CVPR
SEATTLE, WA JUNE 17-21, 2024

复旦大学
FUDAN UNIVERSITY

Highlight

Navigate Beyond Shortcuts: Debiased Learning through the Lens of Neural Collapse

Yining Wang, Junjie Sun, Chenyue Wang, Mi Zhang†, Min Yang

† Corresponding Author

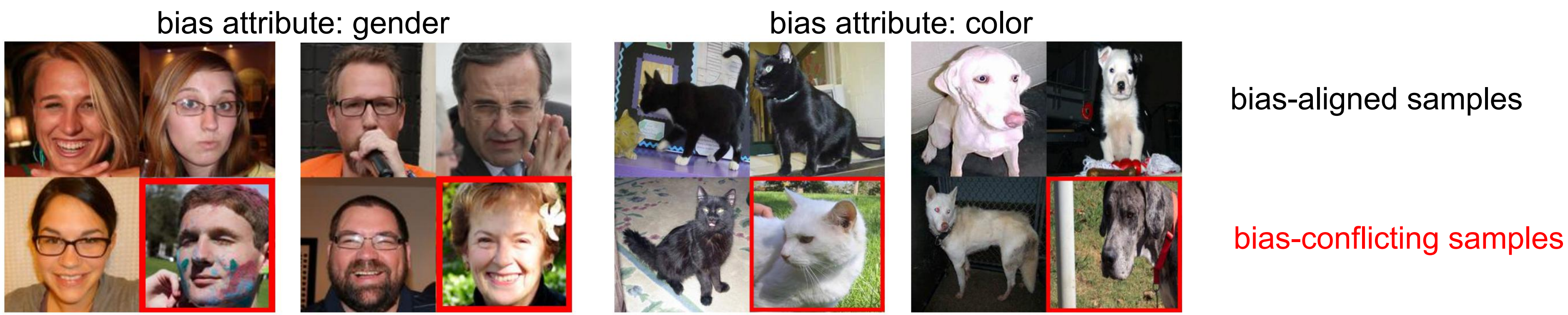
Whitzard-AI Group

School of Computer Science, Fudan University

Poster ID: THU-AM-12065

Debiased Learning

- ***Biased datasets*****:** each class is dominated by some *bias* attributes (e.g., background, texture...)
  - majority of *bias-aligned* samples + minority of *bias-conflicting* samples
- ***Learning shortcuts***: the shortcut correlation between ground-truth labels & bias attributes
  - low generalizability on OOD test samples
- ***Existing debiased learning***: reweight-based, augmentation-based, disentangle-based
  - limited by extra, heavy training expenses

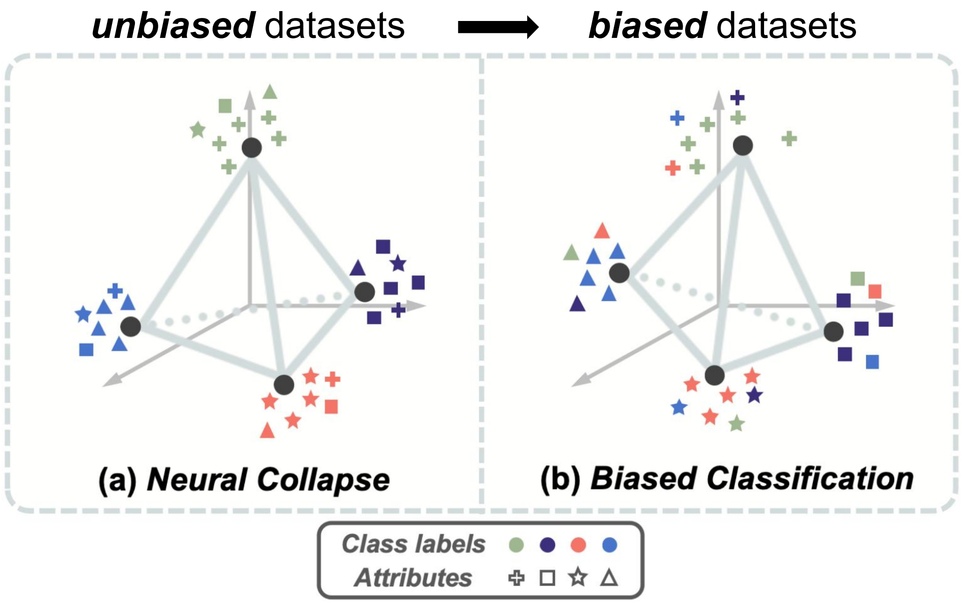


(a) BFFHQ: *young* (b) BFFHQ: *old* (c) Dogs & Cats: *cat* (d) Dogs & Cats: *dog*

e.g., bias-aligned samples and bias-conflicting samples

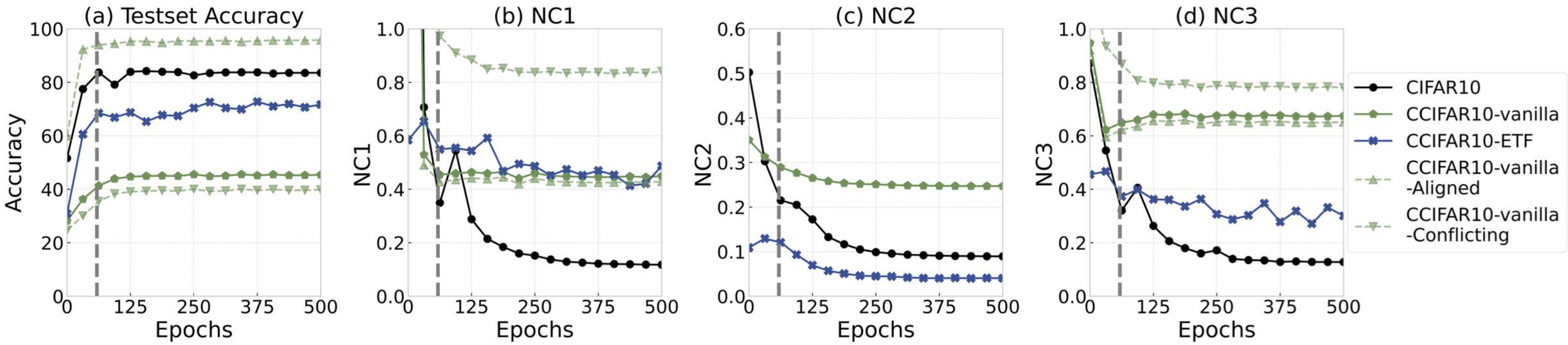
Neural Collapse phenomenon

- ***Neural Collapse phenomenon***
  - discovered by Papyan et al. [PNAS 2020]
  - on ***unbiased*** datasets: last-layer feature space ***converges*** to elegant geometry (simplex ETF)
  - optimal generalization, robustness, and interpretability



Neural Collapse on Biased Datasets

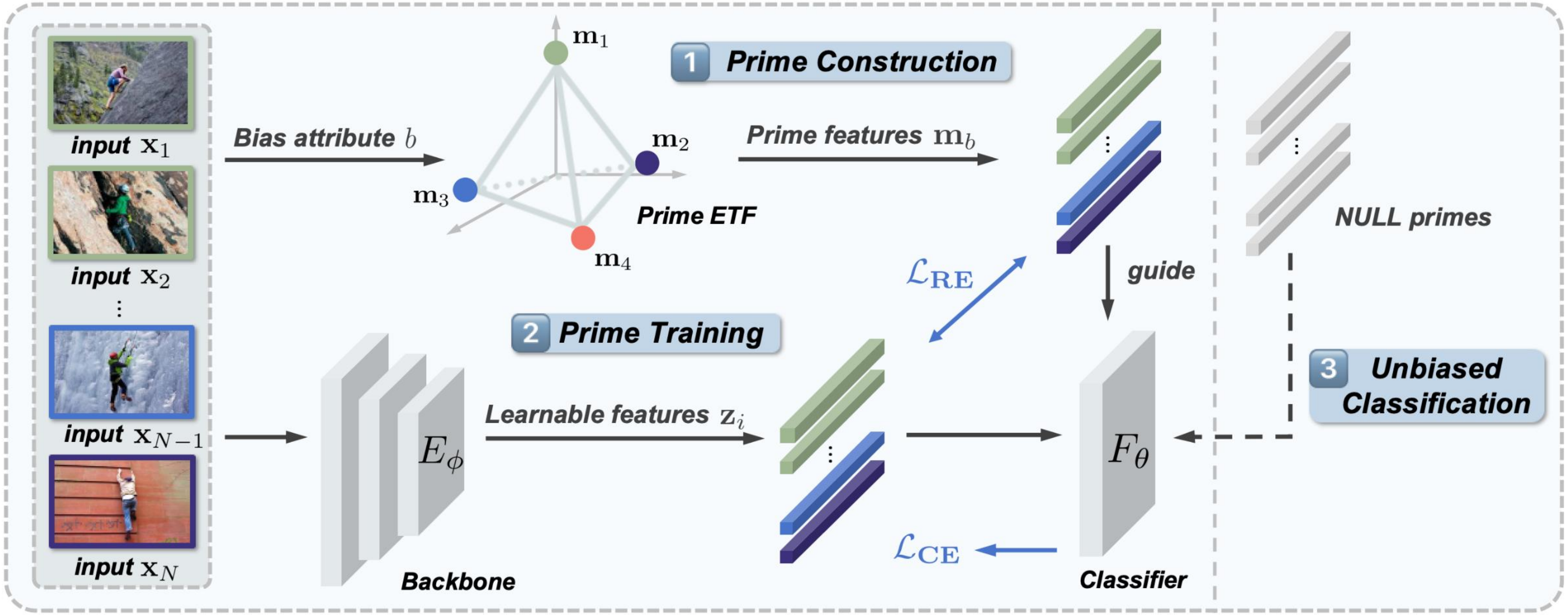
- ***Two Stages of Training***
  - *shortcut learning*: shortcut correlation, converge towards ETF
  - *intrinsic learning*: intrinsic correlation, fail to collapse
  - ➡ non-collapsed, sub-optimal feature space



Comparison of (a) testset accuracy and (b-d) Neural Collapse metrics on unbiased and synthetic biased datasets

Our Method

- ***Avoid-shortcut learning:* ETF-Debias**
  - *ETF as priming*: approximate the "*perfect*" shortcut features
  - skip the active learning of shortcuts, directly persue intrinsic correlations



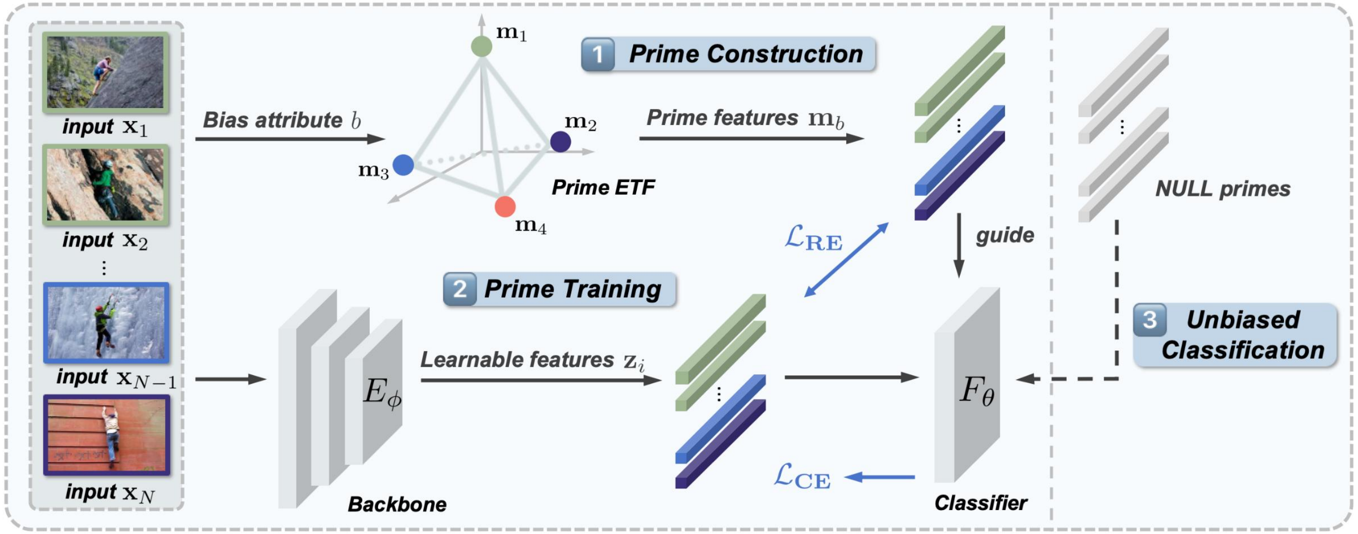
Framework of ETF-Debias

Our Method

- ***Avoid-shortcut learning:* ETF-Debias**
  - *Prime Construction*
  - *Prime Training*: classification objective + prime reinforcement regularization

$$\min_{\phi,\theta} \mathcal{L}_{\mathrm{CE}}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{N} \mathcal{L}(F_\theta(\mathbf{z}_{i,b},\mathbf{m}_b),\mathbf{y}_{i,b}) \qquad \mathcal{L}_{\mathrm{RE}}(\mathbf{x},\mathbf{b}) = \textstyle\sum_{i=1}^{N} \mathcal{L}(F_\theta(\mathbf{z}_{i,b},\mathbf{m}_b) - F_\theta(\mathbf{z}_{i,b},\mathbf{m}_{\mathrm{null}}),\mathbf{b})$$

  - *Unbiased Classification*: rely on intrinsic correlations $\hat{\mathbf{y}} = F_\theta(\mathbf{z}_{i,b},\mathbf{m}_{\mathrm{null}})$



Theoretical Justification

➢ ***Implicit re-weighting of gradients***

- Follow the analysis of Neural Collapse from the perspective of *gradients*
- The pulling & forcing part of gradients are implicitly re-weighted by prime features

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \widetilde{\mathbf{w}}_k} &= \sum_{i=1}^{n_k} -(1-p_k(\widetilde{\mathbf{z}}_{k,i}))\widetilde{\mathbf{z}}_{k,i} + \sum_{k'\neq k}^{K}\sum_{j=1}^{n_{k'}} p_k(\widetilde{\mathbf{z}}_{k',j})\widetilde{\mathbf{z}}_{k',j} \\ &\leq \underbrace{\sum_{i=1}^{n_k} -(1-p_k^{(b)}(\mathbf{m}_{i,b}) - p_k^{(l)}(\mathbf{z}_{k,i}))\widetilde{\mathbf{z}}_{k,i}}_{\text{pulling part}} \\ &\quad + \underbrace{\sum_{k'\neq k}^{K}\sum_{j=1}^{n_{k'}} (p_k^{(b)}(\mathbf{m}_{j,b'}) + p_k^{(l)}(\mathbf{z}_{k',j}))\widetilde{\mathbf{z}}_{k',j}}_{\text{forcing part}}\end{aligned}$$

Gradient *w.r.t* classifier weights

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \widetilde{\mathbf{z}}_{k,i}} &= -(1-p_k(\widetilde{\mathbf{z}}_{k,i}))\mathbf{w}_k + \sum_{k'\neq k}^{K} p_{k'}(\widetilde{\mathbf{z}}_{k,i})\mathbf{w}_{k'} \\ &\leq \underbrace{-(1-p_k^{(b)}(\mathbf{m}_{i,b}) - p_k^{(l)}(\mathbf{z}_{k,i}))\mathbf{w}_k}_{\text{pulling part}} \\ &\quad + \underbrace{\sum_{k'\neq k}^{K} (p_{k'}^{(b)}(\mathbf{m}_{i,b}) + p_{k'}^{(l)}(\mathbf{z}_{k,i}))\mathbf{w}_{k'}}_{\text{forcing part}}\end{aligned}$$

Gradient *w.r.t* features

Experimental Settings

- ***Dataset***
  - 2 synthetic biased datasets: Colored MNIST, Corrupted CIFAR10
  - 3 real-world biased datasets: Biased FFHQ (BFFHQ), BAR, and Dogs & Cats
- ***Baseline***
  - 6 debiasing baselines
  - Reweight-based: LfF (NeurIPS 2020), LfF+BE (AAAI 2023)
  - Disentangle-based: EnD (CVPR 2021), SD (MM 2023)
  - Augmentation-based: DisEnt (NeurIPS 2021), Selecmix (NeurIPS 2022)

Experimental Results

Debiasing Performance

Dataset	Ratio(%)	Vanilla	LfF◇[23]	LfF+BE◇[19]	EnD*[30]	SD*[42]	DisEnt*[18]	Selecmix◇[13]	**ETF-Debias**
Colored MNIST	0.5	32.22±0.13	57.78±0.81	69.69±1.99	35.93±0.40	56.96±0.37	68.83±1.62	70.53±0.46	**71.63**±0.28 (+1.10)
	1.0	48.45±0.06	72.29±1.69	80.90±1.40	49.32±0.58	72.46±0.18	79.49±1.44	**83.34**±0.37	81.97±0.26 (-1.37)
	2.0	58.90±0.12	79.51±1.82	84.90±1.14	65.58±0.46	79.37±0.46	84.56±1.19	85.90±0.23	**86.00**±0.03 (+0.10)
	5.0	74.19±0.04	83.96±1.44	90.28±0.18	80.70±0.17	88.89±0.21	88.83±0.15	91.27±0.31	**91.36**±0.21 (+0.09)
Corrupted CIFAR-10	0.5	17.06±0.12	31.00±2.67	23.68±0.50	14.30±0.10	36.66±0.74	30.12±1.60	33.30±0.26	**40.06**±0.03 (+3.40)
	1.0	21.48±0.55	34.33±1.76	30.72±0.12	20.17±0.19	45.66±1.05	35.28±1.39	38.72±0.27	**47.52**±0.26 (+1.86)
	2.0	27.15±0.46	39.68±1.15	42.22±0.60	30.10±0.54	50.11±0.69	40.34±1.41	47.09±0.17	**54.64**±0.42 (+4.53)
	5.0	39.46±0.58	53.04±0.76	57.93±0.58	45.85±0.21	62.43±0.57	49.99±0.84	54.69±0.29	**65.34**±0.60 (+2.91)

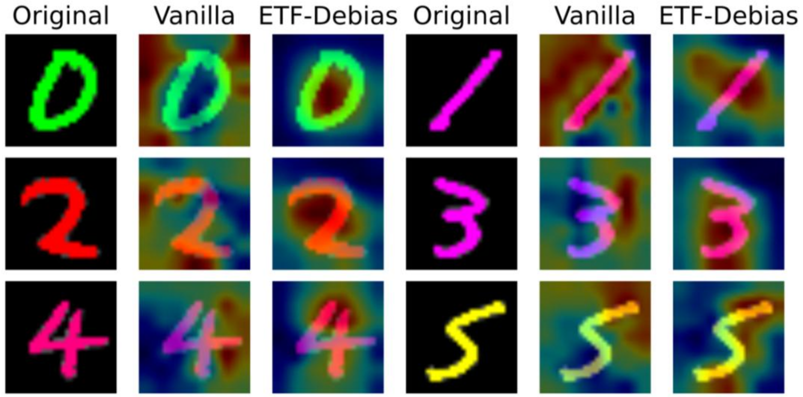
Test set accuracy on 2 synthetic biased datasets

Dataset	Ratio(%)	Vanilla	LfF◇[23]	LfF+BE◇[19]	EnD*[30]	SD*[42]	DisEnt*[18]	Selecmix◇[13]	**ETF-Debias**
Biased FFHQ	0.5	53.27±0.61	65.60±2.27	67.07±2.37	55.93±1.62	65.60±0.20	63.07±1.14	65.00±0.82	**73.60**±1.22 (+6.53)
	1.0	57.13±0.64	72.33±2.19	73.53±1.62	61.13±0.50	69.20±0.20	68.53±2.32	67.50±0.30	**76.53**±1.10 (+3.00)
	2.0	67.67±0.81	74.80±2.03	80.20±2.78	66.87±0.64	78.40±0.20	72.00±2.51	69.80±0.87	**85.20**±0.61 (+5.00)
	5.0	78.87±0.83	80.27±2.02	87.40±2.00	80.87±0.42	84.80±0.20	80.60±0.53	83.47±0.61	**94.00**±0.72 (+6.60)
Dogs & Cats	1.0	51.96±0.90	71.17±5.24	78.87±2.40	51.91±0.24	78.13±1.06	65.13±2.07	54.19±1.61	**80.07**±0.90 (+1.20)
	5.0	76.59±1.27	85.83±1.62	88.60±1.21	79.07±0.28	89.12±0.18	82.47±2.86	81.50±1.06	**92.18**±0.62 (+3.06)
BAR	1.0	68.00±0.43	68.30±0.97	71.70±1.33	68.25±0.19	67.33±0.35	69.30±1.27	69.83±1.02	**72.79**±0.21 (+1.09)
	5.0	79.34±0.19	80.25±1.27	82.00±1.24	78.86±0.36	79.10±0.42	81.19±0.70	78.79±0.52	**83.66**±0.21 (+1.66)

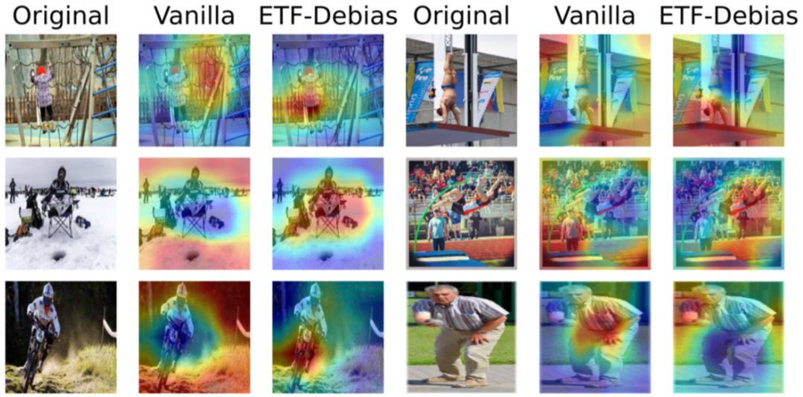
Test set accuracy on 3 real-world biased datasets

Experimental Results

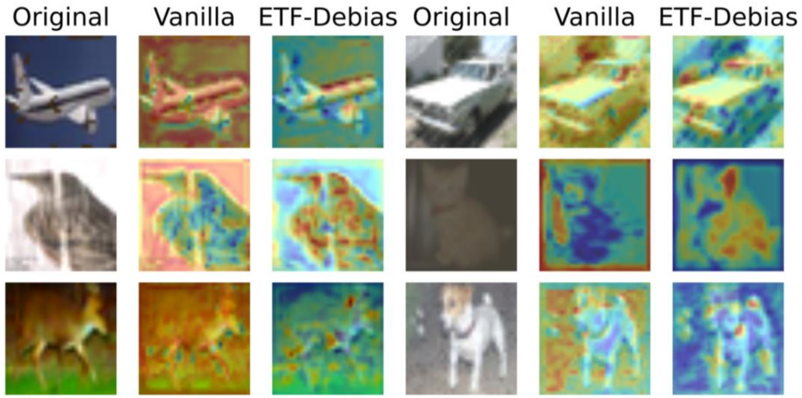
Visualization Comparison



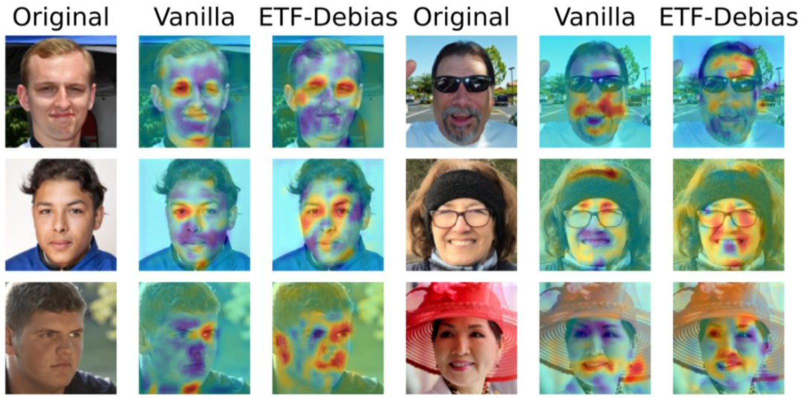
(a) Colored MNIST
bias:color



(b) BAR
bias:place



(c) Corrupted CIFAR-10
bias:corruption



(d) Biased FFHQ
bias:gender

Conclusion

- ***Theoretical Investigation***
  - For the first time, we investigate the Neural Collapse phenomenon on biased datasets
  - Analyze the fundamental issues of biased classification

- ***Avoid-shortcut Learning***
  - Our proposed ETF-Debias achieves SOTA debiasing performance
  - *No* additional training expenses
  - Theoretical & experimental supports

CVPR
SEATTLE, WA JUNE 17-21, 2024

復旦大學 FUDAN UNIVERSITY

Highlight

Thank you for listening!

Navigate Beyond Shortcuts:
Debiased Learning through the Lens of Neural Collapse

If you have any questions, please contact us.
Website of Whitzard-AI Group: https://whitzard-ai.github.io/