shows the overall framework.

2.1 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,

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

where θ is the threshold that represents the discrimination of FB.

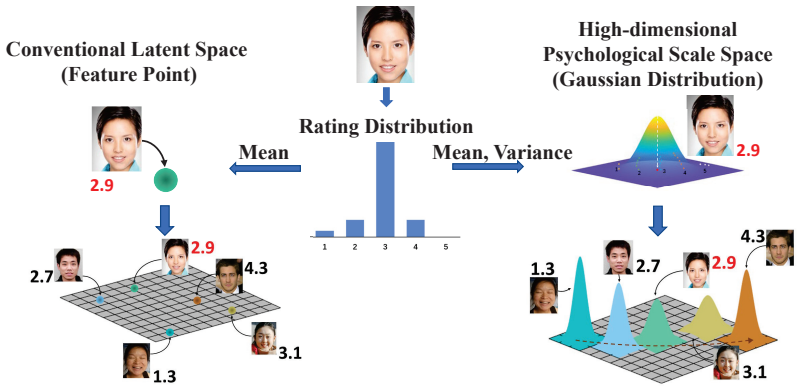


Figure 3: In our UOL, FB features of facial images are modeled as the multi-dimensional Gaussian distributions on psychological scale space instead of fixed points on the conventional latent space.

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 $Y' = f(g(x_1), g(x_2))$, which learns order relation from precise labels of two samples in the pair.

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

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. Such operation can optimize the prediction to be as close as possible to the human ratings,

$$\mathcal{L}_{Dis} = \sum_{m=1}^{M} \eta_m \cdot (\log(\eta_m) - \log(\|\Sigma(x_m)\|_F)), \tag{2}$$

From another perspective, our uncertainty modeling can be considered as a feature-level data augmentation (Li et al., 2021a; Wang et al., 2019), which can improve the robustness of models. The detailed discussion can be found in Appendix B.

There have been a few works to model the uncertainty differently for other tasks. Probout (Gast & Roth, 2018) replaces the intermediate activations with low-dimensional Gaussian distributions by adjusting the activation function, and obtains uncertainty in a lightweight manner instead of traditional Bayesian approaches. GP-DNNOR (Liu et al., 2019) considers the image quality as a distribution rather than a feature point, then models the uncertainty by a low-dimensional Gaussian distribution. However, low-dimensional Gaussian distribution naturally limits the feature expressiveness. By contrast, DUL (Chang et al., 2020) and POEs (Li et al., 2021b) model facial images as multi-dimensional Gaussian distributions, but this uncertainty aims to alleviate the effect of the

inherent noise in the image, so only the "mean" of distribution is used for inference. Also, the information bottleneck loss in POEs emphasizes 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, which is different from the Identity matrix. Thus, both the motivation and modeling approach of UOL have differences from these methods.

2.3 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),

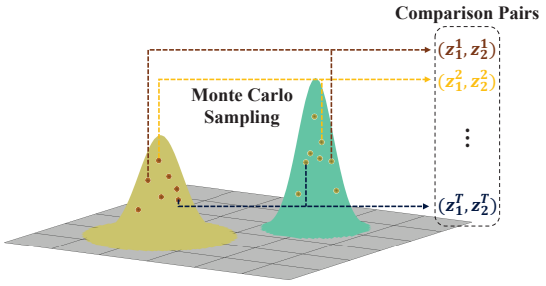


Figure 4: Monte Carlo sampling of our distribution comparison module.

$$\mathcal{L}_{Ord} = \frac{1}{|S|} \sum_{(l,m,n)\in S} max(0, d(z_l, z_m) + \tau - d(z_l, z_n)), \tag{3}$$

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

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 (Liang et al., 2018a), to get the t-th sampling $z_i^{(t)}$ from distribution z_i, where $z_i^{(t)} = \mu(x_i) + diag(\sqrt{\Sigma(x_i)}) \cdot \varepsilon^{(t)}$, $\varepsilon^{(t)} \sim \mathcal{N}(0, I)$. $\mathcal{N}(0, I)$ 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, $Y' = \frac{1}{T} \sum_{t=1}^{T} f(z_1^{(t)}, z_2^{(t)})$.

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

$$\mathcal{L}_{Cls} = -\log \frac{\exp(Y_c')}{\sum_{r=1}^{3} \exp(Y_r')}, \tag{4}$$

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

$$\mathcal{L} = \mathcal{L}_{Cls} + \alpha \mathcal{L}_{Ord} + \beta \mathcal{L}_{Dis}, \tag{5}$$

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

2.4 FB SCORE ESTIMATION

After establishing the order of samples, network cannot predict the FB score yet because the order is independent on the reference base. We apply the maximum consistency (MC) rule (Lim et al., 2019) to compare the input face image with a set of reference images. Figure 5 shows the reference set for SCUT-FBP5500 dataset. These comparisons will find the most similar references to the input. Their precise labels will determine the label of input. Specifically, given a test image x and a reference r_n, $1 \leq n \leq N$, where N is the number of references, our trained comparator will predict one of three relative relations "$>$, $\approx$, $<$". Let s' be an estimate of the true score $s(x)$. Then, the consistency between the comparator result and the estimate is defined as

$$\phi(x, r_n, s') = [x > r_n][s' - s(r_n) > \theta] + [x \approx r_n][|s' - s(r_n)| \leq \theta] + [x < r_n][s' - s(r_n) < -\theta], \tag{6}$$

where $[\cdot]$ is the indicator function. The function $\phi(x, r_n, s')$ returns either 0 for an inconsistent case or 1 for a consistent case. To maximize the consistency with all references, we estimate the FB score of x by

$$\hat{s}_{MC}(x) = \arg\max_{s'} \sum_{n=1}^{N} \phi(x, r_n, s'). \tag{7}$$

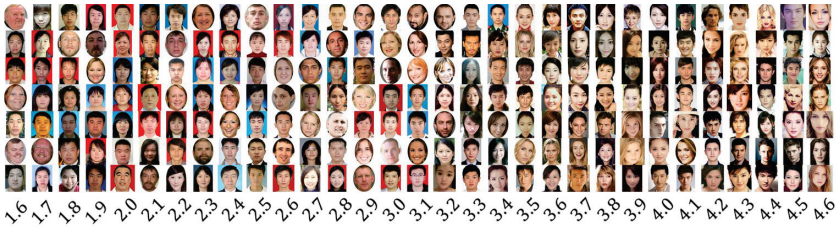


Figure 5: Reference set selected for estimating scores by MC rules (Lim et al., 2019).

3 EXPERIMENT

3.1 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 $0 \sim 1$ 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.

3.2 DATASETS AND EVALUATION METRICS

3.2.1 DATASETS

SCUT-FBP5500 (Liang et al., 2018b) has 5500 frontal facial images, which was scored by 60 volunteers among the range of 1$\sim$5. 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 (Gray et al., 2010) contains 2056 frontal female face images aged 18-40 without constraint on race, lighting, pose or expression. The data was scored by 30 volunteers among the range of -3$\sim$3.

MEBeauty (Lebedeva et al., 2021) 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∼10.

Color FERET (Phillips et al., 2000) 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) (Chen et al., 2017) 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, UOL is only trained on the training set of SCUT-FBP5500, and evaluated on all five datasets without any fine-tuning. We do this for testing the generalization ability of UOL.

3.2.2 Evaluation Metrics

To test the effectiveness of UOL, we follow the evaluation metrics in previous metheds (Lin et al., 2019; Xie et al., 2015; Shi et al., 2019; Lin et al., 2022; Xu et al., 2018; Bougourzi et al., 2022): mean absolute error (MAE) and root mean square error (RMSE).

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 (please refer to Appendix D). The Pearson correlation coefficient (PC) (Benesty et al., 2009) is commonly used in psychological studies of FB for generalization testing. Thus, PC is employed as a metric in this work. 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.

3.3 Comparison with SOTA Methods

To verify the performance and generalization ability of UOL, we compare it with the state-of-the-art methods R^3CNN (Lin et al., 2022), CRNet (Xie et al., 2015), Co-attention (Shi et al., 2019), AaNet (Lin et al., 2019), ComboLoss (Xu et al., 2018) and CNN-ER (Bougourzi et al., 2022).

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

	R^3CNN	CRNet	AaNet	ComboLoss	Co-attention	CNN-ER	**UOL**
PC ↑	0.9142	0.8869	0.9055	0.9199	**0.9260**	0.9250	0.9240
MAE ↓	0.2120	0.2397	0.2236	0.2050	0.2020	0.2009	**0.1975**
RMSE ↓	0.2800	0.3186	0.2954	0.2704	0.2660	0.2650	**0.2633**

3.3.1 Performance Evaluation on SCUT-FBP5500

We first test all methods on the large-scale dataset SCUT-FBP5500 and report the results 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.

3.3.2 Generalization Capability Evaluation

Our method aims to improve the generalization ability of the model by mimicking human cognition. To this end, we train all methods on SCUT-FBP5500 and then test them on unseen datasets (Hot-Or-Not, MEBeauty, Color FERET, MIFS). 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∼5 and compute PC with the estimated scores by each method. Table 2 shows 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 that Co-attention performs worse than its performance on SCUT-FBP5500, which shows it is sensitive to test data when FB standard shifts or image quality varies.

Table 2: Performance of different models on the Hot-Or-Not and MEBeauty.

Methods	Hot-Or-Not			MEBeauty		
	PC ↑	MAE ↓	RMSE ↓	PC ↑	MAE ↓	RMSE ↓
R^3CNN	0.3555	0.5741	0.7140	0.5039	0.5329	0.6691
CRNet	0.3250	0.5811	0.7294	0.4380	0.5645	0.9019
AaNet	0.2893	0.5923	0.7399	0.3746	0.6102	0.7548
ComboLoss	0.3329	0.6154	0.7677	0.5078	0.5481	0.6888
Co-attention	0.2697	0.5613	0.7080	0.4976	0.5476	0.6907
CNN-ER	0.3513	**0.5269**	**0.6653**	0.4911	0.5753	0.6973
UOL	**0.4073**	0.5410	0.6779	**0.5532**	**0.5230**	**0.6489**

(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 seven ratings for each image, the lowest and highest ones are removed, and the mean of the remaining ratings is considered as FB score. Afterward, we do the same process as above and list the results in Table 3. UOL also achieves the highest PC.

Table 3: Comparison of the generalization ability of seven models on Color FERET and the accuracies on MIFS.

	R^3CNN	CRNet	AaNet	ComboLoss	Co-attention	CNN-ER	**UOL**
PC ↑ (Color FERET)	0.8265	0.8564	0.7504	0.9146	0.7067	0.8490	**0.9266**
Acc(%) ↑ (MIFS)	73.91	**96.15**	76.92	92.31	76.92	**96.15**	92.31

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 the volunteers' results. Please refer to Appendix E for the detailed data cleaning process. The volunteers' results are used as the ground truth. We apply each method to images in a pair. If the estimated score of the image with makeup is higher than the one without makeup, the comparison is correct, otherwise incorrect. Table 3 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, in other experiments.

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

3.4 Ablation Studies

3.4.1 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. All versions are trained on the SCUT-FBP5500. The backbone is VGG16. The results under different settings are listed in Table 4. 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

Table 4: The effectiveness evaluation of uncertainty modeling and order learning on SCUT-FBP5500 (**SCUT**), Hot-Or-Not (**HON**) and MEBeauty (**MEB**).

Methods	SCUT			HON	MEB
	PC↑	MAE↓	RMSE ↓	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 5: The effectiveness evaluation of three loss functions on Hot-Or-Not and MEBeauty.

CE	Hinge	KL	Hot-Or-Not			MEBeauty		
			PC ↑	MAE ↓	RMSE ↓	PC ↑	MAE ↓	RMSE ↓
✓			0.4007	0.5419	0.6844	0.5457	0.5252	0.6543
✓	✓		0.4036	0.5410	0.6793	0.5463	0.5230	0.6508
✓		✓	0.4036	0.5426	0.6792	0.5397	0.5290	0.6592
✓	✓	✓	**0.4073**	**0.5410**	**0.6779**	**0.5532**	**0.5230**	**0.6489**

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 4. 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.

3.4.2 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 5. 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.

4 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. 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

Parham Aarabi, Dominic Hughes, Keyvan Mohajer, and Majid 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)*, volume 4, pp. 2644–2647. IEEE, 2001.

Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu. Examples-rules guided deep neural network for makeup recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Hani Altwaijry and Serge Belongie. Relative ranking of facial attractiveness. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 117–124. IEEE, 2013.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.

Andrea Bottino and Aldo Laurentini. The intrinsic dimensionality of attractiveness: A study in face profiles. In *Iberoamerican Congress on Pattern Recognition*, pp. 59–66. Springer, 2012.

F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed. Deep learning based face beauty prediction via dynamic robust losses and ensemble regression. *Knowledge-Based Systems*, 242:108246, 2022.

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5710–5719, 2020.

Cunjian Chen, Antitza Dantcheva, Thomas Swearingen, and Arun Ross. Spoofing faces using makeup: An investigative study. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pp. 1–8. IEEE, 2017.

Yael Eisenthal, Gideon Dror, and Eytan Ruppin. Facial attractiveness: Beauty and the machine. *Neural computation*, 18(1):119–142, 2006.

Jintu Fan, KP Chau, Xianfu Wan, Lili Zhai, and Ethan Lau. Prediction of facial attractiveness from facial proportions. *Pattern Recognition*, 45(6):2326–2334, 2012.

Yang-Yu Fan, Shu Liu, Bo Li, Zhe Guo, Ashok Samal, Jun Wan, and Stan Z Li. Label distribution-based facial attractiveness computation by deep residual learning. *IEEE Transactions on Multimedia*, 20(8):2196–2208, 2017.

Leslie G Farkas and Gwynne Cheung. Facial asymmetry in healthy north american caucasians: an anthropometrical study. *The Angle Orthodontist*, 51(1):70–77, 1981.

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 39(9):1853–1865, 2017.

Junying Gan, Lichen Li, Yikui Zhai, and Yinhua Liu. Deep self-taught learning for facial beauty prediction. *Neurocomputing*, 144:295–303, 2014.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3369–3378, 2018.

Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong. Predicting facial beauty without landmarks. In *European Conference on Computer Vision*, pp. 434–447. Springer, 2010.

Hatice Gunes and Massimo Piccardi. Assessing facial beauty through proportion analysis by image processing and supervised learning. *International journal of human-computer studies*, 64(12): 1184–1199, 2006.

Amit Kagian, Gideon Dror, Tommer Leyvand, Daniel Cohen-Or, and Eytan Ruppin. A humanlike predictor of facial attractiveness. *Advances in Neural Information Processing Systems*, 19, 2006.

Aldo Laurentini and Andrea Bottino. Computer analysis of face beauty: A survey. *Computer Vision and Image Understanding*, 125:184–199, 2014.

Irina Lebedeva, Yi Guo, and Fangli Ying. Mebeauty: a multi-ethnic facial beauty dataset in-the-wild. *Neural Computing and Applications*, pp. 1–15, 2021.

Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q Weinberger. On feature normalization and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12383–12392, 2021a.

Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13936–13944, 2020.

Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13896–13905, 2021b.

Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pp. 689–698, 2018a.

Lingyu Liang, Lianwen Jin, and Xuelong Li. Facial skin beautification using adaptive region-aware masks. *IEEE transactions on cybernetics*, 44(12):2600–2612, 2014.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. Scut-fbp5500: a diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 1598–1603. IEEE, 2018b.

Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *International Conference on Learning Representations*, 2019.

Luojun Lin, Lingyu Liang, Lianwen Jin, and Weijie Chen. Attribute-aware convolutional neural networks for facial beauty prediction. In *IJCAI*, pp. 847–853, 2019.

Luojun Lin, Lingyu Liang, and Lianwen Jin. Regression guided by relative ranking using convolutional neural network (r3cnn) for facial beauty prediction. *IEEE Transactions on Affective Computing*, 13(1):122–134, 2022.

Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan. Wow! you are so beautiful today! *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):1–22, 2014.

Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5301–5309, 2019.

Huiyun Mao, Lianwen Jin, and Minghui Du. Automatic classification of chinese female facial beauty using support vector machine. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pp. 4842–4846. IEEE, 2009.

Yadong Mu. Computational facial attractiveness prediction by aesthetics-aware features. *Neurocomputing*, 99:59–64, 2013.

Xinyu Ou, Si Liu, Xiaochun Cao, and Hefei Ling. Beauty emakeup: A deep makeup transfer system. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 701–702, 2016.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

Thomas L Saaty. A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, 15(3):234–281, 1977.

Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. Computer-suggested facial makeup. In *Computer Graphics Forum*, volume 30, pp. 485–492. Wiley Online Library, 2011.

Kendra Schmid, David Marx, and Ashok Samal. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition*, 41(8):2710–2717, 2008.

Shengjie Shi, Fei Gao, Xuantong Meng, Xingxin Xu, and Jingjie Zhu. Improving facial attractiveness prediction via co-attention learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4045–4049. IEEE, 2019.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69 (1):99–118, 1955.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Davor Sutić, Ivan Brešković, René Huić, and Ivan Jukić. Automatic evaluation of facial attractiveness. In *The 33rd International Convention MIPRO*, pp. 1339–1342. IEEE, 2010.

Louis L Thurstone. Psychophysical analysis. *The American journal of psychology*, 38(3):368–389, 1927a.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927b.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927c.

Jing Wang and Xin Geng. Label distribution learning by exploiting label distribution manifold. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2021.

Shuyang Wang, Ming Shao, and Yun Fu. Attractive or not? beauty prediction with attractiveness-aware encoders and robust late fusion. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 805–808, 2014.

Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019.

R White, A Eden, and M Maire. Automatic prediction of human attractiveness. *UC Berkeley CS280A Project*, 1(2), 2004.

Jacob Whitehill and Javier R Movellan. Personalized facial attractiveness prediction. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–7. IEEE, 2008.

Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. Scut-fbp: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1821–1826. IEEE, 2015.

Lu Xu and Jinhai Xiang. Comboloss for facial attractiveness analysis with squeeze-and-excitation networks. *arXiv preprint arXiv:2010.10721*, 2020.

Lu Xu, Jinhai Xiang, and Xiaohui Yuan. Crnet: Classification and regression neural network for facial beauty prediction. In *Pacific Rim Conference on Multimedia*, pp. 661–671. Springer, 2018.

Haibin Yan. Cost-sensitive ordinal regression for fully automatic facial beauty assessment. *Neurocomputing*, 129:334–342, 2014.

David Zhang, Fangmei Chen, Yong Xu, et al. *Computer models for facial beauty analysis*. Springer, 2016.

A THURSTONE'S THEORY

Our UOL is inspired by the **discriminal dispersion** theory from Thurstone's *Law of Comparative Judgment*. The following are the details that mainly come from Thurstone's papers *Psychophysical Analysis* (Thurstone, 1927a) and *A Law of Comparative Judgment* (Thurstone, 1927b).

A.1 DISCRIMINAL DISPERSION

Thurstone argues that a term is needed for the process by which the organism identifies, distinguishes, discriminates, or reacts to stimuli. In order to avoid any implications, they call the psychological values of psychophysics discriminal processes. In Fig. 6(a), let the circles $R_1, R_2, R_3, \cdots, R_n$ represent a series of stimuli which constitutes a scale with regard to any prescribed stimulus attribute. Psychologically some of these attributes can be measured, while physically the measurement may even be impossible. Thurstone assumes that a series of stimuli have been arranged in a scale according to any attribute about which one can say "more" or "less" and that psychophysics need not be limited to stimuli which have magnitude or size, such as lifted weights and room size. Suppose that each stimulus in the series has a discriminal process which is a psychic or physiological function of the organism. Thus, the stimulus R_n has a discriminal process S_n.

It may be assumed that this relation is not so fixed as might be indicated by Fig. 6(a). It undoubtedly happens that stimulus R_n does not always produce the same discriminal process S_n, shown in Fig. 6(b). The present method of psychophysical analysis rests on the assumption that constant and repeated stimuli are not always associated with exactly the same discriminal process but that there is some qualitative fluctuation from one occasion to the next in this process for a given stimulus. It should be recalled that each of these processes or qualities is identified by that stimulus which most frequently produces it so that S_n is habitually associated with R_n.

A given process S_5 would be associated frequently with R_5 but occasionally also with adjacent and closely similar stimuli in the stimulus continuum such as R_3, R_4, R_5, R_6. Thurstone's research focuses on the second case, namely the qualitative fluctuations in the discriminal processes that are associated with a constant and repeated stimulus.

The psychophysical relations may be summarized, so far, in the following propositions.

(1) A series of stimuli $R_1, R_2, R_3, \cdots, R_n$ can be arranged in a scale, with reference to any prescribed quantitative or qualitative stimulus attribute.

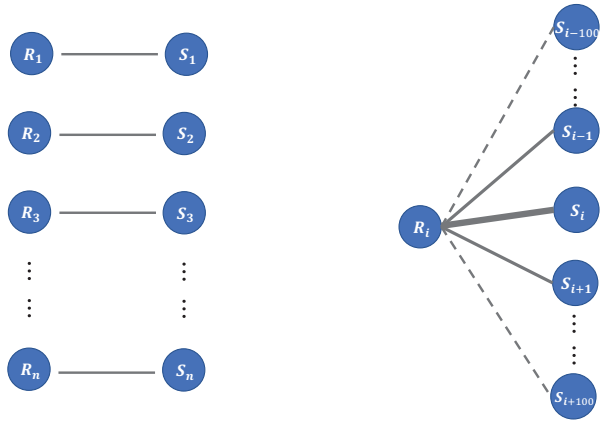
(2) These stimuli are differentiated by processes of the organism of unknown nature and they are designated $S_1, S_2, S_3, \cdots, S_n$, respectively. Every stimulus R_n is identified by the organism with the process S_n. In this discussion they are referred to as the discriminal processes or qualities.

(3) It is assumed that the correspondence $R_n \longrightarrow S_n$ is subject to noticeable fluctuation so that R_n, does not always produce the exact process S_n, but sometimes nearly similar processes S_{n+1} or S_{n-1} and sometimes even S_{n+2} or S_{n-2}. This fluctuation among the discriminal processes for a uniform repeated stimulus will be designated the **discriminal dispersion**.

The relative frequencies of the different processes are shown in Fig. 6(b) for stimulus R_i in a rough diagrammatic way. Thus, there is a thick line connecting R_i with S_i to indicate the relation between the stimulus and its modal discriminal process. There are only thin lines connecting the adjacent processes with the same stimulus R_i and this represents the relatively lower frequency of this association.

Since the assumption of a normal distribution for the discriminal dispersion can be experimentally verified and limited to those stimulus series where its reality can be tested, it will be reasonable to make this assumption subject to verification in every case. The hypothesis can be stated as follows and shown in Fig. 7. **The discriminal dispersion which any given repeated stimulus produces on the psychological continuum is usually normal. The frequencies with which the discriminal processes occur for a given stimulus ordinarily describe a normal distribution when plotted on the psychological continuum as a base.**

According to this hypothesis, assign scale values to the various processes as distances from S_n as an origin. These distances would be assigned in terms of the standard deviation of the distribution of process-frequencies. Psychological measurement depends, then, on the adoption of one of



(a) Fixed discriminal processes.

(b) Relative frequencies of the different processes.

Figure 6: Discriminal process for stimuli.

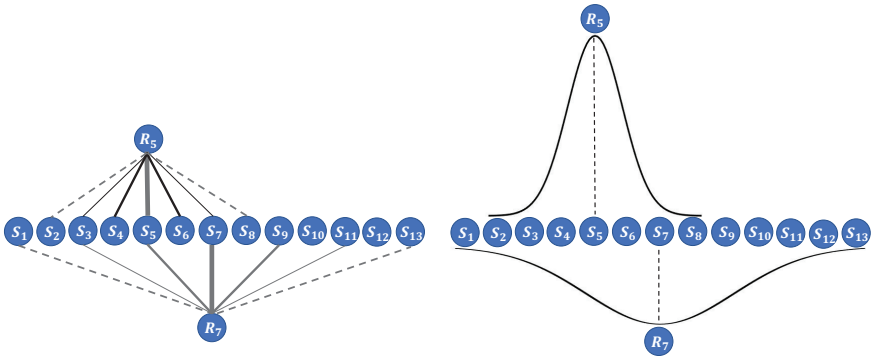


Figure 7: Thurstone models the fluctuation among the discriminal processes for a uniform repeated stimulus (left) as a normal distribution (right) by the name of "discriminal dispersion".

these dispersions as a base, and the use of its standard deviation as a unit of measurement for the psychological continuum under investigation.

A.2 Law of Comparative Judgment

An ambiguous stimulus which is observed at widely different degrees of excellence or beauty on different occasions will have of course a large discriminal dispersion. Some other stimulus or specimen which is provocative of relatively slight fluctuations in discriminal processes will have, similarly, a small discriminal dispersion.

The scale difference between the discriminal processes of two specimens which are involved in the same judgment will be called the *discriminal difference* on that occasion. If the two stimuli be denoted A and B and if the discriminal processes corresponding to them be denoted a and b on any one occasion, then the discriminal difference will be the scale distance $(a-b)$ which varies of course on different occasions. If, in one of the comparative judgments, A seems to be better than B, then, on that occasion, the discriminal difference $(a-b)$ is positive. If, on another occasion, the stimulus B seems to be better than A, then, on that occasion, the discriminal difference $(a-b)$ is negative.

Finally, the scale distance between the modal discriminal process for any two specimens is the separation which is assigned to the two specimens on the scale that their separation is equal to the separation between their respective modal discriminal processes.

Thurstone states the law of comparative judgment as follows:

$$V_1 - V_2 = x_{(1>2)} \cdot \sqrt{\sigma_1^2 + \sigma_2^2 - 2\gamma\sigma_1\sigma_2}, \tag{8}$$

in which V_1 and V_2 are the psychological scale values of the two compared stimuli; $x_{(1>2)}$ is the sigma value corresponding to the proportion of judgments $p_{(1>2)}$, when $p_{(1>2)}$ is greater than 0.5 the numerical value of $x_{(1>2)}$ is positive, otherwise $x_{(1>2)}$ is negative; σ_1 and σ_2 are the discriminal dispersion of stimuli R_1 and R_2; γ is the correlation between the discriminal deviations of R_1 and R_2 in the same judgment.

B Another Perspective of Uncertainty Modeling

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 (Li et al., 2021a; Wang et al., 2019). 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 (Li et al., 2021a; Wang et al., 2019). Our order learning just is a triple classification problem. As a slight feature disturbance does not change the relative relation between instances, feature-level data augmentation is suitable for order learning. In addition, our method uses the mean of multiple comparison results, which statistically lower the impact of extreme augmentations. Moreover, the pairwise comparisons in a batch, which compare any two features in the same batch, can be considered as another form of feature-level augmentation. Therefore, our uncertainty modeling in order learning can achieve better generalization.

C Training Strategy in Methods

C.1 Triplet Construction for Hinge Loss

To model the order on psychological scale space described in **Sec.3.3** of the paper, 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 $i\text{–}th$ instance as the anchor l, then set the instance with index $(i+1)\ mod\ M$ as the second element m. Finally, we select an instance n from the remaining $M-2$ instances as the third element, which satisfies:

$$\arg\min_{n \neq l,m} \left| |y_l - y_m| - |y_l - y_n| \right|, \ \textbf{where} \ |y_l - y_m| \neq |y_l - y_n| \,. \tag{9}$$

Algorithm 1 describes the strategy of this hard triplets selection.

C.2 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

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, ($l = \{l_1, l_2, \ldots, l_M\}$, $m = \{m_1, m_2, \ldots, m_M\}$, $n = \{n_1, n_2, \ldots, n_M\}$).

for $i \leftarrow 1$ **to** M **do**
  $l_i \leftarrow i$;
  $m_i \leftarrow (i+1) \mod M$;
  $dis_{12} \leftarrow |y_{l_i} - y_{m_i}|$;
  $sub = +\infty$;
  for $j \leftarrow 1$ **to** M **do**
    if $j \neq l_i$ **and** $j \neq m_i$ **then**
      $dis_{13} \leftarrow |y_{l_i} - y_j|$;
      $tmp \leftarrow |dis_{12} - dis_{13}|$;
      if $tmp < sub$ **and** $tmp \neq 0$ **then**
        $sub \leftarrow tmp$;
        $n_i \leftarrow j$;
      end
    end
  end
end

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.

D Evaluation Metrics

In the paper, we use Pearson correlation coefficient (PC) as the main evaluation metric due to the following reasons:

(1) PC measures the correlation between prediction and ground truth, which quantifies the degree of interdependence of prediction and ground truth. PC is a well-accepted metric for the evaluation of FB in psychological research.

(2) MAE measures the mean of absolute errors between the predictions and ground truth. RMSE measures the deviation between the predictions and ground truth. Lower MAE and RMSE do not guarantee a better correlation of predictions and ground truth.

(3) Hot-Or-Not, MEBeauty and SCUT-FBP5500 datasets have different score ranges, and were rated by different people. Thus, the score normalization does not guarantee that the same score represents the same FB level across different datasets. Instead, the order of ratings from different people appears more consistent. Thus, we employ PC as a metric to measure the generalization ability of a model trained on SCUT-FBP5500. It also inspires us to improve the generalization ability of the model by learning the order relationship.

E MIFS Dataset Cleaning Process

MIFS has no human rating of FB, but each ID contains at least one facial image with makeup and one without makeup in the dataset. Commonly, face wearing makeup is more attractive than the one wearing no makeup. Therefore, MIFS dataset is used to evaluate the generalization ability of all models after data cleaning.

MIFS cleaning process is 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.

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.

Algorithm 2: Balanced Pairs Selection

Input : FB scores of M (batch size) instances $y = \{y_1, y_2, \ldots, y_M\}$; adjacency list $flag = \{flag[1], flag[2], \ldots, flag[M]\}$, 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 θ.

Output: Pairs of index p, and corresponding order relationships Y.

$p \leftarrow \{\}$;
$Y \leftarrow \{\}$;
for $i \leftarrow 1$ **to** M **do**
  $candidates \leftarrow \{1, 2, \ldots, M\}$;
  // $candidates$ 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($candidates, flag[i][j]$);
  end
  for $j \leftarrow 1$ **to** M **do**
    if $flag[i] > N$ **then**
      DELETE($candidates, j$);
    end
  end
  while $flag[i] < N$ **and** $candidates \neq \{\}$ **do**
    $sim \leftarrow 0$;
    for $j \leftarrow 1$ **to** *LEN*($candidates$) **do**
      $dif \leftarrow |y_i - candidates[j]|$;
      if $dif < \theta$ **then**
        $sim \leftarrow sim + 1$;
        // sim is the number of ``≈" pairs in $candidates$.
      end
    end
    $unsim \leftarrow LEN(candidates) - sim$;
    $prob = \{\}$;
    // $prob$ is the list denotes probability of every instance to be selected.
    for $j \leftarrow 1$ **to** *LEN*($candidates$) **do**
      $dif \leftarrow |y_i - candidates[j]|$;
      if $dif < \theta$ **then**
        INSERT($prob, \frac{1}{3*sim}$);
        // INSERT denotes appending an element to a list.
      else
        INSERT($prob, \frac{2}{3*unsim}$);
      end
    end
    $r \leftarrow$ RANDOM_CHOICE_BY_PROB($candidates, prob$);
    // select an instance's index r from $candidates$ by their probabilities $prob$, we implement this by using numpy.
    INSERT($flag[i], r$);
    INSERT($flag[r], i$);
    INSERT($p, \{i, r\}$);
    $order \leftarrow$ GENLABEL($|y_i - y_r|, \theta$);
    // GENLABEL denotes generating the order label for a pair by their ground truth and threshold.
    INSERT($Y, order$);
    DELETE($candidates, r$);
  end
end