DEEP REPULSIVE CLUSTERING OF ORDERED DATA BASED ON ORDER-IDENTITY DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose deep repulsive clustering (DRC) to decompose a partially ordered dataset into multiple disjoint chains by discovering that each pair of data is comparable or not. Comparability of two distinct data is decided by similarity of attributes which is not relevant to order. To this end, we design a feature decomposition network (FD-Net) to separate order information and attribute information. Then, by jointly optimizing FD-Net and repulsive spherical k-means clustering objective function, we cluster the poset into multiple disjoint chains. We decompose a variety of posets including facial image dataset, aesthetic image dataset, and historical image dataset and improve the performance of order learning by using the decomposed chains. Moreover, the knowledge on each dataset discovered by our clustering is provided.

1 INTRODUCTION

Comparison is one of very useful tools for data assessment and analysis. We often compare with the cases in our memory to assess or analysis something. Hence, in computer vision, various studies (Liu, 2009; Parikh & Grauman, 2011; Lee & Kim, 2019; Lim et al., 2020) have utilized the concept of comparison and shown the effectiveness.

Recently, as a comparison-based deep learning algorithm, order learning was proposed Lim et al. (2020). The basic idea of order learning is that ordering relation between instances is easier to predict than the exact ranks in many cases. For instance, telling the older one between two people is easier than estimating the exact ages of peoples. Thus, for age estimation, order learning learns the age order between facial images, and estimate the exact age class of image by aggregating the predicted ordering relations. In other words, by effectively exploiting comparison of data, order learning casts a classification problem to easier form, order estimation problem. Similarly, it can be applied to regression problems such as aesthetic assessment Lee & Kim (2019).

However, in practical situations, comparison of data is not always effective and meaningful. Figure 1 shows some examples of various applications: aesthetic assessment, age estimation, and photo period classification. For these applications, order of classes can be naturally determined in the contexts (*e.g.* which photograph is aesthetically superior than the other). As shown in Figure 1 (a), in aesthetic assessment, predicting the aesthetic order of close-up food and portrait photographs has no advantage over directly estimating the scores of each. This is because aesthetic criteria, which determines the order of photographs, are divergent with respect to the contents of them. On the other hand, as shown in Figure 1 (b), photographs of same contents are easy to estimate the order of them. Similarly, comparing image pairs in Figure 1 (a) and (b), we find that sharing race (or gender) and content is important condition for effective order learning on age estimation and historical period classification.

From these observations, we see that whole features (or information) of an image can be conceptually dichotomized into two types: order feature and complementary feature. Order feature contains information related to order of data. For example, in age estimation, order feature can be information related to aging process including eye wrinkles, hair color, and skin elasticity. Also, photographic skills can be order feature for aesthetic assessment. On the other hand, complementary feature is the rest of information in an image irrelevant to the order. In other words, same complementary feature can be shared by images of different order classes. Hence, identity of people, contents of photograph can be regarded as complementary feature in age estimation, aesthetic assessment, and

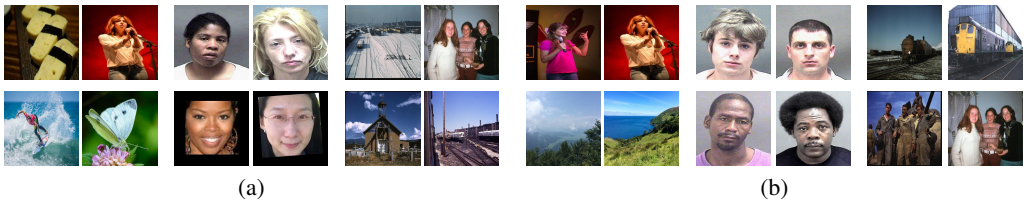


Figure 1: Examples of pairs of data in aesthetic assessment, age estimation, and photo period classification. For all pairs, the right one is {aesthetically superior, looking older, more recently taken} than the left one.

historical period classification. For effective order learning, since the order is naturally determined only for data pairs sharing common complementary features, it is important to find how many dominant complementary features exist in the given dataset, and which complementary features each image contains.

To achieve this, we develop a deep repulsive clustering algorithm (DRC), which jointly optimizes the order-identity decomposition network (ORID) and repulsive spherical k-means clustering. Specifically, ORID encodes order related information and complementary information, separately. Then, repulsive spherical clustering algorithm divide elements in poset into multiple k chains. Section 4 describes our poset decomposition results on various datasets. Also, we show the performances of order learning applications are improved by using decomposed chains. Finally, Section 8 concludes this work.

2 Related Work

2.1 Order Learning

Lim et al. (2020) first proposed the notion of order learning. The objective of order learning is to determine the order graph of classes and classify an object instance into one of the classes. The order graph represents the ordering relationships between classes and it can be either known or unknown depending on the problem situations. Also, instead of directly estimating the label from instance, it trains a comparator to infer the ordering relationship between instances. Then, to estimate the class of a test instance, it compares the test instance with multiple reference instances whose classes are known. By applying the maximum consistency rule (MC rule) on comparison results, the class of test instance is estimated. Similarly, Saaty (1977) proposed the scaling method to estimate absolute priorities from relative priorities. In many cases, As a similar prior work, in aesthetic assessment, Lee & Kim (2019) predict quantized score ratios of pairs of images instead of directly estimating the aesthetic score. The quantized score ratio indicates an image is aesthetically 'superior' or 'similar' or 'inferior' to the other one. Then, they estimate the absolute aesthetic score of an image by using Saaty's scaling method (Saaty, 1977) with predicted quantized score ratios.

2.2 Clustering

Data clustering is a fundamental problem in many areas, such as deep learning, pattern recognition, computer vision, and compression. Given a set of data, the goal of clustering is partition them into disjoint groups, where data in the same group are similar to each other than to those in other groups. Although a large number of clustering methods have been proposed, conventional clustering methods (Hartigan & Wong, 1979; Dhillon & Modha, 2001; Ng et al., 2002; Ester et al., 1996; Kohonen, 1990; Reynolds, 2009) usually held poor performance on high-dimensional data, due to the inefficiency of similarity measure used in these methods. For this reason, dimensionality reduction and feature transform methods have been studied to map the raw data into a new feature space, where the generated feature are easier to be separated. Specifically, linear transform such as PCA (Wold et al., 1987), and non-linear transformations such as kernel methods (Hofmann et al., 2008) and spectral methods (Ng et al., 2002), are proposed.

Recently, since deep neural network has shown its effectiveness as a feature embedding function (LeCun et al., 2015), some approaches (Caron et al., 2018; Yang et al., 2016; Chang et al., 2017) have tried to combine classical clustering algorithms such as k-means, agglomerative clustering, spectral clustering, subspace clustering, with deep neural network. Caron et al. (2018) propose k-means based deep clustering algorithm. It iteratively cluster features from the neural network and trains it using cluster assignments as pseudo-labels. Also, Yang et al. (2016) propose a method that jointly learns feature representations and clusters images by combining the agglomerative clustering and recurrent framework based on neural network. Chang et al. (2017) recast the image clustering task into a binary pairwise classification problem, which is to predict the pair of images belong to either the same clusters or different clusters.

In this work, we provide a deep learning based clustering algorithm of ordered data. As similar to prior studies, we combine the classical spherical k-means algorithm (Dhillon & Modha, 2001) with deep neural network. A distinction should be made between the ordered data and the ordinal data. As a type of categorical data, ordinal data is usually collected from questionnaires, evaluation systems, and so on. Hence, its possible values are a limited number of ordered categories, for example, {poor, neutral, good}.

3 Proposed Algorithm

3.1 Problem Definition

Mathematically, an *order* is a binary relation, often denoted by $\leq$, on a set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ (Schröder, 2003). It should satisfy three properties of reflexivity ($\theta_i \leq \theta_i$ for all i), antisymmetry ($\theta_i \leq \theta_j$ and $\theta_j \leq \theta_i$ imply $\theta_i = \theta_j$), and transitivity ($\theta_i \leq \theta_j$ and $\theta_j \leq \theta_k$ imply $\theta_i \leq \theta_k$). Then, Θ is called a *partially ordered set*. In addition to the three properties, if every pair of elements are comparable ($\theta_i \leq \theta_j$ or $\theta_j \leq \theta_i$ for all i, j), Θ is called a *chain* or *linearly ordered set*.

In practice, an order describes ranks or priorities of classes (Lim et al., 2020). For example, in age estimation, θ_i may represent the age class of i-year-olds. Then, $\theta_{14} \leq \theta_{49}$ represents that 14-year-olds are younger than 49-year-olds. Note that it is less easy to tell the older one between people of different genders than between those of the same gender, as illustrated in Figure 1. An algorithm, hence, may compare a subject with reference subjects of the same gender only. In such a case, the underlying class set Θ is the union of two disjoint chains $\Theta^{\text{female}} = \{\theta_0^{\text{f}}, \theta_1^{\text{f}}, \ldots, \theta_p^{\text{f}}\}$ and $\Theta^{\text{male}} = \{\theta_0^{\text{m}}, \theta_1^{\text{m}}, \ldots, \theta_p^{\text{m}}\}$, where p is the maximum age. There are two age classes θ_i^{f} and θ_i^{m} for each age i, and θ_i^{f} and θ_j^{m} are incomparable with each other for all i, j.

Suppose that a set of n training instances, $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, is given. Also, suppose that there are m ranks and the ground-truth rank of each instance is known. In this sense, $\mathcal{X}$ contains ordered data. The problem is twofold. The first goal is to decompose the class set Θ into k disjoint chains to represent the comparability between instances;

$$\Theta = \bigcup_{l=1}^{k} \Theta^l \tag{1}$$

where $\Theta^i \cap \Theta^j = \emptyset$ for $i \neq j$, each chain is given by $\Theta^l = \{\theta_1^l, \theta_2^l, \ldots, \theta_m^l\}$, and the number k of chains is assumed to be known *a priori*. In other words, we attempt to partition the ordered data in $\mathcal{X}$ into k clusters (or chains), by grouping them according to their characteristics unrelated to their ranks. These characteristics, which tend to remain the same even though an object experiences rank changes, are referred to as 'identity' features in this work. For example, in age estimation, genders or races can be identity features. However, the clustering is performed without any annotation data for identity features. Notice that instances within a chain would be more easily compared than those across chains are, since they have similar identity features. The second goal is to assign an unseen test instance y into one of the chains and determine its rank by comparing it with reference instances within the chain. To achieve these goals, we propose the order-identity decomposition (ORID) and the deep repulsive clustering (DRC).

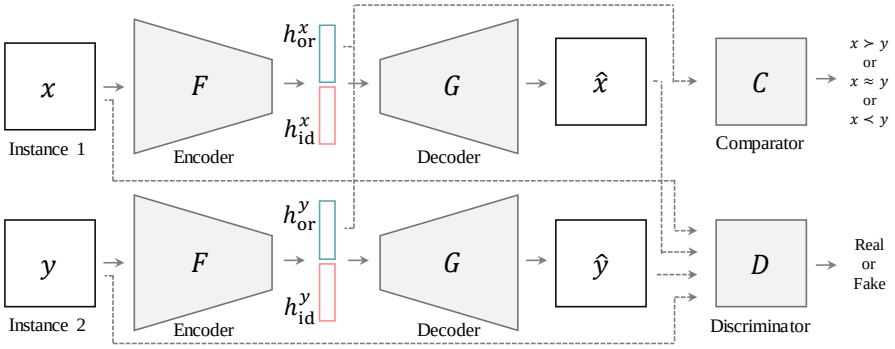


Figure 2: An overview of the ORID network.

3.2 Order-Identity Decomposition

In general, object instances can be compared more easily, as they have more similar identity features, which are irrelevant to order. Therefore, we decompose the information of each object instance into an order feature and an identity feature. To achieve such decomposition, we propose the ORID network in Figure 2, composed of three parts: autoencoder, discriminator, and comparator.

1) Autoencoder: Similarly to various deep clustering algorithms (Yang et al., 2017; Dizaji et al., 2017; Chen et al., 2017a; Ji et al., 2017), we use the autoencoder $G \circ F(\cdot)$, based on a neural network, to extract feature vectors. The encoder $h^x = F(x)$ maps an input vector x to a feature vector h^x, while the decoder $\hat{x} = G(h^x)$ reconstructs $\hat{x}$ from h^x. By minimizing the reconstruction loss $\|x - \hat{x}\|_1$, the encoder F is trained to compactly represent the input instance with as little loss of information as possible. We decompose the overall feature $h^x \in \mathbb{R}^{d_{\text{or}}+d_{\text{id}}}$ into the order feature h^x_{or} and the identity feature h^x_{id}, given by

$$h^x_{\text{or}} = [h^x_1, h^x_2, \ldots, h^x_{d_{\text{or}}}]^T \tag{2}$$

$$h^x_{\text{id}} = \frac{[h^x_{d_{\text{or}}+1}, h^x_{d_{\text{or}}+2}, \ldots, h^x_{d_{\text{or}}+d_{\text{id}}}]^T}{\|[h^x_{d_{\text{or}}+1}, h^x_{d_{\text{or}}+2}, \ldots, h^x_{d_{\text{or}}+d_{\text{id}}}]\|} \tag{3}$$

where d_{or} and d_{id} are the dimensions of h^x_{or} and h^x_{id}. However, without additional control, the output h^x of the deep network F tends to be highly entangled (Higgins et al., 2018). Therefore, to put together order-related information into h^x_{or}, we employ the comparator.

2) Comparator: Using the order features h^x_{or} and h^y_{or} of a pair of instances x and y, we train the comparator, which classifies their ordering relationship into one of three categories:

$$x \succ y \quad \text{if } \theta(x) - \theta(y) > \tau, \tag{4}$$

$$x \approx y \quad \text{if } |\theta(x) - \theta(y)| \leq \tau, \tag{5}$$

$$x \prec y \quad \text{if } \theta(x) - \theta(y) < \tau, \tag{6}$$

where $\theta(\cdot)$ denotes the class of an instance. As in (Lim et al., 2020), '$\succ, \approx, \prec$' represent the ordering relationship between instances, while '$>, =, <$' do the mathematical order between classes. The comparator outputs the softmax probability $p^{xy} = (p^{xy}_{\succ}, p^{xy}_{\approx}, p^{xy}_{\prec})$. It is trained to minimize the cross-entropy between p^{xy} and the ground-truth one-hot vector $q^{xy} = (q^{xy}_{\succ}, q^{xy}_{\approx}, q^{xy}_{\prec})$. Since it is jointly trained with the autoencoder, the information deciding the ordering relationship tends to be encoded into the order features h^x_{or} and h^y_{or}. On the other hand, the remaining information necessary for the reconstruction of $\hat{x}$ and $\hat{y}$ are encoded into the identity features h^x_{id} and h^y_{id}.

3) Discriminator: We adopt the discriminator D that tells real images from synthesized images, generated by the decoder G. Using the GAN loss (Goodfellow et al., 2014), the discriminator helps the decoder to reconstruct more realistic output $\hat{x}$ and $\hat{y}$.

Appendix A provides detailed network structures of these components in ORID.

3.3 DEEP REPULSIVE CLUSTERING

After obtaining the identity features $h_{\text{id}}^{x_1}, h_{\text{id}}^{x_2}, \ldots, h_{\text{id}}^{x_n}$ of all instances $x_i \in \mathcal{X}$ based on ORID, we partition them into k clusters, which correspond to k disjoint chains. Each chain contains instances comparable to one another. The identity features are normalized in Eq. (3) and lie on the unit sphere in $\mathbb{R}^{d_{\text{id}}}$. Thus, the cosine similarity is a natural affinity metric for identity features. Let $\mathcal{C}_j$, $1 \leq j \leq k$, denote the k clusters. Also, let c_j on the unit sphere denote the 'centroid' or the representative vector for the instances in cluster $\mathcal{C}_j$. Then, we define the quality of cluster $\mathcal{C}_j$ as

$$\frac{1}{|\mathcal{C}_j|} \sum_{x \in \mathcal{C}_j} \Big((h_{\text{id}}^x)^t c_j - \alpha \frac{1}{k-1} \sum_{l \neq j} (h_{\text{id}}^x)^t c_l \Big) \tag{7}$$

where $|\mathcal{C}_j|$ is the cardinality of $\mathcal{C}_j$. In Eq. (7), the first term is the similarity of an instance in $\mathcal{C}_j$ to the centroid c_j, the second term with the negative sign quantifies the average dissimilarity of the instance from the other centroids, and α is a weight. For a high quality cluster, instances should be concentrated around the centroid and be far from the other clusters. The second term is referred to as the repulsive term, as its objective is similar to the repulsive rule in (Lee et al., 2015). Next, we measure the overall quality of the clustering by

$$J(\{\mathcal{C}_j\}_{j=1}^k, \{c_j\}_{j=1}^k) = \sum_{j=1}^{k} \frac{1}{|\mathcal{C}_j|} \sum_{x \in \mathcal{C}_j} \Big((h_{\text{id}}^x)^t c_j - \alpha \frac{1}{k-1} \sum_{l \neq j} (h_{\text{id}}^x)^t c_l \Big). \tag{8}$$

We attempt to find the optimum clusters to maximize this objective function J, yet finding the global optimum is NP-complete (Kleinberg et al., 1998; Garey et al., 1982). Hence, we propose an iterative algorithm, called DRC, to find a locally optimum solution, as done in the k-means algorithm (Gersho & Gray, 1991).

1. Centroid rule: After fixing the clusters $\{\mathcal{C}_j\}_{j=1}^k$, we update the centroids $\{c_j\}_{j=1}^k$ to maximize J in Eq. (8). Because the centroids should lie on the unit sphere, we solve the constrained optimization problem:

$$\text{minimize } \; J(\{c_j\}_{j=1}^k) \; \text{ subject to } \; c_j^t c_j = 1 \; \text{ for all } \; j = 1, \ldots, k. \tag{9}$$

   Using Lagrangian multipliers (Bertsekas, 1996), the optimal centroids are obtained as

$$c_j = \frac{\frac{1}{|\mathcal{C}_j|} \sum_{x \in \mathcal{C}_j} h_{\text{id}}^x - \alpha \frac{1}{k-1} \sum_{l \neq j} \frac{1}{|\mathcal{C}_l|} \sum_{x \in \mathcal{C}_l} h_{\text{id}}^x}{\| \frac{1}{|\mathcal{C}_j|} \sum_{x \in \mathcal{C}_j} h_{\text{id}}^x - \alpha \frac{1}{k-1} \sum_{l \neq j} \frac{1}{|\mathcal{C}_l|} \sum_{x \in \mathcal{C}_l} h_{\text{id}}^x \|}. \tag{10}$$

2. Nearest neighbor (NN) rule: On the other hand, after fixing the centroids, we update the membership of each instance to maximize J in Eq. (8). It can be easily derived that the optimal cluster $\mathcal{C}_j$ is given by

$$\mathcal{C}_j = \big\{ x | (h_{\text{id}}^x)^t c_j \geq (h_{\text{id}}^x)^t c_l \text{ for all } 1 \leq l \leq k \big\}. \tag{11}$$

   In other words, an instance should be assigned to $\mathcal{C}_j$ if its nearest centroid is c_j.

3. Regularization: To prevent empty clusters and balance the partitioning, we enforce a regularization constraint so that every cluster contains at least a predefined number of instances.

4. Stopping criterion: We perform steps 1, 2, and 3 iteratively, until the ratio of membership-changing instances is less than a threshold.

Let us describe the impacts of the repulsive term in Eq. (8). Without the repulsive term (*i.e.* $\alpha = 0$), c_j in Eq. (10) is updated by the normalized centroid $\frac{1}{|\mathcal{C}_j|} \sum_{x \in \mathcal{C}_j} h_{\text{id}}^x$, as done in the spherical k-means (Dhillon & Modha, 2001). On the other hand, with a positive α, the objective function J is reduced when the centroids are far from one another. Ideally, the centroid of a cluster should be the opposite of the centroid of all the other clusters;

$$\left(\frac{\sum_{x \in \mathcal{C}_j} h_{\text{id}}^x}{\| \sum_{x \in \mathcal{C}_j} h_{\text{id}}^x \|} \right)^t \left(\frac{\sum_{x \in \mathcal{X} \setminus \mathcal{C}_j} h_{\text{id}}^x}{\| \sum_{x \in \mathcal{X} \setminus \mathcal{C}_j} h_{\text{id}}^x \|} \right) = -1, \quad \text{for all } j = 1, 2, \ldots, k. \tag{12}$$

Note that the ORID network and thus the encoded feature space are trained jointly with the repulsive clustering. As the training goes on, the centroids repel one another, and the clusters are separated more effectively due to the repulsive term.

Algorithm 1 DRC-ORID

Input: Ordered data $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, k = the number of chains

Train ORID network for warm-up epochs to minimize loss $\lambda_{\text{rec}}\ell_{\text{rec}} + \lambda_{\text{com}}\ell_{\text{com}} + \lambda_{\text{gan}}\ell_{\text{gan}}$
Partition $\mathcal{X}$ into $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ using k-means
repeat
  Fine-tune ORID network to minimize loss $\lambda_{\text{rec}}\ell_{\text{rec}} + \lambda_{\text{clu}}\ell_{\text{clu}} + \lambda_{\text{com}}\ell_{\text{com}} + \lambda_{\text{gan}}\ell_{\text{gan}}$
  repeat
    for all $j = 1, 2, \ldots, k$ **do** ▷ Centroid Rule
      Update centroid c_j via Eq. (10)
    end for
    for all $j = 1, 2, \ldots, k$ **do** ▷ NN Rule
      Update chain $\mathcal{C}_j$ via Eq. (11)
    end for
  until convergence or predefined number of iterations
until predefined number of epochs

Output: Chains $\{\mathcal{C}_j\}_{j=1}^k$, centroids $\{c_j\}_{j=1}^k$, ORID network

3.4 DRC-ORID

We jointly optimize the chains (or clusters) and the ORID network parameters, as described in **Algorithm 1**. First, we train the ORID network for warm-up epochs, assuming that every pair of instances x and y are comparable. Then, using the identity features, we partition the input data into k chains by employing k-means. Second, we repeat the fine-tuning of the ORID network and the repulsive clustering alternately. In the fine-tuning, a pair of x and y are constrained to be from the same chain, and the following loss function is employed.

$$\ell = \lambda_{\text{rec}}\ell_{\text{rec}} + \lambda_{\text{clu}}\ell_{\text{clu}} + \lambda_{\text{com}}\ell_{\text{com}} + \lambda_{\text{gan}}\ell_{\text{gan}} \tag{13}$$

where the reconstruction, clustering, comparator, and GAN losses are given by

$$\ell_{\text{rec}} = \frac{1}{2N}\sum_{i=1}^{N}\big(\|x_i - F(G(x_i))\|_1 + \|y_i - F(G(y_i))\|_1\big), \tag{14}$$

$$\ell_{\text{clu}} = -\frac{1}{2N}\sum_{i=1}^{N}\big((h_{\text{id}}^{x_i})^t c_j + (h_{\text{id}}^{y_i})^t c_j\big), \tag{15}$$

$$\ell_{\text{com}} = -\frac{1}{N}\sum_{i=1}^{N}\big(q_{\succ}^{x_i y_i}\log p_{\succ}^{x_i y_i} + q_{\approx}^{x_i y_i}\log p_{\approx}^{x_i y_i} + q_{\prec}^{x_i y_i}\log p_{\prec}^{x_i y_i}\big), \tag{16}$$

$$\ell_{\text{gan}} = -\frac{1}{2N}\sum_{i=1}^{N}\big(\log(1 - D(F(G(x_i)))) + \log(1 - D(F(G(y_i))))\big), \tag{17}$$

respectively. Here, N is the number of image pairs in a minibatch. Also, the weighting parameters are set to $\lambda_{\text{rec}} = 5$, $\lambda_{\text{clu}} = 0.1$, $\lambda_{\text{com}} = 1$, and $\lambda_{\text{gan}} = 1$.

3.5 Estimation

Given an ordered dataset, by applying DRC, we first decide the chains and the chain membership of each instance. Then, instead of directly estimating the class of each instances, we learn pairwise ordering relationship between two instances as in Lim et al. (2020). Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two instances belonging to classes in the same chain. In other words, $\boldsymbol{x}$ and $\boldsymbol{y}$ are considered as a comparable instance pair.

Similar to Lim et al. (2020), we train a pairwise comparator to predict the instance order in (equation **??**). As shown in Figure **??**, our pairwise comparator consists of two parts: a feature extractor and a ternary classifier.

We estimate the class $\theta(\boldsymbol{x})$ of a test instance $\boldsymbol{x}$ by comparing it with reference instances $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$, where M denotes the number of references. The references are selected from training data in the same chain with $\boldsymbol{x}$. Given $\boldsymbol{x}$ and $\boldsymbol{y}_m$, our comparator predicts the probabilities for three possible ordering relation '$\succ, \approx, \prec$.' where $[\cdot]$ is the indicator function.

$$\begin{aligned}\hat{p}_i(\theta) = {} & p([\boldsymbol{x} \succ \boldsymbol{y}] = 1)\hat{p}(\theta \mid [\boldsymbol{x} \succ \boldsymbol{y}] = 1) \\ & + p([\boldsymbol{x} \approx \boldsymbol{y}] = 1)\hat{p}(\theta \mid [\boldsymbol{x} \approx \boldsymbol{y}] = 1) \\ & + p([\boldsymbol{x} \prec \boldsymbol{y}] = 1)\hat{p}(\theta \mid [\boldsymbol{x} \prec \boldsymbol{y}] = 1)\end{aligned} \tag{18}$$

Figure 3: Two chain decomposition results on MORPH II. Each row are examples in the same chain.

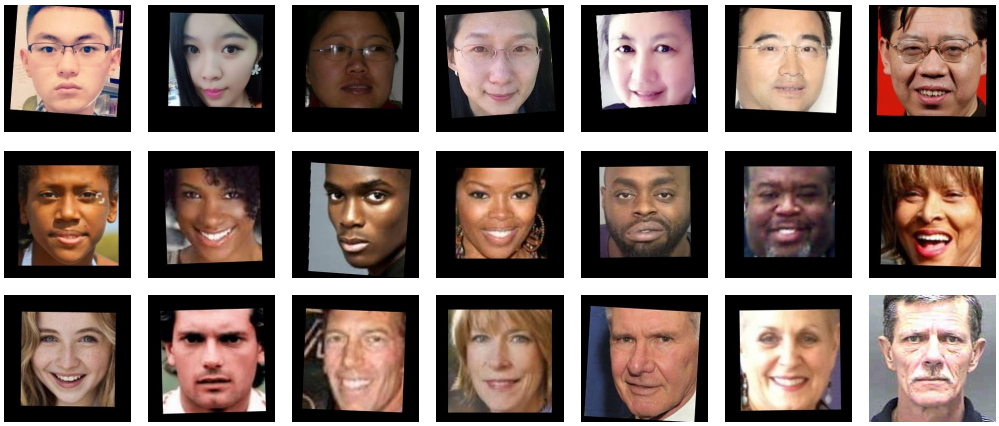


Figure 4: Three chain decomposition results on Balanced dataset. Each row are examples in the same chain.

$$\hat{p}_i(\theta) = \sum_{l=0}^{2} p(q_i = l)\hat{p}(\theta|q_i = l) \tag{19}$$

4 Age Estimation

We evaluate the performance of our DRC algorithm on two age estimation datasets, MORPH II (Rothe et al., 2018) and balanced dataset (Lim et al., 2020). Also, we develop an age estimator based on order learning, and show that the age estimation performance is improved when using disjoint chains.

4.1 Implementation Details

For chain decomposition, we initialize OID-Net parameters with Glorot normal (Glorot & Bengio, 2010). We use Adam optimizer with initial learning rate of 0.0001 and decrease it by a factor of 0.5 every $50,000$ steps. For data augmentation, we only use the random horizontal flip. This is because other popular augmentation methods including brightness or contrast modification may deform important information for clustering such as skin color in facial image dataset. Also, we set α to be 0.1 and decrease it to 0.05 after 200 epoch to avoid the oscillation.

For age estimation, we train the comparator

Table 1: Performance comparison on the MORPH II dataset. The best results are boldfaced, and the second best ones are underlined.

Model	Setting A		Setting B		Setting C		Setting D	
	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)	MAE	CS (%)
Ranking-CNN (Chen et al., 2017b)	-	-	-	-	-	-	2.96	85.0
DMTL (Han et al., 2017)	-	-	-	-	3.00	85.3	-	-
DEX (Rothe et al., 2018)	2.68	-	-	-	-	-	-	-
DRFs (Shen et al., 2018)	2.91	82.9	2.98	-	-	-	2.17	91.3
MO-CNN (Tan et al., 2017)	2.52	85.0	2.70	-	-	-	-	-
MV (Pan et al., 2018)	-	-	-	-	2.80	87.0	2.41	90.0
BridgeNet (Li et al., 2019)	2.38	91.0	2.63	86.0	-	-	-	-
OL (Lim et al., 2020)	2.41	91.7	2.75	88.2	2.68	88.8	2.22	93.3
Proposed (1CH)	2.35	92.4	2.64	88.9	2.64	89.0	2.22	93.3
Proposed (CD)	**2.26**	**93.8**	**2.55**	**89.6**	**2.58**	**89.5**	**2.17**	**93.5**

Table 2: Performance comparison on the balanced dataset. The best results are boldfaced, and the second best ones are underlined.

Model	MAE				CS (%)			
	1CH	2CH	3CH	6CH	1CH	2CH	3CH	6CH
MV (Pan et al., 2018)	4.49	4.52	4.44	4.40	69.9	70.1	70.3	69.6
OL-supervised (Lim et al., 2020)	4.23	4.18	4.19	4.18	73.2	73.4	73.4	74.0
OL-unsupervised	-	4.16	4.17	4.16	-	74.0	73.9	74.0
Proposed	**4.15**	**4.01**	**4.02**		**74.0**	**74.8**	**74.8**	

4.2 DATASETS

MORPH II: This dataset (Rothe et al., 2018) is a collection of about $55,000$ facial images in the age range $[16, 77]$. It also provides gender (female, male) and race (black, white) labels. On this dataset, we employ four different evaluation protocols: setting A, B, C, and D. Detailed descriptions on each of these settings are provided in Appendix.

Balanced dataset: Since most facial age datasets (Rothe et al., 2018; Niu et al., 2016; Zhang et al., 2017; Panis et al., 2016) are biased to specific ethnic groups or genders, Lim et al. (2020) make the balanced dataset by sampling data from popular age estimation datasets, MORPH II (Rothe et al., 2018), AFAD (Niu et al., 2016), and UTK (Zhang et al., 2017). It contains about $6,000$ images for each combination of gender in $\{$female, male$\}$ and ethnic group in $\{$African, Asian, European$\}$.

4.3 CLUSTERING

As in other clustering algorithms (Caron et al., 2018; Xie et al., 2016), we use the Normalized Mutual Information (NMI) for our metric. Specifically, it measure the information shared between two different partitioning A and B of the same data as

$$\text{NMI}(A;B) = \frac{I(A;B)}{\sqrt{H(A)H(B)}} \tag{20}$$

where $I(\cdot)$ denotes the mutual information and $H(\cdot)$ the information entropy.

Results on MORPH II: We apply the DRC to decompose setting 2 of MORPH II into two chains. Setting 2 of MORPH II contains balanced number of black and white face images. Examples of clustering result is shown in Figure 3. We see that chains are well separated according to skin colors. Also, Figure 8 shows that NMI measured using the race labels increases to 0.9 as the training progresses.

Table 1 compares age estimation performance of the proposed algorithm with those of various conventional algorithms. In all settings, the proposed algorithm with chain decomposition (CD) outperforms all conventional algorithms without exception. Compared to single chain hypothesis (1CH), chain decomposition reduces MAE by 0.09 in average.

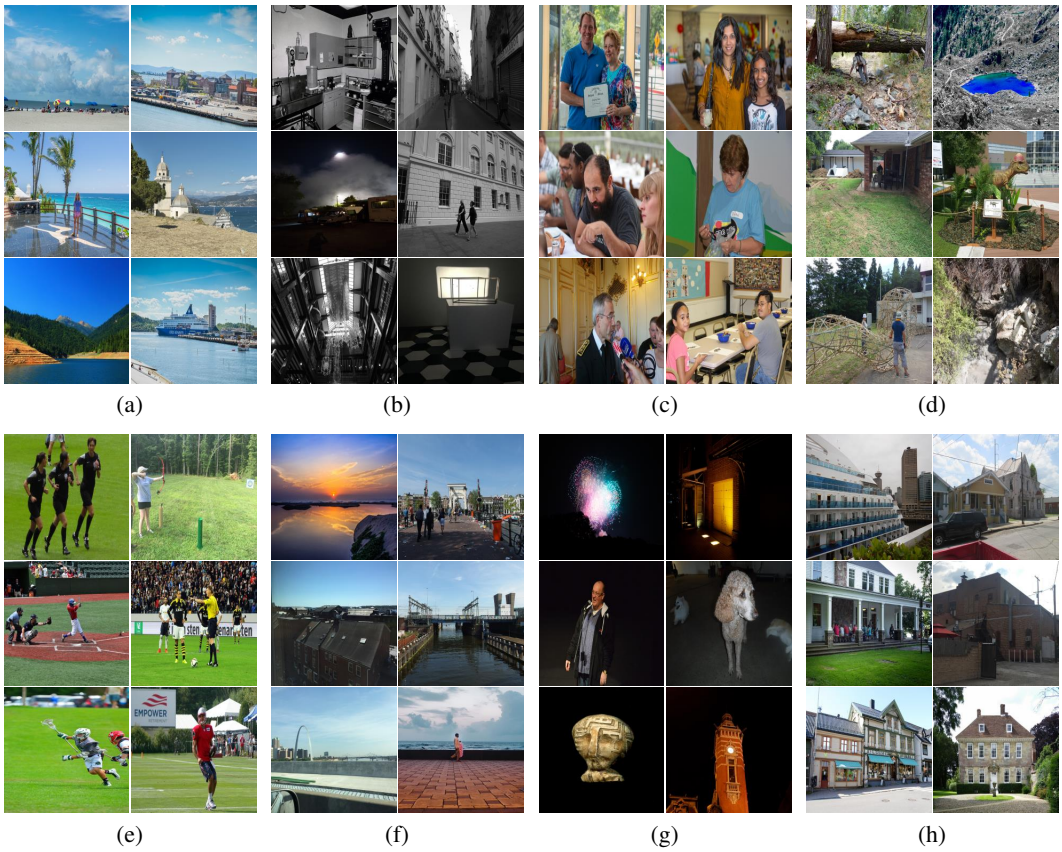


Figure 5: Example images that are grouped into multiple chains by applying DRC algorithm to the AADB dataset. For each chain, we observe the common attribute: (a) blue background, (b) grayscale, (c) people, (d) nature background, (e) people with green background, (f) horizontal decomposition, (g) object with black background, and (h) urban scenery.

Results on Balanced: Figure 4 shows the clustering results of our DRC on the balanced dataset. We divide it into three disjoint chains, which are separated according to race (African, Caucasian, and Asian).

Table 2 compares the age estimation performances on the balanced dataset according to the number of chains.

4.4 Estimation

As in previous studies (Chen et al., 2017b; Han et al., 2017; Rothe et al., 2018; Shen et al., 2018; Tan et al., 2017; Pan et al., 2018; Li et al., 2019; Lim et al., 2020), we measure the performance of age estimation by the mean absolute error (MAE) and the cumulative score (CS). MAE is the average absolute error between estimated and ground-truth ages. Also, CS computes the percentage of test samples whose absolute errors are less than or equal to a tolerance level. In this work, the tolerance level is fixed to 5 as in Han et al. (2017); Shen et al. (2018), and Lim et al. (2020).

5 Aesthetic Assessment

For the feature extractor and comparator, we use the similar network architecture with the baseline Lee & Kim (2019). We detail the network we used for aesthetic assessment in the Appendix including learning rate, data augmentation, network architecture, and hyper parameters of scaling method.

(a) (b) (c) (d)

Figure 6: Example images that are grouped into multiple chains by applying DRC algorithm to the HCI dataset. For each chain, we observe the common attribute: (a) people, (b) portrait, (c) scenery, and (d) urban scenery.

Table 3: Performance comparison on the AADB dataset. The best results are boldfaced.

Model	MAE
Reg-Net (Pan et al., 2018)	0.1268
Lee & Kim (2019)	0.1141
Proposed (1CH)	0.1109
Proposed (CD)	**0.1056**

AADB: The aesthetics and attribute database (AADB) provides $10,000$ photographic images of various themes such as scenery, portrait, and close-up, for aesthetic score assessment task. These images are divided into $8,500$ for training, 500 for validation, and $1,000$ for test. The aesthetic score of each image was evaluated by five annotators and then averaged. Also, aesthetic scores range from 0 to 1 with 1 denoting the highest quality.

Clustering: For AADB dataset, we set $k = 8$ since AADB dataset contains images with various contents. Figure 5 shows qualitative results of our clustering on the AADB dataset.

6 Historical Color Image Classification

HCI: Historical color image dataset (HCI) (Palermo et al., 2012) is for the problem of determining the decade during which a historical color photograph was taken. The dataset provides images collected from five decades, 1930s to 1970s. Each five decade categories has 265 images, and thus 1350 images are provided in total. For fair comparison, the evaluation protocol reported in Palermo et al. (2012) is taken in this study.

7 Analysis

7.1 Feature Decomposition

Age transformation: To evaluate the quality of our feature decomposition, we perform age transformation on MORPH II dataset. Age transformation is synthesizing a person's face in a different age (target age) while preserving the person's identity (Or-El et al., 2020). Given an input image , we first extract order feature and identity feature using ORID encoder. Note that order and identity feature encodes information relevant to age and identity of people, respectively. Hence, by decoding the replacing order feature of input image with that of other image of different age, replace the order feature of the input with the order feature of reference image. We feed the concatenated feature to our decoder. Our decoder fuses the age representation with the identity features and then generates

Table 4: Performance comparison on the historical color image dataset. The best results are bold-faced, and the second best ones are underlined.

Model	Accuracy (%)	MAE
Palermo et al. (Palermo et al., 2012)	44.92±3.69	0.93±0.08
Martin et al. (Martin et al., 2014)	42.76±1.33	0.87±0.05
Frank and Hall (Frank & Hall, 2001)	41.36±1.89	0.99±0.05
Cardoso et al. (Cardoso & da Costa, 2007)	41.31±2.76	0.95±0.04
RED-SVM@8168 (Lin & Li, 2012)	35.92±4.69	0.96±0.06
RED-SVM@deep (Lin & Li, 2012)	25.38±2.34	1.08±0.05
ORCNN (Niu et al., 2016)	44.67±2.54	0.81±0.06
CNNPOR (Liu et al., 2018)	50.12±2.65	0.82±0.05
GP-DNNOR (Liu et al., 2019)	46.60±2.98	0.76±0.05
Proposed (1CH)	44.8	0.82

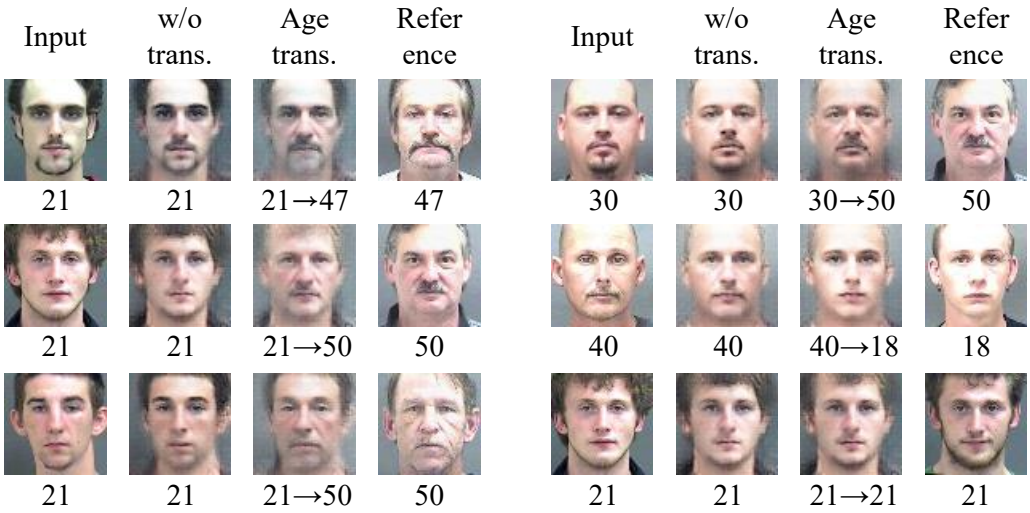


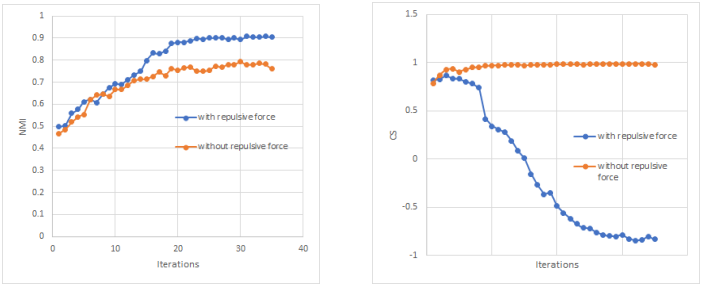
Figure 7: Age transformation results on MOPRH II dataset.

the age transformed facial image. Figure 7 shows some age transformation results. Our transformations to different ages naturally change the skin textures, hair colors, and wrinkles, but not deform the identities of people. This result implicitly shows the reliability of our feature decomposition.

To verify this hypothesis, we conduct experiments that use either order feature or attribute feature to learn to predict . Table shows order feature contains much more information about age, while attribute feature

8 Conclusions

Deep repulsive clustering was proposed in this work. It solves the problem of decomposing posets into multiple disjoint chains by discovering comparable pairs using deep learning base clustering method. Comparability of two distinct data is decided by similarity of attributes which is not relevant to order. To this end, we design a feature decomposition network (OID-Net) to separate order information and attribute information. Then, by jointly optimizing FD-Net and repulsive spherical k-means clustering objective function, we cluster the poset into multiple disjoint chains. We decompose a variety of posets including aesthetic image dataset, facial image dataset, and historical image dataset and improve the performance of order learning by using the decomposed chains. Moreover, the knowledge on each dataset discovered by our clustering is provided.



(a) (b)

Figure 8: Comparison of (a) NMI and (b) CS

REFERENCES

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1996.

Jaime S. Cardoso and Joaquim F. Pinto da Costa. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8:1393–1429, 2007.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *ICCV*, 2017.

Dongdong Chen, Jiancheng Lv, and Zhang Yi. Unsupervised multi-manifold clustering by learning deep representation. In *AAAI Workshops*, 2017a.

Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age estimation. In *CVPR*, 2017b.

Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1–2):143–175, 2001.

Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *ICCV*, 2017.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *ECML-PKDD*, 2001.

M. R. Garey, D. S. Johnson, and Hans S. Witsenhausen. The complexity of the generalized Lloyd-Max problem. *IEEE Trans. Inf. Theory*, 28(2):255–256, 1982.

A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers Norwell, 1991.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Hu Han, Anil K. Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40 (11):2597–2609, 2017.

J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100–108, 1979.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. In *arXiv preprint arXiv:1812.02230*, 2018.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pp. 1171–1220, 2008.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *NIPS*, 2017.

Jon Kleinberg, Christos Papadimitriou, and Prabhakar Raghavan. A microeconomic view of data mining. *Data Mining and Knowledge Discovery*, 2(4):311–324, 1998.

Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Chulwoo Lee, Won-Dong Jang, Jae-Young Sim, and Chang-Su Kim. Multiple random walkers and their application to image cosegmentation. In *CVPR*, 2015.

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*, 2019.

Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. BridgeNet: A continuity-aware probabilistic network for age estimation. In *CVPR*, 2019.

Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2020.

Hsuan-Tien Lin and Ling Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367, 2012.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

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

Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on Gaussian processes. In *ICCV*, 2019.

Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. In *Proc. ACM ICMR*, 2014.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.

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

Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, 2020.

Frank Palermo, James Hays, and Alexei A. Efros. Dating historical color images. In *ECCV*, 2012.

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

Gabriel Panis, Andreas Lanitis, Nicholas Tsapatsoulis, and Timothy F. Cootes. Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*, 5(2):37–46, 2016.

Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011.

Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, 741, 2009.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126(2-4):144–157, 2018.

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

Bernd S. W. Schröder. *Ordered Sets: An Introduction*. Springer, 2003.

Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Yuille. Deep regression forests for age estimation. In *CVPR*, 2018.

Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z. Li. Efficient group-n encoding and decoding for facial age estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11): 2610–2623, 2017.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, 2017.

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, 2016.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017.

A DETAILED NETWORK STRUCTURE OF ORID

B MORE EXPERIMENTS

B.1 SETTINGS

We adopt four widely used testing protocols for evaluation.

Setting A: $5,492$ images of Caucasian race are selected and then randomly divided into two parts: 80% of data for training set and 20% of data for test set. There is no overlapping between training and test sets.

Setting B: $21,000$ images of Africans and Caucasians are randomly to satisfy two constraints: the ratio between Africans and Caucasians to $1:1$ and that between females and males to $1:3$. They are split into three non-overlapped subsets, S1, S2, and S3. The training and testing should be repeated twice in two subsettings: 1) training on S1, testing on S2 + S3, and 2) training on S2, testing on S1 + S3. The average performance of two experiments is reported.

Setting C: This setting is the 5-fold cross-validation on the entire dataset. Images are randomly split into five folds, but the same person's images must belong to only one fold.

Setting D: This setting is called 80-20 protocol.

C PROOFS

Also, centroid similarity (CS) is a measurement for discrimination of clusters defined as

$$\mathrm{CS}(\{\boldsymbol{c}\}_{j=1}^{k}) = \frac{2}{k(k-1)} \sum_{j=1}^{k} \sum_{l=1}^{k} \mathbf{1}_{j>l} \boldsymbol{c}_j^T \boldsymbol{c}_l \tag{21}$$