ClusterFix: A Cluster-Based Debiasing Approach without Protected-Group Supervision

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Giacomo Capitani, Federico Bolelli, Angelo Porrello, Simone Calderara, Elisa Ficarra

University of Modena and Reggio Emilia, Italy

name.surname@unimore.it

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

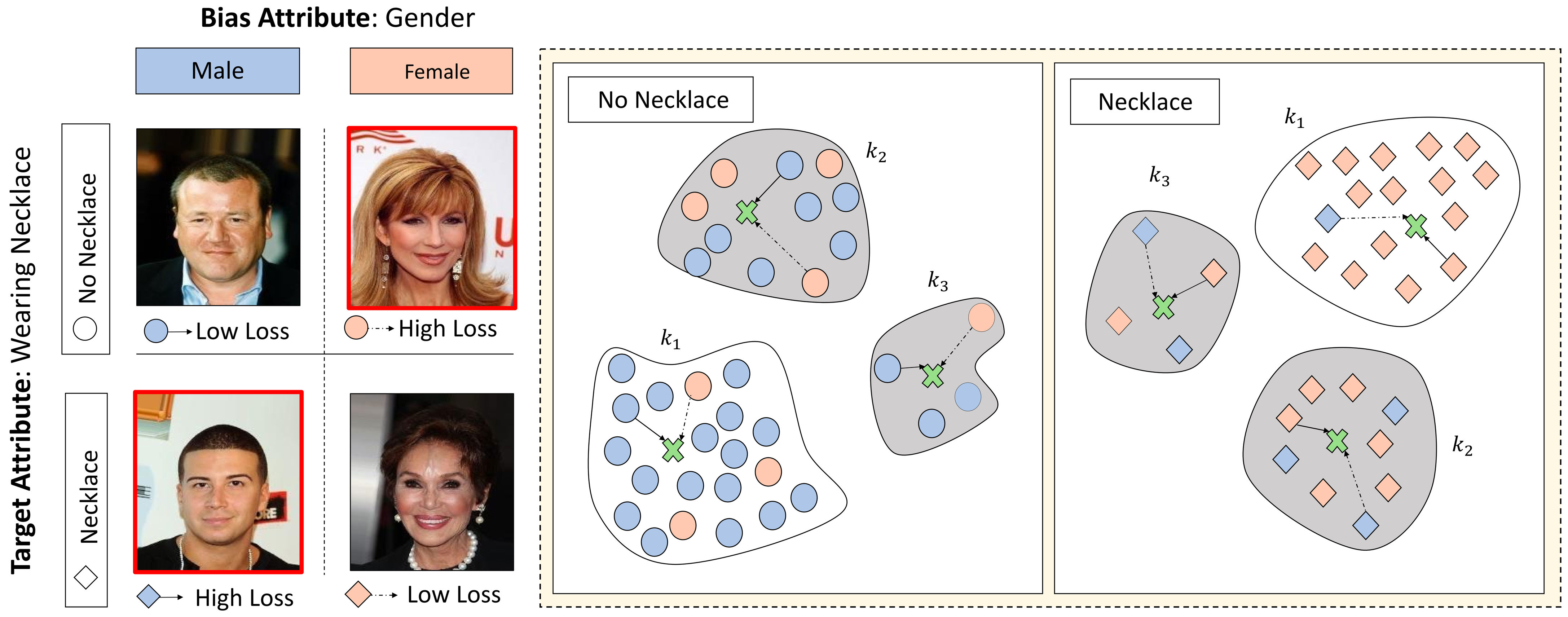
Abstract

- ✗ DNN failures arise from data or algorithm biases [**1**].
- ✗ Current debiasing methods [**2**] **need prior knowledge**, impractical in real scenarios.
- ✓ We suggest a novel debiasing method **without external bias information**.
- ✓ This method extend ERM with per-example weights from the majority class.
- ✓ Weights are based on the model's ability to infer correct pseudo-labels via clustering.

Identifying Bias: Clustering Loss as a Proxy

Main Idea: Samples belonging to clusters with high cluster classification loss can aid in identifying samples prone to spurious correlations, without protected group knowledge.

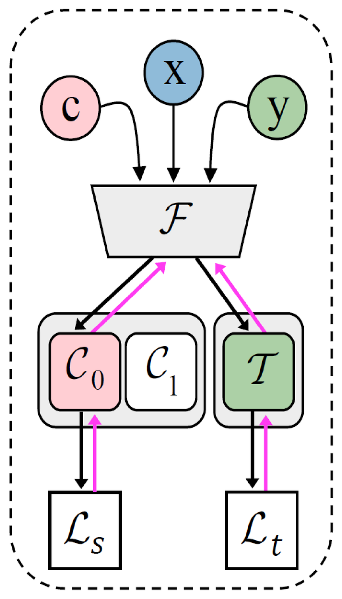
Our Contribution: Our method introduces a reweighting-based approach designed to achieve generalization across groups without requiring their direct supervision.



Method

Cluster-Structural Loss

- Smooth feature space via cluster classification avoids shortcuts in target classification.
- Useful for spotting problematic clusters with high loss and small size.



$$\mathcal{L}_s = \sum_{i=1}^{N} \ell(\mathcal{C} \circ \mathcal{F}(x_i), c_i)$$

Task-weighted Classification Loss

- Sample classification loss weighted by factor w, based on cluster significance.

$$\mathcal{L}_t = w_{c_i} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i)$$

$$w_c = \frac{1}{N_c} \mathbb{E}_{(x,y)\sim P_c}[\ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) + \gamma \ell(\mathcal{C}_{y_i} \circ \mathcal{F}(x_i), c_i)]$$

Experiments and Results

DATASETS: in our experiments with CelebA, the gender attribute has been used as the bias attribute to evaluate the robustness of the proposed method, as in [**3**,**4**].

RESULTS: we show how **ClusterFix** outperforms all competitors across all evaluated scenarios, we report results in terms of **average-group** accuracy and **worst-group** accuracy. For complete results and datasets, check the paper!

AVERAGE ACCURACY

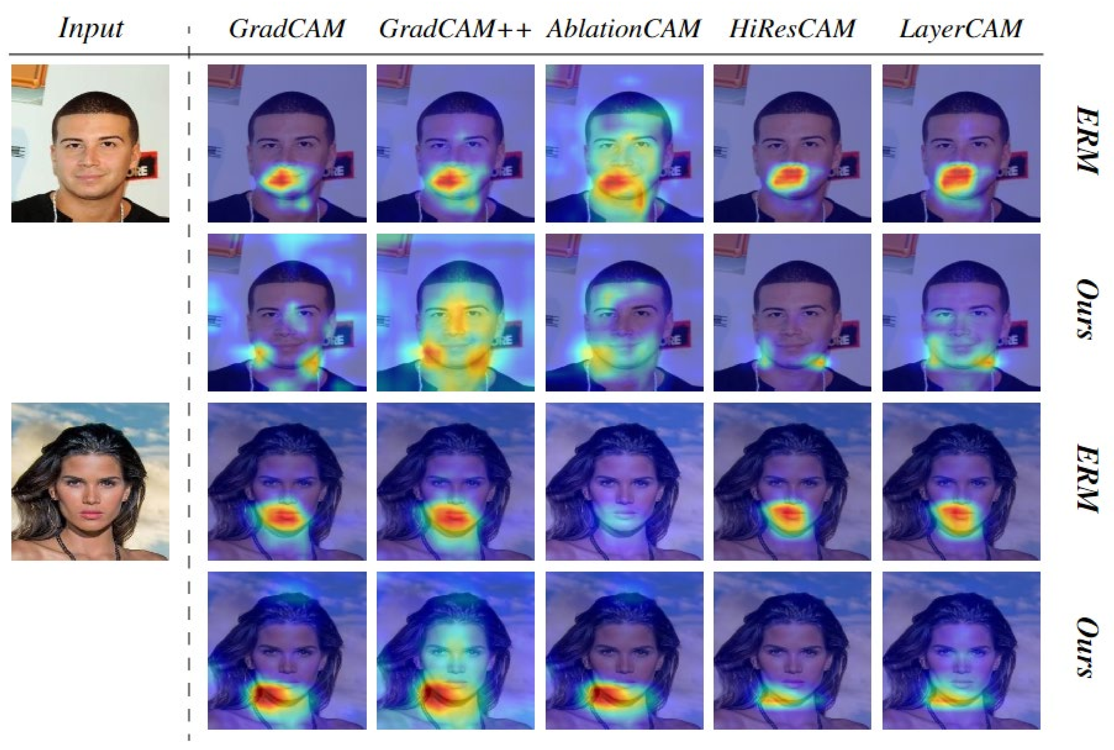
Target	ERM	LfF	George	BPA	Ours	GDRO
Double Chin	64.61	68.47	76.23	82.92	85.13	83.19
Pale Skin	71.5	75.23	78.22	90.06	91.17	90.55
Chubby	67.42	71.56	74.88	83.88	84.16	81.9
Wearing Necklace	55.04	57.21	58.79	68.96	68.99	62.89
Wearing Hat	93.53	94.81	95.72	96.8	97.88	96.84
Big Lips	60.87	62.15	64.99	66.5	65.4	63.7
Bangs	89.04	89.04	92.62	93.94	94.67	94.45
Receding Hairline	69.72	74.58	78.86	84.95	87.0	85.15
Wavy Hair	73.1	74.53	77.39	79.89	79.42	79.65
Brown Hair	78.07	78.93	83.07	83.83	85.3	84.87

WORST GROUP ACCURACY

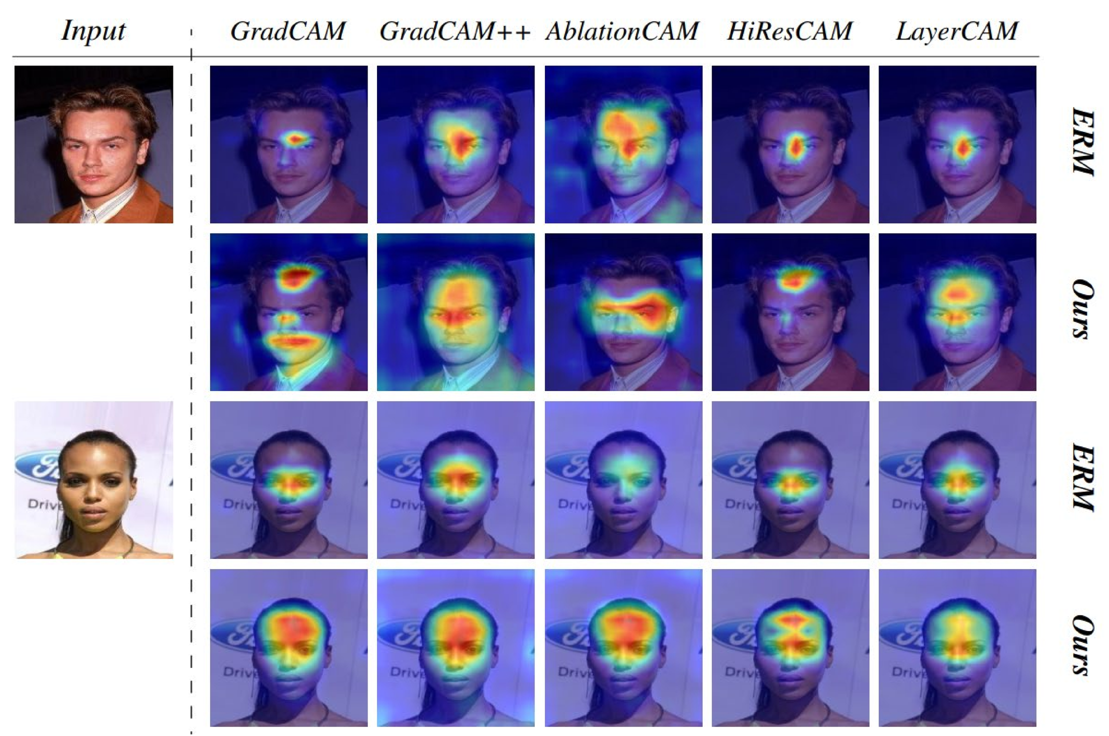
Target	ERM	LfF	George	BPA	Ours	GDRO
Double Chin	21.33	28.24	50.0	67.78	74.26	72.94
Pale Skin	36.64	43.26	62.03	88.6	87.01	87.68
Chubby	24.3	34.09	58.01	72.32	71.01	72.64
Wearing Necklace	2.72	6.67	13.82	41.93	55.56	24.34
Wearing Hat	85.12	88.31	92.93	94.94	96.58	94.67
Big Lips	30.85	38.54	44.51	56.99	57.27	47.55
Bangs	76.91	82.37	85.9	92.21	93.01	92.12
Receding Hairline	35.69	45.53	57.3	79.11	84.15	79.12
Wavy Hair	38.01	45.24	53.17	65.74	69.92	66.79
Brown Hair	59.58	60.68	73.2	71.5	79.18	78.92

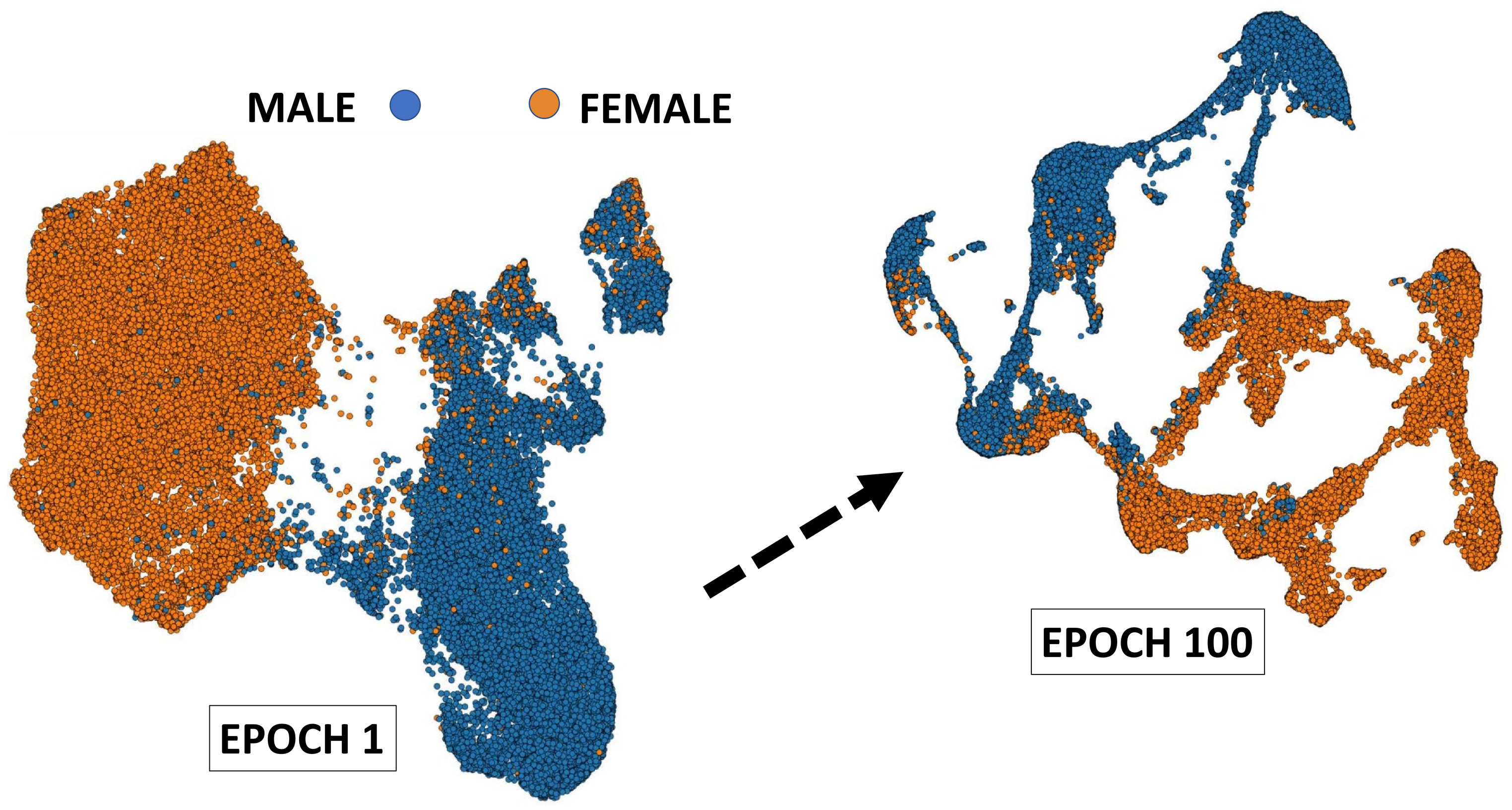
Visualizations

WEARING NECKLACE



RECEDING HAIRLINE





References

[1] Geirhos, Robert, et al., Shortcut learning in deep neural networks, in Nature Machine Intelligence, 2020

[2] Sagawa, Shiori, et al., Distributionally Robust Neural Networks, in ICLR, 2020

[3] Seo ,Seonguk, et al., Unsupervised Learning of Debiased Representations with Pseudo-Attributes, in CVPR, 2022

[4] Sohoni, Nimit, et al., No Subclass Left Behind, in NIPS, 2020