Robust Image Ordinal Regression with Controllable Image Generation

Yi Cheng¹, Haochao Ying²∗, Renjun Hu³, Jinhong Wang⁴, Wenhao Zheng⁴, Xiao Zhang⁵, Danny Chen⁶ and Jian Wu⁷

¹School of Software Technology, Zhejiang University
²School of Public Health, Zhejiang University
³Alibaba Group
⁴College of Computer Science and Technology, Zhejiang University
⁵School of Computer Science and Technology, Shandong University
⁶Department of Computer Science and Engineering, University of Notre Dame
⁷Second Affiliated Hospital School of Medicine

{chengy1, haochaoying, wangjinhong, zhengwenhao, wujian2000}@zju.edu.com, renjun0hu@gmail.com, xiaozhang@sdu.edu.cn, dchen@nd.edu

Abstract

Image ordinal regression has been mainly studied along the line of exploiting the order of categories. However, the issues of class imbalance and category overlap that are very common in ordinal regression were largely overlooked. As a result, the performance on minority categories is often unsatisfactory. In this paper, we propose a novel framework called CIG based on controllable image generation to directly tackle these two issues. Our main idea is to generate extra training samples with specific labels near category boundaries, and the sample generation is biased toward the less-represented categories. To achieve controllable image generation, we seek to separate structural and categorical information of images based on structural similarity, categorical similarity, and reconstruction constraints. We evaluate the effectiveness of our new CIG approach in three different image ordinal regression scenarios. The results demonstrate that CIG can be flexibly integrated with off-the-shelf image encoders or ordinal regression models to achieve improvement, and further, the improvement is more significant for minority categories.

1 Introduction

Ordinal classification, which is also widely known as ordinal regression, is a specific type of classification task in which the categories follow a natural or logical order. Category orders are quite common in computer vision tasks, such as human age, image quality, and disease degrees of lesions. Therefore, image ordinal regression has been extensively applied to a number of diverse scenarios, ranging from image quality ranking [Diaz and Marathe, 2019] and monocular depth estimation [Geiger et al., 2013] to clinical image analysis (e.g., Gleason grading of prostate cancer [Bulten et al., 2022] and embryo stage classification/grading [Lukyanenko et al., 2021; Chen et al., 2022]).

In the literature, image ordinal regression studies have focused on exploiting the order of categories to boost accuracy. Related work can be roughly divided into regression-based, classification-based, and ranking-based methods. Regression methods [Fu and Huang, 2008; Guo and Mu, 2013] treat categorical labels as numerical values and directly use loss functions such as mean absolute/square errors to preserve category orders. But these methods may suffer from the non-stationary characteristics of differences between adjacent categories. Classification methods [Diaz and Marathe, 2019; Li et al., 2021; Shin et al., 2022] cast ordinal regression as multi-classification and leverage strategies such as soft labeling and relative order maintenance to incorporate the category relationships. Ranking methods [Niu et al., 2016; Fu et al., 2018] replace the original problem with multiple binary classifications and aggregate binary labels to derive ordinal labels heuristically. Overall, classification methods perform better than the other methods.

While some progress has been made, existing studies largely neglected the issues of class imbalance and category overlap, which are common in ordinal regression. Note that categories in ordinal regression follow certain orders, and it is often the case that data points are not evenly allocated along the metrics associated with these orders. For instance, there is only a very small fraction of images whose quality is rated as excellent or which corresponds to severe organ lesions. Indeed, we find that the least-sample category in our three datasets only accounts for 5%, 2%, and 0.2% of the total, respectively (Table 1). Moreover, categories in ordinal regression are often empirical rather than by definition, i.e., they are usually divided by rules. This implies the existence of a certain amount of near- or cross-boundary samples due to feature disturbance and inconsistent subjective judgment. As a result, adjacent categories may overlap (e.g., see Fig. 1). The above two issues increase the difficulty for models to learn meaningful category boundaries, especially for the less-represented minority categories. We empir-
We tackle image ordinal regression by directly addressing the class imbalance and category overlap issues, and present one of the first such methods in the literature.

- We propose a new plug-and-play framework CIG for the problem. The main novelty behind CIG is controllable image generation enabled by separating and fusing structural and categorical information of images.

- We verify the effectiveness and improved robustness of CIG on three different image ordinal regression tasks.

2 Related Work

In this section, we review related work on image ordinal regression and briefly overview the ideas of generation networks and self-supervised learning that inspire our work.

**Image ordinal regression.** Existing studies can be classified into regression-, classification-, and ranking-based methods. Regression methods treat categorical labels as numerical values and apply optimization. For instance, in [Fu and Huang, 2008; Guo and Mu, 2013], multiple linear regressions were utilized after dimensionality redundancy of the original image space was reduced with subspace learning. Classification methods cast ordinal regression as multi-classification and emphasize on properly incorporating correlation between categories. SORD [Diaz and Marathe, 2019] replaced the traditional one-hot label encoding with soft probability distributions, which allowed models to learn intra-class and inter-class relationships. POE [Li et al., 2021] represented a data point as a multivariate Gaussian distribution rather than a deterministic point in the latent space, and exploited the ordinal nature of regression via an ordinal distribution constraint. MWR [Shin et al., 2022] leveraged a moving window to refine the prediction of one image based on the supervision of its reference images from adjacent categories. Ranking methods transform an ordinal regression problem into a series of binary classification sub-problems. In [Niu et al., 2016], a multiple output CNN learning algorithm was proposed to collectively solve these sub-problems, and the correlation between these tasks was explored. A ranking-based ordinal loss was developed to ensure that predictions farther from the true label would incur a larger penalty [Fu et al., 2018].

Different from the previous studies, in this paper, we study image ordinal regression from the perspective of directly addressing the class imbalance and category overlap issues. These two issues are very common in ordinal regression but were largely ignored by previous work. Our work also focuses on the robustness of minority categories which we strongly believe is worth more thorough investigations.

**Generation networks** aim to produce images based on feature map vectors. U-Net [Ronneberger et al., 2015] and MAE [He et al., 2022] are two representative generation networks which are commonly used in conjunction with popular CNN-based and transformer-based encoders, respectively. During generation, U-Net connects the encoder and decoder layers that have the same feature map shape. This design makes it possible to contain both high-resolution information and high-dimensional abstract information in the decoder. MAE was originally proposed as an image auto-encoder, and was trained by recovering an image from its masked version.
We note that the lightweight decoder of MAE is very suitable as a generation network given feature maps produced by transformer-based encoders.

Our CIG framework adopts a generation network to produce extra training samples. Differently, we fuse two images to generate the artificial one and require the image generation process to be controllable. This is implemented by a separation-fusion-generation pipeline.

Self-supervision learning can obtain representations to help downstream tasks by learning from some auxiliary tasks. In [Liu et al., 2021], the authors proposed to partition an image into private and invariant domains via a domain-separation network. The separation network was trained based on the self-supervised orthogonality and similarity loss. CycleGAN [Zhu et al., 2017] implemented unpaired image-to-image translation by using a cycle loss that constrained the content consistency between the original and corresponding generated images. Inspired by these methods, our CIG learns to separate structural and categorical information of images by constraining the structural and categorical similarities between the original and generated images.

3 Methodology

In this section, we present our CIG framework with controllable image generation. A framework overview is given in Fig. 2. In a nutshell, CIG generates extra training samples with specific class labels near category boundaries, to facilitate learning of the task. More specifically, to classify an image, referred to as a main image, CIG first samples a reference image from its adjacent categories. Both the main and reference images are passed through the image encoder to extract feature maps. These feature maps are then combined to generate a fusion image through a separation-fusion-generation pipeline, such that the label of the fusion image is the same as the reference image. We exploit both self-supervision and classification supervision to train our CIG.

We first introduce the architecture of CIG in Section 3.1, and then discuss model supervision in Section 3.2.

3.1 Architecture

Sampler. Without loss of generality, we assume that the categories of our target image ordinal regression task are labeled consecutively with integers 1, 2, . . . , K, where K is the total number of categories. Given a main image \(X_m\) with a label \(m\), our CIG first samples a reference image \(X_r\) (with a label \(r\)) from the training set such that \(|m - r| = 1\), i.e., \(X_m\) and \(X_r\) are from adjacent categories. CIG then generates an artificial fusion image \(X_f\) of label \(r\) based on \(X_m\) and \(X_r\). This adjacent sampling assures that the fusion image is near the boundary between the categories \(m\) and \(r\).

We allow the generation process to bias toward minority categories that are originally less represented. We consider two samplers for CIG. The first one is an equal sampler, which samples \(X_r\) from adjacent categories with equal probability. The second one is called an inverse-ratio sampler. Let \(N_{m-1}\) and \(N_{m+1}\) denote the numbers of raw images in categories \(m-1\) and \(m+1\). The probabilities of sampling a reference image from categories \(m-1\) and \(m+1\) are then determined by \(N_{m-1}/N_{adj}\) and \(N_{m-1}/N_{adj}\), respectively, where \(N_{adj} = N_{m-1} + N_{m+1}\). Both samplers can increase the proportion of training samples, i.e., the main and fusion images, for minority categories.

Image encoder and classifier. The main and reference images are then passed through an image encoder to extract feature maps for subsequent generation and classification. Technically, CIG can use all image encoder architectures, and we explore the classic CNN-based architectures (e.g., VGG16 [Simonyan and Zisserman, 2015]) and the more recent Pyramid Vision Transformer (PVT) [Wang et al., 2021] architecture as the encoder in this work. A single fully-connected (FC) layer is adopted as the classification head which takes the extracted feature maps of images as input and outputs the predictive probabilities of all the categories.

Controllable generator. A core of CIG is controllable image generation through which we produce a fusion image \(X_f\) with a label \(r\) using the main and reference images \(X_m\) and \(X_r\). We implement the generation process using a separation-fusion-generation pipeline, and propose to separate the struc-
tural and categorical information of images. Structural and categorical information is the information for determining the overall structure and the category of an image, respectively.

The upper-left part of Fig. 2 illustrates the separation-fusion (S-F) operation. Specifically, we use either two 1 x 1 convolutional layers (for a CNN-based encoder) or two FC layers (for a Transformer-based encoder) as the structural and categorical extractors to extract relevant features from feature maps \( F_m \) and \( F_r \). The two extracted feature maps are concatenated into \( F_{sf} \). The channel numbers (or lengths) of the feature maps from the structural and categorical extractors are set as \( \tau \) or \( 1 - \tau \) times the channel number (or length) of the last feature map \( F^{(4)} \) from the encoder, respectively, where \( \tau \) is a percentage value for controlling the structural information proportion of the concatenated feature map from the S-F module. The computed structural feature map of \( X_m \) and categorical feature map of \( X_r \) are concatenated to generate the fusion image \( X_f \). As such, \( X_f \) is desired to be of class \( r \) while being structurally similar to \( X_m \), i.e., lying near the boundary between categories \( m \) and \( r \). Further, the categorical feature map of \( X_m \) is extracted and concatenated with its structural feature map for regularization (we will explain this below). CIG adopts UNet or the decoder of MAE as the generation network, depending on the encoder architecture (i.e., CNN-based or Transformer-based).

As shown in Fig. 3, in UNet, the concatenated feature map \( F_{sf} \) and the last feature map \( F^{(4)} \) of \( X_m \) are first input into an up-sampling block. This up-sampling process is repeated three times to generate \( X_f \). In the repetitions, the concatenated feature map is replaced by the output of the previous block. In the light-weight MAE decoder, a single Transformer block with a multi-head self-attention (MSA) layer and a feed-forward network (FFN) is used for the generation, and \( F_{sf} \) is repeated four times to ensure size consistency.

We apply three types of self-supervision to ensure the separation-fusion-generation pipeline to behave as we desire.

### 3.2 Model Supervision

CIG exploits self-supervision to guide information separation and image generation, and exploits classification supervision to improve the overall performance.

**Self-supervision.** To ensure the structural extractor works effectively, we require the fusion image \( X_f \) to be more structurally similar to the main image \( X_m \) than \( X_r \). We apply the commonly-used structural similarity index measure (SSIM) to quantify the perception-based similarity between two images \( X \) and \( Y \), as:

\[
SSIM(X, Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{\mu_X^2 + \mu_Y^2 + 2\sigma_{XY} + c_1^2 + c_2^2}
\]

where \( \mu_X \) and \( \sigma_X^2 \) are the pixel sample mean and variance of image \( X \), \( \sigma_{XY} \) is the covariance of images \( X \) and \( Y \), and \( c_1 = (0.01L)^2 \) and \( c_2 = (0.03L)^2 \) are variables for stabilizing the division operation (\( L \) is a dynamic range of the pixel-values of the images). The SSIM values are within \((0, 1]\), with larger values for higher similarity. We then minimize the following structural generation loss in order to enforce structural extraction, as:

\[
L_{SG} = -\frac{1}{2}\left( \log SSIM(X_m, X_f) + \log(1 - SSIM(X_r, X_f)) \right)
\]

On the other hand, the fusion image \( X_f \) should be more categorically similar to the reference image \( X_r \) than \( X_m \). Thus, we require the predicted categorical probability vectors \( P_f \) and \( P_r \) of \( X_f \) and \( X_r \) to be similar. Specifically, we minimize the squared Euclidean distance between the two un-normalized probability vectors to enforce categorical extraction. The corresponding categorical generation loss is:

\[
L_{CG} = ||P_r - P_f||^2,
\]

where \( P_r \in \mathbb{R}^K \) and \( P_f \in \mathbb{R}^K \) are the raw probability vectors of \( X_r \) and \( X_f \) output by the classifier.

Moreover, we desire to use “simple” structural and categorical extractors (the simpler the better). In other words, we hope that the encoder can learn to extract categorical information, instead of relying heavily on the extractors, as the categorical information will benefit the subsequent classification. Therefore, we further optimize a reconstruction loss between the feature map \( F^{(4)} \) and the concatenated vector \( F_{sf} = \text{concat}(h_c(F^{(4)}_m), h_s(F^{(4)}_m)) \) of the main image, as:

\[
L_{RC} = ||F^{(4)}_m - \text{concat}(h_c(F^{(4)}_m), h_s(F^{(4)}_m))||^2,
\]

where \( h_c \) and \( h_s \) stand for the categorical and structural extractors of CIG.

Finally, the overall self-supervised generation loss is computed as a weighted sum of the above three losses:

\[
L_G = \alpha \cdot L_{SG} + \beta \cdot L_{CG} + \lambda \cdot L_{RC}.
\]

**Classification supervision.** We use the traditional cross-entropy (CE) loss to optimize the classification capacity of CIG. The CE loss is evaluated and optimized only on the main and fusion images with labels \( m \) and \( r \), respectively. Let \( P_m \in \mathbb{R}^K \) and \( \hat{P}_f \in \mathbb{R}^K \) be the predicted categorical probability vectors of the main and fusion images outputted from the classifier. With some abuse of notation, the CE loss on \( X_m \) (with label \( m \)) can be expressed as:

\[
L_{CE}(P_m, m) = -\log \frac{\exp(P^h_m)}{\sum_{k=1}^K \exp(P^k_m)},
\]

where \( P^h_m \) denotes the value of the \( h \)-th entry in \( P_m \). The overall classification loss of CIG is a weighted sum of cross-entropy on the main and fusion images, as:

\[
L_C = L_{CE}(P_m, m) + \lambda \cdot L_{CE}(P_f, r).
\]
The complete training process of CIG is summarized as follows. We first train the entire model (i.e., the encoder, classifier, S-F module, and generator) in each batch. The encoder and classifier are supervised by the classification loss in Eq. (7) while the S-F module and generator are optimized by the self-supervision loss in Eq. (5). We find that the generator-related parameters are optimized much more slowly than the other parameters, and hence adopt two optimizers with different learning rates for optimization. After the generation network has been adequately optimized, we continue to train the encoder and classifier alone for another 8,000 to 24,000 batches. Note that with controllable image generation, the encoder is required to also extract structural information, while this information is not used for classification. The continued training allows the model to better focus on category-related features to further boost accuracy.

4 Experiments

We evaluate the effectiveness of our CIG approach on three different image ordinal regression tasks. Four sets of experiments are conducted to evaluate: (i) the overall effectiveness of CIG compared with known state-of-the-art methods, (ii) the robustness for minority categories, (iii) the contribution of each component in CIG, and (iv) the parameter sensitivity.

4.1 Experimental Setup

Datasets. We use three public datasets to evaluate our CIG. (1) Adience [Levi and Hassner, 2015] is a face image dataset from Flickr. Its categories correspond to human ages. (2) DR (Diabetic Retinopathy) [Liu et al., 2018a] contains high-resolution fundus images of patients. These images are classified according to the degrees of retina lesions. (3) Aesthetics [Schifanella et al., 2015] is another Flickr image dataset whose images are rated by the image quality.

Table 1 summarizes some statistics of these datasets, and Fig. 4 illustrates their ordinal categories with example images. Note that image ordinal regression on the three datasets corresponds to human age estimation, diabetic retinopathy diagnosis, and image quality ranking, respectively.

Metrics. We adopt classification accuracy (ACC) and mean average error (MAE) between predicted and ground-truth category probabilities for performance evaluation.

Implementation. Our CIG is implemented using PyTorch [Paszke et al., 2019], which is available at GitHub\(^3\). The inverse-ratio sampler is used by default and image encoders are initialized with the weights pre-trained on ImageNet1K [Deng et al., 2009]. We adopt the default Adam optimizer and a batch size of 18 for model training. The learning rates for the encoder and generator are set as \(1 \times 10^{-4}\) and \(5 \times 10^{-3}\), respectively. We optimize hyper-parameters on Adience with \(\alpha \in \{1, 2, 5\}, \beta \in \{1, 2, 5\}, \lambda \in \{0, 0.1, \ldots, 1\}\), and \(\tau \in \{0.1, 0.2, \ldots, 0.9\}\), and choose \(\alpha = 5, \beta = 2, \lambda = 0.2,\) and \(\tau = 0.2\) for all our tests. We use 5-fold (on Adience and Aesthetics) or 10-fold (on DR) cross-validation, and report the average results. All the experiments are conducted on a machine with 16 Intel(R) Xeon(R) Gold 6226R 2.90GHz CPUs and an NVIDIA RTX 3090 GPU.

4.2 Comparison with Known Methods

We first evaluate the effectiveness of our approach by comparing with seven image ordinal regression methods (including state-of-the-art ones). The results are presented in Table 2, in which CIG-VGG and CIG-PVT denote two variants of our approach based on the CNN and Transformer encoders, respectively. From the results, we observe the following.

Overall, the more recent SORD, POE, and MWR methods perform better than the other baselines. More specifically, MWR gives the best results among these three methods on Adience. This is because MWR further exploits the fine-grained categorical information (e.g., concrete human ages) to refine the results, while such information is not available on the other two datasets. SORD attains competitive classification accuracy due to its soft label design. However, soft labels could introduce noise in ground-truth category probabilities, yielding large MAE by SORD, especially on DR and Aesthetics. On the other hand, POE shows good classification accuracy and MAE by aggregating results that correspond to different samples from an estimated distribution.

Our CIG-PVT model using the Transformer encoder consistently outperforms all the baselines in both ACC and MAE on the three datasets. Indeed, our absolute improvements in ACC (\(\uparrow\)) and MAE (\(\downarrow\)) are (1.8\%, 0.4\%, 0.21\%) and (0.02, 0.008, 0.008) on Adience, DR, and Aesthetics, respectively. The improvement on Adience by CIG is bigger. This is probably because samples in Adience are distributed more evenly in categories and the categorical and structural information of

\(^3\)https://www.kaggle.com/c/diabetic-retinopathy-detection
\(^4\)https://github.com/Ch3ngY1/Controllable-Image-Generation
Table 2: Performance comparison of CIG and known methods. The best and second-best results are marked in bold and underlined, respectively. ‘−’ indicates that we cannot find or reproduce the results due to private implementation of the original papers or inapplicable settings.

Table 3: Effectiveness comparison for minority categories.

4.3 Robustness on Minority Categories

Next, we examine the robustness of different methods for minority categories. Categories on which the classification accuracy is much lower than the highest one are taken as minorities, which are categories {4, 6, 7, 8}, {2, 3, 4, 5}, and {1, 2, 4, 5} in our three datasets, respectively. Based on the results in Table 2, we only compare CIG-PVT with SORD and POE in this set of tests. The overall and detailed results are presented in Table 3 and Fig. 5. We observe the following.

First, the performance on minority categories is worse for all the methods. For instance, the ACCs of (SORD, POE, and our CIG-PVT) are (29%, 20%, 22%), (54%, 42%, 39%), and (56%, 52%, 45%) lower than the overall accuracy on the three datasets, respectively. These results empirically support our hypothesis that the class imbalance issue should be carefully addressed for image ordinal regression tasks.

Second, our CIG-PVT consistently outperforms the other two baselines on minority categories of all the datasets, except for MAE on DR. Note that the soft labels of SORD are prone to decreasing the MAE for the ‘middle’ categories, but in the cost of higher MAE for the leftmost and rightmost categories (see Figs. 5b&5d). Yet, the lower MAE by SORD on cls3 of DR does not increase its corresponding ACC (Fig. 5a).

Third, the performance improvements of CIG-PVT on minority categories are more substantial. Specifically, on DR, CIG-PVT yields better ACC on two of the four minority categories and is on par with the best in one category (Fig. 5a). For the minority categories of Aesthetics, our results in ACC and MAE are consistently the best (Figs. 5c&5d). Overall, the ACC is improved by (8.4%, 4.8%), (20.4%, 5.5%), and (11.2%, 8.9%); and the MAE is decreased by (0.848, 0.064), (−0.272, 0.03), and (0.308, 0.028) compared to SORD and POE on the minority categories of the three datasets, respec-
4.4 Ablation Study

In the third set of tests, we conduct ablation study to empirically verify the rationality of our CIG approach. We consider the following designs/variants of CIG: direct image generation (IG) by simply adding the feature maps of the images, controllable image generation with the separation-fusion module (S-F), and continued training (CT) for the encoder. Moreover, we use CNN- and Transformer-based encoders (i.e., VGG and PVT) as well as an existing image ordinal regression model POE as the backbone to investigate the applicability of CIG. Due to the page limit, we report only the results on Adience in Table 4.

We find that the fusion-based image generation strategy is effective to deal with the class imbalance issue in image ordinal regression. With IG, ACC and MAE of the three backbone models VGG, PVT, and POE are improved by (0.8%, 1.7%, 1.2%) and (0.018, 0.010, 0.010), respectively. Moreover, the S-F module can further improve the performances, by (2.8%, 0.6%, 0.6%) in ACC and (0.047, 0.011, 0.004) in MAE, indicating that our controllable image generation by separating the structural and categorical information is more reliable. Finally, our complete CIG with continued training consistently yields the best results, i.e., CT increases ACC by (0.4%, 0.5%, 0.5%) and decreases MAE by (0.014, 0.013, 0.008). This is probably because the encoder could better focus on extracting category-related information of images in the continued training phase. Overall, we show that our designs of CIG are generally useful, and together they assure the effectiveness of CIG as a whole. In addition, one can see that CIG is plug-and-play and flexible, and off-the-shelf image encoders or models can be readily integrated with CIG to bring further improvement for image ordinal regression.

We also test the effectiveness of two different samplers (i.e., equal and inverse-ratio) for CIG. The results on Adience are reported in Table 5. We find that the inverse-ratio sampler is better since it can supplement more images for minority categories than the equal sampler.

### 4.5 Parameter Sensitivity

Finally, we evaluate the parameter sensitivity of CIG. To examine the impact of \( \tau \), i.e., the length parameter for the structural and categorical information vectors in the separation-fusion module, we vary \( \tau \) from 0.1 to 0.9, fix the other parameters to their default values, and test the ACC and MAE results, as shown in Fig. 6. When increasing \( \tau \), ACC first increases and then decreases in general with the increment of \( \tau \), and MAE first decreases and then increases on the contrary. The best ACC and MAE are attained at \( \tau = 0.2 \). This implies that the image encoder uses more bits (in the categorical information vectors) to maintain categorical information of images. The results for the other hyper-parameters (i.e., \( \alpha \), \( \beta \), and \( \lambda \)) are given in the Supplementary Material.

### 5 Conclusions

In this paper, we focused on the class imbalance and category overlap issues in image ordinal regression which have been largely overlooked. We proposed, to our best knowledge, the first image ordinal regression approach that directly addresses these two issues. We presented a novel framework CIG based on controllable image generation which can generate artificial images to facilitate learning more accurate and robust decision boundaries. Each generated image contains structural information of one image and categorical information of another image from adjacent categories. To achieve such controllable generation, CIG was designed to learn the separation of the structural and categorical information of images using three self-supervised objectives. Extensive experiments on three different image ordinal regression scenarios verified the effectiveness and robustness of CIG compared with state-of-the-art methods. More specifically, CIG based on the Transformer encoder established new best-known performances. We also empirically showed that previous methods incurred considerable robustness issues on minority categories, and our CIG approach yielded higher improvements on such categories. It is expected that our work will inspire further studies on robust image ordinal regression.
Acknowledgments
This research was partially supported by National Key R&D Program of China under grant No. 2018AAA0102100, National Natural Science Foundation of China under grants No. 62106218 and No. 62176231.

References
[He et al., 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15979–15988. IEEE, 2022.


