Diffusing DeBias (DDB): Synthetic Bias Amplification for Model Debiasing

Massimiliano Ciranni* 1, Vito Paolo Pastore* 1,2, Roberto Di Via* 1, Enzo Tartaglione 3, Francesca Odone 1, Vittorio Murino 2,4

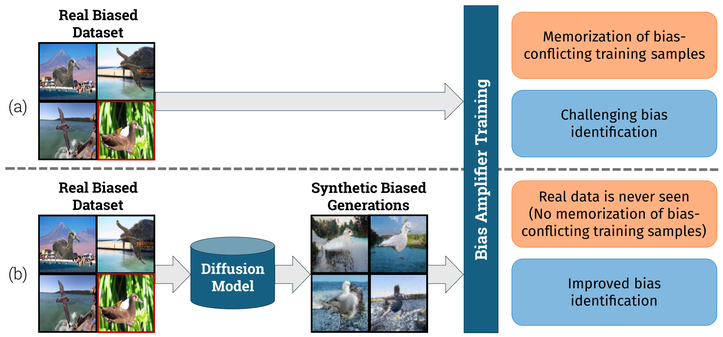
[1]MaLGa–DIBRIS, University of Genoa, Italy · [2] AI For Good (AIGO), Istituto Italiano di Tecnologia, Genoa, Italy

*Equal Contribution [3]Télécom Paris, École Polytechnique, France · [4]Department of Computer Science, University of Verona, Italy

vito.paolo.pastore@unige.it;

Bias-Conflicting Memorization Issue

The few bias-conflicting samples are quickly memorized from the auxiliary models often used in Unsupervised Debiasing methods [3].



Bias Amplification Without Memorization

What if the auxiliary model never sees the real biased dataset, but only a synthetic and bias-amplified version of it?

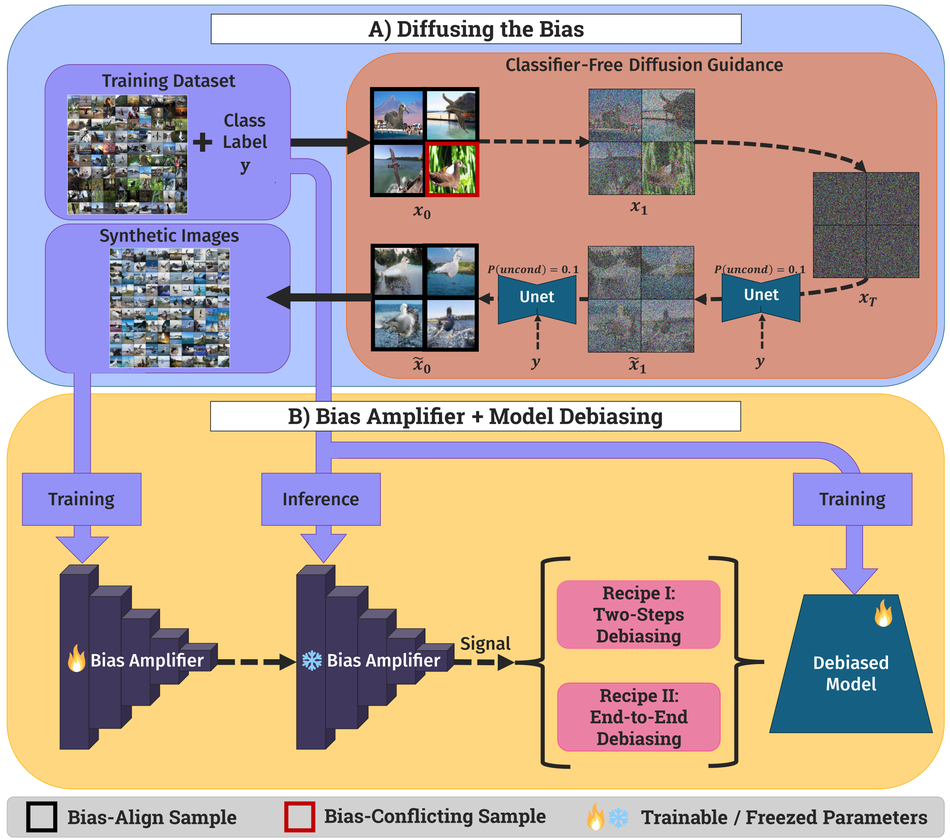
Observation
Conditional Diffusion Probabilistic Models trained on biased data inherently **learn and even amplify** the per-class bias [1,2].

Intuition
Train a highly effective auxiliary model *only* on a synthetic dataset generated by a bias-capturing diffusion model (our **BiasAmplifier, BA).**

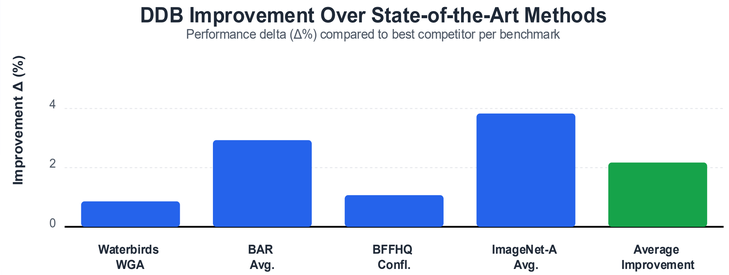
Key Advantage
BA learns bias from a synthetic substitute set, never seeing real samples**.** As such it cannot memorize any bias-conflicting sample, by construction.

Diffusing DeBias: A Plug-in for Unsupervised Debiasing

DDB is a plug-in for unsupervised model debiasing. Hence, it can be integrated as prior step for different debiasing strategies (e.g., [4], [5]).



Results

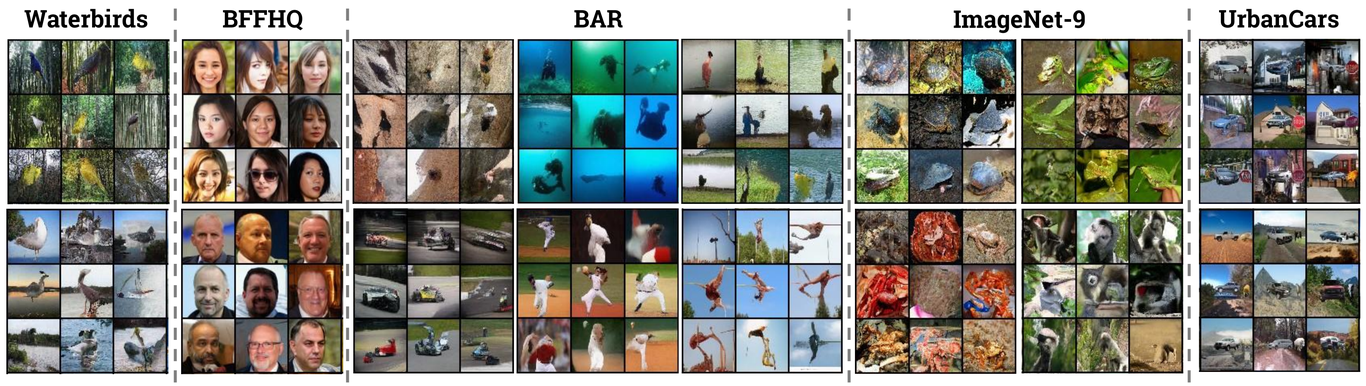


Method	ID.ACC ↑	BG-Gap ↓	CoObj-Gap ↓	BG-CoObj-Gap ↓
GroupDRO [43]	91.60	10.90	3.60	16.40
LLE [32]	96.70	2.10	2.70	5.90
LfF [39]	97.20	11.60	18.40	63.20
JTT [35]	95.90	8.10	13.30	40.10
EIIL [7]	95.50	4.20	24.70	44.90
DebiAN [33]	98.00	14.90	10.50	69.00
DDB-I (ours)	86.39± 0.74	**1.85**± 3.21	**0.52**± 1.38	**0.12**± 1.56
DDB-II (ours)	**98.56**± 0.92	2.30± 0.60	11.10± 1.20	46.70± 2.42

Multiple bias case (UrbanCars).

Debiasing Recipe	BA trained on Synthetic (DDB)	BA trained on Real
Recipe I (BA + G-DRO)	**90.81%**	79.43% (1 epoch)
Recipe II (LfF-style)	**91.56%**	78.45%

Ablation Study on using synthetic biased images for training our BA.



Examples of the synthetic images used to train our Bias Amplifier

Summary of Contributions

We introduce DDB, a novel plug-in framework that learns bias distribution of training data using a conditional diffusion model.

DDB trains a Bias Amplifier on synthetic bias-aligned samples. This uniquely avoids common issues like bias-conflicting sample overfitting and interference.

DDB achieves new state-of-the-art results on popular biased benchmarks (for both single and multiple biases) outperforming previous methods, including those based on Vision-Language Models.

References

[1] D'Incà et al. (2024). "Openbias: Open-set bias detection in text-to-image generative models." In: CVPR 2024
[2] Kim et al. (2024). "Discovering and mitigating visual biases through keyword--explanation.". In: CVPR 2024
[3] Lee et al. (2023). "Revisiting the importance of amplifying bias for debiasing". In: AAAI 2023.
[4] Nam et al. (2020). "Learning from failure: De-biasing classifier from biased classifier." In: NIPS 2020
[5] Sagawa et al. (2020). "Distributionally robust neural networks." In: ICLR 2020