Dual-level Fuzzy Learning with Patch Guidance for Image Ordinal Regression

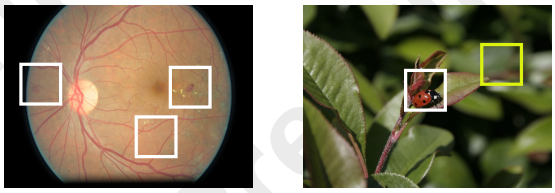
Chunlai Dong[1,3,4,7], **Haochao Ying**[2,3,7,*], **Qibo Qiu**[5], **Jinhong Wang**[1,3,4,7], **Danny Chen**[6] and **Jian Wu**[2,3,4,7,*]
[1]College of Computer Science and Technology, Zhejiang University
[2]School of Public Health, Zhejiang University
[3]State Key Laboratory of Transvascular Implantation Devices, The Second Affiliated Hospital Zhejiang University School of Medicine
[4]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence
[5]China Mobile (Zhejiang) Research & Innovation Institute
[6]College of Engineering, University of Notre Dame
[7]Transvascular Implantation Devices Research Institute
{dongcl, haochaoying, qiuqibo_zju, wangjinhong, wujian2000}@zju.edu.cn, dchen@nd.edu

Abstract

Ordinal regression bridges regression and classification by assigning objects to ordered classes. While human experts rely on discriminative patch-level features for decisions, current approaches are limited by the availability of only image-level ordinal labels, overlooking fine-grained patch-level characteristics. In this paper, we propose a Dual-level Fuzzy Learning with Patch Guidance framework, named DFPG that learns precise feature-based grading boundaries from ambiguous ordinal labels, with patch-level supervision. Specifically, we propose patch-labeling and filtering strategies to enable the model to focus on patch-level features exclusively with only image-level ordinal labels available. We further design a dual-level fuzzy learning module, which leverages fuzzy logic to quantitatively capture and handle label ambiguity from both patch-wise and channel-wise perspectives. Extensive experiments on various image ordinal regression datasets demonstrate the superiority of our proposed method, further confirming its ability in distinguishing samples from difficult-to-classify categories. The code is available at https://github.com/ZJUMAI/DFPG-ord.

1 Introduction

Image ordinal regression, also known as ordinal classification in computer vision, aims to infer the ordinal labels of images. This task resides at a crucial intersection of the fundamental classification and regression paradigms, where class labels display inherent sequential or logical relationships. The image ordinal regression (grading) methodology has exhibited significant utility across diverse computer vision applications, including facial age estimation [Li *et al.*, 2019; Wen *et al.*, 2020], image aesthetic assessment [Kong *et al.*, 2016; Lee and Kim, 2019], historical image dating [Martin *et al.*, 2014], and medical disease grading [Wang *et al.*, 2024].



(a) (b)

Figure 1: Illustrating some influencing factors in human decision-making processes. (a) Doctors focus on specific lesion areas in the DR grading scenario. (b) Evaluators may increase aesthetic scores based on certain regions (*e.g.*, white rectangle) while reducing scores due to blurred areas (*e.g.*, yellow rectangle).

Some known ordinal regression approaches followed either regression or classification paradigms [Yi *et al.*, 2015; Rothe *et al.*, 2015], employing traditional optimization objectives such as mean absolute/square error, or cross-entropy loss. However, these methods did not fully account for the ambiguity and ordinality of ordinal labels. Specifically, for example, in disease grading, the boundary between scores 1 and 2 is ambiguous, and the final value often relies on a doctor's subjective judgment. Meanwhile, the ordinality reflects that a higher score indicates a more severe disease. To address ambiguity, converted the labels of samples into label distributions [Pan *et al.*, 2018; Gao *et al.*, 2017a]. When exploiting the ordinality, some classification-based approaches treated ordinal regression as a multi-class classification task, developing new strategies to assist in learning inter-class ordinal relationships (e.g., soft labeling [Díaz and Marathe, 2019] or binary label sequence prediction [Wang *et al.*, 2023]). Some ranking-based methods compared samples with specific anchors to learn the ordinal relationships [Lee and Kim, 2021a; Shin *et al.*, 2022].

Despite the effectiveness of the aforementioned methods in

*Corresponding Author: Haochao Ying and Jian Wu

enhancing ordinal regression performance [Shin *et al.*, 2022], they neglected an essential phenomenon: when making image grading decisions, it is actually the discriminative patch-level features that guide the human decision-making process. For instance, as shown in Figure 1, in the context of Diabetic Retinopathy (DR) grading, clinicians determine the severity of the disease by focusing only on specific DR lesion features within a small portion of the fundus image regions, such as hemorrhages and soft exudates [Ikram *et al.*, 2024]. Similarly, in the context of image aesthetic assessment, evaluators often base their scores not only on the overall information of an image but also tend to raise their ratings because specific regions align with personal aesthetic preferences [Palmer *et al.*, 2013]. Hence, a key challenge in current research lies in effectively leveraging patch-level granular features within the ordinal regression framework, particularly under the constraint of having only image-level ordinal labels.

In this paper, we propose a Dual-level Fuzzy Learning with Patch Guidance for image ordinal regression, termed DFPG. Specifically, we first train a network annotator offline using solely the available image-level labels, followed by patch-wise division for patch-level pseudo-label inference. Notably, to leverage the inherent ordinality of labels, we propose an Adjacent Category Mixup (ACM) method that enhances the annotator's discriminative capability between similar samples through controlled mixing of adjacent ordinal categories. Next, we posit that ordinal label ambiguity stems from two distinct sources: the inherent ambiguity in the features of constituent patch regions, and the fine-grained ambiguity in patch attribute features. To address this dual nature of ambiguity, we propose a Dual-level Fuzzy Learning (DFL) module that quantitatively analyzes label ambiguity through both patch-wise and channel-wise perspectives. Finally, to further refine the noisy patch-level pseudo-labels, we develop a co-teaching strategy with the Noise-aware Patch Filtering (NPF) paradigm to reduce the negative impact of noisy and redundant patches on training. Specifically, based on the co-teaching strategy, this module can employ two identical models that are trained alternately, providing each other with masking matrices for high-confidence patches. Our main contributions can be summarized as follows:

- We propose DFPG, a novel image ordinal regression framework that emulates the human decision-making process by focusing on discriminative patch-level features for image ordinal regression problem.
- We introduce the DFL module to effectively quantify label ambiguity through finer-grained modeling of patch-wise and channel-wise fuzzification.
- Extensive experiments across diverse datasets demonstrate the consistent superiority of our DFPG framework compared to state-of-the-art approaches, particularly in detail-sensitive scenarios like DR grading.

2 Related Work

2.1 Image Ordinal Regression

The goal of image ordinal regression is to learn a mapping rule that assigns an input image to a specific rank on an ordinal scale. Numerous methods have tackled the direct ordinal label prediction problem with different approaches to effectively leverage the ordinality. One classic approach is K-rank [Frank and Hall, 2001], which trained $K-1$ subclassifiers to rank ordinal categories. Furthermore, some methods [Chen *et al.*, 2017; Niu *et al.*, 2016] used a series of basic CNNs as K-rank classifiers. Other methods adopted an anchor-based comparison scheme. For instance, Order Learning [Lim *et al.*, 2020] designed a pairwise comparator to classify instance relationships, estimating class labels by comparing input instances with reference instances. Building on this, Lee and Kim [Lee and Kim, 2021b] developed deep repulsive clustering and order-identity decomposition methods. Similarly, MWR [Shin *et al.*, 2022] leveraged a moving window approach to refine predictions by comparing input images with reference images from adjacent categories. Recently, Ord2Seq [Wang *et al.*, 2023] treated ordinal regression as a sequence prediction process, transforming each ordinal category label into a unique label sequence, inspired by the dichotomous tree structure. On the other hand, several methods focused on the feature aspect. Some methods use probability distributions to model ordinal relationships between labels, enhancing the model's ability to learn representations. SORD [Díaz and Marathe, 2019] converted one-hot labels into soft probability distributions to train an ordinal regressor. Meanwhile, POEs [Li *et al.*, 2021] represented each data point as a multivariate Gaussian distribution with an ordinal constraint to capture the inherent characteristics of ordinal regression. In contrast, some methods focus on data generation to enhance the model's ability to distinguish subtle category differences. For example, CIG [Cheng *et al.*, 2023] addresses class imbalance and category overlap in image ordinal regression through controllable image generation. OCP-CL [Zheng *et al.*, 2024] disentangles ordinal and non-ordinal content in latent factors, augmenting non-ordinal information to generate diverse images while preserving ordinal content. In recent years, with the development of pre-trained VLMs, researchers have explored borrowing the rank concept from the language domain [Li *et al.*, 2022; Du *et al.*, 2025]. Unlike previous studies, we explicitly utilize patch-level features to uncover key factors influencing human decision-making. In addition, we apply fuzzy logic to quantitatively analyze the inherent ambiguity and ordinality in the feature-label relationships specific to ordinal regression.

2.2 Learning with Label Ambiguity

The general class distinguishability, human annotator heterogeneity, and external factors can all contribute to the ambiguity in the observed labels, which can impair the model's ability to fit the data. Several existing works have tackled this challenge using the label distribution learning (LDL) paradigm [Geng, 2016], which trained models with instances labeled by label distributions. Furthermore, DLDL [Gao *et al.*, 2017b] converts each image label into a discrete label distribution and learns the label distribution by minimizing the Kullback-Leibler divergence between the predicted and true label distributions. OLDL [Wen *et al.*, 2023] further incorporates modeling of the ordinal nature of labels within the LDL paradigm based on spatial, semantic, and temporal or-

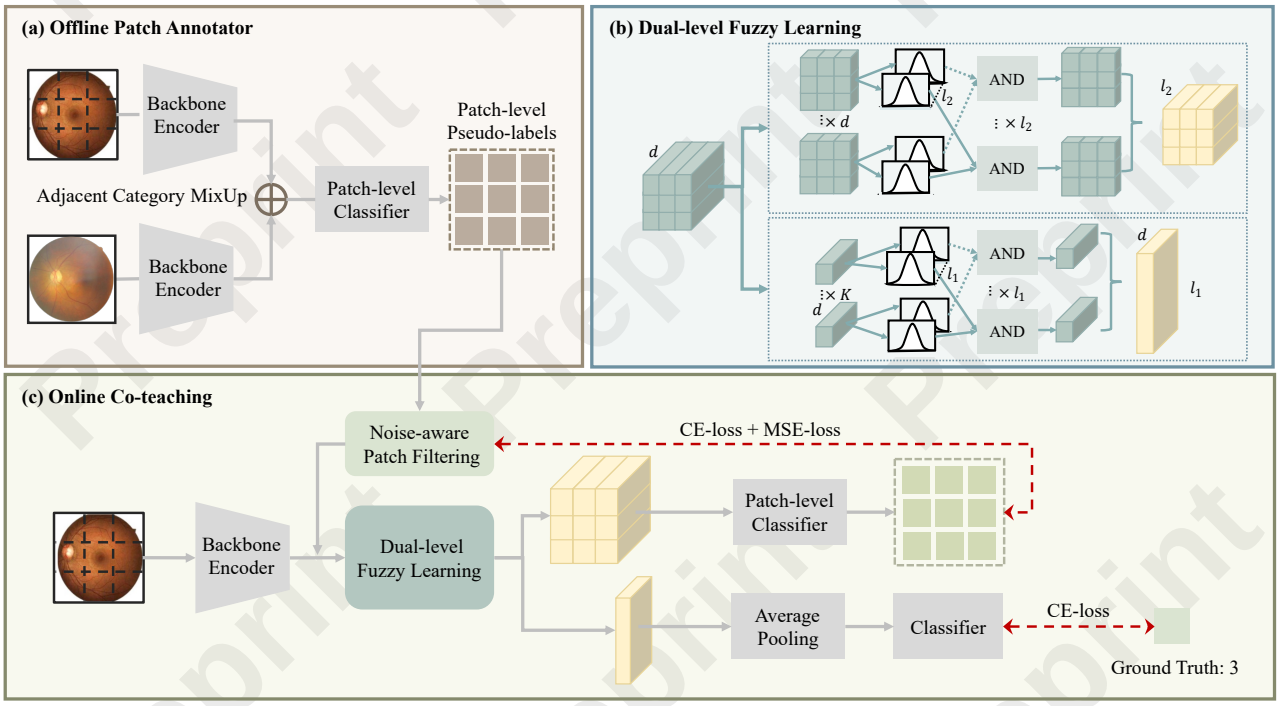


Figure 2: An overview of our DFPG approach. (a) The Patch Annotator module for generating patch-level pseudo-labels. An adjacent category sampling scheme is adopted to preserve ordinal information inherent in the augmented features, thereby enhancing the model's discriminability on samples of adjacent categories. (b) The Dual-level Fuzzy Learning module, in which multiple Gaussian membership functions are used to introduce fuzziness into the precise image representations, effectively capturing the ambiguity in feature-label associations specific to ordinal regression tasks. (c) The overall Co-teaching Strategy of our model, which incorporates a patch-level optimization objective through an unreliable patch filtering method based on the generated pseudo-labels.

der relationships. Another research direction focused on adjusting the representation space based on label ambiguity. In this context, some approaches utilized probabilistic embeddings [Shi and Jain, 2019; Chang *et al.*, 2020], where each sample was represented as a Gaussian distribution rather than a fixed point for classification. Furthermore, POEs [Li *et al.*, 2021] proposed an ordinal distribution constraint to preserve the ordinal relationships in the latent space. In contrast, our DFPG leverages fuzzy logic to model the ambiguity and ordinality of the category labels simultaneously.

3 Methodology

3.1 Overview

Unlike typical methods in the field of image ordinal regression, we focus on capturing discriminative patch-level features while fully considering the fuzziness and ordinality of classification labels during the process. Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{N}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the sets of images and their corresponding categories, respectively, with $|\mathcal{X}| = |\mathcal{Y}| = N$. Each image category $y_i \in \mathcal{C} = \{1, 2, \ldots, C\}$, where C is the total number of categories. Note that, there is directionality between different categories, which means that the difference between samples $\{(x_i, 1), (x_j, 3)\}$ is bigger than that between samples $\{(x_i, 1), (x_k, 2)\}$. This reflects the unique regression characteristics of the problem. The main goal is to obtain a model $F_\theta(\cdot) : \mathcal{X} \to \mathcal{Y}$ to accurately predict ordinal labels, consistent with a straightforward classification paradigm.

3.2 Offline Patch Annotator

With the notation given above, for a dataset $\mathcal{D}$, only the global ordinal label of each image is available. This restricted condition makes it challenging to directly model the impact of regional features on decision-making. Thus, we introduce an offline Patch Annotator module to generate the patch-level pseudo-labels, as illustrated in Figure 2(a).

Backbone Selection. Given an image sample (x, y), it is first processed through an image encoder to extract feature representations. We select the Pyramid Vision Transformer (PVT) [Wang *et al.*, 2021] as the backbone, which has been shown to be effective for ordinal regression in previous studies [Wang *et al.*, 2023]. The encoder represents the raw image x as a patch embedding in the latent space, mapping $x \to \mathbf{H} \in \mathbb{R}^{K \times d}$, where K is the number of patches and d is the embedding dimension. In addition, a fully connected layer and a 1×1 convolution layer are employed as classification heads to process features at different levels.

Patch Annotator. We first use the global labels to train a simple annotator to generate pseudo-labels for each patch, as shown in Figure 2(a). A simple pre-trained annotator is sufficient to generate pseudo-labels at the patch level [Jiang *et al.*, 2021]. Moreover, considering the inherent ordinality of labels, capturing discriminative features that effectively distin-

guish adjacent categories becomes challenging without sufficient variability in the training data. In this context, augmentation methods such as Manifold Mixup [Verma *et al.*, 2019] are helpful by creating diverse samples to enhance model performance. To control the modified samples near category boundaries, we introduce an Adjacent Category Mixup (ACM) scheme. Specifically, we first sample adjacent category pairs (x_i, x_j) with ordinal labels y_i and y_j, respectively, and $|y_i - y_j| = 1$. The images x_i and x_j are initially processed through the backbone encoder, mapping them to hidden representations $\mathbf{H}_i$ and $\mathbf{H}_j$. Subsequently, before passing them into the classification layer, the hidden representations are mixed up for augmentation following Manifold Mixup [Verma *et al.*, 2019]. The augmented samples $(\tilde{\mathbf{H}}, \tilde{y})$ are utilized to train the annotator, which is further used to perform inference on the patch-level hidden representations to generate a patch-level pseudo-label vector $\mathbf{c} \in \mathcal{C}^K$ for each image. The annotator, trained with ACM-augmented data, can capture fine-grained discriminative features, thus enhancing the model's ability to distinguish subtle differences between adjacent category samples. With the supervision of the generated pseudo-labels, the model can explicitly capture regional features that influence grading labels, aligning with the detailed focus that humans use in decision-making processes.

3.3 Dual-level Fuzzy Learning

A key aspect of ordinal regression is in exploiting the inherent ordinality among category labels, whose definition, however, is often ambiguous and depends on multi-view features of the images. To address this, we propose a dual-level fuzzy learning module that models the feature-label relationship in a fine-grained manner, capturing both patch-wise and channel-wise interactions to quantify the ambiguity in category boundaries, as illustrated in Figure 2(b).

Patch-wise Fuzzification. Based on the fuzzy logic, we first approach the problem from a patch perspective, using fuzzy rules to assess the relationships among patch features at different positions. A set of Gaussian membership functions is adopted to fuzzify the k-th input patch features $\mathbf{H}^{(k)}$, converting it into corresponding membership grades, which can be formulated as:

$$g_{mk}(\mathbf{H}^{(k)}) = e^{-\frac{(\mathbf{H}^{(k)} - \mu_{mk})^2}{\sigma_{mk}^2}}, \tag{1}$$

where $m \in \{1, \ldots, \ell_1\}$ represents the m-th rule with ℓ_1 denoting the total number of rules in the fuzzy system, and $k \in \{1, \ldots, K\}$ represents the k-th patch. In this process, every membership function associates the latent representation for each patch $\mathbf{H}^{(k)}$ with a fuzzy linguistic term label. It employs smooth transitions to characterize features at a finer granularity, effectively reducing the ambiguity associated with labels. Following this, the *AND* fuzzy logic operation is applied across all the membership grades of $\mathbf{H}$ for a given rule, as formulated below:

$$\mathbf{f}_m^1 = \prod_{k=1}^{K} g_{mk}(\mathbf{H}^{(k)}). \tag{2}$$

Thus, for a given latent representation $\mathbf{H} \in \mathbb{R}^{K \times d}$, we can obtain a set of activation strengths for the fuzzy rules, where each rule is computed by aggregating the membership values of different patches $\mathbf{H}^{(k)}$. These two operations decompose the modeling process of label ambiguity. First, the features are fuzzified by calculating their membership to various fine-grained terms. Then the fuzzy feature-label relationships are learned based on the aggregated fuzzy rules. In other words, Equations (1) and (2) together form a nonlinear representation projection, mapping $\mathbf{H} \in \mathbb{R}^{K \times d} \rightarrow \mathbf{f}^{\mathbf{1}} \in \mathbb{R}^{\ell_1 \times d}$. The construction of this transformation space aligns with the fuzzy definitions of labels in ordinal regression at a patch-level. Note that, as part of our DFPG, these operations can be regarded as a fuzzy layer, parameterized by a set of Gaussian membership functions, where the mean μ and variance σ are trainable parameters of the shape $\ell_1 \times K$. Additionally, these two parameters are consistent for patches at the same spatial position across different images, but vary for patches at different positions within the same image. This branch enhances the model's ability to recognize fuzzified features across different patches.

Channel-wise Fuzzification. To capture inter-channel relationships within the image attribute representations, we perform fuzzy learning on $\mathbf{H}$ from an alternative perspective, as illustrated in Figure 2(b). This branch is symmetric to that of the patch-level fuzzification, sharing a similar structure. For simplicity, its detailed formalization is omitted here. The difference lies in its implementation of a nonlinear representation projection mapping $\mathbf{H} \in R^{K \times d} \rightarrow \mathbf{f}^2 \in R^{K \times \ell_2}$, with trainable parameters of the shape $\ell_2 \times d$. Different channels of the same image are assigned to different membership functions (*i.e.*, parameters). This setup is based on the rationale that features within the same channel tend to exhibit similar characteristics in the image representations. The activation strengths of all the fuzzy rules in this branch measure interactions across channels, enabling the model to extract discriminative channel-wise features.

The final ordinal label prediction is derived collaboratively from the outputs of both branches. Moreover, a 1×1 convolution layer and a linear layer serve as classification heads for the patch-level and image-level features, respectively, providing the predicted probabilities.

3.4 Online Co-teaching Strategy

Misclassified patch-level pseudo-labels caused by annotator bias introduce erroneous supervisory information during model training. To address this, we propose a co-teaching strategy to conduct noise-aware patch filtering and provide additional supervision at the patch level, as shown in Figure 2(c).

Different from the prior works on learning with noisy labels, which focused on dealing with noise in real-world data and preventing models from overfitting to noisy labels, our goal is to actively filter patch-level pseudo-labels generated by the patch annotator. Hence, we propose a co-teaching approach [Li *et al.*, 2020], which trains two versions of the DFPG model, F_A and F_B, simultaneously. Each model assigns reliable (retained) and unreliable (masked) patches to the other's training dataset based on the patch loss distribution. Deep networks have been shown to learn simple and generalizable patterns more quickly than noisy patterns [Arpit

et al., 2017]. Thus, training samples with smaller loss values are commonly regarded as clean samples. In the patch filtering method, we divide every training epoch into two steps: Mask Matrix Prediction and Pseudo-label Reflection.

Mask Matrix Prediction. A two-component Gaussian Mixture Model (GMM) is initially employed to model the distribution of the cross-entropy loss $\mathcal{L}_{ce}$ across all the patches. By fitting GMM to $\mathcal{L}_{ce}$, it can cluster the patches into two groups based on their loss values. Thus, the credibility probability w_k of each patch $x^{(k)}$ can be determined by calculating the posterior probability $p(g \mid \mathcal{L}_{ce}^k)$, where g is the Gaussian component with the smaller mean in GMM. Based on the credibility probability w_k and a threshold hyperparameter τ, we construct the patch-level mask matrix $\mathbf{M} \in \{0,1\}^K$ for the input image x, as:

$$\mathbf{M}_k = \begin{cases} 1, & \text{if } w_k \geq \tau, \\ 0, & \text{if } w_k < \tau. \end{cases} \tag{3}$$

Pseudo-label Reflection. Note that directly discarding the masked patch could result in a loss of potentially valuable information and reduce the available ordinal context for learning regional features. Hence, reflection on the pseudo-labels is essential to enhance the model performance. For this, we apply a semi-supervised technique to reprocess the pseudo-labels of both the reliable (retained) and unreliable (masked) patches separately. At each epoch, we train the two models F_A and F_B alternately, keeping one fixed while updating the other. For simplicity, the following description takes model F_A as an example. First, for a reliable patch $x^{(k)}$, its pseudo-label $\mathbf{c}_k$ and the model's new prediction probability p_k^A are linearly combined using the corresponding credibility probability w_k, as:

$$\hat{\mathbf{c}}_k = w_k \mathbf{c}_k + (1 - w_k) p_k^A. \tag{4}$$

In contrast, for each unreliable patch, we leverage the average ensemble of the predictions from both models F_A and F_B as the regenerated pseudo-labels with high confidence:

$$\mathbf{c}'_k = (p_k^A + p_k^B)/2, \;\; \bar{\mathbf{c}}_k = \frac{{\mathbf{c}'_k}^{1/\delta}}{\sum_{k|\mathbf{M}_k=0} {\mathbf{c}'_k}^{1/\delta}}, \tag{5}$$

where δ is a hyperparameter that sharpens the regenerated probability distribution, making it more concentrated.

Optimization. In this study, we tackle ordinal regression under the classification paradigm, which utilizes the traditional Cross-Entropy (CE) objective as the main loss to enhance the classification capacity of the model. It is worth noting that we extend it to be applicable to the reliable pseudo-labels $\hat{\mathbf{c}}$ as an auxiliary loss during the training phase, aiming to leverage regional feature supervision, as:

$$\mathcal{L}_{cls} = \mathrm{CE}\big(p(x), y\big) + \beta \frac{1}{\|\mathbf{M}\|_1} \sum_{k|\mathbf{M}_k=1} \mathrm{CE}\big(p(x^{(k)}), \hat{\mathbf{c}}_k\big), \tag{6}$$

where $x^{(k)}$ indicate the k-th patch in image x.

In addition, for the regenerated pseudo-labels in Equation (5), the Mean Squared Error (MSE) loss is employed:

$$\mathcal{L}_{re} = \frac{1}{\|\mathbf{1}_K - \mathbf{M}\|_1} \sum_{k|\mathbf{M}_k=0} \mathrm{MSE}(p(x^{(k)}), \bar{\mathbf{c}}_k). \tag{7}$$

By combining Equations (6) and (7), we optimize DFPG through the minimization of the following objective function with a weight hyperparameter γ.

$$\mathcal{L} = \mathcal{L}_{cls} + \gamma \mathcal{L}_{re}. \tag{8}$$

4 Experiments

In this section, we conduct extensive experiments on datasets under three different scenarios, to evaluate the effectiveness of our proposed DFPG framework.

4.1 Experimental Setup

Datasets. First, we utilize the Image Adience dataset [Levi and Hassncer, 2015] and the Aesthetics dataset [Schifanella *et al.*, 2021] to evaluate our approach in general scenarios. For the Adience dataset, the 26,580 face images are divided into eight age groups (i.e., the images are labeled from 1 to 8 Similarly, in the Aesthetics dataset, each of the 13,706 images was rated by at least five graders across five ranking categories to assess photographic aesthetic quality. The ground truth for each image is determined as the median rank among all the ratings. In addition, to demonstrate the broad applicability of DFPG, we employ a Diabetic Retinopathy (DR) dataset in the medical grading domain. The task is to classify fundus images into five levels of diabetic retinopathy, ranging from 1 to 5, Note that this dataset is highly imbalanced, with 73.5% of samples labeled as 0 (for no DR). Some statistical information about all the considered datasets is summarized in Table 1. Detailed descriptions and example images of these datasets can be found in the Appendix.

Metrics. Due to the nature of ordinal regression as an intermediate problem between classification and regression, we evaluate our DFPG from two perspectives. First, considering the data imbalance in the selected datasets, we find that relying solely on accuracy as a classification metric is insufficient, as accuracy typically reflects the model's performance on the majority categories while overlooking its effectiveness on the minority categories. Thus, we incorporate accuracy, precision, recall, and F1-score to provide a more comprehensive evaluation of the model's classification performance. Notably, to provide an overall evaluation across all the category labels, we calculate the last three metrics using macro averaging. Second, we employ Mean Absolute Error (MAE) between the predicted and ground truth labels, which directly measures the model's capability to capture ordinal relationships among labels.

Implementation Details. We implement DFPG using PyTorch on an NVIDIA GTX 4090 GPU server. For a fair comparison with existing methods [Cheng *et al.*, 2023; Wang *et al.*, 2023], we use the PVT architecture as our backbone. The default Adam optimizer [Kingma and Ba, 2015] is adopted with a batch size of 24, and training is conducted for 50 epochs per stage. We perform 5-fold cross-validation on the Adience and Aesthetics datasets, and 10-fold cross-validation on the DR dataset, reporting the average results.

Dataset	# of Images	Category labels
Adience	26,580	1-8
Aesthetics	13,706	1-5
DR	35,126	1-5

Table 1: Statistical information of the three evaluation datasets.

4.2 Comparison with Known Methods

To ensure a comprehensive comparison, we re-implement and evaluate several cutting-edge ordinal regression models. Table 2 summarizes the experimental results of all methods across the three datasets, from which we draw the following key observations.

Recent ordinal regression works, such as CIG [Cheng *et al.*, 2023] and Ord2Seq [Wang *et al.*, 2023], demonstrate promising results across various datasets. Specifically, CIG and Ord2Seq emphasize the critical importance of distinguishing adjacent categories, highlighting the necessity of explicitly exploring the ambiguous boundaries between neighboring category labels. CIG employs controllable conditional generation to create artificial images based on a base image and its neighboring category samples, which helps the model learn more accurate and robust decision boundaries. Ord2Seq transforms ordinal labels into binary label sequences, using a dichotomy-based sequence prediction method to differentiate adjacent categories through a progressive refinement scheme. Their work promotes us to leverage more fine-grained patch features to resolve the ambiguity issue.

Our model demonstrates superior performance across the three datasets. More concretely, on the Adience dataset, our model achieves improvements of (+1.00%, +2.25%, +1.23%, +0.72%, 0.0014) in all the metrics compared to baseline models. This shows that our model more effectively addresses the challenge of ambiguity in grading boundaries within ordinal regression. Furthermore, the improvements are more pronounced in the precision, recall, and F1-score metrics computed for individual categories. This indicates that our model can effectively distinguish samples across all categories. Furthermore, our DFPG model also achieves significant improvement on the class-imbalanced datasets, Aesthetics and DR. Note that, 65.6% of the samples in Aesthetics dataset are labeled with class 3 (*i.e.*, ordinary), and 73% of the samples in DR dataset are labeled with class 1 (*i.e.*, no DR). Our model demonstrates performance gains of (+2.07%, +0.71%, 0.0043) and (+3.43%, +1.64%, 0.0092) in terms of Recall, F1-score, and MAE on Aesthetics and DR, respectively. Meanwhile, we attribute the relatively lower performance in Precision and Accuracy on these two datasets to DFPG's enhanced ambiguity modeling ability, which allows it to more effectively learn features of the minority class (see Section 4.3 for details). Besides, we observe that DFPG achieves the most significant improvement on the DR dataset. This highlights the effectiveness of patch-level supervision for grading decisions, aligning with clinical practice where physicians prioritize local lesion assessment for diagnosis.

4.3 Minority Class Classification

A typical classification framework using the Cross-Entropy loss often suffers from frequent "passive updates" of minority classes, resulting in low separability among these classes. Hence, we conduct evaluations for each ordinal class on the DR and Adience datasets to examine the robustness and effectiveness of our model under different data distributions. From the results in Table 2, we compare our model with two state-of-the-art methods, Ord2Seq and CIG, which achieved the best performance in their respective baseline methods. Figure 3 presents the detailed evaluation results for each class with these two methods. The corresponding similar observation on the Aesthetics dataset can be found in the Appendix.

On both datasets, there are typical minority or hard-to-distinguish categories, such as class 2 in the DR dataset and class 4 in the Adience dataset. The performance of the three models (Ord2Seq, CIG-PVT, and our DFPG) shows a notable decline in these categories. For example, in level 2 of the DR dataset, the Recall values of the three models are only (3.69%, 1.64%, and 15.98)%, respectively, which is significantly lower than their average recall on this dataset (50.88%, 55.89%, and 59.32%). This is because the limited number of minority category samples restricts the model's ability to effectively capture distinguishing features. This underscores the importance of leveraging features from adjacent categories to aid in classifying minority categories is a valuable research direction for image ordinal regression.

However, despite a performance decline, our model still outperforms the two counterparts in the minority classes. While CIG-PVT leverages controllable generation techniques to supplement minority class samples to improve its performance on these classes, our model enhances minority class recognition by utilizing diverse information from neighboring class samples through DFL module. Consequently, on the DR dataset, our model achieves a notable improvement of (+12.30%, +12.23%) in recall and F1-score for class 2. This is a substantial improvement, as these metrics for this class in the baseline models are typically around 3% and 5%. We attribute this improvement to two factors. First, the limited samples of class 2 in the DR dataset (only 7%) make it challenging for the model to capture distinctive features. Second, the high proportion of adjacent-level samples (73.5% in level 1) and the ordinality of labels causes the model to favor more prevalent neighboring class. This highlights the superiority of DFPG to distinguish minority category samples by leveraging fine-grained, order-related features from adjacent samples in the membership-based latent space.

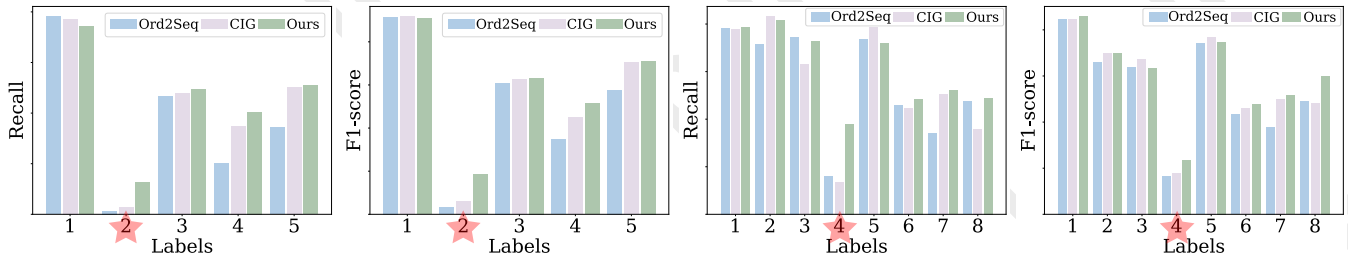
Similarly, our model achieves an improvement of 21.93% in recall and 5.88% in F1-score for class 4 on the Adience dataset. Additionally, across most categories in both datasets, our model shows superior performance. This further demonstrates the robustness of our model, as it consistently enhances performance across various image datasets.

4.4 Ablations

We conduct ablation studies to empirically verify the rationality of DFPG design. We evaluate three main components of the framework, *i.e.*, dual-level fuzzy learning (DFL), patch annotator (PA), and noise-aware patch filtering (NPF). Due to

Dataset	Metric	CNNPOR	SORD	POE	CIG-PVT	Ord2Seq	DFPG (ours)
Adience	Precision	-	0.5430	0.5699	0.5751	0.5804	**0.5904**
	Recall	-	0.5529	0.5636	0.5696	0.5668	**0.5921**
	F1-score	-	0.5363	0.5580	0.5678	0.5603	**0.5801**
	Accuracy	0.5740	0.6097	0.6159	0.6288	0.6244	**0.6360**
	MAE	0.5500	0.4645	0.4713	0.4429	0.4341	**0.4327**
Aesthetics	Precision	-	0.4038	0.4478	**0.4815**	0.4390	0.4512
	Recall	-	0.2773	0.3110	0.3483	0.3484	**0.3691**
	F1-score	-	0.2885	0.3285	0.3763	0.3696	**0.3834**
	Accuracy	0.6748	0.6875	0.6822	**0.6988**	0.6896	0.6966
	MAE	0.3540	0.5248	0.3603	0.3340	0.3230	**0.3187**
Diabetic Retinopathy	Precision	-	0.6025	0.6244	0.6182	**0.6294**	0.6168
	Recall	-	0.4969	0.5248	0.5088	0.5589	**0.5932**
	F1-score	-	0.5241	0.5577	0.5434	0.5844	**0.6008**
	Accuracy	0.8287	0.8034	0.8285	0.8303	0.8310	**0.8339**
	MAE	0.3350	0.2865	0.2557	0.3036	0.2532	**0.2440**

Table 2: Experimental results on the three evaluation datasets. The best and second-best results are marked in **bold** and underlined, respectively. '-' indicates that we could not reproduce the results.



(a) Recall on DR dataset (b) F1-score on DR dataset (c) Recall on Adience dataset (d) F1-score on Adience dataset

Figure 3: Detailed performance for each category on the DR and Adience datasets. We show two evaluation metrics, Recall and F1-score. The star symbol indicates the minority category.

Method			Diabetic Retinopathy		
DFL	PA	NPF	F1-score	Acc	MAE
-	-	-	0.5368	0.8261	0.2636
✓	-	-	0.5808	0.8285	0.2524
✓	✓	-	0.5838	0.8311	0.2485
✓	✓	✓	**0.6008**	**0.8339**	**0.2440**

Table 3: Results of ablation study on the DR dataset.

page limitations, we present an analysis based solely on the metrics in Table 3 for the DR dataset, with analyses of the other two datasets provided in the Appendix.

We consider the PVT network with a linear layer as the fundamental framework, and conduct experiments based on this structure. The results show that our dual-level fuzzy learning (DFL) module is highly effective. By applying fuzzification to the image representations through DFL alone, we improve F1-score, Accuracy, and MAE by (+4.35%, +0.24%, 0.011), respectively. These results indicate that addressing ordinal regression with a traditional classification paradigm is insufficient, and DFL benefits the model in assessing the ordinal label ambiguity. Moreover, with the introduction of additional patch-level supervision, the model achieves improvements of (+0.3%, +0.26%, 0.0112) in the three metrics. In comparison, after filtering the patch-level pseudo-labels, the model achieves improvements of (+2%, +0.54%, 0.0084). This highlights that the filtered patch-level features are effective for grading decisions, while using pseudo-labels from a simple offline annotator yields only limited improvement.

5 Conclusions

In this paper, we presented a novel Dual-level Fuzzy Learning with Patch Guidance (DFPG) framework for image ordinal regression focusing on discriminative patch-level features with only available image-level labels. First, we incorporated the patch annotator and noise-aware filtering paradigm to learn informative patch-level features using only image-level labels. This approach mimics human decision-making by emphasizing discriminative patch-level features for the final prediction. To further explore the ordinal label ambiguity, we developed the dual-level fuzzy learning module that captures ambiguous feature-label relationships via fuzzy rule embeddings from both patch-wise and channel-wise levels. Extensive experimental results on three different image ordinal regression datasets demonstrated the superiority of DFPG compared to state-of-the-art methods. Additionally, we conducted detailed metric evaluations on specific categories to further illustrate the robustness and effectiveness of DFPG.

Acknowledgments

This research was partially supported by National Natural Science Foundation of China under Grant No. 62476246 and No. 92259202, "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C02132, and GuangZhou City's Key R&D Program of China under Grant No. 2024B01J1301.

References

[Arpit *et al.*, 2017] Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 233–242. JMLR.org, 2017.

[Chang *et al.*, 2020] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5709–5718, 2020.

[Chen *et al.*, 2017] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 742–751, 2017.

[Cheng *et al.*, 2023] Yi Cheng, Haochao Ying, Renjun Hu, Jinhong Wang, Wenhao Zheng, Xiao Zhang, Danny Chen, and Jian Wu. Robust image ordinal regression with controllable image generation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Du *et al.*, 2025] Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach clip to develop a number sense for ordinal regression. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 1–17, Cham, 2025. Springer Nature Switzerland.

[Díaz and Marathe, 2019] Raúl Díaz and Amit Marathe. Soft labels for ordinal regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, 2019.

[Frank and Hall, 2001] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, ECML'01, page 145–156, Berlin, Heidelberg, 2001. Springer-Verlag.

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

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

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

[Ikram *et al.*, 2024] Amna Ikram, Azhar Imran, Jianqiang Li, Abdulaziz Alzubaidi, Safa Fahim, Amanullah Yasin, and Hanaa Fathi. A systematic review on fundus image-based diabetic retinopathy detection and grading: Current status and future directions. *IEEE Access*, 12:96273–96303, 2024.

[Jiang *et al.*, 2021] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision Transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18590–18602. Curran Associates, Inc., 2021.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[Kong *et al.*, 2016] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 662–679, Cham, 2016. Springer International Publishing.

[Lee and Kim, 2019] Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1191–1200, 2019.

[Lee and Kim, 2021a] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *International Conference on Learning Representations*, 2021.

[Lee and Kim, 2021b] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *International Conference on Learning Representations*, 2021.

[Levi and Hassncer, 2015] Gil Levi and Tal Hassncer. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015.

[Li *et al.*, 2019] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. BridgeNet: A continuity-aware probabilistic network for age estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[Li *et al.*, 2020] Junnan Li, Richard Socher, and Steven C.H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[Li *et al.*, 2021] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13891–13900, 2021.

[Li *et al.*, 2022] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35313–35325. Curran Associates, Inc., 2022.

[Lim *et al.*, 2020] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *International Conference on Learning Representations*, 2020.

[Martin *et al.*, 2014] Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. ICMR '14, page 447–450, New York, NY, USA, 2014. Association for Computing Machinery.

[Niu *et al.*, 2016] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016.

[Palmer *et al.*, 2013] Stephen E. Palmer, Karen B. Schloss, and Jonathan Sammartino. Visual aesthetics and human preference. *Annual Review of Psychology*, 64(Volume 64, 2013):77–107, 2013.

[Pan *et al.*, 2018] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018.

[Rothe *et al.*, 2015] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep expectation of apparent age from a single image. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 252–257, 2015.

[Schifanella *et al.*, 2021] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):397–406, Aug. 2021.

[Shi and Jain, 2019] Yichun Shi and Anil Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[Shin *et al.*, 2022] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18739–18748, 2022.

[Verma *et al.*, 2019] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR, 09–15 Jun 2019.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021.

[Wang *et al.*, 2023] Jinhong Wang, Yi Cheng, Jintai Chen, TingTing Chen, Danny Chen, and Jian Wu. Ord2Seq: Regarding ordinal regression as label sequence prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5842–5852, 2023.

[Wang *et al.*, 2024] Jinhong Wang, Zhe Xu, Wenhao Zheng, Haochao Ying, Tingting Chen, Zuozhu Liu, Danny Z. Chen, Ke Yao, and Jian Wu. A transformer-based knowledge distillation network for cortical cataract grading. *IEEE Transactions on Medical Imaging*, 43(3):1089–1101, 2024.

[Wen *et al.*, 2020] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *Computer Vision – ECCV 2020: 16th European Conference, Proceedings, Part XXIII*, page 379–395, Berlin, Heidelberg, 2020. Springer-Verlag.

[Wen *et al.*, 2023] Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. Ordinal label distribution learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23424–23434, 2023.

[Yi *et al.*, 2015] Dong Yi, Zhen Lei, and Stan Z. Li. Age estimation by multi-scale convolutional network. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 144–158, Cham, 2015. Springer International Publishing.

[Zheng *et al.*, 2024] Jiyang Zheng, Yu Yao, Bo Han, Dadong Wang, and Tongliang Liu. Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation. In *The Twelfth International Conference on Learning Representations*, 2024.