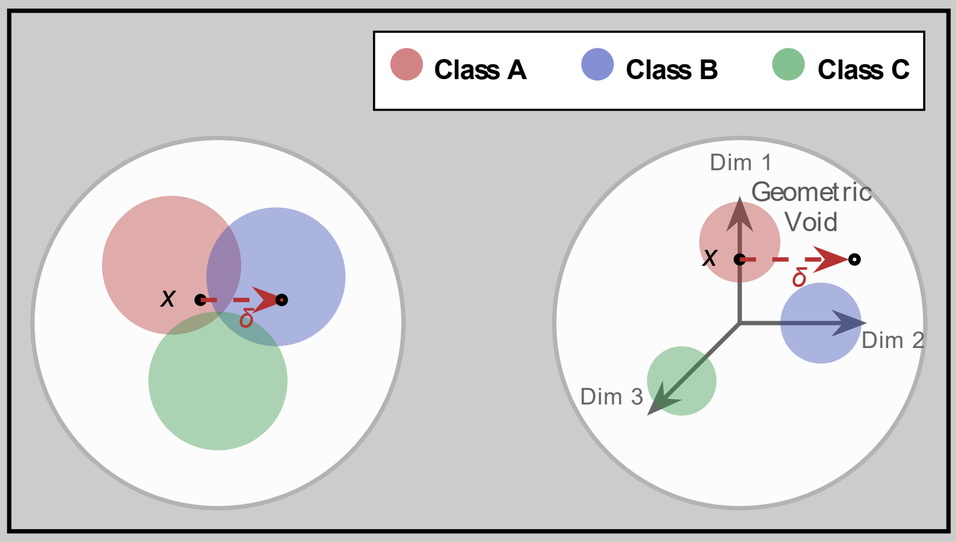


(a) Standard Supervision (b) CCAR (Ours)

Robustness isn't learned. It's just good geometry.

Class-Conditional Activation Regularization (CCAR): Intrinsic Robustness as an Emergent Geometric Property

Akash Samanta*[1], Manish Pratap Singh*[2] and Debasis Chaudhuri*[1]

1 TL;DR

MAIN 3 TAKEAWAYS

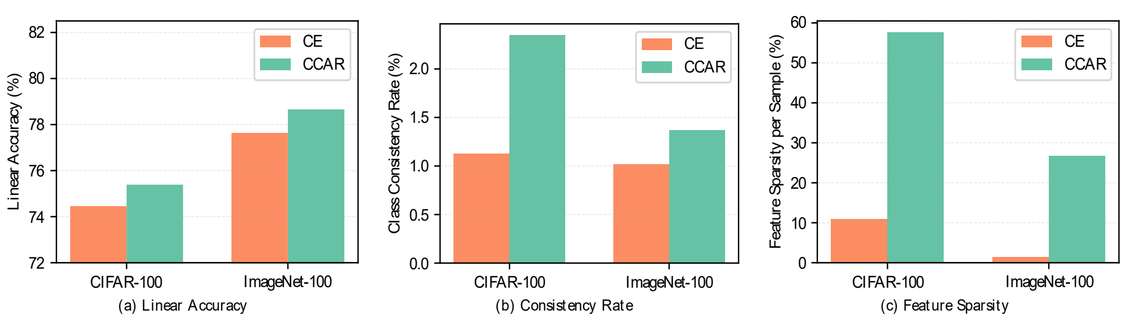
- Standard supervised training yields entangled, brittle features,
- CCAR applies a soft L_2bias to enforce a block-diagonal feature geometry.
- This intrinsic scaffold naturally filters noise, excelling even at 90% label corruption.

2 Background

- Standard training ignores latent geometry, leaving decision boundaries dangerously close to data manifolds.
- Existing defenses (augmentations, adversarial training) treat symptoms rather than structural vulnerabilities.
- Can we directly engineer internal representations to be intrinsically robust without complex pipelines?

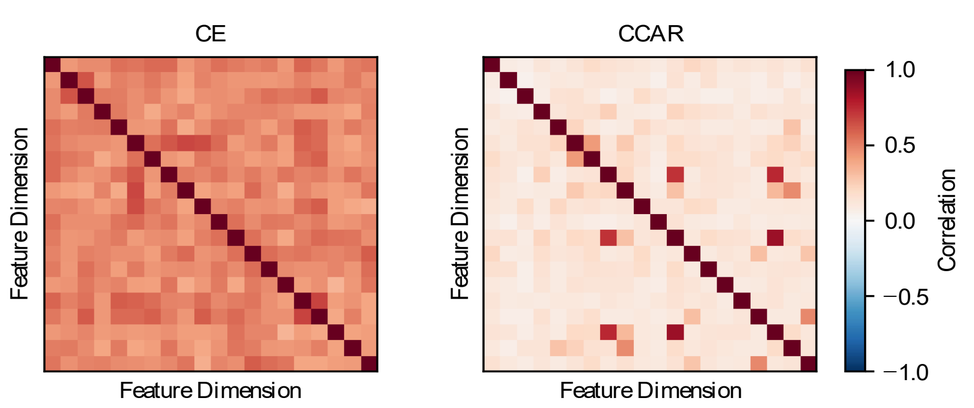
3 Geometric Regularization

- We explicitly partition the high-dimensional feature space, assigning every class its own dedicated, isolated block of active dimensions.
- A soft L_2 penalty ($\mathcal{L}_{\text{CCAR}} \propto \sum (j \in \mathcal{O}_y)\ h_j^2$) actively suppresses feature leakage by driving activations to zero in any dimension belonging to a competing class.
- By compressing intra-class variance in these orthogonal directions, we directly maximize the Fisher Discriminant Ratio.
- The network is forced to build tight, block-diagonal representations, naturally pushing class boundaries far apart.



4 Emergent Robustness & Results

- Orthogonal confinement creates a certified geometric margin against stochastic noise and FGSM attacks.
- Maintains discriminative power under extreme corrupted data.
- At 90% label noise, CCAR outperforms unconstrained baselines trained on entirely clean data.
- Breaks the typical robustness-accuracy compromise, significantly improving semantic feature retrieval.



Noise (%)	Cross-Entropy	Center Loss	Contrastive (CL)	NCL	**CCAR (Ours)**
CIFAR-100					
0	74.46 ± 0.24	75.39 ± 0.15	75.65 ± 0.31	74.21 ± 0.12	**75.69 ± 0.18**
10	72.90 ± 0.11	74.33 ± 0.29	75.26 ± 0.22	73.68 ± 0.09	**75.41 ± 0.25**
20	72.10 ± 0.35	74.19 ± 0.13	75.39 ± 0.16	73.28 ± 0.28	**75.26 ± 0.07**
30	71.52 ± 0.18	73.75 ± 0.32	75.23 ± 0.29	73.10 ± 0.15	**75.26 ± 0.33**
40	70.97 ± 0.27	74.15 ± 0.08	74.94 ± 0.24	73.17 ± 0.36	**75.25 ± 0.12**
50	70.11 ± 0.13	73.58 ± 0.25	74.84 ± 0.10	72.62 ± 0.21	**75.15 ± 0.29**
60	68.59 ± 0.42	73.23 ± 0.33	74.71 ± 0.19	72.61 ± 0.14	**75.11 ± 0.22**
70	66.95 ± 0.21	73.33 ± 0.17	73.94 ± 0.35	72.37 ± 0.26	**75.00 ± 0.11**
80	63.37 ± 0.38	72.21 ± 0.21	73.87 ± 0.13	71.26 ± 0.30	**74.67 ± 0.27**
90	52.74 ± 0.25	69.70 ± 0.41	72.34 ± 0.26	67.27 ± 0.19	**74.21 ± 0.34**
ImageNet-100					
0	77.64 ± 0.15	78.34 ± 0.22	79.84 ± 0.09	79.14 ± 0.31	**80.62 ± 0.14**
10	76.38 ± 0.29	77.66 ± 0.11	79.50 ± 0.25	78.88 ± 0.17	**80.76 ± 0.21**
20	76.22 ± 0.13	77.28 ± 0.34	79.52 ± 0.19	78.68 ± 0.22	**80.60 ± 0.10**
30	76.08 ± 0.36	77.12 ± 0.18	79.36 ± 0.27	79.02 ± 0.13	**80.54 ± 0.28**
40	76.20 ± 0.22	76.86 ± 0.29	79.34 ± 0.14	78.54 ± 0.35	**80.52 ± 0.08**
50	75.68 ± 0.19	77.16 ± 0.12	79.56 ± 0.30	78.56 ± 0.26	**80.42 ± 0.31**
60	75.42 ± 0.31	76.84 ± 0.24	79.54 ± 0.16	78.64 ± 0.09	**80.42 ± 0.23**
70	75.24 ± 0.14	76.92 ± 0.37	79.44 ± 0.21	78.80 ± 0.18	**80.14 ± 0.12**
80	73.66 ± 0.26	75.82 ± 0.15	79.28 ± 0.33	78.56 ± 0.24	**80.28 ± 0.36**
90	69.08 ± 0.39	74.08 ± 0.28	79.50 ± 0.12	78.26 ± 0.38	**79.98 ± 0.64**

5 Limitations and Discussion

- Soft inductive bias allows residual leakage; not a standalone defense against iterative attacks (e.g., PGD).
- Requires feature dimensions to equal or exceed the class count ($D \geq C$).

PAPER

CODE

* Equal contribution
[1] Techno India University, Salt Lake, Kolkata, India
[2] DRDO Young Scientist Laboratory –CT, Chennai, India

akashsamanta.web@gmail.com