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

Vito Paolo Pastore[1,2], Massimiliano Ciranni[1], Vittorio Murino[2,3]
[1]MaLGa – DIBRIS, University of Genoa, Italy [2]AIGO, Istituto Italiano di Tecnologia, Italy [3]University of Verona, Italy

A. Ablation studies on pre-training for the softly biased teacher model

Using GCE loss for training f_{teacher} When training the softly biased teacher model, we employ a GCE loss function, pre-training the model for 5 epochs, across all the considered datasets.

We use a GCE loss function as it provides a weaker learning signal from *harder* samples (i.e. bias-conflicting samples in this context), reducing the risk of memorizing informative samples from the training set, thus being helpful for the training of an intentionally biased model, as shown in [1, 3–5]. The benefit of employing a GCE loss function is confirmed by a dedicated ablation study on BAR ($\rho = 0.99$) in Table 1.

# Loss function	GCE	CE
Accuracy (%)	$\mathbf{75.92} \pm \mathbf{0.93}$	72.58 ± 0.93

Table 1. Ablation study on our choice of using GCE loss in place of Cross-Entropy, for pre-training the softly biased teacher model f_{teacher}. The ablation study is performed on BAR ($\rho = 0.99$).

Number of pre-training epochs for f_{teacher} Regarding the number of epochs for training the teacher model, we would like to specify that we select 5 epochs for all the employed benchmarks, as we found this choice to be reasonable across all our experiments. Other works employing even just 1 epoch, actually tune this value for each examined dataset [2], exploiting a bias-annotated validation set. Here, we provide an ablation study on this matter, showing how our performance varies when the softly-biased teacher model f_{teacher} is trained from 1 epoch, up to 15. Table 2 shows the obtained results. As it can be noticed, too few epochs do not allow for reaching the best performance, while the choice of 5 epochs provides the best trade-off between computational overhead and accuracy, also with lower standard deviations.

# Pre-Training Epochs	15	10	**5**	3	1
Accuracy (%)	74.40 ± 2.03	74.58 ± 1.42	$\mathbf{75.92} \pm \mathbf{0.93}$	72.99 ± 0.90	69.36 ± 2.56

Table 2. Ablation study on the number of pre-training epochs for the teacher model f_{teacher} on BAR ($\rho = 0.99$).

B. Computational time

LoYS's adversarial debiasing scheme requires the joint training of the target classifier head and a bias adversarial head. However, for the latter we employ 5 fully connected layers, requiring only an extra effort of computation, when compared to the ERM regular training.

Considering BFFHQ as an example, when using a batch size of 64 on an RTX4060, a single epoch of standard ERM training takes roughly 18 seconds of compute time. LoYS's debiasing scheme, in comparison, requires roughly 23 seconds. This little overhead is negligible if considering the advantages in terms of generalization brought by our method.

References

[1] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021. 1

[2] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 1

[3] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020. 1

[4] Vito Paolo Pastore, Massimiliano Ciranni, Davide Marinelli, Francesca Odone, and Vittorio Murino. Looking at model debiasing through the lens of anomaly detection. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 2548–2557, 2025.

[5] 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. 1