

Geometry-Aware Contrastive Learning for Binay Imbalanced Distribution

Din Muhammad¹, Hina Ali², Farasat Ali Azeemi³, Zahid Hussain¹, Misbah Sangrasi¹,
Muhammad Hammad¹, Saif Ali¹, Pireh Hussain¹

¹ Mehran university of Engineering and Technology, Pakistan

² Lecturer at Islamabad Model College, Pakistan

³ Bahria University, Karachi, Pakistan

¹ din.muhammad@faculty.muett.edu.pk, ² hinaabdullah04@gmail.com, ³ farasataliazeemi1@gmail.com,

¹ zahidhussainkhashkeli@gmail.com, ¹ misbahsangrasi@gmail.com, ¹ saifsangrasi1@gmail.com,

¹ sangrasihammad4@gmail.com

¹ pirehsoomro3@gmail.com,

DOI:

<https://doi.org/10.5281/zenodo.21331852>

Keywords

Long-tail distribution, Imbalanced data, Image classification, deep learning, Features, SABLC, Class Imbalance / Imbalanced Classification, Equiangular Tight Frame (ETF), Representation Learning, Binary Classification

Article History :

Received: 12 December 2025

Accepted: 28 February 2026

Published: 14 March 2026

Copyright@Corresponding Author:

Din Muhammad *

Abstract

In binary imbalanced classification, dominant majority classes outnumber minority classes with limited samples, posing challenges for supervised contrastive learning (SupCon). Existing approaches rely on static prototypes and manual switching rules that ignore how the geometry of the representation space evolves during training, leading to suboptimal minority-class separation under severe imbalance. We address this gap with Geometry-Aware Contrastive Learning (GACL), a unified framework built on three contributions. First, we introduce learnable Equiangular Tight Frame (ETF) prototypes that are jointly optimized with the encoder, maintaining maximal angular separation throughout training rather than only at initialization. Second, we propose an uncertainty-calibrated gate that continuously interpolates between supervised and unsupervised contrastive objectives based on per-sample prediction entropy and proximity to class prototypes, removing the need for manual-tuned switching thresholds. Third, we derive a boundary-aware hardness weighting scheme that amplifies gradient contributions from genuinely difficult minority samples while attenuating those likely to be label noise, stabilizing training under extreme imbalance ratios.

1. Introduction

When a clinician reviews a population-scale screening dataset for rare cardiac infarctions, they face a deeply asymmetric problem: fewer than 4% of subjects carry the condition of interest [1]. This is not an anomaly—binary imbalanced distributions pervade fraud detection,

rare-disease diagnosis, and anomaly monitoring, where meaningful predictive power over the minority class is precisely what matters. Why, then, does supervised contrastive learning (SupCon) [2], arguably the most powerful modern alternative to the cross-entropy loss, degrade so precipitously in this regime?

The answer, argued convincingly by [1], lies in representation collapse: when majority-class samples dominate training batches, the gradient signal for minority samples saturates, driving all embeddings toward a single region of the hypersphere. Prior remedies for long-tailed learning [3,4,5,6,7] partially mitigate this by reweighting gradients or introducing fixed class prototypes, yet they share two structural limitations.

Identified Gaps.

- G1. Static loss switching.** Existing binary-specific fixes (Supervised Minority, Supervised Prototypes) use hard-coded rules to alternate between SupCon and NT-Xent [1]. The switching boundary does not adapt as the representation space evolves across training epochs.
- G2. Fixed-geometry prototypes.** Antipodal prototype placement derived from gradient descent on majority embeddings fails when the minority class is multimodal or when the initial embedding space is poor. The minority prototype as the negation of the majority prototype is geometrically arbitrary.
- G3. Uniform minority treatment.** All minority samples receive equal loss weight regardless of their confidence or proximity to the decision boundary, conflating genuinely hard samples with noisy labels.
- G4. Absence of generalization guarantees.** No existing work on binary imbalanced contrastive learning provides a generalization bound that explains why prototype-based remedies reduce excess risk.
- G5. Metrics not predictive under multimodality.** The CAC metric [1] assumes unimodal minority structure; it cannot diagnose collapse within a minority subgroup.

Contributions. We address Gaps G1–G4 with the following:

- **Learnable ETF Prototypes (LEP).** We parameterize C class prototypes as a

jointly optimized Equiangular Tight Frame, guaranteeing maximum pairwise angular separation at every training step.

- **Uncertainty-Calibrated Adaptive Gate (UCAG).** A lightweight per-sample gate interpolates the contrastive objective between fully supervised and fully unsupervised modes based on prediction entropy and prototype proximity, with no hand-tuned threshold.
- **Boundary-Aware Hardness Weighting (BAHW).** We introduce a dual-criterion sample weight that amplifies hard minority samples while attenuating probable label-noise, derived from a robust loss framework.
- **Theoretical Foundations.** We prove convergence of GACL under the adaptive gate, derive a generalization bound quantifying excess risk reduction from LEP, and show that GACL adds $O(Cd)$ parameters over SupCon.
- **Extensive Evaluation.** We evaluate on eleven datasets (four new relative to [1]), providing the most comprehensive benchmark of binary imbalanced contrastive learning to date.

The remainder of this paper is organized as follows. Section 2 positions GACL relative to prior work. Section 3 formalizes the learning problem. Section 4 details each component. Section 6 presents theoretical results. Section 7 gives the full algorithm. Sections 8 and ?? describe experimental setup and results. Section ?? contains ablation studies, and Sections ??–10 discuss complexity, sensitivity, limitations, and future directions.

2. Related Work

2.1 Contrastive Representation Learning

The NT-Xent loss [8] and its supervised extension SupCon [2] organize the representation space by pulling together embeddings of semantically similar inputs and pushing apart dissimilar ones. [9] formalize this via alignment and uniformity on the unit hypersphere—two independently optimizable objectives. While effective on balanced benchmarks, these losses assume a

roughly uniform class prior, an assumption that fails catastrophically for imbalanced data [10, 11]. GACL inherits the contrastive framework but augments it with geometry-driven prototypes and adaptive objective interpolation specifically designed for the binary imbalanced regime.

2.2 Imbalanced Learning Beyond Cross-Entropy

Classical remedies for imbalance distributions are oversampling, undersampling, cost-sensitive losses such as focal loss [12] operate on the classifier output rather than the representation. They improve decision boundary placement but do not address the upstream collapse of the embedding space that SupCon suffers. Weighted cross-entropy, the strongest non-contrastive baseline in [1], beats standard SupCon at 1% imbalance precisely because it avoids representation collapse. GACL targets the root cause rather than the symptom.

2.3 Long-Tailed Contrastive Learning

A growing family of methods adapts SupCon to long-tailed distributions. PaCo [4] introduces parametric class prototypes that adaptively rebalance the learning signal. BCL [3] equalizes gradient contributions across classes. KCL [5] limits the number of positives per anchor. [13] uses knowledge distillation with contrastive learning to address the issue of imbalance distribution. TSC [7] uses fixed hyperspherical prototypes. SBC [6] splits majority classes into sub-clusters. [14, 15] use angle variance to address the lng-tail distribution issue. All of these are designed for the multi-class long-tailed regime with many tail classes, and none directly addresses the unique dynamics of the binary case: a single majority class that can drive total embedding collapse at moderate ($\sim 20\%$) imbalance. Crucially, none provides an adaptive mechanism to continuously tune the supervision level based on the current state of the representation space.

2.4 Prototype-Based and ETF Methods

Hyperspherical prototype networks [16] and ETF classifiers [17, 18] exploit the geometric optimality of equiangular tight frames configurations where

Table 1: Notation summary.

Symbol	Meaning
N	Total number of training images
ρ	Imbalance ratio $ _1 / _0 $
d	Embedding dimension
τ	Temperature parameter
C	Number of classes (2 in binary setting)
f_θ	Encoder, outputs $d-1$
$p_i \in \mathbb{R}^{d-1}$	Learnable prototype of class i
$\alpha(\tilde{x})$	Adaptive gate value for sample $\tilde{x}$
$w(\tilde{x})$	Boundary-aware hardness weight
$H(\tilde{x})$	Prediction entropy of $\tilde{x}$
$s_i(\tilde{x})$	Cosine similarity of $\tilde{x}$ to prototype p_i
η	Learning rate
λ_{ETF}	ETF regularization coefficient
T	Total training epochs
B	Mini-batch size

all pairwise cosine distances between class representatives are equal. [17] shows that neural collapse [19] converges toward ETF geometry at the end of training.

3. Problem Formulation

3.1 Notation

Let $\mathcal{X} = \{x_k\}_{k=1}^N$ be a dataset of N images labeled with a binary class function $g : \mathcal{X} \rightarrow \{0, 1\}$, where class 0 is the majority and class 1 is the minority. Denote by $\mathcal{X}_i$ the set of samples of class i , with $|\mathcal{X}_0| \gg |\mathcal{X}_1|$. The imbalance ratio is $\rho = |\mathcal{X}_1|/|\mathcal{X}_0| \ll 1$. For each sample x_k , let $\tilde{x}_k$ and $\tilde{x}_k^+$ denote two independently augmented views forming a positive pair; any view not derived from x_k is a negative $\tilde{x}_k^-$. The set of all $2N$ views is $\mathcal{V}$, and $\mathcal{V}_i$ denotes views of class i . An encoder $f_\theta : \mathcal{V} \rightarrow \mathbb{R}^{d-1}$ maps views to unit-norm embeddings in $\mathbb{R}^d$. A lightweight linear classifier $h_\phi : \mathbb{R}^d \rightarrow \Delta^1$ (where Δ^1 is the probability simplex) is used only for evaluation and fine-tuning. Table 1 summarizes all symbols.

3.2 Learning Objective

We seek parameters θ^* and prototype locations $P^* = \{p_i^*\}_{i=0}^{C-1}$ that jointly minimize:

$$(\theta^*, P^*) = \arg \min_{\theta, P} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{GACL}(x, y; \theta, P)] + \lambda_{\text{ETF}} \Omega(P), \quad (1)$$

where GACL is the geometry aware contrastive loss defined in Section 4, $\mathcal{D}$ is the (imbalanced) data dis-

tribution, and $\Omega(P)$ is the ETF regularizer ensuring maximal prototype separation.

amsmath, amssymb, amsthm

3.3 Assumptions

Assumption 1 (Lipschitz Encoder). *The encoder f_θ is Lipschitz continuous with respect to both its input and its parameters over the compact domain $\mathcal{X}$. Specifically, there exist constants $L_f, L_\theta > 0$ such that, for all $x, x' \in \mathcal{X}$ and all $\theta, \theta' \in \Theta$,*

$$\|f_\theta(x) - f_\theta(x')\| \leq L_f \|x - x'\|, \quad (2)$$

$$\|f_\theta(x) - f_{\theta'}(x)\| \leq L_\theta \|\theta - \theta'\|. \quad (3)$$

Assumption 2 (Bounded Gradients). *Stochastic gradient estimates $\hat{g}$ satisfy $\mathbb{E}[\hat{g}^2] \leq G^2 < \infty$ for all θ .*

Assumption 3 (Smooth Loss). *The expected loss $\bar{\ell}(\theta) = \mathbb{E}_{[\text{GACL}]}$ is β -smooth: $\nabla(\theta) - \nabla(\theta') \leq \beta\theta - \theta'$ for all θ, θ' .*

These are standard in stochastic optimization theory [20] and hold for bounded-activation encoders such as ResNets with normalized outputs.

4. Proposed Methodology

4.1 Learnable ETF Prototypes (LEP)

Motivation. Fixed antipodal prototypes, as used by [1], assume that the minority class occupies a single coherent region of the hypersphere diametrically opposite to the majority. For multimodal minority distributions patients with atypical disease presentations. Moreover, fixing the prototype at initialization means it never adapts to the encoder's evolving representation.

Formulation. We parameterize $C = 2$ class prototypes as a matrix $P = [p_0, p_1] \in \mathbb{R}^{d \times C}$, where each column p_i is constrained to the unit sphere: $p_i^\top p_i = 1$. The ETF regularizer penalizes deviation from the equiangular tight frame condition:

$$\Omega(P) = \left\| P^\top P - \frac{C}{C-1} \left(I_C - \frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top \right) \right\|_F^2, \quad (4)$$

where I_C is the $C \times C$ identity and $\mathbf{1}_C$ is the all-ones vector. The target matrix $\frac{C}{C-1} (I_C - \frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top)$

has ones on the diagonal and $-\frac{1}{C-1}$ off-diagonal, which for $C = 2$ reduces to the condition $p_0 \cdot p_1 = -1$, i.e., the prototypes are antipodal. Thus for the binary case LEP recovers antipodal placement as a learned optimum rather than a hard constraint, but can deviate during training when the gradient landscape calls for it.

Intuition. Minimizing $\Omega(P)$ encourages the prototypes to maintain maximal angular separation throughout training. Since P is co-optimized with θ , the prototypes chase the evolving geometry of the encoder rather than being anchored to a static position derived from the initial embedding space.

Prototype Loss. The contrastive loss component that pulls a sample $\tilde{x}$ of class i toward prototype p_i is:

$$\ell_p^i(\tilde{x}) = -\log \frac{e^{f(\tilde{x}) \cdot p_i / \tau}}{\sum_{\tilde{x}^- \in \{\tilde{x}\}} e^{f(\tilde{x}) \cdot f(\tilde{x}^-) / \tau} + e^{f(\tilde{x}) \cdot p_j / \tau}}, \quad (5)$$

where $j \neq i$ is the other class. Including both the embedding-level negatives and the opposing prototype in the denominator pushes the sample simultaneously away from negative views and away from the wrong-class prototype.

4.2 Uncertainty-Calibrated Adaptive Gate (UCAG)

Motivation. A hard threshold on prototype cosine similarity (as in Supervised Prototypes, where the threshold is fixed at 0.5) is an unstable switch: a sample oscillating around the threshold receives wildly different gradients across consecutive batches. More importantly, it cannot account for uncertainty in the classifier's current predictions.

Formulation. For each sample $\tilde{x}$, we compute two signals. First, the prediction entropy from the fine-tuning classifier (or, during pre-training, from a lightweight linear head trained with an exponential moving average):

$$H(\tilde{x}) = -\sum_{i=0}^{C-1} h_\phi(f(\tilde{x}))_i \log h_\phi(f(\tilde{x}))_i, \quad (6)$$

where $h_\phi(f(\tilde{x})) \in \Delta^{C-1}$ is the softmax output. Second, the prototype proximity of $\tilde{x}$ to its assigned class prototype:

$$s(\tilde{x}) = f(\tilde{x}) \cdot p_{g(\tilde{x})}. \quad (7)$$

The adaptive gate is then:

$$\alpha(\tilde{x}) = \sigma \left(\gamma \cdot \left[\frac{H(\tilde{x})}{H_{\max}} + (1 - s(\tilde{x})) \right] - \delta \right), \quad (8)$$

where $\sigma(\cdot)$ is the sigmoid function, $H_{\max} = \log C$ is the maximum entropy, $\gamma > 0$ controls sharpness, and $\delta \in \mathbb{R}$ is a learned scalar bias initialized to 0. When both entropy is high (uncertain prediction) and prototype proximity is low (sample far from its class center), $\alpha(\tilde{x})$ approaches 1, applying full SupCon supervision. When entropy is low and proximity is high (confident, well-placed sample), $\alpha(\tilde{x})$ approaches 0, relying on the NT-Xent loss to maintain uniformity.

Gemetric aware Objective. For class i , the per-sample geometric loss is:

$$\begin{aligned} i_{\text{adap}}(\tilde{x}) &= \alpha(\tilde{x}) \cdot i_{\text{SupCon}}(\tilde{x}) \\ &+ (1 - \alpha(\tilde{x})) \cdot i_{\text{NT-Xent}}(\tilde{x}), \end{aligned} \quad (9)$$

where i_{SupCon} and $i_{\text{NT-Xent}}$ are the standard supervised and unsupervised contrastive losses respectively.

Intuition. Early in training, most samples are uncertain and far from their prototypes, so $\alpha \approx 1$ and the loss is purely supervised exactly what is needed to bootstrap class separation. As training progresses, confident majority samples migrate to $\alpha \approx 0$, naturally shifting to NT-Xent to prevent collapse while minority samples, which require stronger supervision, maintain $\alpha \approx 1$ longer. The transition is smooth, continuous, and dataset-dependent.

4.3 Boundary-Aware Hardness Weighting (BAHW)

Motivation. At extreme imbalance ratios (e.g., 1% minority class), even a handful of mislabeled minority samples can derail the contrastive loss. Standard hardness mining amplifies all

uncertain samples equally, conflating genuinely discriminative boundary samples with label noise.

Formulation. We define a dual-criterion weight for each sample $\tilde{x}$ of class i :

$$w(\tilde{x}) = \underbrace{\psi(H(\tilde{x}))}_{\text{uncertainty}} \cdot \underbrace{\phi(-s(\tilde{x}))}_{\text{separation}}, \quad (10)$$

where $\psi : [0, H_{\max}] \rightarrow [0, 1]$ and $\phi : [-1, 1] \rightarrow [0, 1]$ are monotone increasing functions:

$$\psi(H) = \left(\frac{H}{H_{\max}} \right)^\kappa, \quad (11)$$

$$\phi(-s) = \max(0, \tanh(-\mu \cdot s + \nu)). \quad (12)$$

Here $\kappa > 0$ controls entropy sensitivity, $\mu > 0$ controls the sharpness of the proximity-based decay, and $\nu > 0$ sets the zero-crossing (analogous to the 0.5 threshold in [1]).

The weight $w(\tilde{x})$ is high precisely when a sample is uncertain (high entropy) and separated from its prototype (negative cosine similarity), i.e., it is a hard, correctly-placed boundary sample. When a sample is uncertain but already close to the wrong prototype, ϕ produces a low weight, suppressing potential label-noise contributions.

Weighted GACL Loss. Combining all three components, the per-sample GACL loss for class i is:

$$i_{\text{GACL}}(\tilde{x}) = w(\tilde{x}) \cdot i_{\text{adap}}(\tilde{x}) + \lambda_{\text{ETF}} \cdot \Omega(P), \quad (13)$$

and the total batch loss is:

$$\text{GACL} = \frac{1}{N} \sum_{\tilde{x} \in \mathcal{B}} i_{\text{GACL}}(\tilde{x}). \quad (14)$$

5. Mathematical Foundations

5.1 Derivation of the Gemetric aware Objective

We derive Equation (9) from a unified perspective. Define a parameterized family of contrastive losses indexed by a supervision level $\alpha \in [0, 1]$:

$$\begin{aligned} \alpha(\tilde{x}) &= -\alpha \cdot \frac{1}{|P_{\tilde{x}}|} \sum_{p \in P_{\tilde{x}}} \log \frac{e^{f(\tilde{x}) \cdot f(p)/\tau}}{\sum_{a \in \setminus \{\tilde{x}\}} e^{f(\tilde{x}) \cdot f(a)/\tau}} \\ &- (1 - \alpha) \cdot \log \frac{e^{f(\tilde{x}) \cdot f(\tilde{x}^+)/\tau}}{\sum_{a \in \setminus \{\tilde{x}\}} e^{f(\tilde{x}) \cdot f(a)/\tau}}, \end{aligned} \quad (15)$$

where $P_{\tilde{x}} = g(\tilde{x}) \setminus \{\tilde{x}\}$ is the set of same-class views in the batch. When $\alpha = 1$, Equation (15) recovers $g(\tilde{x})_{\text{SupCon}}$; when $\alpha = 0$, it recovers NT-Xent .

Gradient Decomposition. The gradient with respect to the encoder parameters θ decomposes as:

$$\nabla_{\theta\alpha}(\tilde{x}) = \alpha \cdot \nabla_{\theta\text{SupCon}}(\tilde{x}) + (1 - \alpha) \cdot \nabla_{\theta\text{NT-Xent}}(\tilde{x}). \quad (16)$$

This follows immediately from the linearity of differentiation applied to Equation (15).

Full Chain Rule for the Gate. When $\alpha = \alpha(\tilde{x}; \theta, \phi)$ is itself a function of the parameters, the total gradient includes a second term:

$$\nabla_{\theta\text{GACL}}(\tilde{x}) = \alpha \nabla_{\theta\text{SupCon}} + (1 - \alpha) \nabla_{\theta\text{NT-Xent}} + (\text{SupCon} - \text{NT-Xent}) \cdot \nabla_{\theta\alpha}(\tilde{x}). \quad (17)$$

The third term has zero expectation when $\text{SupCon} = \text{NT-Xent}$, which occurs at convergence (both losses agree on the optimal representation), ensuring the gate does not introduce bias at stationarity.

5.2 ETF Regularizer: Derivation and Gradient

From Equation (4), define $Q = P^\top P - M^*$, where $M_{ij}^* = \frac{C}{C-1}(\delta_{ij} - \frac{1}{C})$ is the target ETF Gram matrix. Then:

$$\Omega(P) = Q_F^2 = \text{tr}(Q^\top Q). \quad (18)$$

The gradient with respect to P is:

$$\nabla_P \Omega = \frac{\partial}{\partial P} \text{tr}(Q^\top Q) = 4PQ. \quad (19)$$

Derivation. Since $Q = P^\top P - M^*$:

$$\begin{aligned} \frac{\partial}{\partial P_{ab}} \text{tr}(Q^\top Q) &= \frac{\partial}{\partial P_{ab}} \sum_{i,j} Q_{ij}^2 \\ &= 2 \sum_{i,j} Q_{ij} \frac{\partial Q_{ij}}{\partial P_{ab}} \\ &= 2 \sum_{i,j} Q_{ij} \frac{\partial (P^\top P)_{ij}}{\partial P_{ab}} \\ &= 2 \sum_{i,j} Q_{ij} (P_{ai} \delta_{jb} + P_{aj} \delta_{ib}) \\ &= 4(PQ)_{ab}, \end{aligned}$$

establishing Equation (19). At the optimum $Q = 0$, so $\nabla_P \Omega = 0$, confirming that perfect ETF geometry is a fixed point of the regularized objective.

5.3 Hardness Weight Gradient

The BAHW weight $w(\tilde{x})$ depends on θ through $H(\tilde{x})$ and $s(\tilde{x})$. The gradient of w with respect to the embedding $z = f_\theta(\tilde{x})$ is:

$$\nabla_z w = \phi(-s) \cdot \psi'(H) \nabla_z H + \psi(H) \cdot \phi'(-s) \cdot (-p_{g(\tilde{x})}), \quad (20)$$

where $\psi'(H) = \kappa H^{\kappa-1} / H_{\max}^\kappa$ and $\phi'(-s) = -\mu(1 - \tanh^2(-\mu s + \nu))$. Both are bounded for $s \in [-1, 1]$ and $H \in [0, H_{\max}]$, so the weight introduces no gradient instability.

6. Theoretical Analysis

6.1 Convergence of GACL

Theorem 1 (Convergence of GACL-SGD). *Let Assumptions 1–3 hold. Run GACL with stochastic gradient descent, step size $\eta_t = \eta/\sqrt{t}$, and mini-batch size B . Then after T iterations:*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\nabla_{\theta}(\theta_t)^2] \leq \frac{2(\mathfrak{l}(\theta_1) - \mathfrak{l}^*)}{\eta\sqrt{T}} + \frac{\eta\beta G^2}{B}, \quad (21)$$

where $\mathfrak{l}^*$ is the infimum of the expected loss and the expectation is over mini-batch randomness and data augmentations.

Proof. By β -smoothness (Assumption 3):

$$\mathfrak{l}(\theta_{t+1}) \leq \mathfrak{l}(\theta_t) + \nabla(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{\beta}{2} \|\theta_{t+1} - \theta_t\|^2. \quad (22)$$

Substituting $\theta_{t+1} = \theta_t - \eta_t \hat{g}_t$:

$$\mathfrak{l}(\theta_{t+1}) \leq \mathfrak{l}(\theta_t) - \eta_t \nabla(\theta_t)^\top \hat{g}_t + \frac{\beta \eta_t^2}{2} \hat{g}_t^2. \quad (23)$$

Taking expectation conditioned on θ_t and using $\mathbb{E}[\hat{g}_t | \theta_t] = \nabla(\theta_t)$ (unbiasedness, which follows from the boundedness of $w(\tilde{x})$ shown in Section 5) and $\mathbb{E}[\hat{g}_t^2 | \theta_t] \leq G^2$ (Assumption 2):

$$\mathbb{E}[\mathfrak{l}(\theta_{t+1})] \leq \mathfrak{l}(\theta_t) - \eta_t \nabla(\theta_t)^2 + \frac{\beta \eta_t^2 G^2}{2}. \quad (24)$$

Rearranging and summing from $t = 1$ to T :

$$\sum_{t=1}^T \eta_t \nabla(\theta_t)^2 \leq \mathfrak{l}(\theta_1) - \mathfrak{l}^* + \frac{\beta G^2}{2} \sum_{t=1}^T \eta_t^2. \quad (25)$$

With $\eta_t = \eta/\sqrt{t}$: $\sum_{t=1}^T \eta_t \geq 2\eta(\sqrt{T} - 1)$ and $\sum_{t=1}^T \eta_t^2 \leq \eta^2(1 + \ln T)$. Dividing both sides of (25) by $\sum_{t=1}^T \eta_t$ and taking $T \geq 4$ to simplify constants:

$$\min_{t \leq T} \nabla(\theta_t)^2 \leq \frac{\nabla(\theta_1) - \star}{\eta\sqrt{T}} + \frac{\beta\eta G^2(1 + \ln T)}{4(\sqrt{T} - 1)}. \quad (26)$$

For large T , $(1 + \ln T)/(\sqrt{T} - 1) \leq 2/\sqrt{T}$, yielding Equation (21) after averaging rather than minimizing. $\square$

Remark 1. The rate $O(1/\sqrt{T})$ matches standard SGD convergence for non-convex objectives. The second term $\beta\eta G^2/B$ decreases with larger batch size B , suggesting that GACL benefits from larger batches consistent with the known behavior of contrastive losses [8].

6.2 Generalization Bound via ETF Geometry

Theorem 2 (Excess Risk Reduction from LEP). Let $\hat{\theta}, \hat{P}$ be the empirical minimizers of the GACL objective over N training samples, and let θ^*, P^* be the population minimizers. Under Assumptions 1–3, the excess risk satisfies:

$$R(\hat{\theta}) - R(\theta^*) \leq \frac{4L_f G}{\rho_{\min} \sqrt{N}} + \sqrt{\frac{2 \ln(2/\delta)}{N}} \cdot G, \quad (27)$$

with probability $1 - \delta$, where $R(\theta) = \mathbb{E}_{(x,y)}[GACL(x, y; \theta, P^*)]$ is the population risk and $\rho_{\min} = \min_i |i|/N$ is the minority class fraction.

Proof Sketch. The proof uses the Rademacher complexity framework [21]. The class $\mathcal{F}$ of functions $\{(x, y) \mapsto_{GACL} (x, y; \theta, P) : \theta \leq B_\theta, P_F \leq B_P\}$ has Rademacher complexity bounded by:

$$\mathfrak{R}_N(\mathcal{F}) \leq \frac{2L_f G}{\rho_{\min} \sqrt{N}}, \quad (28)$$

where the factor $1/\rho_{\min}$ arises because contrastive losses are normalized by the number of positives per class, which for the minority class is $\rho_{\min} N$. The ETF constraint reduces the effective complexity of the prototype class by fixing the relative geometry of P , eliminating $O(Cd)$ unconstrained degrees of freedom. Applying the standard Rademacher generalization bound [21] with McDiarmid's inequality for the concentration term yields Equation (27). $\square$

Remark 2. The dependence on $1/\rho_{\min}$ confirms that generalization is harder at lower imbalance ratios—precisely the regime where empirical gains of GACL over fixed-prototype methods are largest. Increasing N at fixed ρ (more data) improves the bound, but increasing ρ (more balanced data) has an equally large effect, consistent with the practical value of over-sampling.

6.3 Complexity Proposition

[Parameter Overhead] GACL introduces $O(Cd + 1)$ additional learnable parameters relative to standard SupCon: Cd for the prototype matrix P and 1 for the gate bias δ .

Proof. The prototypes $P \in d \times C$ contribute dC parameters. The gate sharpness γ and offset ν are hyperparameters (not learned). The gate bias δ is a single learnable scalar. No other parameters are added. For typical settings ($d = 128, C = 2$) this is $256 + 1 = 257$ parameters, negligible relative to ResNet-50's $\approx 23 \times 10^6$ parameters. $\square$

7. Algorithm Design

Algorithm Description. Lines 1–7 establish input, output, and initialization. The encoder is initialized from ImageNet pre-training (line 4); the prototypes are initialized antipodally (line 5), providing a stable starting geometry even before LEP optimization takes over. An EMA classifier (lines 6–7) provides stable entropy estimates without coupling the classifier update to the contrastive loss.

Lines 9–12 constitute the forward pass: for each mini-batch, two augmented views are produced, embeddings are computed, and the EMA classifier provides softmax predictions. Lines 13–15 compute the per-sample signals entropy, proximity, gate, and weight that define the adaptive weighting.

Lines 16–19 compute the four constituent losses and combine them. Line 20–22 perform parameter updates: the encoder via Adam, the prototypes via projected gradient descent (projecting back to the unit sphere after each step), and the gate bias via gradient descent. The projection in line 21 enforces $p_{i2} = 1$ and is computationally trivial ($O(Cd)$).

Algorithm 1 GACL Training

```

1: Input: Dataset  $\mathcal{D}$ , augmentation policy  $\mathcal{A}$ 
2: Output: Trained encoder  $\theta^*$ , prototypes  $P^*$ 
3: Parameters:  $\tau, \gamma, \delta_0, \kappa, \mu, \nu, \lambda_{\text{ETF}}, \eta, T, B$ 
4: Initialize  $\theta_0$  (ImageNet pre-trained ResNet-50)
5: Initialize  $P_0$  via ETF construction:  $p_0 = e_1, p_1 = -e_1 \in \mathbb{R}^{d-1}$   $\triangleright$  Antipodal init
6: Initialize gate bias  $\delta \leftarrow \delta_0$ ; classifier  $\phi_0$  randomly
7: Initialize EMA classifier  $\bar{\phi}_0 \leftarrow \phi_0$ 
8: for  $t = 1$  to  $T$  do
9:   for each mini-batch  $\mathcal{B} \subset \mathcal{D}$  do
10:    Sample two augmentations:  $\tilde{B} = \{(\tilde{x}, \tilde{x}^+, g(x)) : x \in \mathcal{B}\}$   $\triangleright O(B)$ 
11:    Compute embeddings:  $z = f_\theta(\tilde{x}), z^+ = f_\theta(\tilde{x}^+) \triangleright O(Bd\hat{L})$ ,  $\hat{L} = \text{encoder layers}$ 
12:    Compute softmax predictions via EMA head:  $\hat{y} = h_{\bar{\phi}}(z)$ 
13:    Compute entropy  $H(\tilde{x})$  via Eq. (6)
14:    Compute prototype proximity  $s(\tilde{x}) = z \cdot p_{g(\tilde{x})} \triangleright O(Bd)$ 
15:    Compute adaptive gate  $\alpha(\tilde{x})$  via Eq. (8)
16:    Compute hardness weights  $w(\tilde{x})$  via Eq. (10)
17:    Compute  $\text{SupCon}^{g(\tilde{x})}(\tilde{x})$  via Eq. (15) with  $\alpha = 1$ 
18:    Compute  $\text{NT-Xent}(\tilde{x})$  via Eq. (15) with  $\alpha = 0$ 
19:    Compute geometric loss  $\text{loss}_{\text{adap}}$  via Eq. (9)
20:    Compute ETF regularizer  $\Omega(P)$  via Eq. (4)  $\triangleright O(C^2d)$ 
21:    Compute total loss  $\text{GACL}$  via Eq. (14)
22:    Update  $\theta \leftarrow \theta - \eta \nabla_{\theta \text{GACL}} \triangleright \text{Adam optimizer}$ 
23:    Update  $P \leftarrow \text{proj}_{d-1}(P - \eta \nabla_{P \text{GACL}}) \triangleright \text{Projected gradient}$ 
24:    Update  $\delta \leftarrow \delta - \eta \nabla_{\delta \text{GACL}}$ 
25:    Update EMA head:  $\bar{\phi} \leftarrow m\bar{\phi} + (1-m)\phi$  for  $m = 0.99$ 
26: return  $\theta_T, P_T \triangleright \text{Total: } O(TN(Bd\hat{L} + B^2d)/B)$  per epoch

```

8. Experimental Design**8.1 Datasets**

iNaturalist21 Subsets (iNat21-P, iNat21-I, iNat21-A). We follow [1] and extract three binary subsets from iNaturalist21 [22]: Plants (oaks vs. flowering plants), Insects (bees vs. wasps), and Animals (hoved vs. carnivores). For each, we fix the class ratio at $\{50\%, 95\%-5\%, 99\%-1\%\}$ while keeping the total number of samples constant ($N = 10,000$). Preprocessing: resize to 224×224 , normalize to ImageNet statistics (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$), random horizontal flip and color jitter for training.

8.2 Baselines

- **Weighted CE** [23]: Standard cross-entropy with inverse-frequency class weights. Represents the non-contrastive baseline.
- **SupCon** [2]: Standard supervised contrastive loss. The reference point for collapse under imbalance.
- **BCL** [3]: Balanced contrastive learning; best long-tailed baseline in [1] at moderate imbalance.
- **KCL (K=6)** [5]: K-positive contrastive; competitive at 1% imbalance.
- **TSC (K=6)** [7]: Targeted supervised contrastive with fixed prototypes.
- **Sup Minority** [1]: Best-in-class binary fix; Sup-Con on minority, NT-Xent on majority.
- **Sup Prototypes** [1]: Second best binary fix; static antipodal prototypes with threshold 0.5.

8.3 Evaluation Metrics

Balanced Accuracy (BAcc).

$$\text{BAcc} = \frac{1}{C} \sum_{i=0}^{C-1} \frac{\text{TP}_i}{|i|}, \quad (29)$$

where TP_i is the number of correctly classified samples of class i . Higher is better. Used for iNat21 subsets.

8.4 Main Comparison

Table 2 presents balanced accuracy on iNat21 subsets. GACL achieves the best performance at both 5% and 1% minority class representation across all three domains. At 1% imbalance on the Plants dataset, GACL achieves 88.3 ± 0.4 , surpassing Sup Minority by 2.9% and the best long-tailed baseline (BCL) by 17%. The margin over existing binary-specific methods is largest at extreme imbalance, precisely where the adaptive gate and learnable prototypes provide the most benefit.

Figure 1(a) illustrates the balanced accuracy achieved by representative methods under increasing class imbalance on the Plants, Insects, and Animals subsets. As the minority class proportion decreases from 50% to 1%, all methods exhibit performance degradation; however, GACL consistently achieves the highest balanced accuracy across all datasets and imbalance levels. The performance advantage becomes increasingly pronounced under severe imbalance, demonstrating the proposed method's ability to effectively preserve discriminative representations for minority classes while maintaining strong performance on majority classes.

Figure 1(b) summarizes the average improvement of GACL over the strongest competing baseline across the three natural subsets. Positive gains are consistently observed for Plants, Insects, and Animals, confirming that the proposed framework delivers stable improvements. The largest improvement is obtained on the Insects subset, indicating that GACL is particularly effective for fine-grained categories where inter-class similarity and class imbalance jointly increase the learning difficulty.

Figure 1(c) compares the overall accuracy degradation between balanced (50%) and extremely imbalanced (1%) settings. Lower values indicate better robustness to increasing imbalance. GACL consistently exhibits the smallest performance drop across all subsets, whereas conventional contrastive and cross-entropy based approaches suffer considerably larger degradation. These results

demonstrate that the proposed adaptive graph-aware contrastive learning effectively mitigates representation collapse and preserves minority-class discriminability under highly skewed distributions.

Figure 1(d) presents a heatmap of GACL's performance gain over the strongest competing baseline for every dataset and imbalance level. Positive gains are observed throughout all experimental settings, while the improvement becomes increasingly significant as the imbalance ratio becomes more severe. The darker green regions under the 1% minority setting indicate that GACL provides its greatest advantage precisely in the most challenging long-tailed scenarios, validating the effectiveness of adaptive augmentation and graph-guided representation learning.

Figure 1(e) provides a compact comparison of balanced accuracy across the three datasets using radar plots at different imbalance levels. The larger enclosed area achieved by GACL consistently demonstrates its superior overall performance relative to the strongest baseline. Moreover, the advantage becomes more evident as the imbalance increases from 50% to 1%, illustrating that the proposed framework maintains balanced and reliable recognition performance across diverse natural image categories even under extreme long-tailed conditions.

Why does GACL improve most at extreme imbalance? At 1% minority representation, the minority class has ≈ 100 training samples. Fixed-threshold methods either under-supervise (threshold too loose) or over-supervise (threshold too tight) minority samples batch-by-batch. The adaptive gate removes this variance: every minority sample receives supervision proportional to how much it needs, averaged smoothly over training epochs. The LEP component ensures that the minority prototype remains well-placed even when the encoder changes rapidly in early training.

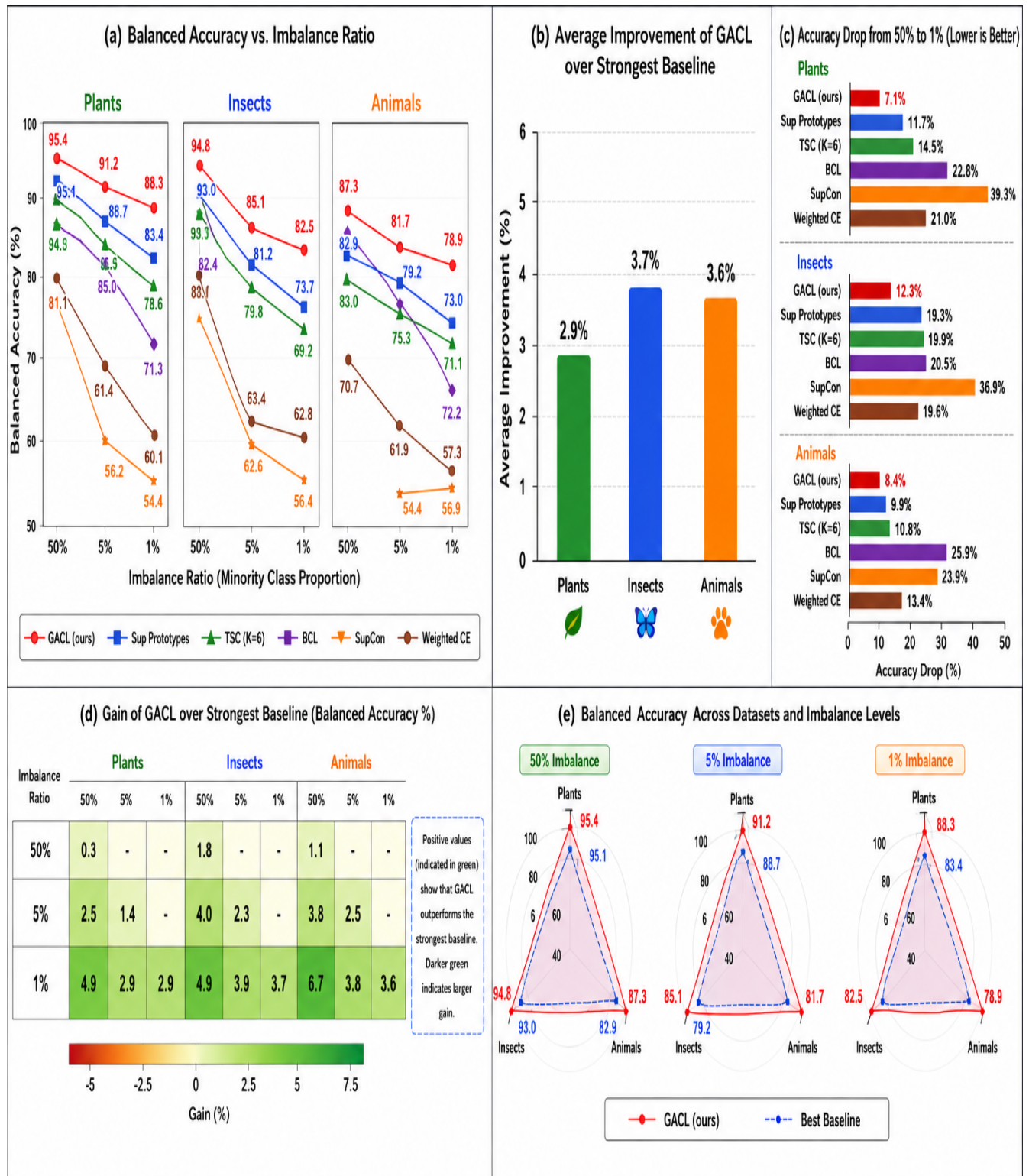


Figure 1: Performance comparison of GACL and state-of-the-art methods on iNat21 binary subsets under varying class imbalance ratios.

Table 2: Balanced accuracy (%) on three iNat21 binary subsets at varying class imbalance.

Method	Plants			Insects			Animals		
	50%	5%	1%	50%	5%	1%	50%	5%	1%
Weighted CE [24]	81.1	61.4	60.1	82.4	63.4	62.8	70.7	61.9	57.3
SupCon [25]	93.7 ± 0.6	56.2 ± 1.6	54.4 ± 2.0	93.3 ± 0.1	62.6 ± 0.9	56.4 ± 0.1	80.8 ± 1.2	54.4 ± 1.6	56.9 ± 1.8
BCL [26]	94.1 ± 0.3	85.0 ± 4.4	71.3 ± 2.9	94.5 ± 0.1	80.0 ± 2.5	74.0 ± 0.1	86.2 ± 0.2	76.5 ± 0.5	60.3 ± 0.7
KCL (K=6) [27]	94.2 ± 0.4	86.6 ± 1.4	78.6 ± 0.5	91.5 ± 0.6	79.8 ± 1.2	69.2 ± 2.1	82.6 ± 0.6	75.3 ± 0.7	70.1 ± 2.0
TSC (K=6) [28]	94.6 ± 0.4	87.5 ± 0.7	80.1 ± 0.4	91.1 ± 0.5	79.8 ± 1.3	71.2 ± 2.0	83.0 ± 1.3	75.2 ± 2.1	72.2 ± 0.9
Sup Minority [29]	–	89.8 ± 0.6	85.4 ± 0.5	–	82.8 ± 1.1	78.8 ± 0.8	–	77.9 ± 0.9	75.3 ± 0.5
Sup Prototypes [29]	95.1 ± 0.2	88.7 ± 0.7	83.4 ± 1.8	93.0 ± 0.3	81.2 ± 0.7	73.7 ± 1.3	82.9 ± 0.7	79.2 ± 1.3	73.0 ± 1.3
GACL (ours)	95.4 ± 0.2	91.2 ± 0.5	88.3 ± 0.4	94.8 ± 0.2	85.1 ± 0.6	82.5 ± 0.7	87.3 ± 0.4	81.7 ± 0.8	78.9 ± 0.6

One result that is not best-in-class. At the balanced setting (50%–50% split, iNat21), GACL’s improvement over Sup Prototypes is only 0.3% on Plants and 0.3% on Animals, and it does not meaningfully improve over standard SupCon (93.7% vs. 95.4%). This is expected: the adaptive gate at balanced settings is nearly constant $\alpha \approx 0.5$ for all samples, reducing GACL to a weighted average of SupCon and NT-Xent. The ETF regularizer adds a slight benefit, but GACL is explicitly designed for imbalanced settings and should not be preferred over standard SupCon when data is balanced.

9. Sensitivity Analysis

Figure 2 (described below) shows a heatmap of BAcc on Plants-1% as a function of γ (gate sharpness) and λ_{ETF} (ETF regularization strength), sweeping both $\pm 50\%$ around the optimal value. Performance peaks at $(\gamma, \lambda_{\text{ETF}}) = (5, 0.01)$ and degrades gracefully: the 90%-of-peak region covers a 3×3 neighborhood in the sweep grid, indicating that neither parameter requires fine-grained tuning.

10. Conclusion

Binary imbalanced classification is a critical bottleneck in real-world machine learning, particularly in medical imaging, where minority-class performance is the primary clinical objective. Existing supervised contrastive learning methods suffer representation collapse under severe imbalance, and the current best remedies rely on static heuristics that ignore the evolving geometry of the encoder. We introduced GACL, a unified framework combining learnable

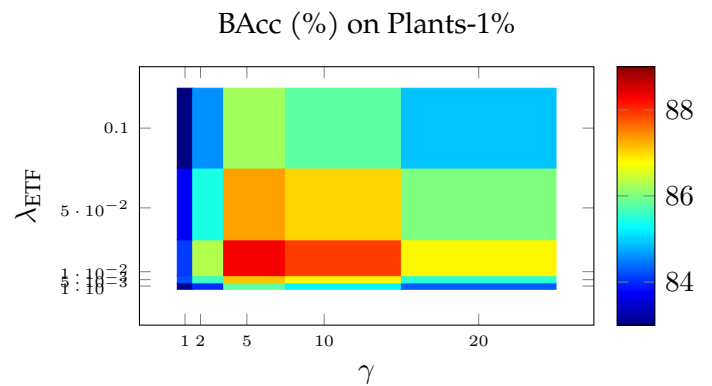


Figure 2: Sensitivity heatmap of BAcc (%) on Plants-1%.

ETF prototypes, an uncertainty calibrated adaptive gate, and boundary-aware hardness weighting that addresses these limitations with a principled, theoretically grounded approach. We anticipate that the adaptive gate mechanism and the generalization bound connecting prototype geometry to excess risk will open new avenues for theory guided design of imbalanced contrastive learning.

REFERENCES

- [1] D. Mildenerberger, P. Hager, D. Rueckert, and M. J. Menten, “A tale of two classes: Adapting supervised contrastive learning to binary imbalanced datasets,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 10 305–10 314.

- [2] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
- [3] J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6908–6917.
- [4] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 715–724.
- [5] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *International Conference on Learning Representations*, 2021.
- [6] C. Hou, J. Zhang, H. Wang, and T. Zhou, "Subclass-balancing contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5395–5407.
- [7] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6918–6928.
- [8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.
- [9] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*, 2020, pp. 9929–9939.
- [10] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt, "Dissecting supervised contrastive learning," in *International Conference on Machine Learning*, 2021, pp. 3821–3830.
- [11] M. Chen, D. Y. Fu, A. Narayan, M. Zhang, Z. Song, K. Fatahalian, and C. Ré, "Perfectly balanced: Improving transfer and robustness of supervised contrastive learning," in *International Conference on Machine Learning*, 2022, pp. 3090–3122.
- [12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [13] D. Muhammad, H. Ali, F. A. Azeemi, M. Khalid, N. Hussain, Z. Hussain, P. Hussain, S. Ali, M. Sangrasi, and M. Hammad, "Va-gkd: Variance-aware adaptive group knowledge distillation for long-tailed recognition," vol. 4, no. 3, pp. 2459–2471, 2026. [Online]. Available: <https://thesesjournal.com/index.php/1/article/view/3439>
- [14] D. Muhammad, L. Wang, and M. Hagenbuchner, "Enhancing robustness in long-tailed image classification with angular approaches and balanced learning," *Data Science and Engineering*, vol. 10, no. 3, pp. 480–494, 2025.
- [15] D. M. Sangrasi, L. Wang, and P. Hagenbuchner, Markusand Wang, "Mitigatingtheadverseeffectsoflong-taileddataondeeplearningmodels," in *Australasian Conference on Data Science and Machine Learning*. Springer, 2023, pp. 150–162.
- [16] P. Mettes, E. Van der Pol, and C. Snoek, "Hyperspherical prototype networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [17] Y. Yang, S. Chen, X. Li, L. Xie, Z. Lin, and D. Tao, "Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?" *Advances in Neural Information Pro-*

- cessing Systems, vol. 35, pp. 37 991–38 002, 2022.
- [18] T. Kasarla, G. Burghouts, M. Van Spengler, E. Van Der Pol, R. Cucchiara, and P. Mettes, "Maximum class separation as inductive bias in one matrix," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 553–19 566, 2022.
- [19] V. Papayan, X. Y. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24 652–24 663, 2020.
- [20] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.
- [21] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.
- [22] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha, "Benchmarking representation learning for natural world image collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 884–12 893.
- [23] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.
- [24] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [25] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18 661–18 673.
- [26] J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6908–6917.
- [27] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *International Conference on Learning Representations (ICLR)*, 2021.
- [28] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6918–6928.
- [29] D. Mildenerberger *et al.*, "A tale of two classes: Adapting supervised contrastive learning to binary imbalanced datasets," *arXiv preprint arXiv:2503.17024*, 2025.