Learning Probabilistic Ordinal Embeddings for Uncertainty-Aware Regression

Wanhua Li[1,2], Xiaoke Huang[1,2,3], Jiwen Lu[1,2,*], Jianjiang Feng[1,2], Jie Zhou[1,2]

[1]Department of Automation, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology, China
[3]School of Artificial Intelligence, Beijing Normal University

li-wh17@mails.tsinghua.edu.cn

xiaokehuang@mail.bnu.edu.cn   {lujiwen,jfeng,jzhou}@tsinghua.edu.cn

Abstract

Uncertainty is the only certainty there is. Modeling data uncertainty is essential for regression, especially in unconstrained settings. Traditionally the direct regression formulation is considered and the uncertainty is modeled by modifying the output space to a certain family of probabilistic distributions. On the other hand, classification based regression and ranking based solutions are more popular in practice while the direct regression methods suffer from the limited performance. How to model the uncertainty within the present-day technologies for regression remains an open issue. In this paper, we propose to learn probabilistic ordinal embeddings which represent each data as a multivariate Gaussian distribution rather than a deterministic point in the latent space. An ordinal distribution constraint is proposed to exploit the ordinal nature of regression. Our probabilistic ordinal embeddings can be integrated into popular regression approaches and empower them with the ability of uncertainty estimation. Experimental results show that our approach achieves competitive performance. Code is available at https://github.com/Li-Wanhua/POEs.

1. Introduction

Regression, as a fundamental machine learning problem, requires to predict a continuous target value y for a given data $\boldsymbol{x}$. It is also extensively studied due to many important applications, such as age estimation [32, 43, 56], historical image dating [37, 41], and image aesthetic assessment [25, 29, 42]. Recent years have witnessed the enormous success [31, 33, 44] of Deep Neural Networks (DNNs). Therefore, dominant approaches [27, 42, 56] for regression employ DNNs to leverage the powerful feature representations. As the most natural choice for regression, the *direct regression* method predicts a scalar value y given an input

---

* Corresponding author
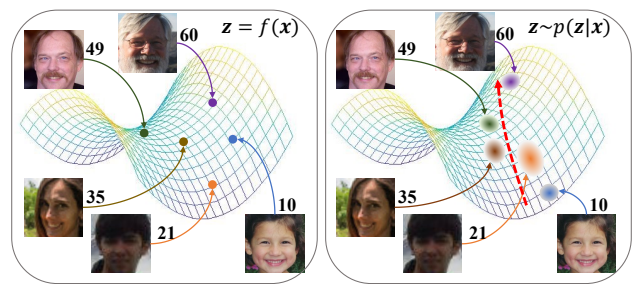
(a) Deterministic Unordered Embeddings   (b) Probabilistic Ordinal Embeddings

Figure 1. The key difference between deterministic unordered embeddings and probabilistic ordinal embeddings. We consider the problem of age estimation and display age labels next to the images. Deterministic unordered embeddings represent each facial image as a point in the latent space without considering the data uncertainty and the ordinal property. By contrast, probabilistic ordinal embeddings learn a distribution following the ordinal constraint in the embedding space. We see that the highly blurred image exhibits high uncertainty and the ordinal relations are preserved with our probabilistic ordinal embeddings.

$\boldsymbol{x}$ and the model is trained with L^1 or L^2 loss.

Modeling data uncertainty is quite important for regression since it gives a clear probabilistic interpretation of the predictions. In fact, humans usually associate confidence in their judgments. For example, if a person is required to estimate the age of a highly blurred facial image, he will give an estimation with high uncertainty. However, DNNs do not necessarily grasp the data uncertainty which captures the noise inherent in the data [36]. Many recent works [12, 20, 36, 55] have been proposed to model the uncertainty, which consider the *direct regression* solutions and replace the point estimation with a probabilistic distribution $p(y|\boldsymbol{x})$ in the output space. For simplicity, they usually consider a certain family of distributions, such as Gaussian or Laplace, which naturally limits the expressiveness.

Meanwhile, many advanced methods [43, 57, 58] have been presented for regression, which significantly surpass the *direct regression* approach. Some of the dominant ap-

proaches include classification based methods [46, 47] and ranking based methods [5, 7, 38]. Classification based method treats the regression problem as a multi-class classification problem by discretizing the target space into C classes. Ranking based methods transform the regression problem into multiple binary classification sub-problems, where each binary classifier aims to predict whether the rank of a sample is larger than a specific value. Modeling data uncertainty within these popular regression methods is quite challenging. Unlike the *direct regression* method which outputs a scalar value, both classification based regression and ranking based approaches have multiple neurons in the output layer. These neurons are highly correlated so it is difficult to choose a suitable distribution to replace them, which makes the above uncertainty-aware strategy unsuitable for these advanced regression methods.

To model the data uncertainty within the present-day regression methods, we propose Probabilistic Ordinal Embeddings (POEs), which estimate a multivariate Gaussian distribution, instead of a fixed point in the latent space. Most existing methods represent each input data as a deterministic point in the embedding space, *i,e.*, $z = f_\theta(x)$ is a point in $\mathbb{R}^D$, where $z \in \mathbb{R}^D$ is the embedding, f is the mapping function and θ is the parameter of f. However, it is difficult to give an accurate point embedding for noisy data, in which larger uncertainty should be exhibited in the embedding space [4, 51]. We consider modeling the data uncertainty in the embedding space and representing each embedding as a random variable: $z \sim p(z|x) \in \mathbb{R}^D$. Meanwhile, the ordinal relation in the target space should be preserved in the embedding space since the uncertainty mainly spreads the probability mass across neighbors. Therefore, an ordinal distribution constraint is further proposed to enforce the ordinal property. Figure 1 shows the key difference between deterministic unordered embeddings and probabilistic ordinal embeddings.

In summary, the main contributions of our paper are threefold: (1) To model the data uncertainty of regression, we propose probabilistic ordinal embeddings, which represent each data as a distribution following the ordinal constraint in the latent space. (2) Our method empowers the present-day regression methods with the ability of uncertainty estimation. A per-exemplar uncertainty score can be used to measure the reliability of predictions. (3) Comprehensive experiments illustrate that uncertainty modeling improves the performance and achieves very competitive performance on three real-world visual tasks.

2. Related Work

Uncertainty Learning: DNNs tend to give overconfident predictions [24, 28] and learning to quantify the predictive uncertainty is essential in many real-world applications [9, 21, 40, 54]. Uncertainty learning for computer vision tasks can be categorized into regression settings such as age estimation [36], and classification settings such as semantic segmentation [19]. There are two major types of uncertainty: *epistemic (model) uncertainty* and *aleatoric (data) uncertainty* [20]. Data uncertainty refers to the noise inherent in the input data which cannot be eliminated with more training data. Our work mainly focuses on data uncertainty. It is usually modeled by placing a distribution over a model's output [20]. Specifically, the direct regression formulation is considered for regression and the scalar output is corrupted with Gaussian random noise. However, it is not suitable for present-day regression methods which usually have multiple neurons in the output layers [7, 38, 46]. Placing a multivariate Gaussian distribution over the vector in output space cannot model the complex relations among the elements in the output vector. Recently, some methods [4, 39, 51] have proposed stochastic embeddings to model the data uncertainty for classification problems. However, these methods are not suitable for regression problems since they fail to model the ordinal nature of regression. In this paper, we propose to learn a distribution following the ordinal constraint in the latent space.

Regression Methods: Supervised regression aims to learn a mapping from the input $x \in \mathcal{X}$ to a continuous target value $y \in \mathcal{Y}$. A variety of methods [25, 32, 41, 43, 47] have been proposed for regression in many computer vision tasks. Here we describe some popular methods. The most natural solution is direct regression, which trains a DNN with a single output neuron to predict the target value. However, the direct regression methods suffer from limited performance and are not very popular in practice [7, 47]. Recently, some classification based methods [47, 57] have been proposed, which divide the output space into several bins and treat each bin as an independent class. Some variants of this method, such as mean-variance loss [43] and label distribution learning [11, 13], have achieved great success in many fields [6, 17, 52]. Another dominant approach is the ranking based method [34, 38], which constructs multiple binary classifiers and trains them with rank labels. An ordinal constraint was used in [18], but it failed to incorporate the data uncertainty on the embedding space. More recent works formulate the regression problem from a probabilistic perspective [15, 20]. For example, He *et al.* [16] placed a Gaussian distribution over the output to model the uncertainty. Mixture Dense Regression [55] further replaced the output with a mixture of Gaussian. While these methods are only compatible with the direct regression methods, we are committed to modeling data uncertainty within dominant regression approaches.

3. Proposed Approach

In this section, we first revisit the formulations of popular regression methods with deterministic embeddings. Then
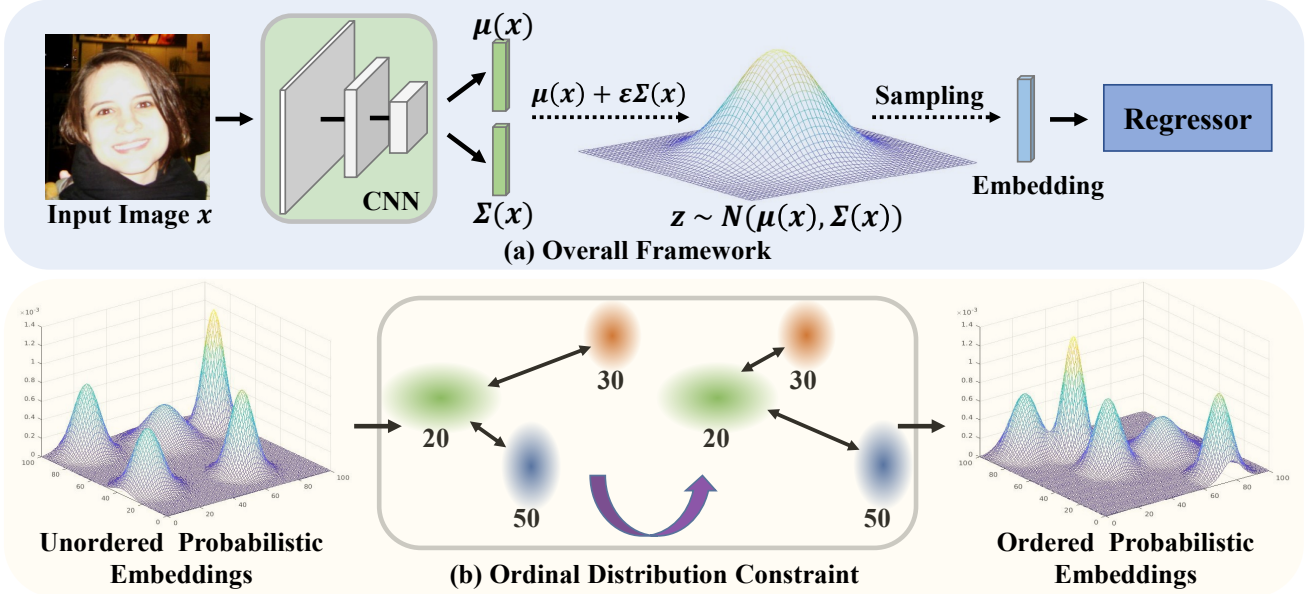
Figure 2. (a) The overall framework of our approach. Given an input image $\boldsymbol{x}$, we first obtain the parameters $\boldsymbol{\mu}(\boldsymbol{x})$ and $\boldsymbol{\Sigma}(\boldsymbol{x})$, which correspond to the mean and diagonal covariance of a multivariate Gaussian distribution respectively. Then we sample embeddings from this distribution with the reparameterization trick. Lastly, the sampled embeddings are sent to a popular regressor for prediction. (b) The idea of ordinal distribution constraint. For three samples with labels of 20, 30, and 50, we enforce the ordinal relations in the embedding space with the symmetrized Kullback-Leibler divergence or 2-Wasserstein distance. In the end, we expect that ordered probabilistic embeddings are obtained by applying the ordinal distribution constraint in the latent space.

we introduce the details of probabilistic embeddings. Lastly, we present the ordinal distribution constraint and show how to integrate the proposed probabilistic ordinal embeddings into different regression methods.

3.1. Preliminaries

Direct Regression Solution: Direct regression method employs learnable parameters $\boldsymbol{w} \in \mathbb{R}^{D \times 1}$ to project the point embedding $\boldsymbol{z} = f_\theta(\boldsymbol{x})$ into a scalar value $\hat{y}$:

$$\hat{y} = \boldsymbol{w}^\top f_\theta(\boldsymbol{x}), \tag{1}$$

where $\cdot^\top$ denotes transposition. Then the model is usually trained with L^1 or L^2 loss. We present the formulation with L^2 loss as follows:

$$\mathcal{L}_{DR}(\boldsymbol{x}) = \left\| y - \boldsymbol{w}^\top f_\theta(\boldsymbol{x}) \right\|^2. \tag{2}$$

Regression by Classification: Many methods discretizing the target space into C classes and treat regression as a classification problem. We obtain the predictions with a classification layer parameterized by $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_C]^\top \in \mathbb{R}^{K \times D}$. Cross entropy loss is utilized to optimize the networks. Specifically, for a sample $\boldsymbol{x}$ with the ground truth c, the loss is defined as follows:

$$\mathcal{L}_{Cls}(\boldsymbol{x}) = -\log p(y = c|\boldsymbol{x}; \theta, \boldsymbol{W})$$
$$= -\log \frac{\exp(\boldsymbol{w}_c^\top f_\theta(\boldsymbol{x}))}{\sum_{r=1}^{C} \exp(\boldsymbol{w}_r^\top f_\theta(\boldsymbol{x}))}. \tag{3}$$

Several strategies can be employed to make predictions from the classification results [46]. A simple choice is to

use the class with the highest probability:

$$\hat{y} = \underset{r \in \{1,2,...,C\}}{\arg\max} \; p(y = r|\boldsymbol{x}; \theta, \boldsymbol{W}). \tag{4}$$

Ranking based Regression: Classification based methods treat different classes independently, which ignores the ordinal relations of different classes. Ranking based methods construct $C - 1$ binary classifiers for C classes, where k-th one predicts whether the rank of the sample $\boldsymbol{x}$ is larger than k. The $C - 1$ labels are constructed as $r_1, r_2, ..., r_{C-1} \in \{0, 1\}$ for a given input $\boldsymbol{x}$. Then all $C-1$ classifiers are jointly trained with the cross-entropy loss:

$$\mathcal{L}_{Rank}(\boldsymbol{x}) = -\sum_{k=1}^{C-1} \log p(b_k = r_k|\boldsymbol{x}; \theta, \boldsymbol{W}_k)$$
$$= -\sum_{k=1}^{C-1} \log \frac{\exp(\boldsymbol{w}_{k,r_k}^\top f_\theta(\boldsymbol{x}))}{\sum_{i=0}^{1} \exp(\boldsymbol{w}_{k,i}^\top f_\theta(\boldsymbol{x}))}, \tag{5}$$

where b_k is the output of the k-th binary classifier, and $\boldsymbol{W}_k = [\boldsymbol{w}_{k,0}, \boldsymbol{w}_{k,1}]^\top \in \mathbb{R}^{2 \times D}$ is the parameters of the k-th binary classifier. The decoding strategy is defined as:

$$\hat{y} = 1 + \sum_{k=1}^{C-1} b_k. \tag{6}$$

3.2. Probabilistic Embeddings for Regression

To model the data uncertainty for different regression methods, we consider treating embeddings as probabilistic

distributions, *i.e.,* $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x})$. Given an input $\boldsymbol{x}$, we can obtain a probabilistic prediction by marginalizing over the embedding distribution:

$$p(y|\boldsymbol{x}) = \int p(y|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{z}. \tag{7}$$

It is difficult to give a closed-form expression for the integral in (7). Therefore, we perform Monte-Carlo sampling from $\boldsymbol{z}^{(t)} \sim p(\boldsymbol{z}|\boldsymbol{x})$ to approximate it:

$$p(y|\boldsymbol{x}) \approx \frac{1}{T} \sum_{t=1}^{T} p(y|\boldsymbol{z}^{(t)}), \tag{8}$$

where T is the number of samples. We set T to 50 in our experiments for a good trade-off.

To sample from the distribution $p(\boldsymbol{z}|\boldsymbol{x})$, we define the embedding $\boldsymbol{z}$ as a multivariate Gaussian distribution for simplicity:

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x})), \tag{9}$$

where the mean $\boldsymbol{\mu}(\boldsymbol{x})$ and the diagonal covariance $\boldsymbol{\Sigma}(\boldsymbol{x})$ are input-dependent: $\boldsymbol{\mu}(\boldsymbol{x}) = f_{\theta_1}(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\Sigma}(\boldsymbol{x})) = f_{\theta_2}(\boldsymbol{x})$, and θ_1 and θ_2 refer to the corresponding model parameters. We implement $f_{\theta_1}(\boldsymbol{x})$ and $f_{\theta_2}(\boldsymbol{x})$ with a shared backbone and two head branches. The mean $\boldsymbol{\mu}(\boldsymbol{x})$ can be regarded as the most likely embedding values while the diagonal covariance $\boldsymbol{\Sigma}(\boldsymbol{x})$ can be interpreted as the data uncertainty. Naturally, larger variance means higher uncertainty. While sampling operation is not differentiable, we consider the reparameterization trick [23] to enable backpropagation:

$$\boldsymbol{z}^{(t)} = \boldsymbol{\mu}(\boldsymbol{x}) + \mathrm{diag}(\sqrt{\boldsymbol{\Sigma}(\boldsymbol{x})}) \cdot \boldsymbol{\epsilon}^{(t)}, \boldsymbol{\epsilon}^{(t)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{10}$$

Therefore, we first sample noise from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and then obtain $\boldsymbol{z}$ following (10) instead of directly sampling from $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$.

To integrate our method into existing regression approaches, we treat the $p(y|\boldsymbol{z})$ as the estimator with different formulations and rewrite the loss functions accordingly. For the direct regression solution, our method minimizes the following loss:

$$\mathcal{L}_{PDR}(\boldsymbol{x}) = \frac{1}{T} \sum_{t=1}^{T} \left\| y - \boldsymbol{w}^{\top} \boldsymbol{z}^{(t)} \right\|^2 \tag{11}$$

For classification based regression, the objective function is defined as follows:

$$\begin{aligned}\mathcal{L}_{PCls}(\boldsymbol{x}) &= -\frac{1}{T} \sum_{t=1}^{T} \log p(y = c|\boldsymbol{z}^{(t)}; \theta_1, \theta_2, \boldsymbol{W}) \\ &= -\frac{1}{T} \sum_{t=1}^{T} \log \frac{\exp(\boldsymbol{w}_c^{\top} \boldsymbol{z}^{(t)})}{\sum_{r=1}^{C} \exp(\boldsymbol{w}_r^{\top} \boldsymbol{z}^{(t)})}.\end{aligned} \tag{12}$$

Furthermore, the loss function for ranking based meth-

ods is given by:

$$\begin{aligned}\mathcal{L}_{PRank}(\boldsymbol{x}) &= -\frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{C-1} \log p(b_k = r_k|\boldsymbol{z}^{(t)}; \theta_1, \theta_2, \boldsymbol{W}_k) \\ &= -\frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{C-1} \log \frac{\exp(\boldsymbol{w}_{k,r_k}^{\top} \boldsymbol{z}^{(t)})}{\sum_{i=0}^{1} \exp(\boldsymbol{w}_{k,i}^{\top} \boldsymbol{z}^{(t)})}.\end{aligned} \tag{13}$$

In this way, we can train the present-day regression methods with probabilistic embeddings. We show the above pipeline in Figure 2(a). Note that the Monte-Carlo sampling is ONLY used during training since we observe no significant performance difference during inference.

3.3. Ordinal Distribution Constraint

The target space of regression is continuous and ordinal, which means that the uncertainty mainly spreads probability mass across the locations that correspond to the neighbors of ground truth value. For example, if we estimate the age of a person as 30, then his uncertainty in the embedding space depicts that he is around 30. Therefore, we expect that the ordinal relation in the target space can be preserved in the embedding space. We consider applying a set of ordinal distribution constraints to enforce the ordinal property in the embedding space. For a given triplet $(\boldsymbol{x}_l, \boldsymbol{x}_m, \boldsymbol{x}_n)$, and its associated label (y_l, y_m, y_n) and probabilistic embedding $(\boldsymbol{z}_l, \boldsymbol{z}_m, \boldsymbol{z}_n)$, we aim to preserve the following constraint:

$$d(\boldsymbol{z}_l, \boldsymbol{z}_m) < d(\boldsymbol{z}_l, \boldsymbol{z}_n) \Leftrightarrow |y_l - y_m| < |y_l - y_n|, \tag{14}$$

where $d(\cdot)$ is a distance metric. Note that the triplet is not the conventional triplet since the positive and negative samples are not fixed for a given anchor sample. More details about our triplet are provided in the supplementary material.

We let $\mathcal{S} = \{(l, m, n) \big| |y_l - y_m| < |y_l - y_n|\}$, then the ordinal distribution constraint is formulated with hinge loss:

$$\mathcal{L}_{Ord} = \frac{1}{|\mathcal{S}|} \sum_{(l,m,n) \in \mathcal{S}} \max(0, d(\boldsymbol{z}_l, \boldsymbol{z}_m) + \delta - d(\boldsymbol{z}_l, \boldsymbol{z}_n)), \tag{15}$$

where δ is the margin. Note that our ordinal distribution constraint in (15) is non-trivial, given that metric $d(\cdot)$ measures the distance between two distributions rather than that between two points. Therefore, some common choices for $d(\cdot)$, such as the Euclidean distance, are not suitable.

Kullback-Leibler divergence is commonly used as the metric to measure the difference between two distributions in many applications, such as label distribution learning [10, 11, 50]. However, it is an asymmetric measure and does not satisfy the triangle inequality. In this paper, we consider a symmetrized Kullback-Leibler divergence $SKL(P, Q) = KL(P\|Q) + KL(Q\|P)$ and ignore the triangle inequality issue. Furthermore, We can obtain the

closed-form expression for symmetrized Kullback-Leibler divergence between two multivariate Gaussian distributions. For example, the symmetrized KL divergence between $\boldsymbol{z}_l$ and $\boldsymbol{z}_m$ is given by:

$$
\begin{aligned}
SKL(\boldsymbol{z}_l, \boldsymbol{z}_m) &= KL\left(\boldsymbol{z}_l \| \boldsymbol{z}_m\right) + KL\left(\boldsymbol{z}_m \| \boldsymbol{z}_l\right) \\
&= \frac{1}{2} \sum_{j=1}^{D} \Big[\frac{(\boldsymbol{\sigma}_l^j)^2}{(\boldsymbol{\sigma}_m^j)^2} + \frac{(\boldsymbol{\sigma}_m^j)^2}{(\boldsymbol{\sigma}_l^j)^2} - 2 \\
&\quad + \frac{(\boldsymbol{\mu}_l^j - \boldsymbol{\mu}_m^j)^2}{(\boldsymbol{\sigma}_l^j)^2} + \frac{(\boldsymbol{\mu}_l^j - \boldsymbol{\mu}_m^j)^2}{(\boldsymbol{\sigma}_m^j)^2} \Big],
\end{aligned}
\tag{16}
$$

where $\boldsymbol{\mu}_l^j, \boldsymbol{\mu}_m^j, \boldsymbol{\sigma}_l^j$, and $\boldsymbol{\sigma}_m^j$ are the j-th dimension of $\boldsymbol{\mu}(\boldsymbol{x}_l), \boldsymbol{\mu}(\boldsymbol{x}_m), \mathrm{diag}(\sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_l)})$, and $\mathrm{diag}(\sqrt{\boldsymbol{\Sigma}(\boldsymbol{x}_m)})$ respectively. We can easily obtain $SKL(\boldsymbol{z}_l, \boldsymbol{z}_n)$ following a similar formulation.

On the other hand, Wasserstein distance has attracted increasing attention due to its success in generative adversarial networks [2, 3, 14]. Wasserstein distance can be viewed as the minimum cost of turning one pile into the other. In this paper, we also consider utilizing the Wasserstein metric to measure the distance of two Gaussian distributions. Specifically, the 2-Wasserstein distance is employed since it gives a closed-form solution for two normal distributions:

$$
\begin{aligned}
W_2(\boldsymbol{z}_l, \boldsymbol{z}_m)^2 &= W_2(\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m))^2 \\
&= \sum_{j=1}^{D} (\boldsymbol{\mu}_l^j - \boldsymbol{\mu}_m^j)^2 + (\boldsymbol{\sigma}_l^j - \boldsymbol{\sigma}_m^j)^2.
\end{aligned}
\tag{17}
$$

With the formulations in (16) and (17), we can easily calculate the ordinal distribution constraint loss in (14) under different distance metrics. We further depict the idea of the ordinal distribution constraint in Figure 2(b).

Although the data uncertainty can be modeled with the introduced $\boldsymbol{\Sigma}(\boldsymbol{x})$, our model may predict small $\boldsymbol{\Sigma}(\boldsymbol{x})$ for all samples and approximately degenerate into deterministic embeddings. To address this issue, we borrow the idea of the variational information bottleneck [1] and introduce a regularization term. The information bottleneck principle aims to learn a latent encoding which captures the task-related parts of $\boldsymbol{x}$ while forgetting the irrelevant parts of the input. Concretely, we introduce a KL divergence term to constrain our probabilistic ordinal embeddings $\boldsymbol{z}$:

$$
\begin{aligned}
\mathcal{L}_{VIB}(\boldsymbol{x}) &= KL\left(\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x})) \| \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\right) \\
&= \frac{1}{2} \sum_{j=1}^{D} ((\boldsymbol{\mu}^j)^2 + (\boldsymbol{\sigma}^j)^2 - \log(\boldsymbol{\sigma}^j)^2 - 1),
\end{aligned}
\tag{18}
$$

where $\boldsymbol{\mu}^j$ and $\boldsymbol{\sigma}^j$ denote the j-th dimension of $\boldsymbol{\mu}(\boldsymbol{x})$ and $\mathrm{diag}(\sqrt{\boldsymbol{\Sigma}(\boldsymbol{x})})$ respectively.

In the end, we combine the above three ingredients to train our method. For example, we train the probabilistic

ordinal embeddings for classification based regression with the following loss function:

$$
\mathcal{L}_{POE} = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \mathcal{L}_{PCls}(\boldsymbol{x}) + \alpha \mathcal{L}_{Ord} + \frac{\beta}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \mathcal{L}_{VIB}(\boldsymbol{x}),
\tag{19}
$$

where $\mathcal{D}$ denotes the training set, α and β are trade-off hyper-parameters which control the importance of three ingredients. We can easily infer the total loss functions for other present-day regression solutions.

4. Experiments

In this section, we conducted extensive experiments to validate the effectiveness of the proposed approach.

4.1. Age Estimation

Datasets: Age estimation attempts to predict the age for a given facial image. In our experiments, we apply our method to two widely used age datasets: the MORPH II [45] dataset and the Adience [30] dataset. MORPH II dataset includes 55,134 facial images of 13,618 individuals. Each image is labeled with the exact age value ranging from 16 to 77 years. We employ the popular protocol as in [32, 46, 49], which selects 5,492 images from Caucasian descent to avoid the cross-race influence. Then these images are divided into two subsets: 80% for training and 20% for testing. Adience dataset contains 26,580 Flickr photos of 2,284 subjects. The age is annotated with eight groups: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60 years old. For evaluation, we adopt the standard five-fold, subject-exclusive cross-validation protocol as used in [35].

Experimental Settings: For the MORPH II database, most existing methods [32, 53, 56] utilized a VGG-16 network that was pre-trained on the large-scale IMDB-WIKI dataset [46] as the feature extractor. For a fair comparison, the same strategy was adopted in our experiments. Meanwhile, the ImageNet pre-trained VGG-16 network was used for the Adience face database since most competing methods [8, 36] used this model. We directly employed the penultimate fully connected layer of VGG-16 to predict the mean $\boldsymbol{\mu}(\boldsymbol{x})$ with embedding dimensions of 4096. An additional head branch was used to output the diagonal covariance $\boldsymbol{\Sigma}(\boldsymbol{x})$. We implemented it with the architecture: `FC-BN-exp`. All experiments in this paper employed this architecture to model the data uncertainty. For both age databases, we trained models for 50 epochs with a batch size of 32. For optimization, Adam optimizer was utilized with a learning rate of 0.0001. For all experiments in this paper, we set the hyper-parameters α and β to 0.0001 and 0.00001 respectively. For the MORPH II database, we employed the MAE metric, which calculated the mean absolute error between ground truths and predictions. For the Adience face database, we also reported the classification accuracy for a

Table 1. Results of symmetrized KL divergence with different margins on the MORPH II dataset.

Margin	1	2	5	10	20	50
MAE	2.38	**2.35**	**2.35**	2.37	2.41	2.41

Table 2. Results of 2-Wasserstein distance with different margins on the MORPH II dataset.

Margin	10	20	50	100	200	500
MAE	2.40	2.41	2.37	**2.36**	2.38	2.39

Table 3. Ablation experiments on the MORPH II dataset.

Choice	Regression Methods									
	Classification					Ranking				
BL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
P-Emb		✓	✓	✓	✓		✓	✓	✓	✓
$\mathcal{L}_{Ord}$			✓		✓			✓		✓
$\mathcal{L}_{VIB}$				✓	✓				✓	✓
MAE	2.64	2.45	2.37	2.39	2.35	2.62	2.53	2.45	2.48	2.43

Table 4. The comparisons between our method and other state-of-the-art methods on the MOPRH II dataset.

Methods	MAE	Year
DRFs [49]	2.91	2018
AGEn [53]	2.52	2018
BridgeNet [32]	2.38	2019
AVDL [56]	2.37	2020
D-Regression	2.74	-
D-Regression + POE	2.50	-
Classification	2.64	-
Classification + POE	**2.35**	-
Ranking	2.62	-
Ranking + POE	2.43	-

Table 5. Results on the Adience face dataset.

Methods	Accuracy(%)	MAE
CNNPOR [35]	57.4 ± 5.8	0.55 ± 0.08
GP-DNNOR [36]	57.4 ± 5.5	0.54 ± 0.07
SORD [8]	59.6 ± 3.6	0.49 ± 0.05
D-Regression	56.3 ± 4.9	0.56 ± 0.07
D-Regression + POE	57.6 ± 4.2	0.54 ± 0.05
Classification [35]	54.0 ± 6.3	0.61 ± 0.08
Classification + POE	$\mathbf{60.5 \pm 4.4}$	$\mathbf{0.47 \pm 0.06}$
Ranking [38]	56.7 ± 6.0	0.54 ± 0.08
Ranking + POE	60.3 ± 4.4	0.48 ± 0.07

more comprehensive comparison. To report the classification accuracy for the direct regression method and ranking based method, the closest class of the regression result was taken as the predicted class.

Parameters Analysis: We first investigate the influence of different distance metrics and different margins δ on the MORPH II database. Table 1 shows the results with symmetrized KL divergence. We see that our method attains the best performance when the margin δ is set to 2 or 5. Furthermore, the performance varies slightly with different margins δ, which shows our method is robust to the hyperparameter δ. We further list the results with 2-Wasserstein distance in Table 2. We see that our method is also robust to the margin δ for 2-Wasserstein distance. Both metrics can achieve similar best results using appropriate margins, which demonstrates the robustness of our method to distance metrics. In the following experiments, we employ the symmetrized KL divergence with margin $\delta = 5$.

Ablation Study: To validate the effectiveness of the proposed probabilistic ordinal embeddings, we conducted ablation studies on the MORPH II database with two popular regression methods. Table 3 shows the comparison results in different settings, where BL stands for baseline that learns unordered deterministic embeddings, and P-Emb means learning probabilistic embeddings. It is clear that the probabilistic embeddings boost the baseline methods, improving the MAE by 0.19 and 0.09 for the classification based method and ranking based method respectively. We find that the ordinal distribution constraint $\mathcal{L}_{Ord}$ and the regularization term $\mathcal{L}_{VIB}$ can further improve the performance while the improvement brought by the $\mathcal{L}_{Ord}$ is more significant. Note that the VIB loss is used to stabilize the training

process rather than improve performance, so for the $\mathcal{L}_{Ord}$ method, further combining with $\mathcal{L}_{VIB}$ leads to marginal improvement. However, the $\mathcal{L}_{Ord}$ effectively outperforms the P-Emb method, showing the superiority of the proposed ordinal distribution constraint.

Comparison with Other Methods: Table 4 shows the experimental results on the MORPH II database. We list the results of three different regression methods and their corresponding probabilistic ordinal embeddings versions. The D-Regression denotes the direct regression method. Equipped with the proposed POE, the direct regression approach, classification based method, and ranking based solution achieve an MAE of 2.50, 2.35, and 2.43 respectively, which outperform the versions of deterministic embeddings by a large margin. By modeling the data uncertainty in the embedding space, our method essentially learns a complex probability model for regression results, which leads to consistent performance improvement. We also compare our method with some state-of-the-art methods [53, 56] which are specially designed for age estimation. For example, DRFs [49] extended the deep neural decision forests [26] to the regression problem and proposed deep regression forests to learn a Gaussian mixture model. AVDL [56] learned an adaptive variance based distribution for labels with a meta-learning framework and achieved excellent results. Compared with these advanced approaches, POE based methods
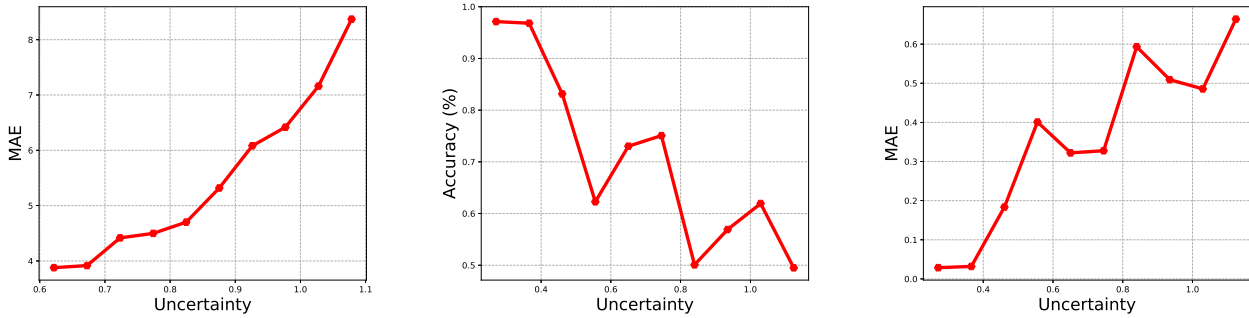
| (a) MAE vs uncertainty on MORPH II dataset | (b) Accuracy vs uncertainty on Adience dataset | (c) MAE vs uncertainty on Adience dataset |

Figure 3. Correlations between the uncertainty score and different performance metrics on two corrupted test sets. To quantity the correlation, we adopt the Kendall's tau correlations [22]. For the curves of three sub-figures, the results are 0.956, -0.733, and 0.778, respectively.

achieve superior or competitive results, which illustrates the effectiveness of our method.

We also conducted experiments on the Adience face database. Table 5 tabulates the results. We observe that the proposed probabilistic ordinal embeddings consistently improve the performance of three commonly used regression methods. For example, our method improves the ranking based regression solution by 3.6% for accuracy and 0.06 years for MAE. With the proposed POEs, the classification based method attains an accuracy of 60.5% and an MAE of 0.47. We see that it also outperforms the previous state-of-the-art methods, which shows the superiority of our probabilistic ordinal embeddings.

Uncertainty Estimation: Knowing what a model does not know is critical for many real-world applications. We aim to estimate when an input image can be reliably predicted or not. Since the data uncertainty is modeled in the diagonal covariance $\Sigma(x)$, we give a per-exemplar uncertainty score using the harmonic mean of the predicted variance: $D / \sum_{j=1}^{D} (\sigma^j)^{-1}$.

To validate the utility of our uncertainty score, we conducted experiments on the MOPRH II database and the Adience face database. We first construct corrupted test sets for both databases using gaussian blur with various radiuses. For each test set, we sort all samples based on the defined uncertainty scores which are calculated with our best model. Then we divide all sorted test samples into 10 bins. For the samples in a bin, we calculate the mean absolute error to show the correlations. For the Adience face database, we also employ the classification accuracy metric. We plot the correlations between performance metrics and our uncertainty score in Figure 3. We see that the proposed uncertainty score correlates well with MAE on the MORPH II database, which demonstrates that our uncertainty score can be regarded as an indicator of the reliability of model predictions. We observe the same trend with some jitters on the Adience face database, which further verifies the utility of our uncertainty score.

4.2. Image Aesthetics Assessment

Dataset: The Aesthetics dataset [48] is used to estimate the aesthetic grades, where 15,687 Flickr image URLs are provided and 13,929 images are available. Four categories are considered in this dataset: nature, animal, urban, and people. The 5-point absolute category rating scale is employed to evaluate the photographic quality: unacceptable, flawed, ordinary, professional, and exceptional. Each image receives at least five judgments from different graders. The ground truth is defined as the median rank. We randomly divide all images into three non-overlapped subsets: 75% for training, 5% for validation, and 20% for testing. Following [8, 35], we perform five-fold cross-validation for a fair comparison.

Experimental Settings: Since most previous works [8, 35, 36] employed the ImageNet pre-trained VGG-16 as the backbone, we also adopted it to extract image features. For optimization, we used Adam optimizer with a batch size of 32. The learning rate was set to 0.001 for the last layer and 0.0001 for other layers. For data preprocessing, we first resized all training image into 256×256, and then randomly cropped them to 224×224 patches. Random horizontal flipping was further performed for data augmentation. Both classification accuracy and MAE were utilized as the evaluation metrics. For the direct regression method and ranking based method, we obtained the predicted class with the same strategy used in age estimation databases.

Results: Table 6 summarizes the results on the Image Aesthetics dataset for each nominal category. We observe that our probabilistic ordinal embeddings significantly boost existing regression methods. For example, the POE version of the ranking based method achieves an overall accuracy of 72.33% and an overall MAE of 0.289, outperforming the original ranking method by 3.37% for accuracy and 0.037 for MAE. The consistent performance improvements across all nominal categories demonstrate the effectiveness of our approach. What's more, the classification

Table 6. Results on the Image Aesthetics dataset. We report the accuracy and MAE for each nominal category.

Methods	Accuracy(%) - higher is better					MAE - lower is better				
	Nature	Animal	Urban	People	Overall	Nature	Animal	Urban	People	Overall
CNNPOR [35]	71.86	69.32	69.09	69.94	70.05	0.294	0.322	0.325	0.321	0.316
SORD [8]	73.59	70.29	**73.25**	70.59	72.03	0.271	0.308	**0.276**	0.309	0.290
D-Regression	71.52	70.72	71.22	69.72	70.80	0.378	0.397	0.387	0.400	0.390
D-Regression + POE	73.44	70.44	72.26	71.02	71.79	0.360	0.386	0.376	0.385	0.376
Classification [35]	70.97	68.02	68.19	71.63	69.45	0.305	0.342	0.374	0.412	0.376
Classification + POE	73.62	**71.14**	72.78	**72.22**	**72.44**	0.273	**0.299**	0.281	**0.293**	**0.287**
Ranking [38]	69.81	69.10	66.49	70.44	68.96	0.313	0.331	0.349	0.312	0.326
Ranking + POE	**74.40**	71.12	72.22	71.58	72.33	**0.268**	0.301	0.288	0.298	0.289

Table 7. Results of our models and state-of-the-art methods on the historical image benchmark. We report the accuracy and MAE.

Methods	Accuracy(%)	MAE
Palermo *et al.* [41]	44.92 ± 3.69	0.93 ± 0.08
CNNPOR [35]	50.12 ± 2.65	0.82 ± 0.05
GP-DNNOR [36]	46.60 ± 2.98	0.76 ± 0.05
D-Regression	42.24 ± 2.91	0.79 ± 0.03
D-Regression + POE	46.04 ± 2.87	0.72 ± 0.02
Classification [35]	48.94 ± 2.54	0.89 ± 0.06
Classification + POE	**54.68 ± 3.21**	0.67 ± 0.04
Ranking [38]	44.67 ± 4.24	0.81 ± 0.06
Ranking + POE	50.01 ± 4.12	**0.66 ± 0.05**

based method with the proposed probabilistic ordinal embeddings achieves the state-of-the-art results, which further shows the importance of data uncertainty modeling.

4.3. Historical Image Dating

Dataset: The historical color image dataset [41] is collected for the task of automatically estimating the age of historical color photos. Each image is annotated with its associated decade, where five decades from the 1930s to 1970s are considered. There are 265 images for each category. Following the setting used in [35, 36], the standard train/val/test split uses 210 images for training, 5 images for validation, and 50 images for testing in each decade. Therefore, the total size of training, validation, and testing sets are 1,050, 25, and 250 images, respectively. Since this benchmark is small scale, we adopt ten-fold cross-validation as used in [35, 41] and report the mean values of the results.

Experimental Settings: All experiments in this dataset utilized VGG-16 as the backbone, which was initialized with the ImageNet pre-trained weights for a fair comparison. We set the initial learning rate to 0.0001 except for the last fully-connected layer, which used a higher learning rate of 0.001 to accelerate convergence. We trained the network for 50 epochs with Adam optimizer. To avoid overfitting, data augmentation was applied in our experiments. Concretely, we performed random horizontal flipping and random cropping for each training image. The prediction was obtained with a central crop during testing.

Results: We show the results on the historical color image dataset in Table 7. It is observed that the performance of three different regression approaches is greatly improved by the proposed probabilistic ordinal embeddings. Specifically, the classification based method with our POEs achieves an accuracy of 54.68% and an MAE of 0.67, which outperforms the deterministic embeddings version by 5.74% for accuracy and 0.22 for MAE. Combined with POEs, the ranking based method achieves the lowest MAE of 0.66 and outperforms the state-of-the-art method CNNPOR [35] by 0.1 years, which illustrates the superiority of our proposed probabilistic ordinal embeddings.

5. Conclusion

In this paper, we have presented the probabilistic ordinal embeddings for uncertainty-aware regression. Unlike the existing methods that only consider the direct regression solution and model the data uncertainty in the output space with a certain family of probabilistic distributions, our method models the uncertainty within the popular regression methods. We treat each data as a multivariate Gaussian distribution in the embedding space and apply the ordinal distribution constraint to enforce the ordinal relations in the latent space. Our method provides a per-exemplar uncertainty score, which can be used to measure the reliability of model predictions. Extensive experiment results on three visual tasks including age estimation, image aesthetics assessment, and historical image dating demonstrate that our method significantly boosts the present-day regression methods and achieves state-of-the-art performance.

Acknowledgement

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. 5

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017. 5

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 5

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *CVPR*, pages 5710–5719, 2020. 2

[5] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, pages 585–592, 2011. 2

[6] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, pages 13984–13993, 2020. 2

[7] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *CVPR*, pages 5183–5192, 2017. 2

[8] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, pages 4738–4747, 2019. 5, 6, 7, 8

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 2

[10] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *TIP*, 26(6):2825–2838, 2017. 4

[11] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, pages 712–718, 2018. 2, 4

[12] Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *CVPR*, pages 3369–3378, 2018. 1

[13] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *TPAMI*, 35(10):2401–2412, 2013. 2

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017. 5

[15] Fredrik K Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B Schön. Energy-based models for deep probabilistic regression. In *ECCV*, 2020. 2

[16] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, pages 2888–2897, 2019. 2

[17] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. S2gan: Share aging factors across ages and share aging trends among individuals. In *ICCV*, pages 9440–9449, 2019. 2

[18] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*, 88(2):115–132, 2000. 2

[19] Po-Yu Huang, Wan-Ting Hsu, Chun-Yueh Chiu, Ting-Fan Wu, and Min Sun. Efficient uncertainty estimation for semantic segmentation in videos. In *ECCV*, pages 520–535, 2018. 2

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017. 1, 2

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. 2

[22] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 7

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[24] Lingkai Kong, Jimeng Sun, and Chao Zhang. Sde-net: Equipping deep neural networks with uncertainty estimates. In *ICML*, 2020. 2

[25] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, pages 662–679, 2016. 1, 2

[26] Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulo. Deep neural decision forests. In *ICCV*, pages 1467–1475, 2015. 6

[27] Felix Kuhnke and Jorn Ostermann. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *ICCV*, pages 10164–10173, 2019. 1

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, pages 6402–6413, 2017. 2

[29] Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *ICCV*, pages 1191–1200, 2019. 1

[30] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *CVPRW*, pages 34–42, 2015. 5

[31] Wanhua Li, Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Graph-based social relation reasoning. In *ECCV*, pages 18–34, 2020. 1

[32] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *CVPR*, pages 1145–1154, 2019. 1, 2, 5, 6

[33] Wanhua Li, Yingqiang Zhang, Kangchen Lv, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Graph-based kinship reasoning network. In *ICME*, pages 1–6, 2020. 1

[34] Yanzhu Liu, Adams Wai-Kin Kong, and Chi Keong Goh. Deep ordinal regression based on data relationship for small datasets. In *IJCAI*, pages 2372–2378, 2017. 2

[35] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *CVPR*, pages 831–839, 2018. 5, 6, 7, 8

[36] Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *ICCV*, pages 5301–5309, 2019. 1, 2, 5, 6, 7, 8

[37] Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. In *ICMR*, pages 447–450, 2014. 1

[38] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age

estimation. In *CVPR*, pages 4920–4928, 2016. 2, 6, 8

[39] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *ICLR*, 2019. 2

[40] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NIPS*, pages 13991–14002, 2019. 2

[41] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *ECCV*, pages 499–512, 2012. 1, 2, 8

[42] Bowen Pan, Shangfei Wang, and Qisheng Jiang. Image aesthetic assessment assisted by attributes through adversarial learning. In *AAAI*, pages 679–686, 2019. 1

[43] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, pages 5285–5294, 2018. 1, 2

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2016. 1

[45] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, pages 341–345, 2006. 5

[46] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126(2-4):144–157, 2018. 2, 3, 5

[47] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPRW*, pages 2074–2083, 2018. 2

[48] Rossano Schifanella, Miriam Redi, and Luca Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *ICWSM*, 2015. 7

[49] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L Yuille. Deep regression forests for age estimation. In *CVPR*, pages 2304–2313, 2018. 5, 6

[50] Wei Shen, Kai Zhao, Yilu Guo, and Alan L Yuille. Label distribution learning forests. In *NIPS*, pages 834–843, 2017. 4

[51] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, pages 6902–6911, 2019. 2

[52] Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *AAAI*, pages 5008–5015, 2019. 2

[53] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z Li. Efficient group-n encoding and decoding for facial age estimation. *TPAMI*, 40(11):2610–2623, 2018. 5, 6

[54] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020. 2

[55] Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *CVPR*, pages 13086–13095, 2020. 1, 2

[56] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *ECCV*, 2020. 1, 5, 6

[57] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *CVPR*, pages 1087–1096, 2019. 1, 2

[58] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stage-wise regression network for age estimation. In *IJCAI*, page 7, 2018. 1