PRIME: Deep Imbalanced Regression with Proxies

Jongin Lim [1] **Sucheol Lee** [1] **Daeho Um** [1] **Sung-Un Park** [1] **Jinwoo Shin** [2]

Abstract

Data imbalance remains a fundamental challenge in real-world machine learning. However, most existing work has focused on classification, leaving imbalanced regression underexplored despite its importance in many applications. To address this gap, we propose PRIME, a framework that leverages learnable proxies to construct a balanced and well-ordered feature space for imbalanced regression. At its core, PRIME arranges proxies to be uniformly distributed in the feature space while preserving the ordinal structure of regression targets, and then aligns each sample feature to its corresponding proxy. By using proxies as reference points, PRIME induces the desired structure of learned representations, promoting better generalization, especially in underrepresented target regions. Moreover, since proxy-based alignment resembles classification, PRIME enables the seamless application of class imbalance techniques to regression, facilitating more balanced feature learning. Extensive experiments demonstrate the effectiveness and broad applicability of PRIME, achieving state-of-the-art performance on four real-world regression benchmark datasets across diverse target domains.

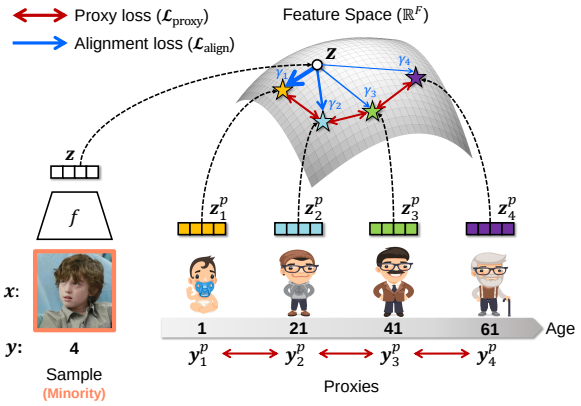


Figure 1. **An overview of PRIME.** Given a sample, PRIME leverages synthetic reference points, termed proxies, to facilitate feature learning. These proxies provide global guidance for effective positioning in the feature space, even for minority samples, enabling the model f to learn balanced and well-ordered representations.

1. Introduction

Data imbalance is a prominent, yet long-standing challenge in most real-world machine learning scenarios (Buda et al., 2018), where certain target values are significantly underrepresented. This imbalance hinders deep models from effectively generalizing to minority groups with limited training samples, driving extensive research efforts to address this challenge (Liu et al., 2019b; Tang et al., 2020; Menon et al., 2021; Zhang et al., 2023b). However, most studies have primarily focused on classification setups, leaving deep imbalanced regression (DIR) underexplored (Yang et al., 2021), despite its significance in various applications.

Unlike classification, regression deals with continuous targets, making it challenging to apply the notion of class imbalance directly. Early works on DIR adapted techniques from imbalanced classification, such as re-weighting (Yang et al., 2021; Steininger et al., 2021) or logit adjustment (Ren et al., 2022), with minor modifications to handle continuous targets. Although intuitive, these methods mainly focus on adjusting loss functions for the final predictions without considering the underlying feature representations. As a result, the learned representations are often fragmented (Zha et al., 2023) and fail to reflect the ordinal relationships of target values (Gong et al., 2022), limiting their effectiveness in real-world applications (Zhang et al., 2023a).

To tackle these issues, recent studies (Gong et al., 2022; Keramati et al., 2024) have explored representation learning approaches for DIR. Specifically, to better reflect the continuous nature of regression targets, these methods impose additional feature regularization terms that encourage samples closer in target space to be positioned closer in feature space. While demonstrating promising results, previous representation learning methods suffer from inherent limitations, as they rely solely on sample relationships within

[1]AI Center, Samsung Electronics [2]Korea Advanced Institute of Science and Technology (KAIST). Correspondence to: Jongin Lim <jonny.lim@samsung.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

individual batches. Due to data imbalance, batches predominantly contain samples with majority targets, causing the learned representations to be biased toward the majority while overlooking or misrepresenting samples with minority targets. Furthermore, representations of minority samples often collapse into those of the majority, which hampers generalization for minority targets. In short, existing representation learning methods remain insufficient for mitigating data imbalance in regression tasks.

In this paper, we propose **P**roxy-based **R**epresentation learning for **IM**balanced r**E**gression (PRIME), a novel representation learning scheme for DIR that effectively addresses the aforementioned limitations. Figure 1 provides an overview of PRIME. Proxies (Movshovitz-Attias et al., 2017) are learnable, synthetic features that serve as representatives of the global feature distribution. The key idea of PRIME is to use proxies as explicit anchors for the desired feature distribution—balanced (*i.e.*, preserving minority features) and well-ordered (*i.e.*, reflecting the ordinality of target values)—and to align sample features with these proxies. To this end, we propose two novel loss functions: proxy loss ($\mathcal{L}_{\text{proxy}}$) and alignment loss ($\mathcal{L}_{\text{align}}$). Specifically, $\mathcal{L}_{\text{proxy}}$ structures the proxies in the feature space to reflect the ordinal relationships of the targets while maintaining sufficient separation to enhance their representative power. Meanwhile, $\mathcal{L}_{\text{align}}$ promotes feature alignment with the corresponding proxy based on target similarity. Unlike prior representation learning methods (Gong et al., 2022; Keramati et al., 2024), PRIME leverages rich sample-proxy relationships to provide holistic supervision for effective feature positioning. By using proxies as reference points, PRIME steers features toward the intended structure for both majority and minority targets, resulting in more generalizable representations.

Furthermore, aligning each feature with its corresponding proxy can be viewed as a classification task, where each proxy serves as a class prototype. This perspective enables PRIME to leverage advances in imbalanced classification to promote balanced feature learning. Indeed, by integrating class imbalance techniques, PRIME further enhances its effectiveness in DIR, bridging the gap between imbalanced regression and classification. To demonstrate its general applicability, we incorporate three widely used methods in imbalanced classification into PRIME: Proxy-wise Re-Weighting (PRW) (Huang et al., 2016), Class-Balanced (CB) loss (Cui et al., 2019), and Label-Distribution-Aware Margin (LDAM) loss (Cao et al., 2019), all of which consistently improve performance on minority targets.

In summary, our contributions are as follows: **(i)** We propose PRIME, a simple yet effective method for learning balanced and well-ordered representations. To the best of our knowledge, PRIME is the first to introduce proxies for imbalanced regression. **(ii)** PRIME enables the application of class imbalance techniques to regression setups, bridging imbalanced regression and classification. **(iii)** We theoretically demonstrate that PRIME provides a bound on the generalization error under balanced test criteria. **(iv)** Extensive experiments demonstrate the effectiveness and broad applicability of PRIME, achieving up to 9.0%, 2.0%, 4.5%, and 3.7% lower regression error on minority targets compared to state-of-the-art methods on AgeDB-DIR, IMDB-WIKI-DIR, NYUD2-DIR, and STS-B-DIR, respectively.

2. Related Work

Imbalanced regression. Early studies (Yang et al., 2021; Steininger et al., 2021) estimate effective label density using kernel density estimation and re-weight samples accordingly. Balanced MSE (Ren et al., 2022) modifies MSE in a manner similar to logit adjustment, while VIR (Wang & Wang, 2023) introduces probabilistic re-weighting to capture prediction uncertainty. However, these methods only focus on the final predictions, which are complementary to our work. RankSim (Gong et al., 2022) and ConR (Keramati et al., 2024) regularize feature representations for imbalanced regression, but are limited to intra-batch relationships, which hinders effective learning of minority features. HCA (Xiong & Yao, 2024) formulates regression as hierarchical classification, incurring higher computational cost and quantization errors. Recently, IM-Context (Nejjar et al., 2024) employs in-context learning with large-scale models such as GPT to handle data imbalance in regression tasks.

Representation learning for regression. Several studies have explored representations tailored for regression. Rank-N-Contrast (Zha et al., 2023) ranks samples and contrasts them based on their relative rankings. Ordinal Entropy (Zhang et al., 2023a) promotes higher-entropy feature space. In addition, contrastive learning approaches (Dufumier et al., 2021a;b; Wang et al., 2022; Schneider et al., 2023; Barbano et al., 2023) have been actively studied. However, these methods overlook the imbalanced target distribution. In contrast, PRIME uses proxies to directly address data imbalance and promote balanced representations.

Proxy learning. Proxies (or prototypes) have been widely studied in deep metric learning (Movshovitz-Attias et al., 2017; Kim et al., 2020; Teh et al., 2020; Lim et al., 2022) and few-shot learning (Snell et al., 2017; Gao et al., 2019; Pan et al., 2019), where each proxy serves as a class representative. Similarly, learnable class centers (Cui et al., 2021; Wang et al., 2021) have been proposed for imbalanced classification, but they do not extend naturally to regression, where handling target-wise proxies is inherently complex. While several studies (Mettes et al., 2019; Dufumier et al., 2021a;b) have explored proxies for continuous values, these methods rely on fixed proxies rather than learning adaptive ones, which distinguishes our approach.

3. Proposed Method

3.1. Problem Definition

We consider a regression problem that predicts the target $\mathbf{y} \in \mathcal{Y}$ based on the input $\mathbf{x} \in \mathcal{X}$, where the underlying data distribution $\mathcal{D}$ is imbalanced. Specifically, we consider the imbalanced training dataset $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ drawn *i.i.d.* from $\mathcal{D}$, where the target distribution $p(\mathbf{y})$ significantly deviates from uniformity. Given $\mathcal{S}$, we aim to train a neural network model $h : \mathcal{X} \to \mathcal{Y}$ composed of a feature encoder $f : \mathcal{X} \to \mathcal{Z}$ and a predictor $g : \mathcal{Z} \to \mathcal{Y}$, where $\mathcal{Z}$ represents the feature space. We denote $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^F$ as the feature of $\mathbf{x}$ and $\hat{\mathbf{y}} = g(\mathbf{z}) \in \mathbb{R}^T$ as the prediction of $\mathbf{y}$. Typically, the encoder f and the predictor g are learned by minimizing a regression loss (*e.g.*, L_1 loss) to ensure the prediction $\hat{\mathbf{y}}$ aligns with the target $\mathbf{y}$. However, the imbalanced target distribution in $\mathcal{S}$ causes the predictions to be biased towards the majority targets. Specifically, the features of minority targets often collapse into those of the majority, leading to higher test errors for minority targets (Yang et al., 2021).

To address this problem, we aim to construct well-ordered feature representations, ensuring that samples closer in $\mathcal{Y}$ are mapped closer in $\mathcal{Z}$. Formally, we define the features $\{\mathbf{z}_i\}_{i=1}^N$ as well-ordered if the following condition holds for all $i, j, k \in [1, N]$: if $d_t(\mathbf{y}_i, \mathbf{y}_j) \leq d_t(\mathbf{y}_i, \mathbf{y}_k)$, then $d_f(\mathbf{z}_i, \mathbf{z}_j) \leq d_f(\mathbf{z}_i, \mathbf{z}_k)$, where $d_t(\cdot, \cdot)$ and $d_f(\cdot, \cdot)$ denote distance metrics defined over $\mathcal{Y}$ and $\mathcal{Z}$, respectively. By encouraging f to encode well-ordered features, we can prevent the features of minority targets from collapsing into those of the majority, enhancing the minority performance.

To this end, we propose PRIME, a simple and effective representation learning scheme for imbalanced regression that introduces synthetic reference points, referred to as proxies (Movshovitz-Attias et al., 2017; Kim et al., 2020). At the core of PRIME, we design proxies to represent a balanced (*i.e.*, uniform target distribution) and well-ordered feature distribution, serving as anchors for representation learning (§3.2). Then, we align features with proxies in accordance with target similarity, providing global guidelines to structure the desired feature space (§3.3). Lastly, we demonstrate that PRIME enables the application of class imbalance techniques to imbalanced regression tasks (§3.4).

3.2. Proxy for Imbalanced Regression

We first introduce proxies, defined as synthetic data points in the product space $\mathcal{Z} \times \mathcal{Y}$. Concretely, we define C proxies as $\mathcal{P} = \{(\mathbf{z}_i^p, \mathbf{y}_i^p)\}_{i=1}^C$, where $\mathbf{z}_i^p \in \mathcal{Z}$ denotes a feature point and $\mathbf{y}_i^p \in \mathcal{Y}$ its corresponding target. Our goal is to design $\mathcal{P}$ to represent a balanced and well-ordered feature distribution. To achieve this, we distribute $\{\mathbf{y}_i^p\}_{i=1}^C$ uniformly across the target values. Specifically, for a scalar target, we compute the minimum (y_{min}) and maximum (y_{max}) values of the targets from $\mathcal{S}$ and define $\{\mathbf{y}_i^p\}_{i=1}^C$ as the $(C+1)$-quantiles of the range $[y_{\text{min}}, y_{\text{max}}]$. For a multi-dimensional target, we can employ K-means clustering to define $\{\mathbf{y}_i^p\}_{i=1}^C$ as the cluster centers. Once $\{\mathbf{y}_i^p\}_{i=1}^C$ are determined, the corresponding feature points $\{\mathbf{z}_i^p\}_{i=1}^C$ are randomly initialized and jointly learned as part of the model parameters.

Now, we formalize our proxy loss $\mathcal{L}_{\text{proxy}}$, which ensures that $\{\mathbf{z}_i^p\}_{i=1}^C$ are well-ordered according to $\{\mathbf{y}_i^p\}_{i=1}^C$. Motivated by stochastic neighbor embedding (Hinton & Roweis, 2002), we define two probability distributions, P and Q, which represent pairwise similarities among $\{\mathbf{y}_i^p\}_{i=1}^C$ and $\{\mathbf{z}_i^p\}_{i=1}^C$, respectively. We then optimize $\{\mathbf{z}_i^p\}_{i=1}^C$ by aligning these two distributions. Specifically, we define the probability distribution $P \in \mathbb{R}^{C \times C}$ to represent pairwise similarities within $\{\mathbf{y}_i^p\}_{i=1}^C$, with its (i, j)-th element defined as:

$$p_{ij} = \frac{e^{-\tau_t d_t(\mathbf{y}_i^p, \mathbf{y}_j^p)}}{\sum_{k \neq l} e^{-\tau_t d_t(\mathbf{y}_k^p, \mathbf{y}_l^p)}}, \tag{1}$$

where $\tau_t > 0$ is the temperature hyperparameter, and we set $p_{ii} = 0$. For nearby targets, p_{ij} is relatively high, while for distant targets, p_{ij} is small. Similarly, we define the probability distribution $Q \in \mathbb{R}^{C \times C}$ to represent pairwise similarities within $\{\mathbf{z}_i^p\}_{i=1}^C$, with its (i, j)-th element defined as:

$$q_{ij} = \frac{e^{-\tau_f d_f(\mathbf{z}_i^p, \mathbf{z}_j^p)}}{\sum_{k \neq l} e^{-\tau_f d_f(\mathbf{z}_k^p, \mathbf{z}_l^p)}}. \tag{2}$$

As before, $\tau_f > 0$ is the temperature hyperparameter, and we set $q_{ii} = 0$. Then, we minimize the Kullback-Leibler divergence between P and Q:

$$D_{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{3}$$

By minimizing (3), $\{\mathbf{z}_i^p\}_{i=1}^C$ are positioned to reflect similarity orders of $\{\mathbf{y}_i^p\}_{i=1}^C$. However, since $\mathcal{Z}$ typically has higher dimensions than $\mathcal{Y}$, trivial solutions (*e.g.*, appending zeros to $\{\mathbf{y}_i^p\}_{i=1}^C$) may arise. To prevent trivial solutions and promote diversity in $\{\mathbf{z}_i^p\}_{i=1}^C$, we introduce a regularization term that encourages features to spread apart. Specifically, we increase the cosine distance between $\mathbf{z}_i^p$ and $\mathbf{z}_j^p$ proportionally to $d_t(\mathbf{y}_i^p, \mathbf{y}_j^p)$. Hence, $\mathcal{L}_{\text{proxy}}$ is defined as:

$$\mathcal{L}_{\text{proxy}} = \sum_{i \neq j} \left\{ p_{ij} \log \frac{p_{ij}}{q_{ij}} - w_{ij} (1 - \cos \theta_{\mathbf{z}_i^p, \mathbf{z}_j^p})^2 \right\}, \tag{4}$$

where $w_{ij} = \alpha d_t(\mathbf{y}_i^p, \mathbf{y}_j^p)$ with $\alpha > 0$. In $\mathcal{L}_{\text{proxy}}$, the first term ensures that proxies are well-ordered, while the second term promotes feature space uniformity (Wang & Isola, 2020), encouraging expressive representations that fully utilize the entire feature space.

3.3. Proxy-based Representation Learning

We leverage the proxy set $\mathcal{P} = \{(\mathbf{z}_i^p, \mathbf{y}_i^p)\}_{i=1}^C$ as explicit anchors to provide informative associations during representation learning. Concretely, given a training sample $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}$, its feature $\mathbf{z}$ is associated with $\{\mathbf{z}_i^p\}_{i=1}^C$. These feature associations are represented by the association vector $\mathbf{A} \in \mathbb{R}^C$, where the j-th element is defined as:

$$\mathbf{A}_j = \frac{e^{-\tau_f d_f(\mathbf{z}, \mathbf{z}_j^p)}}{\sum_{k=1}^C e^{-\tau_f d_f(\mathbf{z}, \mathbf{z}_k^p)}}. \tag{5}$$

Here, $\mathbf{A}_j$ quantifies the association between $\mathbf{z}$ and $\mathbf{z}_j^p$, representing the likelihood that $\mathbf{z}$ would select $\mathbf{z}_j^p$ as its neighbor based on feature similarity. We aim to ensure that such associations are stronger for proxies closer in $\mathcal{Y}$ and weaker for those farther away. To this end, we define $\mathbf{T} \in \mathbb{R}^C$ to represent the target associations between $\mathbf{y}$ and $\{\mathbf{y}_i^p\}_{i=1}^C$, where the j-th element is defined as:

$$\mathbf{T}_j = \frac{e^{-\tau_t d_t(\mathbf{y}, \mathbf{y}_j^p)}}{\sum_{k=1}^C e^{-\tau_t d_t(\mathbf{y}, \mathbf{y}_k^p)}}. \tag{6}$$

To align the feature associations $\mathbf{A}$ with the target associations $\mathbf{T}$, we formalize our alignment loss $\mathcal{L}_{\text{align}}$. Specifically, $\mathcal{L}_{\text{align}}$ is defined as the cross-entropy between $\mathbf{A}$ and $\mathbf{T}$:

$$\mathcal{L}_{\text{align}} = -\sum_{j=1}^C \mathbf{T}_j \log \mathbf{A}_j. \tag{7}$$

In essence, $\mathcal{L}_{\text{align}}$ aligns the sample feature with proxies by pulling $\mathbf{z}$ closer to proxies with similar targets while pushing it away from proxies with dissimilar targets. Since $\{\mathbf{y}_i^p\}_{i=1}^C$ balance the target distribution and $\{\mathbf{z}_i^p\}_{i=1}^C$ are well-ordered with maximal representative power, the proxies serve as global guidelines for structuring the desired feature space, enabling the encoder f to learn more generalizable representations for imbalanced regression.

Finally, the overall loss function $\mathcal{L}_{\text{PRIME}}$ is defined as:

$$\mathcal{L}_{\text{PRIME}}(\mathbf{x}, \mathbf{y}; h, \mathcal{P}) = \mathcal{L}_{\text{reg}} + \lambda_p \mathcal{L}_{\text{proxy}} + \lambda_a \mathcal{L}_{\text{align}}, \tag{8}$$

where $\mathcal{L}_{\text{reg}}$ is the task-specific regression loss, and $\lambda_p > 0$ and $\lambda_a > 0$ are trade-off hyperparameters for $\mathcal{L}_{\text{proxy}}$ and $\mathcal{L}_{\text{align}}$, respectively. Note that the model parameters of h and the proxy features $\{\mathbf{z}_i^p\}_{i=1}^C$ are jointly optimized by minimizing $\mathcal{L}_{\text{PRIME}}$. Furthermore, our PRIME is orthogonal to other imbalanced regression methods and can be seamlessly integrated with existing approaches by simply adding $\mathcal{L}_{\text{proxy}}$ and $\mathcal{L}_{\text{align}}$ to the respective regression loss $\mathcal{L}_{\text{reg}}$.

3.4. Leveraging Class Imbalance Techniques

Although proxies represent a balanced feature distribution, the alignment process in (7) still faces challenges due to sample imbalance. Minority samples, occurring less frequently, often struggle to align properly with their proxies, leading to suboptimal feature representations. Notably, we tackle this issue by leveraging class imbalance techniques.

Attentive readers may notice that PRIME naturally aligns with classification, where each proxy acts as a class center, and $\mathcal{L}_{\text{align}}$ in (7) functions as a classification loss. Hence, any loss-based class imbalance techniques can be seamlessly integrated into our framework. Here, we showcase the application of three widely used techniques, Proxy-wise Re-Weighting (PRW) (Huang et al., 2016), Class-Balanced (CB) loss (Cui et al., 2019), and Label-Distribution-Aware Margin (LDAM) loss (Cao et al., 2019), into PRIME.

PRIME + PRW. Re-weighting (Huang et al., 2016; Wang et al., 2017), which assigns adaptive weights to different classes inversely proportional to their frequency, is the most fundamental approach to addressing class imbalance. PRW adapts this to the proxy setting. Specifically, for each batch, we define the scaling variable s_j for the j-th proxy as $s_j = \frac{C}{N_b} \sum_{i=1}^{N_b} \mathbf{T}_j|_{\mathbf{y}=\mathbf{y}_i}$, where N_b denotes the batch size. Intuitively, s_j represents the proxy frequency, *i.e.*, the number of samples in the batch associated with the j-th proxy. Consequently, the alignment loss with PRW is defined as:

$$\mathcal{L}_{\text{align-PRW}} = -\sum_{j=1}^C \frac{1}{\hat{s_j}} \mathbf{T}_j \log \mathbf{A}_j, \tag{9}$$

where $\hat{s_j} = \max(s_j, \delta_{\text{min}})$, with $\delta_{\text{min}} > 0$ as a hyperparameter to truncate excessively small s_j for stable training.

PRIME + CB. CB loss(Cui et al., 2019) is another seminal work on class imbalance that introduces re-weighting based on the inverse effective number of samples. The alignment loss with CB is given as follows:

$$\mathcal{L}_{\text{align-CB}} = -\sum_{j=1}^C \frac{1-\beta}{1-\beta^{n_j}} \mathbf{T}_j \log \mathbf{A}_j, \tag{10}$$

where $\beta \in [0, 1)$ denotes the effective number parameter and n_j represents the total number of samples belonging to the j-th proxy. We compute n_j by assigning each training sample to the proxy with the closest target. Following (Cui et al., 2019), we set $\beta = 0.99$ for all experiments.

PRIME + LDAM. Margin-based loss functions have also been extensively studied for class imbalance. As the feature association in (5) corresponds to the classification logit, margin-based losses can also be applied without algorithmic changes. LDAM loss (Cao et al., 2019) is one of the most popular margin-based losses, encouraging larger margins for minority classes. The alignment loss formulated with LDAM is defined as follows:

$$\mathcal{L}_{\text{align-LDAM}} = -\sum_{j=1}^C \mathbf{T}_j \log \frac{e^{f_j - \Delta_j}}{e^{f_j - \Delta_j} + \sum_{k \neq j} e^{f_k}}, \tag{11}$$

where $f_j = -\tau_f d_f(\mathbf{z}, \mathbf{z}_j^p)$, and $\Delta_j = M/n_j^{1/4}$, with M as a hyperparameter, denotes the margin for the j-th proxy.

Remark. The use of class imbalance techniques ensures that minority samples receive sufficient alignment focus, ultimately leading to more balanced feature learning. Importantly, PRIME is generalizable and facilitates the use of a wide range of class imbalance techniques, which lays the foundation for future research exploring additional methods. Further discussion is provided in Appendix C.4.

4. Theoretical Analysis

In imbalanced regression, the ultimate goal is to learn a model $h = g \circ f$ that minimizes the expected regression error (or risk) under balanced test criteria, denoted as $\mathcal{R}_{\text{bal}}(h)$. In this section, we prove that optimizing our loss function $\mathcal{L}_{\text{PRIME}}$ in (8) bounds the balanced risk $\mathcal{R}_{\text{bal}}(h)$, supporting the effectiveness of our PRIME.

The balanced risk associated with $\mathcal{L}_{\text{PRIME}}$ is defined as:

$$\mathcal{R}_{\text{bal}}^{\mathcal{L}}(h) = \mathbb{E}_{\text{bal}}\left[\mathcal{L}_{\text{PRIME}}(\mathbf{x}, \mathbf{y}; h, \mathcal{P})\right], \tag{12}$$

where $\mathbb{E}_{\text{bal}}[\cdot]$ represents the expectation over the balanced distribution. Note that, as $\mathcal{L}_{\text{PRIME}}$ accounts for both the regression error and the feature alignment error, it is straightforward to show that $\mathcal{R}_{\text{bal}}(h) \leq \mathcal{R}_{\text{bal}}^{\mathcal{L}}(h)$. Unfortunately, since the balanced distribution is unknown, we can only minimize the empirical risk based on the imbalanced training set $\mathcal{S}$. The empirical risk $\widehat{\mathcal{R}}_{\mathcal{S}}(h)$ is defined as:

$$\widehat{\mathcal{R}}_{\mathcal{S}}(h) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{PRIME}}(\mathbf{x}_i, \mathbf{y}_i; h, \mathcal{P}). \tag{13}$$

Let $\xi : \Omega \to \{1, \ldots, C\}$ be a random variable representing the index of proxy. Note that ξ is a hypothetical random variable with the probability distribution is defined as:

$$p(\xi) = \int q(\xi|\mathbf{y}) p(\mathbf{y}) d\mathbf{y}, \tag{14}$$

where $q(\xi|\mathbf{y})$ represents our probabilistic model for $p(\xi|\mathbf{y})$, and the target association $\mathbf{T}_j$ in (6) offers a natural way to define $q(\xi = j|\mathbf{y})$. We then formalize the skewness of the underlying distribution $\mathcal{D}$, as follows:

$$\mathfrak{C}_{\mathcal{D}} = \max\left\{\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}, \max_{j} \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right\}, \tag{15}$$

where $p_{\text{bal}}(\mathbf{y})$ and $p_{\text{bal}}(\xi)$ denote the balanced distributions (*i.e.*, uniform distributions) over $\mathbf{y}$ and ξ, respectively. The term $\mathfrak{C}_{\mathcal{D}} \geq 1$ quantifies the imbalance in $\mathcal{D}$, taking larger values as the imbalance becomes more severe, and equals 1 when $\mathcal{D}$ is perfectly balanced.

Theorem 4.1. *For any positive $\delta \ll 1$, with probability at least $1 - 2\delta$, the following generalization bound holds for all $h \in \mathcal{H}$ and $f \in \mathcal{F}$:*

$$\mathcal{R}_{\text{bal}}^{\mathcal{L}}(h) \leq \mathfrak{C}_{\mathcal{D}}\left[\widehat{\mathcal{R}}_{\mathcal{S}}(h) + \Phi_{\mathcal{S}}(\mathcal{H}, \delta) + \lambda_a \Phi_{\mathcal{S}}(\mathcal{F}, \delta)\right], \tag{16}$$

where $\Phi_{\mathcal{S}}(\mathcal{H}, \delta)$ and $\Phi_{\mathcal{S}}(\mathcal{F}, \delta)$ represent the empirical Rademacher complexities of $\mathcal{H}$ and $\mathcal{F}$ with some additional terms, respectively.

Proof. Please refer to Appendix A.2. □

Theorem 4.1 confirms that optimizing $\mathcal{L}_{\text{PRIME}}$ provides a bound on the generalization error under a balanced testing distribution. Further analysis can be found in Appendix A.

5. Experiments

5.1. Experimental Setup

Datasets. We conduct experiments on four real-world imbalanced regression benchmarks introduced by (Yang et al., 2021): (i) **AgeDB-DIR** is a facial age estimation dataset derived from AgeDB (Moschoglou et al., 2017). (ii) **IMDB-WIKI-DIR** is an age estimation dataset constructed from IMDB-WIKI (Rothe et al., 2018). (iii) **NYUD2-DIR** is derived from the NYU Depth Dataset V2 (Silberman et al., 2012) for depth prediction from RGB indoor scenes. (iv) **STS-B-DIR** is a natural language dataset based on STS-B (Cer et al., 2017; Wang, 2018), providing continuous similarity scores between pairs of sentences. Detailed descriptions of these datasets are provided in Appendix B.1.

Evaluation metrics. For each dataset, we adopt metrics from (Gong et al., 2022; Keramati et al., 2024). For AgeDB-DIR and IMDB-WIKI-DIR, we use Mean Absolute Error (MAE) and Geometric Mean (GM). For NYUD2-DIR, we adopt Root Mean Squared Error (RMSE) and Threshold Accuracy (δ_1). For STS-B-DIR, we employ Mean Squared Error (MSE) and Pearson correlation. For all datasets, we report results for four subsets: *All*, *Many*, *Median*, and *Few*. *All* refers to the entire test set. Based on the number of training samples per label, *Many* includes labels with over 100 samples, *Median* covers those with 20 to 100 samples, and *Few* consists of labels with fewer than 20 samples.

Baselines. We compare our method against state-of-the-art approaches for DIR, including re-weighting (SQInv and Inv) (Yang et al., 2021), Label Distribution Smoothing (LDS) (Yang et al., 2021), Feature Distribution Smoothing (FDS) (Yang et al., 2021), Balanced MSE (Ren et al., 2022), RankSim (Gong et al., 2022), VIR (Wang & Wang, 2023), ConR (Keramati et al., 2024), HCA (Xiong & Yao, 2024), and IM-Context (Nejjar et al., 2024). For all methods, we use the official implementations when available.

Table 1. **Comparison with state-of-the-art methods on AgeDB-DIR.** † indicates that the results are quoted from the original paper, as the code is not publicly available. The **best** results are marked in bold, while the second best are underlined.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
SQInv (MAE)	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
LDS (Yang et al., 2021)	7.51±0.08	6.93±0.04	8.43±0.22	10.40±0.52	4.80±0.05	4.44±0.05	5.50±0.26	6.98±0.58
FDS (Yang et al., 2021)	7.45±0.09	6.84±0.10	8.52±0.17	10.21±0.22	4.75±0.10	4.37±0.09	5.57±0.27	6.69±0.51
LDS + FDS (Yang et al., 2021)	7.40±0.08	6.82±0.06	8.26±0.17	10.45±0.45	4.70±0.09	4.29±0.06	5.58±0.20	6.97±0.60
Balanced MSE (Ren et al., 2022)	7.60±0.19	7.00±0.29	8.08±0.15	11.96±0.30	4.83±0.15	4.43±0.20	5.34±0.25	8.42±0.08
RankSim (Gong et al., 2022)	7.10±0.05	6.48±0.03	8.19±0.14	10.32±0.14	4.53±0.07	4.10±0.06	5.46±0.16	6.95±0.16
VIR (Wang & Wang, 2023)	7.39±0.05	6.73±0.05	8.42±0.16	10.86±0.28	4.66±0.06	4.22±0.09	5.51±0.14	7.51±0.24
ConR (Keramati et al., 2024)	7.34±0.07	6.74±0.04	8.34±0.31	10.26±0.25	4.73±0.11	4.34±0.07	5.59±0.40	6.80±0.47
HCA† (Xiong & Yao, 2024)	7.45	6.86	8.22	10.90	-	-	-	-
PRIME	7.09±0.08	**6.38**±0.11	8.39±0.26	10.13±0.36	**4.39**±0.08	**3.91**±0.10	5.58±0.22	6.57±0.49
PRIME + PRW	**7.06**±0.09	6.67±0.09	**7.27**±0.25	9.91±0.16	**4.39**±0.08	4.14±0.09	**4.69**±0.20	6.39±0.16
PRIME + CB	7.12±0.09	6.61±0.09	8.07±0.11	**9.29**±0.68	4.47±0.05	4.16±0.08	5.23±0.07	5.81±0.46
PRIME + LDAM	7.24±0.06	6.85±0.14	7.84±0.31	**9.29**±0.44	4.47±0.07	4.26±0.12	4.89±0.25	**5.60**±0.54

Table 2. **Comparison with state-of-the-art methods on IMDB-WIKI-DIR.** † indicates that the results are quoted from the original paper, as the code is not publicly available. The **best** results are marked in bold, while the second best are underlined.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
SQInv (MAE)	7.57±0.04	6.98±0.04	12.23±0.14	23.21±0.13	4.23±0.03	3.99±0.03	6.94±0.13	15.25±0.99
LDS (Yang et al., 2021)	7.75±0.05	7.15±0.05	12.70±0.17	22.77±0.43	4.39±0.06	4.13±0.05	7.43±0.18	14.14±0.67
FDS (Yang et al., 2021)	7.58±0.03	6.98±0.04	12.50±0.12	23.05±0.18	4.25±0.01	3.99±0.01	7.41±0.12	14.89±0.74
LDS + FDS (Yang et al., 2021)	7.75±0.08	7.16±0.08	12.47±0.18	22.80±0.30	4.39±0.08	4.15±0.08	7.17±0.24	14.47±0.23
Balanced MSE (Ren et al., 2022)	7.95±0.12	7.39±0.14	12.27±0.29	23.35±0.67	4.57±0.12	4.34±0.12	7.03±0.27	15.04±0.75
RankSim (Gong et al., 2022)	7.43±0.04	6.85±0.03	12.06±0.24	22.77±0.29	4.14±0.03	3.91±0.02	6.80±0.27	13.47±1.22
VIR (Wang & Wang, 2023)	7.51±0.07	6.90±0.09	12.49±0.46	23.34±0.59	4.16±0.09	3.90±0.10	7.35±0.36	15.73±0.99
ConR (Keramati et al., 2024)	7.45±0.05	6.87±0.04	12.07±0.25	22.78±0.77	4.15±0.05	3.92±0.04	6.77±0.31	14.61±1.41
HCA† (Xiong & Yao, 2024)	7.54	6.91	12.69	22.96	-	-	-	-
PRIME	**7.36**±0.05	**6.73**±0.06	12.48±0.23	23.01±0.92	**3.98**±0.04	**3.73**±0.05	7.17±0.24	14.38±1.16
PRIME + PRW	7.37±0.03	6.74±0.03	**12.04**±0.28	22.34±0.23	4.00±0.05	3.76±0.07	**6.67**±0.34	13.45±1.32
PRIME + CB	7.48±0.01	6.90±0.02	12.05±0.15	22.71±0.42	4.15±0.03	3.91±0.02	6.74±0.16	13.91±0.50
PRIME + LDAM	7.49±0.04	6.91±0.04	12.23±0.18	**22.32**±0.26	4.17±0.06	3.94±0.05	6.94±0.29	**13.44**±0.60

Implementation details. For all experiments, we adopt the benchmark settings of (Yang et al., 2021). To ensure fair comparisons, we use the same backbones and training settings as prior work (*e.g.*, RankSim (Gong et al., 2022) and ConR (Keramati et al., 2024)), tuning only the hyperparameters of PRIME. For AgeDB-DIR and IMDB-WIKI-DIR, we use ResNet50 (He et al., 2016) as the backbone, while for NYUD2-DIR, we adopt the ResNet50-based encoder-decoder architecture (Hu et al., 2019). For STS-B-DIR, we employ BiLSTM + GloVe (Pennington et al., 2014) word embeddings as the feature extractor. Our proxy features $\{\mathbf{z}_i^p\}_{i=1}^C$ are randomly initialized using He initialization (He et al., 2015) and learned jointly with the model parameters. All results are reported as mean and standard deviation over five independent runs. The complete implementation details are provided in Appendix B.2.

5.2. Main Results

Age estimation. Tables 1 and 2 show the overall results on AgeDB-DIR and IMDB-WIKI-DIR, respectively. For fair comparisons, LDS, FDS, RankSim, ConR, and our methods all use the square root inverse (SQInv) re-weighted MAE loss as the regression loss, following the convention of (Yang et al., 2021). Notably, PRIME itself already achieves state-of-the-art performance on both datasets, highlighting the effectiveness of the proposed proxy-based approach. Furthermore, PRW, CB, and LDAM consistently improve performance in the *Median* and *Few* categories.

To further validate the effectiveness of PRIME, we additionally compare it against IM-Context (Nejjar et al., 2024), a recent method that leverages large-scale models (*e.g.*, GPT2 (Garg et al., 2022) and PFN (Müller et al., 2022)) for

Table 3. **Comparison with state-of-the-art methods on NYUD2-DIR.** † indicates that the results are quoted from the original paper, as the code is not publicly available. The **best** results are marked in bold, while the second best are underlined.

Method	RMSE (↓)				δ_1 (↑)			
	All	Many	Median	Few	All	Many	Median	Few
Inv (RMSE)	1.314±0.022	0.751±0.050	0.894±0.056	1.801±0.037	0.687±0.016	0.666±0.033	0.740±0.025	0.688±0.017
LDS	1.386±0.038	**0.690**±0.038	0.887±0.010	1.952±0.067	0.668±0.027	**0.701**±0.030	0.730±0.011	0.612±0.040
FDS	1.343±0.019	0.727±0.044	0.883±0.043	1.865±0.037	0.685±0.014	0.686±0.026	0.749±0.032	0.660±0.023
LDS + FDS	1.335±0.056	0.691±0.051	0.883±0.022	1.865±0.110	0.686±0.010	0.699±0.025	0.743±0.018	0.666±0.036
Balanced MSE	1.307±0.021	0.819±0.027	0.881±0.042	1.761±0.037	0.672±0.014	0.595±0.015	0.808±0.012	0.698±0.020
ConR	1.326±0.030	0.837±0.038	0.885±0.063	1.784±0.079	0.677±0.006	0.604±0.019	0.812±0.021	0.690±0.026
HCA†	1.475	-	-	-	**0.689**	-	-	-
PRIME	1.292±0.020	0.782±0.022	0.881±0.019	1.752±0.044	0.687±0.004	0.624±0.008	0.810±0.005	0.704±0.015
PRIME + PRW	**1.272**±0.032	0.837±0.011	0.920±0.020	**1.682**±0.061	**0.689**±0.003	0.607±0.007	0.814±0.012	**0.724**±0.016
PRIME + CB	1.295±0.032	0.823±0.020	0.900±0.034	1.734±0.066	0.685±0.005	0.605±0.008	0.819±0.021	0.712±0.027
PRIME + LDAM	1.302±0.009	0.807±0.030	**0.871**±0.032	1.758±0.037	0.682±0.002	0.613±0.011	**0.822**±0.013	0.698±0.017

Table 4. **Comparison with state-of-the-art methods on STS-B-DIR.** The **best** results are in bold, while the second best are underlined.

Method	MSE (↓)				Pearson correlation (↑)			
	All	Many	Median	Few	All	Many	Median	Few
Inv (MSE)	1.298±0.072	1.300±0.099	1.281±0.090	1.319±0.068	0.628±0.016	0.603±0.019	0.596±0.015	0.663±0.016
LDS	0.990±0.038	0.931±0.052	1.270±0.048	0.954±0.020	0.742±0.013	0.703±0.015	0.701±0.018	0.766±0.007
FDS	1.262±0.091	1.254±0.147	1.274±0.217	1.316±0.064	0.606±0.015	0.592±0.027	0.612±0.014	0.665±0.007
LDS + FDS	0.974±0.007	0.929±0.008	**1.161**±0.030	0.983±0.051	0.747±0.003	0.709±0.003	0.709±0.003	0.755±0.017
RankSim	0.980±0.014	0.928±0.024	1.208±0.088	0.985±0.025	0.745±0.002	0.707±0.004	0.702±0.014	0.756±0.009
PRIME	0.970±0.004	0.894±0.012	1.325±0.062	0.930±0.035	0.750±0.003	0.712±0.003	0.710±0.010	0.773±0.010
PRIME + PRW	**0.967**±0.004	**0.885**±0.010	1.351±0.061	0.925±0.017	**0.753**±0.002	**0.715**±0.003	**0.711**±0.011	0.775±0.006
PRIME + CB	0.980±0.008	0.906±0.010	1.335±0.079	0.922±0.028	0.748±0.001	0.708±0.001	**0.711**±0.004	0.777±0.009
PRIME + LDAM	0.975±0.016	0.893±0.006	1.366±0.084	**0.919**±0.053	0.751±0.003	0.712±0.003	0.709±0.014	**0.778**±0.015

in-context learning in regression. Following IM-Context, we adopt the pre-trained CLIP image encoder (ViT-B/32) (Radford et al., 2021) as the backbone, and fine-tune it jointly with a two-layer MLP regression head using our PRIME loss. To ensure robustness, we report the average performance of PRIME over five independent runs. As shown in Table 5, PRIME substantially outperforms both PFN-localized and GPT2-localized across all evaluation metrics. These results confirm that PRIME remains effective even on top of a strong pre-trained backbone model, highlighting its compatibility with powerful feature extractors.

Depth estimation. Table 3 presents the depth estimation results on NYUD2-DIR, a more challenging setting where the high-dimensional target space exhibits non-linear relationships. Following ConR, we measure target similarity based on the difference between the average depth values and use Balanced MSE as the regression loss. PRIME outperforms Balanced MSE and ConR, demonstrating its effectiveness even for complex targets. Notably, PRIME significantly improves performance in the *Few* category, underscoring its ability to mitigate data imbalance. Furthermore, incorporating PRW, CB, and LDAM further enhances minority performance, achieving state-of-the-art results across the *All*, *Median*, and *Few* categories.

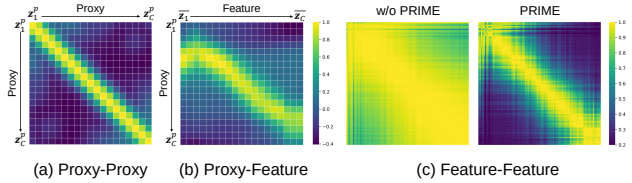
Text similarity estimation. Table 4 reports the performance on STS-B-DIR. Since STS-B-DIR exhibits a highly discrete target distribution (Yang et al., 2021), we smooth it using LDS and employ inverse (INV) re-weighted MSE loss as the regression loss, following RankSim. Overall, PRIME and its variants achieve state-of-the-art results, demonstrating their effectiveness across diverse target domains.

5.3. Analysis

PRIME facilitates effective feature learning. We investigate the effect of proxies on feature learning. Figure 2 illustrates feature space similarities between the learned proxies and the features of the test samples in AgeDB-DIR. For clarity, data points are sorted by their target values, with the expectation that matrix values gradually decrease from the diagonal to the periphery. As shown in Figure 2(a), thanks to $\mathcal{L}_{\text{proxy}}$, the proxies are well-ordered in the feature space according to their target values. Figure 2(b) confirms that features are generally well aligned with their corresponding

Table 5. **Comparison with IM-Context on AgeDB-DIR and IMDB-WIKI-DIR.** IM-Context results are taken from the original paper. Both PRIME and the two IM-Context variants (PFN-localized and GPT2-localized) use the CLIP image encoder (ViT-B/32) as their feature extractor. Under the same backbone, PRIME achieves consistently superior performance.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
Results for AgeDB-DIR:								
PFN-localized (Nejjar et al., 2024)	6.58	5.61	8.49	10.49	4.29	3.58	6.30	8.19
GPT2-localized (Nejjar et al., 2024)	6.05	5.67	6.71	7.83	3.79	3.59	4.17	4.90
PRIME	**5.47**±0.03	**5.46**±0.08	**5.48**±0.23	**5.57**±0.35	**3.48**±0.05	**3.45**±0.07	**3.64**±0.13	**3.35**±0.27
Results for IMDB-WIKI-DIR:								
PFN-localized (Nejjar et al., 2024)	8.96	8.71	10.79	16.33	5.26	5.17	6.00	9.42
GPT2-localized (Nejjar et al., 2024)	7.76	7.35	11.15	17.71	4.29	4.13	5.96	11.00
PRIME	**6.42**±0.03	**5.98**±0.05	**9.92**±0.32	**16.28**±0.57	**3.49**±0.02	**3.33**±0.04	**5.17**±0.32	**9.41**±0.51



(a) Proxy-Proxy (b) Proxy-Feature (c) Feature-Feature

Figure 2. **Feature space similarities on AgeDB-DIR.** (a) Similarity matrix among proxies. (b) Similarity matrix between proxies and the means of their associated features. (c) Similarity matrices among features, with (right) and without (left) PRIME.

proxies[1]. The plot on the left of Figure 2(c) shows that, without PRIME, the features are poorly ordered, and the model fails to learn effective representations—particularly for minority targets (*i.e.*, those at both ends of the matrix diagonal). As shown in the right plot of Figure 2(c), incorporating proxies allows PRIME to guide the features towards the intended structure for both majority and minority targets, resulting in more balanced and well-ordered representations.

PRIME ensures well-ordered representations in the *Few* category. To assess how well PRIME captures the ordinality of target values, we evaluate the Spearman correlation between feature and label similarity matrices on the AgeDB-DIR test set. A higher correlation suggests that the learned features more faithfully reflect the ordinal structure of the label space, which is an essential property for effective regression. As shown in Table 6, PRIME achieves consistently strong correlations across *All* samples and, notably, maintains a high correlation in the *Few* category. In contrast, RankSim and ConR exhibit marked degradation in this underrepresented regime. These results underscore the strength of our proxy-based formulation, which offers holistic guidance for feature positioning and enables minority samples to align with the overall label structure.

[1]See Appendix C.1 for a discussion on the twisted pattern in the top left of Figure 2(b).

Table 6. **Spearman correlation between feature and label similarities on AgeDB-DIR.** A higher correlation indicates better alignment between learned features and targets, implying more well-ordered representations.

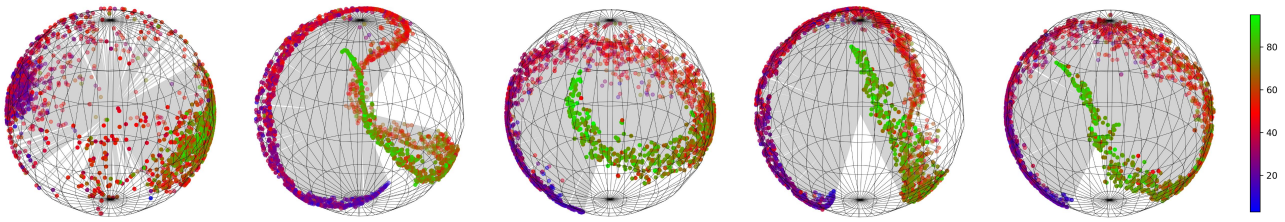
Method	All	Few
RankSim (Gong et al., 2022)	0.804±0.008	0.587±0.036
ConR (Keramati et al., 2024)	0.790±0.024	0.614±0.043
PRIME	**0.942**±0.008	**0.828**±0.020

Table 7. **Comparison with representation learning methods for general regression.** See Appendix C.1 for the complete results.

Method	MAE (↓)	GM (↓)
HPN (Mettes et al., 2019)	7.38±0.08	4.66±0.09
Ordinal Entropy (Zhang et al., 2023a)	7.33±0.08	4.68±0.07
Rank-N-Contrast (Zha et al., 2023)	7.27±0.05	4.69±0.04
PRIME	**7.09**±0.08	**4.39**±0.08

Comparison with recent representation learning methods for general regression. To further demonstrate the effectiveness of PRIME, we compare it with three recent techniques proposed for general regression: HPN (Mettes et al., 2019), Ordinal Entropy (Zhang et al., 2023a), and Rank-N-Contrast (Zha et al., 2023). As shown in Table 7, PRIME achieves clear margins over the compared baselines. In particular, compared to HPN, which uses fixed prototypes on a hypersphere, PRIME achieves significantly better performance, verifying the effectiveness of our proxy design tailored for DIR. Notably, PRIME outperforms Ordinal Entropy and Rank-N-Contrast—two leading representation learning methods for general regression—highlighting its ability to learn robust and reliable representations under imbalanced target distributions.

Effectiveness of our proxy formulation. To validate the effectiveness of our proxy formulation, we compare PRIME



(a) ConR (b) PRIME (c) PRIME + PRW (d) PRIME + CB (e) PRIME + LDAM

Figure 3. **Feature visualization with t-SNE on AgeDB-DIR.** By leveraging proxies as global reference points, PRIME clearly demonstrates well-ordered features with fewer minority feature collapses, effectively capturing the continuity of target values.

Table 8. **Comparison with proxy-based alternatives on AgeDB-DIR.** See Appendix C.1 for the complete results.

Method	MAE (↓)	GM (↓)
ProxyNCA (Movshovitz-Attias et al., 2017)	7.33±0.08	4.64±0.06
Non-learnable (centroid)	7.10±0.05	4.53±0.07
PRIME	**7.09**±0.08	**4.39**±0.08

Table 9. **Ablation study on AgeDB-DIR.** See Appendix C.1 for the complete results.

	Init. $\mathbf{y}_i^p$	$\mathcal{L}_{\text{proxy}}$	Align	MAE (↓)	GM (↓)
M1	Random	-	One-Hot	7.34±0.08	4.66±0.07
M2	Random	✓	One-Hot	7.24±0.09	4.61±0.12
M3	Unif.	✓	One-Hot	7.23±0.12	4.55±0.12
M4	Unif.	✓	Eq. (6)	**7.09**±0.08	**4.39**±0.08

with two proxy-based alternatives: ProxyNCA (Movshovitz-Attias et al., 2017) and a non-learnable variant of PRIME. For ProxyNCA, we adapt the original method to the regression setup by assigning proxies so that associated targets are uniformly distributed, as in PRIME. For the non-learnable variant, proxy features are updated as the centroids of sample features assigned to each proxy, rather than learned. As shown in Table 8, PRIME consistently outperforms both ProxyNCA and the non-learnable variant, attributed to its formulation that optimizes learnable proxies to preserve the ordinal structure of the target space. This leads to more stable and effective feature representations, particularly in data-sparse regions.

Ablation study. As shown in Table 9, we ablate three key design choices: (i) whether to randomly initialize $\{\mathbf{y}_i^p\}_{i=1}^C$ or assign them uniformly in the target space, (ii) whether to include $\mathcal{L}_{\text{proxy}}$, and (iii) whether to perform feature alignment as in (6) or align features to their nearest proxy using a one-hot strategy. Comparing the ablation models, M1 to M2 shows a significant performance gain, demonstrating the effectiveness of $\mathcal{L}_{\text{proxy}}$. From M2 to M3, assigning proxies uniformly in the target space provides a slight improvement over random initialization. Finally, from M3 to M4 (PRIME), performance further improves, as regression deals with continuous targets, making distance proportional assignment as in (6), more effective than one-hot encoding.

Feature visualization. Figure 3 presents t-SNE (Van der Maaten & Hinton, 2008) visualizations of the learned representations from the AgeDB-DIR test set. As shown in Figure 3(a), ConR is biased toward learning discriminative representations only for majority targets (red), failing to capture the continuity of target values. Moreover, minority features (blue and green) collapse into the majority (red), leading to poor performance for minority targets. In Figure 3(b), PRIME produces well-ordered features with fewer minority feature collapses, effectively capturing the continuity of target values. Figure 3(b)-(d) shows that incorporating class imbalance techniques further promotes balanced feature learning, leading to more structured representations.

Computational efficiency (Appendix C.2). We confirm that PRIME offers training efficiency comparable to other imbalanced regression methods.

Hyperparameter sensitivity (Appendix C.3). We analyze the impact of PRIME's hyperparameters (C, λ_p, λ_a, τ_f, τ_t, and α). Overall, PRIME demonstrates reliable and robust performance across a wide range of hyperparameter choices.

6. Conclusion

We introduce PRIME, a novel representation learning framework for imbalanced regression that leverages proxies to learn balanced and well-ordered feature representations. By using proxies as global reference points, PRIME facilitates effective feature learning for both majority and minority targets. Theoretical analysis and extensive experiments on four benchmark datasets spanning diverse target domains demonstrate its effectiveness. Furthermore, PRIME seamlessly integrates with any loss-based class imbalance technique. We believe our work provides a flexible and unified framework for incorporating various class imbalance techniques into regression problems, introducing a new paradigm for addressing imbalanced regression.

Impact Statement

This paper presents work whose goal is to advance the field of imbalanced regression. Imbalanced regression can have societal implications, particularly in applications where accurate predictions across the entire target range are critical. For instance, in social sciences, models trained on imbalanced data may disproportionately favor well-represented groups while yielding less reliable predictions for underrepresented populations. Such biases can exacerbate existing inequalities, leading to unfair decision-making. Our work can mitigate these issues and contribute to more equitable predictive modeling.

References

Barbano, C. A., Dufumier, B., Duchesnay, E., Grangetto, M., and Gori, P. Contrastive learning for regression in multi-site brain age prediction. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2023.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Cao, Y., Wu, Z., and Shen, C. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Dufumier, B., Gori, P., Victor, J., Grigis, A., and Duchesnay, E. Conditional alignment and uniformity for contrastive learning with continuous proxy labels. *arXiv preprint arXiv:2111.05643*, 2021a.

Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., Mcdonald, C., Piguet, C. M., et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pp. 58–68. Springer, 2021b.

Gao, T., Han, X., Liu, Z., and Sun, M. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6407–6414, 2019.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Gong, Y., Mori, G., and Tung, F. Ranksim: Ranking similarity regularization for deep imbalanced regression. In *International Conference on Machine Learning*, pp. 7634–7649. PMLR, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G. E. and Roweis, S. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Hu, J., Ozay, M., Zhang, Y., and Okatani, T. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1043–1051. IEEE, 2019.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384, 2016.

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

Kim, S., Kim, D., Cho, M., and Kwak, S. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3238–3247, 2020.

Lim, J., Yun, S., Park, S., and Choi, J. Y. Hypergraph-induced semantic tuplet loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 212–222, 2022.

Liu, L., Lu, H., Xiong, H., Xian, K., Cao, Z., and Shen, C. Counting objects by blockwise classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30 (10):3513–3527, 2019a.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019b.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Mettes, P., Van der Pol, E., and Snoek, C. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019.

Mohri, M. Foundations of machine learning, 2018.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 51–59, 2017.

Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., and Singh, S. No fuss distance metric learning using proxies. In *Proceedings of the IEEE international conference on computer vision*, pp. 360–368, 2017.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

Nejjar, I., Ahmed, F., and Fink, O. Im-context: In-context learning for imbalanced regression tasks. *Transactions on Machine Learning Research*, 2024.

Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.-W., and Mei, T. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2239–2247, 2019.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ren, J., Zhang, M., Yu, C., and Liu, Z. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7926–7935, 2022.

Rothe, R., Timofte, R., and Van Gool, L. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 10–15, 2015.

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

Schneider, S., Lee, J. H., and Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Steininger, M., Kobs, K., Davidson, P., Krause, A., and Hotho, A. Density-based weighting for imbalanced regression. *Machine Learning*, 110:2187–2211, 2021.

Tang, K., Huang, J., and Zhang, H. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in neural information processing systems*, 33:1513–1524, 2020.

Teh, E. W., DeVries, T., and Taylor, G. W. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 448–464. Springer, 2020.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, A. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Wang, P., Han, K., Wei, X.-S., Zhang, L., and Wang, L. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 943–952, 2021.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., and Li, T. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19376–19385, 2022.

Wang, Y.-X., Ramanan, D., and Hebert, M. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

Wang, Z. and Wang, H. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. *Advances in Neural Information Processing Systems*, 36:30429–30452, 2023.

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

Yang, Y., Zha, K., Chen, Y., Wang, H., and Katabi, D. Delving into deep imbalanced regression. In *International conference on machine learning*, pp. 11842–11851. PMLR, 2021.

Zha, K., Cao, P., Son, J., Yang, Y., and Katabi, D. Rank-n-contrast: Learning continuous representations for regression. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 17882–17903. Curran Associates, Inc., 2023.

Zhang, S., Yang, L., Mi, M. B., Zheng, X., and Yao, A. Improving deep regression with ordinal entropy. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=raU07GpP0P.

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023b.

Appendix

The Appendix includes additional descriptions, experimental results, and analyses omitted from the main manuscript due to space constraints. In Section A, we present a detailed theoretical analysis. In Section B, we provide further details on the experimental setup. In Section C, we report additional experimental results and analyses.

A. Detailed Theoretical Analysis

In this section, we provide theoretical justifications for the proposed method. In §A.1, we provide a complete description of the definitions and notations used in this analysis. In §A.2, we present the proof of Theorem 4.1 from the main manuscript. In §A.3, we present an extended theoretical analysis of Theorem 4.1 under non-optimal proxy settings. In §A.4, we provide further theoretical evidence for the claim that the use of class imbalance techniques promotes balanced feature learning.

A.1. Definitions and Notations

We first rigorously clarify the underlying probabilistic distributions used throughout the paper. Specifically, while $\mathbf{x}$ and $\mathbf{y}$ are random variables with a given probability density function $p(\mathbf{x}, \mathbf{y})$, the corresponding proxy index ξ is a hypothetical random variable for which we need to define a probability distribution. Therefore, we define our joint probability density function as follows:

$$p(\mathbf{x}, \mathbf{y}, \xi) := p(\mathbf{x}|\mathbf{y})q(\xi|\mathbf{y})p(\mathbf{y}), \tag{17}$$

where $q(\xi|\mathbf{y})$ is our probabilistic modeling function of $p(\xi|\mathbf{y})$. For instance, in PRIME, we take

$$q(\xi = j|\mathbf{y}) = \mathbf{T}_j = \frac{e^{-\tau_t d_t(\mathbf{y}, \mathbf{y}_j^p)}}{\sum_{k=1}^{C} e^{-\tau_t d_t(\mathbf{y}, \mathbf{y}_k^p)}} \qquad \text{for } j = 1, \ldots, C. \tag{18}$$

Next, the balanced distributions for $\mathbf{y}$ and ξ are defined as follows:

$$p_{\text{bal}}(\mathbf{x}, \mathbf{y}) := p(\mathbf{x}|\mathbf{y})p_{\text{bal}}(\mathbf{y}), \tag{19}$$

$$p_{\text{bal}}(\mathbf{x}, \mathbf{y}, \xi) := p(\mathbf{x}, \mathbf{y}|\xi)p_{\text{bal}}(\xi). \tag{20}$$

Here, $p_{\text{bal}}(\mathbf{y})$ and $p_{\text{bal}}(\xi)$ correspond to the uniform distributions over the spaces $\mathcal{Y}$ and $\{1, \ldots, C\}$, respectively.

We assume that the proxies are optimally positioned[2], *i.e.*, $\mathcal{L}_{\text{proxy}} = 0$. This assumption is justified, as the proxies can be pre-optimized using (4) prior to model training. Hence, $\mathcal{L}_{\text{PRIME}}$ in (8) simplifies to $\mathcal{L}_{\text{reg}} + \lambda_a \mathcal{L}_{\text{align}}$. Then, our balanced risk is defined as the weighted sum of the balanced regression risk and the balanced alignment risk:

$$\mathcal{R}_{\text{bal}}^{\mathcal{L}}(h) := \underbrace{\int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p_{\text{bal}}(\mathbf{x}, \mathbf{y})}_{\text{balanced regression risk } (i.e., \mathcal{R}_{\text{bal}}(h))} + \lambda_a \underbrace{\int (-\log p_\theta(\xi|\mathbf{x}))\, p_{\text{bal}}(\mathbf{x}, \mathbf{y}, \xi) d\mathbf{x} d\mathbf{y} d\xi}_{\text{balanced alignment risk}}, \tag{21}$$

where $p_\theta(\xi|\mathbf{x})$ is our feature association with proxy, which we take

$$p_\theta(\xi = j|\mathbf{x}) = A_j = \frac{e^{-\tau_f d_f(\mathbf{z}, \mathbf{z}_j^p)}}{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z}, \mathbf{z}_k^p)}} \qquad \text{for } j = 1, \ldots, C \tag{22}$$

in our PRIME model. Note that the balanced risk with respect to $\mathcal{R}_{\text{bal}}^{\mathcal{L}}(h)$ is always greater than the balanced regression risk $\mathcal{R}_{\text{bal}}(h)$ because the balanced alignment risk is always positive.

[2]We later relax this assumption and extend the analysis to non-optimal proxy settings; see Section A.3.

A.2. Proof of Theorem 4.1

Theorem A.1 (Full description of Theorem 4.1)**.** *For any positive $\delta \ll 1$, with probability at least $1-2\delta$, the following generalization bound holds for all $h \in \mathcal{H}$ and $f \in \mathcal{F}$:*

$$\mathcal{R}^{\mathcal{L}}_{bal}(h) \leq \mathfrak{C}_{\mathcal{D}} \Bigg[\underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{reg}(g \circ f(\mathbf{x}_i), \mathbf{y}_i) - \lambda_a \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} q(\xi = j|\mathbf{y}_i) \log p_\theta(\xi = j|\mathbf{x}_i)}_{\widehat{\mathcal{R}}_{\mathcal{S}}(h)} + \underbrace{2\mu_{\mathcal{L}_{reg}} \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) + 3M_{\mathcal{L}_{reg}} \sqrt{\frac{\log \frac{2}{\delta}}{2n}}}_{\Phi_{\mathcal{S}}(\mathcal{H},\delta)} + \lambda_a \underbrace{2\mu_L \hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3M_L \sqrt{\frac{\log \frac{2}{\delta}}{2n}}}_{\Phi_{\mathcal{S}}(\mathcal{F},\delta)} \Bigg],$$

where $\mathfrak{C}_{\mathcal{D}} := \max\left\{\sup_{\mathbf{y}} \frac{p_{bal}(\mathbf{y})}{p(\mathbf{y})}, \max_j \frac{p_{bal}(\xi=j)}{p(\xi=j)}\right\}$, $\mu_{\mathcal{L}_{reg}}$ and μ_L are Lipschitz continuity of $\mathbf{y} \mapsto \mathcal{L}_{reg}(\mathbf{y}, \mathbf{y}')$ for all $\mathbf{y} \in \mathcal{Y}$ for any fixed $\mathbf{y}' \in \mathcal{Y}$ and $\mathbf{x} \mapsto \sum_{j=1}^{C} -q(\xi = j|\mathbf{y}') \log p_\theta(\xi = j|\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ for any fixed $\mathbf{y}' \in \mathcal{Y}$, respectively, $M_{\mathcal{L}_{reg}}$ and M_L are constants satisfying $\mathcal{L}_{reg}(\mathbf{y}, \mathbf{y}') < M_{\mathcal{L}_{reg}}$ for all $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$ and $\sum_{j=1}^{C} -q(\xi = j|\mathbf{y}) \log p_\theta(\xi = j|\mathbf{x}) < M_L$ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, respectively, and $\widehat{\mathfrak{R}}_{\mathcal{S}}$ denotes the Rademacher complexity.

Proof.

$$\begin{aligned}
\mathcal{R}^{\mathcal{L}}_{\text{bal}}(h) &= \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p_{\text{bal}}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + \lambda_a \int \left(-\log p_\theta(\xi|\mathbf{x})\right) p_{\text{bal}}(\mathbf{x}, \mathbf{y}, \xi) d\mathbf{x} d\mathbf{y} d\xi \\
&= \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}|\mathbf{y}) p_{\text{bal}}(\mathbf{y}) d\mathbf{x} d\mathbf{y} + \lambda_a \int \left(-\log p_\theta(\xi|\mathbf{x})\right) p(\mathbf{x}, \mathbf{y}|\xi) p_{\text{bal}}(\xi) d\mathbf{x} d\mathbf{y} d\xi \\
&\leq \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \int \left(-\log p_\theta(\xi|\mathbf{x})\right) p(\mathbf{x}, \mathbf{y}|\xi) p(\xi) d\mathbf{x} d\mathbf{y} d\xi \\
&= \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \int -\log p_\theta(\xi|\mathbf{x}) p(\mathbf{x}, \mathbf{y}, \xi) d\mathbf{x} d\mathbf{y} d\xi \\
&= \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \int -q(\xi|\mathbf{y}) \log p_\theta(\xi|\mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} d\xi \\
&= \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \int -\sum_{\xi} q(\xi|\mathbf{y}) \log p_\theta(\xi|\mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y},
\end{aligned}$$

where the second-to-last equality follows from the definition, $p(\mathbf{x}, \mathbf{y}, \xi) = p(\mathbf{x}|\mathbf{y}) q(\xi|\mathbf{y}) p(\mathbf{y}) = p(\mathbf{x}, \mathbf{y}) q(\xi|\mathbf{y})$. Hence, by applying Theorem 11.3 in (Mohri, 2018) to the regression risk and alignment risk separately, the following inequality holds

with probability at least $1-2\delta$,

$$R^{\mathcal{L}}_{\text{bal}}(h) \leq \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{reg}}(g\circ f(\mathbf{x}_i),\mathbf{y}_i) + 2\mu_{\mathcal{L}_{\text{reg}}}\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}) + 3M_{\mathcal{L}_{\text{reg}}}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right)$$
$$+\left(\max_{j}\frac{p_{\text{bal}}(\xi=j)}{p(\xi=j)}\right)\lambda_a\left(-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}q(\xi=j|\mathbf{y}_i)\log p_\theta(\xi=j|\mathbf{x}_i) + 2\mu_L\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 3M_L\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right).$$

Then, the rest of the proof follows directly from the definition of $\mathfrak{C}_{\mathcal{D}}$. □

Remark. Theorem 4.1 covers various algorithms, such as PRIME, PRIME + PRW, PRIME+CB, and PRIME+LDAM, by modeling $q(\xi|\mathbf{y})$ and $p_\theta(\xi|\mathbf{x})$ appropriately.

A.3. Extension to Non-optimal Proxy Settings

We now extend the theoretical analysis to settings with non-optimal proxies, where the learned proxy features deviate from their optimal positions due to approximation errors.

Let $\{\tilde{\mathbf{z}}^p_j\}_{j=1,\dots,C}$ denote the optimal proxy features that minimize $\mathcal{L}_{\text{proxy}}$, and define the corresponding feature association as $\tilde{p}_\theta(\xi|\mathbf{x}) := \frac{e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_j)}}{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k)}}$ for all $j=1,\dots,C$. Then, we define the learned proxy features as $\mathbf{z}^p_j := \tilde{\mathbf{z}}^p_j + \epsilon_j$ for all $j=1,\dots,C$, where ϵ_j represents the estimation error, and $p_\theta(\xi|\mathbf{x}) := \frac{e^{-\tau_f d_f(\mathbf{z},\mathbf{z}^p_j)}}{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z},\mathbf{z}^p_k)}}$ denotes the corresponding feature association with respect to the learned proxies.

To analyze the non-optimal case, we revisit the balanced alignment risk term in (21), originally derived under the assumption of optimally positioned proxies, and rewrite $-\log\tilde{p}_\theta(\xi|\mathbf{x})$ using the following identity:

$$-\log\tilde{p}_\theta(\xi|\mathbf{x}) = -\log p_\theta(\xi|\mathbf{x}) + \left(\log p_\theta(\xi|\mathbf{x}) - \log\tilde{p}_\theta(\xi|\mathbf{x})\right).$$

Based on this formulation, the first term, $-\log p_\theta(\xi|\mathbf{x})$, follows the same derivation as in the proof of Theorem 4.1. The second term, $\log p_\theta(\xi|\mathbf{x}) - \log\tilde{p}_\theta(\xi|\mathbf{x})$, quantifies the discrepancy arising from the deviation between the learned and optimal proxies. This discrepancy term can be bounded using the following inequality:

$$\begin{aligned}
\log p_\theta(\xi|\mathbf{x}) - \log\tilde{p}_\theta(\xi|\mathbf{x}) =& \log\frac{e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi+\epsilon_\xi)}}{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k+\epsilon_k)}} - \log\frac{e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi)}}{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k)}}\\
=&\tau_f\left(d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi) - d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi+\epsilon_\xi)\right) + \log\frac{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k)}}{\sum_{k=1}^{C} e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k+\epsilon_k)}}\\
\leq&\tau_f\left(d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi) - d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi+\epsilon_\xi)\right) + \log\max_k\left\{\frac{e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k)}}{e^{-\tau_f d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k+\epsilon_k)}}\right\}\\
=&\tau_f\left(d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi) - d_f(\mathbf{z},\tilde{\mathbf{z}}^p_\xi+\epsilon_\xi)\right) + \max_k\tau_f\left(d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k+\epsilon_k) - d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k)\right)\\
\leq&2\tau_f\max_k\left|d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k+\epsilon_k) - d_f(\mathbf{z},\tilde{\mathbf{z}}^p_k)\right|.
\end{aligned}$$

Consequently, we obtain the following upper bound on the desired balanced risk:

$$
\begin{aligned}
\tilde{\mathcal{R}}^{\mathcal{L}}_{\text{bal}}(h) &:= \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p_{\text{bal}}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + \lambda_a \int \left(-\log \tilde{p}_\theta(\xi|\mathbf{x})\right) p_{\text{bal}}(\mathbf{x}, \mathbf{y}, \xi) d\mathbf{x} d\mathbf{y} d\xi \\
&\leq \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \int -\sum_\xi q(\xi|\mathbf{y}) \log \tilde{p}_\theta(\xi|\mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&= \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \int \sum_\xi q(\xi|\mathbf{y}) \left(-\log p_\theta(\xi|\mathbf{x}) + (\log p_\theta(\xi|\mathbf{x}) - \log \tilde{p}_\theta(\xi|\mathbf{x}))\right) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\leq \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \left(\int -\sum_\xi q(\xi|\mathbf{y}) \log p_\theta(\xi|\mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \right. \\
&\qquad \left. + \int 2\tau_f \max_k |d_f(\mathbf{z}, \tilde{\mathbf{z}}^p_k + \epsilon_k) - d_f(\mathbf{z}, \tilde{\mathbf{z}}^p_k)| \, p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \right).
\end{aligned}
$$

If d_f is a norm, we can apply the triangle inequality to simplify the last term of the inequality:

$$
\begin{aligned}
\tilde{\mathcal{R}}^{\mathcal{L}}_{\text{bal}}(h) &\leq \left(\sup_{\mathbf{y}} \frac{p_{\text{bal}}(\mathbf{y})}{p(\mathbf{y})}\right) \int \mathcal{L}_{\text{reg}}(g \circ f(\mathbf{x}), \mathbf{y}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&\quad + \left(\max_j \frac{p_{\text{bal}}(\xi = j)}{p(\xi = j)}\right) \lambda_a \left(\int -\sum_\xi q(\xi|\mathbf{y}) \log p_\theta(\xi|\mathbf{x}) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + 2\tau_f \max_k d_f(\tilde{\mathbf{z}}^p_k + \epsilon_k, \tilde{\mathbf{z}}^p_k) \right).
\end{aligned}
$$

Remark. Importantly, as training progresses and the proxies become more accurate (*i.e.*, $\boldsymbol{\epsilon}_k$ becomes smaller), the residual term decreases accordingly, resulting in a tighter bound. Empirically, we also observe that PRIME performs robustly even when the proxies are randomly initialized.

A.4. Further Theoretical Insight

Another perspective offered by Theorem 4.1 is that $\mathcal{R}^{\mathcal{L}}_{\text{bal}}(h)$ is bounded by a constant multiple of $\mathfrak{C}_{\mathcal{D}}$. A higher value of $\mathfrak{C}_{\mathcal{D}}$ leads to a larger deviation between $\mathcal{R}^{\mathcal{L}}_{\text{bal}}(h)$ and $\widehat{\mathcal{R}}_{\mathcal{S}}(h)$. Intuitively, incorporating data skewness into an effective balancing of the loss function facilitates a more direct estimation of the balanced risk. Specifically, as PRIME focuses on aligning features with proxies, we direct our analysis to the risk associated with $\mathcal{L}_{\text{align}}$.

Theorem A.2. *For any loss function $\mathcal{L}(\mathbf{x}, \mathbf{y}, \xi)$, we have*

$$
\int \mathcal{L}(\mathbf{x}, \mathbf{y}, \xi) p_{bal}(\mathbf{x}, \mathbf{y}, \xi) d\mathbf{x} d\mathbf{y} d\xi = \int \mathcal{L}_r(\mathbf{x}, \mathbf{y}, \xi) p(\mathbf{x}, \mathbf{y}, \xi) d\mathbf{x} d\mathbf{y} d\xi,
$$

where the reweighted loss, $\mathcal{L}_r$, is defined as

$$
\mathcal{L}_r := \frac{p_{bal}(\xi)}{\int q(\xi|\mathbf{y}) p(\mathbf{y}) d\mathbf{y}} \mathcal{L}.
$$

Proof.

$$\begin{aligned}\int \mathcal{L}_r(\mathbf{x},\mathbf{y},\xi)p(\mathbf{x},\mathbf{y},\xi)d\mathbf{x}d\mathbf{y}d\xi &= \int \mathcal{L}(\mathbf{x},\mathbf{y},\xi)\frac{p_{\text{bal}}(\xi)}{\int q(\xi|\mathbf{y})p(\mathbf{y})d\mathbf{y}}p(\mathbf{x},\mathbf{y},\xi)d\mathbf{x}d\mathbf{y}d\xi \\ &= \int \mathcal{L}(\mathbf{x},\mathbf{y},\xi)\frac{p_{\text{bal}}(\xi)}{p(\xi)}p(\mathbf{x},\mathbf{y}|\xi)p(\xi)d\mathbf{x}d\mathbf{y}d\xi \\ &= \int \mathcal{L}(\mathbf{x},\mathbf{y},\xi)p(\mathbf{x},\mathbf{y}|\xi)p_{\text{bal}}(\xi)d\mathbf{x}d\mathbf{y}d\xi \\ &= \int \mathcal{L}(\mathbf{x},\mathbf{y},\xi)p_{\text{bal}}(\mathbf{x},\mathbf{y},\xi)d\mathbf{x}d\mathbf{y}d\xi.\end{aligned}$$

□

Remark. If $\mathcal{L} = -T_\xi \log A_\xi$ and $q(\xi|\mathbf{y}) = T_\xi$, $\mathcal{L}$ reduces to our alignment loss $\mathcal{L}_{\text{align}}$.

Theorem A.2 confirms that employing $\mathcal{L}_r$ allows the balanced alignment risk to be minimized directly. Note that the weighting term $p_{\text{bal}}(\xi)/p(\xi)$ is related to the proxy frequency. Since PRW, CB, and LDAM perform re-weighting, balancing, and margin control based on the proxy frequency, they empirically approximate $\mathcal{L}_r$. Specifically, for PRW, using $q(\xi|\mathbf{y}) = T_\xi$ and applying batch-wise Monte Carlo, we can derive s_j as follows:

$$\frac{p_{\text{bal}}(\xi=j)}{p(\xi=j)} = \frac{p_{\text{bal}}(\xi=j)}{\int q(\xi=j|\mathbf{y})p(\mathbf{y})d\mathbf{y}} \approx \frac{\frac{1}{C}}{\frac{1}{N_b}\sum_{i=1}^{N_b} q(\xi=j|\mathbf{y}=\mathbf{y}_i)} = \frac{1}{s_j}. \tag{23}$$

A similar deriviation is possible for CB, LDAM by appropriately modeling $q(\xi|\mathbf{y})$ and $p_\theta(\xi|\mathbf{x})$. Finally, we conclude that incorporating class imbalance techniques approximately induces the balanced alignment risk, leading to balanced feature learning.

B. Detailed Experimental Setup

This section presents additional details of our experimental setup. In §B.1, we first provide a detailed explanation of the datasets used in our experiments. In §B.2, we provide the implementation details of PRIME.

B.1. Dataset Details

In this work, we conduct experiments on four real-world imbalanced regression benchmarks introduced by (Yang et al., 2021): AgeDB-DIR, IMDB-WIKI-DIR, NYUD2-DIR, STS-B-DIR. For a fair and meaningful comparison with existing methods, we evaluate the proposed method using the same experimental setup, following the previous state-of-the-art methods for each dataset (Yang et al., 2021; Gong et al., 2022; Keramati et al., 2024). Table 10 provides the overall statistics of the four datasets. Please refer to (Yang et al., 2021) for more details.

Table 10. **Overall dataset statistics.**

Dataset	Target type	Target range	Bin size	Max bin	Min bin	# Training	# Val.	# Test
AgeDB-DIR	Age	$[0, 101]$	1	353	1	12,208	2,140	2,140
IMDB-WIKI-DIR	Age	$[0, 186]$	1	7,149	1	191,509	11,022	11,022
NYUD2-DIR	Depth	$[0.7, 10]$	0.1	1.46×10^8	1.13×10^6	50,688	-	654
STS-B-DIR	Text similarity	$[0, 5]$	0.1	428	1	5,249	1,000	1,000

B.2. Implementation Details

For fair comparisons, we follow the benchmark settings of (Yang et al., 2021) for all baselines and our method. Specifically, we use the same backbones and training details as in existing methods and tune only the hyperparameters of PRIME. Tables 11, 12, 13, and 14 summarize the implementation details for AgeDB-DIR, IMDB-WIKI-DIR, NYUD2-DIR, and STS-B-DIR, respectively. Overall, PRIME is easy to implement and can be integrated into existing regression methods by simply adding $\mathcal{L}_{\text{proxy}}$ and $\mathcal{L}_{\text{align}}$ to the regression loss $\mathcal{L}_{\text{reg}}$. We will release the code after publication.

PRIME. The number of proxies C is empirically determined for each dataset. Proxy embeddings $\{\mathbf{z}_i^p\}_{i=1}^C$ are initialized with He initialization (He et al., 2015) and trained jointly with the model. The *Proxy lr* refers to the multiplication factor applied to the learning rate of the proxy. The hyperparameters λ_p, λ_a, τ_f, τ_t, and α are set empirically. Especially, we use high τ_f values following (Movshovitz-Attias et al., 2017; Kim et al., 2020; Teh et al., 2020; Lim et al., 2022).

Class imbalance techniques. The implementations of class imbalance techniques follow their original codes, with hyperparameters set as in the respective papers. In PRW, the truncation threshold δ_{min} for excessively small s_j values is set as a scaled multiple of the median value s_{med}. In CB, $\beta = 0.99$ is used for all experiments, while in LDAM, the max margin value defines the upper limit of the enforced margin. DRW, short for Deferred Re-weighting (Cao et al., 2019), is optionally applied to all three methods, meaning that re-weighting is applied only after the specified DRW epoch.

Training details. For AgeDB-DIR and IMDB-WIKI-DIR, we use ResNet50 (He et al., 2016) as the backbone and the square root inverse (SQInv) re-weighted MAE loss as the regression loss $\mathcal{L}_{\text{reg}}$. For NYUD2-DIR, we adopt a ResNet50-based encoder-decoder architecture (Hu et al., 2019) and Balanced MSE (Ren et al., 2022) as $\mathcal{L}_{\text{reg}}$. For STS-B-DIR, we employ BiLSTM with GloVe (Pennington et al., 2014) word embeddings as the feature extractor and use LDS + inverse (Inv) re-weighted MSE loss as $\mathcal{L}_{\text{reg}}$. The training hyperparameters, including epoch, batch size, learning rate, weight decay, optimizer, and scheduler, are primarily chosen based on the previous state-of-the-art methods (Gong et al., 2022; Ren et al., 2022; Keramati et al., 2024) for each dataset.

Table 11. **Implementation details for experiments on AgeDB-DIR.**

Module	Name	PRIME	+ PRW	+ CB	+ LDAM
PRIME	# Proxy	20	20	20	20
	Proxy lr	1	1	1	1
	λ_p	5	10	5	10
	λ_a	25	50	25	50
	τ_f	5	10	5	10
	τ_t	5	2	5	1
	α	0.005	0.0005	0.001	0.001
PRW	δ_{min}	-	$0.05 \times s_{\text{med}}$	-	-
	DRW	-	✗	-	-
CB	β	-	-	0.99	-
	DRW	-	-	✗	-
LDAM	Max margin	-	-	-	0.5
	DRW	-	-	-	40
Training	Backbone	ResNet50	ResNet50	ResNet50	ResNet50
	$\mathcal{L}_{\text{reg}}$	SQInv (MAE)	SQInv (MAE)	SQInv (MAE)	SQInv (MAE)
	Epoch	80	80	80	80
	Batch size	64	64	64	64
	Learning rate	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}
	Weight decay	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
	Optimizer	Adam	Adam	Adam	Adam
	Scheduler	StepLR (60/0.1)	StepLR (60/0.1)	StepLR (60/0.1)	StepLR (40/0.5)

Table 12. **Implementation details for experiments on IMDB-WIKI-DIR.**

Module	Name	PRIME	+ PRW	+ CB	+ LDAM
PRIME	# Proxy	40	40	40	40
	Proxy lr	5	5	1	1
	λ_p	5	5	5	5
	λ_a	25	25	25	25
	τ_f	10	10	5	10
	τ_t	1	2	5	2
	α	0.001	0.001	0.001	0.001
PRW	δ_{min}	-	$1.0 \times s_{\text{med}}$	-	-
	DRW	-	60	-	-
CB	β	-	-	0.99	-
	DRW	-	-	✗	-
LDAM	Max margin	-	-	-	0.5
	DRW	-	-	-	✗
Training	Backbone	ResNet50	ResNet50	ResNet50	ResNet50
	$\mathcal{L}_{\text{reg}}$	SQInv (MAE)	SQInv (MAE)	SQInv (MAE)	SQInv (MAE)
	Epoch	80	80	80	80
	Batch size	64	64	64	64
	Learning rate	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}
	Weight decay	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
	Optimizer	Adam	Adam	Adam	Adam
	Scheduler	StepLR (60/0.1)	StepLR (60/0.1)	StepLR (60/0.1)	StepLR (60/0.5)

Table 13. **Implementation details for experiments on NYUD2-DIR.**

Module	Name	PRIME	+ PRW	+ CB	+ LDAM
PRIME	# Proxy	10	10	10	10
	Proxy lr	1	1	1	1
	λ_p	0.1	0.1	0.1	0.1
	λ_a	0.5	0.5	0.5	0.5
	τ_f	5	10	10	10
	τ_t	1	2	2	2
	α	0.0001	0.0005	0.0005	0.0005
PRW	δ_{min}	-	$0.05 \times s_{\text{med}}$	-	-
	DRW	-	✗	-	-
CB	β	-	-	0.99	-
	DRW	-	-	✗	-
LDAM	Max margin	-	-	-	0.5
	DRW	-	-	-	✗
Training	Backbone	ResNet50 E-D	ResNet50 E-D	ResNet50 E-D	ResNet50 E-D
	$\mathcal{L}_{\text{reg}}$	Balanced MSE	Balanced MSE	Balanced MSE	Balanced MSE
	Epoch	20	20	20	20
	Batch size	64	64	64	64
	Learning rate	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
	Weight decay	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
	Optimizer	Adam	Adam	Adam	Adam
	Scheduler	StepLR (5/0.1)	StepLR (5/0.1)	StepLR (5/0.1)	StepLR (5/0.1)

Table 14. **Implementation details for experiments on STS-B-DIR.**

Module	Name	PRIME	+ PRW	+ CB	+ LDAM
PRIME	# Proxy	26	26	26	26
	Proxy lr	1	1	1	1
	λ_p	1×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-5}
	λ_a	5×10^{-5}	1×10^{-4}	1×10^{-4}	5×10^{-5}
	τ_f	5	5	5	5
	τ_t	5	5	5	5
	α	0.001	0.01	0.01	0.01
PRW	δ_{min}	-	$3.0 \times s_{\text{med}}$	-	-
	DRW	-	✗	-	-
CB	β	-	-	0.99	-
	DRW	-	-	✗	-
LDAM	Max margin	-	-	-	0.5
	DRW	-	-	-	✗
Training	Backbone	BiLSTM + GloVe	BiLSTM + GloVe	BiLSTM + GloVe	BiLSTM + GloVe
	$\mathcal{L}_{\text{reg}}$	LDS + Inv (MSE)	LDS + Inv (MSE)	LDS + Inv (MSE)	LDS + Inv (MSE)
	Epoch	300	300	300	300
	Batch size	16	16	16	16
	Learning rate	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}
	Optimizer	Adam	Adam	Adam	Adam
	Patience	100	100	100	100

C. Further Analyses

In this section, we present additional experimental results and analyses. In §C.1, we present additional details and complete results for the experiments discussed in the manuscript. In §C.2, we evaluate the computational efficiency of PRIME. In §C.3, we conduct a sensitivity analysis on the hyperparameters of PRIME. Lastly, in §C.4, we discuss the differences between PRIME and Regression-as-Classification approaches, which reformulate regression as a classification problem.

C.1. Additional Results

C.1.1. PRIME FACILITATES EFFECTIVE FEATURE LEARNING

The twisted line in Figure 2(b) appears due to suboptimal alignment between features and their corresponding proxies in the *Few* category. Although the proxies represent a balanced feature distribution, the alignment process in (7) still faces challenges under sample imbalance. Minority samples, which occur infrequently, often fail to align properly with their proxies, resulting in distorted feature–proxy alignment. The use of class imbalance techniques (*e.g.*, PRW, CB, and LDAM) provides better alignment focus on minority samples, mitigating this issue. To empirically validate their effect, we conduct an additional analysis on the AgeDB-DIR dataset, measuring the Spearman correlation between the proxy–feature similarity matrix (as visualized in Figure 2(b)) and the label similarity matrix. A higher correlation indicates better alignment and reduced distortion in the learned feature space. Table 15 reports the Spearman correlation values when PRIME is combined with various class imbalance techniques. Results are averaged over five runs. Incorporating class imbalance techniques significantly improves the correlation, confirming their effectiveness in facilitating better alignment, particularly for samples in the *Few* category.

Table 15. **Spearman correlation between proxy-feature and label similarity matrices on AgeDB-DIR.** A higher correlation indicates better alignment between learned features and the corresponding proxies.

Method	ρ (↑)
PRIME	$0.722_{\pm 0.020}$
PRIME + PRW	$0.802_{\pm 0.021}$
PRIME + CB	$0.800_{\pm 0.023}$
PRIME + LDAM	$0.837_{\pm 0.015}$

C.1.2. COMPARISON WITH RECENT REPRESENTATION LEARNING METHODS FOR GENERAL REGRESSION

In Table 7 of the manuscript, only the results for the entire test set (*i.e.*, *All*) are reported. In Table 16 below, we present the complete results. HPN shows slightly better performance than PRIME in *Few*. However, since HPN uses fixed prototypes on a hypersphere, it cannot effectively represent the entire dataset, leading to suboptimal performance in *Many* and *Median*. Notably, PRIME outperforms state-of-the-art representation learning methods designed for general regression, such as Rank-N-Contrast and Ordinal Entropy, demonstrating its ability to learn effective representations that are robust to data imbalance. Furthermore, PRIME exhibits the flexibility to incorporate various class imbalance techniques for balanced feature learning in regression problems. Applying these techniques effectively enhances the performance of minority targets.

Table 16. **Comparison with representation learning methods for general regression on AgeDB-DIR.** The **best** results are marked in bold, and the second best are underlined.

Method	**MAE (↓)**				**GM (↓)**			
	All	Many	Median	Few	All	Many	Median	Few
SQInv (MAE)	$7.42_{\pm 0.06}$	$6.78_{\pm 0.12}$	$8.55_{\pm 0.18}$	$10.71_{\pm 0.31}$	$4.77_{\pm 0.08}$	$4.37_{\pm 0.14}$	$5.73_{\pm 0.23}$	$7.39_{\pm 0.36}$
+ HPN (Mettes et al., 2019)	$7.38_{\pm 0.08}$	$6.78_{\pm 0.11}$	$8.44_{\pm 0.16}$	$10.10_{\pm 0.53}$	$4.66_{\pm 0.09}$	$4.30_{\pm 0.10}$	$5.46_{\pm 0.16}$	$6.41_{\pm 0.39}$
+ Ordinal Entropy (Zhang et al., 2023a)	$7.33_{\pm 0.08}$	$6.77_{\pm 0.13}$	$8.24_{\pm 0.13}$	$10.16_{\pm 0.38}$	$4.68_{\pm 0.07}$	$4.32_{\pm 0.09}$	$5.39_{\pm 0.19}$	$6.82_{\pm 0.38}$
+ Rank-N-Contrast (Zha et al., 2023)	$7.27_{\pm 0.05}$	$\underline{6.52}_{\pm 0.03}$	$8.62_{\pm 0.15}$	$10.66_{\pm 0.41}$	$4.69_{\pm 0.04}$	$4.23_{\pm 0.02}$	$5.73_{\pm 0.15}$	$7.12_{\pm 0.37}$
+ PRIME	$\underline{7.09}_{\pm 0.08}$	$\mathbf{6.38}_{\pm 0.11}$	$8.39_{\pm 0.26}$	$10.13_{\pm 0.36}$	$\mathbf{4.39}_{\pm 0.08}$	$\mathbf{3.91}_{\pm 0.10}$	$5.58_{\pm 0.22}$	$6.57_{\pm 0.49}$
+ PRIME + PRW	$\mathbf{7.06}_{\pm 0.09}$	$6.67_{\pm 0.09}$	$\mathbf{7.27}_{\pm 0.25}$	$\underline{9.91}_{\pm 0.16}$	$\mathbf{4.39}_{\pm 0.08}$	$4.14_{\pm 0.09}$	$\mathbf{4.69}_{\pm 0.20}$	$6.39_{\pm 0.16}$
+ PRIME + CB	$7.12_{\pm 0.09}$	$6.61_{\pm 0.09}$	$8.07_{\pm 0.11}$	$\mathbf{9.29}_{\pm 0.68}$	$\underline{4.47}_{\pm 0.05}$	$4.16_{\pm 0.08}$	$5.23_{\pm 0.07}$	$\underline{5.81}_{\pm 0.46}$
+ PRIME + LDAM	$7.24_{\pm 0.06}$	$6.85_{\pm 0.14}$	$\underline{7.84}_{\pm 0.31}$	$\mathbf{9.29}_{\pm 0.44}$	$\underline{4.47}_{\pm 0.07}$	$4.26_{\pm 0.12}$	$\underline{4.89}_{\pm 0.25}$	$\mathbf{5.60}_{\pm 0.54}$

C.1.3. Effectiveness of our proxy formulation

Table 8 in the manuscript reports only the results for the entire test set (*i.e.*, *All*). In Table 17 below, we present the complete results. For ProxyNCA, we adapt the original method to the regression setting. Similar to PRIME, proxy assignment is designed to ensure that the associated targets are uniformly distributed in the target space. However, unlike PRIME, which explicitly optimizes proxy features to be well-ordered in the feature space via $\mathcal{L}_{\text{proxy}}$, ProxyNCA relies solely on feature-proxy alignment. As shown in our results, PRIME significantly outperforms ProxyNCA, demonstrating the advantage of our DIR-specific proxy formulation. For the non-learnable variant of PRIME, proxy features are computed as the centroids of sample features assigned to each proxy. To enable proper backpropagation of the proxy and alignment losses, these proxy features are updated within each mini-batch based on the current sample-to-proxy assignments. While the centroid-based method achieves slightly better performance than the learnable proxy in the Median category, it suffers from notable performance degradation in the other regions. In particular, we observe a significant performance drop in the Few category, indicating that the centroid-based proxies struggle under severe data sparsity. This performance gap stems from their inherent limitations: centroid quality depends on the number of assigned samples and becomes unstable when only a few are available. In contrast, learnable proxies are global parameters updated via backpropagation, offering greater stability and robustness under sparse conditions.

Table 17. **Comparison with proxy-based alternatives on AgeDB-DIR.** The **best** results are marked in bold, and the second best are underlined.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
ProxyNCA (Movshovitz-Attias et al., 2017)	7.33±0.08	6.52±0.09	8.69±0.27	11.14±0.21	4.64±0.06	4.12±0.07	5.80±0.28	7.60±0.53
Non-learnable (centroid)	7.21±0.09	6.57±0.10	**8.20**±0.13	10.89±0.33	4.67±0.11	4.24±0.12	**5.42**±0.15	7.64±0.22
PRIME	**7.09**±0.08	**6.38**±0.11	8.39±0.26	**10.13**±0.36	**4.39**±0.08	**3.91**±0.10	5.58±0.22	**6.57**±0.49

C.1.4. Ablation study

Table 9 in the manuscript reports only the results for the entire test set (*i.e.*, *All*). In Table 18 below, we present the complete results. Comparing the ablation models, the transition from M1 to M2 leads to a significant performance improvement across *All*, *Many*, and *Few*, highlighting the effectiveness of $\mathcal{L}_{\text{proxy}}$. From M2 to M3, assigning proxies uniformly in the target space leads to a notable performance gain in *Few* compared to random initialization, demonstrating the effectiveness of ensuring that proxies represent the target space in a balanced manner. Finally, the transition from M3 to M4 (PRIME) further enhances performance, as regression involves continuous targets, making distance-proportional assignment, as described in (6), more effective than one-hot encoding.

Table 18. **Performance comparison of ablation models on AgeDB-DIR.** The **best** results are marked in bold.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
M1	7.34±0.08	6.49±0.04	8.84±0.29	11.24±0.36	4.66±0.07	4.11±0.06	5.96±0.24	7.88±0.48
M2	7.24±0.09	6.47±0.11	8.40±0.18	11.32±0.47	4.61±0.12	4.09±0.12	5.71±0.25	7.96±0.60
M3	7.23±0.12	6.45±0.11	8.52±0.29	10.96±0.51	4.55±0.12	4.01±0.11	5.74±0.23	7.76±0.36
M4 (PRIME)	**7.09**±0.08	**6.38**±0.11	**8.39**±0.26	**10.13**±0.36	**4.39**±0.08	**3.91**±0.10	**5.58**±0.22	**6.57**±0.49

C.2. Computational Efficiency

To analyze the computational efficiency of PRIME, we compute the average wall-clock training time (in seconds) using four NVIDIA Tesla V100 GPUs. Table 19 presents comparisons with existing representation learning methods on AgeDB-DIR. For fair comparisons, we apply the same training details (*e.g.*, epochs, batch size, optimizer, etc.) to all methods. The training time of PRIME is considerably lower than that of FDS and ranking-based methods (RankSim and Rank-N-Contrast) while remaining comparable to the time complexity of ProxyNCA, Ordinal Entropy, and ConR. Although incorporating class imbalance techniques introduces a slight computational overhead, it remains comparable to other well-established methods. Overall, these results demonstrate that PRIME is an effective representation learning framework that does not compromise efficiency.

Table 19. **Average wall-clock training time (in seconds) on AgeDB-DIR.**

Method	**Training time (sec) (↓)**	Δ
SQInv (MAE)	1818.0 ± 58.7	-
+ ProxyNCA (Movshovitz-Attias et al., 2017)	1934.8 ± 77.4	(+ 116.8)
+ HPN (Mettes et al., 2019)	1833.8 ± 13.2	(+ 15.8)
+ FDS (Yang et al., 2021)	3380.0 ± 21.5	(+ 1562.0)
+ RankSim (Gong et al., 2022)	2254.4 ± 20.9	(+ 436.4)
+ Ordinal Entropy (Zhang et al., 2023a)	2067.6 ± 79.9	(+ 249.6)
+ Rank-N-Contrast (Zha et al., 2023)	2122.6 ± 33.7	(+ 304.6)
+ ConR (Keramati et al., 2024)	1952.6 ± 23.0	(+ 134.6)
+ PRIME	1936.6 ± 45.4	(+ 118.6)
+ PRIME + PRW	2094.6 ± 20.6	(+ 276.6)
+ PRIME + CB	1990.2 ± 41.4	(+ 172.2)
+ PRIME + LDAM	1987.6 ± 18.7	(+ 169.6)

C.3. Sensitivity Analysis

We analyze the impact of PRIME's hyperparameters using the AgeDB-DIR dataset. Specifically, we examine the effects of the number of proxies (C), the trade-off hyperparameters for $\mathcal{L}_{\text{proxy}}$ and $\mathcal{L}_{\text{align}}$ in (8) (λ_p and λ_a), the temperature hyperparameters (τ_f and τ_t), and the coefficient for the regularization term in (4) (α). We evaluate their influence by varying the values as follows: $C \in \{10, 20, 30, 40\}$, $\lambda_p \in \{0.5, 2.5, 5.0, 10.0\}$, $\lambda_a \in \{5, 10, 25, 50\}$, $\tau_f \in \{0.5, 1.0, 2.0, 5.0\}$, $\tau_t \in \{0.5, 1.0, 2.0, 5.0\}$, and $\alpha \in \{0, 0.0001, 0.0005, 0.001, 0.005\}$. Tables 20, 21, 22, 24, 23, and 25 summarize the results. Overall, PRIME demonstrates reliable and robust performance across different hyperparameter choices. In particular, PRIME consistently outperforms the w/o PRIME baseline in almost all cases (SQInv (MAE) serves as the baseline), further demonstrating its effectiveness. As mentioned in Table 11, for the main results, we set $C = 20$, $\lambda_p = 5$, $\lambda_a = 25$, $\tau_f = 5$, $\tau_t = 5$, and $\alpha = 0.005$, which are highlighted in gray in the tables.

Table 20. **Effect of the number of proxies (C) on AgeDB-DIR.** The gray-highlighted value indicates the selected setting used in our experiment. The **best** results are marked in bold.

Method	**MAE (↓)**				**GM (↓)**			
	All	Many	Median	Few	All	Many	Median	Few
w/o PRIME	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
PRIME ($C = 10$)	7.35±0.10	6.49±0.06	8.80±0.40	11.47±0.46	4.65±0.05	4.09±0.05	5.95±0.28	7.90±0.47
PRIME ($C = 20$)	**7.09**±0.08	6.38±0.11	8.39±0.26	**10.13**±0.36	**4.39**±0.08	**3.91**±0.10	5.58±0.22	**6.57**±0.49
PRIME ($C = 30$)	7.14±0.09	6.41±0.14	**8.31**±0.23	10.61±0.32	4.48±0.06	3.98±0.09	**5.55**±0.13	7.31±0.50
PRIME ($C = 40$)	7.16±0.07	**6.37**±0.05	8.41±0.18	11.20±0.28	4.53±0.10	4.02±0.08	**5.55**±0.30	7.94±0.28

Table 21. **Effect of λ_p on AgeDB-DIR.** The gray-highlighted value indicates the selected setting used in our experiment. The **best** results are marked in bold.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
w/o PRIME	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
PRIME ($\lambda_p = 0.5$)	7.19±0.10	6.42±0.07	8.46±0.25	10.97±0.54	4.53±0.08	4.00±0.06	5.73±0.23	7.79±0.78
PRIME ($\lambda_p = 2.5$)	7.21±0.08	6.46±0.11	**8.33**±0.29	10.84±0.35	4.44±0.07	4.05±0.06	**5.50**±0.38	**6.55**±0.57
PRIME ($\lambda_p = 5.0$)	**7.09**±0.08	**6.38**±0.11	8.39±0.26	**10.13**±0.36	**4.39**±0.08	**3.91**±0.10	5.58±0.22	6.57±0.49
PRIME ($\lambda_p = 10.0$)	7.26±0.06	6.49±0.04	8.58±0.17	10.91±0.49	4.63±0.10	4.09±0.08	5.83±0.35	6.83±0.64

Table 22. **Effect of λ_a on AgeDB-DIR.** The gray-highlighted value indicates the selected setting used in our experiment. The **best** results are marked in bold.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
w/o PRIME	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
PRIME ($\lambda_a = 5$)	7.20±0.14	6.54±0.12	**8.37**±0.29	10.25±0.40	4.57±0.14	4.16±0.13	**5.46**±0.31	6.78±0.48
PRIME ($\lambda_a = 10$)	7.18±0.08	6.41±0.11	8.51±0.17	10.80±0.62	4.63±0.05	4.10±0.08	5.83±0.19	7.68±0.63
PRIME ($\lambda_a = 25$)	**7.09**±0.08	**6.38**±0.11	8.39±0.26	**10.13**±0.36	**4.39**±0.08	**3.91**±0.10	5.58±0.22	**6.57**±0.49
PRIME ($\lambda_a = 50$)	7.21±0.07	6.42±0.07	8.56±0.26	10.94±0.45	4.54±0.07	4.02±0.08	5.72±0.24	7.65±0.70

Table 23. **Effect of τ_f on AgeDB-DIR.** The gray-highlighted value indicates the selected setting used in our experiment. The **best** results are marked in bold.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
w/o PRIME	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
PRIME ($\tau_f = 0.5$)	7.20±0.09	6.61±0.16	**8.18**±0.28	**10.10**±0.39	4.59±0.13	4.20±0.18	**5.41**±0.23	6.68±0.37
PRIME ($\tau_f = 1.0$)	7.24±0.07	6.51±0.07	8.58±0.33	10.45±0.39	4.56±0.09	4.06±0.11	5.82±0.30	6.96±0.36
PRIME ($\tau_f = 2.0$)	7.22±0.10	6.45±0.02	8.59±0.36	10.68±0.47	4.62±0.11	4.09±0.06	5.90±0.38	7.39±0.62
PRIME ($\tau_f = 5.0$)	**7.09**±0.08	**6.38**±0.11	8.39±0.26	10.13±0.36	**4.39**±0.08	**3.91**±0.10	5.58±0.22	**6.57**±0.49

Table 24. **Effect of τ_t on AgeDB-DIR.** The gray-highlighted value indicates the selected setting used in our experiment. The **best** results are marked in bold.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
w/o PRIME	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
PRIME ($\tau_t = 0.5$)	7.20±0.09	6.42±0.10	8.46±0.27	11.07±0.26	4.43±0.12	**3.90**±0.14	5.70±0.35	7.95±0.27
PRIME ($\tau_t = 1.0$)	7.16±0.08	6.44±0.06	**8.33**±0.17	10.67±0.46	4.52±0.08	4.04±0.07	**5.57**±0.11	7.38±0.61
PRIME ($\tau_t = 2.0$)	7.20±0.10	6.40±0.12	8.60±0.25	10.85±0.22	4.59±0.04	4.06±0.07	5.79±0.31	7.63±0.30
PRIME ($\tau_t = 5.0$)	**7.09**±0.08	**6.38**±0.11	8.39±0.26	**10.13**±0.36	**4.39**±0.08	3.91±0.10	5.58±0.22	**6.57**±0.49

Table 25. **Effect of α on AgeDB-DIR.** The gray-highlighted value indicates the selected setting used in our experiment. The **best** results are marked in bold.

Method	MAE (↓)				GM (↓)			
	All	Many	Median	Few	All	Many	Median	Few
w/o PRIME	7.42±0.06	6.78±0.12	8.55±0.18	10.71±0.31	4.77±0.08	4.37±0.14	5.73±0.23	7.39±0.36
PRIME ($\alpha = 0$)	7.22±0.11	6.44±0.10	8.44±0.08	11.16±0.40	4.56±0.09	4.02±0.10	5.76±0.06	7.77±0.40
PRIME ($\alpha = 0.0001$)	7.17±0.06	6.44±0.08	**8.27**±0.38	11.03±0.17	4.55±0.07	4.04±0.02	5.62±0.38	7.79±0.24
PRIME ($\alpha = 0.0005$)	7.22±0.07	6.47±0.05	8.34±0.26	11.19±0.34	4.60±0.07	4.10±0.03	5.66±0.25	7.76±0.29
PRIME ($\alpha = 0.001$)	7.28±0.06	6.52±0.03	8.39±0.17	10.33±0.40	4.53±0.03	4.12±0.04	5.70±0.18	6.87±0.53
PRIME ($\alpha = 0.005$)	**7.09**±0.08	**6.38**±0.11	8.39±0.26	**10.13**±0.36	**4.39**±0.08	**3.91**±0.10	**5.58**±0.22	**6.57**±0.49

C.4. Discussion on Regression-as-Classification Approaches

Although our PRIME shares a classification-like perspective, we highlight two key differences from Regression-as-Classification approaches (Rothe et al., 2015; Cao et al., 2017; Liu et al., 2019a; Xiong & Yao, 2024) that quantize continuous targets into discrete bins and treat each bin as a class: **(i)** As samples with different target values are grouped under the same class, previous methods suffer from quantization errors. In contrast, PRIME assigns proxies based on target associations derived from target distances, as in (6), effectively mitigating quantization error. **(ii)** Moreover, rather than directly predicting proxy indices (*i.e.,* classes), PRIME optimizes the model to minimize the feature distance to the corresponding proxy. As our proxy loss $\mathcal{L}_{\text{proxy}}$ in (4) enforces proxies to be well-ordered, it facilitates better regression-specific representation learning.