Enhancing Contrastive Learning for Ordinal Regression via Ordinal Content Preserved Data Augmentation

Jiyang Zheng[1,2] **Yu Yao**[3,4] **Bo Han**[5] **Dadong Wang**[2] **Tongliang Liu**[1]*
[1]Sydney AI Center, The University of Sydney [2]CSIRO, Data61
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]Carnegie Mellon University [5]Hong Kong Baptist University
jzhe5740@uni.sydney.edu.au yu.yao@mbzuai.ac.ae
dadong.wang@data61.csiro.au tongliang.liu@sydney.edu.au

Abstract

Contrastive learning, while highly effective for a lot of tasks, shows limited improvement in ordinal regression. We find that the limitation comes from the predefined strong data augmentations employed in contrastive learning. Intuitively, for ordinal regression datasets, the discriminative information (ordinal content information) contained in instances is subtle. Existing strong augmentations can easily overshadow this ordinal content information. As a result, when contrastive learning is used to extract common features between weakly and strongly augmented images, the derived features often lack this essential ordinal content, rendering them less useful in training models for ordinal regression. To improve contrastive learning's utility for ordinal regression, we propose a novel augmentation method to replace the predefined strong argumentation based on the principle of *minimal change*. Our method is designed in a generative manner that can effectively generate images with different styles but contains desired ordinal content information. Extensive experiments validate the effectiveness of our proposed method, which serves as a plug-and-play solution and consistently improves the performance of existing state-of-the-art methods for ordinal regression.

1 Introduction

Contrastive learning encourages feature consistency between strongly and weakly augmented versions of data (Chen et al., 2020a). This method has demonstrated its effectiveness in both supervised and unsupervised learning contexts (Oord et al., 2018; He et al., 2019; Khosla et al., 2020).

However, the efficacy of contrastive learning diminishes when applied to ordinal regression, which aims to predict ordered categories from instances. A primary reason for this limited success is that contrastive learning does not fully account for the unique characteristics of ordinal regression data. Unlike conventional classification data, instances in ordinal regression datasets are inherently characterized by their subtle ordinal content information, which is essential for predicting ordinal labels. This subtlety presents a significant challenge for contrastive learning to effectively learn discriminative features, primarily due to the predefined strong data augmentations such as color jittering and color dropping (Chen et al., 2020a; Von Kügelgen et al., 2021; Xiao et al., 2021). These augmentations can easily overshadow or diminish the vital ordinal content information.

Take an ordinal regression task of age estimation as an example. While images used in this task may be rich in pixels and colors, age-discriminate features, such as wrinkles or gray hair, are often localized to small regions. This localization renders the age-related content information subtle, making it susceptible to being overshadowed by strong augmentations. As shown in Figure 1, the commonly used strong augmentations can distort or even erase these essential features in ordinal regression data, compromising the performance of contrastive learning in such tasks. Similarly, in

[*] Corresponding Author

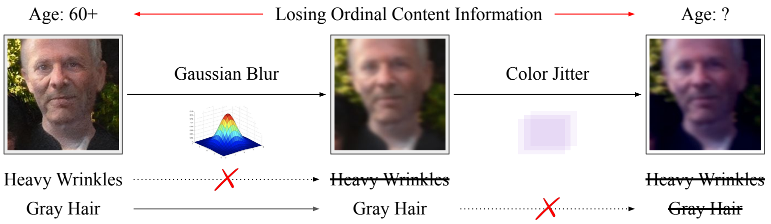


Figure 1: Ordinal content information in data can be easily distorted by standard augmentations in contrastive learning. As illustrated in the example, the age-related features: wrinkles and hair color are eliminated after Gaussian blurring and color jitter, making the age become unidentifiable.

the task of diabetic retinopathy grading, the parts in the retina images that indicate the severity of the condition, such as microaneurysms or hemorrhages (Salz et al., 2016), are often localized and subtle, which can be simply removed by augmentation operations such as Gaussian blurring.

To enhance contrastive learning for ordinal regression, we propose a novel augmentation method as an alternative to predefined strong augmentations. Our method is in a generative manner which can effectively preserve ordinal content information. After training with ordinal datasets, it can be leveraged to generate new instances with the same ordinal content information but different styles. To achieve this, we employ a generative model combined with the principle of *minimal change*.

Specifically, a Generative Adversarial Network (GAN) is employed to generate new instances. However, this generative model does not naturally preserve ordinal content information during the generation process. As indicated by existing work (Hyvarinen et al., 2019; Khemakhem et al., 2020; Von Kügelgen et al., 2021), without any constraints, learned latent factors entangle both ordinal content and non-ordinal information. Such entanglement compromises the controllability of the generative process, making it challenging to produce instances with specific ordinal information. To address this, we introduce the *minimal change* principle (Xie et al., 2022; Kong et al., 2022), which claims that *changes to latent factors should have minimal impact on the generation of instances*.

By leveraging the minimal change, we effectively separate ordinal content factors from non-ordinal ones. In our approach, we partition the latent factors into two sets: $\hat{z}_o$ and $\hat{z}_n$. The former, $\hat{z}_o$, serves as a "container" for ordinal content information. The latent factors in this set are essential for determining the ordinal category and are generated in alignment with ordinal labels. Minimal change is applied to $\hat{z}_o$, thereby limiting its influence during the generation of new instances. As a result, once the generative model attains a minimal reconstruction error, $\hat{z}_o$ solely captures essential ordinal content information, while $\hat{z}_n$ encompasses the remaining non-ordinal information (information that does not contribute to determining ordinal labels). This results in the successful disentanglement of ordinal and non-ordinal information. A more detailed explanation of this disentanglement, along with the generative process introduced by our method, is provided in Section 3.

Our method enhances existing ordinal regression methods by harnessing the benefits of contrastive learning through ordinal content-preserving augmentations. After learning the latent factors $\hat{z}_o$ and $\hat{z}_n$ with the minimal change principle, our method creates augmented instances that belong to a specific ordinal label. To achieve it, we fix $\hat{z}_o$ and randomly sample $\hat{z}_n$, feeding them into our trained generator. This process produces instances with varied styles but consistent ordinal information. These instances can serve as strong augmented views to the natural images. By utilizing these instances with a contrastive learning loss, our method can be seamlessly integrated into existing ordinal regression methods.

The remaining paper is structured as follows: Section 2 provides an overview of the related works in ordinal regression and contrastive learning. Section 3 details our proposed methodology and its practical implementations. Section 4 includes the experimental validation of our approach on different ordinal regression tasks. Section 5 concludes our paper.

2 Related Works

Method for Ordinal Regression. Recent advancements often frame ordinal regression as a classification task (Niu et al., 2016; Liu et al., 2017; Beckham & Pal, 2017). Liu et al. (2018b) introduced a constrained optimization formulation for ordinal regression. This approach minimizes the negative log-likelihood across multiple classes while simultaneously preserving the inherent order relationship between instances. Diaz & Marathe (2019) leveraged the natural ordinal relationships between targets, imparting them as prior knowledge to the model through soft labels. Liu et al. (2019) addressed the task from a probabilistic modeling perspective, where a Gaussian Process model is attached to the output layer of the deep neural network to model uncertainty. Li et al. (2021) proposed a framework that employs probabilistic embeddings to model data uncertainty. Their method enforces a constraint between the learned embedding distributions and pre-defined ordinal distributions, ensuring that the learned embedding space remains ordered. Shin et al. (2022) proposed a regression-based rank estimation algorithm that learns to model the order relationship between instances. Cheng et al. (2023) propose a data fusion approach to address the class-imbalance issue in ordinal regression datasets. While most of the previous studies primarily focused on aligning the model's final predictions with the target, our approach emphasizes the importance of preserving ordinal content information when augmenting ordinal regression data. Our method is orthogonal to end-to-end trainable ordinal regression model, which can serve as a plug-and-play solution to improve the performance of existing state-of-the-art ordinal regression frameworks.

Contrastive Learning and Data Augmentations. Contrastive learning (CL) extracts discriminative information from data by organizing samples into similar and dissimilar pairs. It amplifies the similarity of similar pairs and increases the difference between dissimilar pairs in the feature space. In a self-supervised setting (He et al., 2019; Chen et al., 2020a), similar pairs are generated through a data augmentation module. Given a reference image, this module introduces modifications such as random scaling, cropping, color jittering, blurring, and flipping to generate new perspectives of the image. The original image and its augmented versions constitute a similar pair. In contrast, other images in the batch are treated as dissimilar samples, forming dissimilar pairs. Khosla et al. (2020) expanded CL to a supervised setting. In addition to the augmented versions of the image, samples from the same class also become part of a similar pair. Zha et al. (2022) introduced a supervised CL framework for regression tasks, ensuring that the order of representations in the feature space corresponds to their target values. The results demonstrate that existing regression methods consistently benefit from a CL module for extracting discriminative features from data.

Furthermore, intensive data augmentations have been found crucial for the success of the contrastive learning framework across all settings (Chen et al., 2020a;b; Khosla et al., 2020; Li et al., 2023; 2022a; Huang et al., 2021; Zheng et al., 2022). However, such aggressive augmentations can compromise an image's content. Xiao et al. (2021) addressed this breach of the invariance assumption (i.e., data augmentations altering the data's semantic information) by decomposing a compound series of augmentations into individual operations and creating distinct heads for each single augmentation. This method is effective when data is sensitive to a few specific augmentations within the full augmentation sequence but maintains its invariance assumption with others. Given that content information in ordinal data is particularly susceptible, many augmentation techniques can potentially distort an image's semantics, thereby limiting the method's efficacy. Compared to their approach, our method does not depend on any predetermined augmentation method. Instead, we guide the model to discern which aspects of the augmented data should be preserved as content variables and which can be modified to introduce diversity as style variables.

3 Ordinal Content Preserving Contrastive Learning (OCP-CL)

In this section, we introduce our Ordinal Content Preserving Contrastive Learning (OCP-CL) framework. Specifically, to improve the utility of contrastive learning for ordinal regression, we propose a novel ordinal content-preserving augmentation method that replaces the predefined strong augmentations in a contrastive learning framework. First, we explain the generative process of ordinal regression data, which serves as the foundational understanding of our proposed generative model. Then, we introduce our approach for disentangling ordinal content and non-ordinal factors via minimal change, and detail the implementation of the generative model. Next, we describe the process of generating content-preserving data augmentations through interventions on non-ordinal latent factors. Finally, we present the contrastive learning formulation with our generated augmentations from the original instances. The contrastive learning objective can be integrated into any existing end-to-end trainable deep ordinal regression methods to form a joint objective.

Data Generative Process. We first explain the causal data generative process (Glymour & Zhang, 2019; Yao et al., 2023) as illustrated in Figure 2. The graph outlines the generative process for ordinal regression data, segregating latent factors into different functional groups based on their relationships to the observed variables.

Specifically, $\boldsymbol{z}_{\mathrm{v}}$ denotes a set of invariant ordinal factors that capture all ordinal content relevant features across different ordinal categories. For example, in age estimation, $\boldsymbol{z}_{\mathrm{v}}$ encompasses a comprehensive collection of age-specific attributes such as the severity of wrinkles or variations in skin texture across age groups. The ordinal label y serves as a constraining variable, ensuring that $\boldsymbol{z}_{\mathrm{o}}$ selectively retains only those features from $\boldsymbol{z}_{\mathrm{v}}$ that pertain to its corresponding ordinal category. Similarly, $\boldsymbol{z}_{\mathrm{n}}$ denotes the set of non-ordinal, style-related factors. Together, $\boldsymbol{z}_{\mathrm{o}}$ and $\boldsymbol{z}_{\mathrm{n}}$ collaboratively generate x, the observed data instance. Our primary objective for the generative model is to disentangle the learned latent factors $\hat{\boldsymbol{z}}$ into $\hat{\boldsymbol{z}}_{\mathrm{o}}$ and $\hat{\boldsymbol{z}}_{\mathrm{n}}$, in alignment with the proposed SCM. In this setup, $\hat{\boldsymbol{z}}_{\mathrm{o}}$ is approximated to only contain ordinal content information that determines the ordinal category, whereas $\hat{\boldsymbol{z}}_{\mathrm{n}}$ holds styling information. Generating content-preserving samples hinges on accurately recovering the true joint distribution of image and class, denoted as $P(\boldsymbol{X}, \boldsymbol{Y})$. Based on our proposed generative process, the causal factorization of the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}_{\mathrm{o}}, \boldsymbol{Z}_{\mathrm{v}}, \boldsymbol{Z}_{\mathrm{n}})$ is:

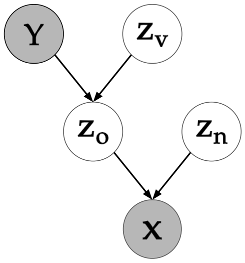


Figure 2: The data generative process employed by our method. The shaded variables are observable and the unshaded variables are latent.

$$P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}_{\mathrm{o}}, \boldsymbol{Z}_{\mathrm{v}}, \boldsymbol{Z}_{\mathrm{n}}) = P(\boldsymbol{Y})P(\boldsymbol{Z}_{\mathrm{v}})P(\boldsymbol{Z}_{\mathrm{n}})P(\boldsymbol{Z}_{\mathrm{o}}|\boldsymbol{Y}, \boldsymbol{Z}_{\mathrm{v}})P(\boldsymbol{X}|\boldsymbol{Z}_{\mathrm{o}}, \boldsymbol{Z}_{\mathrm{n}}). \tag{1}$$

Our method is designed to fulfill the this generative process by modelling each probability in Eq. 1.

3.1 Philosophy of Disentangling Ordinal Content Factors.

The principle of minimal change is pivotal in our generative process, aiding in the effective disentanglement of non-ordinal factors from ordinal content factors. To underscore the importance of this principle, we first discuss the limitations of relying solely on ordinal labels y. Subsequently, we explore scenarios that incorporate the *minimal change* principle. We consider a generative model (Zhou et al., 2023; Yao et al., 2021) that aligns with the generative process depicted in Figure 2 and infers the factors $\hat{\boldsymbol{z}}_{\mathrm{n}}$, $\hat{\boldsymbol{z}}_{\mathrm{o}}$, and $\hat{\boldsymbol{z}}_{\mathrm{v}}$. After the learning phase, in which the reconstruction error is minimized, our goal is to align $\hat{\boldsymbol{z}}_{\mathrm{n}}$ with non-ordinal factors $\boldsymbol{z}_{\mathrm{n}}$; align $\hat{\boldsymbol{z}}_{\mathrm{o}}$ with ordinal content factors $\boldsymbol{z}_{\mathrm{o}}$; and align $\hat{\boldsymbol{z}}_{\mathrm{v}}$ with the set of invariant ordinal factors $\boldsymbol{z}_{\mathrm{v}}$. Note that $\boldsymbol{z}_{\mathrm{n}}$, $\boldsymbol{z}_{\mathrm{o}}$ and $\boldsymbol{z}_{\mathrm{v}}$ are the true latent factors in the data generative process.

If we rely solely on ordinal labels y, the generation of $\hat{\boldsymbol{z}}_{\mathrm{o}}$ is influenced by both y and $\hat{\boldsymbol{z}}_{\mathrm{v}}$. This relationship can be mathematically expressed as: $\hat{\boldsymbol{z}}_{\mathrm{o}} = g(y, \hat{\boldsymbol{z}}_{\mathrm{v}}) + \epsilon$. In essence, this generation mechanism allows $\hat{\boldsymbol{z}}_{\mathrm{o}}$ to assimilate information from both y and $\hat{\boldsymbol{z}}_{\mathrm{v}}$. Given the generative model's typical assumption that both $\hat{\boldsymbol{z}}_{\mathrm{n}}$ and $\hat{\boldsymbol{z}}_{\mathrm{v}}$ follow a standard Gaussian distribution, non-ordinal information can feasibly reside in either $\hat{\boldsymbol{z}}_{\mathrm{v}}$ or $\hat{\boldsymbol{z}}_{\mathrm{n}}$. These factors are fundamentally similar, and their differences are purely notational. Hence, the choice of where to store non-ordinal information does not influence the reconstruction error. As a result, $\hat{\boldsymbol{z}}_{\mathrm{o}}$, being influenced by both y and $\hat{\boldsymbol{z}}_{\mathrm{v}}$, inevitably contains non-ordinal information. Thus, the disentanglement can not be achieved solely with y.

To address this, we apply the minimal change principle in the disentanglement process. The minimal change principle serves as a constraint on image generation, ensuring that the influence of certain factors during instance generation remains minimal. Specifically, when applied to $\hat{\boldsymbol{z}}_{\mathrm{o}}$, this principle limits the factor's influence. Consequently, the information within $\hat{\boldsymbol{z}}_{\mathrm{o}}$ remains minimal and focused. By introducing an additional constraint on the generative function g, we ensure that $\hat{\boldsymbol{z}}_{\mathrm{v}}$ inherently carries information related to Y. Assuming the reconstruction error is minimized, this suggests that ordinal content information is primarily included within the latent factors $\hat{\boldsymbol{z}}_{\mathrm{o}}$. As $\hat{\boldsymbol{z}}_{\mathrm{n}}$ is generated independently of the ordinal label Y, it will not contain ordinal content information. Therefore, all ordinal content information becomes localized within $\hat{\boldsymbol{z}}_{\mathrm{o}}$. Furthermore, by minimizing the influence of $\hat{\boldsymbol{z}}_{\mathrm{o}}$, it becomes exclusively representative of ordinal content information.

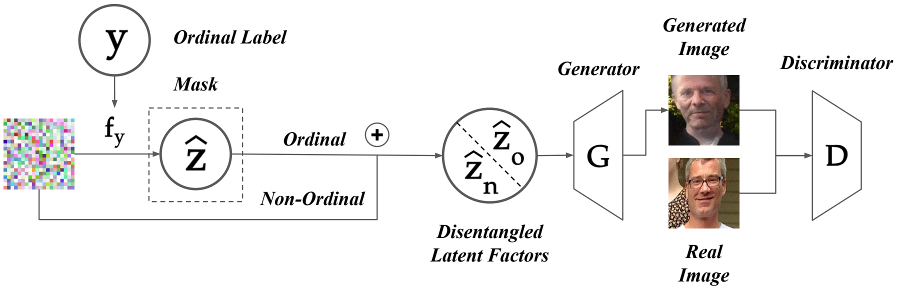


Figure 3: **Architecture of Our Generative Model.** The class label $\boldsymbol{y}$ is leveraged for both disentangling latent factors and enforcing minimal change. An ordinal head is appended to the discriminator to preserve the ordinal distribution of generated samples in relation to their class.

3.2 Ordinal Content and Non-Ordinal Information Disentanglement via Minimal Change

Our primary objective is to achieve *minimal change* in the generation process, specifically by constraining the influence of ordinal content factors $\hat{\boldsymbol{z}}_{\text{o}}$. This is realized by limiting the number of these factors. To this end, we introduce a mask operation, consistent with our data generative process. The essence of this mask is to regulate the quantity of ordinal content factors. A sparser mask translates to fewer ordinal content factors. To promote this sparsity, we impose an L1 loss on the mask, represented as $\mathcal{L}_{\text{sp}} = \|M\|_1$. Let's define our latent factors as $\hat{\boldsymbol{z}} := [\hat{z}_o, \hat{z}_n]$ and $\hat{\boldsymbol{z}}_{on} := [\hat{z}_v, \hat{z}_n]$. The mask operation is then given by:

$$\hat{\boldsymbol{z}}_{\boldsymbol{on}} = \hat{\boldsymbol{z}} + \boldsymbol{M} \odot f_{\boldsymbol{y}}(\hat{\boldsymbol{z}}), \;\; \hat{z} \sim \mathcal{N}(0, \mathbf{I}), \;\; y \sim P(Y). \tag{2}$$

In the above equation, $\hat{\boldsymbol{z}}$ encapsulates both the invariant ordinal content factors $\hat{\boldsymbol{z}}_{\text{v}}$ and the non-ordinal factors $\hat{\boldsymbol{z}}_{\text{n}}$. The label distribution $P(Y)$ is derived empirically by counting the occurrences of each label within the dataset and then normalizing these counts to form a probability distribution. We utilize the Deep Sigmoidal Flow (Huang et al., 2018) for the function $f_{\boldsymbol{y}}$. It is a type of normalization flow characterized by the use of small neural networks with sigmoid units. These units introduce inflection points in the transformation function, enabling the modeling of complex probability distributions. Within our context, it serves as a component-wise transformation function that transforms $\hat{z}_{on}$ in a component-wise manner. This function, when applied, acts as a label influence. Although it impacts all elements in $\hat{\boldsymbol{z}}$, each element undergoes an independent transformation. The mask operation on $\hat{\boldsymbol{z}}$ rejects certain transformed elements. By adding $\hat{\boldsymbol{z}}$ back, we ensure that certain elements remain uninfluenced by the label Y, effectively distinguishing them as non-ordinal factors.

It's worth noting that for elements unaffected by the mask but influenced by y, the addition of elements from $\hat{\boldsymbol{z}}$ is inconsequential. Given that $\hat{\boldsymbol{z}}$ is sampled from a high-dimensional Gaussian distribution, adding it to these elements is akin to introducing random Gaussian noise, ensuring our method remains consistent with the proposed data generative process.

To make latent factors, we employ a Generative Adversarial Network (GAN) model (Mirza & Osindero, 2014). The architecture of this model is illustrated in Figure 3. The GAN comprises two main components: a generator G_θ and a discriminator D_ϕ, each parameterized by their respective learnable parameters θ and ϕ. The generator's role is to craft realistic instances, while the discriminator endeavors to differentiate between genuine and generated instances. The GAN loss, vital for image reconstruction, is formally articulated as:

$$\mathcal{L}_{\text{gan}} = \mathbb{E}[\log(D_\phi(\boldsymbol{x}))] + \mathbb{E}[\log(1 - D_\phi(G_\theta(\hat{\boldsymbol{z}}_{\boldsymbol{on}})))]. \tag{3}$$

In the given formulation, $D_\phi(\boldsymbol{x})$ represents the discriminator's estimated probability that the instance $\boldsymbol{x}$ is sampled from the real data distribution. The generator, denoted by G_θ, aims to produce instances that the discriminator D_ϕ perceives as real, maximizing the likelihood of them being classified as genuine. Conversely, the discriminator D_ϕ endeavors to distinguish real instances from those generated by G_θ, classifying them accurately as either real or fake.

The objective function for disentangling ordinal content and non-ordinal information is combined with the GAN loss and the sparsity loss, i.e.,

$$\arg \min_{\{\phi,\theta,f_y,\boldsymbol{M}\}} \mathcal{L}_{\text{Aug}} = \arg \min_{\{\phi,\theta,f_y,\boldsymbol{M}\}} \mathcal{L}_{\text{gan}} + \lambda \cdot \mathcal{L}_{\text{sp}}. \tag{4}$$

where λ is the coefficient to control the contribution of the sparsity loss to the overall objective.

3.3 Content-Preserving Augmentation for Ordinal Regression

Our method is crafted to complement existing ordinal regression techniques, leveraging the strengths of contrastive learning. After training, we could have a generator $G_{\hat{\theta}}$. To generate an instance x' corresponding to a specific ordinal label, we employ Eq. 2. For instance, generating an example for the ordinal label $Y = 1$ can be achieved by:

$$\boldsymbol{x'_i} = G_{\hat{\theta}}(\hat{z}_{on}), \;\; \hat{z}_{on} = \hat{\boldsymbol{z}} + \boldsymbol{M} \odot f_{\boldsymbol{1}}(\hat{\boldsymbol{z}}), \;\; \hat{z} \sim \mathcal{N}(0, \mathbf{I}), \;\; Y = 1. \tag{5}$$

Specifically, we start by sampling $\hat{z}_{on}$. By setting the ordinal label $Y = 1$, we utilize $f_{\boldsymbol{1}}(\hat{\boldsymbol{z}})$ to generate $\hat{z}_{on}$ specific to the ordinal label $Y = 1$. By sampling different $\hat{z}_{on}$ values and maintaining the ordinal label $Y = 1$ constant, we can generate diverse instances that, while exhibiting stylistic variations, consistently belong to the label Y.

To integrate with existing ordinal regression methods, x' can be employed as the strongly augmented data. We then incorporate the supervised contrastive loss (Khosla et al., 2020) as a regularization term for the prevailing method. This loss emphasizes intra-class similarities while concurrently maximizing inter-class disparities. Let's define $\mathbf{X}$ as the feature space of x. We introduce h_ψ as a model with learnable parameters ψ, which is used by the ordinal regression method. For a given instance x, $h_\psi(x) = z$ outputs the latent representation z used for ordinal regression. The contrastive loss on h_ψ is defined as:

$$\mathcal{L}^{\text{con}} = \sum_{i \in I} \mathcal{L}_i^{\text{con}} = -\sum_{i \in I} \frac{1}{|\mathcal{S}(h_\psi(\boldsymbol{x_i}))|} \sum_{h_\psi(\boldsymbol{x_s}) \in \mathcal{S}(h_\psi(\boldsymbol{x_i}))} \log \frac{\exp\left(h_\psi(\boldsymbol{x_i}) \cdot h_\psi(\boldsymbol{x'_i})/\tau\right)}{\sum_{b \in B} \exp\left(h_\psi(\boldsymbol{x_i}) \cdot h_\psi(\boldsymbol{x_b})/\tau\right)}. \tag{6}$$

In the equation above, I denotes the set of sample indices in the batch. For each instance $\boldsymbol{x_i}$ belonging to y, its strongly augmented counterpart x'_i is generated using our method by setting the ordinal label $Y = y$ during the generation process.

4 Experiments

In this section, we evaluate our method across three real-world applications within the domain of ordinal regression: age estimation, diabetic retinopathy rating, and weather condition prediction. Due to space constraints, we include qualitative analyses of our generative model in Appendix D.

Baselines. We employ five state-of-the-art deep learning-based ordinal regression methods as our baselines. OR-CNN (Niu et al., 2016) utilizes a series of binary classifiers and optimizes the model through the one-hot encoding of labels. CNNPOR (Liu et al., 2018b) reduces the multi-class negative log-likelihood while concurrently maintaining the intrinsic ordinal relationship among instances. SORD (Diaz & Marathe, 2019) employs a soft labeling strategy during training. POE (Li et al., 2021) captures data uncertainty via probabilistic embeddings. MWR (Shin et al., 2022) leverages an auxiliary set of reference images to model ordinal relationships. All these models are end-to-end trainable. We seamlessly integrate our contrastive learning objective into their original loss formulations, augmented with ordinal content-preserving data transformations.

Experimental Settings. For the generative model, we use StyleGAN2 (Karras et al., 2020) as the base model, λ_1 is set to 1e-4 across all settings. For all ordinal regression methods, we use VGG16 (Simonyan & Zisserman, 2014) the base deep neural network architecture, with ImageNet (Deng et al., 2009) pre-trained weight for initialization. We employ an embedding layer before the final output layer in the model to extract feature embeddings. The dimension of feature embedding is set to 128. The ratio of contrastive loss is consistently set to 1e-4 for OR-CNN, CNNPOR and POE, and 1e-5 for SORE and MWR. For the three datasets, the input images are resized

Adience	w/o OCP-CL		w/ OCP-CL	
	Accuracy (↑)	MAE (↓)	Accuracy (↑)	MAE (↓)
OR-CNN (Niu et al., 2016)	54.6 ± 5.5	0.60 ± 0.09	57.1 ± 5.1 *(+2.5)*	0.56 ± 0.06 *(+0.04)*
CNNPOR (Liu et al., 2018b)	55.1 ± 6.0	0.60 ± 0.08	57.7 ± 4.2 *(+2.6)*	0.55 ± 0.07 *(+0.05)*
SORD (Diaz & Marathe, 2019)	57.8 ± 4.9	0.53 ± 0.06	59.9 ± 5.0 *(+2.1)*	0.49 ± 0.06 *(+0.04)*
POE (Li et al., 2021)	60.5 ± 4.8	0.47 ± 0.08	63.7 ± 4.6 *(+3.2)*	0.43 ± 0.07 *(+0.04)*
MWR (Shin et al., 2022)	62.6 ± 5.0	0.45 ± 0.08	63.6 ± 4.7 *(+1.0)*	0.43 ± 0.07 *(+0.02)*

Diabetic Retinopathy	w/o OCP-CL		w/ OCP-CL	
	Accuracy (↑)	MAE (↓)	Accuracy (↑)	MAE (↓)
OR-CNN (Niu et al., 2016)	71.9 ± 1.3	0.42 ± 0.01	72.8 ± 0.7 *(+0.9)*	0.41 ± 0.00 *(+0.01)*
CNNPOR (Liu et al., 2018b)	71.3 ± 1.1	0.42 ± 0.02	72.6 ± 1.0 *(+1.3)*	0.41 ± 0.01 *(+0.01)*
SORD (Diaz & Marathe, 2019)	69.1 ± 1.0	0.45 ± 0.01	69.9 ± 1.1 *(+0.8)*	0.44 ± 0.01 *(+0.01)*
POE (Li et al., 2021)	73.6 ± 1.0	0.40 ± 0.01	74.8 ± 0.8 *(+1.2)*	0.38 ± 0.00 *(+0.02)*
MWR (Shin et al., 2022)	74.5 ± 1.1	0.38 ± 0.02	75.1 ± 1.1 *(+0.6)*	0.37 ± 0.01 *(+0.01)*

SkyFinder	w/o OCP-CL		w/ OCP-CL	
	Accuracy (↑)	MAE (↓)	Accuracy (↑)	MAE (↓)
OR-CNN (Niu et al., 2016)	60.3 ± 2.1	0.48 ± 0.03	62.1 ± 2.3 *(+1.8)*	0.46 ± 0.04 *(+0.02)*
CNNPOR (Liu et al., 2018b)	57.6 ± 1.6	0.52 ± 0.03	59.7 ± 1.5 *(+2.1)*	0.49 ± 0.03 *(+0.03)*
SORD (Diaz & Marathe, 2019)	58.2 ± 1.9	0.51 ± 0.06	60.5 ± 2.0 *(+2.3)*	0.48 ± 0.04 *(+0.03)*
POE (Li et al., 2021)	61.9 ± 1.7	0.46 ± 0.05	64.1 ± 1.6 *(+2.2)*	0.42 ± 0.05 *(+0.04)*
MWR (Shin et al., 2022)	62.4 ± 1.8	0.45 ± 0.05	63.2 ± 1.9 *(+0.8)*	0.44 ± 0.06 *(+0.01)*

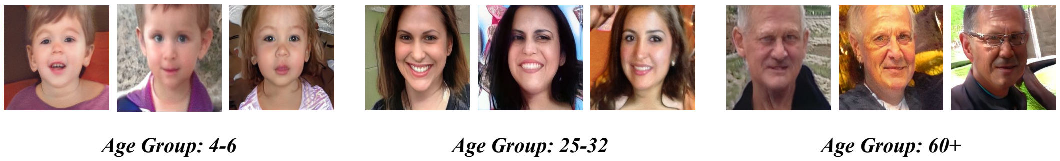
Table 1: Accuracy (%) and MAE comparison on Adience dataset (Eidinger et al., 2014), Diabetic Retinopathy dataset (Liu et al., 2018a) and SkyFinder dataset (Mihail et al., 2016).

into 256 × 256 and center cropped into a sub-region of 224 × 224. Adam (Kingma & Ba, 2014) optimizer is used for all baseline methods, with a base learning rate of 1e-4. We uniformly train all baseline models for 200 epochs with a batch size of 256 for all baselines except MWR. For MWR, the batch size is set to 128 due to memory constraints. We report the results via the accuracy and mean absolute error (MAE) metrics. For the other parameters in the baselines, we adhere to the original settings designed in the papers unless specified in our experimental settings. While we employ our techniques for strong augmentations, weak augmentations are achieved solely through resizing, center cropping, and normalizing the original instances. No other augmentation methods are applied to the data. All experiments are conducted in on two 48GB NVIDIA RTX A6000 GPUs.

4.1 Age Estimation

Dataset. Age estimation is the task of predicting age groups based on facial images. The Adience dataset (Eidinger et al., 2014) comprises 26,580 photos from Flickr, featuring 2,284 subjects. These photos are annotated across eight age groups: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60 years. The dataset adheres to a standard five-fold, subject-exclusive cross-validation protocol, as widely utilized in previous studies (Rothe et al., 2018; Shen et al., 2018; Li et al., 2019; 2021). The generative model is trained in accordance with the training fold of this protocol. For each instance in the dataset, we generate 3 augmented views using the generative model, the augmented views and the original instances are jointly trained by the models.

Results. We present the experimental results in Table 1 (Top). Employing our proposed OCP-CL method, OR-CNN experiences a 4.58% boost in accuracy and a 6.67% reduction in MAE. CNNPOR benefits from a 4.72% increase in accuracy and an 8.33% improvement in MAE. SORD's performance is uplifted by 3.63% in accuracy and 7.55% in MAE. POE sees the largest accuracy improvement of 5.29% and an MAE reduction of 8.51%. Lastly, MWR has a modest 1.6% increase in accuracy and a 4.44% decrease in MAE. This consistent improvement across multiple ordinal regression methods validate the efficacy of our OCP-CL approach for the task of age estimation. Additionally, we visualise the generative augmentations in Figure 1. We observe that the augmentations have preserved the ordinal content information for their respective age groups, capturing details such as the sparse eyebrows of children, silky skin texture of young adults, and the pronounced wrinkles of seniors, thereby allowing the ordinal regression methods and the contrastive learning framework to effectively learn the critical ordinal content information.


Age Group: 4-6

Age Group: 25-32

Age Group: 60+

Figure 4: Generated augmentations for the age estimate task, the collections corresponds to the age group of (4-6), (25-32), and 60+ respectively.

4.2 Diabetic Retinopathy Rating

Dataset. The Diabetic Retinopathy dataset[1] is utilized for predicting the severity stages of Diabetic Retinopathy based on high-resolution RGB retina images. The dataset consists of 35,126 individual instances, each annotated into one of five ordinal categories representing increasing levels of severity (*i.e.*, *No DR*, *Mild*, *Moderate*, *Severe*, and *Proliferative DR*). The dataset is partitioned into training, validation, and testing sets, which constitute 80%, 5%, and 15% of the total dataset, respectively. The dataset contains 25,810, 2443, 5292, 873 and 708 images for each category, respectively. Account for the imbalances between adjacent categories, we generate augmented views dynamically depending on the ground truth label, which mitigates the class imbalance issue. Specifically, the number of augmentations for instances from each increasing level of severity is set as [1, 3, 2, 5, 5].

Results. In Table 1 (Middle), we evaluate the performance of various ordinal regression methods on the Diabetic Retinopathy dataset, with and without the incorporation of our proposed OCP-CL (Ordinal Content-Preserving Contrastive Learning) module. Remarkably, all the compared methods exhibit improvement in both accuracy and MAE upon integration with the OCP-CL module. Practicularly, the incorporation of the OCP-CL module results in accuracy improvements of 1.25%, 1.82%, 1.16%, 1.63%, and 0.81% for OR-CNN, CNNPOR, SORD, POE, and MWR, respectively. Concurrently, the MAE reduces by 2.38%, 2.38%, 2.22%, 5.00%, and 2.63%, respectively. These results collectively indicate that the introduction of the OCP-CL module consistently enhances the performance across a diverse set of ordinal regression models. This validates the generalizability and efficacy of our proposed OCP-CL approach in boosting performance for ordinal regression tasks.

4.3 Weather Condition Prediction

Dataset. The SkyFinder Dataset (Mihail et al., 2016) comprises 94,804 labeled outdoor images, sourced from 53 static webcams affiliated with the Archive of Many Outdoor Scenes (AMOS). These images encapsulate a broad spectrum of weather and lighting conditions. A specialized subset of 62,988 images, specifically featuring the weather conditions of *Clear*, *Partly Cloudy*, and *Mostly Cloudy*, has been curated to create a weather prediction dataset. This subset is further partitioned into training, validation, and testing sets, constituting 80%, 5%, and 15% of the dataset, respectively. For each instance in the dataset, we generate 3 augmented views using the generative model, the augmented views and the original instances are jointly trained by the models.

Results. Table 1 (Bottom) presents the results of our experiments, highlighting the performance improvements achieved by all baseline models upon the incorporation of the contrastive module. Specifically, the accuracy improvements for the baselines are 2.9%, 3.5%, 3.8%, 3.4%, and 1.1% for OR-CNN, CNNPOR, SORD, POE, and MWR, respectively. Similarly, the improvements in MAE for the baselines are 4.2%, 5.8%, 5.9%, 8.7%, and 2.2%, respectively. With an average improvement of 2.94% in accuracy and 5.42% in MAE, these results demonstrate the efficacy of our method in enhancing the performance of deep-learning-based ordinal regression models on the weather condition estimation task.

4.4 Analysis

Transfer Learning. In this section, we evaluate the transfer learning performance of our contrastive learning approach. Initially, we pre-train the encoder using a contrastive learning objective

[1] Accessible from https://www.kaggle.com/competitions/diabetic-retinopathy-detection

Dataset	SupMoCo (He et al., 2019)		SupCon (Khosla et al., 2020)		S-LooC (Xiao et al., 2021)		SupCReg (Zha et al., 2022)		OCP-CL (Ours)	
	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy (↑)	MAE (↓)
DR	63.6	0.52	60.5	0.55	63.0	0.52	62.7	0.53	**65.2**	**0.50**
Adience	51.9	0.65	51.7	0.67	52.4	0.64	51.2	0.63	**53.1**	**0.61**
SkyFinder	56.1	0.55	53.9	0.59	55.6	0.56	55.1	0.55	**57.5**	**0.52**

Table 2: Linear evaluation on supervised contrastive learning frameworks. Accuracy (%) and MAE are reported for various ordinal datasets including Diabetic Retinopathy dataset, Adience (Levi & Hassner, 2015) and SkyFinder dataset (Mihail et al., 2016).

for feature extraction. Following this, we freeze the trained encoder and employ the extracted features as input to a single-layer MLP predictor, which is then fine-tuned on the training data. We assess the efficacy of our approach against recent state-of-the-art supervised contrastive learning frameworks across three different tasks. To mitigate performance degradation due to parameter settings, we dynamically adopt the recommended configurations from the original papers. However, for SupCon (Khosla et al., 2020), a batch size of 1024 is unfeasible for image instances of size 224 by 224. To ensure convergence, we reduced the image size to 64 by 64 and the batch size to 512. The results are presented in Table 2. Notably, our method significantly outperforms all existing approaches in the task of ordinal regression. The benefits of ordinal content-preserving data augmentation become evident when benchmarked against SupMoCo (He et al., 2019). Specifically, we adopt the SupMoCo framework as the baseline contrastive learning framework and integrate our augmentation strategy by replacing the original data augmentation modules. This aids in evaluating transfer learning performance. The contrastive loss formulation in SupMoCo aligns with our Eq. 6, and an additional momentum encoder is incorporated to ensure training convergence. The MoCo strategy is not employed in other experiments.

Minimal Change in Image Generation. We study the effect of minimal change on ordinal data generation by manipulating the mask hyperparameter. When this hyperparameter is set to zero, sparsity is not enforced, effectively removing the minimal change constraint from the generative process. As illustrated in Figure 5 ($\lambda_{\text{mask}} = 0 \rightarrow$ no minimal change), the absence of minimal change leads to the inclusion of non-ordinal features, such as hair, which do not contribute to identifying infant age groups and should be considered non-ordinal factors. However, in the case where minimal change is not applied, these features are learned as ordinal content factors and appear in all generated images, thereby demonstrating poor disentanglement performance. By enforcing minimal change, these non-ordinal elements are suppressed, enhancing the overall quality of the generated instances.

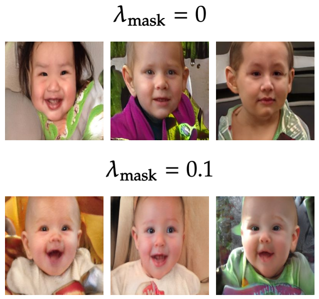


Figure 5: Influence of *minimal change* in image generation.

5 Conclusion

In this paper, we address the open challenge of applying contrastive learning to ordinal regression tasks. We find that the strong data augmentations in the contrastive learning frameworks often diminish the intrinsic discriminative semantic information associated with ordinal labels. Consequently, when contrastive learning is used to identify invariant features between weakly and strongly augmented views, the extracted features frequently lack the essential ordinal content information. To mitigate this issue, we introduce a novel augmentation method grounded in the principle of *minimal change*. This generative approach ensures that the images retain the essential ordinal content information during the data augmentation process. As a result, our method enhances the applicability of contrastive learning to ordinal regression tasks. Extensive experiments validate the efficacy of this approach in improving the performance of existing ordinal regression models. This work not only broadens the scope of contrastive learning in ordinal regression but also provides valuable insights for future research aimed at preserving crucial task-specific information during data augmentation.

6 ACKNOWLEDGEMENT

The authors would like to thank Muyang Li and Xiaobo Xia for their valuable feedback throughout the project. Jiyang Zheng is supported by the CSIRO Next Generation Graduates and AI for Missions PhD program. Tongliang Liu is partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031. BH was supported by the NSFC General Program No. 62376235, Guangdong Basic and Applied Basic Research Foundation Nos. 2022A1515011652, 2024A1515012399, HKBU Faculty Niche Research Areas No. RC-FNRA-IG/22-23/SCI/04, and HKBU CSD Departmental Incentive Scheme.

REFERENCES

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Christopher Beckham and Christopher Pal. Unimodal probability distributions for deep ordinal classification. In *International Conference on Machine Learning*, pp. 411–419. PMLR, 2017.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

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. *arXiv preprint arXiv:2305.04213*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4738–4747, 2019.

Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

Clark Glymour and Kun Zhang. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:418407, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FYKVPOHCpE.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pp. 2078–2087. PMLR, 2018.

Zhuo Huang, Chao Xue, Bo Han, Jian Yang, and Chen Gong. Universal semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:26714–26725, 2021.

Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. *arXiv preprint arXiv:2207.03162*, 2022.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*, pp. 11455–11472. PMLR, 2022.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–42, 2015.

Muyang Li, Runze Wu, Haoyu Liu, Jun Yu, Xun Yang, Bo Han, and Tongliang Liu. Instant: Semi-supervised learning with instance-dependent thresholds. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022a.

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*, pp. 1145–1154, 2019.

Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13896–13905, 2021.

Yewen Li, Chaojie Wang, Xiaobo Xia, Tongliang Liu, Bo An, et al. Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae. *Advances in Neural Information Processing Systems*, 35:7383–7396, 2022b.

Yexiong Lin, Yu Yao, Xiaolong Shi, Mingming Gong, Xu Shen, Dong Xu, and Tongliang Liu. Cs-isolate: Extracting hard confident examples by content and style isolation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zinan Lin, Kiran Koshy Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *arXiv preprint arXiv:1906.06034*, pp. 60, 2019.

Xiaofeng Liu, Yang Zou, Yuhang Song, Chao Yang, Jane You, and BV K Vijaya Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018a.

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

Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 831–839, 2018b.

Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5301–5309, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Radu Paul Mihail, Scott Workman, Zach Bessinger, and Nathan Jacobs. Sky segmentation in the wild: An empirical study. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–6, 2016. doi: 10.1109/WACV.2016.7477637.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4920–4928, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

D. Salz, Talisa E de Carlo, M. Adhi, Eric M. Moult, WhooJhon Choi, C. Baumal, A. Witkin, J. Duker, J. Fujimoto, and N. Waheed. Select features of diabetic retinopathy on swept-source optical coherence tomographic angiography compared with fluorescein angiography and normal eyes. *JAMA Ophthalmology*, 2016. doi: 10.1001/jamaophthalmol.2016.0600. URL https://dx.doi.org/10.1001/jamaophthalmol.2016.0600.

Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L Yuille. Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2304–2313, 2018.

Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18760–18769, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. *Advances in neural information processing systems*, 30, 2017.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*, 2022.

Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6721–6730, 2021.

Xiaobo Xia, Wenhao Yang, Jie Ren, Yewen Li, Yibing Zhan, Bo Han, and Tongliang Liu. Pluralistic image completion with gaussian mixture models. *Advances in Neural Information Processing Systems*, 35:24087–24100, 2022.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CZ8Y3NzuVzO.

Shaoan Xie, Lingjing Kong, Mingming Gong, and Kun Zhang. Multi-domain image generation and translation with identifiability guarantees. In *The Eleventh International Conference on Learning Representations*, 2022.

Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021.

Yu Yao, Mingming Gong, Yuxuan Du, Jun Yu, Bo Han, Kun Zhang, and Tongliang Liu. Which is better for learning with noisy labels: the semi-supervised method or modeling label noise? In *International Conference on Machine Learning*, pp. 39660–39673. PMLR, 2023.

Kaiwen Zha, Peng Cao, Yuzhe Yang, and Dina Katabi. Supervised contrastive regression. *arXiv preprint arXiv:2210.01189*, 2022.

Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3961–3970, 2022.

Guanglin Zhou, Lina Yao, Xiwei Xu, Chen Wang, Liming Zhu, and Kun Zhang. On the opportunity of causal deep generative models: A survey and future directions. *arXiv preprint arXiv*, 2301, 2023.

Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5861–5870, 2021.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

Appendix

A Ablation Analysis

Number of Augmented Instances. We conduct an experiment to assess the sensitivity of the models to the number of augmented instances. Specifically, we adjust the number of views from the set [1,3,5,10] and "dynamic", and exam the performance on POE w/ OCP-CL. As illustrated in Figure 6, the models' performance remains relatively stable when the number of augmented instances ranges between 3 and 5. However, over-supplementing the data with augmented instances can lead to a degradation in model performance. Interestingly, we find that a dynamic number of augmentations depending on the class could benefit the models. This is particularly relevant for ordinal regression datasets that suffer from class imbalance. Specifically, by increasing the number of instances in underrepresented classes, we observe an improvement in the overall performance of the methods.

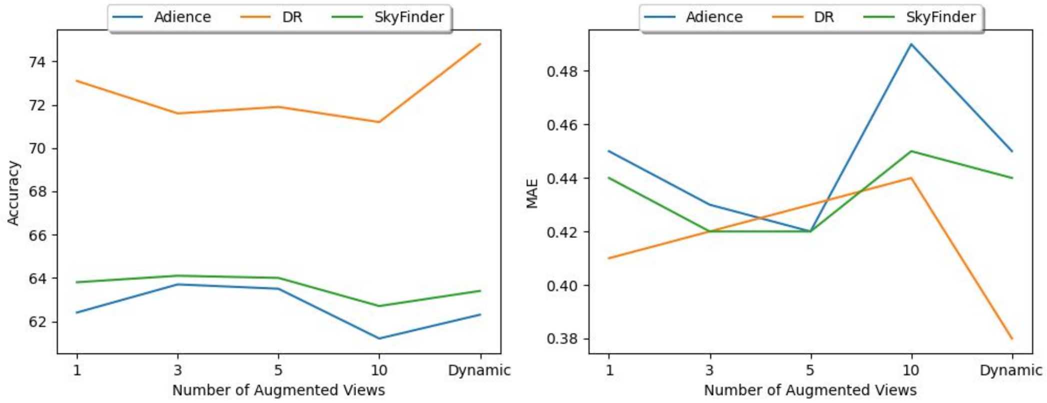


Figure 6: Ablation Study on the Number of Augmented Views.

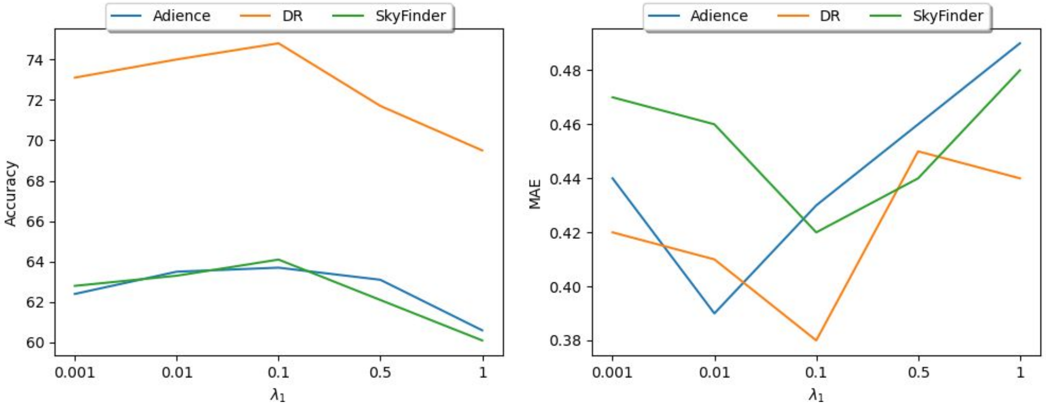


Figure 7: Sensitivity Analysis on λ_1 ratio.

Sensitivity Analysis on λ_1. We conduct experiments to assess the impact of the mask sparsity ratio, denoted as λ_1, on the performance of the ordinal regression model across three downstream tasks. For this purpose, we utilize POE w/ OCP-CL to examine sensitivity to changes in λ_1. We test five different sparsity ratios within the range of 1e-3 and 1. The results, presented in Figure 7, indicate that the ordinal regression model achieves optimal performance when λ_1 is set to 0.1 for all downstream tasks, and its performance decreases linearly with increases in the ratio beyond 0.1.

B Additional Related Works

In this section, we provide additional background discussions relevant to our work. Specifically, we discuss recent advancements in Generative Data Augmentation, Disentangled Representation Learning, and Nonlinear ICA and explore their relationship with our research.

Generative Models for Data Augmentations. Instead of generating data augmentations using predefined transformations, Generative Data Augmentation (GDA) employs an alternative approach that leverages Deep Latent Variable Models (DLVMs) to generate new synthetic views from existing samples, based on conditional generative processes. Antoniou et al. (2017) and Tran et al. (2017) propose the use of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Xia et al., 2022) to create a broader set of augmented data. More recently, Diffusion Models (DMs) (Sohl-Dickstein et al., 2015) have been utilized to alter high-level semantic attributes, thereby addressing the problem of lack of diversity along key semantic axes in data augmentation (Trabucco et al., 2023). While these approaches can generate impressive results that appear both realistic and novel, most of them are not guaranteed to maintain the invariance of the original data. Our proposed method fulfills the need for controllable generative data augmentation, offering a more trustworthy GDA approach.

Disentangled Representation Learning. The objective of Disentangled Representation Learning (DRL) is to construct a model proficient in recognizing and isolating the latent factors concealed within observable data (Wang et al., 2022). This isolation into semantically meaningful factors enhances the model's ability to produce interpretable data representations, thereby simulating the cognitive processes humans employ in understanding objects or relationships. In the context of generative modeling, Higgins et al. (2017) introduce a β-penalty coefficient for the KL divergence term in the evidence lower bound of a Variational Autoencoder (VAE) (Kingma & Welling, 2013; Li et al., 2022b; Huang et al., 2022; Hong et al., 2024; Lin et al., 2023) to balance latent channel capacity and independence constraints with reconstruction accuracy. Subsequently, various modifications to VAE have been introduced to improve its capability for disentanglement (Chen et al., 2018; Kumar et al., 2017). These include the incorporation of either implicit or explicit inductive biases as well as the utilization of diverse regularization techniques. On the other hand, InfoGAN (Chen et al., 2016) was the first to address the problem of disentangling latent factors in Generative Adversarial Networks, introducing an extra variational regularization of mutual information. Lin et al. (2019) introduced InfoGAN-CR, an unsupervised extension of InfoGAN that includes a contrastive regularizer to infer latent dimensions. Zhu et al. (2021) present PS-SC GAN, which builds upon InfoGAN and features a Spatial Constriction (SC) strategy to extract significant areas influenced by each latent dimension, along with a Perceptual Simplicity (PS) approach to make the latent factors more unambiguous. Wei et al. (2021) propose a method known as Orthogonal Jacobian Regularization (OroJaR) aimed at enhancing disentanglement in generative models. OroJaR uses the Jacobian matrix to examine how output alterations correspond to changes in input variables, specifically the latent dimensions. Our methods are parallel to GAN-based DRL methods, wherein we disentangle the latent factors by introducing the principle of minimal change.

Nonlinear ICA. Nonlinear independent component analysis (ICA) theoretically addresses the problem of disentangling latent factors when a nonlinear invertible transformation function exists, mapping independent samples to the latent space Hyvarinen & Morioka (2016; 2017). Recent developments Locatello et al. (2020); Zimmermann et al. (2021); Xie et al. (2022); Kong et al. (2022) indicate that, within a conditional generative process, the true latent factors might become identifiable when auxiliary information is provided. Khemakhem et al. (2020) demonstrate that the joint data and latent space distributions can be recovered, up to a simple transformation in the latent space, provided the generative process conditions on a variable observed alongside the data. Von Kügelgen et al. (2021) employ two views of the same image to disentangle the latent factors into ordinal content and non-ordinal components, with only the content component associating with the image's semantics. Our generative model leverages Nonlinear ICA theories to theoretically justify the disentanglement of latent factors. By manipulating the ordinal content variable, while keeping the ordinal content factors consistent, our model can generate augmented views that preserve ordinal content.

C Intuition of Why Augmenting non-ordinal factors

Here, we provide further discussion on why augmenting non-ordinal factors in images can benefit the training of ordinal regression/classfication models. In general, data augmentation aims to modify the styling factors in original examples that are not related to the predictive objectives of downstream tasks (Von Kügelgen et al., 2021).

In computer vision tasks, by changing the styles in images, we add more variety to the training data. This helps the model not to focus too much on the specific styles it sees in the training images. It teaches the model to recognize objects or features in images, no matter how the style of the image changes. This is important because in the real world, images can come in many different styles. So, adding style changes in training helps the model perform well on all kinds of images

In the context of ordinal regression, style information is referred to as non-ordinal information, governed by underlying non-ordinal factors. Our method also aims to enrich style diversity by altering non-ordinal information. This is achieved by randomly sampling non-ordinal factors while maintaining the ordinal content factors, then generating examples based on these factors. By preserving the ordinal content factors, we can change the image's style while keeping its ordinal content unchanged, thereby generating synthetic (counterfactual) images not seen in the training data. This approach enables neural networks to access more samples with diverse styles, thereby improving the generalization capabilities of ordinal regression models for unseen samples. As demonstrated in Figure 7, by randomly sampling non-ordinal factors while maintaining the ordinal content factors, we can alter various aspects of the image's style, such as people's dressing, background, and camera angles, etc., ensuring the ordinal content remains unchanged.

It is also important to emphasize two major advantages of our data augmentation methods: Firstly, existing image augmentation strategies do not guarantee the preservation of ordinal information during the augmentation process. For example, color jittering can change an image's color, potentially altering white hair to yellow, which could obscure the age of the person in the image. Secondly, our proposed data augmentation method is general. Our approach can be broadly applied to automatically infer ordinal content from other types of information and generate new examples with guarantees. While primarily tested on image data, our method's framework should be adaptable to non-image data. This adaptability is not achievable with traditional data augmentation methods, which mainly focus on image data. For instance, applying rotations to non-image data is not feasible.

D Visualisation of Data Augmentation

We provide additional visualizations of the augmentation results. As shown in Figure 8, we conducted experiments on three ordinal regression datasets. Our findings indicate that our augmentation method effectively retains age-related and weather-related features, while simultaneously introducing significant stylistic variations. In the case of diabetic retinopathy instances, the differences between the original and augmented views are subtle. Without domain-specific knowledge, it is challenging to conclusively determine whether the ordinal content has been preserved. Figure 9 demonstrates the generative results for altering the ordinal factors. For each instance, we fix the non-ordinal factors and replace the ordinal factors with age-specific ordinal factors (i.e., representing different age groups). The age-specific ordinal factor is extracted from training images of the corresponding age group. By visualizing the results, we can observe that the age of the individuals has changed following augmentation, while the styling information from non-ordinal factors remains similar. This effectively illustrates the efficacy of our method in disentangling ordinal and non-ordinal factors. We also present image generation results from a conventional GAN (Karras et al., 2020) in Figure 10. It is important to note an advantage of our method over conventional GANs: the inability of conventional GANs to disentangle ordinal factors from non-ordinal factors. This means they cannot guarantee the preservation of an image's semantic information. Additionally, our model focuses more on fine-grained details when constructing novel samples. This is evident in the detailed modeling of age-related components in facial images. While conventional GANs can achieve high generative quality, they sometimes fail to accurately represent certain age-related features in the images, such as generate hair for infants and child face for seniors.

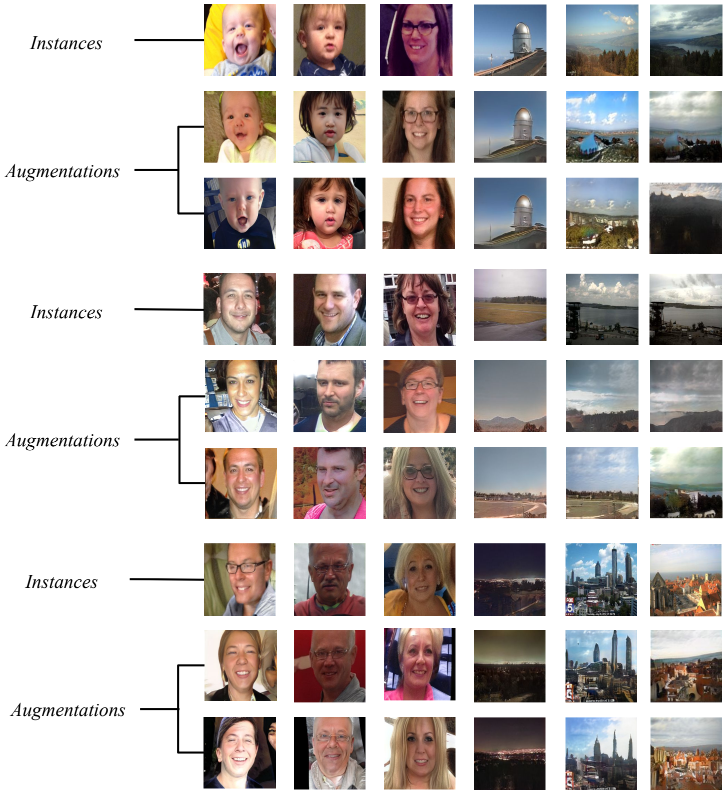


Figure 8: Generated augmentations by augmenting the non-ordinal factors $\hat{z}_n$.

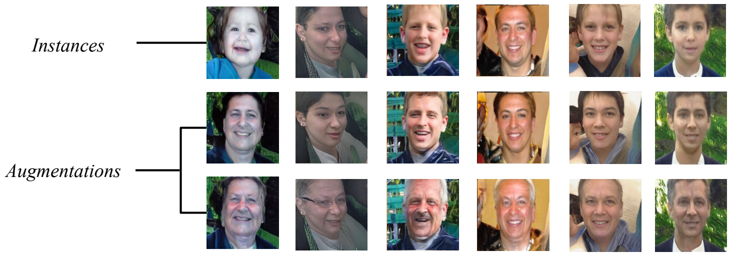


Figure 9: Generated augmentations by augmenting the ordinal factors $\hat{z}_o$ with age-specific factors.

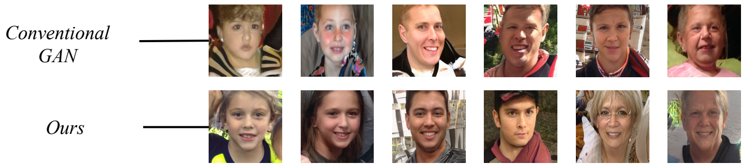


Figure 10: Unconditional image generation results of conventional GAN (the first Row) and Our method (the second Row).