Parameterized Diffusion Optimization enabled Autoregressive Ordinal Regression for Diabetic Retinopathy Grading

Qinkai Yu[1], Wei Zhou[2], Hantao Liu[2], Yanyu Xu[3], Meng Wang[4], Yitian Zhao[5], Huazhu Fu[6], Xujiong Ye[1], Yalin Zheng[7], and Yanda Meng[1](✉)

[1] Computer Science Department, University of Exeter, Exeter, UK
[2] School of Computer Science and Informatics, University of Cardiff, Cardiff, UK
[3] The Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, China.
[4] Centre for Innovation and Precision Eye Health & Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
[5] Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
[6] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore.
[7] Eye and Vision Sciences Department, University of Liverpool, Liverpool, UK
Y.M.Meng@exeter.ac.uk

Abstract. As a long-term complication of diabetes, diabetic retinopathy (DR) progresses slowly, potentially taking years to threaten vision. An accurate and robust evaluation of its severity is vital to ensure prompt management and care. Ordinal regression leverages the underlying inherent order between categories to achieve superior performance beyond traditional classification. However, there exist challenges leading to lower DR classification performance: 1) The uneven distribution of DR severity levels, characterized by a long-tailed pattern, adds complexity to the grading process. 2)The ambiguity in defining category boundaries introduces additional challenges, making the classification process more complex and prone to inconsistencies. This work proposes a novel autoregressive ordinal regression method called AOR-DR to address the above challenges by leveraging the clinical knowledge of inherent ordinal information in DR grading dataset settings. Specifically, we decompose the DR grading task into a series of ordered steps by fusing the prediction of the previous steps with extracted image features as conditions for the current prediction step. Additionally, we exploit the diffusion process to facilitate conditional probability modeling, enabling the direct use of continuous global image features for autoregression without relearning contextual information from patch-level features. This ensures the effectiveness of the autoregressive process and leverages the capabilities of pre-trained large-scale foundation models. Extensive experiments were conducted on four large-scale publicly available color fundus datasets, demonstrating our model's effectiveness and superior performance over six recent state-of-the-art ordinal regression methods. The implementation code is available at https://github.com/Qinkaiyu/AOR-DR.

Keywords: Autoregressive ordinal regression · Diabetic retinopathy.

1 Introduction

Ordinal regression not only learns categories but also explicitly models the ordinal relationships between them [2,18] [23] [14,15,13] [10], which makes it ideal for classifying diabetic retinopathy (DR) [9] into five ordinal groups: No DR, Mild, Moderate, Severe, and Proliferative. However, two major challenges arise: 1) Long-tailed data, where categories like Proliferative are underrepresented, weakening ordinal relationships; 2) Ambiguity of class boundaries, where differences in annotators' definitions lead to inconsistent boundaries. We propose that autoregressive methods can address these challenges by breaking the ordinal regression task into binary classification steps, where each prediction conditionally depends on the previous one. This approach inherently requires more inference steps to determine the outcome for the categories positioned further along the ordinal scale. In other words, these 'further' categories (*e.g.* severe, proliferative) will contribute more significantly to the overall loss optimization compared with other categories (*e.g.* no DR, mild). This helps to compensate for the deficiencies caused by the under-representation of these 'further' categories in the long-tailed DR distributions. Meanwhile, the autoregressive mechanism effectively leverages the transition characteristics between adjacent categories, enabling a gradual and stepwise prediction that enhances discrimination at the ambiguous boundaries between classes.

However, ordinal regression methods based on the autoregressive paradigm remain largely unexplored. This is mainly because the autoregressive mechanism typically uses patch-level features extracted by a backbone network, converting image patches into discrete tokens [3]. These features are then fed into a downstream decoder to capture contextual information across patches, a process commonly employed in object detection [4] and image segmentation [21]. However, backbone networks used in image classification tasks are often specifically designed to extract continuous global features [6], which is contradictory to some extent. Ord2Seq [20] follows the above approach by converting class labels into sequences with ordinal information and adopts a masked prediction strategy reminiscent of the BERT [5] training paradigm, using a transformer to predict the sequence. Such an approach fails to fully leverage pre-trained models. It is worth noting that, the autoregressive paradigm has not been proposed for ordinal regression tasks. [12] breaks through aforementioned contradiction by enabling autoregressive training without vector quantization via their proposed diffusion process, enabling continuous global features to be utilized in downstream autoregressive tasks.

In this work, we propose a novel ordinal regression model called **AOR-DR** to address the challenges of long-tail distribution and ambiguity of class boundaries in diabetic retinopathy grading tasks. In particular, we introduce feature fusion to jointly condition image features with previous predictions. To this end, we adopt two widely used strategies: affine fusion and cross-attention fusion. Our

proposed method represents the first attempt at an autoregressive paradigm for ordinal regression, exploiting diffusion-based optimization to model conditional probability distributions effectively without tokenization. Additionally, our work is highly compatible with mainstream backbone architectures, benefiting from large-scale pre-trained foundation models, such as RETFound [24].

The contributions of this work are as follows:

- We decompose the DR grading task into a series of ordered steps by fusing the prediction of the previous steps with extracted image features as conditions for the current prediction step.
- We exploit the diffusion process to facilitate conditional probability modeling, enabling the direct use of continuous global image features for autoregression without relearning contextual information from patch-level features.
- Extensive experiments were conducted on four large-scale publicly available datasets, demonstrating our superior performance over six recently proposed state-of-the-art ordinal regression methods. We also observe that the classic autoregressive mechanism can bring a considerable performance boost solely on simple cross-entropy loss. However, its robustness may diminish across datasets with varying long-tailed distributions. In contrast, our proposed model consistently achieves optimal performance across all four dataset configurations.

2 Methods

An overview of the proposed AOR-DR framework is shown in Figure 1. The details of label preprocessing, model structure, and other proposed components are elaborated below.

2.1 Preliminaries

For a classification task involving K classes with images $I \in \mathcal{R}^{C\times H\times W}$, our autoregressive framework decomposes the ordinary process into *K-1* binary classification tasks. It is represented as $f : I \rightarrow y^{1:k-1}$, where $y^{1:k-1}$ denotes $\{y^1, y^2, ..., y^{K-1}\}$ and each $y^i \in \{0, 1\}$ is a binary variable. For example, the corresponding labels of 'Severe' in DR grading (fourth class, the original label is 3) would be $\{1, 1, 1, 0\}$. To align with the autoregressive paradigm, we introduce a start token $<BOS>$ and an end token $<EOS>$ at the beginning and end positions. The complete label in our training process becomes $Y = \{<BOS>, Y^1, Y^2, ..., Y^4, <EOS>\}$. Thus, for the given label Y, our proposed autoregressive model reformulates the classic ordinal regression as the following optimization objective:

$$\max p(y^1, ..., y^4) = \max \prod_{i=1}^{4} p(y^i | y^1, ..., y^{i-1}), \tag{1}$$

where $p(\cdot)$ represents the probability that the model predicts the correct outcomes.

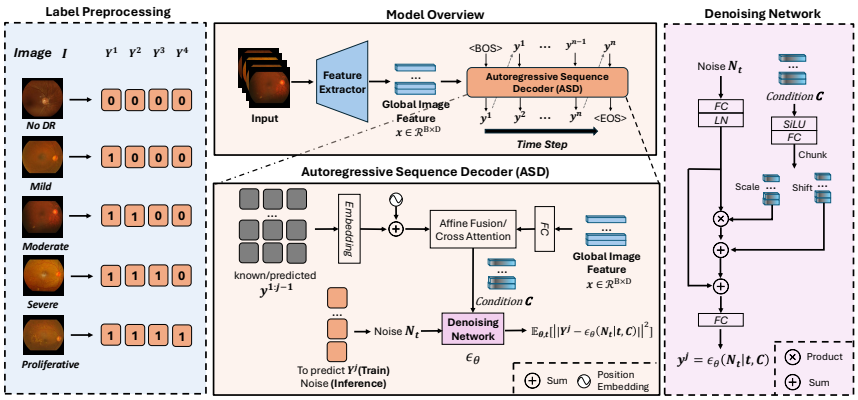

Fig. 1: Overview of the proposed AOR-DR framework. The Label Preprocessing part divides the DR grading task into a series of ordered steps. Our model architecture consists of two main modules: feature extractor and autoregressive sequence decoder. The image feature extractor outputs global image features, denoted as $x \in \mathcal{R}^{B\times D}$. The autoregressive sequence decoder takes these global features x along with the prediction from previous steps $y^{1:k-1}$ as input. The architecture of the denoising network is referenced by the final layer designed in DiT [17]. The condition C is linearly transformed and decomposed into learnable scaling and shift, then merged with the noise N_t through residual connections.

2.2 Model structure

Feature Extractor. Our model is compatible with commonly used mainstream backbones. In this work, we adopt the RETFound [24], a large foundation model pre-trained on 1.6 million color fundus images, as the feature extractor backbone. Specifically, we use vanilla ViT [6] (including ViT-B and ViT-L). The class token obtained by the ViT backbone is the global image feature x. The dimension D of x is 768 in ViT-B and 1024 in ViT-L. Even with such a powerful backbone network, our proposed autoregressive mechanism can still bring a large margin improvement over four large-scale datasets.

Autoregressive Sequence Decoder. In this component, image features are fused with predictions from the previous steps, and diffusion-based optimization is adopted to model continuous fusion features as conditions to the current step prediction. Inspired by the approach proposed by *Li et al.* [12], we use an MLP with multiple residual structures as the denoising network in the diffusion process. Differently, we introduce feature fusion to jointly condition image features with previous predictions. To this end, we adopt two widely used strategies: affine fusion and cross-attention fusion. To ensure the effectiveness of the autoregressive process, we introduce the position embedding to represent the current step in the sequence, similar to those in vanilla ViT [6]. The prediction from the previous steps is passed through an embedding layer h, then combined with the

corresponding position embedding e. We can define the embedded predictions y_{embd}^{j-1} donate in $j-1$ step as:

$$y_{embd}^{j-1} = h(y^{1:j-1}) + e^{j-1}. \tag{2}$$

Meanwhile, the global image features x, extracted from the backbone, are projected through a fully connected layer to match the same dimension as the embedded predictions from the previous step. The final fused features are fed into the denoising network as condition C for the current step's prediction. The fusion process of affine fusion and cross-attention fusion is formulated as follows:

$$C = \text{Condition}(x, y_{embd}^{j-1}) = FC(x) \cdot y_{embd}^{j-1} + FC(x), \tag{3}$$

and

$$C = \text{Condition}(x, y_{embd}^{j-1}) = \text{Softmax}\frac{(W_Q \cdot FC(x))(W_K \cdot y_{embd}^{j-1})}{\sqrt{d_K}}(W_V \cdot y_{embd}^{j-1}), \tag{4}$$

where W_Q, W_K, and W_V represent learnable weight matrices used to project the input image feature x and the previous embedding y_{embd}^{j-1} into query, key, and value spaces, respectively. The term $\sqrt{d_K}$ is a scaling factor.

2.3 Modeling Conditional Probabilities via Diffusion Process

Denoising diffusion networks are considered capable of modeling arbitrary distributions [7,19]. A previous study [22] demonstrated the effectiveness of using diffusion models for modeling entropy, achieving remarkable performance in classification tasks. In our case, the diffusion models are employed to represent the distribution of each ordinal step prediction in the autoregressive process. Let y^j be the prediction in step j. We reformulate the conditional probability optimization objective in Equation 1 as: max $p(y^j|C)$ at step j, where C combines previous predictions and image features. Here, we present the training and inference process in step j as an example.

Training. The y^j is fed into the denoising network after adding random Gaussian noise: $N_t = \sqrt{\alpha_t}y^j + \sqrt{1-\alpha_t}\epsilon$, where ϵ is the noise sampled from $N(0,1)$; α defines the noise schedule and t is the diffusion time step of the noise schedule. Modeling the conditional probability distribution $p(y^j|C)$ can be formulated with the following loss function:

$$L(C, y^j) = \mathbb{E}_{\epsilon,t}\left[\|Y^j - \epsilon_\theta(N_t|t, C)\|^2\right], \tag{5}$$

where ϵ_θ represents the denoising network with parameter θ. $\epsilon_\theta(N_t|t, C)$ means the network takes Noise N_t as input and is conditional on both t and C. In the end, Equation 5 is used as a diffusion-based parameterized optimization function during training. In this way, our proposed autoregressive mechanism directly leverages continuous global features to model conditional probabilities in continuous space, avoiding the limitations of traditional autoregressive paradigms

that involve introducing additional transformer decoders to relearn contextual information from patch-level features, which may hinder the effective utilization of pre-trained models.

Inference. In the inference phase, predictions are obtained by sampling the distribution $p(y^j|C)$. To be specific, the sampling follows the reverse diffusion procedure, starting with $y_t^j \sim N(0,1)$, ending with $y_0^j \sim p(y^j|C)$. Thus we can define $t-1$ step y_{t-1}^j of the reverse diffusion process as:

$$y_{t-1}^j = \frac{1}{\sqrt{\alpha_t}}(y_t^j - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(y_t^j|t,C)) + \sigma_t\delta, \tag{6}$$

where δ is sampled from the Gaussian distribution $N(0,1)$, and σ_t is the noise level at the diffusion time step t. Finally, we normalize the sampled y_0^j to get the binary prediction y^j.

3 Datasets and Implementation Details

Evaluation Settings. We conducted extensive experiments on four large-scale publicly available diabetic retinopathy grading datasets with various long-tail distribution characteristics, namely APTOS [8], Messidor [1], DDR [11], DeepDR [16]. The datasets were pre-split by the providers, and we used 20% of the training set for validation. Performance was evaluated using Accuracy (ACC), Macro F1-Score (F1), Sensitivity (Sen), and Specificity (Spec), with the highest scores in **bold** and the second-highest in underline. We compare the experimental results with the re-implemented state-of-the-art ordinal regression models with their publicly available code, including CORAL [2], CORN [18], PoEs [13], GoL [10], Ord2Seq [20], and CLIP-DR [23].

Implementation Details. To ensure the fairness of the comparison, all compared models use the same backbone network ViT-B/16 [6] with the input images resized to 224×224. To mitigate the randomness introduced by diffusion predictions, we perform five times predictions for each sample and use their average as the final results. Image augmentations only include horizontal flipping. All models were trained with a batch size of 32 for the same number of epochs. For the diffusion model hyperparameter, the total number of diffusion steps was set to 1000 for training, while 100 steps were used for inference sampling to improve efficiency.

4 Results

Main Results. Table 1 shows that our proposed model AOR-DR consistently outperforms other state-of-the-art methods across all evaluated datasets. Classic ordinal regression methods, such as k-rank classifiers-based models CORN [18] and CORAL [2] struggle with the long-tailed distribution of DR data, leading to suboptimal performance. Furthermore, other state-of-the-art methods cannot obtain consistent promising performance across the four datasets. For instance,

Table 1: Comparison with recent state-of-the-art ordinal regression models. * represents replacing the backbone as ViT-B/16. The best results are **bolded**, with underline indicating the second highest.

Model	APTOS				Messidor				DDR				DEEPDR			
	ACC	*F1*	*Sen*	*Spec*	*ACC*	*F1*	*Sen*	*Spec*	*ACC*	*F1*	*Sen*	*Spec*	*ACC*	*F1*	*Sen*	*Spec*
ViT-B/16	69.9	38.3	36.4	68.6	63.9	41.9	43.2	72.7	57.4	26.0	22.3	63.6	62.5	39.1	39.8	71.7
CORAL* [2]	66.4	30.4	27.3	67.8	62.2	18.7	29.0	72.3	54.1	26.6	24.7	67.2	57.7	37.3	36.9	71.2
CORN* [18]	69.6	31.4	27.9	60.2	62.2	18.6	24.9	72.3	54.7	26.3	23.9	66.5	56.2	31.4	28.6	71.1
PoEs* [13]	67.2	32.4	30.4	68.2	44.7	28.1	31.8	71.2	51.3	28.0	41.6	68.8	55.5	31.3	34.2	71.8
GoL* [10]	74.2	52.2	60.7	71.4	60.8	33.8	41.4	72.6	63.6	40.9	46.1	69.3	62.7	51.2	56.7	72.8
Ord2Seq [20]	70.2	30.4	27.7	64.8	**64.7**	34.1	44.5	72.6	64.7	34.1	36.9	65.5	61.2	43.2	52.4	71.3
CLIP-DR* [23]	71.8	45.5	44.5	71.1	63.3	36.8	45.8	72.7	68.4	45.6	50.5	69.2	64.5	50.6	57.1	72.5
Ours (Affine Fusion)	74.5	52.5	58.2	71.0	58.3	37.6	**52.2**	72.7	**70.3**	**46.5**	**51.6**	68.3	**65.7**	50.5	47.6	72.5
Ours (Cross Attention)	**77.1**	**58.6**	**62.6**	**71.9**	62.5	**45.1**	50.6	**73.0**	65.0	43.9	49.5	**69.7**	65.0	**51.9**	**57.7**	**72.9**

the relative ordinal estimate-based algorithm, GoL [10], struggled to maintain good performance on the Messidor [1] test set due to the absence of certain categories. Similarly, PoEs [13] performed poorly when confronted with the ambiguity of class boundaries (*e.g.* 'No DR' and 'Mild' in APTOS [8] test dataset). Although Ord2Seq [20] and CLIP-DR [23] demonstrated relatively robust grading performance, they fall short of achieving optimal performance compared to our proposed model. Specifically, our model outperforms Ord2Seq [20] 42 % on *F1* and 51 % on *Sen* in average and outperforms CLIP-DR [23] 13 % on *F1* and 13 % on *Sen* in average. To further demonstrate our model's superior ability to handle long-tail distributions and class ambiguities in DR grading, we compared its accuracy in predicting correct, adjacent, and other classes against vanilla ViR-B/16 [6], GoL [10] and CLIP-DR [23] on the APTOS [8] dataset. As shown in Figure 2, our model demonstrates superior performance across all categories, particularly notable improvements in the underrepresented classes, *e.g.* severe and proliferative. Additionally, Figure 3 shows the t-SNE visualization results of our model's representation before the final *FC* layer on the DDR [11] test dataset. Clear decision boundaries separating neighboring classes are observed at each step, where the expanding red area represents the set of samples whose ordinal grade has been conclusively determined, validating the effectiveness of the proposed autoregressive model. A potential concern with this autoregressive setup is the possibility of generating invalid ordinal sequences (e.g., [0, 1, 0, 1]). However, since the training dataset contains no such invalid constructs, the model implicitly learns to avoid them. Empirically, we did not observe any invalid sequences in the test set after the model was fully trained.

Ablation study on different backbones. Our proposed method can be integrated into widely used backbone architectures. We conduct ablation study experiments and evaluate its performance on the ViT [6] series, including ViT-B/16, and RETFound (ViT-L/16)[24], shown in Table 2. Notably, for RETFound [24], the backbone is frozen during training to demonstrate that our method can leverage the performance of pre-trained models, while for ViT-B, full parameters fine-tuning is performed. In all test dataset scenarios, *Ours* demonstrates supe-

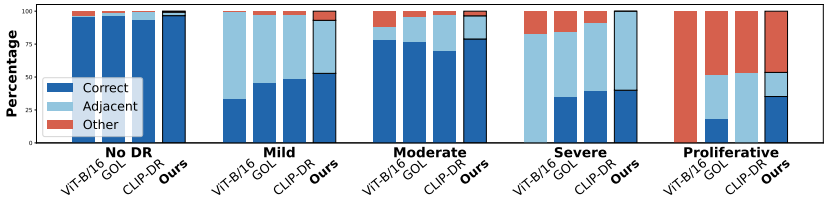

Fig. 2: The performance of the vanilla ViT-B/16 [6], GOL [10], CLIP-DR [23], and *ours* on the APTOS [8] dataset is analyzed for each category, presenting the percentage of test dataset that classify **correctly**, misclassified into **adjacent** categories, and misclassified into **other** categories.

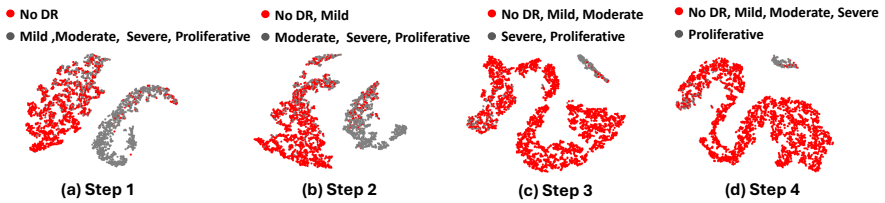

Fig. 3: The t-SNE visualization illustrates our model's feature distribution before the final FC layer on the DDR [11] test dataset.

rior adaptability and performance, consistently achieving promising performance under all four metrics. Relatively modest performance improvement is observed for the comparison with RETFound [24] due to the frozen backbone weights.

Ablation study on different optimization functions. To verify the effectiveness of the diffusion-based optimization function, we remain the model structure unchanged and replace the proposed parameterized diffusion loss with other commonly used loss functions, such as cross-entropy loss. The fusion strategy was selected based on the best-performing approach for each test dataset in Table 1. Results in Table 2 indicate that on the APTOS [8] and DDR [11] datasets, the cross-entropy based autoregressive method outperforms ViT-B/16. However, on the Messidor [1] and DEEPDR [16] datasets, which have severe long-tailed distributions with highly underrepresented classes, cross-entropy loss underperforms the baseline. This is likely because the autoregressive approach becomes trapped in repetitive patterns, hindering the learning of rare class features. Additionally, when training with the Retfound backbone frozen, the cross-entropy optimization fails to effectively leverage pretrained weights. In contrast, our proposed paradigm benefits from a diffusion-based optimization procedure that operates in continuous space, allowing for more effective conditional probability modeling of the global continuous features and thereby achieving superior performance.

Table 2: DR grading performance with different backbone network structures and different optimization functions. † represents freezing the backbone during training. The best results are **bolded**, with underline indicating the second highest.

Method	APTOS				Messidor				DDR				DEEPDR			
	ACC	F1	Sen	Spec	ACC	F1	Sen	Spec	ACC	F1	Sen	Spec	ACC	F1	Sen	Spec
ViT-B/16	69.9	38.3	36.4	68.6	63.9	41.9	43.2	72.7	57.4	26.0	22.3	63.6	62.5	39.1	39.8	71.7
AR+ce loss (*w*/ViT-B/16)	73.4	54.4	56.7	71.1	59.1	35.3	55.2	72.9	67.4	43.8	48.3	**69.8**	50.7	30.2	33.2	70.4
Ours (*w*/ViT-B/16)	**77.1**	**58.6**	**62.6**	**71.9**	**62.5**	**45.1**	**50.6**	**73.0**	**70.3**	**46.5**	**51.6**	68.3	**65.0**	**51.6**	**57.7**	**72.9**
RETFound† [24]	**82.1**	65.5	**69.2**	72.6	64.2	45.4	57.0	72.9	71.7	**57.2**	50.3	67.3	75.7	57.3	59.8	73.9
AR+ce loss (*w*/RETFound†)	63.0	40.3	45.8	70.9	58.3	37.4	38.0	72.8	64.6	46.5	50.2	**69.8**	49.75	34.8	38.9	71.4
Ours (*w*/RETFound†)	80.3	**65.7**	66.5	**72.8**	**64.7**	**54.0**	**62.4**	**73.9**	**72.8**	55.9	**55.3**	69.1	**75.9**	**58.1**	**59.9**	**73.9**

5 Conclusion

We propose a novel ordinal regression model to address the challenges of long-tail distribution and inter-class ambiguity in the diabetic retinopathy grading task. Our work is the first attempt to introduce an autoregressive mechanism to ordinal regression, combined with a parameterized diffusion optimization process, enabling the direct utilization of continuous image features without traditional tokenization. Extensive experiments were conducted on four large-scale publicly available datasets, demonstrating the effectiveness and superior performance of our proposed approach, compared with six recent state-of-the-art ordinal regression methods.

Acknowledgments This work is supported by Y. Meng's The Royal Society Fund (IEC\NSFC\242172), United Kingdom, and by H. Fu's Agency for Science, Technology and Research (A*STAR) Central Research Fund ("Robust and Trustworthy AI system for Multi-modality Healthcare"). Qinkai Yu thanks the PhD studentship funded by Liverpool Centre for Cardiovascular Science and University of Exeter.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abràmoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M.: Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative ophthalmology & visual science **57**(13), 5200–5206 (2016)
2. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters **140**, 325–331 (2020). https://doi.org/https://doi.org/10.1016/j.patrec.2020.11.008, http://www.sciencedirect.com/science/article/pii/S016786552030413X
3. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International conference on machine learning. pp. 1691–1703. PMLR (2020)

4. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021)
5. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Karthik, M., Dane, S.: Aptos 2019 blindness detection. Kaggle https://kaggle.com/competitions/aptos2019-blindness-detection Go to reference in p. 5 (2019)
9. Kempen, J.H., O'Colmain, B.J., Leske, M.C., Haffner, S.M., Klein, R., Moss, S.E., Taylor, H.R., Hamman, R.F., et al.: The prevalence of diabetic retinopathy among adults in the united states. Archives of ophthalmology (Chicago, Ill.: 1960) **122**(4), 552–563 (2004)
10. Lee, S.H., Shin, N.H., Kim, C.S.: Geometric order learning for rank estimation. Advances in Neural Information Processing Systems **35**, 27–39 (2022)
11. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
12. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. arXiv preprint arXiv:2406.11838 (2024)
13. Li, W., Huang, X., Lu, J., Feng, J., Zhou, J.: 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)
14. Li, W., Huang, X., Zhu, Z., Tang, Y., Li, X., Zhou, J., Lu, J.: Ordinalclip: Learning rank prompts for language-guided ordinal regression. Advances in Neural Information Processing Systems **35**, 35313–35325 (2022)
15. Liang, D., Xie, J., Zou, Z., Ye, X., Xu, W., Bai, X.: Crowdclip: Unsupervised crowd counting via vision-language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2893–2903 (2023)
16. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)
17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
18. Shi, X., Cao, W., Raschka, S.: Deep neural networks for rank-consistent ordinal regression based on conditional probabilities (2021)
19. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
20. Wang, J., Cheng, Y., Chen, J., Chen, T., Chen, D., Wu, J.: Ord2seq: Regarding ordinal regression as label sequence prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5875 (2023)
21. Wang, S., Wang, C., Gao, F., Su, L., Zhang, F., Wang, Y., Yu, Y.: Autoregressive sequence modeling for 3d medical image representation. arXiv preprint arXiv:2409.08691 (2024)
22. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 95–105. Springer (2023)

23. Yu, Q., Xie, J., Nguyen, A., Zhao, H., Zhang, J., Fu, H., Zhao, Y., Zheng, Y., Meng, Y.: Clip-dr: Textual knowledge-guided diabetic retinopathy grading with ranking-aware prompting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 667–677. Springer (2024)
24. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)