Ordinal Label Distribution Learning

Changsong Wen* Xin Zhang* Xingxu Yao Jufeng Yang†
VCIP & TMCC & DISSec, College of Computer Science, Nankai University
downdric@163.com, zhangxin.nk@mail.nankai.edu.cn, yxx_hbgd@163.com, yangjufeng@nankai.edu.cn

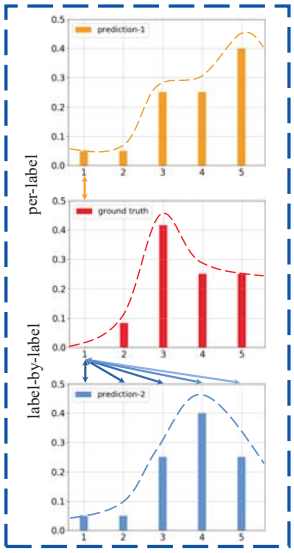
Abstract

Label distribution learning (LDL) is a recent hot topic, in which ambiguity is modeled via description degrees of the labels. However, in common LDL tasks, e.g., age estimation, labels are in an intrinsic order. The conventional LDL paradigm adopts a per-label manner for optimization, neglecting the internal sequential patterns of labels. Therefore, we propose a new paradigm, termed ordinal label distribution learning (OLDL). We model the sequential patterns of labels from aspects of spatial, semantic, and temporal order relationships. The spatial order depicts the relative position between arbitrary labels. We build cross-label transformation between distributions, which is determined by the spatial margin in labels. Labels naturally yield different semantics, so the semantic order is represented by constructing semantic correlations between arbitrary labels. The temporal order describes that the presence of labels is determined by their order, i.e. five after four. The value of a particular label contains information about previous labels, and we adopt cumulative distribution to construct this relationship. Based on these characteristics of ordinal labels, we propose the learning objectives and evaluation metrics for OLDL, namely CAD, QFD, and CJS. Comprehensive experiments conducted on four tasks demonstrate the superiority of OLDL against other existing LDL methods in both traditional and newly proposed metrics. Our project page can be found at https://downdric23.github.io/.

(a) Image

Metric	(gt,pd-1)	(gt,pd-2)
KL↓	0.1379	0.1379
Canberra↓	1.7293	1.7293
Cosine↑	0.2638	0.2638
M_{CAD}↓	0.3670	**0.0217**
M_{QFD}↓	0.1275	**0.1125**
M_{CJS}↓	0.0768	**0.0434**

(c) Metrics



(b) Examples for order-sensitive distributions

Figure 1. An example of image aesthetics rating using scores from 1 to 5. There exist obvious order relations in the label space. (a) shows an example of the image. In (b), compared with prediction-1 (pd-1), prediction-2 (pd-2) obviously fits better to the ground truth label distribution. However, they both have equal performance on the widely-used distances metrics (KL, Canberra, and Cosine) as shown in (c). Our order-sensitive metrics including M_{CAD}, M_{QFD}, and M_{CJS} can reflect the difference between the two predictions.

1. Introduction

In the common machine learning paradigms, single-label learning [27] predicts one specific label for an instance. Multi-label learning [58] predicts multiple labels which can handle some ambiguous cases where an instance is related to more than one class. However, it treats each label equally and only assigns the same degree to the related classes. To tackle this problem, Geng [19] proposes label distribution learning (LDL). This new paradigm models the different relative importance between labels for describing an instance. Due to the characteristics of LDL [19], it has been widely used to solve various real-world ambiguous tasks [57, 66, 74, 34], including age estimation [17, 28], image aesthetics analysis [62], *etc*. Specifically, LDL algorithms assign a number d_x^y less than or equal to 1 to the potential label y for describing an instance x. Note that $\sum_y d_x^y = 1$ is held in order to satisfy the condition that label space is complete.

Generally, LDL tries to learn accurate results at each label. However, in some common LDL tasks, such as age estimation [47] and beauty rating [65], labels are in order. The natural order of labels presents a sequential pattern and is different from other LDL tasks like facial

expression recognition. In this case, existing algorithms in LDL become insufficient to fully explore the distribution difference in OLDL tasks. Concretely, LDL algorithms [19, 60, 48, 49, 23, 37] take mean squared error (MSE), mean absolute error (MAE), and Kullback-Leibler (KL) divergence [7] as learning objectives. These methods mainly concentrate on the distance between distributions at the same labels (per-label) and didn't utilize the implicit orders between them (label-by-label). As demonstrated in Fig. 1, while measuring the prediction of image aesthetics scores for the image in (a), prediction-2 is relatively closer to the ground truth distribution shown in (b). However, prediction-1 and prediction-2 have identical performance when evaluated with the widely used per-label LDL metrics in (c). Meanwhile, there are some previous works studying ordinal labels. Li *et al*. [38] makes a unimodal assumption in age predictions based on the distance relationship between ages. To process sequential words in the text, many works [54, 9] take semantic relations between words into account. Time series analysis [31] is a prevalent direction to explore ordinal information. These theories further inspire the exploration of sequential patterns in labels.

In this paper, we propose a new paradigm, termed ordinal label distribution learning (OLDL). Following the inspiration above, we model the sequential patterns in labels from three aspects: spatial, semantic, and temporal order relationship. 1) From the perspective of spatial order, the sequential pattern in labels describes whether the label is adjacent or distant from the others [39]. Thus, the spatial order can be naturally modeled via distance (margin). In most cases, the labels in OLDL are real numbers, such as scores [44]. So, the margin between label i and label j is reasonably formulized as $|i - j|$. Based on the margin, we introduce the Cumulative Absolute Distance (CAD). It aims to find the minimal cost of transforming one distribution to another label-by-label. 2) From the perspective of semantic order, we transform the sequential pattern between labels into semantic similarities. Motivated by the nearby labels are also close in the semantic space, we design an order-sensitive matrix and integrate it into the calculation of Quadratic Form Distance (QFD) [1]. Each element in this matrix represents the semantic similarity between pairwise labels, which indicates their relations in order. 3) From the perspective of temporal order, because the cumulative density function can cumulate distributions in sequence [45], it intrinsically contains temporal orders between labels. Moreover, based on the divergence that can reflect the difference between two distributions, we propose to use the information theory-based divergence measure, termed Cumulative Jensen-Shannon divergence (CJS) in OLDL.

Distances from these methods are directly adopted as learning objectives. Further, we adopt these three distances from the algorithms as new order-sensitive metrics, denoted as M_{CAD}, M_{QFD}, and M_{CJS}. These metrics are suggested to better assess the prediction results in OLDL.

The contributions of the paper are three-fold:

- We propose a novel paradigm for label distribution learning, termed OLDL, in which the sequential patterns of labels are fully explored to further boost OLDL tasks.
- We explore the sequential patterns of labels from three aspects: spatial order, semantic order and temporal order. The corresponding orders are modeled as margins, semantic similarities, and distribution cumulation. Distances CAD, QFD, and CJS are derived from the methods.
- We conduct comprehensive experiments on five widely used datasets of four vision tasks. Evaluated by both the existing and newly proposed metrics, the results demonstrate the superiority of the proposed OLDL paradigm.

2. Related Work

2.1. Label Distribution Learning

Label distribution learning (LDL) [19] aims at solving the ambiguity problem. LDL [19, 53, 21, 59] exploits real-valued probabilities to stand for the description degree of labels. It has wide applications in downstream tasks like facial landmark prediction [56], head pose estimation [20], facial expression recognition [61, 5], emotion analysis [63, 73, 72], *etc*.

In [19], three strategies are introduced to tackle problems of LDL, including Problem Transform (PT), Algorithm Adaption (AA), and Specialized Algorithms (SA). PT transforms LDL to single-label classification, and the SLL algorithms are applied to the learning process. AA extends the existing algorithms and adapts them to the LDL task. AA-*k*NN and AA-BP are two representative methods. To design algorithms based on the characteristic of LDL, the maximum entropy model [2] is introduced in LDL. Gauss-Newton and quasi-Newton methods are respectively exploited in SA-IIS (IIS-LDL) [22] and SA-BFGS (BFGS-LDL) [19] to optimize the model. Recently, [67] explores a label enhancement method to refine noise annotations from trusted data. To tackle the objective inconsistency in training and testing [60], they apply L_1-norm loss as a learning metric and proposed a re-weighting strategy. [33] designs a new loss function to learn more accurate ranking between labels. [32] proposes to use additional information extracted by a local correlation vector and mine semantics between labels on local samples.

In the recent years, CNNs achieve great progress in many vision tasks [71, 41]. Due to the large labor and time cost

of annotating data with label distributions, some methods are designed to learn label distributions with incomplete label information or generate reliable complete distributions using additional information sources [69, 68, 70]. They utilize available prior knowledge to generate reliable ground-truth [16, 17, 18, 6]. In [69], Xu *et al*. extend logical labels (*i.e*. binary indicators indicating if items are instances of a category or not) to label distributions using so-called graph Laplacian label enhancement. Considering the inaccurate label information of facial landmarks, Su and Geng [57] propose a bi-variant label distribution learning algorithm for tackling soft facial landmark detection tasks.

2.2. Ordinal Regression/Classification

Ordinal regression focuses on the order relations of different labels in classification tasks [75, 10]. They [24, 15] aim to make the prediction fit better to the ground truth by considering the inter-label relations. [14] first uses decision trees to solve ordinal classification problems. [52] introduces the pairwise distance of labels to depict the ordinal relations. Liu *et al*. [42] conduct a pairwise hinge loss on tuples of instances of different categories, and the negative log-likelihoods for different categories are minimized by the order relations. Some ordinal regression methods also introduce LDL to improve the robustness [47] or performance [38]. A mean-variance loss is proposed by Pan *et al*. [47] to penalize the difference of regression results and distributions simultaneously. To estimate the age groups, Hou *et al*. [29] design a hybrid loss composed of cross-entropy and squared earth mover's distance (EMD^2). More recently, Li *et al*. [38] utilize a unimodal-concentrated loss to enforce the predicted distributions to be unimodal and have the highest prediction consistent with the ground-truth label.

Compared with label distribution learning, ordinal regression/classification focus on obtaining an accurate ground-truth label. However, LDL requires the distributions to be consistent. It is more challenging. Moreover, OLDL model the ordinal relation between labels in the distribution perspective, which is not fully explored in ordinal regression task.

3. Methodology

3.1. Problem Formulation of OLDL

In the OLDL paradigms, $p(y|x;\boldsymbol{\theta})$ is defined as a parametric model, where x and y are instance and labels, $\boldsymbol{\theta}$ represent parameters in the model. In the ordinal label space $y = \{y_1, y_2, \ldots, y_C\}$ of C different classes, we aim to predict the description degrees of labels (*i.e*. label distribution) for the given i-th instance x_i in the input space. In this paper, we consider the natural orders in y, $y_1 \prec y_2 \prec \cdots \prec y_C$, where $\prec$ denotes the ordered relation between labels and does not exist in the normal label space. For an instance x_i, its label distribution is denoted as $Q_i = \{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \cdots, d_{x_i}^{y_C}\}$, where $d_{x_i}^{y_j}$ denotes the description degree of y_j to x_i and we define $Q_i(j) = d_{x_i}^{y_j}$. Note that the constraints $Q_i(j) \in [0,1]$ and $\sum_j Q_i(j) = 1$ should be satisfied. Based on the definition, we naturally introduce the sequential patterns of labels from these three aspects as explained above: spatial order, semantic order, and temporal order.

3.2. Spatial Order: Cumulative Absolute Distance

Spatial order between labels describes that one label is adjacent or far from another label. We model this kind of sequential pattern as distance (margin). As stated above, we assume the labels are real numbers and the intervals between labels are equal. So, we define a margin between the j-th and k-th labels as $|j-k|$ to represent their relations in order. Based on the margins between labels, we modify the earth mover's distance (EMD) [51] into OLDL.

the earth mover's distance computes the minimal cost that is needed to transform between two different distributions, and the margin m_{jk} is defined as the cost of transformation between two labels. The orders are modeled via the cost, where transforming between closer labels costs less. Because the distributions in OLDL are all one-dimensional, EMD is equivalent to Mallows distance [36]. For distributions P_i and Q_i with equal length n, Mallows distance is usually reduced to an assignment problem to get the simplified solution: $\frac{1}{n}\sum_{i=1}^{n}|P_{(i)} - Q_{(i)}|$ [36]. The association of the i-th label with a different j-th label can be achieved. However, such a simplified process makes EMD lose the property that the i-th label could associate with multiple labels. Because cumulative distribution integrates information of previous labels in order, the value at each label intrinsically builds ordinal connections between multiple labels. Therefore, we use $\mathrm{CDF}_n(P_i)$ instead of $P_{(i)}$ in the simplified solution. We introduce cumulative absolute distance (CAD) defined as follows:

$$\mathrm{CAD}(P_i, Q_i) = \sum_{n=1}^{C} |\mathrm{CDF}_n(P_i) - \mathrm{CDF}_n(Q_i)|. \qquad (1)$$

where CDF$(\cdot)$ is the cumulative density function and $\mathrm{CDF}_n(P_i) = \sum_{j=1}^{n} P_i(j)$. By introducing CAD, the spatial order relations of ordinal labels are modeled via a distance measure.

3.3. Semantic Order: Quadratic Form Distance

Ordinal labels naturally yield different semantics, and the adjacent labels, *i.e*. ages 26 and 27, are obviously close in the semantic space. We propose to model the semantic order of labels as a sort of semantic relation.

Specifically, we construct an order-sensitive matrix $A \in \mathbb{R}^{C\times C}$, where C represents the number of labels. The

weight in A, denoted as a_{jk}, represents the semantic similarity of the j-th and k-th labels. The closer labels have stronger semantic relations, thus, the value of a_{jk} should be approaching 1. On the contrary, a_{jk} will be proximate to zero if the j-th and k-th labels are far from each other. The weight in A can be formulized as:

$$a_{jk} = 1 - \frac{m_{jk}}{m_{max}}, \quad \text{where} \quad m_{max} = \max_{j,k}(m_{jk}). \quad (2)$$

We integrate this matrix into previous evaluation metrics, *i.e*. Quadratic Form Distance (QFD) [1]. Given ground truth distribution $Q_i = \left\{d_{x_i}^{y_1}, d_{x_i}^{y_2}, \cdots, d_{x_i}^{y_C}\right\}$ and predicted distribution $P_i = \left\{\tilde{d}_{x_i}^{y_1}, \tilde{d}_{x_i}^{y_2}, \cdots, \tilde{d}_{x_i}^{y_C}\right\}$. QFD can be formulized as:

$$\mathrm{QFD}(P_i, Q_i) = \sqrt{(P_i - Q_i)^{\mathrm{T}} \cdot A \cdot (P_i - Q_i)}. \quad (3)$$

The matrix A should be a positive semi-definite matrix to make the inside of the square root hold a non-negative value. In the label distributions, $\sum_{j=0}^{C}(P_i(j) - Q_i(j))$ equals zero and the margin between the j-th and k-th labels is represented as $|j-k|$, $(P_i - Q_i)^{\mathrm{T}} \cdot A \cdot (P_i - Q_i)$ is non-negative [25]. Based on the matrix A, the semantic order relations between each pairwise label are involved in the QFD calculation.

3.4. Temporal Order: Cumulative JS Divergence

To depict the difference between two label distributions, KL and Jensen-Shannon (JS) divergence are widely adopted in previous approaches [19, 22]. Compared with these probability distribution-based divergences in which label $d_{x_i}^{y_j}$ only represents its own value, the distribution value of each label in the cumulative distribution is cumulated through all previous labels. It intrinsically integrates all the temporal order information of previous labels.

We introduce the idea of cumulating label distributions through the ordinal scale and adopt cumulative JS divergence (CJS) [45] to make use of the order relations in divergence measures. We show in detail how to extend JS divergence into the cumulative version. JS divergence between P_i and Q_i is formulized as:

$$D_{js}(P_i||Q_i) = \frac{1}{2}\sum_{j=1}^{C}\left(P_i(j)\log\frac{P_i(j)}{M_i(j)} + Q_i(j)\log\frac{Q_i(j)}{M_i(j)}\right), \quad (4)$$

where $M_i(j) = (P_i(j) + Q_i(j))/2$. Inspired by the cumulative density function, CJS divergence can be formulized as:

$$\mathrm{CJS}(P_i, Q_i) = \sum_{n=1}^{C} D_{js}\Big(\mathrm{CDF}_n(P_i)||\mathrm{CDF}_n(Q_i)\Big). \quad (5)$$

In our method, we use CJS as an ordinal divergence measure.

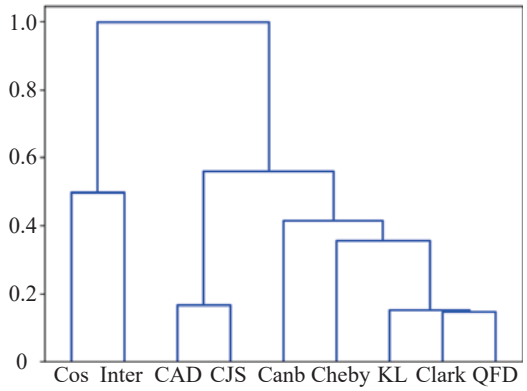


Figure 2. Dendrogram for the distance measures.

3.5. Hierarchical Clustering on Distance Measures

To further investigate the semantic similarity between distance measures, we conduct clustering analysis [4, 19]. We randomly generate $n = 100$ reference distributions $\{r_i\}_{i=1}^{n}$ and query distribution q, the correlation between two distances d_x, d_y is calculated by the formula:

$$Co(d_x, d_y) = \frac{\sum_{i=1}^{n}(d_x(r_i,q) - \overline{d_x}) \cdot (d_y(r_i,q) - \overline{d_y})}{\sqrt{\sum_{i=1}^{n}(d_x(r_i,q) - \overline{d_x})^2 \cdot (d_y(r_i,q) - \overline{d_y})^2}} \quad (6)$$

Where $\overline{d}$ is the average between n distances. We calculate the semantic distance between them by $1 - Co$ [4]. As the results shown in Fig. 2, the clusters of distance measures verify these distances can reflect different aspects of an algorithm.

4. Experiments

4.1. Datasets

We evaluate the effectiveness of our method through experiments on two aspects. We first analyze the results of common LDL tasks, including facial beauty prediction and image aesthetics analysis. Moreover, to show the generality of OLDL algorithms, we further conduct experiments on label distribution-related age estimation and joint acne grading tasks. The detailed dataset information is shown as follows: (1) **To assess facial attractiveness** [13], we evaluate our methods on the SCUT-FBP5500 [39] dataset, which consists of 5,500 images of frontal faces. Each image is rated with scores in the range 1–5 by 60 participants. (2) **For the image aesthetics analysis**, we conduct the experiments on a large-scale AVA [44] dataset. 210 participants rated each image with scores ranging from 1 to 10. We follow the train and test split used in the previous works [43, 35, 44]. (3) **Age estimation** is an OLDL-related ordinal regression task. Experiments are conducted on two age estimation datasets ChaLearn16 [12] and Morph [50]. ChaLearn16 contains 7,591 facial images, they are split into

Table 1. Experimental results on the SCUT-FBP5500 and AVA datasets. The deep methods are based on the pre-trained VGG-16 network, and they can be grouped into L_1-based, L_2-based, and KL-based. The methods of the proposed OLDL are presented in blue background.

Dataset	SCUT-FBP5500							AVA						
Metrics	L_1-based		L_2-based		KL-based			L_1-based		L_2-based		KL-based		
	MAE	CAD	MSE	QFD^2	LRR	JS	CJS	MAE	CAD	MSE	QFD^2	LRR	JS	CJS
Chebyshev↓	0.155±.005	**0.141±.004**	0.171±.005	0.148±.004	0.145±.002	0.150±.004	0.142±.004	0.096	0.092	0.104	0.095	0.101	0.095	**0.091**
Clark ↓	1.317±.019	1.327±.020	**1.294±.016**	1.303±.016	1.311±.032	1.321±.019	1.320±.010	1.307	1.326	1.363	1.322	1.301	1.310	**1.299**
Canberra ↓	2.305±.046	2.300±.056	2.263±.048	**2.235±.016**	2.267±.057	2.303±.045	2.284±.050	3.190	3.221	3.382	3.217	3.163	3.196	**3.139**
KL div ↓	0.148±.020	0.128±.016	0.173±.021	**0.112±.009**	0.135±.036	0.139±.014	0.128±.015	0.122	0.120	0.140	0.121	0.119	0.120	**0.112**
Cosine↑	0.937±.008	0.947±.016	0.930±.009	**0.953±.002**	0.944±.016	0.941±.006	0.947±.006	0.943	0.946	0.937	0.947	0.941	0.944	**0.949**
Intersection ↑	0.826±.006	0.827±.004	0.801±.007	**0.852±.004**	0.839±.025	0.827±.004	0.844±.005	0.821	0.827	0.810	0.831	0.826	0.827	**0.836**
M_{QFD} ↓	0.311±.020	0.281±.017	0.350±.023	**0.266±.018**	0.290±.045	0.302±.013	0.284±.015	0.305	0.295	0.334	0.300	0.309	0.304	**0.289**
M_{CAD} ↓	0.055±.005	0.048±.003	0.066±.005	**0.045±.002**	0.050±.009	0.053±.003	0.049±.003	0.052	0.049	0.059	0.050	0.050	0.052	**0.048**
M_{CJS} ↓	0.024±.003	0.019±.002	0.030±.003	**0.018±.001**	0.021±.006	0.022±.002	0.020±.002	0.031	0.030	0.040	0.030	0.029	0.033	**0.028**

4,113, 1,500, and 1,978 for training, validation, and testing [17]. Morph is the largest released real-age dataset and consists of 55,134 face images. Following the experimental settings used in [17, 46], 80% and 20% images are used for training and testing. (4) **In acne grading**, an acne severity grading dataset termed *ACNE04* [64] is presented, which is labeled by the global acne severity. The images in which the numbers of lesions are 1-5, 6-20, 21-50, and above 50 are labeled as mild, moderate, severe, and very severe, respectively. The dataset contains 1,457 images, which are split into 1,165 for training and 292 for testing.

4.2. Implementation details

All our CNN-based methods are using VGG-16 [55] pre-trained on ImageNet [8]. The original images are resized to 256×256 followed by 224×224 center cropping. The learning rate is initialized as 1.0×10^{-3} and reduces by one-tenth every 60 epochs for facial age estimation; the initial learning rate is set to 1.0×10^{-4} for other tasks with the same reduction setting. We fine-tune all layers of the network for a total of 120 epochs with a batch size of 32. The parameters of the framework are optimized by SGD with a weight decay of 0.0005 and a momentum of 0.9. We use six existing metrics (*i.e*. Chebyshev, Clark, Canberra, KL, Cosine, and Intersection) and three proposed distance metrics ($M_{QFD}, M_{CAD}, M_{CJS}$) to measure the distance between the predicted distribution P and real distribution Q.

In the comparison experiments, we evaluate the performance of the general deep LDL methods (DLDL) [19, 22] with VGG-16 as the baseline. As a representative distance metric, KL loss or JS loss has been widely used as learning objectives for various LDL tasks [40, 3, 49]. Besides, we demonstrate the results of baselines MAE and MSE, the important loss functions for LDL and regression tasks, which have been used in [60, 48]. Note that, we find optimizing with QFD^2 achieves better performance compared with QFD. So, we utilize QFD^2 in the experiments.

4.3. Comparison on the Common OLDL Tasks

In Table 1, we present the experimental results of our algorithms on OLDL tasks, *i.e*. facial beauty prediction (SCUT-FBP5500) and image aesthetics analysis (AVA). For a more convenient description and comparison with the deep methods, we analyze them in groups. Formally, CAD and MAE are L_1-based methods, QFD^2 and MSE are L_2-based methods. Both JS and CJS divergence are KL-based divergences. LRR [33] is the state-of-the-art method developed on KL divergence and ranking-based loss. In general, we present the results of L_1-based, L_2-based, and KL-based methods.

First, our algorithms perform favorably against other conventional learning objectives in both previous and newly proposed metrics. We can observe that QFD^2 and CJS achieve the best results in SCUT-FBP5500 and AVA datasets. Because learning label distributions requires a lot of data [30], and SCUT-FBP5500 is a much smaller dataset than AVA. Fitting the whole distribution by CJS is more difficult in the SCUT-FBP5500 dataset. Second, in these three groups, we achieve performance gains in most metrics. Within L_1-based, L_2-based, and KL-based methods, the superiority of the proposed CAD, QFD^2, and CJS has been demonstrated. It is observed that while comparing any two methods, the traditional metrics may give inconsistent voting results among these methods, but our order-sensitive metrics give consistent comparison results. This shows the reliability of order-sensitive metrics in OLDL tasks. The KL-based method includes LRR, JS, and CJS. Taking ranking between labels into consideration, LRR outperforms JS loss. The relation between labels is an important cue for learning a distribution. Simultaneously, the proposed CJS loss also has obvious improvement compared with JS loss. For LRR and CJS, CJS gets obviously better performance in the AVA and achieves comparable results in SCU-FBP5500 dataset (CJS is better in six of the nine metrics), while QFD^2

Table 2. Experimental results on age estimation datasets, *i.e.* Morph and ChaLearn16.

Method	Morph	ChaLearn16	
	MAE ↓	MAE ↓	ϵ-error ↓
DLDL (MAE, Max)	8.52	14.45	0.68
DLDL (MAE, Exp)	7.99	15.80	0.76
OLDL (CAD, Max)	6.51	8.56	0.54
OLDL (CAD, Exp)	6.41	8.27	0.51
DLDL (MSE, Max)	15.85	19.19	0.88
DLDL (MSE, Exp)	11.92	17.39	0.79
OLDL (QFD^2, Max)	**2.23**	5.74	0.40
OLDL (QFD^2, Exp)	2.24	5.56	0.41
DLDL (KL, Max)	2.96	6.23	0.48
DLDL (KL, Exp)	2.95	5.81	0.43
DLDL (JS, Max)	2.69	6.12	0.44
DLDL (JS, Exp)	2.52	5.79	0.42
OLDL (CJS, Max)	2.45	5.30	0.41
OLDL (CJS, Exp)	2.39	**5.06**	**0.39**

still outperforms LRR in SCU-FBP5500 dataset. As LRR focuses on the ranking of each label, Chebyshev, Clark, and Canberra measure the difference between the labels at each label in two distributions, LRR performs better in these three metrics. Compared to CJS which directly fits the distribution, QFD directly calculates the difference between two labels and relates them with order information, it has better results than LRR. All the analyses above demonstrate that order-sensitive learning objectives significantly improve performance and show the importance of considering order relations in OLDL tasks.

4.4. Comparison on Other OLDL-Related Tasks

4.4.1 Facial Age Estimation

Facial age estimation is a prevalent ordinal regression task, one promising direction is to predict the age distribution. We adopt Mean Absolute Error (MAE) to evaluate the performance, which represents the average distance between the prediction and ground-truth ages:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{7}$$

where $\hat{y}_i$ and y_i are predicted and ground-truth age. In addition, there is also a specific metric used for the ChaLearn16 dataset, ϵ-error, introduced in the ChaLearn competition, which is computed as:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \exp\left(-\frac{(\hat{y}_i - y_i)^2}{2\sigma_i^2}\right)\right), \tag{8}$$

where σ_i is the standard deviation of the i-th testing image. For each method, the results of 'Max' and 'Exp' are provided. To get the final age prediction, 'Max' means that we use the argmax value of the predicted distribution, and 'Exp' is the expectation of $P(i)$, formulized as $\text{Exp} = \sum_i y_i P(i)$.

As shown in Table 2, generally, the proposed OLDL losses outperform the corresponding conventional losses. Within the deep LDL methods, MSE loss performs worst because MSE weakens the penalty when the absolute error is less than 1. The disadvantage increases as the number of labels grows, because the value of the label is close to zero (age ranges from 1 to 85). However, by using the cumulative density function, the problem is largely alleviated. The performance is improved from 15.85 (Max) and 11.92 (Exp) to 2.23 (Max) and 2.24 (Exp) on the Morph dataset, respectively. The similar phenomenon also appears on the ChaLearn16 dataset. Compared with the results of MSE, QFD^2 obviously improves the performance in terms of two metrics, *i.e.*, MAE and ϵ-error. In the MAE-based method and KL-based method, the loss proposed for OLDL performs better. As the symmetrical version of KL loss, JS loss performs better than KL loss. It is mainly because JS loss guarantees a more stable convergence of the model due to its symmetry. Though KL loss and JS loss obtain high performance on age estimation tasks, CJS can further improve the results on both MAE and ϵ-error.

4.4.2 Acne Image Grading

Computer vision plays an increasingly important role in medical disease diagnosis, especially in skin disease diagnosis. Acne vulgaris, a common skin disease, has infected about 80% of adolescents as reported in [11], and they require effective treatment immediately in order to avoid scars and pigmentation. A fast and accurate diagnosis for acne is necessary for subsequent treatment and recovery. Nowadays, the Hayashi criterion [26] has been widely used by dermatologists to grade acne severity. It is a measurement determined by lesion counting and global assessment. Based on the number of lesions, four levels of the severity of acne are graded by the Hayashi criterion, including mild, moderate, severe, and very severe. Obviously, there are natural order relations between severities of acne. Meanwhile, similar levels of severity show similar appearance, resulting in the ambiguity issue which can be tackled by OLDL. We conduct experiments on the *ACNE04* dataset [64], which contains 1,457 images.

In the experiments, we plug our CJS into the algorithms proposed in [64]. In [64], apart from using cross-entropy for grade classification, KL loss is developed for grade label distribution learning and counting distribution learning. We directly replace the KL loss with the proposed CJS to conduct label distribution learning. For the evaluation of methods, in addition to commonly used accuracy and precision, we also select some essential specific metrics in the medical field following [64], including Specificity, Sensi-

Table 3. Experimental results on the ACNE04 dataset. We show the results of the representative label distribution learning methods and state-of-the-art method, *i.e.*, JGC. The values following '±' are plus/minus one standard deviation. The methods of the proposed OLDL are presented in blue background.

Criterion	PT-Bayes	PT-SVM	AA-kNN	AA-BP	SA-BFGS	DLDL	JGC	Ours
Precision ↑	45.30±0.09	44.60±0.07	67.61±0.13	65.36±0.10	73.85±0.03	78.51±0.03	84.37±0.02	**86.32±0.02**
Specificity ↑	79.39±0.03	83.04±0.03	87.73±0.07	87.37±0.02	91.01±0.01	92.24±0.01	93.80±0.00	**94.00±0.04**
Sensitivity ↑	45.06±0.12	46.05±0.05	67.33±0.15	58.65±0.10	72.03±0.03	78.57±0.05	81.52±0.02	**83.50±0.03**
Youden Index ↑	24.44±0.15	29.10±0.08	55.05±0.22	46.02±0.11	63.03±0.04	68.81±0.05	75.32±0.02	**76.89±0.02**
Accuracy ↑	45.38±0.07	48.15±0.11	68.15±0.17	66.44±0.04	76.16±0.03	79.31±0.02	84.11±0.01	**84.80±0.01**

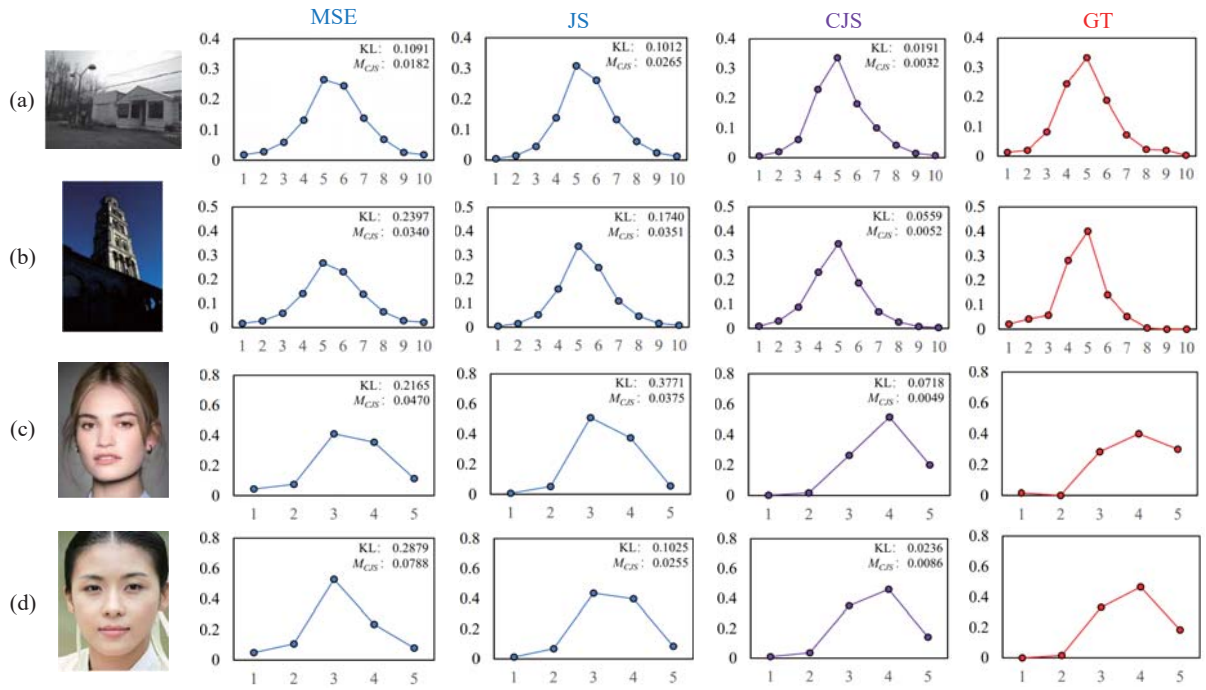


Figure 3. Visualization of predicted label distributions. We show the instances from the AVA dataset ((a) and (b)) and SCUT-FBP5500 dataset ((c) and (d)). The predictions of MSE loss, JS loss, and CJS loss, are presented from the second column to the fourth column. GT represents ground-truth distribution. The quantitative results are also provided. The metrics include KL divergence, and M_{CJS}.

tivity, and the Youden Index.

Specificity reflects the proportion of negatives that are correctly identified and is also termed true negative rate. Sensitivity reflects the true positive rate or recall, representing the proportion of positives that are correctly identified. Youden Index is a more comprehensive metric, which is computed as (Sensitivity + Specificity - 1).

In Table 3, we compare several traditional and CNN-based SOTA methods. As the baseline of CNN-based methods, DLDL performs better than traditional methods, including PT and AA algorithms. Compared with the new state-of-the-art method, *i.e.*, JGC [64], our algorithm further improves the performance, especially on the Sensitivity (recall), which is the proportion of disease that can be successfully diagnosed. Generally, our methods can not only provide a stronger baseline for various ordinal LDL tasks, but also can be used to replace a component in existing state-of-the-art methods.

4.5. Visualization

To qualitatively demonstrate the effectiveness of our proposed method, we provide visualizations of predicted distributions for real instances selected from the AVA and SCUT-FBP5500 datasets shown in Fig. 3. For each sample, we show the results of optimizing the model with conventional MSE loss, JS loss, and newly proposed CJS loss with the

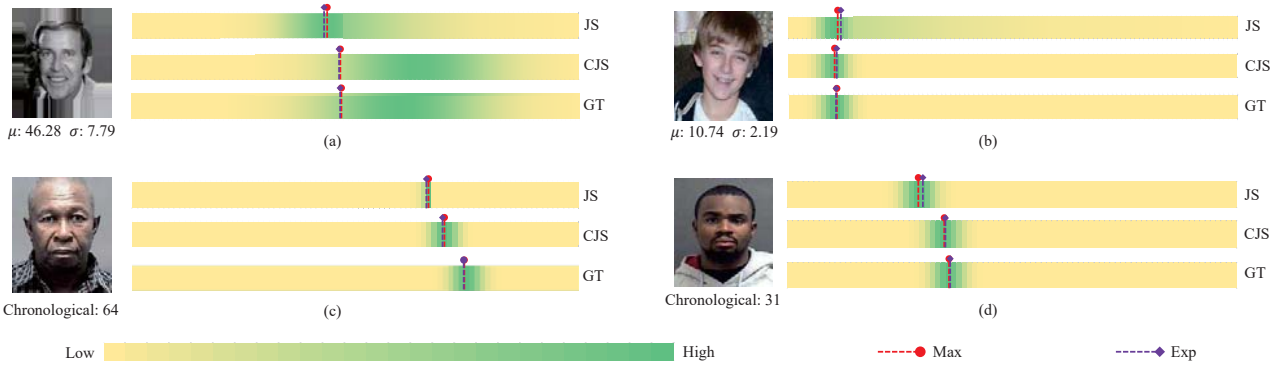


Figure 4. Visualization of predicted distribution of age estimation and corresponding ground truth. (a) and (b) are sampled from Morph datset. (c) and (b) are sampled from ChaLearn16 dataset.

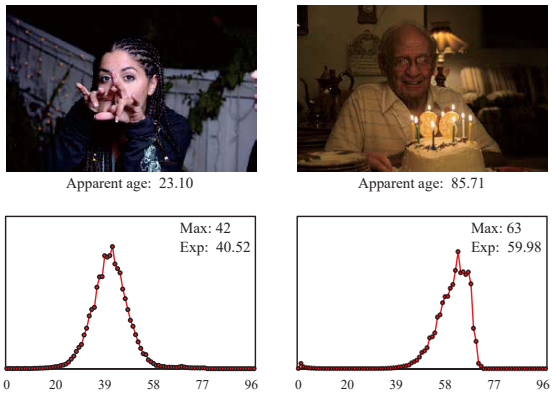


Figure 5. Failure cases of age estimation from CJS algorithm. The two instances are from ChaLearn16 dataset.

best performance. We also demonstrate KL divergence and CJS divergence (M_{CJS}) as metrics. From Fig. 3, we observe that the distributions learned by CJS loss are obviously more similar to the ground truth. Quantitatively, compared with the predictions using JS loss and MSE loss, the performance of the predicted distributions using OLDL losses have significant improvements according to the metrics. The model trained by CJS can capture relations between the labels, so that learning for each label can benefit from the adjacent labels.

For Fig. 3 (a) and (b), when only focusing on each individual label, there is little difference between the ground-truth and the predictions of both JS and CJS losses. However, from a holistic perspective, the distribution learned by CJS loss is overall closer to that of ground-truth, which is reflected by results on M_{CJS}. Therefore, CJS divergence (M_{CJS}) can reflect the ordinal distance, overcoming the deficiency of KL divergence. For Fig. 3 (c) and (d), CJS loss achieves the correct top-1 position of the fourth label (score 4), while the positions of the top-1 predictions learned by JS loss and MSE loss are at position of three.

We further present the predicted label distribution as well as the estimated age of JS and CJS in Fig. 4. First, in the results of CJS, the distance between 'Max' prediction and 'Exp' prediction is smaller than that of JS. Second, the prediction learned by OLDL paradigm is closer to ground truth in all the instances.

In most cases, our algorithms can accurately estimate the age based on human faces. In Fig. 5, we show two failure cases of the OLDL methods. For the first instance, the large error of prediction may result from the occlusion on face. Under the weak light, it is even difficult for human to clearly see the face in second image, so the prediction has a certain bias with ground truth.

5. Conclusion

In this paper, we propose a novel paradigm, named OLDL, which is successfully employed in tasks where labels have naturally ordered relations. Instead of computing the per-label difference by existing methods, we explore the sequential patterns of labels from spatial order, semantic order, and temporal order relationships. Based on these characteristics, we design three algorithms for OLDL, termed CAD, QFD, and CJS to further improve network performance. We also introduce order-sensitive metrics based on the distances to evaluate the predicted distribution more reasonably. The experiment results on five datasets of four tasks demonstrate the effectiveness of our methods.

6. Acknowledgments

This work was supported by the National Key Research and Development Program of China Grant (NO. 2018AAA0100400), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), Fundamental Research Funds for the Central Universities, and Supercomputing Center of Nankai University (NKSC).

References

[1] Christian Beecks, Merih Seran Uysal, and Thomas Seidl. Earth mover's distance vs. quadratic form distance: an analytical and empirical comparison. In *ISM*, 2015. 2, 4

[2] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996. 2

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 5

[4] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical models and Methods in Applied Sciences*, 1(4):300–307, 2007. 4

[5] 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*, 2020. 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*, 2020. 3

[7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019. 2

[10] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, 2019. 3

[11] Brigitte Dreno and Florence Poli. Epidemiology of acne. *Dermatology*, 206(1):7–10, 2003. 6

[12] Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al. ChaLearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *CVPRW*, 2016. 4

[13] 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. 4

[14] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *ECML*, 2001. 3

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 3

[16] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017. 3

[17] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, 2018. 1, 3, 5

[18] Yongbiao Gao, Yu Zhang, and Xin Geng. Label enhancement for label distribution learning via prior knowledge. In *IJCAI*, 2020. 3

[19] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016. 1, 2, 4, 5

[20] Xin Geng, Xin Qian, Zengwei Huo, and Yu Zhang. Head pose estimation based on multivariate label distribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1974–1991, 2020. 2

[21] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. In *AAAI*, 2010. 2

[22] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013. 2, 4, 5

[23] Manuel González, José-Ramón Cano, and Salvador García. Prolsfeo-ldl: Prototype selection and label-specific feature evolutionary optimization for label distribution learning. *Applied Sciences*, 10(9):3089–3104, 2020. 2

[24] Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015. 3

[25] James Hafner, Harpreet S. Sawhney, William Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–736, 1995. 4

[26] Nobukazu Hayashi, Hirohiko Akamatsu, Makoto Kawashima, and Acne Study Group. Establishment of grading criteria for acne severity. *Dermatology*, 35(5):255–260, 2008. 6

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[28] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang. Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image Processing*, 26(8):3846–3858, 2017. 1

[29] Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared Earth Mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016. 3

[30] Peng Hou, Xin Geng, Zeng-Wei Huo, and Jia-Qi Lv. Semi-supervised adaptive label distribution learning for facial age estimation. In *AAAI*, 2017. 5

[31] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019. 2

[32] Xiuyi Jia, Zechao Li, Xiang Zheng, Weiwei Li, and Sheng-Jun Huang. Label distribution learning with label correlations on local samples. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1619–1631, 2021. 2

[33] Xiuyi Jia, Xiaoxia Shen, Weiwei Li, Yunan Lu, and Jihua Zhu. Label distribution learning by maintaining label ranking relation. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1695–1707, 2023. 2, 5

[34] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *CVPR*, 2019. 1

[35] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 4

[36] Elizaveta Levina and Peter Bickel. The Earth Mover's distance is the Mallows distance: Some insights from statistics. In *ICCV*, 2001. 3

[37] Peipei Li, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Deep label refinement for age estimation. *Pattern Recognition*, 100:107178, 2020. 2

[38] Qiang Li, Jingjing Wang, Zhaoliang Yao, Yachun Li, Pengju Yang, Jingwei Yan, Chunmao Wang, and Shiliang Pu. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In *CVPR*, 2022. 2, 3

[39] Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *ICPR*, 2018. 2, 4

[40] Miaogen Ling and Xin Geng. Soft video parsing by label distribution learning. *Frontiers of Computer Science*, 13(2):302–317, 2019. 5

[41] Xin Liu and Jufeng Yang. Progressive neighbor consistency mining for correspondence pruning. In *CVPR*, 2023. 2

[42] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *CVPR*, 2018. 3

[43] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, 2015. 4

[44] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 2, 4

[45] Hoang-Vu Nguyen and Jilles Vreeken. Non-parametric jensen-shannon divergence. In *ECML*, 2015. 2, 4

[46] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *CVPR*, 2016. 5

[47] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, 2018. 1, 3

[48] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 2015. 2, 5

[49] Tingting Ren, Xiuyi Jia, Weiwei Li, and Shu Zhao. Label distribution learning with label correlations via low-rank approximation. In *IJCAI*, 2019. 2, 5

[50] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FG*, 2006. 4

[51] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The Earth Mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. 3

[52] J Sánchez-Monedero, Pedro A Gutiérrez, Peter Tiňo, and C Hervás-Martínez. Exploitation of pairwise class distances for ordinal classification. *Neural Computation*, 25(9):2450–2485, 2013. 3

[53] Wei Shen, Kai Zhao, Yilu Guo, and Alan L Yuille. Label distribution learning forests. In *NeurIPS*, 2017. 2

[54] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3):491–504, 2014. 2

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[56] Kate Smith-Miles and Xin Geng. Revisiting facial age estimation with new insights from instance space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2689–2697, 2020. 2

[57] Kai Su and Xin Geng. Soft facial landmark detection by label distribution learning. In *AAAI*, 2019. 1, 3

[58] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007. 1

[59] Jing Wang and Xin Geng. Classification with label distribution learning. In *IJCAI*, 2019. 2

[60] Jing Wang, Xin Geng, and Hui Xue. Re-weighting large margin label distribution learning for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5445–5459, 2021. 2, 5

[61] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 2

[62] Zhangyang Wang, Ding Liu, Shiyu Chang, Florin Dolcos, Diane Beck, and Thomas Huang. Image aesthetics assessment using Deep Chatterjee's machine. In *IJCNN*, 2017. 1

[63] Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *CVPR*, 2023. 2

[64] Xiaoping Wu, Ni Wen, Jie Liang, Yu-Kun Lai, Dongyu She, Ming-Ming Cheng, and Jufeng Yang. Joint acne image grading and counting via label distribution learning. In *ICCV*, 2019. 5, 6, 7

[65] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. SCUT-FBP: a benchmark dataset for facial beauty perception. In *ICSMC*, 2015. 1

[66] Changdong Xu and Xin Geng. Hierarchical classification based on label distribution learning. In *AAAI*, 2019. 1

[67] Ning Xu, Jia-Yu Li, Yun-Peng Liu, and Xin Geng. Trusted-data-guided label enhancement on noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[68] Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *AAAI*, 2019. 3

[69] Ning Xu, An Tao, and Xin Geng. Label enhancement for label distribution learning. In *IJCAI*, 2018. 3

[70] Xue-Qiang Zeng, Su-Fen Chen, Run Xiang, Guo-Zheng Li, and Xue-Feng Fu. Incomplete label distribution learning based on supervised neighborhood information. *International Journal of Machine Learning and Cybernetics*, 11(1):111–121, 2020. 3

[71] Zhicheng Zhang, Song Chen, Zichuan Wang, and Jufeng Yang. Planeseg: Building a plug-in for boosting planar region segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[72] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR*, 2023. 2

[73] Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, 2022. 2

[74] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *ACM MM*, 2015. 1

[75] Haiping Zhu, Yuheng Zhang, Guohao Li, Junping Zhang, and Hongming Shan. Ordinal distribution regression for gait-based age estimation. *Science China Information Sciences*, 63(2):1–14, 2020. 3