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

Giacomo Capitani Federico Bolelli Angelo Porrello
Simone Calderara Elisa Ficarra

Università degli Studi di Modena e Reggio Emilia, Italy
{*name.surname*}@unimore.it

1. Supplementary Material

1.1. Additional Experiments

Shortcuts Surface in the Early Epochs. In Fig. 1, we present a comprehensive analysis of accuracy trends across different demographic groups for the *Wearing Necklace* target (Fig. 1c and Fig. 1d represent male individuals, while Fig. 1b and Fig. 1d represent individuals wearing necklaces). Our findings indicate that CFix effectively mitigates spurious correlations from the early epochs and demonstrates a consistent improvement in accuracy. In contrast, the ERM classifier capitalizes on these spurious correlations and shows negligible improvement over time. We further extend this analysis by providing a visual comparison between the biased classifier and our CFix approach in Fig. 3, Fig. 4, Fig. 5, and Fig. 6. These figures reveal that the issue of lower accuracy is not confined to underrepresented groups but is often a byproduct of the learned model itself.

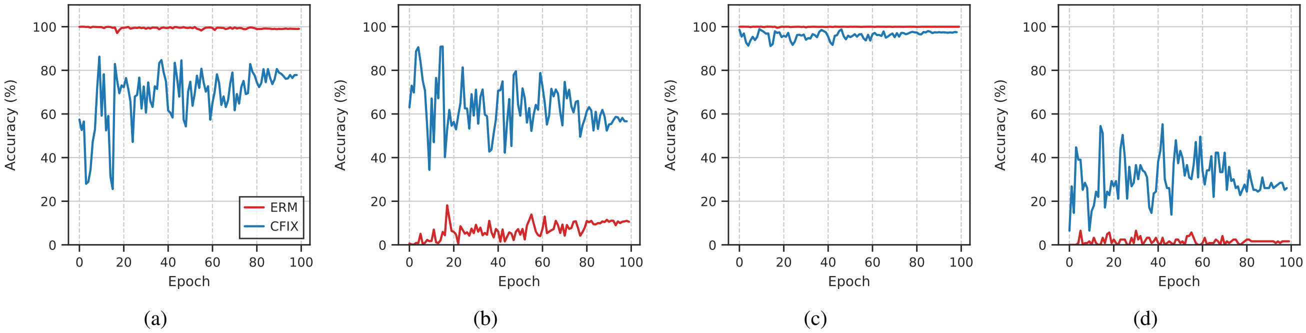
On the Number of Clusters. We conducted an ablation study using the CelebA dataset to scrutinize the sensitivity of ClusterFix to the number of clusters used for partitioning the feature space. The objective was to ascertain how ClusterFix's performance varies with the number of clusters. Fig. 2a displays the results, which indicate that the average accuracy across groups plateaus when $K \geq 8$. This suggests that increasing the number of clusters beyond this point does not yield significant improvements in performance. However, we observed a decline in the accuracy of the worst-performing group as the number of clusters increased. This decline can be attributed to the inherent class imbalance in the dataset, particularly in the positive partition (*Wearing Necklace* = True), which is less populated than the negative partition (*Wearing Necklace* = False). Increasing the number of clusters in the positive partition is likely to create smaller, noisier clusters that include outliers, making the model more sensitive to these outliers and thereby reducing the accuracy for the worst-performing group. To mitigate this issue, we propose determining the optimal number of clusters for each partition, as suggested by George [7].

Table 1. Unbiased accuracy (%) on Colored MNIST dataset.

Target	**Bias**	**ERM**	**LfF** [4]	**BPA** [6]	**CFix**	**GDRO** [5]
Digit	Color	74.48	85.15	85.26	**87.07**	85.88

Sensitivity Analysis of the γ Hyperparameter. We performed a sensitivity analysis focusing on the γ hyperparameter in the context of *Wearing Necklace* classification. Fig. 2b elucidates the impact of varying γ values, as defined in the main paper. We specifically examined the disparity in accuracy between the worst and average performing groups. Our results indicate that the accuracy of the worst-performing group remains relatively stable across different γ values, particularly on real-world datasets like CelebA. Additionally, Tab. 2 presents an ablation study that explores the role of γ in the context of ERM pre-training.

Entropy Correlation. To further validate the insights discussed in the introduction of the paper, which form the foundation of our re-weighting strategy, we present a rigorous empirical proof that establishes a correlation between the membership of samples in the *critical* group, denoted as z, and the output entropy of the cluster classifiers $\mathcal{C}_{1,\dots,Y}$ as defined in Eq. (2). Specifically, the output of the cluster classifier $\mathcal{C}_y$ for each sample (x, y) is determined by the softmax function applied to the composition of $\mathcal{C}_y$ and the feature extractor $\mathcal{F}(x)$, which is then divided by the temperature parameter t as shown in Eq. (1).



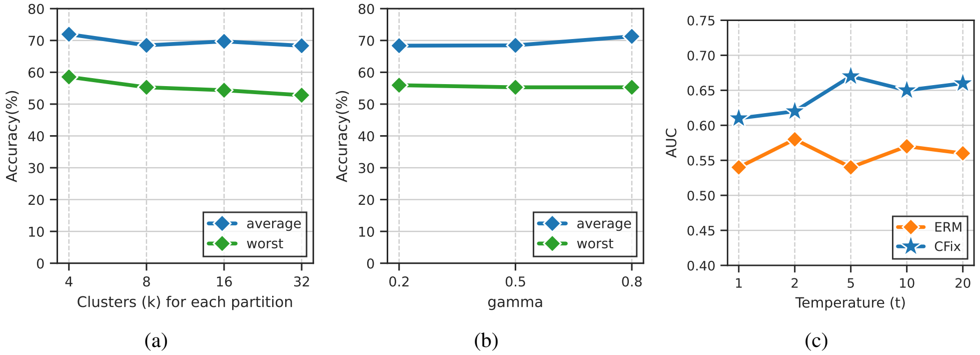
(a)

(b)

(c)

(d)

Figure 1. Temporal accuracy variations on the test set using ERM and CFix, observed for each group.



(a)

(b)

(c)

Figure 2. Ablation study on number of clusters (a) and gamma sensitivity (b); Entropy and critical samples correlation (c).

$$\mathcal{U} = \text{softmax}(\mathcal{C}_y \circ \mathcal{F}(x)/t) \tag{1}$$

$$H(x) = -\sum_{i=1}^{n} \mathcal{U} \log_2 \mathcal{U} \tag{2}$$

To quantify the correlation between the output entropy $H(x)$ and the membership in the critical group z, we compute the Area Under the Curve (AUC) for the relationship between $H(x)$ and z. Our experiments on the Waterbirds dataset reveal that this correlation is not only present in models trained using Empirical Risk Minimization but is notably pronounced in models trained using the ClusterFix approach. This correlation tends to strengthen with higher values of the temperature parameter t, as shown in Fig. 2c. These findings suggest that the cluster classifier's uncertainty is also correlated with the critical samples, thereby reinforcing the efficacy of cluster-based reweighting strategies.

Colored MNIST with ERM Pre-Training. In addition to our primary experiments, we also conducted tests on the Colored MNIST dataset, following the methodologies outlined in [4] and [6]. To ensure a fair comparison and to emphasize the role of cluster classification, we employed ERM pre-training for clustering, akin to other methods listed in the table. Consequently, the only modification introduced was in the reweighting procedure, which was handled by our ClusterFix algorithm.

Comparison with Other Attributes and Methods in CelebA. In Tab. 3 and Tab. 4, we present a comprehensive comparison focusing on various attributes such as "blond hair," "heavy makeup," and "gender," which have also been utilized by other methods for evaluation. Our results clearly demonstrate that ClusterFix outperforms all other methods across these attributes, thereby establishing its superiority in mitigating biases.

On the Backbone Choice on Waterbirds. Tab. 5 illustrates a comparison between using a ResNet50 architecture versus a ResNet18 architecture on the Waterbirds dataset. Our findings indicate that the choice of backbone architecture does not significantly impact the performance, thereby validating the robustness of our ClusterFix algorithm.

Additional Examples of Explainability on CelebA. Fig. 3, Fig. 4, Fig. 5, and Fig. 6 showcase additional examples of regions of interest identified by both the biased ERM and our debiased ClusterFix methods. These exam-

Table 2. Unbiased accuracy (%) on the Colored MNIST dataset varying with the parameter γ.

Target	**Bias**	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 5$	$\gamma = 10$	$\gamma = 20$	$\gamma = 50$
Digit	Color	77.42	80.17	82.14	83.44	85.37	86.05	**87.07**

Table 3. Unbiased accuracy (%) on additional CelebA attributes.

Target	**Bias**	**ERM**	**PGI** [1]	**EIIL** [2]	**DebiAN [3]**	**BPA** [6]	**CFix**
Blond Hair	Gender	79.80 ± 0.30	82.00 ± 1.10	81.60 ± 0.30	84.00 ± 0.14	90.18 ± 0.23	**91.27 ± 0.06**
Gender	Heavy Makeup	85.10 ± 0.01	85.40 ± 3.40	84.00 ± 1.20	87.80 ± 1.30	-	**90.18 ± 0.70**
Gender	Wearing Leapstick	86.60 ± 0.40	86.90 ± 3.10	86.30 ± 1.00	**88.50 ± 1.11**	-	88.06 ± 0.12
Heavy Makeup	Gender	71.90 ± 0.37	-	-	-	73.78 ± 0.25	**76.65 ± 0.68**

Table 4. Worst accuracy (%) on additional CelebA attributes.

Target	**Bias**	**ERM**	**PGI** [1]	**EIIL** [2]	**DebiAN [3]**	**BPA** [6]	**CFix**
Blond Hair	Gender	37.90 ± 1.10	46.10 ± 4.90	40.90 ± 6.40	52.90 ± 4.70	82.54 ± 1.22	**85.51 ± 0.62**
Gender	Heavy Makeup	45.40 ± 0.01	46.90 ± 13.10	40.90 ± 4.50	56.00 ± 5.20	-	**68.18 ± 0.70**
Gender	Wearing Leapstick	53.90 ± 1.20	56.00 ± 11.70	52.40 ± 3.20	61.70 ± 4.20	-	**65.96 ± 1.10**
Heavy Makeup	Gender	17.35 ± 4.60	-	-	-	39.84 ± 2.28	**47.22 ± 1.99**

Table 5. Results on the Waterbirds dataset using ResNet18 (CFix) and ResNet50 (CFix*).

		Unbiased Accuracy (%)						**Worst-Group Accuracy (%)**					
		Unsupervised					**Sup.**	**Unsupervised**					**Sup.**
Target	**Bias**	**ERM**	**LfF**	**BPA**	**CFix**	**CFix***	**GDRO**	**ERM**	**LfF**	**BPA**	**CFix**	**CFix***	**GDRO**
Object	Place	84.63	84.57	87.05	86.29	**87.36**	88.99	62.39	61.68	71.39	**74.03**	72.27	80.82

ples further validate the efficacy of ClusterFix in identifying meaningful features while mitigating biases.

1.2. Details on the Experimental Setting

Experimental Setup. For our experiments, we followed a rigorous evaluation protocol. We selected the best-performing model based on both metrics from the validation set and used it as a checkpoint for evaluating the test set, in line with the methodology described in [6]. The detailed ClusterFix learning procedure is provided in Algorithm 1.

On the Choice of CelebA Targets. In our experiments, we followed the guidelines set by [6] and focused on gender as the fixed bias attribute. We excluded 8 out of the 40 attributes due to the limited number of samples in the test set. Among the remaining 32 attributes, 26 exhibited a significant correlation with gender, as evidenced by a classification accuracy gap exceeding 5% when compared to unbiased accuracy [5]. To ensure a comprehensive evaluation, we selected the top 5 attributes with the highest accuracy gap and the bottom 5 attributes with the lowest gap, as identified in [6].

References

[1] Faruk Ahmed, Yoshua Bengio, Harm Van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2020.

[2] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

[3] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022.

[4] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: Training Debiased Classifier from Biased Classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Algorithm 1 ClusterFix Procedure

```
Require: learning rate $\eta_\theta$, momentum $m$, training steps $T$, batch size $B$, number of clusters $K$, balance loss factor $\lambda$, k-means cluster assignment $\mathcal{A}$, pre-trained feature extractor $\mathcal{F}$, whole network $\theta = \{\mathcal{F}, \mathcal{T}, \mathcal{C}_{1,..Y}\}$.

Step 1: Cluster Assignment
for $k = 1, ..., K$ do
    $P_k = \{\mathcal{A}(\mathcal{F}(x_i)) = k\}$
    $N_k = |P_k|$

Step 2: Debiased Training
for $t = 1, ..., T$ do
    for $i = 1, ..., B$ do
        Sample $(x_i, y_i) \sim P$
        $\alpha_i \leftarrow \omega_k$
        $\alpha \leftarrow (\alpha_1, ..., \alpha_B)$
        $\alpha \leftarrow \frac{\alpha}{\sum_{i=1}^{B} \alpha_i}$
        $\theta \leftarrow \theta - \eta_\theta \frac{1}{B} \sum_{i=1}^{B} \alpha_i \nabla \mathcal{L}_t + \lambda \nabla \mathcal{L}_s$
        for $k = 1, ..., K$ do
            $\omega_k \leftarrow (1 - m)\omega_k + \frac{m}{N_k} \sum_{(x,y) \in P_k} \mathcal{L}_t + \lambda \mathcal{L}_s$
```

[5] Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2020.

[6] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16742–16751, 2022.

[7] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

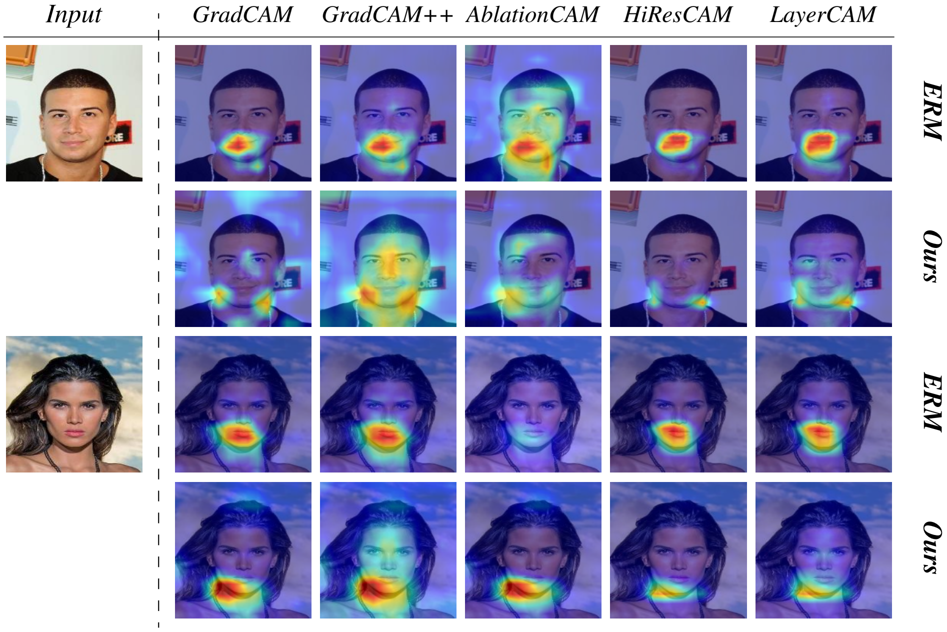


Figure 3. Visualizing activation maps for *Wearing Neacklace* target.

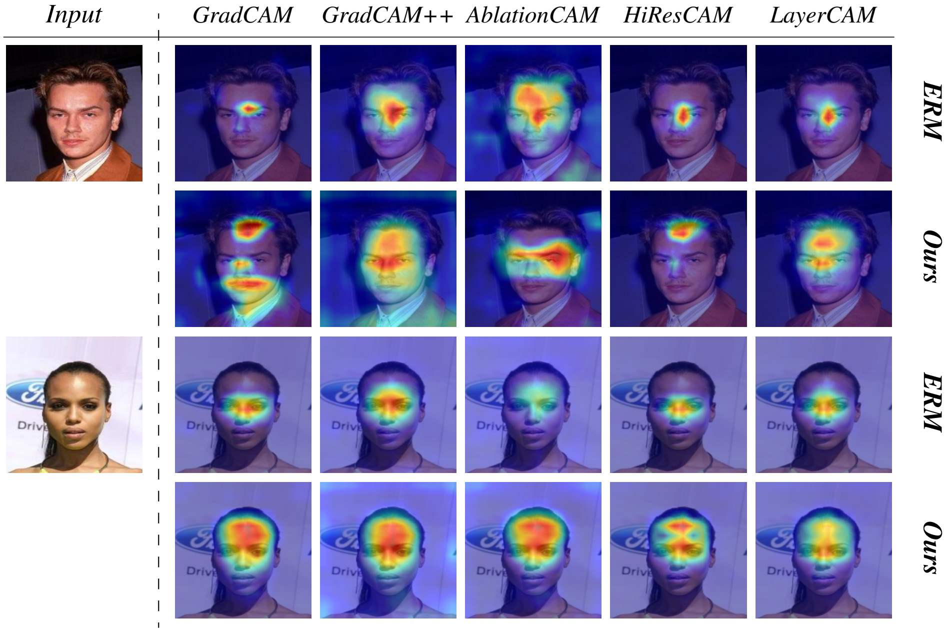


Figure 4. Visualizing activation maps for *Receding Hairline* target.

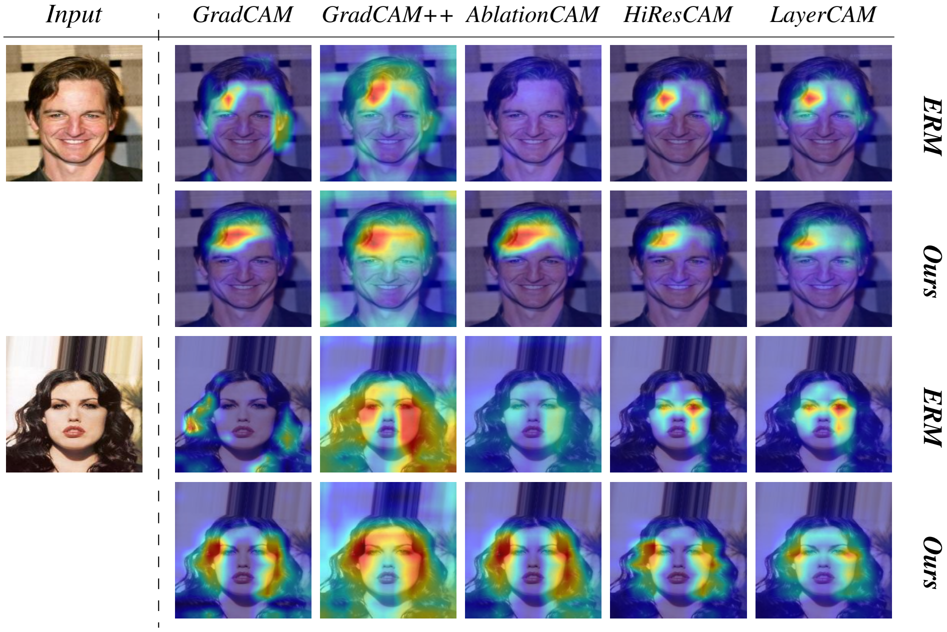


Figure 5. Visualizing activation maps for *Wavy Hair* target.

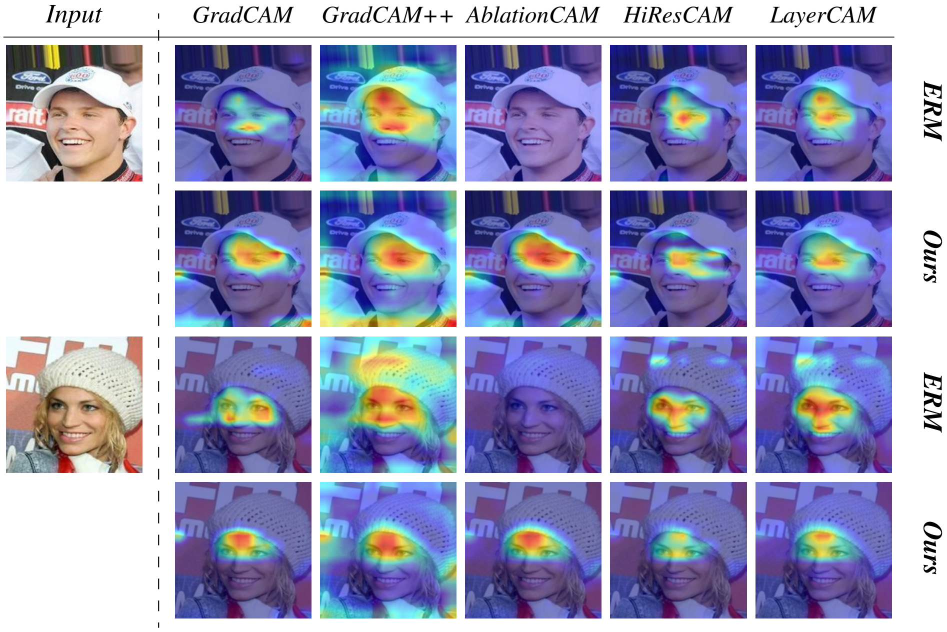


Figure 6. Visualizing activation maps for *Wearing Hat* target.