RETHINKING FAIR ANOMALY DETECTION VIA REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair anomaly detection (AD) methods typically trade detection accuracy for group fairness, especially under severe group imbalance. We argue that this trade-off is not inherent to fully unsupervised AD but stems from how fairness constraints distort learned representations. We identify two failure modes: *vanishing protected-group influence* under imbalance and *dimensional collapse* caused by overly restrictive fairness regularization. We show theoretically that many existing fairness regularizations dilute the protected group's gradients as imbalance grows and, even with the AD loss included, suppress all group-predictive features as the fairness coefficient increases. We propose FADIG, which combines an adaptively re-balanced autoencoder with a fairness-aware contrastive objective, and prove that it removes the dependence of group contributions on the imbalance ratio and makes collapsed representations suboptimal. On four image and tabular datasets, FADIG achieves higher accuracy with lower group disparity than existing methods.

1 INTRODUCTION

Anomaly detection (AD) aims to identify instances that deviate significantly from the majority (Chandola et al., 2009) and is widely used in high-stakes domains such as financial fraud detection (Huang et al., 2018; Rezapour, 2019), medical pathology analysis (Faust et al., 2018; Shvetsova et al., 2021), and cybersecurity intrusion detection (Liao et al., 2013; Ahmad et al., 2021; Ahmed et al., 2016). Despite strong detection performance, modern AD methods (Sohn et al., 2021; Li et al., 2023; Yan et al., 2021; Wang et al., 2023) often exhibit systematic bias across sensitive groups, motivating fairness-aware AD methods that add explicit fairness constraints (Shekhar et al., 2021; Fioresi et al., 2023; Song et al., 2021; Zhang & Davidson, 2021).

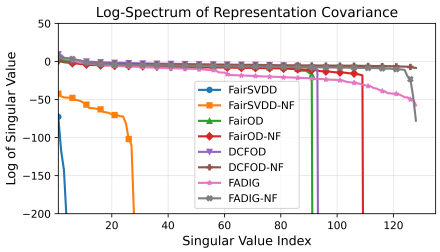


Figure 1: Singular value spectrum of embedding space on MNIST-USPS dataset. '-NF' denotes that the fairness regularizer is removed. Each embedding vector has a dimension of 128. A number of singular values drop to zero, indicating the dimensional collapse.

Existing fair AD methods (Shekhar et al., 2021; Song et al., 2021; Zhang & Davidson, 2021) typically improve fairness at the expense of detection accuracy. While this trade-off is widely observed, a fundamental question remains unexplored: *What causes it, and is it inherent to fair AD, or a consequence of how fairness constraints are imposed?* We argue that existing approaches overlook the representation-level consequences of fairness constraints, in particular *dimensional collapse* (Jing et al., 2022; He et al., 2024; Chen et al., 2024). Figure 1 shows the singular value spectrum of embeddings learned with and without fairness constraints. With the constraints, the spectrum of existing methods decays much more rapidly, indicating that representations concentrate in a low-dimensional subspace, whereas fairness-unaware models ('-NF') preserve variance across many dimensions. Since AD relies on modeling fine-grained deviations from normal patterns (Audibert et al., 2020; Yan et al., 2021; Wang et al., 2023), this collapse degrades detection, particularly for minority-pattern anomalies.

In this paper, we trace the trade-off between performance and fairness to two mechanisms. *(1) Vanishing protected-group influence:* when a fairness loss aggregates over samples without correcting for imbalance, the protected group's gradient contribution is diluted by a factor of order $\sqrt{n/m}$,

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

where n and m are the protected and unprotected group sizes. *(2) Dimensional collapse:* common fairness regularizers are minimized by a collapsed representation, and as the fairness coefficient grows they suppress group-dependent structure even when the AD loss is included. Both mechanisms stem from how fairness is imposed, suggesting that the trade-off is not inherent to fair AD.

Motivated by these insights, we propose FADIG, a fairness-aware contrastive learning-based anomaly detection method under group imbalance scenario designed from a principled representation-learning perspective. Instead of enforcing fairness by suppressing group information, FADIG explicitly preserves representation dimensionality while aligning group semantics. FADIG consists of two key components: (1) *Fairness-Aware Contrastive Learning Module* aligns group representations without collapsing feature dimensions by maximizing inter-group similarity while preserving intra-group diversity, thereby maintaining the expressiveness required for effective anomaly detection. (2) *Adaptively Re-balanced Autoencoder Module* reweights the reconstruction losses of the protected and unprotected groups using weights computed from group-wise losses without learnable parameters. Under these weights, the group-size factor cancels exactly in the reconstruction gradients, preventing majority dominance and mitigating vanishing minority influence during representation learning. Together, these components enable FADIG to mitigate fairness-induced dimensional collapse, achieving strong anomaly detection performance with low group disparity, even under severe imbalance. Our main contributions are summarized as follows:

- We identify vanishing protected-group influence and dimensional collapse as two representation-level causes of the performance–fairness trade-off in fully unsupervised, group-imbalanced AD.
- We show that sample-additive fairness regularizers dilute protected-group gradients by a factor of order $\sqrt{n/m}$, and that regularizers minimized by collapse suppress group-dependent structure as the fairness coefficient grows, even with the AD loss included.
- We propose FADIG and prove that its group contributions do not depend on the imbalance ratio and that collapsed representations are suboptimal for its contrastive objective.
- On four image and tabular datasets, FADIG achieves the best detection accuracy with low group disparity.

2 Preliminaries

In this paper, we explore the fairness issue in the unsupervised anomaly detection task. Among the various fairness definitions proposed, there is no consensus about the best one to use. In this work, we focus on group fairness notions which usually pursue the equity of certain metrics among the groups. Without loss of generality, we consider the groups here to be the protected group and the unprotected group (e.g., Black and Non-Black in race). We are given a dataset $D = P \cup U$, where $P = \{x_i^P, y_i^P\}_{i=1}^n$ are examples from the protected group, $U = \{x_i^U, y_i^U\}_{i=1}^m$ are examples from the unprotected group, and x_i^P, x_i^U are sampled i.i.d. from distributions $\mathcal{P}_P, \mathcal{P}_U$ over the input space $\mathbb{R}^d$ respectively. Let $G = g_d \circ f_e$ be an autoencoder, where $z = f_e(x)$ denotes the latent representation and $G(x) = g_d(z)$. The ground-truth labels $y_i^P, y_i^U \in \mathcal{Y} = \{0, 1\}$ represent whether the example is an anomaly $(y = 1)$ or not, which are given by deterministic labeling functions $a_P, a_U : \mathbb{R}^d \to \mathcal{Y}$, respectively. We study the *fully unsupervised* setting: training data are unlabeled and may contain anomalies, unlike one-class classification, which trains on curated normal samples. Sensitive attributes are assumed available, as they are typically recorded for compliance in domains such as lending, healthcare, and criminal justice, even when anomaly labels are not. The task of unsupervised anomaly detection is to find a hypothesis $h : \mathbb{R}^d \to \mathcal{Y}$ which identifies a subset $\mathcal{A} \subset D$ whose elements deviate significantly from the normal examples in D, relying only on intrinsic properties of the data such as its distribution or distances between examples. For clarity, we summarize the notations in Table 3.

Fairness-regularized AD. Let $s_i \in \{0, 1\}$ denote the group membership of x_i ($s_i = 1$ for P). Most fair AD methods train the encoder f_θ by minimizing

$$J_\alpha(\theta) = \mathcal{L}_{\mathrm{AD}}(\theta) + \alpha\, R(\theta), \tag{1}$$

where $\mathcal{L}_{\mathrm{AD}}$ is the anomaly-detection loss (e.g., reconstruction or one-class) and $R(\theta) \geq 0$ penalizes dependence between the group and the representation or anomaly score.

3 Limitations of Existing Fairness Regularizers

In this section, we analyze existing fairness regularizers for unsupervised AD from a representation-learning perspective and identify two interconnected failure modes, *vanishing protected-group*

influence under group imbalance and *dimensional collapse*, which together explain the commonly observed performance–fairness trade-off.

3.1 Vanishing Protected-Group Influence under Group Imbalance

A fundamental challenge in fairness-aware representation learning arises when the protected group is under-represented. Many fairness regularization terms (Song et al., 2021; Zhang & Davidson, 2021; Fioresi et al., 2023) aggregate losses or gradients across samples without explicitly correcting for group imbalance, causing the optimization dynamics to be dominated by the majority group. To formalize this phenomenon, we consider *sample-additive* regularizers, $R(\theta) = \frac{1}{M}\sum_{i=1}^{M} \ell_i(\theta)$, where $\ell_i(\theta)$ is the per-sample regularization loss with respect to the encoder parameters θ.

Assumption 3.1. Assume that for each group $A \in \{P, U\}$, the per-sample regularization losses $\{\ell_i(\theta)\}_{i\in A}$ are i.i.d. random variables. Each loss is differentiable with respect to the encoder parameters θ and satisfies: (1) $\|\nabla_\theta \ell_i(\theta)\| \le k$ almost surely for some constant $k > 0$, independent of the group size M_A. (2) $\|\mathbb{E}[\nabla_\theta \ell_i(\theta)]\| = 0$.

Assumption 3.2. For the unprotected group U, there exists a constant $c > 0$ independent of m such that the sum of per-sample gradients satisfies $\mathbb{E}\Big[\|\nabla_\theta R_U(\theta)\|\Big] \ge c\sqrt{m}$, where $R_U(\theta) = \sum_{i\in U} \ell_i(\theta)$ and m is the size of unprotected group.

Under these assumptions, the protected group's influence on the encoder gradients vanishes as the group imbalance grows:

Lemma 3.3. *Let $R(\theta)$ be a sample-additive fairness regularizer decomposed by group as*

$$R(\theta) = \frac{1}{M}\big(R_U(\theta) + R_P(\theta)\big), \qquad R_A(\theta) = \sum_{i\in A} \ell_i(\theta),$$

where U and P denote the unprotected and protected groups with sizes m and n ($M = m + n$), respectively. Under Assumptions 3.1 and 3.2, the expected ratio of gradient magnitudes satisfies

$$\mathbb{E}\left[\frac{\|\nabla_\theta R_P(\theta)\|}{\|\nabla_\theta R_U(\theta)\|}\right] = \mathcal{O}\Big(\sqrt{\frac{n}{m}}\Big).$$

Consequently, as $m/n \to \infty$, the influence of the protected group vanishes in expectation.

The proof is provided in Appendix C.1. This result highlights a statistical limitation shared by all sample-additive fairness regularization schemes: when group imbalance is severe, minority-group signals are overwhelmed by majority-group gradients. As a result, the learned representations primarily capture majority-group patterns, leading to systematic under-representation of protected-group features.

3.2 Dimensional Collapse under Fairness Constraints

Beyond imbalance-induced gradient dominance, fairness regularization can also distort the geometry of learned representations: objectives that enforce strong group invariance may reduce feature diversity, resulting in *dimensional collapse* (Jing et al., 2022; He et al., 2024; Chen et al., 2024). This applies to a broad class of regularizers R in Eq. (1) that penalize dependence between the group label s and the representation or anomaly score. In adversarial regularization, as in FairSVDD (Zhang & Davidson, 2021), a discriminator $\mathcal{D}_{\theta_d}$ predicts s_i from $z_i = f_e(x_i)$ by minimizing the cross-entropy $\mathcal{L}_D(\theta, \theta_d) = \frac{1}{M}\sum_{i=1}^{M} \ell_i(\theta, \theta_d)$, where θ denotes the encoder parameters and $M = m + n$. The encoder is penalized by $R(\theta) = H(s) - \min_{\theta_d} \mathcal{L}_D(\theta, \theta_d) \ge 0$, where $H(s)$ is the entropy of the group label. Other examples include distribution-matching penalties between the groups' embeddings and penalties on the difference in anomaly scores between groups (Shekhar et al., 2021). We first analyze adversarial regularization (Lemma 3.4) and then generalize (Proposition 3.5).

Lemma 3.4. *Assume the discriminator $\mathcal{D}_{\theta_d}$ in FairSVDD has sufficient capacity. Then, any embedding function f_θ satisfying $f_\theta(x) = c$ for all $x \in \mathcal{X}$, where $c \in \mathbb{R}^d$ is a constant vector, is a global maximizer of the adversarial objective: $\max_\theta \min_{\theta_d} \mathcal{L}_D(\theta, \theta_d)$, or equivalently, a global minimizer of $R(\theta)$.*

The proof is given in Appendix C.2. Intuitively, collapsing all embeddings to a single point removes all group-discriminative information, perfectly satisfying the fairness constraint.

Lemma 3.4 concerns the fairness term in isolation: it does not distinguish a collapsed encoder from an informative group-invariant one, so the anomaly-detection loss $\mathcal{L}_{\mathrm{AD}}$ is the only force in Eq. (1) that opposes collapse. The following result characterizes the full objective as the fairness coefficient grows, for any regularizer that, like the adversarial one, is minimized by a collapsed encoder.

Proposition 3.5. *Suppose $0 \le \mathcal{L}_{\mathrm{AD}}(\theta) \le B$ for all θ, and the fairness regularizer satisfies $R(\theta) \ge 0$ with $R(\theta_c) = 0$ for a collapsed encoder θ_c in the parameter space. Then any global minimizer θ^*_α of J_α in Eq. (1) satisfies $R(\theta^*_\alpha) \le B/\alpha$. For adversarial regularization with a discriminator of sufficient capacity, $R(\theta) = I\big(f_\theta(x); s\big)$, so $I\big(f_{\theta^*_\alpha}(x); s\big) \le B/\alpha$, where $I(\cdot;\cdot)$ denotes mutual information.*

The proof is given in Appendix C.3. The condition $R(\theta_c) = 0$ holds for adversarial, distribution-matching, and score-parity regularizers alike, since a collapsed encoder carries no group information and assigns identical scores to both groups. Hence, as α grows, each such regularizer suppresses group-dependent structure in the representation, which the AD loss can resist only up to the bounded amount B. In fair AD, the groups' normal patterns typically differ (e.g., MNIST vs. USPS digits), so the features needed to model them are themselves predictive of s. Suppressing them removes informative dimensions of z, consistent with the reduced-rank spectra in Figure 1, although this rank reduction is an empirical observation rather than a consequence of the proposition. Because AD relies on modeling subtle deviations from normal patterns (Audibert et al., 2020; Yan et al., 2021; Wang et al., 2023), losing these dimensions degrades detection, particularly for minority-pattern anomalies. Section 5 shows that both failure modes also appear beyond adversarial regularization: FairOD, which penalizes statistical parity of anomaly scores rather than training a discriminator, likewise shows unstable protected-group gradients and a reduced effective rank. The two failure modes are intertwined: as minority-group gradients vanish under imbalance, fairness constraints further push toward invariant representations, amplifying feature compression. This motivates an objective that explicitly preserves feature dimensionality while ensuring balanced group contributions.

4 Proposed Method

Our proposed FADIG mainly consists of two modules: a Fairness-aware Contrastive Learning Module and an Adaptively Re-balanced Autoencoder Module.

4.1 Fairness-aware Contrastive Learning

Contrastive loss has demonstrated strong benefits in learning robust and transferable representations across diverse groups, domains, and classes (Sohn et al., 2021; Chen et al., 2020). It operates by pulling similar samples closer in the embedding space while pushing dissimilar ones apart with data augmentation techniques (Khosla et al., 2020; Zhang et al., 2023). Wang & Isola (2020) show that contrastive learning encourages uniformity with negative estimation that prefers a feature distribution preserving maximal information, and Jing et al. (2022) further demonstrate the capability of contrastive learning in addressing dimensional collapse. Inspired by this, we propose a *fairness-aware contrastive loss* tailored for AD under the group imbalance scenario. Specifically, rather than relying on data augmentations, we maximize the similarity between representations of samples from the protected and unprotected groups to encourage fair and shared pattern learning. At the same time, to prevent representation collapse and preserve anomaly discrimination, we promote intra-group uniformity by pushing apart representations within each group. Formally, our fairness-aware contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{FAC}} = \underbrace{-\log\left(\frac{\sum_j \sum_k \mathrm{sim}\left(z_j^P, z_k^U\right)}{mn}\right)}_{\mathcal{L}_{\mathrm{fair}}} + \underbrace{\log\left(\frac{\sum_{j\neq k} \mathrm{sim}\left(z_j^U, z_k^U\right)}{m(m-1)} + \frac{\sum_{j\neq k} \mathrm{sim}\left(z_j^P, z_k^P\right)}{n(n-1)}\right)}_{\mathcal{L}_{\mathrm{unif}}} \tag{2}$$

The first term $\mathcal{L}_{\mathrm{fair}}$ promotes alignment across groups, encouraging the model to learn shared representations for both protected and unprotected groups to ensure fairness. The second term $\mathcal{L}_{\mathrm{unif}}$ pushes apart all representations within each group. Because each group contains unlabeled normal and anomalous samples alike, this prevents within-group representations from collapsing and preserves the diversity needed to separate anomalies from normal samples. Next, we aim to theoretically show how the fairness-aware contrastive loss alleviates two issues. First, we formalize these properties under a mild regularity assumption:

Assumption 4.1. Let $z_i = f_\theta(x_i)$ be the learned representation. Assume that (1) there exists $C > 0$ such that $\|\nabla_\theta z_i\| \le C$ for all i; (2) the similarity function $\mathrm{sim}(z_i, z_j)$ is smooth and bounded, and the gradient of $\mathrm{sim}(z_i, z_j)$ with respect to representations is uniformly bounded: $\|\nabla_{z_i}\mathrm{sim}(z_i, z_j)\| \le C'$.

Theorem 4.2. *Consider the fairness-aware contrastive loss defined in Eq. (2). Then the gradient magnitudes contributed by the protected and unprotected groups satisfy* $\left\|\nabla_\theta \mathcal{L}_{\mathrm{FAC}}^{(P)}\right\| = \mathcal{O}(1)$, *and* $\left\|\nabla_\theta \mathcal{L}_{\mathrm{FAC}}^{(U)}\right\| = \mathcal{O}(1)$, *independent of the group imbalance ratio* m/n. *Any collapsed representation* $z_i \equiv c$ *for all samples* i *is* not *an optimal solution of* $\mathcal{L}_{\mathrm{FAC}}$.

Remark. Theorem 4.2 explains why $\mathcal{L}_{\mathrm{FAC}}$ alleviates both failure modes in Section 3. First, in Lemma 3.3 the protected group is diluted by a factor of order $\sqrt{n/m}$ because the regularizer sums over all samples. In contrast, each term of $\mathcal{L}_{\mathrm{FAC}}$ is normalized by its own pair count (mn, $m(m-1)$, and $n(n-1)$), which removes this group-size factor: the bounds on both groups' gradient contributions do not depend on m/n. These are upper bounds, so they rule out domination by group size rather than guaranteeing a fixed lower bound on the protected-group gradient. Vanilla contrastive learning, lacking this normalization, still suffers from vanishing protected-group influence. Second, unlike the regularizers in Proposition 3.5, which are minimized by a collapsed encoder, the uniformity term $\mathcal{L}_{\mathrm{unif}}$ penalizes excessive within-group similarity, making collapsed representations strictly suboptimal. Together, these properties suggest that fairness and representation expressiveness need not conflict when the objective is designed at the representation level. Proof of Theorem 4.2 is provided in Appendix C.4.

4.2 Adaptively Re-balanced Autoencoder

Re-balanced autoencoder. Existing fairness-unaware AD methods overly label more examples from the protected group as anomalies than from the unprotected group, prioritizing learning frequent patterns from the unprotected group while neglecting the under-represented patterns of the protected group. To further investigate the reason, let us first look at the formulation of the reconstruction error for AD methods (Shekhar et al., 2021; Song et al., 2021) with autoencoder-based architecture formulated as follows:

$$\mathcal{L}_{\mathrm{UN-REC}} = \sum_{x_i \in P \cup U} \|x_i - G(x_i)\|^2 = \underbrace{\sum_{i=1}^{n} \|x_i^P - G\left(x_i^P\right)\|^2}_{\mathcal{L}_P} + \underbrace{\sum_{i=1}^{m} \|x_i^U - G\left(x_i^U\right)\|^2}_{\mathcal{L}_U}. \tag{3}$$

where $G = g_d \circ f_e$ is autoencoder, i.e. $G(x) = g_d(f_e(x))$. By decomposing the reconstruction loss into two terms for protected and unprotected groups respectively, we observe that the learning objective in Equation (3) mainly concentrates on learning frequent patterns of the unprotected group (i.e., $\mathcal{L}_U$) due to the more abundant examples in the unprotected group than the protected group in the imbalanced group scenario. As a consequence, it yields higher reconstruction errors for the examples from the protected group and disproportionately labels more examples from the protected group as anomalies than from the unprotected group. Furthermore, when we simply optimize the reconstruction loss in Equation (3) with the fairness loss, the model's capability of learning important frequent patterns is usually diminished by pursuing equal performance for the two groups, as shown by the fair AD baselines in Section 5. This observation inspires us to think about whether we can ensure fairness while maintaining good performance under group imbalance.

To answer this question, we propose to reweigh the reconstruction losses during training to assign different emphasizes to each group. Specifically, we design a re-balanced autoencoder by minimizing the reweighted reconstruction loss as follows:

$$\mathcal{L}_{\mathrm{REC}} = (1-\epsilon)\mathcal{L}_U + \epsilon \mathcal{L}_P, \tag{4}$$

where ϵ is a weight to adjust trade-off between learning frequent and under-represented patterns. Notice that if we set ϵ as a hyperparameter, it would be non-trivial to determine the appropriate range of ϵ. Next, we introduce an adaptive design of ϵ that introduces no learnable parameters and requires no tuning, and dynamically adjusts the trade-off by fitting normal examples from both protected and unprotected groups at each training iteration.

Adaptive weight ϵ. We theoretically characterize a range of ϵ that promotes fairness and accordingly propose an adaptive design of ϵ that is computed from data without tuning. In anomaly detection research based on autoencoders, a common assumption is that the model can effectively reconstruct normal samples well but struggles to reconstruct anomalies (Guo et al., 2023; You et al., 2022; Deng & Li, 2022). As a result, inputs with high reconstruction errors are classified as anomalies. Consider the four subgroups of data samples in the fair AD task: unprotected/protected normal examples (UN/PN) and unprotected/protected anomalies (UA/PA). Ideally, the model should only fit normal

samples, *i.e.*, the two subgroups, UN and PN. We assume that the model is capable of fitting two out of the four subgroups. Let $\mathcal{L}_0^t$ denote the loss of the unfitted model on the subgroup $t \in \{\text{UN}, \text{PN}, \text{UA}, \text{PA}\}$, and let $\mathcal{L}_1^t$ denote the loss of the fitted model on the subgroup t. $\Delta^t = \mathcal{L}_0^t - \mathcal{L}_1^t > 0$ means the difference of loss between the fitted and unfitted models on the subgroup t. We further adopt a *snapshot assumption*: at a given training iteration, the four subgroup losses are treated as separable, i.e., fitting one subgroup does not change the loss on another. This idealization is standard in stepwise reweighting and influence analyses (Ren et al., 2018; Koh & Liang, 2017). Since ϵ is recomputed at every iteration, the assumption only needs to hold locally, not over the whole training trajectory. For the design of ϵ, we have the following lemma:

Lemma 4.3. *Under the snapshot assumption, a proper weight ϵ for model fitting on normal examples in both protected and unprotected groups should be within the range $\frac{\Delta^{UA}}{\Delta^{UA}+\Delta^{PN}} < \epsilon < \frac{\Delta^{UN}}{\Delta^{UN}+\Delta^{PA}}$ such that fitting normal samples of both groups leads to a lower loss compared to fitting abnormal samples from either group. Although $\frac{\Delta^{UA}}{\Delta^{UA}+\Delta^{PN}}$ and $\frac{\Delta^{UN}}{\Delta^{UN}+\Delta^{PA}}$ are unknown, the following ϵ always lies in the desired range:*

$$\epsilon = \frac{\mathcal{L}_0^U - \mathcal{L}_U}{\mathcal{L}_0^U - \mathcal{L}_U + \mathcal{L}_0^P - \mathcal{L}_P} \tag{5}$$

where $\mathcal{L}_0^U = \mathcal{L}_0^{UN} + \mathcal{L}_0^{UA}$ and $\mathcal{L}_0^P = \mathcal{L}_0^{PN} + \mathcal{L}_0^{PA}$.

We provide the proof of Lemma 4.3 in Appendix C.5. Since the losses of the unfitted model are not observable, we instantiate them with *reference losses* computed from a model that reconstructs every sample by its group-mean reconstruction: $\mathcal{L}_0^U = \sum_{i\in U} \|x_i - \overline{G_U(x)}\|^2$, where $\overline{G_U(x)} = \frac{1}{|U|}\sum_{i\in U} G(x_i)$, and $\mathcal{L}_0^P = \sum_{i\in P} \|x_i - \overline{G_P(x)}\|^2$, where $\overline{G_P(x)} = \frac{1}{|P|}\sum_{i\in P} G(x_i)$. Other choices of reference losses are discussed in Appendix C.6.

Remark: According to the design of ϵ in Lemma 4.3, when FADIG fails to effectively learn frequent patterns of the unprotected group, the loss $\mathcal{L}_U$ becomes large and ϵ decreases accordingly, thereby assigning a larger weight to $\mathcal{L}_U$. This adjustment encourages the model to focus more on learning frequent patterns of unprotected group in subsequent iterations. Conversely, if the model overlooks under-represented patterns (*i.e.*, protected group), their corresponding weight ϵ increases, compelling the model to pay more attention to learn under-represented patterns and thus ensuring fairness. If ϵ falls below the lower bound, the unprotected group is over-weighted and its anomalies are reconstructed too well, raising their false-negative rate; above the upper bound, the same failure occurs for protected-group anomalies. The bounds themselves cannot be computed without anomaly labels, which motivates the data-driven choice in Eq. (5). Overall, the adaptive nature of ϵ dynamically prioritizes the optimization of the group that is less well-fitted during training, thus harmonizing anomaly detection with fairness.

Overall Objective. To summarize, the overall training scheme of FADIG is to minimize:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{REC}} + \alpha\mathcal{L}_{\text{FAC}},$$

where α is a hyperparameter to balance the reconstruction loss and the contrastive loss. During inference, we compute the reconstruction error for each example and rank them. The top k examples with the highest reconstruction errors are flagged as anomalies. While FADIG focuses on the binary group, we can naturally extend it to the multi-value multi-group case as discussed in Appendix C.8.

Connecting the analysis to the overall objective. (i) *Group balance.* For $A \in \{P, U\}$, let $\bar{g}_A = \|\nabla_\theta \mathcal{L}_A\|/|A|$ and $\bar{\delta}_A = (\mathcal{L}_0^A - \mathcal{L}_A)/|A|$ be the per-sample gradient scale and loss reduction. Since the adaptive weight in Lemma 4.3 satisfies $\epsilon/(1-\epsilon) = m\bar{\delta}_U/(n\bar{\delta}_P)$, the ratio of the protected to the unprotected gradient contribution is $\frac{n}{m}\cdot\frac{\bar{g}_P}{\bar{g}_U}$ under the unweighted loss in Eq. (3) but $\frac{\bar{\delta}_U}{\bar{\delta}_P}\cdot\frac{\bar{g}_P}{\bar{g}_U}$ under $\mathcal{L}_{\text{REC}}$. The group-size ratio cancels exactly, and the remaining factor up-weights whichever group has improved less per sample. Hence neither term of $\mathcal{L}_{\text{overall}}$ has group contributions that depend on the imbalance ratio. (ii) *No collapse.* The regularizers in Eq. (1) are minimized by a collapsed encoder, leaving $\mathcal{L}_{\text{AD}}$ as the only resistance (Proposition 3.5). In $\mathcal{L}_{\text{overall}}$, neither term favors collapse: a collapsed encoder cannot reconstruct its inputs and is suboptimal for $\mathcal{L}_{\text{FAC}}$ (Theorem 4.2). Fairness and utility therefore do not compete over the dimensionality of z, consistent with the joint gains observed in Section 5.

5 EXPERIMENTS

In this section, we experimentally analyze and compare our proposed FADIG with other anomaly detection methods. We try to answer the following research questions: *Q1: How does* FADIG

Table 1: Performance on image datasets. The best score is marked in bold.

Methods	MNIST-USPS (K=1200)					MNIST-Invert (K=500)				
	AUCROC	Recall@K	Recall Diff	AUPRC	AUPRC Diff	AUCROC	Recall@K	Recall Diff	AUPRC	AUPRC Diff
FairOD	50.00±0.28	12.42±0.30	11.56±0.64	12.42±0.30	5.55±0.93	50.40±0.20	12.21±3.14	8.26±1.27	16.49±5.13	8.38±13.09
DCFOD	50.09±0.27	12.63±0.33	8.99±0.83	12.70±8.72	6.49±1.39	50.54±0.54	6.95±0.91	**7.23±2.02**	6.27±0.35	3.34±1.63
FairSVDD	58.33±1.18	15.62±1.52	13.75±2.56	12.54±0.26	8.37±3.38	49.67±3.98	12.41±0.76	12.46±2.12	6.66±0.60	2.48±1.04
FairAD	69.79±0.40	35.35±2.81	28.96±10.17	29.45±1.84	6.08±2.68	62.33±3.69	16.90±3.97	15.05±22.05	11.02±2.85	10.80±14.06
MCM	80.69±0.74	54.99±0.68	18.61±1.51	56.01±0.70	7.41±0.53	92.21±0.78	60.61±1.11	27.54±7.04	60.45±1.47	8.16±6.84
NSNMF	65.38±0.58	39.16±0.84	62.90±3.84	32.81±0.12	13.33±0.45	74.21±0.34	51.79±0.61	51.07±1.79	50.27±0.73	24.29±0.73
ReContrast	83.46±3.77	64.29±3.18	41.16±5.63	25.28±1.36	9.21±0.59	85.13±5.19	64.22±1.60	56.50±11.23	22.22±0.88	14.20±0.95
FADIG	**91.28±0.46**	**67.19±0.33**	**3.77±2.18**	**66.47±0.33**	**2.52±1.09**	**97.99±0.07**	**71.82±0.63**	9.78±3.10	**72.01±0.20**	**1.72±0.59**

Table 2: Performance on tabular datasets. The best score is marked in bold.

Methods	COMPAS (K=350)					CelebA (K=5000)				
	AUCROC	Recall@K	Recall Diff	AUPRC	AUPRC Diff	AUCROC	Recall@K	Recall Diff	AUPRC	AUPRC Diff
FairOD	50.09±1.28	16.56±2.12	7.97±1.23	11.42±0.99	**0.67±0.41**	49.94±0.12	8.93±0.14	**0.68±0.56**	10.42±0.11	5.88±0.04
DCFOD	49.55±1.21	16.08±1.94	9.81±1.76	17.49±0.48	3.32±2.78	49.92±0.14	9.66±0.69	7.83±1.26	10.40±0.31	6.39±0.51
FairSVDD	52.68±5.29	15.33±2.10	11.57±4.06	17.24±2.00	8.14±7.42	58.40±1.02	10.19±0.50	10.95±1.93	**13.87±0.43**	6.95±0.41
FairAD	54.41±5.62	23.74±4.28	17.40±4.80	23.31±3.04	10.10±3.72	58.07±0.70	11.44±0.94	8.20±0.88	13.01±0.47	2.38±1.06
CFAD	48.66±1.50	14.07±2.52	39.06±34.85	16.75±0.28	2.97±1.55	53.47±1.69	9.84±1.24	28.79±16.27	12.58±1.76	2.80±1.03
MCM	50.97±0.43	19.05±0.92	6.43±3.13	20.66±0.36	8.45±0.32	50.54±2.32	11.00±0.97	10.79±4.74	10.95±0.43	5.58±1.41
NSNMF	60.76±0.31	30.15±0.15	36.04±3.33	25.48±0.50	4.17±1.89	50.45±0.30	10.91±0.54	8.04±1.33	7.93±0.35	4.86±0.29
FADIG	**61.45±0.47**	**34.38±0.36**	**5.97±4.34**	**27.42±1.56**	4.79±3.33	**59.43±0.42**	**11.96±0.49**	4.72±1.26	12.21±0.20	**1.41±0.70**

compare with other baselines on imbalanced datasets? Q2: Does FADIG *shift the performance–fairness frontier, or merely move along it? Q3: Can* FADIG *prevent representation dimensional collapse and vanishing protected-group influence?*

5.1 Experimental Setup

Datasets: We conduct experiments on two image datasets, MNIST-USPS and MNIST-Invert (Zhang & Davidson, 2021), and two tabular datasets, COMPAS (Angwin et al., 2022) and CelebA (Liu et al., 2015). The detailed characteristics of the datasets are provided in Table 6 in Appendix.

Baseline Methods: In our experiments, we compare FADIG with the following fairness-aware AD baselines: FairOD (Shekhar et al., 2021), FairSVDD (Zhang & Davidson, 2021), FairAD (Xiao et al., 2025) and CFAD (Han et al., 2023). We also compare with the following fairness-agnostic AD baselines: MCM (Yin et al., 2024), NSNMF (Ahmed et al., 2021), ReContrast (Guo et al., 2023). Notice that as ReContrast is designed for image data, we only evaluate it on MNIST-USPS and MNIST-Invert datasets. Due to the high complexity of CFAD, we do not evaluate it on two images datasets. All methods are trained on the same unlabeled data, which contains both normal samples and anomalies; anomaly labels are used only for evaluation.

Metrics: To measure the model performance and group fairness, we choose five widely-used metrics (Shekhar et al., 2021; Zhang & Davidson, 2021; Ahmed et al., 2021): (1) Recall@K, which measures the proportion of anomalies found in the top-K ranked examples; (2) AUCROC, which computes the area under the receiver operating characteristic curve; (3) Recall Diff, which measures the absolute value of the recall difference between two groups, *i.e.*, the gap in true positive rates (equal opportunity); (4) AUPRC, which computes the area under the precision-recall curve; (5) AUPRC Diff, which measures the absolute value of the AUPRC difference between two groups.

Additional results include group imbalance and ablation studies (Appendices D.3 and D.4), alternative reweighting strategies comparison (Appendix D.7), and parameter/sensitivity analyses (Appendices D.11 and D.12). Implementation details are provided in Appendix D.2.

5.2 Effectiveness of FADIG (Q1)

We first evaluate the effectiveness of FADIG through comparison with baselines across four datasets by four independent runs. The task performance (*i.e.*, AUCROC, Recall@K and AUPRC), and group fairness measure (*i.e.*, Recall Diff and AUPRC Diff) are presented in Tables 1 and 2 (see Appendix D.5 for results with different K). Two patterns emerge among the baselines. First, the fair AD baselines (*e.g.*, FairOD, DCFOD, and FairSVDD) obtain AUCROC close to 50% on most datasets, *i.e.*, near-random detection. Their disparities are therefore of limited value, since a near-random detector finds a similar fraction of anomalies in both groups. Second, fairness-agnostic baselines achieve higher detection accuracy but large group gaps, *e.g.*, a Recall Diff of 41.16% for ReContrast and 62.90% for NSNMF on MNIST-USPS. In contrast, FADIG achieves the best AUCROC and

Recall@K on all four datasets and the best AUPRC on three, while attaining the lowest disparity in five of the eight dataset–metric pairs. In the remaining cases, the baseline with lower disparity either detects at chance level (AUCROC between 48.66% and 50.54%; *e.g.*, FairOD on CelebA) or differs from FADIG by less than one standard deviation (NSNMF's AUPRC Diff on COMPAS: 4.17% vs. 4.79±3.33%). This joint improvement is expected rather than paradoxical. The detection metrics are computed over a single ranking of all samples. When a detector mainly fits the majority group, normal samples from the protected group receive high reconstruction errors and occupy top-K slots that should go to true anomalies, which lowers overall recall and widens the group gap at the same time. By balancing group contributions, FADIG reduces these false positives, so accuracy and fairness improve together. We report results of more metrics in Appendix D.6, and present comparisons with additional baselines and large-scale datasets in Appendices D.8 and D.10.

5.3 Performance–Fairness Trade-off (Q2)

To examine whether better detection must cost fairness, we vary each method's fairness coefficient over $\{0.1, 1, 10, 100, 1000\}$ on MNIST-USPS and plot performance against group disparity in Figure 2. Since coefficients are not comparable across methods, we compare the regions each method occupies rather than individual points. The baselines show the conventional trade-off or worse. As its coefficient grows, FairAD more than halves its Recall Diff, but its Recall also falls by more than half, while FairSVDD's Recall drops to below 5% without a consistent fairness gain. Both trends match Proposition 3.5: a stronger regularizer suppresses more group-dependent structure, which the AD loss resists only up to a bounded amount. FairOD and DCFOD keep small disparities at every coefficient, but with AUCROC near 50%, reflecting near-random detection in both groups rather than equitable detection. In contrast, FADIG stays in the upper-left region for $\alpha \in [0.1, 100]$, with Recall around 65% and Recall Diff below 6%. Its AUPRC Diff is comparable to that of FairOD, DCFOD, and FairAD, yet its AUPRC is more than twice that of the strongest baseline. FADIG degrades only at $\alpha = 1000$, where the contrastive term dominates reconstruction, and even then its Recall exceeds every baseline configuration. Thus, FADIG shifts the performance–fairness frontier rather than moving along it, consistent with Section 4: neither term of $\mathcal{L}_{\text{overall}}$ favors collapse, so fairness and accuracy do not compete over the representation.

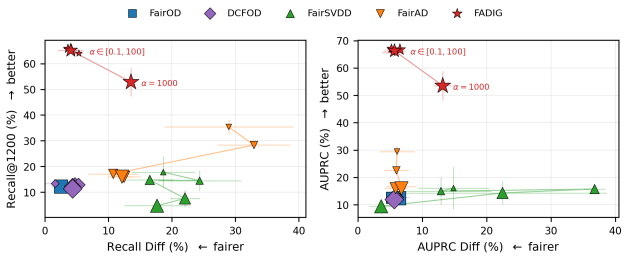
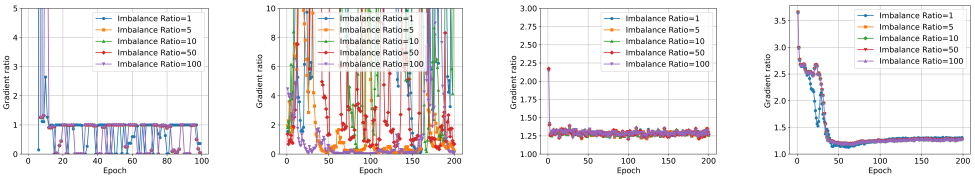


Figure 2: Performance–fairness trade-off on MNIST-USPS as the fairness coefficient of each method varies. Marker size increases with the coefficient.

5.4 Mitigating Two Fundamental Issues (Q3)

To empirically validate the identified failure modes, we design controlled diagnostics that track optimization dynamics during training. To measure *vanishing protected-group influence*, we monitor the feature update magnitude of each group and report the gradient ratio $\|\nabla Z_P\|/\|\nabla Z_U\|$, a scale-invariant indicator of protected-group influence. Figure 3 shows the evolution of this ratio across training epochs on MNIST-USPS dataset. A ratio close to zero indicates negligible protected-group contribution. As the imbalance ratio $\frac{|U|}{|P|}$ increases from 1 to 100, FairSVDD exhibits delayed gradient collapse: the ratio remains high early in training and only fluctuates between 0 and 1 after about epoch 15. In contrast, FairOD displays severe instability, with gradient ratios ranging from 0 to over 200 across imbalance settings. Since FairOD does not use adversarial training, this indicates that the failure mode is not specific to adversarial regularization. By comparison, FADIG and DCFOD maintain stable gradient ratios across all imbalance levels. We attribute the robustness of FADIG to the per-group normalization of $\mathcal{L}_{\text{FAC}}$ (Theorem 4.2) and the group-size cancellation under ϵ, while DCFOD benefits from its sample-wise reweighting scheme.

To assess *dimensional collapse*, we compute the effective rank of the embedding matrix at convergence. Because the coefficients of fairness regularization are method-specific and not directly comparable, we plot effective rank as a function of the achieved fairness level (measured by recall difference) rather than raw regularization strength. We conduct experiments on the MNIST–USPS



(a) FairSVDD (b) FairOD (c) DCFOD (d) FADIG

Figure 3: Visualization of gradient ratio ($\|\nabla Z_P\|/\|\nabla Z_U\|$) across training epochs on MNIST-USPS.

dataset with group ratio $\frac{|U|}{|P|} \approx 4.1$, varying the regularization coefficients of fair AD methods, and report the results in Figure 4. We observe that FairSVDD and FairOD consistently yield lower effective rank than DCFOD and FADIG. For these baselines, lower rank comes with lower recall, revealing the cost of dimensional collapse, while higher rank comes with a higher recall difference: each baseline moves along a trade-off between expressiveness and fairness. These results support the analysis in Section 3 (Proposition 3.5): regularizers that enforce group invariance suppress group-dependent structure and reduce the dimensionality of the representation. In contrast, FADIG remains near full rank across a wide range of contrastive regularization strengths while keeping the recall difference low, and thus escapes this trade-off. DCFOD also retains a relatively high rank, yet its AUCROC stays near 50% (Table 1), indicating that avoiding collapse alone is not sufficient for accurate detection; balanced group contributions and an informative alignment objective are also needed.

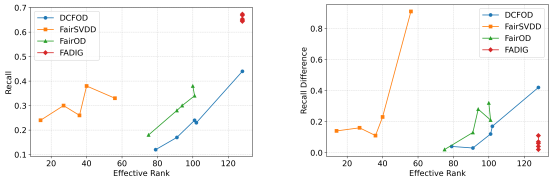


Figure 4: Visualization of dimensional collapse vs Performance and Fairness. **Left:** Rank vs Recall Rate; **Right:** Rank vs Recall Difference.

6 Related Work

In this section, we briefly review the related work on unsupervised anomaly detection and fair machine learning. A more extensive discussion is provided in Appendix A.

Unsupervised Anomaly Detection does not require labeled data and has been widely applied in domains such as cybersecurity, healthcare, and finance (West & Bhattacharya, 2016; Faust et al., 2018; Shvetsova et al., 2021; Ahmad et al., 2021). As surveyed in Pang et al. (2022), most approaches either learn representations of normal patterns using feature extractors (Audibert et al., 2020; Chen et al., 2021; Hou et al., 2021; Yan et al., 2021; Wang et al., 2023), or directly learn scalar anomaly scores in an end-to-end fashion (Sohn et al., 2021; Li et al., 2023; Jiang et al., 2022). Our work extends autoencoder-based reconstruction models, but crucially introduces two fairness-driven innovations: adaptive reweighting that balances groups without additional parameters or hyperparameters, and a contrastive objective for shared group representations without conventional data augmentations.

Fair Machine Learning aims to mitigate bias and ensure equitable outcomes with respect to protected attributes (Zemel et al., 2013; Bolukbasi et al., 2016; Hashimoto et al., 2018; Zhang et al., 2018; Kobren et al., 2019). Within AD, recent works have begun to explore fairness (Shekhar et al., 2021; Zhang et al., 2023; Yin et al., 2024). However, these methods often suppress group-dependent information without considering its role in preserving informative representations, which may lead to dimensional collapse and degraded detection performance. We instead formulate fair AD from a representation-learning perspective, using adaptive reconstruction reweighting and fairness-aware contrastive learning to balance group contributions while preserving representation diversity.

7 Conclusion

We traced the performance–fairness trade-off in unsupervised anomaly detection to two representation-level failure modes of existing fairness regularizers: vanishing protected-group influence under imbalance and dimensional collapse. We proposed FADIG combining a fairness-aware contrastive loss with an adaptively re-balanced autoencoder and proved that its group contributions do not depend on the imbalance ratio and that collapsed representations are suboptimal. Experiments on four datasets show that FADIG achieves the best detection accuracy with low group disparity.

8 AI USE STATEMENT

In this work, we used generative AI tools for improving the writing and readability. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, or interpret results. Assisting with translation and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-generated polishing of the writing and manually verified the content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David Acuna, Guojun Zhang, Marc T Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, pp. 66–75. PMLR, 2021.

Zeeshan Ahmad, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, and Farhan Ahmad. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Trans. Emerg. Telecommun. Technol.*, 32(1), 2021.

Imtiaz Ahmed, Xia Ben Hu, Mithun P. Acharya, and Yu Ding. Neighborhood structure assisted non-negative matrix factorization and its application in unsupervised point-wise anomaly detection. *J. Mach. Learn. Res.*, 22:34:1–34:32, 2021.

Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31, 2016.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28 (1):131–142, 1966.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: unsupervised anomaly detection on multivariate time series. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 3395–3404. ACM, 2020.

Maryam Badar, Sandipan Sikdar, Wolfgang Nejdl, and Marco Fisichella. Fairtrade: Achieving pareto-optimal trade-offs between balanced accuracy and fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10962–10970, 2024.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4349–4357, 2016.

Roel Bouman, Zaharah Bukhsh, and Tom Heskes. Unsupervised anomaly detection algorithms on real-world data: how many do we need? *Journal of Machine Learning Research*, 25(105):1–34, 2024.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Huiyuan Chen, Vivian Lai, Hongye Jin, Zhimeng Jiang, Mahashweta Das, and Xia Hu. Towards mitigating dimensional collapse of representations in collaborative filtering. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 106–115, 2024.

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

Wei Chen, Meng Yuan, Zhao Zhang, Ruobing Xie, Fuzhen Zhuang, Deqing Wang, and Rui Liu. Fairdgcl: Fairness-aware recommendation with dynamic graph contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Xuanhao Chen, Liwei Deng, Feiteng Huang, Chengwei Zhang, Zongquan Zhang, Yan Zhao, and Kai Zheng. DAEMON: unsupervised anomaly detection and interpretation for multivariate time series. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*, pp. 2225–2230. IEEE, 2021.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9737–9746, 2022.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM international conference on data mining*, pp. 594–602. SIAM, 2019.

Jingcan Duan, Siwei Wang, Pei Zhang, En Zhu, Jingtao Hu, Hu Jin, Yue Liu, and Zhibin Dong. Graph anomaly detection via multi-scale contrastive learning networks with augmented view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7459–7467, 2023.

Kevin Faust, Quin Xie, Dominick Han, Kartikay Goyle, Zoya I. Volynskaya, Ugljesa Djuric, and Phedias Diamandis. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinform.*, 19(1):173:1–173:15, 2018.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 13552–13563. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01251.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1705–1714, 2019.

Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *arXiv preprint arXiv:2306.02602*, 2023.

Xiao Han, Lu Zhang, Yongkai Wu, and Shuhan Yuan. Achieving counterfactual fairness for anomaly detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 55–66. Springer, 2023.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1934–1943. PMLR, 2018.

Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. *Advances in Neural Information Processing Systems*, 37:95579–95606, 2024.

Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8771–8780. IEEE, 2021.

Dongxu Huang, Dejun Mu, Libin Yang, and Xiaoyan Cai. Codetect: Financial fraud detection with anomaly feature detection. *IEEE Access*, 6:19161–19174, 2018.

Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Yamei Ju, Xin Tian, Hongjian Liu, and Lifeng Ma. Fault detection of networked dynamical systems: a survey of trends and techniques. *Int. J. Syst. Sci.*, 52(16):3390–3409, 2021.

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

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pp. 180–191. Toronto, Canada, 2004.

Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 1247–1257. ACM, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. Pmlr, 2017.

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Loop: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1649–1652, 2009.

Jundong Li, Xia Hu, Jiliang Tang, and Huan Liu. Unsupervised streaming feature selection in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1041–1050, 2015.

Yuxin Li, Wenchao Chen, Bo Chen, Dongsheng Wang, Long Tian, and Mingyuan Zhou. Prototype-oriented unsupervised anomaly detection for multivariate time series. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19407–19424. PMLR, 2023.

Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.*, 36(1):16–24, 2013.

Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL `http://arxiv.org/abs/1511.00830`.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, pp. 6755–6764. PMLR, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2):38:1–38:38, 2022.

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10389–10398, 2022.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Mahdi Rezapour. Anomaly detection using unsupervised methods: credit card fraud case study. *International Journal of Advanced Computer Science and Applications*, 10(11), 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shubhranshu Shekhar, Neil Shah, and Leman Akoglu. Fairod: Fairness-aware outlier detection. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 210–220, 2021.

Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access*, 9:118571–118583, 2021.

Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hanyu Song, Peizhao Li, and Hongfu Liu. Deep clustering based fair outlier detection. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao (eds.), *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 1481–1489. ACM, 2021. doi: 10.1145/3447548.3467225.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Rui Wang, Chongwei Liu, Xudong Mou, Kai Gao, Xiaohui Guo, Pin Liu, Tianyu Wo, and Xudong Liu. Deep contrastive one-class time series anomaly detection. In Shashi Shekhar, Zhi-Hua Zhou, Yao-Yi Chiang, and Gregor Stiglic (eds.), *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, pp. 694–702. SIAM, 2023.

Ruijia Wang, Xiao Wang, Chuan Shi, and Le Song. Uncovering the structural fairness in graph contrastive learning. *Advances in neural information processing systems*, 35:32465–32473, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757, 2021.

Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: A comprehensive review. *Comput. Secur.*, 57:47–66, 2016.

Feng Xiao, Xiaoying Tang, and Jicong Fan. Fairness-aware anomaly detection via fair projection. *arXiv preprint arXiv:2505.11132*, 2025.

Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 3110–3118. AAAI Press, 2021.

Jiaxin Yin, Yuanyuan Qiao, Zitang Zhou, Xiangchao Wang, and Jie Yang. Mcm: Masked cell modeling for anomaly detection in tabular data. In *The Twelfth International Conference on Learning Representations*, 2024.

Sangwoong Yoon, Yung-Kyun Noh, and Frank Park. Autoencoding under normalization constraints. In *International Conference on Machine Learning*, pp. 12087–12097. PMLR, 2021.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584, 2022.

Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 325–333. JMLR.org, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pp. 335–340. ACM, 2018.

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2022.

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2023.

Hongjing Zhang and Ian Davidson. Towards fair deep anomaly detection. In Madeleine Clare Elish, William Isaac, and Richard S. Zemel (eds.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp. 138–148. ACM, 2021.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

A Related Work

A.1 Unsupervised Anomaly Detection

Anomaly detection (AD) has been studied for decades across applications such as fraud detection (West & Bhattacharya, 2016; Huang et al., 2018), medical diagnosis (Faust et al., 2018; Shvetsova et al., 2021), intrusion detection (Liao et al., 2013; Ahmad et al., 2021), and fault detection in safety-critical systems (Ju et al., 2021). Given various types of anomalies (Breunig et al., 2000; Kriegel et al., 2009; Bouman et al., 2024), the authors of Pang et al. (2022) divide the existing anomaly detection methods into two major branches. The first branch focuses on learning representations of normal samples via feature extractors (Audibert et al., 2020; Chen et al., 2021; Hou et al., 2021; Yan et al., 2021; Wang et al., 2023). For example, Audibert et al. (2020) train an encoder–decoder to amplify reconstruction error through adversarial training, while Chen et al. (2021) propose a GAN-based autoencoder to model multivariate time series. Our method also builds on autoencoder-based reconstruction, but introduces an adaptive reweighting scheme to balance groups and a fairness-aware contrastive regularizer to reduce representation disparity between groups. The second branch aims at learning scalar anomaly scores in an end-to-end fashion (Sohn et al., 2021; Li et al., 2023; Jiang et al., 2022). Notably, the authors of Sohn et al. (2021) combine distribution-augmented contrastive regularization with a one-class classifier to detect anomalies. While Sohn et al. (2021) uses image augmentations, such as rotations, to form positive pairs and negative pairs for contrastive learning, we instead use existing examples for contrastive learning and thus our method is applicable to various types of data, not limited to image data. To be more specific, we design a fairness-aware contrastive learning loss which minimizes the representation disparity of the groups for fairness, and encourages the uniformity within each group for better anomaly detection.

A.2 Fair Machine Learning

Fair Machine Learning aims to amend the biased machine learning models to be fair or invariant regarding specific variables. A surge of research in fair machine learning has been done in the machine learning community (Kobren et al., 2019; Zemel et al., 2013; Bolukbasi et al., 2016; Hashimoto et al., 2018; Zhang et al., 2018; Park et al., 2022). For example, Zemel et al. (2013) presents a learning algorithm for fair classification by enforcing group fairness and individual fairness in the obtained data representation; Hashimoto et al. (2018) develop a robust optimization framework that minimizes the worst case risk over all distributions and preserves the minority group in an imbalanced data set; Zhang et al. (2018) present an adversarial-learning based framework for mitigating the undesired bias in modern machine learning models. Park et al. (2022) propose a fair supervised contrastive loss to train a fair representation model. However, they rely on target labels and need negative samples since their method is based on supervised contrastive learning. Instead, our designed fair contrastive learning loss uses examples from different groups in an unsupervised way to minimize the representation disparity of the groups and encourage the uniformity within each group.

Several studies investigate fairness–utility trade-offs using multi-objective optimization (Badar et al., 2024; Liu & Vicente, 2022; Martinez et al., 2020; Wang et al., 2021). These works characterize Pareto-efficient trade-offs between accuracy and fairness across various tasks. While we do not explicitly optimize Pareto fronts, our method harmonizes fairness and utility in the unsupervised AD setting by automatically rebalancing groups and aligning their latent representations.

In the specific field of fair anomaly detection, Zhang & Davidson (2021) utilizes the adversarial generative nets to ensure group fairness and use one-class classification to detect the anomalies; Song et al. (2021) introduces fairness adversarial training and proposes a dynamic weight to reduce the negative impacts from outlier points. However, these existing fair anomaly detection methods (Song et al., 2021; Zhang & Davidson, 2021; Fioresi et al., 2023) tend to suffer from the representation disparity issue when groups are imbalanced. Our approach directly addresses this by combining (i) a hyperparameter-free, parameter-free reweighted autoencoder that adaptively balances groups with theoretical guarantees, and (ii) a fairness-driven contrastive objective that maximizes similarity between protected and unprotected groups while maintaining within-group uniformity.

While the techniques of loss reweighting and contrastive learning have been extensively studied, we contribute to the novel design of both modules through the lens of fairness under imbalance. For loss reweighting, most existing loss reweighting methods (Song et al., 2021; Zhang & Davidson, 2021)

Table 3: Notation Table.

Symbol	Description
x	Input feature
$\mathcal{P}_P$	Protected group's distribution
$\mathcal{P}_U$	Unprotected group's distribution
P	Protected group's empirical distribution
U	Unprotected group's empirical distribution
n/m	Size of protected/unprotected group
a_P/a_U	labeling functions on protected/unprotected group
ℓ	Loss function
$R_D^\ell(h)$	Risk of hypothesis h over data D
$\hat{R}_D^\ell(h)$	Empirical risk of hypothesis h over data D
$s(x)$	Anomaly score of example x
$\mathfrak{R}_D(\mathcal{F})$	Rademacher complexity of $\mathcal{F}$ given data D
$D_f(P \parallel Q)$	f-divergence between distributions P and Q

introduce a learnable vector to learn the sample-wise weight or introduce hyperparameters for samples from different unprotected and protected groups. Our proposed method presents a hyperparameter-free and parameter-free reweighting scheme by automatically and adaptively weighing the importance of unprotected and protected groups with a theoretical guarantee. Compared with the existing reweighting methods, our method either reduces the number of parameters in the model or gets rid of the hyperparameter for group weight. For contrastive learning, we design the contrastive learning loss from the fairness perspective. The traditional contrastive learning loss uses two augmentation methods to generate two views to form the positive pairs, maximizes the similarity of positive pairs, and minimizes the similarity of negative pairs. Most of the existing fairness methods (Wang et al., 2022; Zhang et al., 2022; Chen et al., 2025) adopt this procedure, which largely rely on augmentations and do not directly encourage fairness. In our method, we maximize the similarity between representations of samples from the protected and unprotected groups (instead of two augmented views) to directly encourage fair and shared pattern learning. Furthermore, our design alleviates the impact of the imbalanced groups.

B Limitations and Reproducibility

This paper proposes a fairness-aware anomaly detection method, which aims to provide fair results when the algorithm is applied to detect anomalies. Our method currently focus on the binary group fairness case. We can naturally extend our framework to the multi-value multi-group case by extending our design of $\mathcal{L}_{\text{FAC}}$ and $\mathcal{L}_{\text{REC}}$. Incorporating individual fairness notions would be an interesting future direction. For Reproducibility, in Appendix D.2 we present the experimental setup, training details, and implementation code. In Appendix C, we include the proofs for all the theoretical results.

C Proofs

C.1 Proof for Lemma 3.3

Proof. Let ℓ_i denote the per-sample discriminator loss. Then the gradient of the discriminator loss decomposes as

$$\nabla_\theta R = \frac{1}{M}\big(\nabla_\theta R_U + \nabla_\theta R_P\big), \qquad \nabla_\theta R_A = \sum_{i \in A} \nabla_\theta \ell_i.$$

By Assumption 3.1, the ℓ_i are i.i.d., and each $\nabla_\theta \ell_i$ is bounded by k. Thus, $\nabla_\theta R_P$ is a sum of n bounded (possibly dependent) random vectors. Define zero-mean vectors

$$X_i := \nabla_\theta \ell_i - \mathbb{E}[\nabla_\theta \ell_i].$$

Then, $\nabla_\theta R_P - n\mathbb{E}[\nabla_\theta \ell_i] = \sum_{i \in P} X_i$ is a sum of independent, bounded vectors $\|X_i\| \leq 2k$. By the matrix Bernstein inequality:

$$\mathbb{P}\left[\Big\| \sum_{i \in P} X_i \Big\| \geq t\right] \leq 2d \exp\left(\frac{-t^2/2}{\sum_{i \in P} \mathbb{E}[\|X_i\|^2] + (2k)t/3}\right), \tag{6}$$

where d is the dimension of θ.

The expected norm can then be written as

$$\mathbb{E}\Big[\Big\| \sum_{i \in P} X_i \Big\|\Big] = \int_0^\infty \mathbb{P}\Big[\Big\| \sum_{i \in P} X_i \Big\| > t\Big] dt.$$

Plugging in the tail bound (6) and integrating yields

$$\mathbb{E}\Big[\Big\| \sum_{i \in P} X_i \Big\|\Big] \leq \mathcal{O}\big(\sqrt{n}k\big).$$

Adding back the mean shift $n\mathbb{E}[\nabla_\theta \ell_i]$ only affects constants by Assumption 3.1, so we have

$$\mathbb{E}\Big[\|\nabla_\theta R_P\|\Big] \leq \mathcal{O}(\sqrt{n}).$$

For the unprotected group, Assumption 3.2 guarantees

$$\mathbb{E}\Big[\|\nabla_\theta R_U\|\Big] \geq c\sqrt{m}.$$

Combining these bounds yields

$$\mathbb{E}\left[\frac{\|\nabla_\theta R_P\|}{\|\nabla_\theta R_U\|}\right] \leq \frac{\mathcal{O}(\sqrt{n})}{c\sqrt{m}} = \mathcal{O}\Big(\sqrt{\frac{n}{m}}\Big).$$

Taking the limit $m/n \to \infty$ completes the proof. □

C.2 Proof for Lemma 3.4

Proof. If $f_\theta(x) = c$, then the discriminator output $\mathcal{D}_{\theta_d}(f_\theta(x)) = \mathcal{D}_{\theta_d}(c)$ is independent of the sensitive attribute Z. For any fixed f_θ, the optimal discriminator minimizes $\mathcal{L}_D$ by predicting the empirical prior:

$$g^*(c) = \mathbb{P}(Z = 1).$$

Let $p = \mathbb{P}(Z = 1)$ denote the marginal probability of the sensitive attribute. Substituting the constant prediction q into loss function, the loss reduces to

$$\mathcal{L}_D(q) = -\Big[p \log q + (1 - p) \log(1 - q)\Big]. \tag{7}$$

The discriminator minimizes $\mathcal{L}_D(q)$ with respect to q. Taking the derivative and setting it to zero yields

$$\frac{\partial \mathcal{L}_D(q)}{\partial q} = -\left(\frac{p}{q} - \frac{1-p}{1-q}\right) = 0,$$

which implies $q^* = p$. Substituting the optimal discriminator prediction $q = p$ into Eq. (7), we obtain

$$\max_{\theta} \mathcal{L}_D(\theta, \theta_d) = -\Big[p \log p + (1-p)\log(1-p)\Big] = H(Z),$$

where $H(Z)$ denotes the Shannon entropy of the sensitive attribute Z. Based on the basic entropy inequality $H(Z|f_\theta(x)) \leq H(Z)$ with equality if and only if $Z \perp\!\!\!\perp f_\theta(x)$. Therefore, when the representation $f_\theta(X)$ contains no information about Z, the discriminator's optimal loss equals the entropy of Z, which is the maximum achievable loss. As a result, collapsed representations are global optima of the adversarial fairness objective

$$\max_{\theta} \min_{\theta_d} \mathcal{L}_D(\theta, \theta_d),$$

which completes the proof. □

C.3 Proof of Proposition 3.5

Proof. Since θ^*_α is a global minimizer of J_α and $\mathcal{L}_{\mathrm{AD}} \geq 0$,

$$\alpha R(\theta^*_\alpha) \leq J_\alpha(\theta^*_\alpha) \leq J_\alpha(\theta_c) = \mathcal{L}_{\mathrm{AD}}(\theta_c) + \alpha R(\theta_c) = \mathcal{L}_{\mathrm{AD}}(\theta_c) \leq B,$$

so $R(\theta^*_\alpha) \leq B/\alpha$. For adversarial regularization, a discriminator of sufficient capacity outputs $\mathcal{D}^*(z) = \Pr(s = 1 \mid z)$, so the minimal cross-entropy is the conditional entropy, $\min_{\theta_d} \mathcal{L}_D(\theta, \theta_d) = H\big(s \mid f_\theta(x)\big)$. Therefore

$$R(\theta) = H(s) - H\big(s \mid f_\theta(x)\big) = I\big(f_\theta(x); s\big) \geq 0,$$

which equals zero for a collapsed encoder (recovering Lemma 3.4), and the bound becomes $I\big(f_{\theta^*_\alpha}(x); s\big) \leq B/\alpha$. □

C.4 Proof for Theorem 4.2

Proof. The contrastive learning loss is formulated as

$$\mathcal{L}_{\mathrm{FAC}} = \mathcal{L}_{\mathrm{fair}} + \mathcal{L}_{\mathrm{unif}},$$

where

$$\mathcal{L}_{\mathrm{fair}} = -\log(\frac{1}{mn}\sum_{i,j} \mathrm{sim}(z_i^P, z_j^U)),$$

and

$$\mathcal{L}_{\mathrm{unif}} = \log(A_U + A_P), \quad A_U = \frac{1}{m(m-1)}\sum_{j \neq k} \mathrm{sim}(z_j^U, z_k^U), \quad A_P = \frac{1}{n(n-1)}\sum_{j \neq k} \mathrm{sim}(z_j^P, z_k^P).$$

(1) To prove gradient scaling under imbalance, we define

$$S_{PU} = \frac{1}{mn}\sum_{i,j} \mathrm{sim}(z_i^P, z_j^U).$$

By Assumption 4.1 and the chain rule, we have

$$\|\nabla_\theta \mathrm{sim}(z_i, z_j)\| \leq \|\nabla_{z_i} \mathrm{sim}\| \cdot \|\nabla_\theta z_i\| + \|\nabla_{z_j} \mathrm{sim}\| \cdot \|\nabla_\theta z_j\| \leq G,$$

uniformly for all pairs (i, j), where $G = 2CC'$.
Next, we want to bound the gradient for protected-group and unprotected-group from $\mathcal{L}_{\mathrm{fair}}$. The fair term is

$$\mathcal{L}_{\mathrm{fair}} = -\log S_{PU}, \quad S_{PU} = \frac{1}{mn}\sum_{i,j} \mathrm{sim}(z_i^P, z_j^U).$$

Its per gradient for protected group is

$$\nabla_\theta \mathcal{L}^{(P)}_{\mathrm{fair}} = -\frac{1}{S_{PU}} \cdot \frac{1}{mn}\sum_{i,j} \nabla_\theta \mathrm{sim}(z_i^P, z_j^U).$$

Since $\mathrm{sim}(z_i, z_j) = \exp(\cos(z_i, z_j)) \in [e^{-1}, e]$, we have $S_{PU} \geq e^{-1}$. As there are mn terms of $\|\nabla_\theta \mathrm{sim}(z_i^P, z_j^U)\|$, each with gradient $\leq G$, we have

$$\|\nabla_\theta \mathcal{L}_{\text{fair}}^{(P)}\| \leq \frac{1}{S_{PU}} \cdot \frac{1}{mn} \cdot mn\, G \leq eG = \mathcal{O}(1).$$

Similarly, the unprotected group satisfies

$$\|\nabla_\theta \mathcal{L}_{\text{fair}}^{(U)}\| = \mathcal{O}(1).$$

Similarly, we can bound the gradient for two groups from $\mathcal{L}_{\text{unif}}$. We have

$$\mathcal{L}_{\text{unif}} = \log(A_P + A_U), \quad \nabla_\theta \mathcal{L}_{\text{unif}}^{(P)} = \frac{1}{A_P + A_U} \nabla_\theta A_P.$$

Since $A_P + A_U \geq 2e^{-1}$ and

$$\|\nabla_\theta A_P\| \leq \frac{1}{n(n-1)} \sum_{i \neq j} \|\nabla_\theta \mathrm{sim}(z_i^P, z_j^P)\| \leq G,$$

it follows that

$$\|\nabla_\theta \mathcal{L}_{\text{unif}}^{(P)}\| \leq \frac{G}{2e^{-1}} = eG/2 = \mathcal{O}(1).$$

Similarly, $\|\nabla_\theta \mathcal{L}_{\text{unif}}^{(U)}\| = \mathcal{O}(1)$.
Finally,

$$\|\nabla_\theta \mathcal{L}_{\text{FAC}}^{(P)}\| \leq \|\nabla_\theta \mathcal{L}_{\text{fair}}^{(P)}\| + \|\nabla_\theta \mathcal{L}_{\text{unif}}^{(P)}\| = \mathcal{O}(1),$$

and

$$\|\nabla_\theta \mathcal{L}_{\text{FAC}}^{(U)}\| = \mathcal{O}(1),$$

independent of the ratio m/n.

(2). To prove the mitigation of feature collapse, consider a collapsed representation $z_i \equiv c$ for all samples i. Then for any $i \neq j$,

$$\mathrm{sim}(z_i, z_j) = \exp(1),$$

which implies

$$A_U = A_P = \exp(1), \quad \mathcal{L}_{\text{unif}} = \log(2\exp(1)).$$

For any non-collapsed perturbation $\{z_i'\}$ such that $z_i' \neq z_j'$ for some $i \neq j$ within either group, cosine similarity strictly decreases, yielding

$$\mathrm{sim}(z_i', z_j') < \exp(1).$$

Consequently, either A_U or A_P strictly decreases, which implies

$$\mathcal{L}_{\text{unif}}(z') < \mathcal{L}_{\text{unif}}(z).$$

Since $\mathcal{L}_{\text{fair}}$ is constant under collapse, the total loss $\mathcal{L}_{\text{con}}$ is strictly reduced by breaking collapse. Hence, collapsed representations cannot be an optimal solution of $\mathcal{L}_{\text{FAC}}$.

Combining both parts completes the proof. □

C.5 Proof of Lemma 4.3

Let us divide the data into four subgroups: unprotected normal examples (UN), protected normal examples (PN), unprotected anomalies (UA), and protected anomalies (PA). For subgroup $t \in \{\text{UN}, \text{PN}, \text{UA}, \text{PA}\}$, let $\mathcal{L}_0^t$ denote the loss of the unfitted model on t, and $\mathcal{L}_1^t$ denote the loss of the fitted model on t. $\Delta^t = \mathcal{L}_0^t - \mathcal{L}_1^t > 0$ means the difference of loss between the fitted model and the unfitted one on the subgroup t. Assuming that the model can only fit two sets of data, to ensure that the model fits the sets of protected normal examples and unprotected normal examples, we need the following 5 inequalities to hold:

$(1-\epsilon)(\mathcal{L}_1^{UN} + \mathcal{L}_0^{UA}) + \epsilon(\mathcal{L}_1^{PN} + \mathcal{L}_0^{PA}) <$

Table 4: Performance of FADIG with different designs of $\mathcal{L}_0^U$ and $\mathcal{L}_0^P$.

Methods	MNIST-USPS (K=1200)			MNIST-Invert (K=500)		
	Recall@K	AUCROC	Rec Diff	Recall@K	AUCROC	Rec Diff
loss1	67.19±0.33	91.28±0.46	3.77±2.18	71.82±0.63	97.99±0.07	9.78±3.10
loss2	66.47±1.73	90.60±0.52	4.78±2.36	72.44±0.74	98.04±0.03	7.22±0.21
loss3	66.31±0.65	91.37±0.88	6.32±1.74	71.39±1.96	97.22±1.42	8.95±0.92
loss4	66.56±2.32	90.88±1.67	2.54±2.11	71.92±3.58	97.01±1.85	8.96±3.23

(1) $(1-\epsilon)(\mathcal{L}_0^{UN}+\mathcal{L}_1^{UA})+\epsilon(\mathcal{L}_1^{PN}+\mathcal{L}_0^{PA})$, implied by $\Delta^{UN}>\Delta^{UA}$ which naturally holds;
(2) $(1-\epsilon)(\mathcal{L}_1^{UN}+\mathcal{L}_0^{UA})+\epsilon(\mathcal{L}_0^{PN}+\mathcal{L}_1^{PA})$, implied by $\Delta^{PN}>\Delta^{PA}$ which naturally holds;
(3) $(1-\epsilon)(\mathcal{L}_0^{UN}+\mathcal{L}_1^{UA})+\epsilon(\mathcal{L}_0^{PN}+\mathcal{L}_1^{PA})$, this case is equivalent to case (1) plus case (2);
(4) $(1-\epsilon)(\mathcal{L}_1^{UN}+\mathcal{L}_1^{UA})+\epsilon(\mathcal{L}_0^{PN}+\mathcal{L}_0^{PA})$, we need $\epsilon>\frac{\Delta^{UA}}{\Delta^{UA}+\Delta^{PN}}$;
(5) $(1-\epsilon)(\mathcal{L}_0^{UN}+\mathcal{L}_0^{UA})+\epsilon(\mathcal{L}_1^{PN}+\mathcal{L}_1^{PA})$, we need $\epsilon<\frac{\Delta^{UN}}{\Delta^{UN}+\Delta^{PA}}$.

Inequalities (1)(2)(3) are implied by our assumptions regardless of ϵ and inequalities (4)(5) give the range of ϵ that $\frac{\Delta^{UA}}{\Delta^{UA}+\Delta^{PN}}<\epsilon<\frac{\Delta^{UN}}{\Delta^{UN}+\Delta^{PA}}$. We design $\epsilon=\frac{\mathcal{L}_0^U-\mathcal{L}_U}{\mathcal{L}_0^U-\mathcal{L}_U+\mathcal{L}_0^P-\mathcal{L}_P}$, and we discuss the following three cases:

- If $\mathcal{L}_U=\mathcal{L}_1^{UN}+\mathcal{L}_0^{UA}, \mathcal{L}_P=\mathcal{L}_1^{PN}+\mathcal{L}_0^{PA}$, then $\epsilon=\frac{\Delta^{UN}}{\Delta^{UN}+\Delta^{PN}}$, which is within the range;
- If $\mathcal{L}_U=\mathcal{L}_1^{UN}+\mathcal{L}_1^{UA}, \mathcal{L}_P=\mathcal{L}_0^{PN}+\mathcal{L}_0^{PA}$, then $\epsilon=1$, it encourages to fit $\mathcal{L}_P$;
- If $\mathcal{L}_U=\mathcal{L}_0^{UN}+\mathcal{L}_0^{UA}, \mathcal{L}_P=\mathcal{L}_1^{PN}+\mathcal{L}_1^{PA}$, then $\epsilon=0$, it encourages to fit $\mathcal{L}_U$.

C.6 Estimations of $\mathcal{L}_0^U$ and $\mathcal{L}_0^P$

We estimate $\mathcal{L}_0^U=\sum_{i\in U}\|x_i-\overline{G_U(x)}\|^2$ where $\overline{G_U(x)}=\frac{1}{|U|}\sum_{i\in U}G(x_i)$, and $\mathcal{L}_0^P=\sum_{i\in P}\|x_i-\overline{G_P(x)}\|^2$ where $\overline{G_P(x)}=\frac{1}{|P|}\sum_{i\in P}G(x_i)$. Let us denote this by loss1. We also test different designs of $\mathcal{L}_0^U$ and $\mathcal{L}_0^P$ as follows:

- loss2: $\mathcal{L}_0^U=\sum_{i\in U}\|x_i\|^2$ and $\mathcal{L}_0^P=\sum_{i\in P}\|x_i\|^2$
- loss3: $\mathcal{L}_0^U=\sum_{i\in U}\|G(x_i)-\overline{x}\|^2$ and $\mathcal{L}_0^P=\sum_{i\in P}\|G(x_i)-\overline{x}\|^2$
- loss4: $\mathcal{L}_0^U=\sum_{i\in U}\|x_i-\overline{x}\|^2$ and $\mathcal{L}_0^P=\sum_{i\in P}\|x_i-\overline{x}\|^2$

We provide results on real-world datasets with these different estimation designs in Table 4. We can see that although the results may vary with different estimation designs, our method always performs better than the baselines in both task performance and fairness.

C.7 Additional Fairness Analysis

In this subsection, we analyze how $\mathcal{L}_{\text{FAC}}$ promotes group fairness.

Before we move forward, let's first define some necessary notations. The risk of a hypothesis h w.r.t. the true labeling function a under distribution $\mathcal{D}$ using a loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_+$ is defined as: $R_{\mathcal{D}}^{\ell}(h,a)\coloneqq\mathbb{E}_{x\sim\mathcal{D}}\left[\ell(h(x),a(x))\right]$. We assume that ℓ satisfies the triangle inequality. For notation simplicity, we denote $R_P^{\ell}(h)\coloneqq R_{\mathcal{P}_P}^{\ell}(h,a_P)$ and $R_U^{\ell}(h)\coloneqq R_{\mathcal{P}_U}^{\ell}(h,a_U)$. The empirical risks over the protected group P and the unprotected group U are denoted by $\hat{R}_P^{\ell}$ and $\hat{R}_U^{\ell}$. Formally, the autoencoder consists of two main components: an encoder $g_e:\mathbb{R}^d\to\mathbb{R}^r$ and a decoder $g_d:\mathbb{R}^r\to\mathbb{R}^d$, where r is the dimensionality of the hidden representations. $g_e(x)$ encodes the input x to a hidden representation z that preserves the important information of the input. Then, $g_d(z)$ aims to recover $x'\approx x$, a reconstruction of the input from the hidden representation z.

Table 5: Popular f-divergences and their conjugate functions.

Divergence	$f(x)$	Conjugate $f^*(t)$	$f'(1)$	Activation func.
Kullback-Leibler (KL)	$x \log x$	$\exp(t-1)$	1	x
Reverse KL (KL-rev)	$-\log x$	$-1-\log(-t)$	-1	$-\exp x$
Jensen-Shannon (JS)	$-x+1\log\frac{1+x}{2}+x\log x$	$-\log(2-e^t)$	0	$\log\frac{2}{1+\exp(-x)}$
Pearson χ^2	$(x-1)^2$	$\frac{t^2}{4}+t$	0	x
Total Variation (TV)	$\frac{1}{2}\lvert x-1\rvert$	$1_{-1/2\le t\le 1/2}$	$[-1/2, 1/2]$	$\frac{1}{2}\tanh x$

Overall, the autoencoder can be written as $G = g_d \circ g_e$, i.e. $G(x) = g_d(g_e(x))$. For a given autoencoder-based framework, the anomaly score for x is computed using the reconstruction error as: $s(x) = \|x - G(x)\|^2$.

The Rademacher complexity for a function class is:

Definition C.1. (Rademacher Complexity (Shalev-Shwartz & Ben-David, 2014)) Given a space $\mathcal{X}$, and a set of i.i.d. examples D = $\{x_1, x_2, ..., x_{|D|}\} \subseteq \mathcal{X}$, for a function class $\mathcal{F}$ where each function $r : \mathcal{X} \to \mathbb{R}$, the empirical Rademacher complexity of $\mathcal{F}$ is given by:

$$\mathfrak{R}_D(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{r\in\mathcal{F}} \left(\frac{1}{|D|} \sum_{i=1}^{|D|} \sigma_i r(z_i) \right) \right] \tag{8}$$

Here, $\sigma_1, ..., \sigma_m$ are independent random variables uniformly drawn from $\{-1, 1\}$.

Recall that we previously defined $R^\ell_{\mathcal{D}}(h) \coloneqq R^\ell_{\mathcal{D}}(h, a) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(h(x), a(x))]$. We introduce a commonly used property of Rademacher complexity:

Lemma C.2. *(Property of Rademacher complexity (Mohri et al., 2018)). For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of an i.i.d. samples D of size $|D|$, the following inequality holds for all $h \in \mathcal{H}$:*

$$|R^\ell_D(h) - \hat{R}^\ell_D(h)| \le 2\mathfrak{R}_D(\ell \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2|D|}} \tag{9}$$

where $\mathfrak{R}_D(\ell \circ \mathcal{H})$ is the Rademacher complexity of the function class $\ell \circ \mathcal{H}$ given data D.

Next we introduce the definition of f-divergence, which enables us to derive an upper bound on the group risk difference.

Definition C.3. (f-divergence (Ali & Silvey, 1966)) Let P and Q be two distribution functions with densities p and q, respectively. Let p be absolutely continuous w.r.t q and both be absolutely continuous with respect to a base measure dx. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex, lower semi-continuous function that satisfies $f(1) = 0$. The f-divergence D_f is defined as:

$$D_f(P \parallel Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx. \tag{10}$$

Many commonly used divergences in machine learning, such as KL divergence and total variation distance, are special cases of f-divergence. Table 5 provides examples. Following (Nguyen et al., 2010), f-divergence can be estimated via a variational formulation:

$$D_f(P \parallel Q) \ge \sup_{T\in\mathcal{T}} \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))] \tag{11}$$

where f^* is the (Fenchel) conjugate function of f defined as $f^*(y) \coloneqq \sup_{x\in\mathbb{R}_+}\{xy - f(x)\}$, $T : \mathcal{X} \to \text{dom } f^*$, and $\mathcal{T}$ is the set of all measurable functions.

Next, with the help of Rademacher complexity, we now provide a fairness bound for the performance difference between the protected and unprotected groups $R^\ell_P(h) - R^\ell_U(h)$:

Theorem C.4. *(Fairness with Rademacher Complexity) Suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, f^* is L-Lipschitz continuous, and $[0, 1] \subseteq$ dom f^*. Let U and P be two empirical distributions corresponding to datasets containing m and n data points sampled i.i.d. from P_U and P_P, respectively. Let us denote $\mathfrak{R}$*

as the Rademacher complexity of a given hypothesis class, and define $\ell \circ \mathcal{H} := \{x \mapsto \ell(h(x), h'(x)) : h, h' \in \mathcal{H}\}$. *Let* h^* *be the ideal joint hypothesis, i.e.,* $h^* = \arg\min_{h \in \mathcal{H}} R_U^\ell(h) + R_P^\ell(h)$. *For any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *we have:*

$$R_P^\ell(h) - R_U^\ell(h) \leq D_f(U\|P) + \hat{R}_U^\ell(h^*) + \hat{R}_P^\ell(h^*) + 4\mathfrak{R}_U(\ell \circ \mathcal{H}) + 2(L+1)\mathfrak{R}_P(\ell \circ \mathcal{H}) \quad (12)$$

$$+ 2\sqrt{\frac{\log \frac{1}{\delta}}{2m}} + 2\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \quad (13)$$

Under the assumption of an ideal joint hypothesis h^*, fairness (*i.e.*, the risk difference between the protected and unprotected groups) can be improved by minimizing the discrepancy between the hidden representation of the examples from two groups and regularizing the model to limit the complexity of the hypothesis class. In Appendix C.7.1, we provide a full proof of Theorem C.4 and analyze how minimizing the objective $\mathcal{L}_{\text{FAC}}$ leads to small empirical f-divergence $D_f(U||P)$ for total variation, and thereby promotes fairness.

C.7.1 Proof of Theorem C.4

Given that $D_f(P \parallel Q)$ involves the supremum over all measurable functions and does not account for the hypothesis class, and that it cannot be estimated from finite examples of arbitrary distributions (Kifer et al., 2004), we further consider a discrepancy which helps relieve these issues based on the variational characterization of f-divergence in Equation (11):

Definition C.5. ($D^f_{h,\mathcal{H}}$ discrepancy (Acuna et al., 2021)) Let f^* be the Fenchel conjugate of a convex, lower semi-continuous function f that satisfies $f(1) = 0$, and let $\hat{T}$ be a set of measurable functions such that $\hat{T} = \{\ell(h(x), h'(x)) : h, h' \in \mathcal{H}\}$ where ℓ is a loss function and $\mathcal{H}$ is the hypothesis space. We define the discrepancy between the two distributions P and Q as:

$$D^f_{h,\mathcal{H}}(P \parallel Q) := \sup_{h' \in \mathcal{H}} |\mathbb{E}_{x\sim P}[\ell(h(x), h'(x))] - \mathbb{E}_{x\sim Q}[f^*(\ell(h(x), h'(x)))]|$$

From the definition we can easily get $D^f_{h,\mathcal{H}}(P \parallel Q) \leq D_f(P \parallel Q)$.

With property of the rademacher complexity in Lemma C.2, we now show that $D^f_{h,\mathcal{H}}$ can be estimated from finite examples:

Lemma C.6. *Suppose* $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, f^* *is L-Lipschitz continuous, and* $[0, 1] \subseteq$ *dom* f^*. *Let* U *and* P *be two empirical distributions corresponding to datasets containing* m *and* n *data points sampled i.i.d. from* P_U *and* P_P, *respectively. Let us denote* $\mathfrak{R}$ *as the Rademacher complexity of a given hypothesis class, and define* $\ell \circ \mathcal{H} := \{x \mapsto \ell(h(x), h'(x)) : h, h' \in \mathcal{H}\}$. *For any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *we have:*

$$|D^f_{h,\mathcal{H}}(P_U\|P_P) - D^f_{h,\mathcal{H}}(U\|P)| \leq 2\mathfrak{R}_{P_U}(\ell \circ \mathcal{H}) + 2L\mathfrak{R}_{P_P}(\ell \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \quad (14)$$

Proof.

$$\begin{aligned}
D^f_{h,\mathcal{H}}(P_U\|P_P) - D^f_{h,\mathcal{H}}(U\|P) &= \sup_{h' \in \mathcal{H}} \{|R_U^\ell(h, h') - R_P^{f^* \circ \ell}(h, h')|\} \\
&\quad - \sup_{h' \in \mathcal{H}} \{|\hat{R}_U^\ell(h, h') - \hat{R}_P^{f^* \circ \ell}(h, h')|\} \\
&\leq \sup_{h' \in \mathcal{H}} \|R_U^\ell(h, h') - R_P^{f^* \circ \ell}(h, h')| - |\hat{R}_U^\ell(h, h') - \hat{R}_P^{f^* \circ \ell}(h, h')\| \\
&\leq \sup_{h' \in \mathcal{H}} |R_U^\ell(h, h') - R_P^{f^* \circ \ell}(h, h') - \hat{R}_U^\ell(h, h') + \hat{R}_P^{f^* \circ \ell}(h, h')| \\
&= \sup_{h' \in \mathcal{H}} |R_U^\ell(h, h') - \hat{R}_U^\ell(h, h')| + |R_P^{f^* \circ \ell}(h, h') - \hat{R}_P^{f^* \circ \ell}(h, h')| \\
&\leq 2\mathfrak{R}_{P_U}(\ell \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} + 2\mathfrak{R}_{P_P}(f^* \circ \ell \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}
\end{aligned}$$

where the last inequality comes from the property of Rademacher complexity. Similarly, by Lemma 5.7 and Definition 3.2 of Mohri et al. (2018) we have: $\mathfrak{R}_{P_P}(f^* \circ \ell \circ \mathcal{H}) \leq L\mathfrak{R}_{P_P}(\ell \circ \mathcal{H})$, with $f^* \circ \ell \circ \mathcal{H} := \{x \mapsto \phi(\ell(h(x), h'(x))) : h, h' \in \mathcal{H}\}$. □

Lemma C.6 shows that the empirical discrepancy $D^f_{h,\mathcal{H}}$ converges to the true discrepancy, and the gap is bounded by the complexity of the hypothesis class and the number of examples. We then present another theorem:

Theorem C.7. *Let h^* be the ideal joint hypothesis, i.e., $h^* = \arg\min_{h\in\mathcal{H}} R^\ell_U(h) + R^\ell_P(h)$. The risk difference between the two groups is upper bounded by:*

$$R^\ell_P(h) - R^\ell_U(h) \leq D^f_{h,\mathcal{H}}(P_U \| P_P) + R^\ell_U(h^*) + R^\ell_P(h^*). \tag{15}$$

Proof. First, notice that by definition, $f^*(t) = \sup_{x\in \mathrm{dom} f}(xt - f(x)) \geq t - f(1) = t$. Then we can prove:

$$\begin{aligned}
R^\ell_P(h, a_P) &\leq R^\ell_P(h, h^*) + R^\ell_P(h^*, a_P) && \text{(triangle inequality } \ell\text{)} \\
&= R^\ell_P(h, h^*) + R^\ell_P(h^*, a_P) - R^\ell_U(h, h^*) + R^\ell_U(h, h^*) \\
&\leq R^{f^*\circ\ell}_P(h, h^*) - R^\ell_U(h, h^*) + R^\ell_U(h, h^*) + R^\ell_P(h^*, a_P) \\
&\leq |R^{f^*\circ\ell}_P(h, h^*) - R^\ell_U(h, h^*)| + R^\ell_U(h, h^*) + R^\ell_P(h^*, a_P) \\
&\leq D^f_{h,\mathcal{H}}(P_U \| P_P) + R^\ell_U(h, h^*) + R^\ell_P(h^*, a_P) \\
&\leq D^f_{h,\mathcal{H}}(P_U \| P_P) + R^\ell_U(h, a_U) + R^\ell_U(h^*, a_U) + R^\ell_P(h^*, a_P) \\
&= D^f_{h,\mathcal{H}}(P_U \| P_P) + R^\ell_U(h) + R^\ell_U(h^*) + R^\ell_P(h^*)
\end{aligned}$$

□

Combining Theorem C.7, Lemma C.6 and the property of Rademacher Complexity, we can easily get:

$$\begin{aligned}
R^l_P(h) - R^l_U(h) \leq\ & D^f_{h,\mathcal{H}}(U \| P) \\
&+ \hat{R}^l_U(h^*) + 4\mathfrak{R}_U(\ell \circ \mathcal{H}) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \\
&+ \hat{R}^l_P(h^*) + 2(L+1)\mathfrak{R}_P(\ell \circ \mathcal{H}) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2n}}
\end{aligned}$$

Since by definition we have $D^f_{h,\mathcal{H}}(U\|P) \leq D_f(U\|P)$, and for $D_f(U\|P) = \mathrm{TV}(U\|P)$, we have:

$$\begin{aligned}
R^l_P(h) - R^l_U(h) \leq\ & \mathrm{TV}(U \| P) \\
&+ \hat{R}^l_U(h^*) + 4\mathfrak{R}_U(\ell \circ \mathcal{H}) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \\
&+ \hat{R}^l_P(h^*) + 2(L+1)\mathfrak{R}_P(\ell \circ \mathcal{H}) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.
\end{aligned} \tag{16}$$

Now we motivate why minimizing the objective $\mathcal{L}_{\mathrm{FAC}}$ leads to small $\mathrm{TV}(U\|P)$. Let U, P be the empirical distributions over the common measurable space $\mathcal{X} := \{z^U_j\}_{j=1}^n \cup \{z^P_k\}_{k=1}^m$ with densities $\hat{p}_U, \hat{p}_P$ that are c_U, c_P-Lipschitz with respect to ℓ_2-norm, respectively. Let $x^* := \arg\min_{x\in\mathcal{X}} |\hat{p}_U(x) - \hat{p}_P(x)|$, $\delta := |\hat{p}_U(x^*) - \hat{p}_P(x^*)|$, and

$$\sigma := \sum_{x\in\mathcal{X}} \|x - x^*\| = \sum_{x\in\mathcal{X}} \sqrt{2 - 2\log \mathrm{sim}(x, x^*)},$$

where the equality is due to law of cosine (and that sim normalizes z_j). We first show how $\mathrm{TV}(U\|P)$ is related to δ and σ.

Lemma C.8.

$$\mathrm{TV}\left(U\|P\right) \leq \frac{1}{2}\left(|\mathcal{X}|\delta + (c_U + c_P)\sigma\right).$$

Proof.

$$\begin{aligned}
\mathrm{TV}\left(U\|P\right) &\coloneqq \frac{1}{2}\sum_{x\in\mathcal{X}} |\hat{p}_U(x) - \hat{p}_P(x)| \\
&\leq \frac{1}{2}\sum_{x\in\mathcal{X}} |\hat{p}_U(x) - \hat{p}_U(x^*)| + |\hat{p}_U(x^*) - \hat{p}_P(x^*)| + |\hat{p}_P(x^*) - \hat{p}_P(x)| \quad \text{(triangle inequality)} \\
&= \frac{1}{2}\left(|\mathcal{X}|\delta + \sum_{x\in\mathcal{X}} |\hat{p}_U(x) - \hat{p}_U(x^*)| + |\hat{p}_P(x) - \hat{p}_P(x^*)|\right) \\
&\leq \frac{1}{2}\left(|\mathcal{X}|\delta + (c_U + c_P)\sum_{x\in\mathcal{X}} \|x - x^*\|\right) \quad \text{(Lipschitz conditions)} \\
&= \frac{1}{2}\left(|\mathcal{X}|\delta + (c_U + c_P)\sigma\right).
\end{aligned}$$

□

Next we motivate why minimizing our objective $\mathcal{L}_{\mathrm{FAC}}$ leads to small δ and σ simultaneously, hence small $\mathrm{TV}(U\|P)$. Recall that our fairness-aware contrastive loss is

$$\mathcal{L}_{\mathrm{FAC}} \coloneqq \mathcal{L}_{\mathrm{fair}} + \mathcal{L}_{\mathrm{unif}},$$

where

$$\begin{aligned}
\mathcal{L}_{\mathrm{fair}} &\coloneqq -\log\left(\sum_{j\in[n]}\sum_{k\in[m]} \mathrm{sim}\left(z_j^U, z_k^P\right)\right), \\
\mathcal{L}_{\mathrm{unif}} &\coloneqq \log\left(\sum_{j\neq k} \mathrm{sim}\left(z_j^U, z_k^U\right) + \sum_{j\neq k} \mathrm{sim}\left(z_j^P, z_k^P\right)\right).
\end{aligned}$$

Intuitively, minimizing $\mathcal{L}_{\mathrm{FAC}}$ leads to small $\mathcal{L}_{\mathrm{fair}}$ and $\mathcal{L}_{\mathrm{unif}}$ simultaneously, which correspond to large $\mathrm{sim}(z_j^U, z_k^P)$ and small $\mathrm{sim}(z_j^U, z_k^U), \mathrm{sim}(z_j^P, z_k^P)$, which in turn correspond to small $\|z_j^U - z_k^P\|$ and large $\|z_j^U - z_k^U\|, \|z_j^P - z_k^P\|$. Hence it is natural to consider the following surrogate losses

$$\begin{aligned}
\mathcal{L}'_{\mathrm{fair}} &\coloneqq \sum_{j,k\in[n]} \|z_j^U - z_k^P\|, \\
\mathcal{L}'_{\mathrm{unif}} &\coloneqq -(\sum_{j\neq k} \|z_j^U - z_k^U\| + \|z_j^P - z_k^P\|).
\end{aligned}$$

Then it follows immediately that $\sigma \leq \mathcal{L}'_{\mathrm{fair}}$, explaining why minimizing our objective $\mathcal{L}_{\mathrm{FAC}}$ (hence $\mathcal{L}'_{\mathrm{fair}}$) leads to small σ.

To see that $\delta \coloneqq |\hat{p}_U(x^*) - \hat{p}_P(x^*)|$ cannot be too large, first consider the extreme case where $\{z_j^U\}_{j=1}^n \cap \{z_k^P\}_{k=1}^n = \emptyset$. Without loss of generality let $\|z_1^U - z_1^P\| = \max_{j,k\in[n]} \|z_j^U - z_k^P\|$. Then adjusting z_1^U, z_1^P to be the unit vector on their angle bisector clearly decreases $\mathcal{L}'_{\mathrm{fair}}$ without affecting $\mathcal{L}'_{\mathrm{unif}}$ by much due to high uniformity within $\{z_j^U\}_{j=1}^n$ and $\{z_k^P\}_{k=1}^n$ respectively. Hence we may assume without loss of generality that $z_1^U = z_1^P = x^*$. Next consider the extreme case where $\hat{p}_U(x^*) = \frac{1}{n}$ and $\hat{p}_P(x^*) = 1$. Then adjusting $z_2^P = \arg\max_{x\neq x^*} \sum_{j\in[n]} \|x - z_j^U\|$ clearly decreases $\mathcal{L}'_{\mathrm{unif}}$ without affecting $\mathcal{L}'_{\mathrm{fair}}$ by much due to high uniformity within $\{z_j^U\}_{j=1}^n$. Hence minimizing our objecive $\mathcal{L}_{\mathrm{FAC}}$ leads to small $\delta \coloneqq |\hat{p}_U(x^*) - \hat{p}_P(x^*)|$.

Table 6: Characteristics of datasets.

Datasets	Unprotected Group		Protected Group		#Features	Sensitive Attribute	Anomaly Definition
	#Instances	#Anomaly	#Instances	#Anomaly			
MNIST-USPS	7,785	882	1,876	323	1,024	Source of the digits	Digit 0 or not
MNIST-Invert	7,344	441	408	38	1,024	Color of the digits	Digit 0 or not
COMPAS	1,839	325	299	39	8	Race	Reoffending or not
CelebA	41,919	4,008	7,300	1,142	39	Gender	Attractive or not

C.8 Extension to multi-value multi-group cases

We can easily extend our method to the multi-value multi-group case by modifying our designed $\mathcal{L}_{\mathrm{FAC}}$ and $\mathcal{L}_{\mathrm{REC}}$. In our loss design, $\mathcal{L}_{\mathrm{REC}}$ is an auto-reweighted reconstruction error, and the weight is in proportion with $\mathcal{L}_0^G - \mathcal{L}_G$, where G denotes a certain group. In the multi-valued, multiple protected attributes case, we can extend the design of $\mathcal{L}_{\mathrm{REC}}$ by weighting each reconstruction error of group G with $\frac{\mathcal{L}_0^G - \mathcal{L}_G}{\sum_g \mathcal{L}_0^g - \mathcal{L}_g}$. Then, in $\mathcal{L}_{\mathrm{FAC}}$, $\mathcal{L}_{\mathrm{fair}}$ encourages the representation similarity between different groups, and we can extend it to the similarity between different combinations of multi-valued, multiple protected attributes. $\mathcal{L}_{\mathrm{unif}}$ encourages the uniformity within each group, and we can add the uniformity term for each group in the multi-valued, multiple group case.

D Additional Experiments

D.1 Dataset Statistics

In Table 6, we provide the statistics of the datasets we used in the experiments.

D.2 Training details and experimental setup

For the COMPAS dataset, we use a two-layer MLP with hidden units of [32, 32]. For celebatab dataset, we use three-layer MLP with hidden units of [10, 10, 10]. For all the other datasets, we use MLP with one hidden layer of dimension 128. For our method, we use the Adam optimizer with learning rate $1e^{-3}$ and we set the hyperparameter $\alpha = 4$ across all the data sets. For the four independent runs, we choose random seeds in [40, 41, 42, 43]. We set ϵ to the group-ratio weight during a short warm-up (the first 100 iterations) until the group-wise losses stabilize, and fall back to this weight whenever finite-sample fluctuations produce $\epsilon \notin (0, 1)$. All the experiments were executed using one Tesla V100 SXM2 GPUs, supported by a 12-core CPU operating at 2.2GHz.

For the baselines, we use the suggested hyperparameter settings in their original papers. For FairOD, we follow the instructions for the configurations as proposed by their authors using grid search. For DCFOD, the author studied the hyperparameters α and β, and we set them as their suggested values. FairSVDD suggested setting $\lambda = 1$ for all datasets. ReContrast suggested setting $\alpha = 1$. Other baselines are tested insensitive to hyperparameters and we set them as the default values in their implementations. For a more comprehensive comparison, we test different hyperparameter settings of the two baseline methods (i.e., ReContrast, FairSVDD) on MNIST-USPS with K=1200. The results are shown in Tables 7 and 8. The experimental results show that our method still outperforms the baselines in terms of both anomaly detection performance and fairness, across all the hyperparameter settings. Notice that determining the hyperparameter for these baseline methods solely based on the experimental results involves data leakage. That is why we follow the default value for these hyperparameters mentioned in these papers to avoid data leakage.

For evaluation, since the task is unsupervised, the train and test sets are the same. Following Shekhar et al. (2021); Zhang & Davidson (2021); Ahmed et al. (2021), to evaluate task performance, we use Recall@K, AUCROC and AUPRC. For fairness evaluation, considering the imbalance between normal examples and anomalies, we focus on Recall Parity in anomaly detection. Given our score-based anomaly detection framework, we would like to state the mathematical formulation of Recall Parity fairness in anomaly detection as: Let anomaly score for example x be $s(x)$ and let t_K be the anomaly score threshold for top-K selection. Then, the predicted normal examples are the ones with $s(x) < t_K$, and the predicted anomalies are those with $s(x) \geq t_K$. The recall parity in anomaly

Table 7: FairSVDD results with different hyperparameter settings.

Hyperparameter	Recall@K	AUCROC	Recall Diff
$\lambda = 1$	15.62±1.52	58.33±1.18	13.75±2.56
$\lambda = 2$	15.77±0.36	58.62±1.32	23.78±31.87
$\lambda = 3$	16.86±0.88	59.24±1.61	34.33±17.76
$\lambda = 4$	15.36±4.20	59.82±2.67	33.33±0.87
$\lambda = 5$	15.05±0.01	55.52±2.03	2.26±1.50

Table 8: ReContrast results with different hyperparameter settings.

Hyperparameter	Recall@K	AUCROC	Recall Diff
$\alpha = -2$	63.29±2.33	82.80±2.40	9.55±2.69
$\alpha = -1$	64.18±2.80	84.91±3.04	23.81±2.79
$\alpha = 0$	64.25±4.89	84.37±5.86	19.38±2.62
$\alpha = 1$	64.29±3.18	83.46±3.77	41.16±5.63
$\alpha = 2$	63.04±5.98	70.56±7.69	18.31±7.65

detection requires that $P(s(x) \geq t_K | x \in U, y = 1) = P(s(x) \geq t_K | x \in P, y = 1)$. We use the absolute value of their difference, *i.e.*, Recall Diff, to evaluate the fairness level. We also use the AUPRC Diff as another fairness evaluation, which is not related to K selection.

Our code is at: `https://anonymous.4open.science/r/FADIG_ICLR35426/`

D.3 Group Imbalance Study

To further study the performance of FADIG under group imbalance, we vary the ratio between unprotected and protected groups on the MNIST-USPS dataset. We report the average results over four independent runs in Table 9. As the imbalance grows from $1:1$ to $4:1$, the fair AD baselines degrade: FairOD's Recall drops from 20.47% to 11.90% and FairSVDD's from 40.04% to 10.98%. This is consistent with Lemma 3.3, under which a sample-additive regularizer increasingly dilutes the protected group as $|U|/|P|$ grows. DCFOD keeps small fairness gaps, but its AUCROC stays at chance level (49.66–52.77%). In contrast, FADIG achieves the best AUCROC, Recall, and AUPRC at every ratio, with Recall of 61.15%, 60.94%, and 63.65% and AUCROC of 87.06%, 87.85%, and 93.98%, and the lowest Recall Diff at $1:1$ and $4:1$ (0.93% and 2.60%). Two baselines attain lower disparity in some settings: DCFOD has the lowest AUPRC Diff at every ratio and the lowest Recall Diff at $2:1$, but at chance-level detection, and MCM has a lower AUPRC Diff than FADIG at $2:1$ and $4:1$, but a 2–10$\times$ larger Recall Diff. Notably, FADIG's detection performance does not degrade as imbalance increases, and its AUCROC is highest at $4:1$. We attribute this to the adaptive weight ϵ, under which the group-size factor cancels in the reconstruction gradients, so the protected group's influence does not shrink as $|U|/|P|$ grows. We also do the data imbalance study on the COMPAS dataset. Results in Table 10 also show that most fair AD methods (e.g., FairOD, DCFOD and FairSVDD) tend to suffer from the imbalanced groups scenario when the unprotected group gradually becomes dominant.

D.4 Ablation Study (Q3)

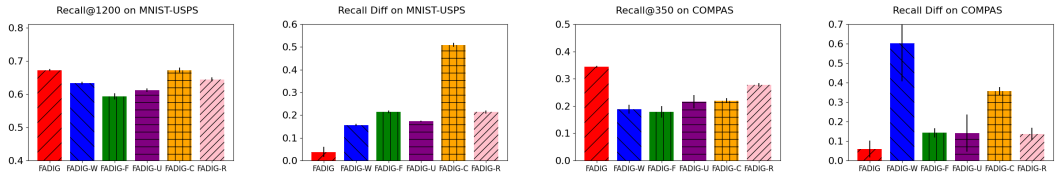


Figure 5: Ablation Study of variants of FADIG on MNIST-USPS and COMPAS datasets.

To assess the contribution of each component in FADIG, we conduct an ablation study to demonstrate the necessity of each component of FADIG on the MNIST-USPS and COMPAS datasets. The

Table 9: Performance on MNIST-USPS with different ratios. The best score is marked in bold.

Ratio	Metrics	FairOD	DCFOD	FairSVDD	FairAD	MCM	NSNMF	ReContrast	FADIG
$\|U\| : \|P\| = 1 : 1$	AUCROC	60.69±1.11	52.77±2.61	61.40±11.45	60.57±0.33	75.38±0.29	79.00±0.29	77.03±0.41	**87.06±0.89**
	Recall@650	20.47±0.63	32.97±0.87	40.04±15.55	29.62±1.82	57.12±0.87	48.38±0.74	45.51±0.67	**61.15±0.29**
	Recall Diff	11.67±1.07	3.76±0.59	47.26±9.44	39.84±9.40	8.51±1.95	40.40±1.80	54.90±3.13	**0.93±0.87**
	AUPRC	15.43±0.38	17.50±0.08	28.40±18.09	27.03±2.15	56.94±0.91	48.27±1.44	29.97±0.49	**60.69±0.18**
	AUPRC Diff	2.34±1.33	**0.75±0.94**	17.02±3.07	4.89±3.33	2.25±0.85	2.51±2.41	11.13±2.40	1.26±0.48
$\|U\| : \|P\| = 2 : 1$	AUCROC	56.82±0.81	49.66±0.39	52.99±4.38	54.42±4.42	76.17±0.59	79.02±2.59	77.46±2.73	**87.85±0.43**
	Recall@1000	15.08±1.24	22.05±1.20	22.60±3.47	23.91±1.10	56.37±1.07	41.90±2.80	46.54±4.33	**60.94±0.17**
	Recall Diff	9.29±1.61	**1.55±1.17**	15.02±21.45	27.81±6.45	10.33±1.14	53.83±2.63	41.80±2.95	5.32±1.71
	AUPRC	14.20±0.45	17.16±0.07	18.33±1.14	19.55±3.90	56.54±0.81	43.00±1.83	28.32±1.83	**57.73±1.16**
	AUPRC Diff	9.60±0.33	**0.85±0.32**	19.22±30.59	3.87±2.19	2.70±0.47	3.96±3.39	10.55±2.53	3.51±1.61
$\|U\| : \|P\| = 4 : 1$	AUCROC	50.67±0.57	50.62±0.69	51.20±8.28	48.60±1.41	79.52±1.19	84.39±0.50	80.99±0.92	**93.98±0.11**
	Recall@1200	11.90±3.71	13.58±1.03	10.98±6.66	16.16±1.86	53.92±1.38	38.87±0.89	41.05±1.60	**63.65±0.57**
	Recall Diff	11.78±5.30	3.03±1.87	21.06±8.04	7.75±2.38	25.45±1.91	62.38±1.68	42.97±7.86	**2.60±0.43**
	AUPRC	10.23±2.95	13.67±0.19	14.60±4.15	15.47±0.98	54.25±1.66	41.70±0.84	24.86±1.10	**66.43±0.49**
	AUPRC Diff	7.85±0.41	**0.67±0.42**	9.29±15.88	4.58±2.03	2.46±0.32	3.77±0.57	9.73±3.06	3.76±0.64

Table 10: Performance on COMPAS dataset with different ratios. The best score is marked in bold.

Methods	$\|U\| : \|P\| = 1 : 1$ (K=80)			$\|U\| : \|P\| = 2 : 1$ (K=120)			$\|U\| : \|P\| = 5 : 1$ (K=240)		
	Recall@K	AUCROC	Rec Diff	Recall@K	AUCROC	Rec Diff	Recall@K	AUCROC	Rec Diff
FairOD	13.68±2.67	50.10±0.85	11.97±1.48	13.11±0.50	50.11±0.74	6.60±0.97	12.54±1.37	49.58±0.87	7.68±0.72
DCFOD	11.54±4.62	48.50±2.69	7.69±4.445	15.95±3.00	53.28±0.75	10.68±2.67	12.96±2.02	49.76±1.16	6.36±0.70
FairSVDD	16.24±2.18	52.34±1.38	6.84±3.20	14.53±1.84	51.69±2.15	7.69±3.77	14.10±4.53	50.04±4.98	14.87±7.54
MCM	18.38±0.60	40.77±0.25	7.69±3.63	16.24±0.01	40.42±0.12	10.26±4.80	18.81±0.60	44.04±0.15	5.76±2.31
NSNMF	20.08±0.74	53.86±0.42	14.53±10.36	19.09±1.31	53.28±0.75	10.68±2.67	20.09±2.22	53.86±1.28	10.77±5.40
FADIG	**29.91±0.74**	**61.87±1.89**	**3.42±1.48**	**28.42±0.43**	**57.39±2.84**	**1.92±1.72**	**29.77±1.31**	**58.05±1.34**	**4.83±0.78**

experimental results are presented in Figure 5. Specifically, FADIG-W refers to a variant of our method replacing the re-balancing autoencoder with the unweighted reconstruction loss in Equation (3); FADIG-F and FADIG-U remove $\mathcal{L}_{\text{fair}}$ and $\mathcal{L}_{\text{unif}}$ in Equation (2), respectively; FADIG-C substitutes the proposed fair contrastive loss $\mathcal{L}_{\text{FAC}}$ with the traditional contrastive loss (Chen et al., 2020); FADIG-R removes the entire $\mathcal{L}_{\text{FAC}}$. We have the following observations: (1) FADIG greatly outperforms FADIG-F and FADIG-U, highlighting the importance of both $\mathcal{L}_{\text{fair}}$ and $\mathcal{L}_{\text{unif}}$ in achieving accurate and fair anomaly detection. (2) While FADIG-C can sometimes achieve comparable recall to FADIG, it consistently yields a higher recall difference. This suggests that simply using a standard contrastive loss is insufficient for group fairness, validating the need for our fairness-aware regularization. (3) Compared with FADIG, FADIG-W has a lower recall rate and a higher recall difference. It suggests that classical unweighted reconstruction loss tends to mainly focus on learning the frequent patterns of the unprotected group while ignoring the protected group, consistent with the n/m dilution of the unweighted loss derived in Section 4, and validates the effectiveness of our re-balanced autoencoder. (4) While FADIG-R has relatively high recall rates and low recall differences on the datasets, FADIG still surpasses it in all metrics, showing the necessity of $\mathcal{L}_{\text{FAC}}$.

D.5 Effectiveness validation of FADIG under different choices of K

For the datasets we use, we approximate K as a number that is close to the exact number of the anomalies in hundreds or fifties. We also conduct experiments on the four datasets with different choices of K, and the results are in Table 11 and Table 12. The AUCROC scores are the same as in the main paper. We can also tell from the tables that accuracy difference is inadequate for measuring group fairness in the imbalanced setting.

For potential users, with knowledge of the dataset (i.e., exact numbers of the anomalies), you can choose the number in hundreds closest to the the exact number of anomalies. If you do not know the exact number, with some domain knowledge, you can set K according to the approximated proportion of anomalies in the dataset. E.g., you can set K as 10% of the total sample number if you expect such an anomaly rate in this dataset.

Table 11: Performance on image datasets with different Ks.

Methods	MNIST-USPS (K=1000)			MNIST-Invert (K=400)		
	Recall@K	Acc Diff	Rec Diff	Recall@K	Acc Diff	Rec Diff
FairOD	10.46±1.16	4.35±0.33	13.21±1.43	6.05±0.21	2.70±0.15	9.99±1.18
DCFOD	10.24±0.82	4.79±1.12	8.40±1.83	5.57±1.70	2.69±0.37	8.78±2.31
FairSVDD	13.75±1.83	5.73±5.64	13.49±2.55	10.57±0.92	5.38±3.12	14.25±2.96
MCM	34.38±0.32	29.81±0.84	52.46±0.94	22.48±0.54	8.32±1.10	64.37±1.66
NSNMF	33.56±0.70	22.26±0.40	65.12±2.36	43.91±0.84	4.54±0.20	55.20±0.92
Recontrast	45.73±2.74	10.59±2.62	29.62±2.40	52.00±4.86	13.81±4.30	54.96±13.77
FADIG	61.60±2.50	6.50±0.89	7.95±5.94	62.28±3.24	1.62±1.32	7.02±4.48

Table 12: Performance on tabular datasets with different Ks.

Methods	COMPAS (K=300)			CelebA (K=4500)		
	Recall@K	Acc Diff	Rec Diff	Recall@K	Acc Diff	Rec Diff
FairOD	14.20±1.83	3.92±1.63	10.75±0.90	7.95±0.21	4.94±0.25	2.26±1.06
DCFOD	13.10±1.35	3.57±2.29	7.23±2.82	8.64±0.79	4.98±0.40	9.24±1.12
FairSVDD	13.02±1.66	3.90±2.43	9.45±3.80	8.82±0.61	2.21±0.40	10.22±2.33
MCM	16.87±1.14	4.10±1.98	10.17±1.64	9.26±0.48	7.21±5.98	28.69±12.14
NSNMF	17.29±1.42	3.60±1.93	33.57±1.22	8.90±1.09	5.66±0.54	40.51±1.54
FADIG	19.14±2.29	9.35±3.00	4.75±3.69	10.56±1.11	13.04±0.30	5.10±1.52

Table 13: Results of additional metrics on the image datasets.

Methods	MNIST-USPS (K=1200)			MNIST-Invert (K=500)		
	Precision@K	Precision Diff	AUCROC Diff	Precision@K	Precision Diff	AUCROC Diff
FairOD	12.25±1.10	6.35±1.79	9.10±4.25	7.33±0.66	10.89±2.07	2.78±1.74
DCFOD	12.39±0.65	8.25±0.62	10.17±1.49	6.60±0.86	2.75±0.90	2.69±0.41
FairSVDD	12.39±1.55	26.07±24.50	22.34±12.91	10.07±1.47	10.24±1.25	23.33±22.45
MCM	38.92±0.34	34.67±1.98	19.68±2.36	24.52±1.78	19.60±5.13	16.95±7.40
NSNMF	39.19±0.81	37.54±2.09	19.65±0.93	49.60±5.89	9.24±0.47	20.14±0.74
ReContrast	65.11±3.03	9.91±2.67	18.77±13.31	68.13±5.47	26.25±11.11	29.29±1.80
FADIG	67.00±4.33	7.54±1.63	9.07±1.89	69.27±0.76	9.97±3.64	0.40±0.14

Table 14: Results of additional metrics on the tabular datasets.

Methods	COMPAS (K=350)			CelebA (K=5000)		
	Precision@K	Precision Diff	AUCROC Diff	Precision@K	Precision Diff	AUCROC Diff
FairOD	17.23±2.20	6.76±4.76	5.59±2.66	10.30±0.22	6.01±1.44	0.66±0.47
DCFOD	16.57±1.99	5.07±6.09	6.14±3.01	10.07±0.65	6.06±2.62	2.14±0.96
FairSVDD	15.81±3.03	9.35±3.97	8.84±5.36	11.60±1.65	7.40±2.59	5.73±3.92
MCM	28.68±1.71	7.66±2.04	8.85±4.32	12.17±0.81	6.20±2.67	8.25±3.16
NSNMF	31.43±0.23	6.83±3.77	10.55±0.98	11.31±0.64	4.09±0.63	1.99±0.13
FADIG	35.14±0.49	3.82±0.96	4.49±2.77	12.96±0.22	6.09±0.37	1.43±0.04

D.6 Results of additional Metrics

To provide a more comprehensive assessment of fairness, we now include additional metrics: precision, precision difference@K, and AUCROC diff which is not related with K selection. The results on the image and tabular datasets are presented in Tables 13 and 14.

We can observe that FADIG still achieves the highest precision across all the datasets while maintaining the low precision difference and AUCROC difference.

Table 15: Comparison of our designed rebalancing strategy with group ratio weighting.

Datasets	FADIG				Group Ratio Reweighting			
	Recall@K	AUCROC	Rec Diff	Time(s)	Recall@K	AUCROC	Rec Diff	Time(s)
MNIST-USPS	67.19±0.33	91.28±0.46	3.77±2.18	122.84	62.35±0.10	87.61±0.47	11.87±3.90	75.28
MNIST-Invert	71.82±0.63	97.99±0.07	9.78±3.10	52.28	68.33±0.24	89.91±0.02	74.22±0.26	146.45
COMPAS	34.38±0.36	61.45±0.47	5.97±4.34	17.94	33.42±1.61	60.49±4.20	5.85±5.75	15.95
CelebA	11.96±0.49	59.43±0.42	4.72±1.26	52.81	12.75±0.62	57.23±0.25	13.54±0.89	48.12

Table 16: Additional results on image datasets. The best score is marked in bold.

Methods	MNIST-USPS (K=1200)			MNIST-Invert (K=500)		
	Recall@K	AUCROC	Rec Diff	Recall@K	AUCROC	Rec Diff
FairOD	12.42±0.30	50.00±0.28	11.56±0.64	12.21±3.14	50.40±0.20	8.26±1.27
DCFOD	12.63±0.33	50.09±0.27	8.99±0.83	6.95±0.91	50.54±0.54	**7.23±2.02**
FairSVDD	15.62±1.52	58.33±1.18	13.75±2.56	12.41±0.76	49.67±3.98	12.46±2.12
MCM	54.99±0.68	80.69±0.74	18.61±1.51	60.61±1.11	92.21±0.78	27.54±7.04
NSNMF	39.16±0.84	65.38±0.58	62.90±3.84	51.79±0.61	74.21±0.34	51.07±1.79
Recontrast	64.29±3.18	83.46±3.77	41.16±5.63	64.22±1.60	85.13±5.19	56.50±11.23
Normalized	50.38±1.21	89.54±0.20	45.81±1.65	12.16±3.16	54.66±33.11	80.74±26.14
MemAE	31.12±6.69	66.07±6.69	31.19±14.81	27.97±7.73	77.27±5.08	42.06±41.58
VFAE	17.43±0.25	50.49±0.38	23.80±0.34	10.99±0.67	56.15±1.61	10.54±1.74
FADIG	**67.19±0.33**	**91.28±0.46**	**3.77±2.18**	**71.82±0.63**	**97.99±0.07**	9.78±3.10

D.7 Comparison with reweighting heuristic

We compare our designed hyperparameter free and parameter free re-balanced autoencoder with a simple heuristic of setting ϵ as the scaled majority group size, and the results on the four datasets are shown in Table 15. We can observe that compared with the heuristic weight, our designed learnable ϵ can effectively better enhance task performance and meanwhile promote fairness.

D.8 Comparison with additional baselines on additional datasets

We compare our method with three variants of the vanilla auto-encoders, (1) Normalized autoencoder (Yoon et al., 2021), (2) MemAE (Gong et al., 2019) and (3) VFAE (Louizos et al., 2016). Furthermore, we also compare our method with two classical baselines dealing with data imbalance. (4) Separate: An intuitive baseline which trains two separate anomaly detectors for two groups. Since we do not know the ratio of anomalies in protected and unprotected group, during the test time, we set the K for top-K selection in proportion to the group ratio (i.e., set top-K as $K\times$ protected group ratio for protected group model, and set top-K for unprotected group model as $K\times$ unprotected group ratio). (5) SMOTE (Chawla et al., 2002), a classical method handling data imbalance for tabular data.

We also add a commonly-used large-scale tabular dataset ACSIncome (Ding et al., 2021). For imbalanced anomaly detection setting, we do random sampling in ACSIncome. The resulting dataset has 64794 unprotected samples where 9580 in them are anomalies, and 13778 protected samples where 1222 in them are anomalies. The sensitive attribute is sex and the label is income. The results are shown in Tables 16 and 17. We can see that FADIG still outperforms almost all the baselines in both task performance and fairness level.

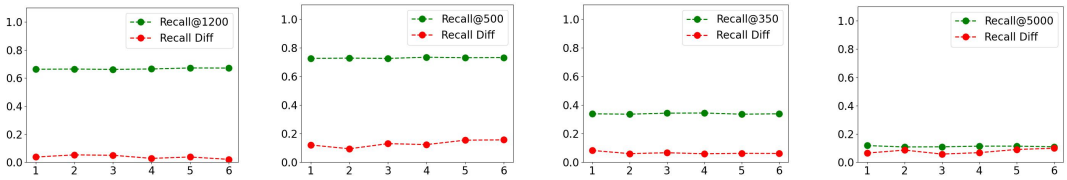
We further add the comparison on a comprehensive image benchmark VisA (Zou et al., 2022) with an advanced backbone EfficientNet-B0 (Tan & Le, 2019). To fit our group-imbalance anomaly detection setting, we select all of Macaronis 1 examples as the unprotected group which contains 1000 normal examples and 100 anomalies. We randomly sample 20% of Macaronis 2 examples as the protected group, which then contains 200 normal examples and 20 anomalies. We show the results in Table 18. We can observe that FADIG always outperforms all the baselines in both utility and fairness, which further shows the effectiveness of our method.

Table 17: Additional results on tabular datasets. The best score is marked in bold.

Methods	COMPAS (K=350)			CelebA (K=5000)			ACSIncome(K=12000)		
	Recall@K	AUCROC	Rec Diff	Recall@K	AUCROC	Rec Diff	Recall@K	AUCROC	Rec Diff
FairOD	16.56±2.12	50.09±1.28	7.97±1.23	8.93±0.14	49.94±0.12	**0.68±0.56**	8.43±0.21	46.92±0.21	6.52±1.02
DCFOD	16.08±1.94	49.55±1.21	9.81±1.76	9.66±0.69	49.92±0.14	7.83±1.26	8.68±0.72	47.04±0.32	9.16±0.84
FairSVDD	15.33±2.10	52.68±5.29	11.57±4.06	10.19±0.50	58.40±1.02	10.95±1.93	10.37±1.92	57.04±4.58	8.61±1.33
MCM	19.05±0.92	50.97±0.43	6.43±3.13	11.00±0.97	50.54±2.32	10.79±4.74	10.97±0.20	58.05±0.81	13.78±2.21
NSNMF	30.15±0.15	60.76±0.31	36.04±3.33	10.91±0.54	50.45±0.30	8.04±1.33	11.53±0.46	58.91±0.29	19.08±5.46
Normalized	26.37±0.55	58.73±0.89	22.15±6.36	6.39±3.40	34.92±23.01	9.79±3.46	10.12±0.78	57.63±1.17	20.12±4.38
MemAE	20.15±2.13	55.59±2.89	20.10±5.88	8.93±1.29	52.85±0.41	14.25±4.61	11.08±0.54	58.77±0.32	22.16±4.89
VFAE	17.21±2.42	53.71±2.32	13.43±3.18	8.62±0.07	48.11±0.49	10.00±0.09	9.54±0.82	54.41±0.96	12.17±1.48
Separate	15.02±2.03	44.48±3.09	18.60±6.97	8.22±0.50	44.27±0.91	0.91±0.54	8.16±0.34	43.21±0.58	9.54±1.10
SMOTE	29.92±2.62	60.45±4.18	6.96±5.75	8.14±0.40	45.39±0.35	5.07±1.43	10.81±0.38	56.79±1.66	6.72±2.91
FADIG	**34.38±0.36**	**61.45±0.47**	**5.97±4.34**	**11.96±0.49**	**59.43±0.42**	4.72±1.26	**12.47±0.41**	**60.52±0.59**	**5.47±3.12**

Table 18: Performance on the VisA data set with EfficientNet.

Methods	Recall@120	AUCROC	Rec Diff
FairOD	9.17±1.85	55.60±9.39	10.83±3.44
FairSVDD	10.81±1.97	58.16±3.66	11.21±2.90
MCM	14.78±0.56	63.21±0.68	28.72±1.96
NSNMF	16.11±0.81	59.23±0.42	19.75±2.29
ReContrast	16.34±3.58	67.04±4.12	24.31±8.77
FADIG	17.50±0.23	71.71±0.49	6.28±0.52



(a) MNIST-USPS (b) MNIST-Invert (c) COMPAS (d) CelebA

Figure 6: Parameter analysis of α on four datasets. The x-axis is α and the y-axis is for the values of recall and recall difference.

D.9 Different anomaly types

We extend our experimental setup to analyze how our method performs on different types of anomalies. In MNIST-USPS and MNIST-Invert, the normal samples are digit 0 and the anomalies are the digits 1-9. In COMPAS and CelebA, we use whether the sample is reoffending / attractive or not to define normal and abnormal samples. Compared with the two image datasets, the anomalies in the tabular ones are more clustered. Thus, we sample more clustered anomalies on the image data set MNIST-USPS by selecting only digit 1 as the anomalies with the same anomaly amount. The results are shown in Table 19. We can observe that our FADIG achieves the best recall rate and the second-best AUCROC score, with a relatively low recall difference. Notably, the baselines with extremely low recall differences are showing "fake" fairness since their task performances are very poor.

D.10 Graph Tasks

We also compare FADIG on graph tasks with two graph anomaly detection baselines, DOMINANT (Ding et al., 2019) and GRADATE (Duan et al., 2023). We adapt our method on the graph dataset Flickr (Li et al., 2015), replacing the backbone with GCN. The results are shown in Table 20. We can observe that FADIG outperforms DOMINANT in both task performance and fairness. While GRADATE has better task performance compared with our method, it may be because we have not

Table 19: Performance on more clustered anomaly detection.

Methods	Recall@K	AUCROC	Rec Diff
FairOD	12.03±0.42	50.04±0.33	3.99±3.89
DCFOD	12.40±1.23	49.94±0.72	4.08±6.02
FairSVDD	18.67±1.73	54.05±9.09	29.72±20.39
MCM	4.01±0.46	14.28±0.60	3.81±1.57
NSNMF	4.64±0.11	45.92±0.78	12.46±0.62
ReContrast	16.34±1.78	51.82±2.26	40.81±18.42
FADIG	21.04±1.27	53.91±1.22	14.39±0.43

Table 20: Performance on the Flickr dataset.

Methods	Recall@K	AUCROC	Rec Diff
DOMINANT	21.34±0.48	61.72±0.59	20.56±3.32
GRADATE	24.96±0.62	66.54±1.12	35.63±5.34
FADIG	23.10±0.61	63.89±1.12	5.33±1.52

optimized our framework specifically for graph data. In addition, our method achieves a much lower recall difference than both baselines.

D.11 Parameter Analysis

In this section, we conduct the parameter analysis on the four datasets. The experiments are repeated four times and the mean of the recall rate and recall difference are reported. Figure 6 shows the parameter analysis for the parameter α on the four datasets, respectively. The parameter α is used to balance the importance between the reconstruction error and the fair contrastive loss. We can observe that FADIG is robust to the choice of α, which may be a benefit from our re-balanced autoencoder.

D.12 Sensitivity of the Adaptive Weight ϵ

Table 21: Sensitivity of FADIG to perturbations of the adaptive weight on MNIST-USPS. $\hat{\epsilon}$ is the data-driven weight from Eq. (5) and δ is the perturbation. Unlike the main text, Recall Diff here is *signed* (unprotected minus protected recall), so its sign indicates which group is favored.

δ	Unprotected Recall (%)	Protected Recall (%)	Recall Diff (%)
-0.6	35.87±1.02	90.71±3.75	-54.85±4.77
-0.4	52.34±0.40	79.98±0.18	-27.64±0.42
-0.2	57.03±0.57	74.20±1.53	-17.17±1.12
-0.1	60.51±0.85	69.87±1.25	-9.36±0.79
0	35.03±2.08	6.60±0.71	28.43±1.74
0.2	33.94±1.98	7.74±1.07	26.20±2.97
0.4	34.54±2.27	7.22±0.47	27.32±2.69

Lemma 4.3 characterizes an interval of valid weights, but its endpoints depend on the four subgroup losses (UN, PN, UA, PA) and on counterfactual fitted models, so they cannot be computed without anomaly labels. We therefore study how FADIG responds when the weight is moved away from the data-driven value, training with $\epsilon_\delta = \mathrm{clip}(\hat{\epsilon} + \delta, 0.01, 0.99)$ for $\delta \in \{-0.6, -0.4, -0.2, -0.1, 0, 0.2, 0.4\}$. Note that clipping replaces the fallback rule used in the main experiments, where $\hat{\epsilon}$ is replaced by the group-ratio weight whenever it falls outside $(0, 1)$. Table 21 reports the per-group recall. The results follow the two failure modes predicted by Lemma 4.3. For negative δ, the weight on $\mathcal{L}_U$ increases, so the model also learns to reconstruct unprotected-group anomalies, which lowers their reconstruction error and hence the unprotected recall; the protected group is under-fitted, and its recall rises. The gap grows with the size of the perturbation, from -9.36% at $\delta = -0.1$ to -54.85% at $\delta = -0.6$. For positive δ, the perturbed weight saturates

at the upper clipping value 0.99 in almost every epoch, so the objective reduces to approximately $0.01\,\mathcal{L}_U + 0.99\,\mathcal{L}_P$. The model then reconstructs protected-group anomalies well, and protected recall falls to about 7%. Because the protected group has few training samples, this extreme weighting also provides too little signal to fit unprotected normal patterns, and unprotected recall drops to about 34%. This abrupt degradation is caused by saturation at the clipping boundary rather than by a gradual performance–fairness trade-off. Overall, moving ϵ in either direction shifts detection errors toward one group, consistent with the valid range in Lemma 4.3, and large perturbations degrade detection for both groups.

E Impact Statements

This work contributes to reducing bias and discrimination in AI applications by embedding fairness into anomaly detection algorithms, ensuring that technologies serve diverse populations equitably. In sectors such as finance, healthcare, and law enforcement, where anomaly detection plays a crucial role in identifying fraud, diseases, and criminal activities, incorporating fairness principles can prevent the perpetuation of historical biases and protect vulnerable groups from unjust outcomes. Furthermore, by advancing fairness in AI, this research aligns with global efforts to promote ethics in technology development, fostering trust between AI systems and their users.