

# CONTRASTIVE ORDER LEARNING FOR ORDINAL REGRESSION

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

We propose a novel contrastive learning algorithm for ordinal regression, called contrastive order learning (ConOrd), which combines the strengths of order learning and contrastive learning effectively. While contrastive learning excels at leveraging all samples in a batch, it often overlooks the inherent order among rank labels. Conversely, order learning superbly captures label ordinality but typically relies on local margin-based comparisons, limiting its global consistency and representation power. ConOrd addresses these limitations by introducing a contrastive order loss, which adopts soft affinity and disparity weights based on rank differences, enabling fine-grained modeling of ordinal relationships across all sample pairs in a batch. Moreover, to improve the embedding space further, we incorporate a center loss with learnable reference points, which guide compact clustering and ordinal alignment. Extensive experiments on various ordinal regression tasks — including facial age estimation, blind image quality assessment, and blind video quality assessment — show that the proposed ConOrd consistently yields state-of-the-art results and generalizes well to diverse ordinal regression scenarios.

## 1 INTRODUCTION

Ordinal regression is a task to estimate the discrete or continuous rank of an object instance. For example, facial age estimation aims to predict a person’s age given their facial photograph, while image quality assessment predicts the quality score for an image. It is a fundamental problem frequently arising in many real-world scenarios, including facial age estimation (Moschoglou et al., 2017), health status scoring (Engemann et al., 2022), image and video quality assessment (Ying et al., 2021; Hosu et al., 2017), and gaze direction estimation (Wang et al., 2022).

Despite its wide applicability, ordinal regression poses inherent challenges: there is no clear distinction between adjacent ranks, and the semantic gap between neighboring labels can be subtle or ambiguous. It is hence difficult for a machine to learn discriminative representations reflecting the underlying ordinal structure accurately. To face these challenges, various methods have been proposed (Li & Lin, 2007; Rothe et al., 2018; Geng et al., 2013; Diaz & Marathe, 2019). Recently, order learning techniques (Lim et al., 2020; Lee & Kim, 2021; Lee et al., 2022) have achieved notable success. Among these, geometric order learning (GOL) (Lee et al., 2022) emerged as an effective scheme for learning representations of ordinal data, which enforces metric and order constraints to arrange instances according to their ranks in an embedding space. Despite its proficiency in capturing ordinal structure, GOL is based on a margin-based triplet loss, which adopts a fixed threshold on the distance between instances. Once the margin is satisfied, no gradient is generated, hindering the learning of fine-grained ordinal relationships. Moreover, as a pairwise approach, it cannot fully exploit the rich ordinal context available in an entire batch.

Meanwhile, supervised contrastive learning (Khosla et al., 2020) has been extended to ordinal regression tasks, *e.g.*, in the RnC algorithm (Zha et al., 2023). Whereas supervised contrastive learning uses categorical labels to define positive and negative pairs in a batch, RnC constructs rank-aware pairs. Specifically, RnC first selects an anchor and a positive sample in a batch. Then, it declares other samples, whose rank differences from the anchor are bigger than the rank difference between the anchor and the positive, as negatives. However, because of this construction, RnC may fail to exploit informative samples with small rank differences, resulting in under-utilization of data.

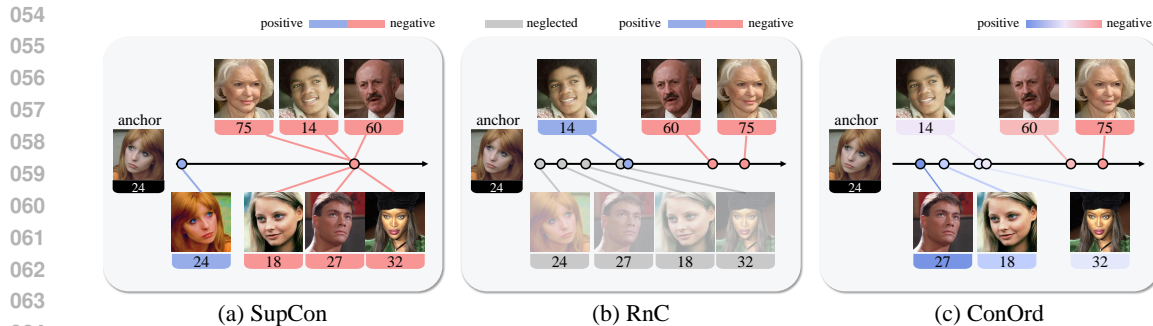


Figure 1: Comparison of three contrastive learning techniques: (a) supervised contrastive learning (SupCon) (Khosla et al., 2020), (b) Rank-N-Contrast (RnC) (Zha et al., 2023), and (c) the proposed ConOrd. SupCon considers the augmented view of an anchor as the only positive and treats all the others as negatives. RnC selects one sample as a positive and defines negatives as those with rank differences greater than the anchor-positive pair. Different from these existing techniques, ConOrd compares *all* samples in a batch by defining positive and negative samples in a *soft* manner.

In this paper, we propose contrastive order learning (ConOrd), which combines the advantages of both order learning and contrastive learning and alleviates their limitations. Order learning can explicitly model the ordinal nature of data in an embedding space, *i.e.*, attraction of samples with similar ranks and repulsion of those with dissimilar ranks (Lee et al., 2022). However, it uses pairwise losses or margin losses, which rely on limited comparisons such as pairs or triplets. This local supervision may be insufficient for building a globally consistent embedding space. Contrastive learning, on the other hand, excels at leveraging the relationships among all samples within a batch, but it may fail to exploit the ordered nature of the labels fully, sometimes discarding informative samples with nearby ranks. The proposed ConOrd brings together the complementary strengths of both approaches. As illustrated in Figure 1, ConOrd incorporates a dynamic affinity mechanism into a contrastive learning framework that adjusts attraction or repulsion between samples based on their rank differences. Specifically, it introduces affinity and disparity weights to define positive and negative samples in a soft manner. Then, by comparing all samples in a batch, ConOrd not only learns the absolute positions of those samples in an embedding space but also captures the relative relationships between the ranks in a globally consistent manner. Extensive experiments demonstrate that ConOrd provides excellent performance in a variety of ordinal regression tasks.

Our contributions are summarized as follows:

- We propose the ConOrd algorithm, a novel approach that combines the strengths of both order learning and contrastive learning for effective ordinal regression.
- ConOrd overcomes key limitations of existing contrastive ordinal regression methods by contrasting *all* pairs in a batch *softly* using affinity and disparity weights, enabling the construction of a more discriminative embedding space for ordinal data.
- ConOrd achieves state-of-the-art performance on various benchmark tasks, including face age estimation, blind image quality assessment (BIQA), and blind video quality assessment (BVQA). Furthermore, it performs excellently on other regression problems, such as temperature prediction and gaze direction estimation, validating its generalization capability across diverse domains.

## 2 RELATED WORK

### 2.1 ORDINAL REGRESSION

Ordinal regression aims to predict the ordinal labels or ranks of object instances, which have an inherent order. It has been used in various applications, including medical diagnosis (Wu et al., 2019; Liu et al., 2018), depth estimation (Fu et al., 2018), and facial age estimation (Rothe et al., 2018; Zhu et al., 2021). Early approaches either reformulated the problem as multiple binary classification tasks (Frank & Hall, 2001; Li & Lin, 2007) or dealt with it as a regression problem by adapting traditional classification loss functions (Rennie & Srebro, 2005; Rothe et al., 2018). To better exploit

108 the ordinal properties of labels, subsequent methods have employed label distribution learning (Geng  
109 et al., 2013), mean-variance loss (Pan et al., 2018), soft ordinal labels (Diaz & Marathe, 2019), and  
110 probabilistic embeddings (Li et al., 2021).

111 However, many of these methods fail to capture inter-class relationships effectively, which may  
112 limit performance (Niu et al., 2016; Chen et al., 2017). To address this issue, distance-aware label  
113 embeddings (Shi et al., 2016), rank learning (Chen et al., 2017), and monotonic loss functions  
114 (Zhu et al., 2021) have been proposed. Also, Zhang et al. (2023) introduced an ordinal entropy  
115 regularizer, which promotes higher-entropy feature spaces while maintaining ordinal relationships.  
116 Meanwhile, evaluation metrics and loss designs that consider class proximity have been explored  
117 (Amigó et al., 2020), along with methods that tackle class imbalance (Nachmani et al., 2025). Through  
118 this progression, ordinal regression has come to be recognized — distinct from ordinary classification  
119 — as a predictive task that requires modeling both the order and the relative distances between classes.

## 120 121 2.2 ORDER LEARNING

122 Recently, order learning (Lim et al., 2020) has emerged as a promising approach to ordinal regression  
123 or rank estimation. In order learning, ordering relationships between object instances are learned,  
124 and the rank of an instance is estimated by comparing it with reference instances of known ranks.  
125 For reliable reference selection, Lee & Kim (2021) decomposed object embeddings into order and  
126 identity features and selected references with similar identity features. Shin et al. (2022) proposed a  
127 regression-based formulation to estimate a continuous relative rank between two references. Lee &  
128 Kim (2022) extended order learning to a weakly-supervised setting to cope with limited annotations,  
129 and Lee et al. (2024) proposed unsupervised order learning, which optimizes ordered clustering and  
130 embedding space construction alternately.

131 However, the direct comparison methods (Lim et al., 2020; Lee & Kim, 2021; Shin et al., 2022)  
132 should compare a test instance with multiple references, demanding considerable testing complexity,  
133 and do not consider metric relations between instances. To overcome these issues, Lee et al. (2022)  
134 proposed GOL that exploits metric, as well as order, relations to construct an embedding space and  
135 enables efficient rank estimation through a simple  $k$ -NN search in the embedding space.

## 136 137 2.3 CONTRASTIVE LEARNING

138 Contrastive learning aims to learn discriminative representations by modeling similarities and dissim-  
139 ilarities between object instances. It encourages the representations of similar (or positive) pairs to be  
140 pulled closer in an embedding space, while those of dissimilar (or negative) pairs to be pushed apart.  
141 Early methods (Chen et al., 2020; He et al., 2020) were largely explored in self-supervised scenarios,  
142 in which positive and negative pairs were constructed without requiring explicit labels.

143 To further leverage label information, supervised contrastive learning (Khosla et al., 2020) was  
144 introduced. Instead of solely relying on data augmentations to construct positive pairs, it also defines  
145 positive pairs from samples of the same class. The extension of contrastive learning to the fully  
146 supervised setting allows the learned feature space to reflect semantic structures more effectively,  
147 improving performance significantly in downstream tasks, such as image classification.

148 Although contrastive learning has shown remarkable performance in various tasks, such as semantic  
149 segmentation (Liu et al., 2021), object detection (Xie et al., 2021), and medical imaging (Basak &  
150 Yin, 2023), its applications to regression tasks have been less successful. As it was primarily designed  
151 for classification tasks, it fails to capture the underlying continuous order between samples. To  
152 overcome this limitation, Zha et al. (2023) proposed a contrastive framework that learns continuous  
153 representations through rank-based pairings. However, as their method incorporates ordinal informa-  
154 tion through relative ranking, it does not consider the magnitude of rank differences. More recently,  
155 OCL (Baek et al., 2024) modeled ordinal severity using adaptive temperatures, but its temperature  
156 scaling requires careful balancing to prevent embedding collapse or over-dispersion. Also, Zheng  
157 et al. (2024) improved ordinal regression performance through data augmentation. CLOC (Pitawela  
158 et al., 2025) introduced a multi-margin loss for greater flexibility, yet its joint margin optimization  
159 adds training complexity and risks unstable learning. Building on these advances, we present a  
160 soft-weighted contrastive scheme that explicitly employs rank differences to achieve more stable  
161 training and superior performance across diverse ordinal regression benchmarks.

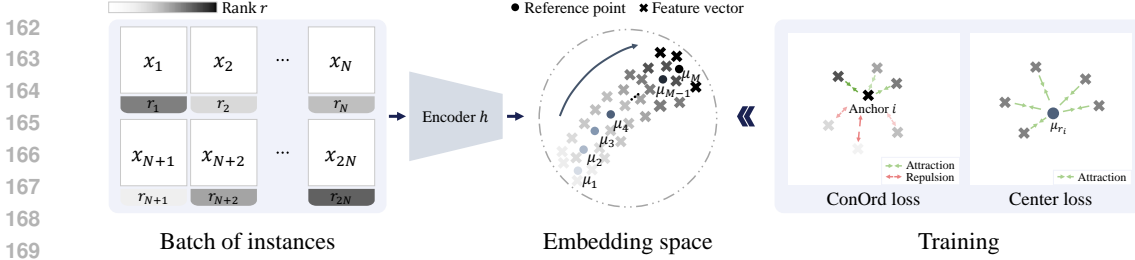


Figure 2: Overview of the proposed ConOrd algorithm.

### 3 PROPOSED ALGORITHM

#### 3.1 PROBLEM FORMULATION AND OVERVIEW

The objective of ordinal regression is to estimate the rank  $r$  of a given instance  $x$ . Unlike classification, in ordinal regression, rank labels are naturally ordered with inherent distances (Lee et al., 2022). For example, let us consider three instances  $x_i, x_j$ , and  $x_k$  with ranks  $r_i = 10, r_j = 12$ , and  $r_k = 30$ . Then, there is a natural ordering  $r_i < r_j < r_k$ , and the distance between  $r_i$  and  $r_j$  is smaller than that between  $r_j$  and  $r_k$ , *i.e.*,  $|r_i - r_j| < |r_j - r_k|$ .

We aim to construct an embedding space in which instances are arranged according to their ordinal relationships. To this end, we employ an encoder  $h$  to map each instance  $x \in \mathcal{X}$  into a feature vector  $z = h(x)$  in the embedding space. Note that we normalize the features to locate them on the unit sphere. Suppose that there are  $M$  ranks in a training set  $\mathcal{X}$ . Then, we introduce  $M$  learnable reference points  $\{\mu_m\}_{m=1}^M$ , where each  $\mu_m$  represents the centroid of instances with rank  $m$ . To train the encoder, we develop the ConOrd loss. Together with the center loss, the ConOrd loss guides the encoder to produce representations that are both order-aware and well-clustered. An overview of the proposed algorithm is in Figure 2.

#### 3.2 MOTIVATION

The motivation behind ConOrd is grounded in three foundational principles in order learning, each guiding a key component of our formulation.

- The observation that relational supervision is often more reliable than absolute predictions (Lim et al., 2020) motivates ConOrd to adopt an all-pairs contrastive objective that fully exploits pairwise relationships.
- While Lim et al. (2020) model pairwise order relations using a ternary classifier (“larger,” “similar,” “smaller”), Shin et al. (2022) extend this perspective by treating the magnitude of the rank difference itself as meaningful supervision. Their formulation distinguishes not only whether two samples differ in rank, but by how much, motivating ConOrd to incorporate continuous weights based on rank differences into its contrastive objective.
- The explicit use of attraction and repulsion forces in the embedding space, introduced by Lee et al. (2022), provides a geometric mechanism for enforcing ordinal structure. While their approach relies on margin-based forces, we extend this idea to a smooth contrastive formulation that assigns continuous attraction and repulsion to all pairwise relationships.

These principles naturally lead to the smooth, fully pairwise ordinal contrastive formulation presented in Section 3.3. Before introducing the formulation, we first revisit the SupCon loss (Khosla et al., 2020) to interpret it probabilistically and to clarify how it can be adapted for ordinal regression.

We refer to a set of  $2N$  instances used for training,  $\{x_i\}_{i=1}^{2N}$ , as a batch. Note that while Khosla et al. (2020) sample  $N$  instances and augment them to form  $2N$  instances, thereby obtaining positive pairs, the proposed algorithm does not rely on such data augmentation. Thus, in the proposed algorithm, a batch consists of  $2N$  randomly sampled instances.

Let  $I = \{1, \dots, 2N\}$  be the index set of samples within a batch, and  $z_i$  be the embedding vector of  $x_i$ . Also,  $\kappa_{ij}$  denotes the similarity between  $z_i$  and  $z_j$  (*e.g.*,  $\kappa_{ij} = z_i^T z_j$ ), and  $\tau$  is a temperature

Table 1: Several configurations that could be used for the contrastive order loss in (3).

Method	$\kappa_{ij}$	$a_{ij}$	$b_{ij}$	Method	$\kappa_{ij}$	$a_{ij}$	$b_{ij}$
I	$z_i^T z_j$	$( r_i - r_j  + \epsilon)^{-1}$	1	V	$-\ z_i - z_j\ _2^2$	$( r_i - r_j  + \epsilon)^{-1}$	1
II	$z_i^T z_j$	$( r_i - r_j  + \epsilon)^{-1}$	$ r_i - r_j $	VI	$-\ z_i - z_j\ _2^2$	$( r_i - r_j  + \epsilon)^{-1}$	$ r_i - r_j $
III	$z_i^T z_j$	$((r_i - r_j)^2 + \epsilon)^{-1}$	1	VII	$-\ z_i - z_j\ _2^2$	$((r_i - r_j)^2 + \epsilon)^{-1}$	1
IV	$z_i^T z_j$	$((r_i - r_j)^2 + \epsilon)^{-1}$	$(r_i - r_j)^2$	VIII	$-\ z_i - z_j\ _2^2$	$((r_i - r_j)^2 + \epsilon)^{-1}$	$(r_i - r_j)^2$

parameter controlling the sharpness of the similarity. Then, the SupCon loss is given by

$$L_{\text{SupCon}} = -\frac{1}{2N} \sum_{i=1}^{2N} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\kappa_{ip}/\tau)}{\sum_{j \in A(i)} \exp(\kappa_{ij}/\tau)} \quad (1)$$

where  $i$  is the index of an anchor,  $A(i) = I - \{i\}$  is the set of the remaining indices, and  $P(i)$  is the set of the indices of positive samples in the batch distinct from  $i$ .

In (1), the ratio  $e^{\kappa_{ip}/\tau} / \sum_j e^{\kappa_{ij}/\tau}$  can be interpreted as the probability that the anchor  $i$  is matched to the positive sample  $p$  among all other samples in the embedding space. By minimizing the negative logarithm in (1), the matching probability is maximized. This suits the purpose of classification. On the other hand, in ordinal regression, it is desirable to match the anchor  $i$  to another sample  $j$  such that the rank estimation error  $|r_i - r_j|$  is minimized. Thus, we can define the mean absolute error (MAE) loss for ordinal regression as

$$L_{\text{MAE}} = \frac{1}{2N} \sum_{i=1}^{2N} \frac{1}{|A(i)|} \log \frac{\sum_{j \in A(i)} |r_i - r_j| \exp(\kappa_{ij}/\tau)}{\sum_{j \in A(i)} \exp(\kappa_{ij}/\tau)}. \quad (2)$$

Note that  $L_{\text{MAE}}$  involves the positive logarithm because it attempts to minimize the expected rank estimation error  $\sum_j |r_i - r_j| \times (e^{\kappa_{ij}/\tau} / \sum_k e^{\kappa_{ik}/\tau})$ .

### 3.3 CONTRASTIVE ORDER LEARNING

The naive loss  $L_{\text{MAE}}$  in (2), however, makes the encoder training less reliable, for its positive logarithm tends to increase the magnitudes of gradients as the training goes on. To address this training issue while retaining its goal of matching the anchor to another sample of a similar rank in the embedding space, we design the ConOrd loss as follows:

$$L_{\text{ConOrd}} = -\frac{1}{2N} \sum_{i=1}^{2N} \frac{1}{|A(i)|} \log \frac{\sum_{j \in A(i)} a_{ij} \exp(\kappa_{ij}/\tau)}{\sum_{j \in A(i)} b_{ij} \exp(\kappa_{ij}/\tau)} \quad (3)$$

where the affinity weight  $a_{ij}$  is designed to promote the similarity between a pair of samples with similar ranks. Thus,  $a_{ij}$  is defined as a monotonically decreasing function of the rank gap  $|r_i - r_j|$ . Specifically, we set

$$a_{ij} = \frac{1}{(r_i - r_j)^2 + \epsilon} \quad (4)$$

where  $\epsilon$  is a small constant to prevent division by zero. On the other hand, in (3), the disparity weight  $b_{ij}$  controls the separation of samples with a large rank gap. It is a monotonically increasing function of the rank gap, set as

$$b_{ij} = (r_i - r_j)^2 \quad (5)$$

in this work. Also, we set  $\kappa_{ij} = -\|z_i - z_j\|_2^2$ . In Table 1, we suggest more configurations of  $\kappa_{ij}$ ,  $a_{ij}$ , and  $b_{ij}$ , the performances of which are compared in Section 4.4 and Appendix C.3.

The proposed ConOrd loss has the following properties.

- **Attraction and repulsion induced by affinity and disparity weights:** In Appendix A, the gradient of  $L_{\text{ConOrd}}$  is derived as

$$\frac{\partial L_{\text{ConOrd}}}{\partial z_i} = \frac{1}{N\tau|A(i)|} \sum_{j \in A(i)} \exp(\kappa_{ij}/\tau) \left( \frac{a_{ij}}{\alpha_i} - \frac{b_{ij}}{\beta_i} \right) (z_i - z_j) \quad (6)$$

where  $\alpha_i = \sum_{j \in A(i)} a_{ij} \exp(\kappa_{ij}/\tau)$  and  $\beta_i = \sum_{j \in A(i)} b_{ij} \exp(\kappa_{ij}/\tau)$ . The factor  $(\frac{a_{ij}}{\alpha_i} - \frac{b_{ij}}{\beta_i})$  in (6) determines whether gradient descent moves  $z_i$  toward or away from  $z_j$ . Since the affinity weight  $a_{ij}$  is large for small rank gaps and the disparity weight  $b_{ij}$  is large for large rank gaps, ConOrd attracts samples with similar ranks and repels samples with distant ranks, as illustrated in Figure 3. This provides a smooth ordinal interaction mechanism that generalizes the margin-based attraction and repulsion forces in GOL (Lee et al., 2022).

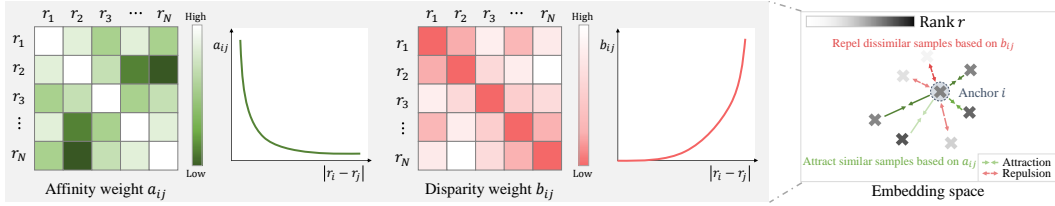


Figure 3: The ConOrd loss  $L_{\text{ConOrd}}$  encourages the attraction of similar samples and the repulsion of dissimilar samples in the embedding space, by employing affinity weights  $a_{ij}$  and disparity weights  $b_{ij}$ , respectively.

- Locality induced by the exponential kernel:** The factor  $\exp(\kappa_{ij}/\tau)$  in (6) further modulates each pairwise contribution according to embedding proximity, because  $\kappa_{ij} = -\|z_i - z_j\|_2^2$ . Pairs that are already close in the embedding space receive stronger weights, whereas distant pairs are exponentially suppressed. Consequently, ConOrd focuses its updates on the most informative ordinal relations — pairs that are close in rank and close in the embedding space — leading to fine-grained ordinal discrimination.
- Contrasting all samples with soft weights:** A key strength of  $L_{\text{ConOrd}}$  lies in its ability to contrast all samples in a batch through soft weighting, rather than relying on hard assignment of positives and negatives. Traditional contrastive losses (Chen et al., 2020; He et al., 2020; Khosla et al., 2020; Caron et al., 2020) select a positive sample and contrast it with all other samples in a batch, as in (1). In contrast, ConOrd employs all samples as both positives and negatives simultaneously and introduces continuous weights  $a_{ij}$  and  $b_{ij}$  that vary smoothly with the rank difference  $|r_i - r_j|$ . This formulation allows every pair  $(i, j)$  to contribute to the loss, with its influence modulated by the degree of ordinal similarity or dissimilarity. By retaining all pairwise comparisons, ConOrd can exploit the complete range of ordinal similarities present in the batch, from the most closely ranked samples to those farthest apart.
- Connection to RnC loss:**  $L_{\text{ConOrd}}$  is conceptually related to the RnC loss in (Zha et al., 2023), which also incorporates rank information into a contrastive framework. However, similar to the traditional contrastive losses, RnC makes hard decisions to select positive and negative samples and employs binary weights. Specifically, the RnC loss is given by

$$L_{\text{RnC}} = -\frac{1}{2N} \sum_{i=1}^{2N} \frac{1}{|A(i)|} \sum_{j \in A(i)} \log \frac{\exp(\kappa_{ij}/\tau)}{\sum_{k: |r_i - r_k| \geq |r_i - r_j|} \exp(\kappa_{ik}/\tau)}. \quad (7)$$

RnC contrasts each positive sample  $j$  against the negative samples  $k$  with  $|r_i - r_k| \geq |r_i - r_j|$ . However, those negatives are aggregated equally in the denominator, regardless of the rank gaps. In other words, although RnC ensures that negatives are farther from an anchor than a positive is, it does not distinguish how much farther those negatives are from the anchor. This prevents RnC from exploiting the fine-grained ordinal structure in the data fully. On the contrary, the proposed ConOrd uses all samples as positives simultaneously but combines them with the affinity weights  $a_{ij}$ . Furthermore, the same samples are also used as negatives with the disparity weights  $b_{ij}$ . Thus, these soft weights  $a_{ij}$  and  $b_{ij}$  in ConOrd can be seen as a relaxation of the hard rank thresholds in RnC, allowing finer-grained ordinal supervision.

- Choice of affinity and disparity weights:** The affinity and disparity weights depend only on the rank gap  $d_{ij} = |r_i - r_j|$ . We use the quadratic forms  $a_{ij} = 1/(d_{ij}^2 + \epsilon)$  and  $b_{ij} = d_{ij}^2$  as simple smooth, symmetric, and strictly monotonic functions of  $d_{ij}$ . This choice provides a natural balance between attraction and repulsion — small rank gaps receive strong attractive weights, while large gaps contribute more strongly to the repulsive term. Other monotonic forms are also viable, and we observe that ConOrd is not highly sensitive to this choice.

More properties and the gradient analysis of  $L_{\text{ConOrd}}$  are provided in Appendix A.

### 3.4 TRAINING AND INFERENCE

In addition to  $L_{\text{ConOrd}}$ , we introduce the center loss (Nguyen et al., 2018) as a regularization term to further structure the embedding space. The center loss, which seeks to locate each reference point

Table 2: MAE comparison on age estimation datasets.

Algorithm	Backbone	MORPH II	CLAP2015	AgeDB	UTK	CACD	Adience
POE (Li et al., 2021)	VGG-16	2.35	3.75	6.41	4.64	4.68	0.47
PML (Deng et al., 2021)	VGG-16	2.15	2.91	6.78	4.63	4.87	0.47
MWR-G (Shin et al., 2022)	VGG-16	2.24	2.82	6.18	4.49	4.76	0.46
GOL (Lee et al., 2022)	VGG-16	2.17	3.38	6.21	4.35	4.52	0.43
OrdinalCLIP (Li et al., 2022b)	VGG-16	2.32	3.20	5.85	4.53	4.36	0.47
RankNContrast (Zha et al., 2023)	ResNet-18	2.47	4.72	6.14	4.74	5.14	0.40
RankSim (Gong et al., 2022)	ResNet-50	3.10	5.55	6.51	4.93	4.68	0.60
OrdinalEntropy (Zhang et al., 2023)	ResNet-50	3.08	5.66	6.47	4.95	4.73	0.53
CLOC (Pitawela et al., 2025)	ResNet-50	2.84	4.14	6.87	4.81	4.66	0.41
L2RCLIP (Wang et al., 2023)	ViT-B	2.13	2.62	-	-	-	0.36
NumCLIP (Du et al., 2024)	ViT-B	<u>2.08</u>	<u>2.55</u>	<u>5.42</u>	<u>4.24</u>	<b>4.11</b>	<u>0.31</u>
ConOrd (Proposed)	ViT-B	<b>1.96</b>	<b>2.46</b>	<b>5.15</b>	<b>3.92</b>	<u>4.18</u>	<b>0.26</b>

Table 3: Performance comparison on five BIQA datasets.

Algorithm	Backbone	BID		CLIVE		KonIQ10k		SPAQ		FLIVE	
		SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC
BMPRI (Min et al., 2018)	Handcrafted	0.515	0.458	0.487	0.523	0.658	0.655	0.750	0.754	0.274	0.315
SFA (Li et al., 2018)	ResNet-50	0.826	0.840	0.812	0.833	0.888	0.897	0.906	0.907	0.542	0.626
DB-CNN (Zhang et al., 2018)	S-CNN	0.845	0.859	0.844	0.862	0.878	0.887	0.910	0.913	0.554	0.652
PaQ-2-PiQ (Ying et al., 2020)	ResNet-18	-	-	0.840	0.850	0.870	0.880	-	-	0.571	0.623
HyperIQA (Su et al., 2020)	ResNet-50	0.869	0.878	0.859	0.882	0.906	0.917	0.916	0.919	0.535	0.623
UNIQUE (Zhang et al., 2021)	ResNet-34	0.858	0.873	0.854	0.890	0.896	0.901	-	-	-	-
MUSIQ (Ke et al., 2021)	ViT-L	-	-	-	-	0.905	0.919	0.917	0.920	0.640	0.721
TReS (Golestaneh et al., 2022)	ResNet-50	-	-	0.846	0.877	0.915	0.928	-	-	0.554	0.625
Madhusudana et al. (2022)	ResNet-50	-	-	0.845	0.857	0.894	0.906	0.914	0.919	0.580	0.641
Re-IQA (Saha et al., 2023)	ResNet-50	-	-	0.840	0.854	0.914	0.923	0.918	0.925	<u>0.645</u>	0.733
QPT (Zhao et al., 2023)	Resnet50	0.842	0.852	0.857	0.881	0.912	0.927	0.916	0.921	0.551	0.635
LQMamba-B (Guan et al., 2024)	Mamba-B	-	-	0.837	0.875	0.895	0.913	0.912	0.914	-	-
LoDa (Xu et al., 2024)	Swin-B	-	-	<u>0.876</u>	<u>0.899</u>	0.932	0.944	<u>0.925</u>	<u>0.928</u>	0.578	0.679
QCN (Shin et al., 2024)	ConvNeXt-B	<u>0.892</u>	<u>0.890</u>	0.875	0.893	<u>0.934</u>	<u>0.945</u>	0.923	<u>0.928</u>	0.644	<u>0.741</u>
ConOrd (Proposed)	ViT-B	<b>0.913</b>	<b>0.925</b>	<b>0.900</b>	<b>0.921</b>	<b>0.947</b>	<b>0.958</b>	<b>0.927</b>	<b>0.931</b>	<b>0.651</b>	<b>0.752</b>

$\mu_m$  at the center of all instances with rank  $m$ , is defined as

$$L_{\text{center}} = \sum_i \|z_i - \mu_{r_i}\|_2. \quad (8)$$

This encourages the encoder to produce a compact cluster in the embedding space for each rank.

Overall, to design an embedding space in which instances are well aligned according to the ordinal characteristics of ranks, we optimize the encoder parameters and reference points by minimizing the total loss function, given by

$$L_{\text{total}} = L_{\text{ConOrd}} + L_{\text{center}}. \quad (9)$$

In the inference phase, we estimate the rank of an unseen test instance based on the  $k$ -NN rule. We first extract the feature vector  $z_t = h(x_t)$  of a test instance  $x_t$ . Then, in the embedding space, we find a set  $\mathcal{N}$  of its  $k$  nearest neighbors among all training instances in  $\mathcal{X}$ . Finally, the rank of  $x_t$  is estimated by

$$\hat{r}_t = \frac{1}{k} \sum_{i: x_i \in \mathcal{N}} r_i. \quad (10)$$

## 4 EXPERIMENTAL RESULTS

We apply ConOrd to three different ordinal regression tasks: facial age estimation, image quality assessment, and video quality assessment. Due to the space limitation, datasets and implementation details for each task are specified in Appendix B. Results on additional regression tasks, including temperature prediction and gaze direction estimation, are also presented in Appendix C.

### 4.1 FACIAL AGE ESTIMATION

**Datasets:** We use six datasets of MORPH II (Ricanek & Tesafaye, 2006), CLAP2015 (Escalera et al., 2015), AgeDB (Moschoglou et al., 2017; Yang et al., 2021), UTK (Zhang et al., 2017c), CACD (Chen et al., 2015), and Adience (Levi & Hassner, 2015), as detailed in Appendix B.1.

**Comparison with state-of-the-art methods:** Recent state-of-the-art age estimators (Wang et al., 2023; Du et al., 2024) have adopted the ViT-B network in (Radford et al., 2021), owing to its

Table 4: Comparison of BVQA results in the intra-dataset evaluation on LSVQ and in the cross-dataset evaluation on KonViD-1k, LIVE-VQC, CVD2014, and YouTube-UGC.

Algorithm	Backbone	LSVQ-test		LSVQ-1080p		KonViD-1k		LIVE-VQC		CVD2014		YouTube-UGC	
		SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC
TVLQM (Korhonen, 2019)	Handcrafted	0.772	0.774	0.589	0.616	0.732	0.724	0.670	0.691	-	-	-	-
VSFA (Li et al., 2019)	ResNet-50	0.801	0.796	0.675	0.704	0.810	0.811	0.753	0.795	0.756	0.760	0.718	0.721
VIDEVAL (Fu et al., 2021)	Handcrafted+CNN	0.794	0.783	0.545	0.554	0.751	0.741	0.630	0.640	-	-	-	-
PVQ (Ying et al., 2021)	PaQ2PiQ+3D ResNet-18	0.827	0.828	0.711	0.739	0.791	0.795	0.770	0.807	-	-	-	-
Li22 (Li et al., 2022a)	ResNet-50	0.852	0.854	0.772	0.788	0.843	0.835	0.793	0.811	0.817	0.811	0.802	0.792
SimpleVQA (Sun et al., 2022)	ResNet-50+SlowFast	0.866	0.863	0.750	0.793	0.826	0.820	0.749	0.789	0.780	0.802	0.802	0.806
FastVQA (Wu et al., 2022)	Swin-T	0.876	0.877	0.779	0.814	0.859	0.855	0.823	0.844	0.805	0.814	0.730	0.747
DOVER (Wu et al., 2023)	Inflated-ConvNeXt-T+Video Swin-T	0.888	0.889	0.795	0.830	0.884	0.883	0.832	0.855	0.829	0.832	-	-
KSVQE (Lu et al., 2024)	ViT-B+3D Swin Transformer	0.886	0.888	0.790	0.823	-	-	-	-	-	-	-	-
ModularBVQA (Wen et al., 2024)	ViT-B+SlowFast+ResNet-18	0.895	0.895	0.809	0.844	0.878	0.884	0.806	0.844	0.823	0.839	0.788	0.804
ConOrd (Proposed)	ViT-B+SlowFast	<b>0.904</b>	<b>0.904</b>	<b>0.818</b>	<b>0.851</b>	<b>0.889</b>	<b>0.892</b>	<b>0.836</b>	<b>0.865</b>	<b>0.839</b>	<b>0.845</b>	<b>0.805</b>	<b>0.828</b>

	CLAP2015	LSVQ-1080p	
	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
$L_{\text{SupCon}}$ in (3) (Khosla et al., 2020)	2.625	0.614	0.682
$L_{\text{OC}}$ (Baek et al., 2024)	2.597	0.697	0.687
$L_{\text{MMNP}} + L_{\text{CE}}$ (Pitawela et al., 2025)	2.777	0.716	0.727
$L_{\text{RnC}}$ in (7) (Zha et al., 2023)	2.531	0.812	0.843
$L_{\text{ConOrd}} + L_{\text{Center}}$ in (9)	<b>2.461</b>	<b>0.818</b>	<b>0.851</b>

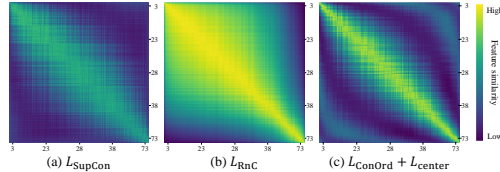


Table 5: Comparison of different contrastive losses on the CLAP2015 and LSVQ-1080p datasets.

Figure 4: Visualization of feature ordinality (Zha et al., 2023) on the CLAP2015 dataset.

strong representational capacity. Following this trend, the proposed ConOrd employs the same ViT-B backbone as the encoder  $h$  to ensure fair comparison. In Table 2, we compare ConOrd with conventional algorithms using the MAE metric, which computes the average absolute error between predicted and ground-truth ages. It is observed that ConOrd achieves the best performance across all datasets except CACD, which highlights both its generalizability and robustness in diverse age estimation scenarios.

Notably, ConOrd even outperforms OrdinalCLIP and NumCLIP, which also rely on the same backbone but incorporate both visual and textual information to guide the learning process. In other words, despite the fact that these methods leverage text embeddings (age-related textual descriptions or labels) to improve their performance, ConOrd outperforms them without relying on the additional modality. This demonstrates the efficacy of the proposed contrastive learning scheme, tailored for ordinal data, in discriminating different ages in an embedding space.

## 4.2 BLIND IMAGE QUALITY ASSESSMENT

**Datasets:** We conduct BIQA experiments on five datasets of BID (Ciancio et al., 2010), CLIVE (Ghadiyaram & Bovik, 2015), KonIQ10k (Hosu et al., 2020), SPAQ (Fang et al., 2020), and FLIVE (Ying et al., 2020). The details are available in Appendix B.2.

**Comparison with state-of-the-art methods:** We use Spearman’s rank order correlation coefficient (SRCC) and Pearson’s linear correlation coefficient (PCC) to assess perceptual ranking and linearity. In Table 3, ConOrd outperforms the existing methods in terms of both metrics on all five datasets with no exception. Note that recent transformer-based techniques QPT (Zhao et al., 2023), LoDa (Xu et al., 2024), and QCN (Shin et al., 2024) achieve high performance through large-scale pretraining or sophisticated network designs. Nevertheless, the proposed algorithm consistently outperforms these techniques meaningfully, confirming the effectiveness of contrastive order learning.

## 4.3 BLIND VIDEO QUALITY ASSESSMENT

**Datasets:** We evaluate ConOrd on five widely used BVQA benchmarks — LSVQ (Ying et al., 2021), KonViD-1k (Hosu et al., 2017), LIVE-VQC (Sinno & Bovik, 2018), CVD2014 (Nuutinen et al., 2016), and YouTube-UGC (Wang et al., 2019). For training, the LSVQ dataset is used. For evaluation, intra-dataset tests on LSVQ-test and LSVQ-1080p, as well as cross-dataset evaluations on the remaining four datasets, are conducted. More details on the datasets are in Appendix B.3.

**Comparison with state-of-the-art methods:** Table 4 presents a comprehensive comparison of ConOrd with state-of-the-art BVQA methods. Performance is measured using SRCC and PCC, as in

Method	$L_{\text{ConOrd}}$	$L_{\text{center}}$	CLAP2015		LSVQ-1080p	
			MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	
I	✓		2.509	0.815	0.848	
II		✓	2.842	0.753	0.793	
III	✓	✓	2.461	0.818	0.851	

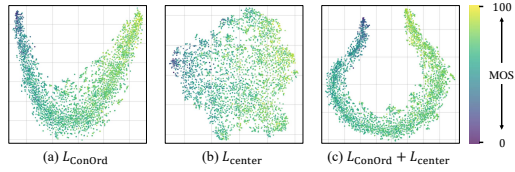
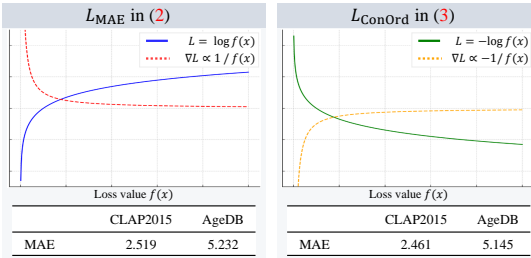


Table 6: Ablation study for the loss function in (9) on CLAP2015 and LSVQ-1080p. Figure 5: t-SNE (Van der Maaten & Hinton, 2008) plots of the embedding spaces on LSVQ-1080p.



Method	$\kappa_{ij}$	$a_{ij}$	$b_{ij}$	MAE ( $\downarrow$ )
I	$z_i^T z_j$	$( r_i - r_j  + \epsilon)^{-1}$	1	2.516
II	$z_i^T z_j$	$( r_i - r_j  + \epsilon)^{-1}$	$ r_i - r_j $	2.494
III	$z_i^T z_j$	$((r_i - r_j)^2 + \epsilon)^{-1}$	1	2.483
IV	$z_i^T z_j$	$((r_i - r_j)^2 + \epsilon)^{-1}$	$(r_i - r_j)^2$	2.518
V	$-\ z_i - z_j\ _2^2$	$( r_i - r_j  + \epsilon)^{-1}$	1	2.497
VI	$-\ z_i - z_j\ _2^2$	$( r_i - r_j  + \epsilon)^{-1}$	$ r_i - r_j $	2.490
VII	$-\ z_i - z_j\ _2^2$	$((r_i - r_j)^2 + \epsilon)^{-1}$	1	2.472
VIII	$-\ z_i - z_j\ _2^2$	$((r_i - r_j)^2 + \epsilon)^{-1}$	$(r_i - r_j)^2$	<b>2.461</b>
IX	$-\ z_i - z_j\ _2^2$	$(\Delta^2 + \epsilon)^{-1}$	$\Delta^2$	2.536
X	$-\ z_i - z_j\ _2^2$	Learnable $a_{ij}$	Learnable $b_{ij}$	2.880

Figure 6: Comparison of the losses in (2) and (3) on CLAP2015 and AgeDB. Table 7: Comparison of alternative configurations of  $L_{\text{ConOrd}}$  in (3) on CLAP2015.

BIQA. Again, ConOrd consistently outperforms all conventional methods across all datasets, achieving the best SRCC and PCC scores. While ModularBVQA uses three distinct backbone networks of ResNet-18, SlowFast, and ViT-B, ConOrd relies on only two backbones of SlowFast and ViT-B. Despite the usage of fewer backbones, ConOrd outperforms ModularBVQA meaningfully. This indicates that ConOrd is capable of extracting more discriminative features with fewer computational resources, demonstrating its potential for practical deployment in real-world BVQA applications.

#### 4.4 ABLATIONS AND ANALYSES

**Comparison of contrastive learning schemes:** In Table 5, we conduct a comparative analysis of both general-purpose (Khosla et al., 2020) and ordinal-regression-oriented (Zha et al., 2023; Baek et al., 2024; Pitawela et al., 2025) contrastive losses on the CLAP2015 and LSVQ-1080p datasets.  $L_{\text{SupCon}}$  yields weak performance, as it disregards the ordinal property of rank labels.  $L_{\text{OC}}$  shows slight improvements but remains limited, as it was primarily designed for medical diagnostic datasets with relatively few ordinal levels. It does not scale effectively to tasks with many ordinal levels, e.g., age estimation or BVQA.  $L_{\text{MMNP}} + L_{\text{CE}}$  achieves moderate correlation gains; however, its performance is inconsistent across metrics due to the complexity of joint margin optimization.  $L_{\text{RnC}}$  benefits from its ordinal-aware contrastive formulation and substantially outperforms these other existing losses. Finally, the proposed loss in (9) achieves the best results overall, showing that the all-pairs comparison through soft weighting helps learn better representations for ordinal regression.

These findings are further supported by the feature similarity matrices in Figure 4. As in Zha et al. (2023), the matrices are computed using the negative L2 norm between learned features on the CLAP2015 dataset. Representations are sorted by ground-truth ranks, so entries farther from the diagonal indicate larger rank differences. Compared to  $L_{\text{SupCon}}$  and  $L_{\text{RnC}}$ , the proposed loss produces a clearer high-similarity band along the diagonal, confirming that the learned representations reflect the underlying ordinal structure more faithfully.

**Ablation study of loss function:** To assess the contribution of each component in (9), we conduct an ablation study in Table 6 and visualize the embedding space for each ablated method in Figure 5.  $L_{\text{center}}$  alone fails to capture the ordinal nature of the task. While  $L_{\text{ConOrd}}$  alone provides better results, its combination with  $L_{\text{center}}$  further improves the results as the intra-class compactness is also considered. Both quantitatively and visually, the combined loss yields the most favorable results.

**Alternative choices for loss function:** We compare the behavior of the naive loss formulation  $L_{\text{MAE}}$  in (2) with that of the proposed  $L_{\text{ConOrd}}$  in (3). As shown in Figure 6, the positive logarithm in  $L_{\text{MAE}}$  may make the training less reliable, since the gradient magnitude may increase as the loss decreases. It means that the optimization process may become unstable over time, ultimately

486 degrading the effectiveness of learning. Empirically,  $L_{\text{ConOrd}}$  outperforms  $L_{\text{MAE}}$  on CLAP2015 and  
487 AgeDB, confirming its improved stability.  
488

489 Also, to investigate the impacts of different design choices in (3), we perform an ablation study on  
490 CLAP2015 by varying the components of  $L_{\text{ConOrd}}$ . The following observations can be made from  
491 the results in Table 7. First, negative squared Euclidean distance for  $\kappa_{ij}$  consistently outperforms  
492 cosine similarity, suggesting that explicit distance-based measures better reflect ordinal relations in  
493 the embedding space. Second, method VIII that incorporates squared differences into both  $a_{ij}$  and  
494  $b_{ij}$  achieves the lowest MAE of 2.461. This indicates that emphasizing larger ordinal gaps during  
495 the contrastive optimization enhances model sensitivity to ordinal labels and improves regression  
496 accuracy. Additional insights can be drawn from other configurations. Method IX sets  $\Delta$  to 0  
497 if  $r_i = r_j$ , and to a positive threshold of 5 otherwise. These coarser approximations of ordinal  
498 differences are less effective than the proposed fine-grained modeling of rank gaps. Method X  
499 employs fully learnable  $a_{ij}$  and  $b_{ij}$ . It yields poor results, suggesting that explicit encoding of ordinal  
500 structure is more effective than unconstrained learnable parameters. An extended table with additional  
501 configurations and results on more datasets is provided in Appendix C.3.

## 502 5 CONCLUSIONS

503  
504 In this paper, we proposed a novel contrastive learning algorithm for ordinal regression, referred to as  
505 ConOrd. By bringing together the strengths of both order learning and contrastive learning, ConOrd  
506 can represent the fine-grained ordinal relationships across all samples within a batch faithfully in  
507 an embedding space. To this end, we developed a contrastive order loss that leverages soft affinity  
508 and disparity weights based on rank differences. Additionally, we integrated the center loss based  
509 on learnable reference points to further improve the intra-class compactness of the embedding  
510 space. Extensive experiments on diverse ordinal regression tasks, including facial age estimation,  
511 BIQA, and BVQA, demonstrated that ConOrd consistently outperforms existing methods, achieving  
512 state-of-the-art results on most benchmark datasets.  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

## REFERENCES

- 540  
541  
542 E. Amigó, J. Gonzalo, S. Mizzaro, and J. Carrillo-de Albornoz. An effectiveness metric for ordinal  
543 classification: Formal properties and experimental results. *arXiv preprint arXiv:2006.01245*, 2020.  
544 3
- 545 Seunghun Baek, Jaeyoon Sim, Guorong Wu, and Won Hwa Kim. OCL: Ordinal contrastive learning  
546 for imputating features with progressive labels. In *Proc. MICCAI*, 2024. 3, 8, 9  
547
- 548 Himadri Basak and Zihan Yin. Pseudo-label guided contrastive learning for semi-supervised medical  
549 image segmentation. In *Proc. CVPR*, 2023. 3
- 550 Axel Berg, Magnus Oskarsson, and Mark O’Connor. Deep ordinal regression with label diversity. In  
551 *Proc. ICPR*, 2021. 18  
552
- 553 Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin.  
554 Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. NeurIPS*,  
555 2020. 6
- 556 Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost  
557 sensitivities for age estimation. In *Proc. CVPR*, 2011. 18  
558
- 559 Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Face recognition and retrieval using cross-age  
560 reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 2015. 7, 18
- 561 Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age  
562 estimation. In *Proc. CVPR*, 2017. 3  
563
- 564 Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for  
565 contrastive learning of visual representations. In *Proc. ICML*, 2020. 3, 6  
566
- 567 Wei-Ta Chu, Kai-Chia Ho, and Ali Borji. Visual weather temperature prediction. In *Proc. WACV*,  
568 2018. 20
- 569 Alexandre Ciancio, André Luiz N Targino Targino da Costa, Eduardo A. B. da Silva, Amir Said,  
570 Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on  
571 multifeature classifiers. *IEEE Trans. Image Proc.*, 2010. 8, 18
- 572 Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. PML:  
573 Progressive margin loss for long-tailed age classification. In *Proc. CVPR*, 2021. 7  
574
- 575 Ramon Diaz and Arjun Marathe. Soft labels for ordinal regression. In *Proc. CVPR*, 2019. 1, 3  
576
- 577 Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach CLIP to develop a number sense for  
578 ordinal regression. In *Proc. ECCV*, 2024. 7
- 579 Denis A Engemann, Ayoub Mellot, Richard Höchenberger, Hubert Banville, David Sabbagh, Lukas  
580 Gemein, and Alexandre Gramfort. A reusable benchmark of brain-age prediction from M/EEG  
581 resting-state signals. *NeuroImage*, 2022. 1  
582
- 583 Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi González, Hugo J. Escalante, Dusan  
584 Misevic, Ulrich Steiner, and Isabelle Guyon. ChaLearn looking at people 2015: Apparent age and  
585 cultural event recognition datasets and results. In *Proc. ICCV Workshops*, 2015. 7, 18
- 586 Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of  
587 smartphone photography. In *Proc. CVPR*, 2020. 8, 18  
588
- 589 Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video  
590 recognition. In *Proc. ICCV*, 2019. 19
- 591 E. Frank and M. Hall. A simple approach to ordinal classification. In *Proc. ECML*, 2001. 2  
592
- 593 Huizhong Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep  
ordinal regression network for monocular depth estimation. In *Proc. CVPR*, 2018. 2

- 594 Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label  
595 distribution learning. In *Proc. IJCAI*, 2018. 20
- 596
- 597 X. Geng, C. Yin, and Z. H. Zhou. Facial age estimation by learning from label distributions. *IEEE*  
598 *Trans. Pattern Anal. Mach. Intell.*, 2013. 1, 3
- 599
- 600 Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective  
601 picture quality. *IEEE Trans. Image Proc.*, 2015. 8, 18
- 602
- 603 S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via  
604 transformers, relative ranking, and self-consistency. In *Proc. WACV*, 2022. 7
- 605
- 606 Yu Gong, Greg Mori, and Frederick Tung. RankSim: Ranking similarity regularization for deep  
607 imbalanced regression. In *Proc. ICML*, 2022. 7
- 608
- 609 F. Guan, X. Li, Z. Yu, Y. Lu, and Z. Chen. Q-Mamba: On first exploration of vision Mamba for  
610 image quality assessment. *arXiv preprint arXiv:2406.09546*, 2024. 7, 27
- 611
- 612 Fredrik K Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B Schön. Energy-based models  
613 for deep probabilistic regression. In *Proc. ECCV*, 2020. 18
- 614
- 615 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
616 recognition. In *Proc. CVPR*, 2016. 20
- 617
- 618 Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for  
619 unsupervised visual representation learning. In *Proc. CVPR*, 2020. 3, 6
- 620
- 621 Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common  
622 corruptions and perturbations. In *ICLR*, 2019. 24
- 623
- 624 Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and  
625 Dietmar Saupe. The Konstanz natural video database (KoNViD-1k). In *Proc. QoMEX*, 2017. 1, 8,  
626 19
- 627
- 628 Vlad Hosu, Tingting Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid  
629 database for deep learning of blind image quality assessment. *IEEE Trans. Image Proc.*, 2020. 8,  
630 18
- 631
- 632 Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger.  
633 Snapshot ensembles: Train 1, get m for free. In *Proc. ICLR*, 2017. 18
- 634
- 635 Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image  
636 quality transformer. In *Proc. ICCV*, 2021. 7
- 637
- 638 Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron  
639 Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proc. NeurIPS*, 2020.  
640 1, 2, 3, 4, 6, 8, 9
- 641
- 642 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*,  
643 2015. 18
- 644
- 645 Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans.*  
646 *Image Proc.*, 2019. 8
- 647
- 648 Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity  
649 decomposition. In *Proc. ICLR*, 2021. 1, 3
- 650
- 651 Seon-Ho Lee and Chang-Su Kim. Order learning using partially ordered data via chainization. In  
652 *Proc. ECCV*, 2022. 3
- 653
- 654 Seon-Ho Lee, Nyeong-Ho Shin, and Chang-Su Kim. Geometric order learning for rank estimation.  
655 In *Proc. NeurIPS*, 2022. 1, 2, 3, 4, 5, 7
- 656
- 657 Seon-Ho Lee, Nyeong-Ho Shin, and Chang-Su Kim. Unsupervised order learning. In *Proc. ICLR*,  
658 2024. 3

- 648 Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In  
649 *Proc. CVPR*, 2015. 7, 18
- 650
- 651 Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality  
652 of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Trans. Circuits*  
653 *Syst. Video Technol.*, 2022a. 8
- 654
- 655 Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: the clear  
656 blue sky or a blurry animal? *IEEE Trans. Multimedia*, 2018. 7
- 657
- 658 Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proc.*  
659 *ACM Multimedia*, 2019. 8
- 660
- 661 Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Proc. NeurIPS*,  
662 2007. 1, 2
- 663
- 664 Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal  
665 embeddings for uncertainty-aware regression. In *Proc. CVPR*, 2021. 3, 7
- 666
- 667 Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. OrdinalCLIP:  
668 learning rank prompts for language-guided ordinal regression. In *Proc. NeurIPS*, 2022b. 7
- 669
- 670 Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its  
671 application to age estimation. In *Proc. ICLR*, 2020. 1, 3, 4
- 672
- 673 Wei Liu, Denis Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Do-  
674 main adaptation for semantic segmentation via patch-wise contrastive learning. *arXiv preprint*  
675 *arXiv:2104.11056*, 2021. 3
- 676
- 677 X. Liu, Y. Zou, Y. Song, C. Yang, J. You, and B. K. Vijaya Kumar. Ordinal regression with neuron  
678 stick-breaking for medical diagnosis. In *Proc. ECCV*, 2018. 2
- 679
- 680 Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv*  
681 *preprint arXiv:1608.03983*, 2016. 18
- 682
- 683 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*  
684 *arXiv:1711.05101*, 2017. 18
- 685
- 686 Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo  
687 Chen. KVQ: Kwai video quality assessment for short-form videos. In *Proc. CVPR*, 2024. 8
- 688
- 689 Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality  
690 assessment using contrastive learning. *IEEE Trans. Image Proc.*, 2022. 7
- 691
- 692 Razvan-Paul Mihail, Scott Workman, Zachary Bessinger, and Nathan Jacobs. Sky segmentation in  
693 the wild: an empirical study. In *Proc. WACV*, 2016. 20
- 694
- 695 Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. Blind image quality  
696 estimation via distortion aggravation. *IEEE Trans. Broadcasting*, 2018. 7
- 697
- 698 Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia,  
699 and Stefanos Zafeiriou. AgeDB: The first manually collected, in-the-wild age database. In *Proc.*  
700 *CVPR*, 2017. 1, 7, 18
- 701
- 702 I. Nachmani, B. Genossar, C. Scharf, R. Shraga, and A. Gal. SLACE: A monotone and balance-  
703 sensitive loss function for ordinal regression. In *Proc. AAAI*, 2025. 3
- 704
- 705 Bac Nguyen, Carlos Morell, and Bernard De Baets. Distance metric learning for ordinal classification  
706 based on triplet constraints. *Knowledge-Based Systems*, 2018. 6
- 707
- 708 Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple  
709 output CNN for age estimation. In *Proc. CVPR*, 2016. 3, 20

- 702 Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka  
703 Häkkinen. CVD2014—A database for evaluating no-reference video quality assessment algorithms.  
704 *IEEE Trans. Image Proc.*, 2016. 8, 19
- 705  
706 H. Pan, H. Han, S. Shan, and X. Chen. Mean-variance loss for deep age estimation from a face. In  
707 *Proc. CVPR*, 2018. 3
- 708 Dileepa Pitawela, Gustavo Carneiro, and Hsiang-Ting Chen. CLOC: Contrastive learning for ordinal  
709 classification with multi-margin n-pair loss. In *Proc. CVPR*, 2025. 3, 7, 8, 9
- 710  
711 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
712 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.  
713 Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 7,  
714 18
- 715 J. D. Rennie and N. Srebro. Loss functions for preference levels: regression with discrete ordered  
716 labels. In *Proc. IJCAI Workshop on Advances in Preference Handling*, 2005. 2
- 717  
718 Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult  
719 age-progression. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2006. 7, 18
- 720 Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep expectation of apparent age from a  
721 single image. In *Proc. ICCV*, 2015. 20
- 722  
723 Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a  
724 single image without facial landmarks. *Int. J. Comput. Vis.*, 2018. 1, 2
- 725 Avinab Saha, Sandeep Mishra, and Alan C Bovik. Re-IQA: Unsupervised learning for image quality  
726 assessment in the wild. In *Proc. CVPR*, 2023. 7, 27
- 727  
728 Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal  
729 regression based on conditional probabilities. *Pattern Analysis and Applications*, 2023. 20
- 730  
731 Y. Shi, W. Li, and F. Sha. Metric learning for ordinal data. In *Proc. AAAI*, 2016. 3
- 732 Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach  
733 to ordinal regression. In *Proc. CVPR*, 2022. 3, 4, 7, 18
- 734  
735 Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Blind image quality assessment based on  
736 geometric order learning. In *Proc. CVPR*, 2024. 7, 8, 18, 27
- 737 Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Trans.*  
738 *Image Proc.*, 2018. 8, 19
- 739  
740 Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly  
741 assess image quality in the wild guided by a self-adaptive hyper network. In *Proc. CVPR*, 2020. 7
- 742  
743 Wei Sun, Xionghuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality  
744 assessment model for UGC videos. In *Proc. ACM Multimedia*, 2022. 8
- 745  
746 Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland,  
747 Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications*  
748 *of the ACM*, 2016. 18, 19
- 749  
750 Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. UGC-VQA:  
751 Benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image*  
752 *Proc.*, 2021. 8
- 753  
754 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine*  
755 *learning research*, 2008. 9
- 756  
757 Rui Wang, Pei Pei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-Rank  
758 meets language: Boosting language-driven ordering alignment for ordinal classification. In *Proc.*  
759 *NeurIPS*, 2023. 7

- 756 Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube UGC dataset for video compression research.  
757 In *Proc. MMSP*, 2019. 8, 19
- 758 Yujing Wang, Yuxiao Jiang, Jie Li, Bingbing Ni, Weilin Dai, Chenglong Li, and Tao Li. Contrastive  
759 regression for domain adaptation on gaze estimation. In *Proc. CVPR*, 2022. 1
- 760 Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. Modular blind video  
761 quality assessment. In *Proc. CVPR*, 2024. 8
- 762 B. Wu, X. Sun, L. Hu, and Y. Wang. Learning with unsure data for medical image diagnosis. In *Proc.*  
763 *ICCV*, 2019. 2
- 764 Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and  
765 Weisi Lin. FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In  
766 *Proc. ECCV*, 2022. 8
- 767 Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun,  
768 Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from  
769 aesthetic and technical perspectives. In *Proc. ICCV*, 2023. 8
- 770 Enze Xie, Jianwei Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, and Ping Luo. DetCo:  
771 Unsupervised contrastive learning for object detection. In *Proc. ICCV*, 2021. 3
- 772 K. Xu, L. Liao, J. Xiao, C. Chen, H. Wu, Q. Yan, and W. Lin. Boosting image quality assessment  
773 through efficient transformer adaptation with local feature enhancement. In *Proc. CVPR*, 2024. 7,  
774 8, 27
- 775 Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbal-  
776 anced regression. In *Proc. ICML*, 2021. 7
- 777 Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik.  
778 From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proc.*  
779 *CVPR*, 2020. 7, 8, 18
- 780 Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-VQ: ‘patching  
781 up’ the video quality problem. In *Proc. CVPR*, 2021. 1, 8, 19
- 782 Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-N-Contrast: Learning  
783 continuous representations for regression. In *Proc. NeurIPS*, 2023. 1, 2, 3, 6, 7, 8, 9, 20
- 784 Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep  
785 regression with ordinal entropy. In *Proc. ICLR*, 2023. 3, 7
- 786 Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment  
787 using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.*,  
788 2018. 7
- 789 Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality  
790 assessment in the laboratory and wild. *IEEE Trans. Image Proc.*, 2021. 7
- 791 Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-world dataset  
792 and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017a. 20
- 793 Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It’s written all over your face:  
794 Full-face appearance-based gaze estimation. In *Proc. CVPR Workshops*, 2017b. 20
- 795 Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial  
796 autoencoder. In *Proc. CVPR*, 2017c. 7, 18
- 797 Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for  
798 blind image quality assessment. In *Proc. CVPR*, 2023. 7, 8, 18
- 799 Jiyang Zheng, Yu Yao, Bo Han, Dadong Wang, and Tongliang Liu. Enhancing contrastive learning  
800 for ordinal regression via ordinal content preserved data augmentation. In *Proc. ICLR*, 2024. 3
- 801 Hao Zhu, Hongming Shan, Yuyin Zhang, Liu Che, Xiaowei Xu, Jian Zhang, and Fei-Yue Wang.  
802 Convolutional ordinal regression forest for image ordinal estimation. *IEEE Trans. Neural Netw.*  
803 *Learn. Syst.*, 2021. 2, 3

## A PROPERTIES AND GRADIENT ANALYSIS OF $L_{\text{CONORD}}$ IN (3)

We analyze the gradient of  $L_{\text{ConOrd}}$  with respect to the embedding  $z_i$ . Using the chain rule, we have

$$\frac{\partial L_{\text{ConOrd}}}{\partial z_i} = \sum_{j \in A(i)} \frac{\partial L_{\text{ConOrd}}}{\partial \kappa_{ij}} \cdot \frac{\partial \kappa_{ij}}{\partial z_i}. \quad (11)$$

For simplicity, let  $\alpha_i = \sum_{j \in A(i)} a_{ij} \exp(\kappa_{ij}/\tau)$  and  $\beta_i = \sum_{j \in A(i)} b_{ij} \exp(\kappa_{ij}/\tau)$  in (3). Then,

$$L_{\text{ConOrd}} = -\frac{1}{2N} \sum_{i=1}^{2N} \frac{1}{|A(i)|} \log \frac{\alpha_i}{\beta_i}. \quad (12)$$

Computing the first term in the chain rule expression in (11),

$$\frac{\partial L_{\text{ConOrd}}}{\partial \kappa_{ij}} = -\frac{1}{2N|A(i)|} \left( \frac{1}{\alpha_i} \frac{\partial \alpha_i}{\partial \kappa_{ij}} - \frac{1}{\beta_i} \frac{\partial \beta_i}{\partial \kappa_{ij}} \right) \quad (13)$$

$$= -\frac{1}{2N|A(i)|} \left( \frac{1}{\alpha_i} \cdot \frac{a_{ij}}{\tau} \exp(\kappa_{ij}/\tau) - \frac{1}{\beta_i} \cdot \frac{b_{ij}}{\tau} \exp(\kappa_{ij}/\tau) \right) \quad (14)$$

$$= -\frac{1}{2N\tau|A(i)|} \exp(\kappa_{ij}/\tau) \left( \frac{a_{ij}}{\alpha_i} - \frac{b_{ij}}{\beta_i} \right). \quad (15)$$

Assuming that the similarity  $\kappa_{ij}$  is defined as the negative squared Euclidean distance, *i.e.*,  $\kappa_{ij} = -\|z_i - z_j\|_2^2$ ,

$$\frac{\partial \kappa_{ij}}{\partial z_i} = -2(z_i - z_j). \quad (16)$$

Thus, we have the gradient

$$\frac{\partial L_{\text{ConOrd}}}{\partial z_i} = \sum_{j \in A(i)} -\frac{1}{2N\tau|A(i)|} \exp(\kappa_{ij}/\tau) \left( \frac{a_{ij}}{\alpha_i} - \frac{b_{ij}}{\beta_i} \right) \cdot -2(z_i - z_j) \quad (17)$$

$$= \frac{1}{N\tau|A(i)|} \sum_{j \in A(i)} \exp(\kappa_{ij}/\tau) \left( \frac{a_{ij}}{\alpha_i} - \frac{b_{ij}}{\beta_i} \right) \cdot (z_i - z_j). \quad (18)$$

The following observations can be made from this gradient expression.

- Attraction and repulsion induced by affinity and disparity weights:** The scalar factor  $\left(\frac{a_{ij}}{\alpha_i} - \frac{b_{ij}}{\beta_i}\right)$  in (18) determines whether gradient descent pulls  $z_i$  toward or pushes it away from  $z_j$ . Since the affinity weight  $a_{ij}$  is large for small rank differences and the disparity weight  $b_{ij}$  is large for large rank differences, pairs with similar ranks tend to yield positive factors (attraction), whereas pairs with distant ranks tend to yield negative factors (repulsion).
- Locality induced by the exponential kernel:** The factor  $\exp(\kappa_{ij}/\tau)$  in (18) modulates each pairwise contribution according to the current embedding distance, since  $\kappa_{ij} = -\|z_i - z_j\|_2^2$ . Pairs that are already close in the embedding space receive larger weights, while distant pairs are exponentially downweighted. Consequently, the gradient places greater emphasis on sample pairs that are both close in rank and close in the embedding space, enabling ConOrd to capture fine-grained ordinal distinctions more effectively.
- Selective emphasis on informative pairwise relations:** Because strong forces arise only when two samples are close in the embedding space, the optimization primarily updates  $z_i$  using pairs that the model already considers relevant. Combined with the affinity and disparity weights, this suppresses uninformative interactions with very distant ranks and allows ConOrd to adjust embeddings where ordinal information is most meaningful, without introducing instability as training progresses.
- Choice of affinity and disparity weights:** The affinity and disparity weights depend only on the rank gap  $d_{ij} = |r_i - r_j|$ . We adopt the quadratic forms  $a_{ij} = 1/(d_{ij}^2 + \epsilon)$  and

864  $b_{ij} = d_{ij}^2$  because they are the simplest smooth, symmetric, and strictly monotonic functions  
865 of the ordinal distance. Quadratic growth provides a natural balance between near and far  
866 ranks — small gaps yield strong attractive weights, whereas large gaps contribute more  
867 to the repulsive term. While alternative monotonic choices (*e.g.*, linear or higher-order  
868 functions) are possible, we found that ConOrd is not highly sensitive to the exact form of  
869 these weights. Across the variants evaluated in Table 14, the quadratic form offers a good  
870 trade-off between stability and performance and performs consistently well across datasets.  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

## B EXPERIMENTAL DETAILS

### B.1 FACIAL AGE ESTIMATION

**Datasets:** Here, we provide more descriptions of each facial age estimation dataset.

- MORPH II (Ricanek & Tesafaye, 2006): As in Chang et al. (2011), we use 5,492 Caucasian images divided into training and test sets with a ratio of 8:2.
- CLAP2015 (Escalera et al., 2015): This dataset provides 4,691 facial images in total that are split into 2,476 for training, 1,136 for validation, and 1,079 for testing.
- AgeDB (Moschoglou et al., 2017): It contains 12.2K images for training, and the validation and test sets are balanced with 2.1K images. The age value ranges from 0 to 101.
- UTK (Zhang et al., 2017c): It consists of 20,000 facial images in a wide age range of [0,116]. We adopt the evaluation protocol in (Gustafsson et al., 2020; Berg et al., 2021).
- CACD (Chen et al., 2015): It provides 160k images of 2000 celebrities, which have an age range of [14, 62]. We use the training set specified in (Shin et al., 2022).
- Adience (Levi & Hassner, 2015): It has 26,580 facial images that are grouped into 8 ordinal classes: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60-year-olds.

**Implementation details:** For age estimation, ViT-B in the CLIP algorithm (Radford et al., 2021) is employed as the encoder  $h$ . The model is trained using the Adam optimizer (Kingma & Ba, 2015) with a weight decay of 0.0005. We employ a cosine annealing scheduler (Huang et al., 2017) to adjust the learning rate. For data augmentation, only random horizontal flipping is applied.

### B.2 BLIND IMAGE QUALITY ASSESSMENT

**Datasets:** We evaluate the performance of the proposed ConOrd algorithm on the BIQA task using five datasets. For all datasets except FLIVE, we randomly split each dataset into train and test sets with a ratio of 4:1. Then, we repeat the training and evaluation over 10 different splits and report the median evaluation scores as done in previous methods (Zhao et al., 2023; Shin et al., 2024). For FLIVE, we employ the same evaluation protocol as in (Ying et al., 2020), where 30K images are used for training and 1.8K for testing.

- BID (Ciancio et al., 2010): It contains 586 images degraded by various types of realistic distortion (*e.g.*, motion blur, out-of-focus), with mean opinion scores (MOS) in the range [0, 5].
- CLIVE (Ghadiyaram & Bovik, 2015): It consists of 1,169 natural images collected in diverse environments, annotated with MOS in the range [1, 100].
- KonIQ10k (Hosu et al., 2020): This dataset provides 10,073 images sampled from YFCC100M (Thomee et al., 2016), with MOS ranging from 1 to 100.
- SPAQ (Fang et al., 2020): Smartphone Photography Attribute and Quality (SPAQ) database includes 11,125 images taken with 66 different smartphones. Each image is assigned image attribute scores, but we use only the overall image quality scores in the range [0, 100].
- FLIVE (Ying et al., 2020): It is a large-scale BIQA dataset comprising approximately 40,000 images and 120,000 patches in the range [0, 100]. Following Ying et al. (2020), we only use the full-resolution images for training and testing, not the patches.

**Implementation details:** For BIQA, we also employ ViT-B from the CLIP algorithm (Radford et al., 2021) as the encoder  $h$ . The AdamW optimizer (Loshchilov & Hutter, 2017) is used with a weight decay of  $2 \times 10^{-3}$ . We use a cosine annealing scheduler (Loshchilov & Hutter, 2016) with a 5-epoch warm-up phase, during which the learning rate increases gradually to five times the initial value. For data augmentation, we use top-left, bottom-right, and center crops during training and use the average feature of the three cropped images.

### B.3 BLIND VIDEO QUALITY ASSESSMENT

**Datasets:** The following five BVQA datasets are used to train and evaluate the proposed algorithm.

- LSVQ (Ying et al., 2021): It is one of the largest datasets consisting of 39K videos, split into 28K for training and 11K for testing. The resolutions of videos are from 99p to 4K.
- KoNViD-1k (Hosu et al., 2017): It contains 1200 videos selected from YFCC-100M (Thomee et al., 2016) to cover various contents and distortions. All videos have a resolution of 540p.
- LIVE-VQC (Sinno & Bovik, 2018): It consists of 585 videos with various resolutions from 240p to 1080p.
- CVD2014 (Nuutinen et al., 2016): It is composed of 234 videos in five distinct scene types, filmed using 78 different cameras.
- YouTube-UGC (Wang et al., 2019): It provides about 1K video samples from user-generated contents on YouTube. The video resolutions range from 360p to 4K.

**Implementation details:** Recent state-of-the-art approaches in the BVQA task employ multiple backbone networks to capture diverse aspects of a video signal. Following this trend, we adopt a dual-backbone architecture in Figure 7. Specifically, we utilize ViT-B for the spatial feature extractor, which encodes  $N$  sampled frames  $I_i^1, I_i^2, \dots, I_i^N$  into spatial feature vectors  $z_i^{I_1}, z_i^{I_2}, \dots, z_i^{I_N}$ . These are then averaged to obtain a compact spatial representation  $z_i^s$ . For the temporal feature extractor, we adopt the Fast pathway network of the SlowFast video recognizer (Feichtenhofer et al., 2019) to extract temporal feature maps  $z_i^{P_1}, z_i^{P_2}, \dots, z_i^{P_M}$  from  $M$  sampled clips  $P_i^1, P_i^2, \dots, P_i^M$ . To enhance the temporal representations, we further refine the extracted temporal feature maps using a transformer module. This module performs inter-frame and intra-frame attention for each temporal feature map, yielding refined temporal features  $\tilde{z}_i^{P_1}, \tilde{z}_i^{P_2}, \dots, \tilde{z}_i^{P_M}$ . The refined features are then averaged to obtain the final temporal representation  $z_i^t$ . Finally, the spatial and temporal representations are fused to generate the final feature vector  $z_i$ , which is used to compute the proposed ConOrd loss.

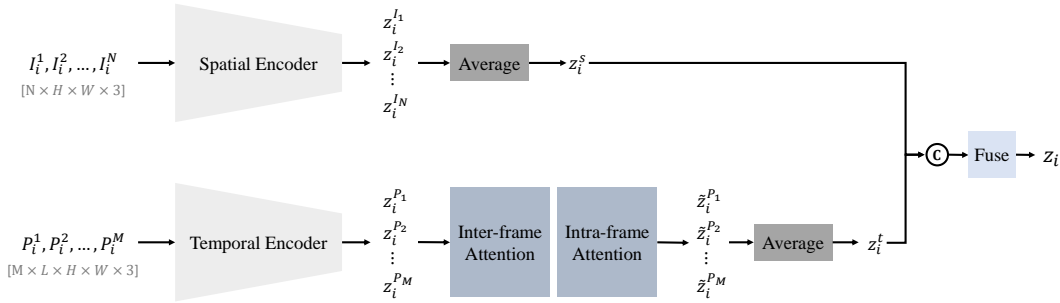


Figure 7: Network architecture for BVQA.

### B.4 TRAINING AND INFERENCE CONFIGURATIONS

Table 8 summarizes the training and inference configurations for the proposed ConOrd algorithm.

Table 8: Training and inference configurations.

Dataset	MORPH II	CLAP2015	AgeDB	UTK	CACD	Adience	BID	CLIVE	KonIQ10k	SPAQ	FLIVE	LSVQ
Learning rate	$5 \times 10^{-6}$	$5 \times 10^{-6}$	$5 \times 10^{-6}$	$5 \times 10^{-7}$	$5 \times 10^{-7}$	$2 \times 10^{-6}$	$2 \times 10^{-6}$	$2 \times 10^{-6}$	$2 \times 10^{-6}$	$2 \times 10^{-6}$	$1 \times 10^{-6}$	$5 \times 10^{-6}$
Batch size	128	128	128	128	128	128	32	32	64	64	64	16
$\tau$ in (3)	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
$\epsilon$ in (3)	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$	$10^{-7}$
$k$ in (10)	4	4	60	60	60	60	10	10	10	10	10	30

## C MORE EXPERIMENTAL RESULTS

### C.1 MORE COMPARISON ON ADDITIONAL BENCHMARKS

**Datasets:** We evaluate the performance of ConOrd on two additional regression datasets previously used in the evaluation of RnC (Zha et al., 2023).

- SkyFinder (Mihail et al., 2016; Chu et al., 2018): It is a dataset for predicting ambient temperatures from outdoor webcam images. It consists of 35,417 images taken by 44 different cameras under diverse weather and lighting conditions. The corresponding temperature values range from  $-20^{\circ}\text{C}$  to  $49^{\circ}\text{C}$ . Following Zha et al. (2023), we split the dataset into 28,373 training and 3,522 test images.
- MPIIFaceGaze (Zhang et al., 2017a;b): It is a dataset for estimating gaze directions from face images, containing 213,659 face images collected from 15 participants during natural laptop use. We divide the dataset to construct a training set of 33,000 images and a test set of 6,000 images, ensuring no participant overlap across splits. Each image is annotated with a 2D gaze vector of pitch and yaw angles, where pitch angles range from  $-40^{\circ}$  to  $10^{\circ}$ , and yaw from  $-45^{\circ}$  to  $45^{\circ}$ .

For a fair comparison, we adopt the ResNet-18 backbone (He et al., 2016) as done in RnC (Zha et al., 2023). Table 9 shows that ConOrd achieves the best performance, outperforming all prior methods on both datasets.

Table 9: Performance comparison on the SkyFinder and MPIIFaceGaze datasets.

Algorithm	SkyFinder	MPIIFaceGaze
	MAE ( $\downarrow$ )	Angular MAE ( $\downarrow$ )
DEX (Rothe et al., 2015)	3.58	5.72
OR (Niu et al., 2016)	2.92	5.86
DLDL-v2 (Gao et al., 2018)	2.99	5.47
CORN (Shi et al., 2023)	3.24	5.88
RnC (Zha et al., 2023)	<u>2.86</u>	<u>5.27</u>
ConOrd (Proposed)	<b>2.65</b>	<b>4.98</b>

**Regression examples:** We provide examples of regression results on the SkyFinder and MPIIFaceGaze datasets in Figure 8 and Figure 9, respectively.

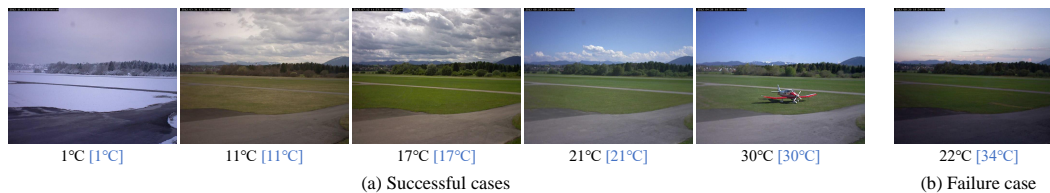


Figure 8: Examples of regression results on the SkyFinder dataset. The estimated and ground-truth values are specified under each image: estimated [true].



Figure 9: Examples of regression results on the MPIIFaceGaze dataset. The estimated and ground-truth values are specified under each image: estimated [true].

## C.2 HYPERPARAMETER ANALYSIS

**Performance according to temperature  $\tau$  in (3):** Table 10 reports the performance of the proposed algorithm according to the temperature parameter  $\tau$  in (3), which controls the smoothness of the representation distribution. We observe that the performance is relatively stable within a range  $0.05 \leq \tau \leq 0.10$ , achieving the best results when  $\tau = 0.07$ , with the minimal MAE on CLAP2015 and consistently high correlation scores on BID and LSVQ. With larger values (*e.g.*,  $\tau \geq 1.0$ ), the performance degrades gradually across all datasets, indicating that excessively large temperature values reduce the effectiveness of the proposed contrastive learning objective.

Table 10: Performance of the proposed algorithm according to  $\tau$ .

$\tau$	CLAP2015 (Age)	BID (BIQA)		LSVQ-test (BVQA)	
	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
0.05	2.475	0.910	0.925	0.904	0.904
0.06	2.462	0.912	0.926	0.903	0.903
0.07	2.461	0.913	0.925	0.904	0.904
0.08	2.480	0.913	0.931	0.903	0.904
0.09	2.466	0.911	0.921	0.903	0.903
0.10	2.530	0.911	0.924	0.903	0.903
0.50	2.576	0.901	0.913	0.897	0.893
1.00	2.591	0.888	0.905	0.897	0.894
1.50	2.589	0.883	0.901	0.895	0.892
2.00	2.770	0.888	0.900	0.892	0.890

We further visualize how the performance varies with the temperature parameter  $\tau$ . As shown in Figure 10, both SRCC and PCC remain stable within a practical range ( $0.05 \leq \tau \leq 0.10$ ) and begin to deteriorate only when  $\tau$  becomes excessively large. The corresponding training curves also indicate consistent convergence behavior across all settings.

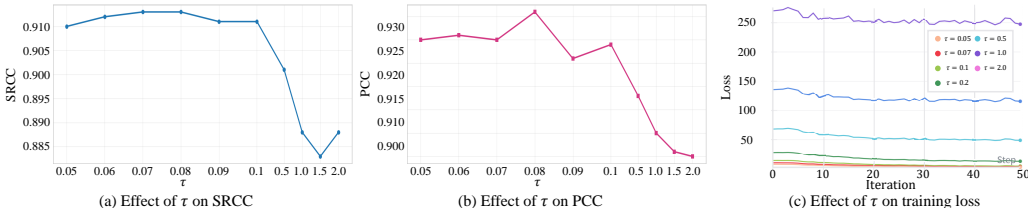


Figure 10: Sensitivity of ConOrd to the temperature parameter  $\tau$  on the BID dataset.

**Performance according to  $\epsilon$  in (4):** Table 11 presents the results for varying values of  $\epsilon$  in (4), which is used to prevent division by zero in the affinity weight computation. The model yields stable performance across a wide range of  $\epsilon$  values, indicating that it is robust to the choice of  $\epsilon$ .

Table 11: Performance of the proposed algorithm according to  $\epsilon$ .

$\epsilon$	CLAP2015 (Age)	BID (BIQA)		LSVQ-test (BVQA)	
	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
$10^{-3}$	2.494	0.915	0.928	0.904	0.904
$10^{-4}$	2.483	0.925	0.926	0.904	0.904
$10^{-5}$	2.485	0.914	0.929	0.903	0.903
$10^{-6}$	2.472	0.914	0.930	0.902	0.903
$10^{-7}$	2.461	0.921	0.925	0.904	0.904
$10^{-8}$	2.484	0.910	0.928	0.903	0.903
$10^{-9}$	2.483	0.911	0.925	0.904	0.904

To complement the quantitative results in Table 11, we additionally visualize the effect of  $\epsilon$  in Figure 11, showing its influence on BID performance and training behavior. SRCC and PCC remain nearly unchanged across a wide range of  $\epsilon$ , indicating strong robustness to this parameter. Training loss curves also show consistent convergence for all tested values.

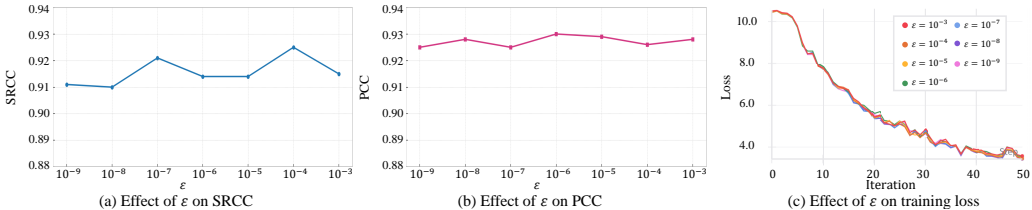


Figure 11: Sensitivity of ConOrd to  $\epsilon$  on the BID dataset.

**Performance according to  $k$  in (10):** Table 12 reports the results for varying values of  $k$  used during  $k$ -NN inference. The performance remains stable across a broad range of  $k$ , with only marginal fluctuations across all datasets. The detailed configuration of the  $k$  values is provided in Table 8.

Table 12: Performance across different values of  $k$  used for  $k$ -NN inference.

$k$	CLAP2015 (Age)	AgeDB (Age)	BID (BIQA)		LSVQ-test (BVQA)	
	MAE ( $\downarrow$ )	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
4	2.461	5.232	0.910	0.924	0.894	0.892
10	2.483	5.159	0.913	0.925	0.901	0.900
20	2.496	5.158	0.909	0.923	0.903	0.903
30	2.503	5.158	0.910	0.924	0.904	0.904
40	2.496	5.154	0.910	0.923	0.904	0.904
50	2.487	5.158	0.910	0.923	0.904	0.904
60	2.486	5.145	0.910	0.923	0.904	0.904

### C.3 ADDITIONAL ANALYSIS

**Performance according to loss balancing factors in (9):** Table 13 reports the impact of varying the weights of  $L_{center}$  relative to  $L_{ConOrd}$  across three benchmark datasets. The results indicate that the model is relatively robust to changes in the loss balancing factor, with only minor variations in MAE, SRCC, and PCC. Notably, setting both weights to 1.0 yields the best MAE on CLAP2015, suggesting that equal emphasis on the ordinal contrastive term and the intra-class compactness term provides the most effective trade-off for stable and accurate learning.

Table 13: Performance according to loss balancing factors.

Loss combination	CLAP2015 (Age)	BID (BIQA)		LSVQ-test (BVQA)	
	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
$1.0 \times L_{ConOrd} + 0.5 \times L_{center}$	2.484	0.9127	0.9182	0.903	0.904
$1.0 \times L_{ConOrd} + 0.8 \times L_{center}$	2.488	0.9023	0.9170	0.903	0.903
$1.0 \times L_{ConOrd} + 1.2 \times L_{center}$	2.487	0.9005	0.9158	0.904	0.904
$1.0 \times L_{ConOrd} + 1.5 \times L_{center}$	2.485	0.9036	0.9173	0.902	0.903
$1.0 \times L_{ConOrd} + 1.0 \times L_{center}$	2.461	0.9025	0.9165	0.904	0.904

**Additional configurations for loss in (3):** Table 14 extends the analysis presented in Table 7 by including additional configurations of  $\kappa_{ij}$ ,  $a_{ij}$ , and  $b_{ij}$ , as well as results on datasets beyond CLAP2015. These additional configurations explore alternative formulations, such as square-root and logarithmic scaling (methods XI and XII) and truncated weights (method XIII). The table provides a comprehensive view of how different design choices influence results. Overall, it reaffirms the effectiveness of the proposed configuration (method VIII) in accurately capturing ordinal relationships.

Table 14: Performance according to different configurations of  $\kappa_{ij}$ ,  $a_{ij}$ ,  $b_{ij}$  in (3). Note that method IX sets  $\Delta$  to 0 if  $r_i = r_j$ , and to a positive threshold of 5 otherwise.

Method	$\kappa_{ij}$	$a_{ij}$	$b_{ij}$	CLAP2015 (Age)	BID (BIQA)		LSVQ-test (BVQA)	
				MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
I	$z_i^T z_j$	$\frac{1}{ r_i - r_j  + \epsilon}$	1	2.516	0.9095	0.9263	0.902	0.901
II	$z_i^T z_j$	$\frac{1}{ r_i - r_j  + \epsilon}$	$ r_i - r_j $	2.494	0.9044	0.9180	0.904	0.904
III	$z_i^T z_j$	$\frac{1}{(r_i - r_j)^2 + \epsilon}$	1	2.483	0.9078	0.9235	0.903	0.903
IV	$z_i^T z_j$	$\frac{1}{(r_i - r_j)^2 + \epsilon}$	$(r_i - r_j)^2$	2.518	0.9088	0.9241	0.904	0.903
V	$-\ z_i - z_j\ _2^2$	$\frac{1}{ r_i - r_j  + \epsilon}$	1	2.497	0.9086	0.9246	0.903	0.902
VI	$-\ z_i - z_j\ _2^2$	$\frac{1}{ r_i - r_j  + \epsilon}$	$ r_i - r_j $	2.490	0.9096	0.9257	0.903	0.902
VII	$-\ z_i - z_j\ _2^2$	$\frac{1}{(r_i - r_j)^2 + \epsilon}$	1	2.472	0.9105	0.9242	0.902	0.902
VIII	$-\ z_i - z_j\ _2^2$	$\frac{1}{(r_i - r_j)^2 + \epsilon}$	$(r_i - r_j)^2$	2.461	0.9128	0.9250	0.904	0.904
IX	$-\ z_i - z_j\ _2^2$	$\frac{1}{(\Delta^2 + \epsilon)^{-1}}$	$\Delta^2$	2.536	0.6951	0.6948	0.834	0.830
X	$-\ z_i - z_j\ _2^2$	Learnable $a_{ij}$	Learnable $b_{ij}$	2.880	0.6679	0.6854	0.882	0.881
XI	$-\ z_i - z_j\ _2^2$	$\frac{1}{\sqrt{ r_i - r_j  + \epsilon}}$	$\sqrt{ r_i - r_j }$	2.473	0.9099	0.9240	0.901	0.900
XII	$-\ z_i - z_j\ _2^2$	$\frac{1}{\log(1 +  r_i - r_j ) + \epsilon}$	$\log(1 +  r_i - r_j )$	2.500	0.9038	0.9199	0.901	0.900
XIII	$-\ z_i - z_j\ _2^2$	$\begin{cases} \frac{1}{ r_i - r_j  + \epsilon}, & \text{if }  r_i - r_j  < 10, \\ \frac{1}{10}, & \text{otherwise,} \end{cases}$	$\begin{cases}  r_i - r_j , & \text{if }  r_i - r_j  < 10, \\ 10, & \text{otherwise.} \end{cases}$	2.500	0.8958	0.9143	0.900	0.899

**Effect of  $L_{\text{center}}$ :** We evaluate the impact of the center loss by applying  $L_{\text{center}}$  to  $L_{\text{SupCon}}$ ,  $L_{\text{RnC}}$ , and  $L_{\text{ConOrd}}$ . As shown in Table 15, ConOrd performs strongly even without the center term, confirming that its gains mainly arise from the loss design itself. Adding  $L_{\text{center}}$  yields small improvements and does not alter the relative ranking among methods. Overall, the center loss acts as a mild stabilizer rather than a key performance factor.

Table 15: Ablation of the effect of the center loss  $L_{\text{center}}$ .

Method	CLAP2015 (Age)	BID (BIQA)		LSVQ-1080p (BVQA)	
	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
$L_{\text{SupCon}}$	2.625	0.819	0.876	0.614	0.682
$L_{\text{SupCon}} + L_{\text{center}}$	2.610	0.815	0.875	0.622	0.689
$L_{\text{RnC}}$	2.531	0.892	0.906	0.812	0.843
$L_{\text{RnC}} + L_{\text{center}}$	2.745	0.892	0.906	0.808	0.839
$L_{\text{ConOrd}}$	2.509	0.909	0.925	0.815	0.848
$L_{\text{ConOrd}} + L_{\text{center}}$	2.461	0.913	0.925	0.818	0.851

**Initialization of reference points  $\mu_m$ :** We assess several initialization strategies for  $\mu_m$ . On the BID dataset, random, zero, truncated normal, and Kaiming normal initializations yield nearly identical results, indicating that the method is largely insensitive to initialization.

Table 16: Performance across different initialization schemes for  $\mu_m$  on BID.

Init. method	Random	Zeros	Trunc. Normal	Kaiming Normal
SRCC	0.913	0.910	0.913	0.910
PCC	0.925	0.922	0.925	0.925

We further test initializing  $\mu_m$  with the per-rank mean feature. Although this variant offers a small performance gain, it incurs a substantial overhead due to the extra dataset pass. Considering the minimal improvement, random initialization is a more efficient choice.

Table 17: Random initialization versus mean-feature initialization.

Method	SRCC	PCC	Time (s)
Random init.	0.913	0.925	$1.21 \times 10^{-3}$
Mean-feature init.	0.915	0.932	5.25

**Resilience to reduced training data:** To assess robustness under limited ordering information, we progressively subsample the AgeDB training set and compare the resulting MAE performance — evaluated on the full test set — of ConOrd with SupCon and RnC. As shown in Table 18, ConOrd consistently outperforms both baselines across all sampling ratios. The advantage is most pronounced in low-data settings (*e.g.*, ratios of 0.1 and 0.3), suggesting that ConOrd learns more sample-efficient and stable ordinal representations when supervision is scarce.

Table 18: MAE results on AgeDB under different training-set sampling ratios, where each ratio denotes the proportion of training data used.

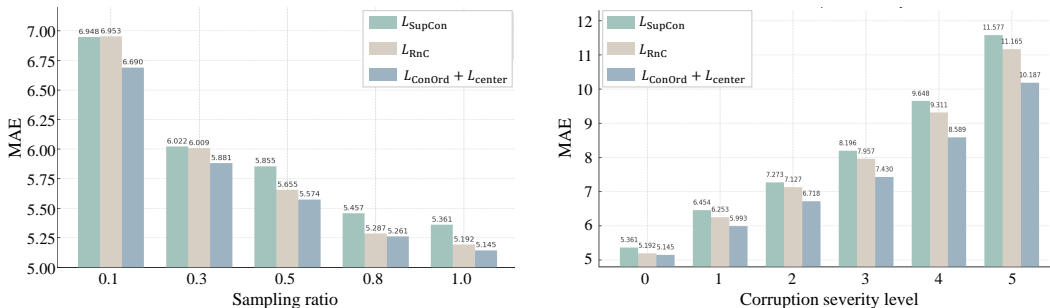
Sampling ratio	$L_{\text{SupCon}}$	$L_{\text{RnC}}$	$L_{\text{ConOrd}} + L_{\text{center}}$
0.1	6.948	6.953	6.690
0.3	6.022	6.009	5.881
0.5	5.855	5.655	5.574
0.8	5.457	5.287	5.261
1.0	5.361	5.192	5.145

**Robustness to data corruption:** We adopt the corruption process defined in the ImageNet-C protocol (Hendrycks & Dietterich, 2019) and apply it to the AgeDB test set to evaluate robustness under data degradation. All methods are trained on the clean AgeDB training data, and MAE is measured on corrupted versions of the test images across 19 corruption types and severity levels 0–5. As shown in Table 19, ConOrd achieves the best performance under clean conditions (severity 0) and exhibits the slowest degradation as severity increases. Even at the highest corruption level (severity 5), ConOrd maintains a lower MAE than SupCon and RnC, indicating stronger robustness to corrupted inputs.

Table 19: MAE results on AgeDB under test-time data corruptions.

Corruption severity level	$L_{\text{SupCon}}$	$L_{\text{RnC}}$	$L_{\text{ConOrd}} + L_{\text{center}}$
0	5.361	5.192	5.145
1	6.454	6.253	5.993
2	7.273	7.127	6.718
3	8.196	7.957	7.430
4	9.648	9.311	8.589
5	11.577	11.165	10.187

While Tables 18 and 19 summarize the numerical results, the corresponding visualizations in Figure 12 help illustrate the relative performance trends. In both reduced-data and corruption scenarios, ConOrd shows a consistently favorable margin over the baselines.



(a) Resilience to reduced training data.

(b) Robustness to data corruption.

Figure 12: Comparison of SupCon, RnC, and ConOrd under reduced supervision and test-time corruptions.

**Standard deviation of performance:** To assess the reliability of ConOrd, we report the mean and standard deviation across multiple random seeds. We use five seeds for the age estimation (CLAP2015) and BVQA (LSVQ-test) benchmarks, and ten seeds for the BIQA (BID) task. Table 20 summarizes the resulting variability. ConOrd exhibits stable performance across datasets.

Table 20: Mean and standard deviation of ConOrd across multiple random seeds.

	CLAP2015 (Age)	BID (BIQA)		LSVQ-test (BVQA)	
	MAE ( $\downarrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )	SRCC ( $\uparrow$ )	PCC ( $\uparrow$ )
$L_{\text{SupCon}}$	$2.6058 \pm 0.0121$	$0.8182 \pm 0.0052$	$0.8754 \pm 0.0035$	$0.7576 \pm 0.0034$	$0.7660 \pm 0.0032$
$L_{\text{RnC}}$	$2.5324 \pm 0.0202$	$0.8901 \pm 0.0260$	$0.9041 \pm 0.0237$	$0.9024 \pm 0.0006$	$0.9004 \pm 0.0011$
ConOrd	$2.4698 \pm 0.0122$	$0.9118 \pm 0.0243$	$0.9255 \pm 0.0183$	$0.9040 \pm 0.0007$	$0.9036 \pm 0.0009$

**t-SNE visualization of learned embeddings:** Figure 13 provides a qualitative t-SNE comparison of embeddings learned by  $L_{\text{SupCon}}$ ,  $L_{\text{RnC}}$ , and  $L_{\text{ConOrd}} + L_{\text{center}}$  on the BID dataset. While t-SNE does not preserve global geometry and should be interpreted with caution, some overall trends can be observed. SupCon yields largely mixed points with no apparent ordering, and RnC shows a coarse progression but with some overlap between neighboring quality levels. ConOrd produces a more coherent progression of points along a smooth trajectory, suggesting that its embedding space reflects ordinal structure more clearly under this visualization.

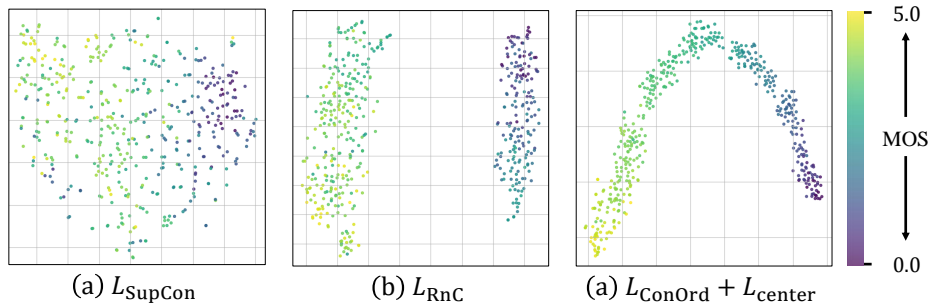


Figure 13: t-SNE visualization of embeddings learned by SupCon, RnC, and ConOrd on the BID dataset. Colors denote MOS scores.

**Grad-CAM visualization:** Figure 14 presents Grad-CAM maps computed independently for a test image and its top- $k$  nearest neighbors in the embedding space. Each heatmap highlights the regions that influence the encoder’s representation for that individual image, without implying any direct

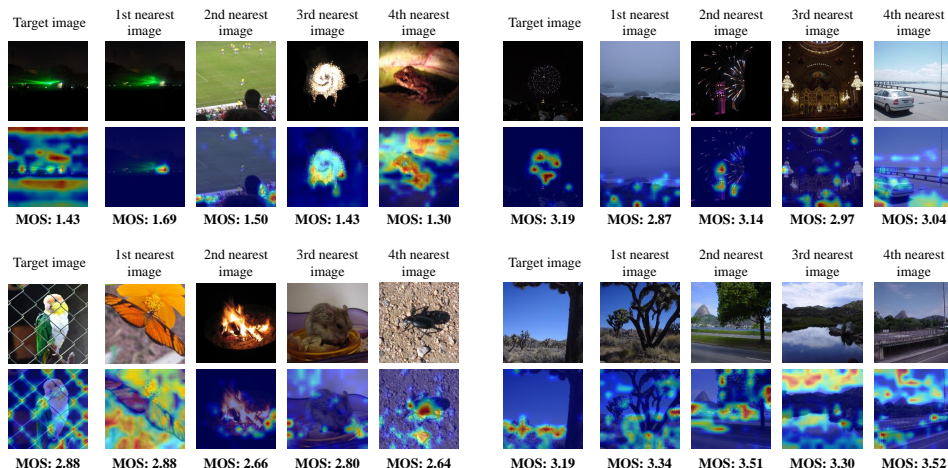
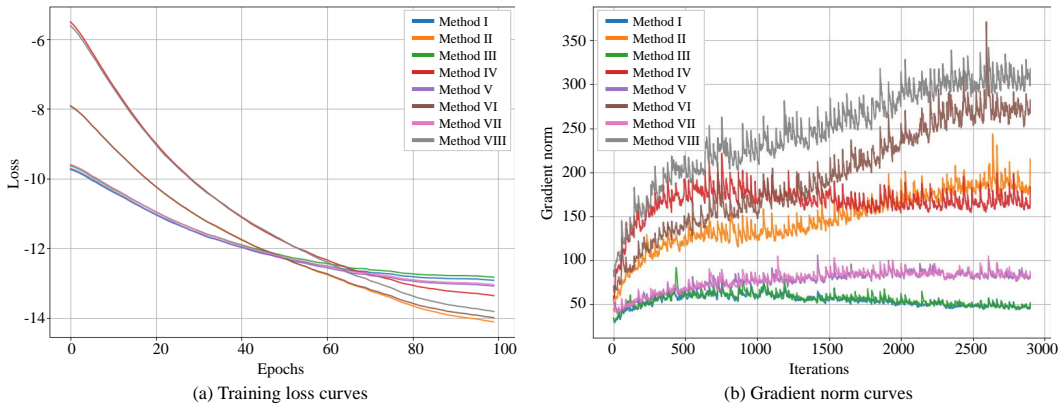


Figure 14: Grad-CAM maps computed independently for a test image and its top- $k$  nearest neighbors in the embedding space on the BID dataset.

1350 explanation of the pairwise similarity. Across the examples, ConOrd tends to activate qualitatively  
 1351 meaningful areas related to perceptual quality. Although Grad-CAM offers only a qualitative view of  
 1352 model focus, these patterns suggest that ConOrd forms representations that align with perceptually  
 1353 relevant content.

1354 **Training stability and gradient dynamics:** Figure 15 compares the optimization behavior of the  
 1355 eight weighting configurations listed in Table 1. All configurations show smooth, monotonic decreases  
 1356 in training loss without signs of instability or divergence, indicating that the ConOrd formulation  
 1357 remains robust across a wide range of affinity-disparity designs. The gradient norms differ moderately  
 1358 in scale but remain consistently bounded and settle quickly into steady ranges. These results confirm  
 1359 that all eight variants exhibit well-behaved gradients and stable training dynamics, demonstrating the  
 1360 resilience of the proposed contrastive order formulation.



1375 Figure 15: Training loss and gradient-norm dynamics across different weighting configurations,  
 1376 showing stable convergence and controlled gradients in all cases.

#### 1378 C.4 COMPLEXITY

1380 We use PyTorch and NVIDIA GeForce RTX 4090 GPUs for all experiments.

1381 **Training time:** Table 21 compares the average training times required for training one epoch on the  
 1382 SPAQ dataset. The proposed algorithm achieves the fastest training time per epoch. This efficiency  
 1383 is attributed to its design, which eliminates the need for data augmentation and pairwise sample  
 1384 construction. Unlike  $L_{\text{SupCon}}$  and  $L_{\text{RnC}}$ , which should generate augmented sample pairs during  
 1385 training, the proposed ConOrd eliminates this step and improves efficiency. The RnC loss incurs  
 1386 the longest training time because it needs to dynamically select negative samples based on label  
 1387 distances for each anchor-positive pair. This conditional filtering introduces computational overhead  
 1388 and hinders parallelization. In contrast, ConOrd uses fixed weight masks, allowing more efficient and  
 1389 parallel computations. Thus, the proposed ConOrd loss requires the shortest computation time.

1390 Table 21: Comparison of processing times required for training one epoch on SPAQ.

Algorithm	Time (s)
$L_{\text{SupCon}}$ in (1)	41.7
$L_{\text{RnC}}$ in (7)	65.5
$L_{\text{ConOrd}} + L_{\text{center}}$ in (9)	31.8

1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398 **Testing time on SPAQ:** We also report the average testing time on the SPAQ dataset. The whole  
 1399 process takes only  $5.0 \times 10^{-3}$  seconds to test an image on average: 10.0 seconds for the full test  
 1400 feature extraction, and  $1.5 \times 10^{-4}$  seconds for the score estimation. Hence, ConOrd provides a  
 1401 computationally efficient solution for practical deployment.

**Testing time on CLAP2015, BID, and LSVQ-test:** To further evaluate the reliability of the  $k$ -NN inference, we measure the end-to-end test-time latency for processing each full test dataset. We repeat this measurement over multiple runs and report the mean and standard deviation in Table 22. Compared with the state-of-the-art baselines for each task — NumCLIP for age estimation, LoDa for BIQA, and DOVER for BVQA — ConOrd achieves consistently faster and stable test-time performance across all datasets.

Table 22: Mean and standard deviation of end-to-end test-time latency, measuring the variability of  $k$ -NN inference, on the CLAP2015, BID, and LSVQ-test datasets.

Method	CLAP2015 (Age)	BID (BIQA)	LSVQ-test (BVQA)
SOTA baseline	NumCLIP: $5.15 \pm 0.42s$	LoDa: $5.63 \pm 0.31s$	DOVER: $1222.36 \pm 9.04s$
ConOrd ( $k$ -NN)	$2.16 \pm 0.14s$	$0.88 \pm 0.05s$	$253.70 \pm 7.65s$

**Loss complexity:** To provide a clearer comparison of efficiency, we report the per-batch loss computation time and the GPU memory usage associated with computing  $L_{ConOrd}$ ,  $L_{RnC}$ , and the BIQA loss used in QCN. As summarized in Tables 23 and 24,  $L_{ConOrd}$  is the most computationally efficient in both runtime and memory consumption.

Table 23: Loss-level computation time on BID (batch size = 32).

Method	$L_{ConOrd}$	$L_{RnC}$	QCN
Loss computation time (batch size = 32)	4.6ms	8.7ms	8.5ms

Table 24: GPU memory usage for loss computation on BID.

Algorithm	Memory
$L_{ConOrd}$	2.17MB
$L_{RnC}$	1.17MB
QCN	273.78MB

**Model efficiency:** Table 25 compares the model complexity of the proposed BIQA algorithm with those of other recent algorithms. The proposed algorithm adopts ViT-B as the encoder, resulting in a complexity of 86M. While this is not the smallest among the compared models, the proposed algorithm consistently outperforms others across multiple BIQA benchmarks. In particular, it achieves better performance than LoDa, which uses a larger model, and LQMamba, which employs the same encoder architecture.

Table 25: Comparison with BIQA algorithms in terms of network complexity.

Algorithm	# parameters (M)
ReIQA (Saha et al., 2023)	47
LQMamba (Guan et al., 2024)	86
QCN (Shin et al., 2024)	30
LoDa (Xu et al., 2024)	95
ConOrd (Proposed)	86

1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

## D REGRESSION RESULTS

Figures 16, 17, and 18 show regression results of the proposed ConOrd on the facial age estimation, BIQA, and BVQA tasks, respectively.

### D.1 FACIAL AGE ESTIMATION



Figure 16: (a) Success and (b) failure cases of regression results on the facial age estimation datasets. Under each image, the estimated age is specified with the ground-truth in brackets.

1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

D.2 BLIND IMAGE QUALITY ASSESSMENT

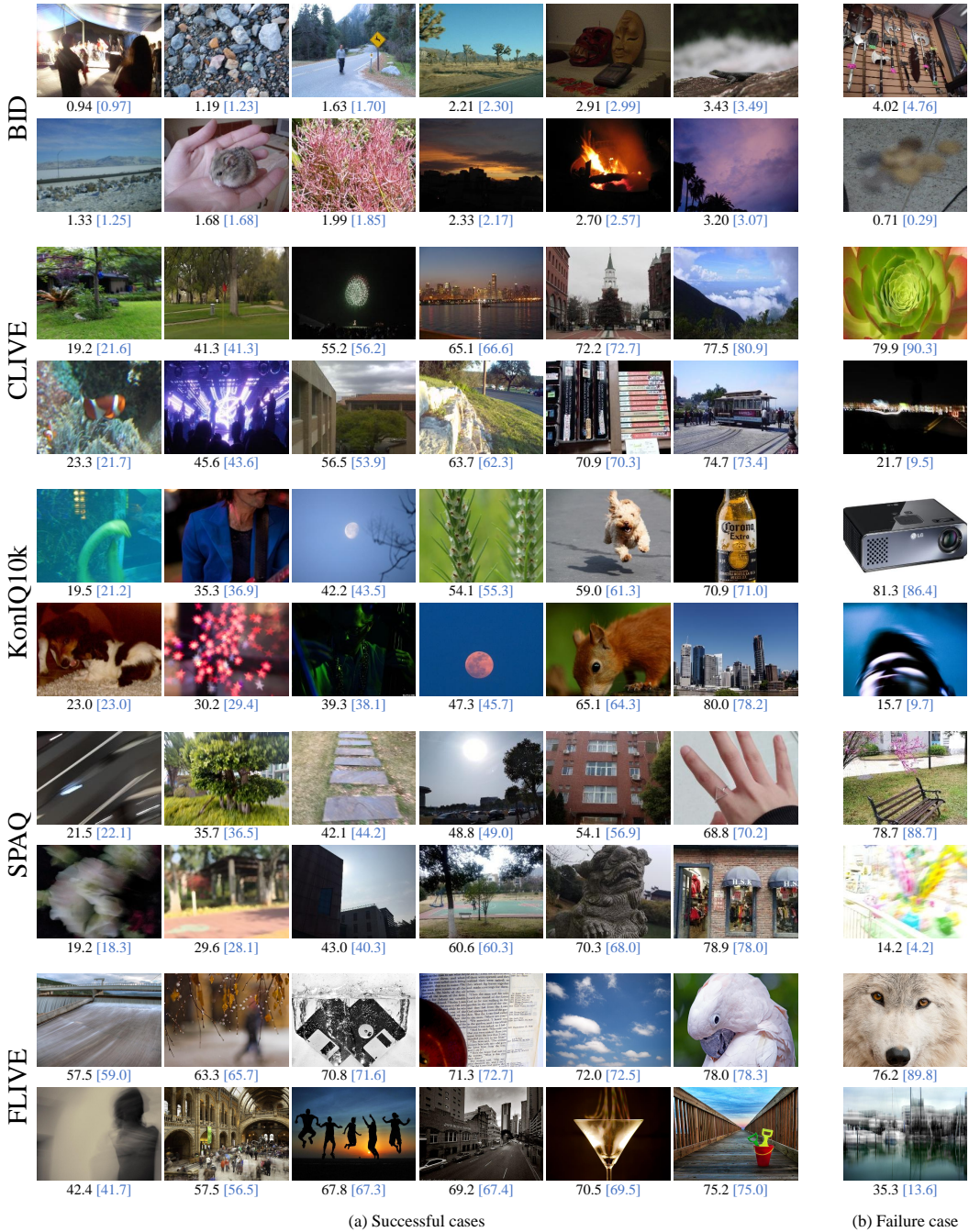


Figure 17: (a) Success and (b) failure cases of regression results on the BIQA datasets. Under each image, the estimated quality score is specified with the ground-truth in brackets.

1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

### D.3 BLIND VIDEO QUALITY ASSESSMENT

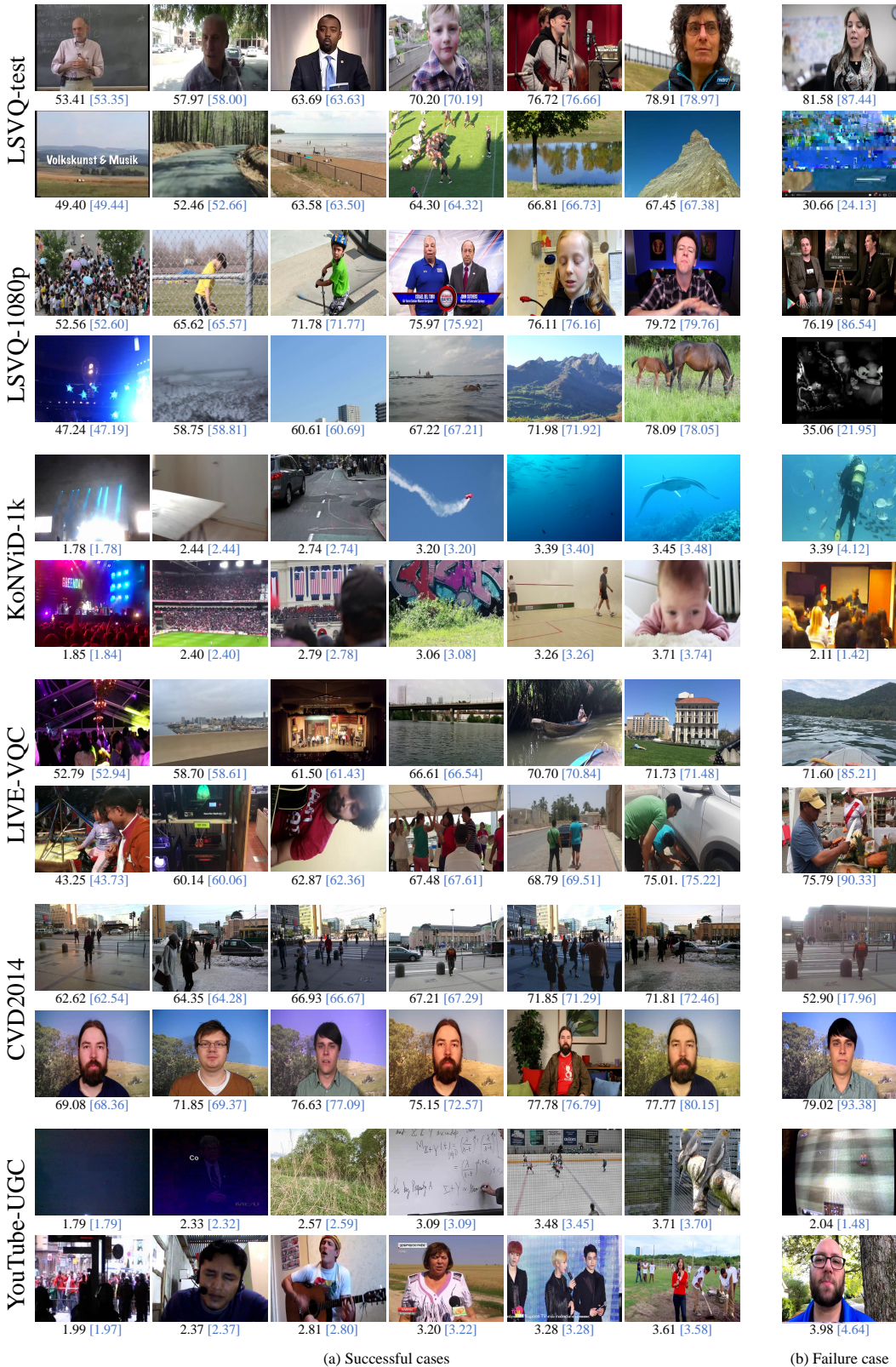


Figure 18: (a) Success and (b) failure cases of regression results on the BVQA datasets. Under each image, the estimated quality score is specified with the ground-truth in brackets.

## E LIMITATIONS

As shown in Appendix D, the proposed algorithm generally demonstrates strong predictive performance across a variety of regression tasks and dataset types. However, along with successful cases, failure cases are also illustrated in Figures 16(b), 17(b), and 18(b). It is observed from Figure 16(b) that for the task of facial age estimation, the model tends to under-estimate the age of older people with smooth skin or strong lighting and over-estimate the age of younger individuals when the images are shadowed. From the BIQA and BVQA results in Figures 17(b) and 18(b), we observe that prediction errors are more frequent when MOS values are extremely low or high. This appears to stem from the limited representation of such samples in the training data, suggesting that the performance could be further improved by adopting learning strategies that better handle imbalanced data distributions.

## F BROADER IMPACTS

This work proposes a general method for ordinal regression that demonstrates strong performance across tasks, including facial age estimation, blind image quality assessment, and blind video quality assessment. However, caution is needed when deploying such models in sensitive domains. If used without appropriate safeguards, predictions involving human or facial attributes may raise ethical concerns, especially in the presence of biases within the training data. We recommend that future deployments incorporate fair evaluations and that the model be used as a decision-support tool rather than as an autonomous system.