Lose Your Self (LoYS): an adversarial entropy-based unsupervised approach for model debiasing

WACV 2026 (Tucson, AZ, USA)
March 6–10, 2026

Vito Paolo Pastore [1,2] Massimiliano Ciranni [1] Vittorio Murino [2,3]
vito.paolo.pastore@unige.it

[1] MaLGa – DIBRIS, University of Genoa, Italy
[2] AIGO, Istituto Italiano di Tecnologia, Italy
[3] University of Verona, Italy

Problem

- Deep classifiers can latch onto **spurious correlations** (“shortcuts”) in biased training data, **failing to generalize** to bias-conflicting test data.
- Debiasing methods aim to mitigate dependency on bias when making predictions;
- Bias-supervised methods assume **bias attributes are known**—often unrealistic in practice.

Goal

Debias **without** bias annotations or bias-annotated validation sets.

Our Idea

- Use model **confidence** as a signal: biased models tend to be *over-confident* on shortcut-driven predictions.

- Exploiting the confidence, train a debiasing model that **keeps task accuracy** while **discouraging bias-related features**

Lose Your Self (LoYS): adversarial, entropy-based bias-unsupervised debiasing.

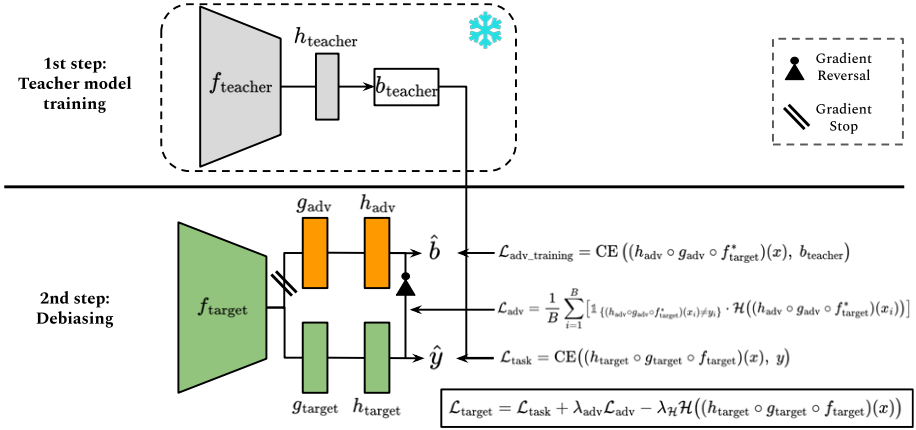
Contributions

1. **LoYS**: an unsupervised debiasing framework that does not require bias attributes nor bias-annotated validation sets.

2. An adversarial mechanism that **penalizes confidence** of a head trained to approximate a biased teacher.

3. Extensive evaluation on synthetic and single and multi-bias realistic biased benchmarks, reaching SOTA or competitive performance

LoYS at a Glance



Two-stage recipe: (1) train a softly biased teacher; (2) adversarially train target encoder to make bias prediction uncertain.

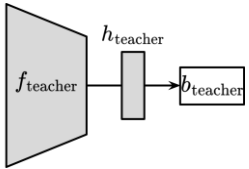
Step 1: Softly Biased Teacher

- Train f_{teacher} for a few epochs with **Generalized Cross-Entropy (GCE)** to capture shortcut/bias early.
- Use its outputs as pseudo bias labels b_{teacher}.

Why

A *noisy* bias estimate can still provide enough signal to push a *contrary* (debiasing) model.

1st step: Teacher model training



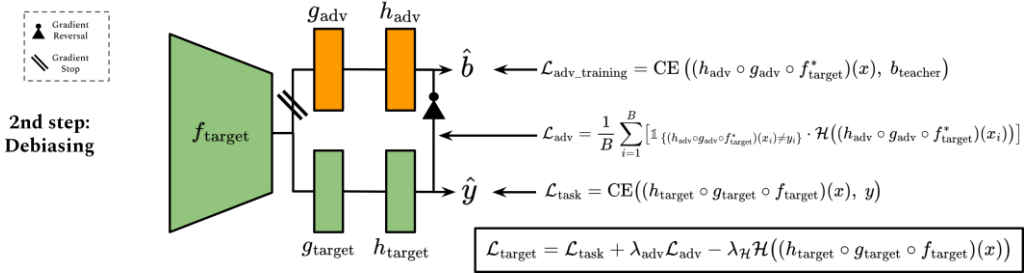
Teacher stage.

Step 2: Debiasing

- Add an adversarial head $(g_{\text{adv}}, h_{\text{adv}})$ on top of the target features.
- Freeze the encoder and train the adversary to match the teacher pseudo-labels:

Interpretation

The adversary learns “bias-predictive” margins *given* the current representation.



Debiasing stage.

Datasets: Visual Examples

- **Corrupted CIFAR-10**: synthetic additive corruption biases with varying
- **BFFHQ**: age classification with gender shortcut ($\rho = 0.995$).
- **BAR**: action recognition with context shortcuts ($\rho \in \{0.99, 0.95\}$).
- **ImageNet-9/A**: shape/texture shortcuts; robustness on ImageNet-A.
- **UrbanCars**: multiple biases (background + co-occurring object).

Results

Realistic benchmarks

Dataset	Metric	LoYS
BFFHQ (ρ = 0.995)	Conflicting Acc.	**68.50 (+ 0.17)**
BAR (ρ = 0.990)	Conflicting Acc.	**75.92 (+ 1.13)**
BAR (ρ = 0.995)	Conflicting Acc.	**85.00 (+ 0.04)**
ImageNet-9/A	Average Acc.	**41.92 (+ 5.95)**

(See full table in the paper.)

Multiple Bias (UrbanCars)

Method	ID.ACC ↑	BG-Gap ↓	CoObj-Gap ↓	BG-CoObj–Gap ↓
ERM [23]	97.30	15.30	11.20	69.20
LLE [23]	96.70	2.10	2.70	5.90
LfF [30]	97.20	11.60	18.40	63.20
JTT [26]	95.90	8.10	13.30	40.10
EIIL [7]	95.50	**4.20**	24.70	44.90
DebiAN [22]	**98.00**	14.90	10.50	69.00
LoYS (ours)	97.46 ± 0.28	8.08 ± 0.90	$\mathbf{9.86 \pm 3.34}$	$\mathbf{39.98 \pm 2.86}$

Syntethic benchmark (Corrupted CIFAR-10)

Method	BS	Corrupted CIFAR-10 $\rho = 0.995$	$\rho = 0.990$	$\rho = 0.980$	$\rho = 0.950$
ERM+CE	–	21.43 ± 0.57	25.82 ± 0.33	30.37 ± 1.37	39.88 ± 0.70
EnD [40]	✓	22.89 ± 0.27	25.46 ± 0.41	31.31 ± 0.35	40.26 ± 0.85
ReBias [2]	✓	22.27 ± 0.41	25.72 ± 0.20	31.66 ± 0.43	43.43 ± 0.41
BiaSwap [16]	✗	$29.11 \pm --$	$32.54 \pm --$	$35.25 \pm --$	$41.62 \pm --$
LfF [31]	✗	28.81 ± 0.44	33.07 ± 0.77	40.66 ± 0.70	50.72 ± 1.31
DFA [19]	✗	29.95 ± 0.71	36.49 ± 1.79	41.78 ± 2.29	51.13 ± 1.28
LC [27]	✗	34.56 ± 0.69	37.34 ± 1.26	47.81 ± 2.00	54.55 ± 1.26
MoDAD [32]	✗	27.26 ± 0.47	–	41.27 ± 0.26	50.48 ± 0.42
DeNetDM [41]	✗	$\mathbf{38.93 \pm 1.16}$	$\mathbf{44.20 \pm 0.77}$	47.35 ± 0.70	56.30 ± 0.42
LoYS Ours	✗	28.94 ± 1.01	38.91 ± 0.37	$\mathbf{48.55 \pm 0.07}$	$\mathbf{63.53 \pm 0.55}$

Conclusion

- **LoYS** debiases models **without** bias annotations via adversarial **entropy** regularization.
- A softly biased teacher provides enough signal to learn more robust, less shortcut-dependent features.
- Works well on realistic biased datasets and shows robustness benefits.

TL;DR

LoYS is an Unsupervised Debiasing methods that reduces shortcut reliance by **making bias prediction uncertain** in an adversarial scheme, **while learning stronger features for the task** at the same time.

Lose Your Self (LoYS): an adversarial entropy-based unsupervised approach for model debiasing

Vito Paolo Pastore [1, 2] Massimiliano Ciranni [1] Vittorio Murino [2, 3]

vito.paolo.pastore@unige.it

[1] MaLGa – DIBRIS, University of Genoa, Italy
[2] AIGO, Istituto Italiano di Tecnologia, Italy
[3] University of Verona, Italy

WACV 2026 — Tucson, AZ, USA — March 6–10, 2026