ENHANCING FEW-SHOT SAMPLES IN DEEP IMBALANCED REGRESSION VIA FROBENIUS NORM

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Imbalanced Regression (DIR) aims to learn a regression model where the training label distribution is extremely imbalanced. Recent studies in DIR have shown that aligning feature ordinality with that of labels can effectively enhance the prediction outcomes. However, these gains are largely confined to many-shot samples, while medium- and few-shot regions benefit little. This suggests that preserving feature-label ordinality alignment alone may be insufficient. Empirically, we observe that the feature Frobenius norm also exhibits a clear ordinal structure primarily shaped by well-fitted many-shot samples. Specifically, samples whose feature intensities deviate from this structure consistently exhibit higher test errors, whereas those that better conform to it tend to achieve lower errors. Therefore, this motivates us to investigate the feature intensity to further improve the model performance. In this paper, we theoretically reveal that conforming to the ordinality of the Frobenius norm can lead to good training performance across the labels. Furthermore, by characterizing the generalization properties through the lens of feature Frobenius norm across labels, we show that it should neither grow unboundedly nor be directly minimized. This motivates us to learn an appropriate scale for the feature Frobenius norm to address the imbalance in DIR. Based on this insight, we propose Frobenius-norm Enhanced Inference (FEI), a plug-and-play post-training pipeline that requires only a three-layer MLP to distill the feature-norm prior from a pretrained model and exploits the learned ordinal property to improve performance on medium- and few-shot samples. Extensive experiments on real-world datasets show that FEI improves medium- and few-shot performance by over 5% at the cost of less than 2% on many-shot samples in average, which further demonstrates the effectiveness of our method.

1 INTRODUCTION

Deep Imbalance Regression (DIR) Yang et al. (2021), where certain labels are less frequently observed or even missing when training Deep Neural Networks (DNNs) for regression tasks, is one of the most intriguing problems in machine learning. Due to the imbalanced training distribution and its continuous label space, DIR poses new challenges for accurately regressing target values on unseen, balanced test distributions.

Recently, modern solutions have attempted to solve DIR from several perspectives. For instance, Yang et al. (2021) incorporated smoothing into both the label and feature spaces to re-weight the loss functions. Gong et al. (2022) utilized ranking regularization to align the order of the feature space with the labels. Zha et al. (2023); Keramati et al. (2024) further introduced contrastive learning to maintain the ordinality of the feature space. Pintea et al. (2023); Xiong & Yao (2024b); Pu et al. (2025) leveraged classification to assist with DIR. Dong et al. (2025) utilized a geometric constraint to improve the quality of the representation feature for DIR. However, while previous works have achieved remarkable progress by leveraging the feature ordinality in tackling DIR, we can observe a clear performance divergence between the different shots of the data [1] in Fig. 1 Left.

Specifically, we observe that the overall performance is largely dominated by the many-shot region. In the meantime, many-shot samples consistently achieve substantially better performance than their

[1] Following Yang et al. (2021), the labels with <20 samples as the few-shot samples (Few.), >100 samples as the many-shot samples (Maj.), and others as the median-shot samples (Med.).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

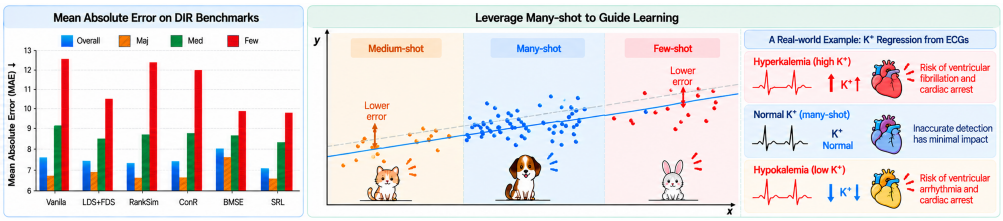

Figure 1: Left) Per-shot performance between different methods. Right) Motivation for leveraging the many-shot sample to help the other shots (Dashed line is the previously fitted model, while the solid line is the improved model).

medium- and few-shot counterparts. In contrast, medium- and few-shot samples exhibit considerably higher prediction errors, yet their degradation can be obscured by the aggregate metric due to their limited numbers in the training distribution. This discrepancy suggests that underrepresented regions remain insufficiently learned compared with the well-fitted many-shot region. Therefore, simply maintaining the feature-label ordinality becomes insufficient in guaranteeing per-label performance and the performance divergence between the different shots.

Meanwhile, as pointed out by Nie et al. (2025), medium- and few-shot samples can be particularly crucial in real-world applications, such as potassium (K^+) concentration regression from ECG signals (as illustrated in the right panel of Fig. 1.). To investigate, considering the label distribution shift from the imbalanced training distribution to the unseen balanced test distribution in DIR, DIR can be treated as a unique case of Domain Adaptation Regression (DAR) Chen et al. (2021); Nejjar et al. (2023); Adachi et al. (2025) where the test label distribution is balanced, and the test samples are not available during training Ren et al. (2022). Therefore, inspired by Chen et al. (2021); Adachi et al. (2025) that the scale of the Frobenius norm significantly affects the performance of regression under distribution shifts, we examine the feature Frobenius norm for each label and associate it with the corresponding test Mean Absolute Error (MAE) throughout the training process.

By calculating the Frobenius norm of each training label from its feature prototypes, we observe in Fig.2 that existing methods exhibit an approximately ordinal pattern in the feature Frobenius norms across the many-shot region (Fig.2). Meanwhile, these labels consistently correspond to substantially lower per-label test MAE in Fig.2. This observation suggests that the abundant supervision in the many-shot region enables the model to establish a well-structured feature-norm pattern, which helps to adequately fit these labels during training.

In contrast, the feature Frobenius norms of medium- and few-shot labels deviate noticeably from the approximately ordinal pattern established in the many-shot region. As shown in Fig.2, these deviations are accompanied by higher test MAEs than in the many-shot region. Together, these observations suggest that insufficient supervision prevents underrepresented labels from following the well-fitted feature-norm pattern observed in the many-shot region. Therefore, this motivates us to exploit the ordinality of the feature Frobenius norm that characterizes feature magnitude, to leverage the well-established pattern in many-shot regions to guide the learning of the under-fitted medium- and few-shot regions (as shown in Fig. 1 Right).

In this paper, we propose Frobenius-norm Enhanced Inference (FEI), a plug-and-play post-training pipeline to enhance the medium- and few-shot regions in DIR. We first provide a theoretical analysis of the role of the feature Frobenius norm in per-label learning in DIR. Our analysis reveals that effective fitting of the target labels requires a label-dependent feature magnitude. This suggests that the feature Frobenius norm should maintain an appropriate scale across different labels rather than being directly minimized or allowed to grow unboundedly. Motivated by this insight, we develop FEI, a two-phase training scheme that progressively enhances per-label learning performance for underrepresented medium- and few-shot samples by distilling knowledge from the well-trained many-shot samples, thereby enhancing overall regression performance.

Specifically, in the first phase, we train a tiny network to capture the well-established feature-norm pattern from the well-fitted many-shot region and use the learned relationship to predict appropriate feature norms for medium- and few-shot samples. In the second phase, we align the representations of underrepresented samples with their predicted feature norms and leverage the aligned prototypes to further retrain the regressor, thereby improving the learning of underrepresented regions by uti-

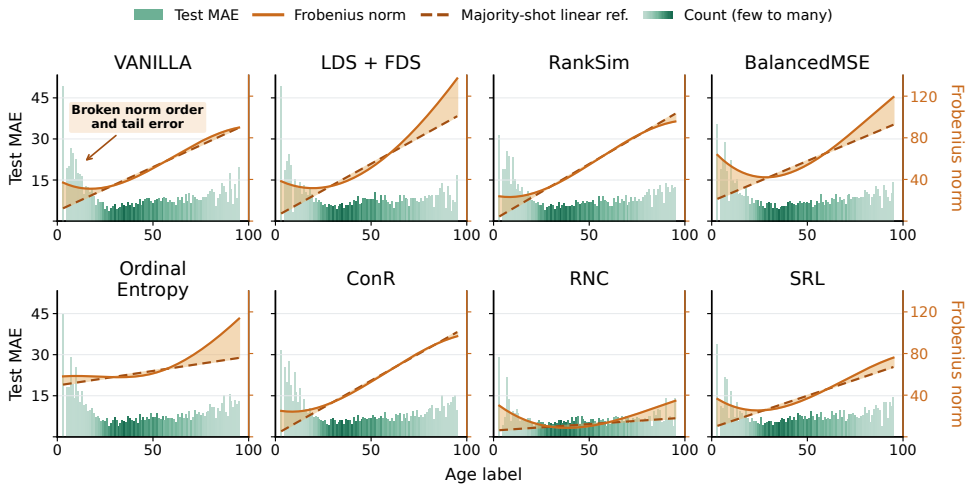

Figure 2: Illustration of the MAE and Frobenius norm between different DIR solutions.

lizing the knowledge acquired from the many-shot region. Extensive experiments on real-world benchmarks demonstrate that our method achieves over 4% and 6% improvements in the medium- and few-shot regions, respectively, while incurring less than a 1% performance degradation in the many-shot region.

In summary, we conclude our contributions as follows:

- We empirically observe the correlation between the Frobenius norm and the test MAE in DIR, and we theoretically demonstrate that learning towards an ordinal Frobenius norm across the labels can enhance training performance.
- We propose FEI, which incorporates the feature Frobenius norm of the many-shot samples to train a Frobenius linear network to guide the learning of median and few-shot samples.
- We leverage feature prototypes rescaled by the predicted feature Frobenius norm to retrain the regressor, which further helps to transfer the knowledge from the many-shot region.
- Extensive experiments on three real-world datasets validate the effectiveness of our method.

2 Methodology

In this section, we first provide a simple analysis to illustrate how the feature Frobenius norm can facilitate improved predictive performance. Then, we introduce FEI, a lightweight plug-and-play module that distills knowledge from well-fitted many-shot region to improve learning for medium- and few-shot samples.

2.1 Preliminary and Notations

In this paper, we have a N_{tr}-sample training set $D_{tr} = \{x_i, y_i\}_{i=1}^{N}$ where $x_i \in \mathcal{X}$ is the input data and $\mathcal{X}$ is the input space, and $y_i \in \mathcal{Y} = \mathbb{R}$ denotes the target label while $\mathcal{Y}$ is the label space [2]. Following Yang et al. (2021), we divide the continuous label space $\mathcal{Y}$ into B bins with equal intervals, and each bin index represents a target label. Therefore, the maximum target label would be y^B. In DIR, the training set D_{tr} follows an underlying highly skewed distribution $D_{tr} \sim P_{tr}$, while the testing set D_{te} is balanced. We denote the deep neural network as $\{f, r\}$, where f is the feature extractor and r is the regressor. We first extract the feature representation $z_i = f(x_i)$, where $z_i \in \mathcal{R}^d$ with d dimensions, and then we forward the feature to the regressor to obtain the target

[2]Following Ren et al. (2022); Zha et al. (2023), we discuss the target regression label with 1 dimension in this paper for easier interpretation, but this can still be generalized to multiple-label scenarios.

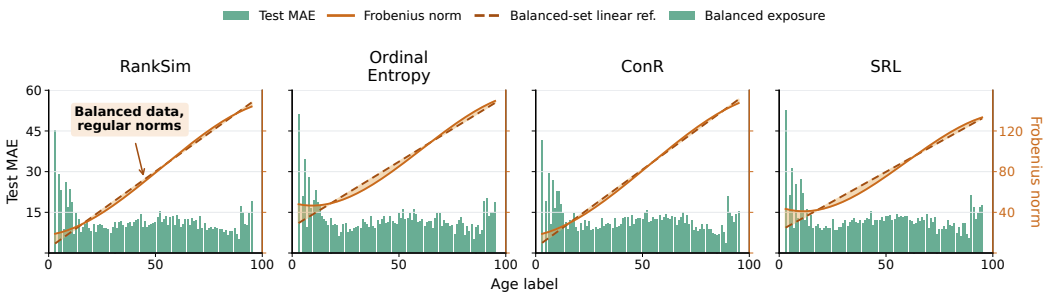

Figure 3: Illustration of the MAE and Frobenius norm between different DIR solutions under a balanced training set.

prediction $\hat{y}_i = r(z_i)$. Following Yang et al. (2021), we employ Mean Squared Error (MSE) as the loss function in the baseline.

2.2 Motivation: A Simple Analysis of DIR

For a feature representation z and a scalar target label. For $\lambda \geq 0$, considering the ridge objective for the regression task $w \in \arg\min_{w \in \mathcal{R}^d} (z^\top w - y)^2 + \lambda \|w\|_2^2$, we can obtain the closed form of the minimized w as $w = \frac{y}{\|z\|_2^2 + \lambda} z$. Since for a vector z, the Frobenius norm is equivalent to the Euclidean norm, we have $w = \frac{y}{\|z\|_F^2 + \lambda} z$. Therefore, for a regressor r that aims to regress features from all labels to their corresponding w, the scale of the feature Frobenius norm would significantly affect w as y grows. Meanwhile, this also coincides with the empirical observation in Fig. 2, where the ordinal Frobenius norm of the features in each label would lead to considerably low training errors, which encourages us to regularize the Frobenius norm across the labels to tackle the DIR. More specifically, we theoretically reveal an inherent relationship between the target magnitude and the feature Frobenius norm required for effective fitting in DIR as follows:

Theorem 2.1. *Let $\mathcal{L}$. be the M-Lipschitz loss function of the DIR. If the samples in this region are perfectly fitted, e.g.,*

$$Z^{(k)} w = y^k \mathbf{1}_{n_k}, \tag{1}$$

Then, the feature Frobenius norm necessarily satisfies:

$$\|Z^{(k)}\|_F \geq \frac{y^k \sqrt{n_k}}{A} \tag{2}$$

where n_k is the number of samples in the target label y^k, and A is the upper bound on the $\ell_2 - norm$ of the regression head's weight vector w. Moreover, for an M-Lipschitz loss ℓ and the hypothesis class of the regression model $\mathcal{F}$, the empirical Rademacher complexity of the corresponding loss class satisfies:

$$\widehat{\mathfrak{R}} n_k(\ell \circ \mathcal{F}) \leq \frac{MA}{n_k} \|Z^{(k)}\|_F. \tag{3}$$

Thus, under a shared bounded regression head, the feature Frobenius norm required to fit a label region scales at least as $\Omega(|y^k|\sqrt{n_k})$, while the feature Frobenius norm also directly enters the region-wise generalization bound. Proofs can be found in the Supplementary Material.

Remark: The above theorem reveals that the feature Frobenius norm required for effective fitting is jointly related to the number of training samples n_k and the magnitude of the corresponding target y^k. Specifically, since $|Z^{(k)}|_F^2 = \sum_{i=1}^{n_k} |\mathbf{z}_i^{(k)}|_2^2$, the matrix-level Frobenius norm naturally aggregates the feature norms of individual samples (we can have $\frac{\|Z^{(k)}\|_F}{\sqrt{n_k}} \geq \frac{y^k}{A}$, where $\frac{\|Z^{(k)}\|_F}{\sqrt{n_k}} = \sqrt{\frac{1}{n_k} \sum_i \|z_i^k\|_2^2}$) for an idea fit. This connects our theoretical analysis to the sample-level empirical

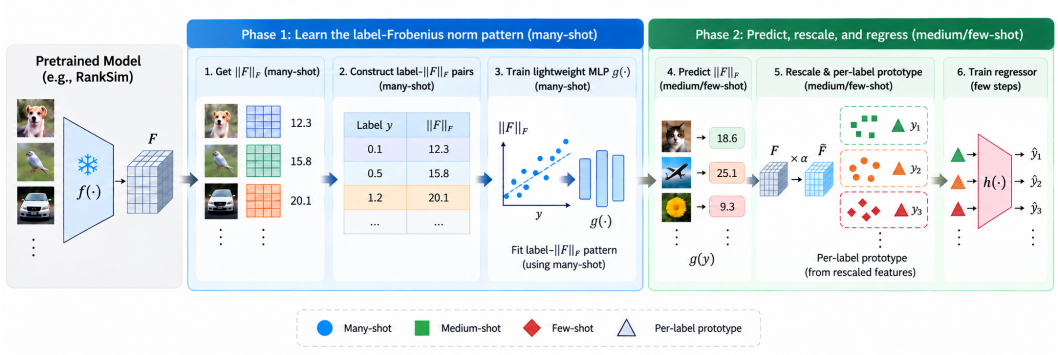

Figure 4: The Pipeline of FEI

observations in practice. Moreover, this suggests that the lower bound is primarily determined by the target magnitude and thus exhibits an ordinal relationship with y^k (Fig. 2). In DIR, the highly non-uniform sample label distribution introduces the additional factor $\sqrt{n_k}$ across the labels, which can distort this ordinal relationship across target regions. Therefore, the feature Frobenius norm should be encouraged to grow consistently with y^k.

Meanwhile, the upper bound on $\widehat{\mathfrak{R}}$ indicates that the feature Frobenius norm should remain controlled rather than grow without bound. As a result, the feature Frobenius norm should maintain an appropriate scale across different labels, rather than being directly minimized or allowed to grow unboundedly. Together, these observations motivate learning an appropriately scaled and ordinal feature Frobenius norm for effective regression under label imbalance. Interestingly, we can also observe from Fig. 3 that under the balanced training set, the feature Frobenius norm exhibits a clear ordinal relationship with the regression target, which is also associated with lower MAE. Therefore, we leverage the feature Frobenius norm of well-fitted many-shot samples to enhance the feature in under-fitted medium- and few-shot regions. Subsequently, we use the enhanced features to retrain the regressor to tackle the label imbalance in DIR.

2.3 FEI: Frobenius-norm Enhanced Inference

Motivated by our theoretical analysis, we enhance the medium- and few-shot samples through the lens of the feature Frobenius norm. To leverage knowledge from well-fitted many-shot region, we introduce a Frobenius-norm network to capture the underlying correspondence between target labels and their feature Frobenius norms. Intuitively, if a regression model were sufficiently trained across a balanced label space, as it typically is within the many-shot regions of DIR, the feature Frobenius norm would exhibit an ordinal relationship with the target label as we observed in Fig. 3. Therefore, we propose a two-phase training pipeline to distill the knowledge from the many-shot region to help the medium- and few-shot samples.

Phase 1: In phase 1, assuming we already have a pretrained model from any arbitrary method, e.g., RankSim Gong et al. (2022), ConR Keramati et al. (2024), SRL Dong et al. (2025), and so on. We first infer the many-shot samples in the training set and obtain their feature Frobenius norm (e.g., $F^k = \sqrt{\sum_i^d (e_i^k)^2}$ where e_i^k denotes the i-th element of the feature representation z^k for a sample with target label k in the many-shot region, and d denotes the feature dimensionality). Then, we construct a set of label-feature Frobenius norm pairs (e.g., $\{k : F^k\}$) for all the samples in the many-shot region to build a feature Frobenius set D_F^{Many}.

To fully exploit the knowledge from the well-fitted many-shot region, we use D_F^{Many} to train a 3-layer MLP g that captures the underlying ordinal relationship between the target labels and their corresponding feature Frobenius norms. Specifically, we take the target label as input to the MLP and use the corresponding feature Frobenius norm as the target label. In this way, we transform the complex input-to-target regression of DIR into a lightweight auxiliary regression problem that captures the underlying target–feature norm relationship. Therefore, we can distill the knowledge

encoded in the well-fitted many-shot region to mitigate the label imbalance in the medium- and few-shot regions.

Phase 2: In phase 2, we use the MLP to predict the feature Frobenius norm of samples in medium- and few-shot regions. Since we can observe from Fig. 2 that the feature Frobenius norm of the samples in medium- and few-shot regions is not reliable, we adopt the predicted feature Frobenius norm to rescale their feature representation z as follows:

$$\hat{z}^k = \frac{g(y^k)}{\|z^k\|_2} z^k = \alpha^k z^k \tag{4}$$

where we use α^k to denote the rescaled parameter of the target label k, which encodes the knowledge from the many-shot region to enhance the feature representations of other shot samples.

However, since feature rescaling does not alter the underlying imbalanced label distribution. To further exploit the improved representations after rescaling, we construct feature prototypes from the rescaled features and use them to retrain the regressor. Specifically, we obtain the feature prototypes of the same label and construct a prototype set D^{proto} as follows:

$$\bar{z}^k = \frac{1}{n^k} \sum \hat{z}^k \rightarrow D^{proto} = \{\bar{z}^k\}_{k=1}^{|Y|} \tag{5}$$

where $|Y|$ denotes the magnitude of the label space [3]. We use D^{proto} to train a regressor $\tilde{r}$ for the downstream tasks. The pipeline of FEI can be found in Fig. 4. During test time, when a new sample arrives, the deep neural network is constructed as $f, \tilde{r}$ for regression tasks.

3 Experiment

3.1 Implementation

We implement our experiments on three real-world datasets, AgeDB-DIR, IMDB-WIKI-DIR, and WIND. AgeDB-DIR and IMDB-WIKI-DIR involve tasks of age regression from facial images, while WIND is a wind speed estimation dataset. In this paper, we adopt the MAE as the loss penalty for the regression prediction, which is also consistent with Yang et al. (2021). The reported results are averaged over five experiments with different random seeds. We adopted the following baselines to make the comparison with our proposed method: LDS and FDS Yang et al. (2021), BMSE Ren et al. (2022), RankSim Gong et al. (2022) , OE Zhang et al. (2023), Con-R Keramati et al. (2024), RNC Zha et al. (2023), and SRL Dong et al. (2025).

3.2 Datasets

We validate the effectiveness of our proposed method on three public real-world datasets as follows:

AgeDB-DIR is a real-world dataset that aims to regress age from facial images. It was constructed by Moschoglou et al. (2017) and re-organized by Yang et al. (2021) for DIR. It contains 12.2K images of training data, 2.1K images of validation data, and 2.1K images of testing data. The bin length is set to 1 year (the minimum resolution), the minimum age is 0, and the maximum age is set to 101 (following Yang et al. (2021) the centroids of the bins are set to be the target labels).

IMDB-WIKI-DIR is another real-world dataset aimed at regressing ages from facial images. It was constructed by Rothe et al. (2018) and re-organized by Yang et al. (2021) for DIR. It contains a total of 235K face images. Following Yang et al. (2021), 191.5K facial images are used for training, 11K images are used for validation, and 11K images are used for the testing set. Among them, the validation set and testing set are considerably more balanced than the training set. The minimum resolution is also set to 1 year for the AgeDB-DIR.

WIND-DIR is a real-world satellite image dataset to estimate the wind speed. It was constructed by Maskey et al. (2020) and re-organized with kernel density estimation by Jiang et al. (2025) for DIR. It contains 54K images for training, 14K images for validation, and 44K images for testing, and the wind speed ranges from 0 to 185 (We leave the results and analysis in the supplementary materials).

[3] Although regression problems are typically defined over a continuous and potentially unbounded label space, the observed label range is generally bounded in practical applications.

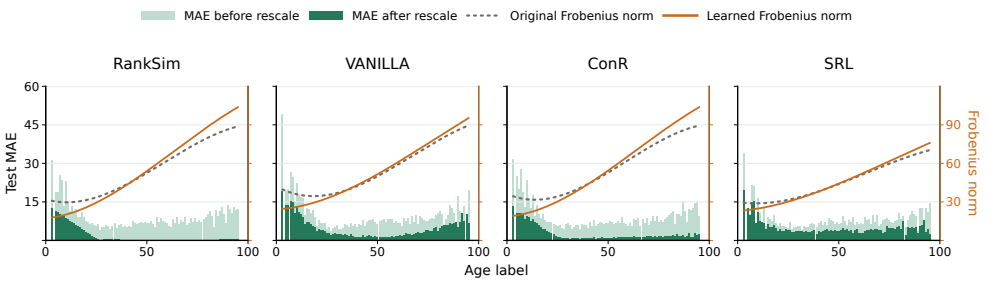

Figure 5: Performance of four DIR solutions enhanced by FEI (Dot line is the previous feature and solid is the scaled feature representation).

Shot / Method	MAE↓ All	MAE↓ Many.	MAE↓ Med.	MAE↓ Few.	GM↓ All	GM↓ Many.	GM↓ Med.	GM↓ Few.
VANILLAYang et al. (2021)	7.77	6.62	9.55	13.67	4.94	4.26	6.54	9.14
VANILLA+Ours	7.65	6.93	8.64	11.81	4.82	4.30	5.99	7.77
LDS+FDS	7.67	6.99	8.68	11.27	4.87	4.46	5.78	6.96
LDS+FDS+Ours	7.67	7.01	8.61	11.22	4.86	4.48	5.56	7.16
RankSim Gong et al. (2022)	6.92	6.34	7.79	9.91	4.28	3.91	4.84	7.01
RankSim+Ours	6.91	6.40	7.70	9.52	4.34	3.99	4.98	6.53
BalancedMSE Ren et al. (2022)	7.51	6.62	9.21	11.18	4.71	4.20	5.82	7.78
BalancedMSE + Ours	7.49	6.70	8.92	10.96	4.70	4.25	5.74	7.03
Ordinal Entropy Zhang et al. (2023)	7.59	6.59	9.16	12.72	4.88	4.18	6.49	9.81
Ordinal Entropy + Ours	7.46	6.84	8.21	11.27	4.70	4.20	5.57	8.39
ConR Keramati et al. (2024)	7.43	6.51	8.75	12.45	4.79	4.19	5.84	9.81
ConR+Ours	7.40	6.56	8.63	11.92	4.75	4.18	5.78	9.37
UVote Jiang et al. (2025)	7.34	6.77	12.01	21.22	4.26	3.97	4.86	5.85
UVote+Ours	7.26	6.76	11.32	20.60	4.31	4.05	4.77	5.82
SRL Dong et al. (2025)	7.22	6.64	8.28	9.81	4.53	4.17	5.33	6.35
SRL+Ours	7.19	6.72	7.94	9.59	4.43	4.15	4.98	5.94

Table 1: Evaluation on AgeDB-DIR.

3.3 Analysis of AgeDB-DIR

As shown in Table 1, our method consistently improves the performance of underrepresented regions (medium- and few-shot regions) across various DIR baselines. In the meantime, the method also largely preserves the performance of the many-shot region. The improvement is particularly evident for the VANILLA method, where our method reduces the medium- and few-shot MAE by approximately 9.5% and 13.6%, respectively, together with reductions of 8.4% and 15.0% in GM. Similar improvements can also be observed for Ordinal Entropy Zhang et al. (2023), with around 10.4% and 11.4% reductions in medium- and few-shot MAE and 14.2% and 14.5% reductions in GM. Meanwhile, integrating our method with RankSim Gong et al. (2022), BalancedMSE Ren et al. (2022), ConR Keramati et al. (2024), and SRL Dong et al. (2025) also consistently improves the medium- or few-shot regions, with relative error reductions reaching approximately 9.6%.

Notably, these improvements generally introduce only minor degradation in many-shot performance, indicating that our method enhances underrepresented regions without substantially sacrificing the well-fitted many-shot region, which can be observed from Fig. 5 and is comparable to that of Fig. 3. Also, we can observe from Fig. 5 that the rescaled feature representations are more consistent with the many-shot region ordinal pattern, leading to a considerable performance improvement compared to the previous ones. Although the overall improvement can therefore appear modest, the considerably larger gains on medium- and few-shot samples demonstrate that our method effectively alleviates the performance disparity induced by label imbalance. The consistent improvements across diverse DIR baselines further demonstrate its complementary and plug-and-play nature.

Method \ Shot	MAE↓				GM↓			
	All	Many.	Med.	Few.	All	Many.	Med.	Few.
VANILLAYang et al. (2021)	8.04	7.21	15.18	25.89	4.53	4.13	10.77	18.90
VANILLA+Ours	7.93	7.18	14.19	25.07	4.54	4.20	9.38	17.10
LDS+FDS	7.68	7.06	12.78	21.87	4.33	4.08	7.48	12.71
LDS+FDS+Ours	7.65	7.08	12.40	21.50	4.31	4.08	7.05	12.70
RankSim Gong et al. (2022)	7.42	6.84	12.13	22.17	4.11	3.89	6.76	13.14
RankSim+Ours	7.41	6.85	11.88	21.93	4.09	3.87	6.59	13.55
BalancedMSE Ren et al. (2022)	8.07	7.51	12.43	23.22	4.61	4.38	7.20	13.71
BalancedMSE + Ours	8.08	7.53	12.39	23.21	4.65	4.41	7.26	14.71
Ordinal Entropy Zhang et al. (2023)	7.93	7.12	14.62	26.97	4.46	4.09	9.91	20.79
Ordinal Entropy + Ours	7.90	7.16	13.94	26.20	4.42	4.09	8.81	19.34
ConR Keramati et al. (2024)	7.77	7.00	14.28	25.43	4.37	4.01	9.60	18.88
ConR+Ours	7.66	6.95	13.52	24.88	4.37	4.04	8.73	17.34
UVote Jiang et al. (2025)	6.81	6.38	7.58	8.66	4.02	3.82	6.27	10.71
UVote+Ours	6.80	6.41	7.49	8.59	3.96	3.80	5.62	11.49
SRL Dong et al. (2025)	7.68	7.12	12.30	21.46	4.33	4.12	6.60	12.47
SRL+Ours	7.67	7.13	12.11	21.24	4.34	4.13	6.64	11.77

Table 2: Evaluation on IMDB-WIKI-DIR.

3.4 Analysis of IMDB-WIKI-DIR

On IMDB-WIKI-DIR, our method also consistently improves the MAE of the underrepresented regions across most baselines in Tab. 2. The medium-shot MAE is reduced by 6.5% for VANILLA, 3.0% for LDS+FDS Yang et al. (2021), 2.1% for RankSim Gong et al. (2022), 0.3% for BalancedMSE Ren et al. (2022), 4.7% for Ordinal Entropy Zhang et al. (2023), 5.3% for ConR Keramati et al. (2024). Improvements are also observed on few-shot MAE for all eight baselines, ranging up to 3.2%. In contrast, the many-shot MAE changes by less than 0.8% for every baseline. As a result, the overall MAE is maintained or improved for seven out of eight methods, with improvements of up to 1.4%. The clear difference between the changes in many-shot and underrepresented regions shows that our method mainly affects the medium- and few-shot samples, rather than obtaining the overall improvement by further fitting the already well-learned many-shot region.

A similar pattern can be found in GM, although the results show larger variations than MAE. In the medium-shot region, GM is reduced by 12.9% for VANILLA, 5.7% for LDS+FDS Yang et al. (2021), 2.5% for RankSim Gong et al. (2022), 11.1% for Ordinal Entropy Zhang et al. (2023), 9.1% for ConR Keramati et al. (2024), and 10.4% for UVote Jiang et al. (2025). For the few-shot region, notable reductions include 9.5% for VANILLA, 7.0% for Ordinal Entropy, 8.2% for ConR, and 5.6% for SRL Dong et al. (2025), indicating that the improvement is less uniform under GM. Meanwhile, many-shot GM varies up to 1.7%, further showing that the original many-shot performance is largely retained. Together with the AgeDB-DIR results, these observations support the design of our method that the sufficiently trained many-shot samples provide a relatively reliable label–feature Frobenius norm relationship, and transferring this relationship to medium- and few-shot regions mainly improves the representations where the original model is less sufficiently fitted.

3.5 Motivation for Leveraging Many-Shot Feature Frobenius norm

As shown in Fig. 6 Left, while many-, medium-, and few-shot samples remain largely overlapped in the feature space, their parameter gradients exhibit pronounced shot-dependent discrepancies. This demonstrates that different shot regions contribute unequally to model optimization. In particular, the rich supervision signal from many-shot samples enables more sufficient and stable optimization, allowing their representations to establish a more reliable relationship between the target label and feature Frobenius norm. In contrast, medium- and few-shot samples receive substantially fewer optimization opportunities, making their feature intensities more susceptible to the imbalance-induced optimization bias and thus deviate from this well-fitted relationship. This motivates us to treat the feature Frobenius norms learned from the well-optimized many-shot region as reliable knowledge, and we can further leverage their underlying label–norm relationship to guide the feature intensities of medium- and few-shot samples to improve the learning of underrepresented regions.

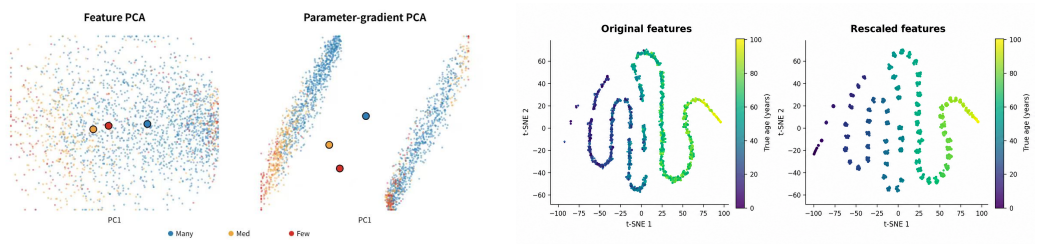

Figure 6: Right) PCA of the feature and parameter-gradient matrix. Left) Comparison of t-SNE between the original feature (RankSim) and the rescaled feature (Ours).

Method	Runtime		Model Size		
	Per-epoch time (s)	Per-epoch MLP time (ms)	Parameter size	MLP size	Percentage added
BalancedMSE Ren et al. (2022)	28.75	0.94	23510081	4353	0.018%
ConR Keramati et al. (2024)	79.23	0.87	23510081	1153	0.0049%
Ordinary Entropy Zhang et al. (2023)	29.99	1.07	23510081	16,897	0.071%
SRL Dong et al. (2025)	33.91	0.83	23772353	1153	0.0049%

Table 3: Evaluation of runtime and parameter size for AgeDB-DIR

3.6 Analysis on per-label feature t-SNE

We visualize one of the DIR solutions (RankSim Gong et al. (2022)) and our rescaled features using t-SNE. As shown in Fig. 6 Right, the feature representations in RankSim exhibit noticeable entanglement across different target regions and continuous ordinality. However, our feature rescaling makes the feature representation more structured with respect to the target. Moreover, samples with similar ages are more consistently organized into neighboring regions, while different target ranges become better separated. More importantly, the rescaled features exhibit a clearer progression along the continuous target space, which suggests that the predicted feature Frobenius norm does not merely alter feature magnitude. These observations provide qualitative evidence that our feature rescaling improves the representation of underrepresented target regions while maintaining the continuous structure required for DIR. These observations also coincide with the conclusion from Pintea et al. (2023) that introducing classification can help the regression tasks.

3.7 Efficiency Analysis

Tab. 3 evaluates the additional computational and parameter overhead introduced by our method on AgeDB-DIR. We adopt the same three-layer MLP architecture for all methods, with only the hidden-layer width adjusted among 32, 64, and 128 neurons. Specifically, we use 32 hidden neurons for the DIR methods ConR Keramati et al. (2024) and SRL Dong et al. (2025), while 128 neurons are used for the regression method Ordinal Entropy Zhang et al. (2023). Despite these variations, training the MLP requires only around 1 ms per epoch across different benchmarks, which is negligible compared with the tens of seconds required for the original model training. In terms of model size, the additional MLP parameters account for less than 0.1% of the original model across all evaluated methods, and as little as approximately 0.005% for ConR Keramati et al. (2024) and SRL Dong et al. (2025). These results also demonstrate that our method introduces negligible computational and storage overhead, making it a lightweight plug-and-play module for existing DIR methods.

4 Conclusion

In this paper, we theoretically investigate the role of the feature Frobenius norm in DIR. Motivated by our empirical observations and theoretical insight, we propose a two-phase training pipeline to leverage the feature Frobenius norm in well-fitted many-shot region to help learning in medium- and few-shot regions, which is not only lightweight but also effective in practice. Extensive experiments demonstrate the effectiveness and efficiency of our approach. Overall, our study establishes the feature Frobenius norm as a new perspective for understanding feature representations under label imbalance and provides a simple yet effective direction for addressing DIR.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with language polishing, grammar correction, and improving the clarity and readability of the manuscript. We have not used generative AI tools for ideas, methodology, experiments, analysis, and conclusions, and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we did not use generative AI tools for anything else. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve personally identifiable information or other sensitive data. We do not identify any specific ethical concerns associated with this research beyond those generally applicable to machine learning research.

REPRODUCIBILITY STATEMENT

Proofs of the theoretical results and the implementation code are provided in the supplementary material. We will publicly release the official code upon acceptance.

REFERENCES

Kazuki Adachi, Shin'ya Yamaguchi, Atsutoshi Kumagai, and Tomoki Hamagami. Test-time adaptation for regression by subspace alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SXtl7NRyE5.

Samir Brahim Belhaouari, Ashhadul Islam, Khelil Kassoul, Ala Al-Fuqaha, and Abdesselam Bouzerdoum. Oversampling techniques for imbalanced data in regression. *Expert Systems with Applications*, 252:124118, 2024.

Paula Branco, Luis Torgo, and Rita P. Ribeiro. Rebagg: Resampled bagging for imbalanced regression. In Luís Torgo, Stan Matwin, Nathalie Japkowicz, Bartosz Krawczyk, Nuno Moniz, and Paula Branco (eds.), *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 94 of *Proceedings of Machine Learning Research*, pp. 67–81. PMLR, 10 Sep 2018. URL https://proceedings.mlr.press/v94/branco18a.html.

Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *ICML*, pp. 1749–1759, 2021.

Zijian Dong, Yilei Wu, Chongyao Chen, Yingtian Zou, Yichi Zhang, and Juan Helen Zhou. Improve representation for imbalanced regression through geometric constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5082–5091, 2025.

Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. *arXiv preprint arXiv:2205.15236*, 2022.

Yuchang Jiang, Vivien Sainte Fare Garnot, Konrad Schindler, and Jan Dirk Wegner. Uncertainty voting ensemble for imbalanced deep regression. In Daniel Cremers, Zorah Lähner, Michael Moeller, Matthias Nießner, Björn Ommer, and Rudolph Triebel (eds.), *Pattern Recognition*, pp. 329–343, Cham, 2025. Springer Nature Switzerland.

Mahsa Keramati, Lili Meng, and R. David Evans. Conr: Contrastive regularizer for deep imbalanced regression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIuevDSK5V.

Jangho Lee. Toward a balanced feature space for the deep imbalanced regression. *IEEE Access*, 11: 92565–92574, 2023.

Jongin Lim, Sucheol Lee, Daeho Um, Sung-Un Park, and Jinwoo Shin. PRIME: Deep imbalanced regression with proxies. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hvaPL2DMpQ.

Dekun Lin, Tailai Peng, Rui Chen, Xinran Xie, and Zhe Cui. Let multi-classification help deep imbalanced regression. In *International Conference on Artificial Neural Networks*, pp. 430–447. Springer, 2024.

Yang Liu, Siru Wang, Wei Xia, Aaron Fenster, Haitao Gan, and Ran Zhou. Ldw-rs loss: Label density-weighted loss with ranking similarity regularization for imbalanced deep fetal brain age regression. In *International Conference on Neural Information Processing*, pp. 125–137. Springer, 2023.

Manil Maskey, Rahul Ramachandran, Muthukumaran Ramasubramanian, Iksha Gurung, Brian Freitag, Aaron Kaulfus, Drew Bollinger, Daniel J Cecil, and Jeffrey Miller. Deepti: Deep-learning-based tropical cyclone intensity estimation system. *IEEE journal of selected topics in applied Earth observations and remote sensing*, 13:4271–4281, 2020.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1997–2005, 2017. doi: 10.1109/CVPRW.2017.250.

Ismail Nejjar, Qin Wang, and Olga Fink. Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11744–11754, 2023.

Guangkun Nie, Gongzheng Tang, and Shenda Hong. Dist loss: Enhancing regression in few-shot region through distribution distance constraint. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YeSxbRrDRl.

Silvia L Pintea, Yancong Lin, Jouke Dijkstra, and Jan C van Gemert. A step towards understanding why classification helps regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19972–19981, 2023.

Ruizhi Pu, Gezheng Xu, Ruiyi Fang, Bing-Kun Bao, Charles Ling, and Boyu Wang. Leveraging group classification with descending soft labeling for deep imbalanced regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19978–19985, 2025.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7926–7935, 2022.

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, 04 2018. doi: 10.1007/s11263-016-0940-3.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110:1–25, 08 2021. doi: 10.1007/s10994-021-06023-5.

Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. Smote for regression. In *Portuguese conference on artificial intelligence*, pp. 378–389. Springer, 2013.

Haipeng Xiong and Angela Yao. Deep imbalanced regression via hierarchical classification adjustment, 2024a.

Haipeng Xiong and Angela Yao. Deep imbalanced regression via hierarchical classification adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23721–23730, 2024b.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pp. 11842–11851. PMLR, 2021.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=raU07GpP0P.

Shihao Zhang, Kenji Kawaguchi, and Angela Yao. Deep regression representation learning with topology. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=HbdeEGVfEN.

Shihao Zhang, Yuguang Yan, and Angela Yao. Improving deep regression with tightness. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dkoiAGjZV9.

A Related Works

A.1 Deep Imbalanced Regression

DIR was first studied in Yang et al. (2021) and has been widely discussed in recent research. SMOTER Torgo et al. (2013) was first proposed to use synthetic data to address the imbalance. Branco et al. (2018); Belhaouari et al. (2024) leveraged the oversampling strategy for imbalanced regression tasks. However, they mostly handle low-dimensional imbalance cases. Steininger et al. (2021); Yang et al. (2021) incorporated kernel-density estimation (KDE) to assign smoothed weights to different labels. Liu et al. (2023) extended the work from Yang et al. (2021) in estimating label density to deal with the fetal brain age regression task under the label. BMSE Ren et al. (2022) reviewed DIR from a statistical perspective and proposed a balanced-MSE loss function to accommodate the imbalanced training label distribution, but its performance relies heavily on a balanced testing distribution.

Moreover, modern research has also investigated leveraging representation learning to address the DIR. RankSim Gong et al. (2022) was the first work proposed to tackle the DIR by maintaining the relationship between the label space and the feature space. Meanwhile, Liu et al. (2023) introduced a ranking regularization for the feature representation to handle label imbalance. Zha et al. (2023) incorporated contrastive learning to pull the samples against each other with different scales based on their relative distances. Lee (2023) attempted to learn a feature space with a more balanced structure. Con-R Keramati et al. (2024) modeled global and local label similarities in the feature space and pushed the negative samples from the anchors with more penalties. Dong et al. (2025) also tried to use geometric constraints in the feature representation space (e.g., unit hypersphere) to tackle the distortion caused by target label imbalance. Lim et al. (2025) leveraged learnable proxies to construct a balanced feature space for DIR. Although they have achieved impressive results in DIR, they often encounter heavy computational burdens. Furthermore, Zhang et al. (2023; 2024; 2025) proposed entropy-based methods to regularize the feature representation. However, they mostly focused on balanced regression tasks.

Specifically, it has been explored by recent research Pu et al. (2025); Xiong & Yao (2024a); Pintea et al. (2023) that treating regression as a classification task can effectively help address the imbalance. Moreover, Lin et al. (2024) adopted classification as an auxiliary task to provide additional gradients for median and few-shot samples through multi-grained classifiers. Lim et al. (2025) leveraged the discretized distance measure to the label density in the continuous space to facilitate learning. Jiang et al. (2025) leveraged an uncertainty-driven multi-expert system to dynamically address the DIR. However, the inherent difference between classification and regression would still hinder the effectiveness of classification in solving the DIR (e.g., unbounded label space).

Method \ Shot	MAE↓				GM↓			
	All	Many.	Med.	Few.	All	Many.	Med.	Few.
VANILLAYang et al. (2021)	7.66	7.55	13.43	24.25	4.56	4.50	8.89	19.44
VANILLA+Ours	7.62	7.53	12.10	20.89	4.50	4.46	7.34	14.37
LDS+FDS	7.83	7.74	12.69	21.63	4.69	4.63	8.48	16.57
LDS+FDS+Ours	7.78	7.69	12.37	20.66	4.68	4.63	8.07	15.69
RankSim Gong et al. (2022)	7.79	7.70	12.40	21.88	4.61	4.56	7.63	14.87
RankSim+Ours	7.81	7.73	12.07	21.33	4.63	4.59	7.60	14.81
BalancedMSE Ren et al. (2022)	7.60	7.50	13.05	22.47	4.56	4.51	8.91	16.79
BalancedMSE + Ours	7.57	7.48	12.20	20.98	4.53	4.48	8.17	15.64
Ordinal Entropy Zhang et al. (2023)	7.74	7.66	11.78	22.86	4.56	4.52	7.29	16.05
Ordinal Entropy + Ours	7.75	7.68	11.35	22.01	4.58	4.54	6.88	14.71
ConR Keramati et al. (2024)	7.62	7.57	10.33	14.19	4.63	4.60	6.28	9.49
ConR+Ours	7.81	7.79	9.16	11.77	4.77	4.76	5.30	8.14
UVote Jiang et al. (2025)	7.28	7.23	11.09	15.42	4.33	4.28	7.41	12.25
UVote+Ours	7.30	7.23	11.02	15.33	4.33	4.29	7.29	12.06
SRL Dong et al. (2025)	7.81	7.71	12.85	21.71	4.39	4.35	7.39	18.47
SRL+Ours	7.91	7.83	11.98	19.87	4.67	4.64	7.10	16.02

Table 4: Evaluation on WIND-DIR.

B Experiment and analysis on WIND-DIR

We also conduct our experiments on the WIND-DIR Jiang et al. (2025). As shown in Table 4, incorporating our method consistently improves the performance in the underrepresented medium- and few-shot regions across almost all baselines. For the VANILLA method, our method reduces the medium- and few-shot MAE from 13.43 and 24.25 to 12.10 and 20.89, respectively, while the corresponding GM decreases from 8.89 and 19.44 to 7.34 and 14.37. More substantial improvements can be observed for ConR Keramati et al. (2024), where the medium- and few-shot MAE decreases from 10.33 and 14.19 to 9.16 and 11.77. This corresponds to relative reductions of approximately 11.3% and 17.1%, respectively. Similarly, more improvements are also observed in GM, which decreases from 6.28 to 5.30 for medium-shot samples. Meanwhile, it reduces from 9.49 to 8.14 for few-shot samples. These results demonstrate that the enhanced features according to the pattern learned from well-fitted many-shot regions can effectively improve the representations of under-fitted regions.

Our method also exhibits broad compatibility with different DIR approaches, including LDS+FDS Yang et al. (2021), RankSim Gong et al. (2022), BalancedMSE Ren et al. (2022), Ordinal Entropy Zhang et al. (2023), ConR Keramati et al. (2024), UVote Jiang et al. (2025), and SRL Dong et al. (2025). In particular, improvements in medium- and few-shot regions can be consistently observed even when the underlying methods adopt substantially different strategies in DIR. Meanwhile, the many-shot performance is generally maintained with only limited variation. Methods such as ConR and SRL exhibit a moderate trade-off between the many-shot and underrepresented regions. Consequently, improvements in medium- and few-shot regions do not always lead to better overall performance, since the aggregate metrics are dominated by the abundant many-shot samples. This observation demonstrates that our method specifically improves the under-fitted medium- and few-shot regions rather than merely optimizing the aggregate performance.

C Implementation details

We directly adopt the pre-trained models provided by the benchmark methods and perform our performance enhancement on top of them, without retraining the original models. We present a set of detailed training parameters for different methods. To make a fair comparison between our method and other works, following Yang et al. (2021), we use the MAE and Geometric Mean (GM) to evaluate the performance on the datasets. More specifically, we show the parameter setup for different methods in AgeDB-DIR in Tab. 5. We report results for the four subsets: All, Many, Median, and Few, for each dataset to identify the effectiveness of our method.

Method	MLP			Regressor	
	MLP hidden neurons	learning rate MLP	training time MLP (s)	learning rate regressor	training time regressor (s)
Vanilla Yang et al. (2021)	32	3e-3	2.3	3e-5	0.10
Ranksim Gong et al. (2022)	32	3e-3	0.76	1e-3	0.67
ConR Keramati et al. (2024)	32	3e-2	0.23	1e-3	0.69
SRL Dong et al. (2025)	32	3e-3	2.38	3e-6	0.72
BalancedMSE Ren et al. (2022)	64	1e-3	0.24	3e-6	0.13
Ordinal Entropy Zhang et al. (2023)	128	3e-3	0.97	3e-5	0.04
UVOTE Jiang et al. (2025)	128	1e-3	0.1	3e-5	0.01

Table 5: Parameter setup for different methods in AgeDB-DIR

D PROOFS

Proof. We first prove the 1-Lipschitz loss function case, and we show why regularizing the feature Frobenius norm is crucial in DIR. Under the standard Rademacher Complexity Shalev-Shwartz & Ben-David (2014), for a 1-Lipschitz loss function (e.g., MAE), the generalization error can be upper-bounded as follows: With probability 1-δ, for a class of linear regressors $f \in \mathcal{F}, f = wz^\top$ where w is the weight vector shared across the entire label space and z is the feature vector, following Shalev-Shwartz & Ben-David (2014), we have:

$$\mathcal{L}(f) - \mathcal{L}(\tilde{f}) \leq 2\mathcal{R}_n(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{n}} \tag{6}$$

where n is the number of samples, and $\mathcal{R}_\setminus$ is the empirical Rademacher complexity.

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}_\sigma[\sup_{\|w\|_2 \leq A, f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(z_i)] \\ &= \frac{1}{n}\mathbb{E}_\sigma[\sup_{\|w\|_2 \leq A, f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i w z_i^\top] \end{aligned} \tag{7}$$

where σ are independent random variables. Since w is shared across the labels, we have:

$$\sum_{i=1}^{n} \sigma_i w z_i^\top = w \sum_{i=1}^{n} \sigma_i z_i^\top \tag{8}$$

Following the Cauchy-Schwarz inequality, we take the w over the expectation, and we have:

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{A}{n}\mathbb{E}_\sigma[\| \sum_{i=1}^{n} \sigma_i z_i \|_2] \tag{9}$$

Following Jensen's Inequality, we have:

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &\leq \frac{A}{n}\mathbb{E}_\sigma[\| \sum_{i=1}^{n} \sigma_i z_i \|_2] \\ &\leq \frac{A}{n}\sqrt{\mathbb{E}_\sigma[\| \sum_{i=1}^{n} \sigma_i z_i \|_2^2]} \\ &\leq \frac{A}{n}\sqrt{\mathbb{E}_\sigma[\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_i \sigma_j z_i z_j^\top]} \end{aligned} \tag{10}$$

Since σ are independent random variables that are uniformly chosen from the set $\{-1, 1\}$, they have zero mean and unit variance. Therefore, $\mathbb{E}[\sigma_i \sigma_j] = 0$ for $i \neq j$, $\mathbb{E}[\sigma_i \sigma_j] = 1$ for $i = j$, and we have $\sum_{i=1}^{n} z_i z_i^\top = \sum_{i=1}^{n} \|z_i\|_2^2$. Since the squared Frobenius norm of the matrix is equivalent to the sum of the squared L2 norms of the feature vectors, we have:

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{A}{n}\|Z\|_F \tag{11}$$

where Z is the feature matrix consisting of $z_i, i \in [n]$ for the target label k. For brevity, for each target label k, we present the generalization error as ϵ^k, and we have $\epsilon^k \leq \frac{A}{n_k}\|Z\|_F^k + C^k$, where

C^k is a constant for each target label with respect to the number of training samples n_k, as stated in Eq. 6. However, this does not encourage us to regularize the $\|Z\|_F$ directly. Since A serves as an upper bound for the weight norm and the model needs to satisfy the task $y_i = wz_i$, we have $A \geq \frac{y_i}{\|z_i\|_F}$, when $\|z_i\|_F$ approaches to 0 (minimization), A would grow to infinity, which makes the generalization error bound trivial.

Therefore, we further investigate the $\|Z\|_F$. As we stated before, the model needs to satisfy the regression task, e.g., $< w, z^k > \approx y^k$. Following the Cauchy-Schwarz Inequality, we can derive:

$$|y^k| \leq \|w\|_w \|z\| \leq A \frac{\|Z^k\|_F}{\sqrt{n_k}} \tag{12}$$

To prove this, we have:

$$y^k = \frac{1}{n_k} \sum_{i=1}^{n_k} w z_i^\top = \frac{w}{n_k} \sum_{i=1}^{n_k} z_i \tag{13}$$

Applying the Cauchy-Schwarz inequality and taking the absolute value on both sides, we have:

$$|y^k| = |\frac{w}{n_k} \sum_{i=1}^{n_k} z_i| \leq \|w\|_2 \|\frac{1}{n_k} \sum_{i=1}^{n_k} z_i\|_2 = A \|\frac{1}{n_k} \sum_{i=1}^{n_k} z_i\|_2 \tag{14}$$

Applying the triangle inequality, we have:

$$\|\frac{1}{n_k} \sum_{i=1}^{n_k} z_i\|_2 \leq \frac{1}{n_k} \sum_{i=1}^{n_k} \|z_i\|_2 \tag{15}$$

Applying the Cauchy-Schwarz inequality again, we have:

$$\sum_{i=1}^{n_k} (\nVdash \cdot \|z_i\|_2) \leq \sqrt{\sum_{i=1}^{n_k} 1^2} \sqrt{\sum_{i=1}^{n_k} \|z_i\|_2^2} \tag{16}$$

where $\nVdash$ is the all 1 vector has the same length with z. By taking the definition of the feature Frobenius norm into account, we have:

$$\sum_{i=1}^{n_k} \cdot \|z_i\|_2 \leq \sqrt{n_k} \|Z^k\|_F \tag{17}$$

Therefore, we have:

$$|y^k| \leq A \frac{\|Z^k\|_F}{\sqrt{n_K}} \tag{18}$$

where we build the connection between the feature Frobenius norm and the target label as:

$$\|Z^k\|_F \geq \frac{|y^k|}{A} \sqrt{n_k} \tag{19}$$

This shows that the feature Frobenius norm must grows linearly with the label magnitude, which demonstrate the $\|Z^k\|_F \propto ||y^k \sqrt{n_k}$.

Furthermore, to prove the M-Lipschitz loss function (e.g., MSE). The empirical Rademacher complexity of the loss function l would become:

$$\mathcal{R}_n(l \circ \mathcal{F}) \leq M \mathcal{R}_n(\mathcal{F}) \tag{20}$$

By taking it back, we have:

$$\mathcal{R}_n(l \circ \mathcal{F}) \leq \frac{MA \|Z^k\|_F}{n_k} \tag{21}$$

Therefore, we can have a proportional relationship (lower bound) between the feature Frobenius norms and the number of training samples, with the magnitude of the target labels.

Here we complete our proof. □

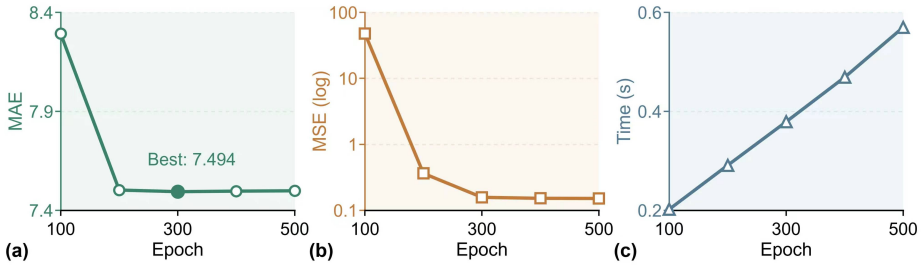

Figure 7: Ablation study on training epochs

D.1 Discussion on the Domain adaptation regression with DIR

Domain adaptation regression (DAR) Chen et al. (2021); Nejjar et al. (2023) is a regression task that aims to transfer knowledge from the source domain to the target domain. DAR has shown that regression tasks are sensitive to the scale of the feature Frobenius norm. Therefore, it stabilizes the scale of the average feature Frobenius norm between the source and the target. Different from our method, DIR focuses on the distribution shift between the training and testing distributions. Meanwhile, the testing data is still not available during the training process in DIR. Therefore, we cannot access the target data during the training phase.

In this paper, we study the feature Frobenius norm in a per-label manner instead of the global average. Our method does not violate the DAR Chen et al. (2021) that stabilizes the feature Frobenius norm, but we want to maintain the same ordinality of the feature Frobenius norm between the source (imbalanced training distribution) and target (balanced testing distribution) as that of the labels, which is an implicit stabilization of the feature Frobenius norm between the source and the target domains. To achieve this goal, we use the feature Frobenius norm regularization. Notably, our feature Frobenius norm regularization does not directly regularize the features. Instead, we use the feature Frobenius norm of the majority-shot samples to regularize the median and few-shot samples' feature Frobenius norm. To the best of our knowledge, our work is the first to directly regularize on the feature Frobenius norm to address the DIR, which demonstrates the novelty of our work.

E Ablation study on training epoch

We conduct an ablation study on the training epochs of the AgeDB-DIR in Fig. 7. As shown in Fig. 7(a), the downstream MAE decreases substantially when the training epochs increase from 100 to 200, dropping from approximately 8.3 to 7.5. Beyond 200 epochs, the performance becomes highly stable, with only marginal variations, and the best MAE of 7.494 is achieved at 300 epochs. This indicates that the MLP can quickly learn the underlying label–feature Frobenius norm relationship and that additional optimization after convergence provides little performance gain. The reason why we need this many epochs is that the training sample size is small and the training pair is easy. It is basically a one-dimensional to one-dimensional mapping, which makes the task easy to learn.

Fig. 7(b) further confirms this convergence behavior from the optimization perspective. The MSE decreases by more than two orders of magnitude between 100 and 300 epochs. Subsequently, it remains nearly unchanged, suggesting that the label-to-feature-norm mapping has been sufficiently fitted. Meanwhile, Fig. 7(c) shows that the computational cost grows approximately linearly with the number of epochs, from around 0.20 s at 100 epochs to 0.57 s at 500 epochs. Therefore, we adopt 300 epochs as the default setting, which achieves the best downstream performance while avoiding unnecessary additional optimization. Overall, these results demonstrate that the proposed MLP fitting procedure is insensitive to the number of training epochs once convergence is reached, while introducing negligible computational overhead.

Method	Extending on the epochs				Base Method MAE
	+5 epoch MAE	+10 epoch MAE	+15 epoch MAE	+20 epoch MAE	
Vanilla Yang et al. (2021)	7.85	7.80	7.84	7.82	7.76
Ranksim Gong et al. (2022)	7.37	7.39	7.37	7.41	7.38
ConR Keramati et al. (2024)	7.49	7.44	7.49	7.46	7.42
SRL Dong et al. (2025)	7.43	7.54	7.46	7.51	7.40

Table 6: Comparison between the different extended epochs MAE

F Ablation study on Performance

To further validate the effectiveness of our method, we extend the training epochs of the previous methods. As shown in Tab. 6, simply extending the training of existing DIR methods does not improve their performance and can even lead to performance degradation. For instance, the MAE of ConR Keramati et al. (2024) increases from 7.42 to 7.49 after 15 additional epochs, while that of SRL Dong et al. (2025) increases from 7.40 to 7.54 after 10 additional epochs. Similar fluctuations or degradation can also be observed for Vanilla and RankSim Gong et al. (2022), indicating that additional optimization under the original training objectives cannot effectively address the remaining label imbalance . Furthermore, it may further reinforce the bias already learned from the imbalanced training distribution. In contrast, our method explicitly leverages the feature Frobenius norm learned from well-fitted many-shot regions to enhance the representations of underrepresented regions, allowing the additional training stage to provide effective performance gains. Therefore, the improvement of our method is attributed to the proposed feature enhancement mechanism rather than simply using more training epochs.

G Code

The code can be found in Here.