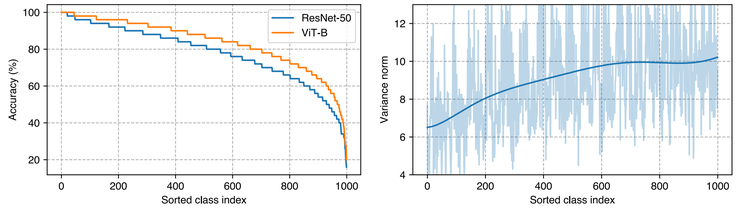
Reducing Class-Wise Performance Disparity via Margin Regularization

Beier Zhu[1] Kesen Zhao[2] Jiequan Cui[3] Qianru Sun[4] Yuan Zhou[2] Xun Yang[1] Hanwang Zhang[2]

[1]USTC [2]NTU [3]HFUT [4]SMU

Paper Code

1. Why this matters



(a) Per-class accuracy on ImageNet (b) Per-class variance norm on ImageNet

Figure 1. Classes are sorted in descending order of per-class accuracy.

- Training on balanced data does not eliminate class-wise accuracy disparity.
- Harder classes tend to exhibit larger feature variability.
- We reduce disparity via margin regularization in both *logit* and *representation* spaces.

2. Our Method: MR^2

Setup. We consider a classifier

$$f(\mathbf{x}, y) = \mathbf{w}_y^\top \phi(\mathbf{x}),$$

where $\phi : \mathcal{X} \to \mathbb{R}^d$ is the feature encoder and $\{\mathbf{w}_y\}_{y=1}^K$ are class-specific classifier weights.

Per-class statistics. For each class y, let $\mathcal{D}_y$ denote its training samples, with $N_y = N/K$. We compute the empirical class mean and feature spread as

$$\hat{\boldsymbol{\mu}}_y = \frac{1}{N_y} \sum_{\mathbf{x} \in \mathcal{D}_y} \phi(\mathbf{x}), \qquad \|\hat{\mathbf{s}}_y\|_2^2 = \frac{1}{N_y} \sum_{\mathbf{x} \in \mathcal{D}_y} \|\phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_y\|_2^2.$$

During training, we maintain exponential moving averages of $\{\|\hat{\boldsymbol{\mu}}_y\|_2^2\}_{y=1}^K$ and $\{\|\hat{\mathbf{s}}_y\|_2^2\}_{y=1}^K$.

Logit margin. Given the logit vector $\mathbf{z} = [f(\mathbf{x}, 1), \ldots, f(\mathbf{x}, K)]^\top$ for an input $\mathbf{x}$ with label y, we define the class-wise margin cross-entropy loss as

$$\ell_{\gamma,\mathtt{ce}}(f, \mathbf{x}, y) = -\mathbf{1}_y^\top \ln\left[\mathrm{softmax}(\mathbf{z}/\gamma_y)\right],$$

where

$$\gamma_y = \frac{\bar{c}\, K \left(\|\hat{\boldsymbol{\mu}}_y\|_2^2 + \|\hat{\mathbf{s}}_y\|_2^2\right)^{1/3}}{\sum_{k=1}^K \left(\|\hat{\boldsymbol{\mu}}_k\|_2^2 + \|\hat{\mathbf{s}}_k\|_2^2\right)^{1/3}}.$$

Here, $\bar{c} > 0$ controls the average margin. Intuitively, classes with larger feature variability receive larger margins, which improves their generalization and benefits hard classes.

2. Our Method: MR^2

Representation margin. Let $\bar{s} = \frac{1}{K}\sum_{k=1}^K \|\hat{\mathbf{s}}_k\|_2^2$ denote the average class-wise feature spread. We define the representation margin loss as

$$\ell_{\bar{s}}(f, \mathbf{x}, y) = \ln\left(1 + \sum_{\mathbf{x}^+ \in \mathcal{D}_y \setminus \{\mathbf{x}\}} \exp\left(\|\phi(\mathbf{x}) - \phi(\mathbf{x}^+)\|_2^2 - 2\bar{s}\right)\right).$$

This loss encourages intra-class compactness by penalizing overly large within-class pairwise distances, with $2\bar{s}$ serving as a soft margin.

Overall objective combines regularization in both logit and representation spaces:

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{(\mathbf{x},y) \in \mathcal{D}} \left[\ell_{\gamma,\mathtt{ce}}(f, \mathbf{x}, y) + \lambda\, \ell_{\bar{s}}(f, \mathbf{x}, y)\right].$$

3. Theoretical Analysis

Our key result shows that both the overall generalization risk and the per-class errors depend on the class-wise feature statistics and the assigned margins.

Theorem 1. For $f(\mathbf{x}, y) = \mathbf{w}_y^\top \phi(\mathbf{x})$ with $\|\mathbf{w}_y\|_2 \le \Lambda$, the expected risk satisfies

$$\mathcal{R}(f) \;\le\; \frac{1}{\ln 2} \widehat{\mathcal{R}}_{\mathcal{D}}^{\gamma,\mathtt{ce}}(f) + \frac{4\sqrt{2}\Lambda K}{\sqrt{N}} \sqrt{\sum_{k=1}^K \frac{\|\hat{\boldsymbol{\mu}}_k\|_2^2 + \|\hat{\mathbf{s}}_k\|_2^2}{\gamma_k^2}} + \mathcal{O}(1/\sqrt{N}).$$

Interpretation. The complexity term is *class-sensitive*: classes with larger feature variability $\|\hat{\boldsymbol{\mu}}_k\|_2^2 + \|\hat{\mathbf{s}}_k\|_2^2$ incur larger complexity and thus worse generalization unless compensated by a larger margin γ_k.

Corollary 1 (Optimal class-wise margin). Under a fixed average margin budget $\bar{c} = \frac{1}{K}\sum_k \gamma_k$, the complexity term is minimized by

$$\gamma_k = \frac{\bar{c} K \left(\|\hat{\boldsymbol{\mu}}_k\|_2^2 + \|\hat{\mathbf{s}}_k\|_2^2\right)^{1/3}}{\sum_{j=1}^K \left(\|\hat{\boldsymbol{\mu}}_j\|_2^2 + \|\hat{\mathbf{s}}_j\|_2^2\right)^{1/3}}.$$

Why does MR^2 work?

- **Logit-level regularization:** assigning $\gamma_k \propto (\|\hat{\boldsymbol{\mu}}_k\|_2^2 + \|\hat{\mathbf{s}}_k\|_2^2)^{1/3}$ gives larger margins to high-variability (harder) classes, which balances their error bounds.
- **Representation-level regularization:** reducing $\|\hat{\mathbf{s}}_k\|_2^2$ directly tightens the complexity term, improves generalization, and further reduces class-wise disparity.

5. Experiments

Table 1. **Comparison against existing methods** in reducing class-wise accuracy disparity on CIFAR-100 and ImageNet. ResNet-32 and CLIP ResNet-50 backbones are adopted respectively.

(a) CIFAR-100

Method	Overall	Easy	Medium	Hard
ERM	70.9	84.5	71.0	56.7
LfF	69.1	83.6 (-0.9)	70.1 (-0.9)	53.7 (-3.0)
JTT	70.6	84.3 (-0.2)	70.8 (-0.2)	56.2 (-0.5)
EQL	70.7	84.4 (-0.1)	70.9 (-0.1)	56.4 (-0.3)
LDAM	71.1	84.7 (+0.2)	71.2 (+0.2)	57.0 (+0.3)
LGM	70.8	84.0 (-0.5)	71.4 (+0.4)	56.6 (-0.1)
CMIC-DL	71.1	85.1 (+0.6)	70.9 (-0.1)	56.8 (+0.1)
SAM	71.0	84.6 (+0.1)	71.1 (+0.1)	56.9 (+0.2)
DFL	71.3	84.8 (+0.3)	71.3 (+0.3)	57.1 (+0.4)
SupCon	71.5	85.0 (+0.5)	72.1 (+1.1)	57.5 (+0.8)
DRL	71.9	85.5 (+1.0)	72.5 (+1.5)	57.7 (+1.0)
CSR	71.2	85.1 (+0.6)	71.3 (+0.3)	57.1 (+0.4)
DRO	71.6	85.1 (+0.6)	72.2 (+1.2)	57.2 (+0.5)
FairDRO	72.0	85.7 (+1.2)	72.5 (+1.5)	57.7 (+1.0)
MR^2 (ours)	**73.9**	**85.9** (+1.4)	**73.8** (+2.8)	**61.9** (+5.2)

(b) ImageNet

Method	Overall	Easy	Medium	Hard
ERM	75.2	91.1	78.3	56.4
LfF	74.4	90.2 (-0.9)	77.8 (-0.5)	55.2 (-1.2)
JTT	74.8	90.7 (-0.4)	77.9 (-0.4)	55.7 (-0.4)
EQL	75.3	91.3 (+0.2)	78.4 (+0.3)	56.2 (-0.2)
LDAM	75.4	91.5 (+0.4)	78.6 (+0.3)	56.1 (-0.3)
LGM	75.3	90.9 (-0.2)	78.3 (+0.0)	56.6 (+0.2)
CMIC-DL	75.2	91.0 (-0.1)	78.2 (-0.1)	56.5 (+0.1)
SAM	75.6	91.4 (+0.3)	78.8 (+0.5)	56.6 (+0.2)
DFL	75.8	91.7 (+0.6)	78.6 (+0.3)	57.1 (+0.7)
SupCon	75.7	91.4 (+0.3)	78.9 (+0.6)	56.7 (+0.3)
DRL	75.5	91.4 (+0.3)	78.5 (+0.2)	56.6 (+0.2)
CSR	75.1	90.8 (-0.3)	77.9 (-0.4)	56.6 (+0.2)
DRO	75.9	91.0 (-0.1)	79.3 (+1.0)	57.2 (+0.8)
FairDRO	76.1	91.5 (+0.4)	79.6 (+1.3)	57.3 (+0.9)
MR^2 (ours)	76.9	91.5 (+0.4)	79.7 (+1.4)	59.6 (+3.2)

Table 2. **Evaluation across diverse model architecture.** We perform **linear probing** on MoCov2 ResNet-50 using only $\ell_{\gamma,\mathtt{ce}}$. CLIP models are **cosine classifier** models. RN: ResNet. W-RN: WideResNet. TfS: Train from Scratch.

(a) CIFAR-100

Model	Overall	Easy	Medium	Hard
ResNet-20	68.7	83.3	70.3	53.0
+ MR^2	70.9	84.4 (+1.1)	72.2 (+1.9)	56.7 (+3.7)
PreAct RN-20	69.1	83.8	70.5	53.4
+ MR^2	71.3	84.7 (+0.9)	73.2 (+2.7)	56.6 (+3.2)
PreAct RN-32	71.2	85.0	71.7	56.7
+ MR^2	73.3	86.1 (+1.1)	73.9 (+2.2)	60.1 (+3.4)
W-RN-22-10	78.4	89.6	80.0	65.4
+ MR^2	81.2	90.7 (+1.1)	82.5 (+2.5)	70.2 (+4.8)

(b) ImageNet

Model	Overall	Easy	Medium	Hard
RN-50 TfS	71.7	88.5	74.1	52.6
+ MR^2	74.2	89.9 (+1.4)	76.9 (+2.8)	55.9 (+3.3)
MoCov2 RN-50	71.1	89	73.5	50.7
+ MR^2 ($\ell_{\gamma,\mathtt{ce}}$)	72.4	89.6 (+0.6)	74.8 (+1.3)	52.7 (+2.0)
CLIP ViT-B/32	75.6	91.4	78.4	57.0
+ MR^2	77.1	91.7 (+0.3)	79.9 (+1.5)	59.8 (+2.8)
MAE ViT-B/16	80.4	94.5	83.9	62.7
+ MR^2	82.0	94.8 (+0.3)	85.3 (+1.4)	66.1 (+3.4)

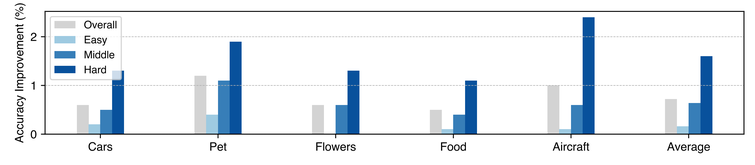


Figure 2. **Relative improvements on fine-grained datasets**: StanfordCars, OxfordPets, Flowers, Food and FGVCAircraft with CLIP ResNet-50.