Gradient Extrapolation for Debiased Representation Learning

Ihab Asaad[1], Maha Shadaydeh[1], Joachim Denzler[1]

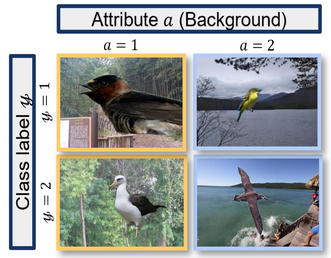
[1]Computer Vision Group, Friedrich Schiller University Jena, Germany

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Background & Motivation

Spurious correlations are unintended associations between non-causal features and labels. ERM models often exploit them when prevalent and easier to learn , rather than relying on the causal features, leading to biased predictions and poor generalization, especially when these correlations are absent in test.

Motivation: Despite extensive research, current debiasing methods still struggle on highly biased datasets. This calls for more general and effective approaches that promote causal representation and avoid spurious correlation.



Example: Waterbirds Classification
Most Waterbirds appear on water background and most landbirds on land. ERM models exploit this correlation, relying on backgrounds to predict the labels rather than intrinsic bird features.

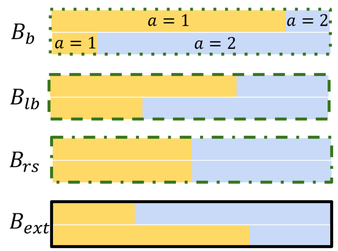
Contributions

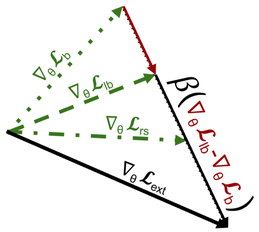
GERNE, A novel training strategy that:

- Steers representation learning away from spurious correlations.
- **General debiasing framework**, with ERM and Resampling as special cases.
- Optimizes **Group-Balanced or Worst-Group Accuracy**.
- Achieves **SOTA** on diverse vision & NLP benchmarks, in both known and unknown attributes.

https://gerne-debias.github.io/

GERNE: Gradient Extrapolation for Debiased Representation Learning





$$\nabla_\theta \mathcal{L}_{ext} = \nabla_\theta \mathcal{L}_{lb} + \beta \cdot (\nabla_\theta \mathcal{L}_{lb} - \nabla_\theta \mathcal{L}_b)$$

GERNE samples two types of batches with different levels of spurious correlations B_b, B_{lb} and computes their respective losses. The difference between the gradients of these losses indicates a **debiasing direction**. GERNE then **linearly extrapolates** these gradients toward the batch with fewer spurious correlations to form a **target gradient**, which—controlled by an extrapolation factor β—is used to update the model parameters.

Bounds on $\boldsymbol{\beta}$ are determined by keeping the target conditional attribute distribution $p_{ext}(a|y) = \alpha_{ya} + c \cdot (\beta + 1) \cdot \left(\frac{1}{A} - \alpha_{ya}\right) \in [0,1]$ for all $(y,a) \in \mathcal{G}$; $\alpha_{ya} = p_b(a|y)$ and c is the bias reduction factor between B_b, B_{lb}.

GERNE as a General Debiasing Framework:

- $\beta = -1$, GERNE reduces to ERM in loss expectation.
- $c = 1, \beta = 0$, GERNE reduces to Resampling.
- $c \cdot (\beta + 1) = 1$, GERNE reduces to Resampling in loss expectation.
- $c \cdot (\beta + 1) > 1$, GERNE oversamples the minority groups while keep controlling the loss's variance.

The connection between β and the risk for **the worst-case group** (L_g is group g's risk)**:**

$$\mathcal{L}_{ext} = \frac{1}{K} \cdot \sum_{g=(y,a)\in\mathcal{G}} p_{ext}(a|y)(\beta) \cdot L_g \,,$$

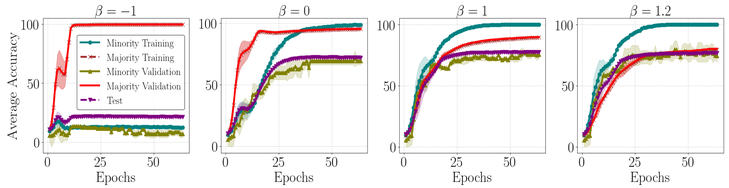
where K is the number of classes.

GERNE for Unknown Attributes:

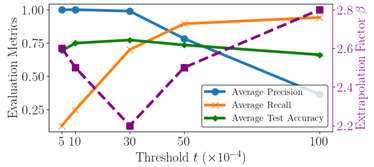
- Train an ERM model to obtain a biased predictor $\tilde{f}$.
- Split samples by class into easy vs. difficult based on confidence threshold t using $\tilde{f}$.
- Form **pseudo-groups** combining classes and difficulty.
- Apply GERNE as in the known attributes case.

Ablation Study

Ablation 1-The effect of tuning $\boldsymbol{\beta}$: C-MNIST, 0.5% bias-conflicting ratio, known attributes. By tuning β, Group-Balanced accuracy can be maximized.



Ablation 2- The effect of t on β for best model performance: C-MNIST, 0.5% bias-conflicting ratio, unknown attributes. Accurate minority identification leads to best performance.



Experiments & Discussion

Experiment 1: GERNE without data augmentation.

- **Metrics:** Group-Balanced accuracy (C-MNIST, C-CIFAR-10) and minority-group accuracy (bFFHQ).

Methods	Group Info	C-MNIST 0.5	C-MNIST 5	C-CIFAR-10 0.5	C-CIFAR-10 5	bFFHQ
Group DRO	✓	63.12	84.20	33.44	57.32	-
Resampling	✓	77.68±0.89	91.98±0.08	45.10±0.60	62.16±0.05	72.13±0.90
GERNE (ours)	✓	**77.79±0.90**	**92.16±0.10**	**45.34±0.60**	**62.40±0.27**	**85.20±0.86**
ERM	×	35.19±3.49	82.17±0.74	23.08±1.25	39.42±0.64	56.70±2.70
JTT	×	53.03±3.89	84.03±1.10	24.73±0.60	42.20±0.31	65.30±2.50
LfF	×	52.50±2.43	84.79±1.09	28.57±1.30	50.27±1.56	62.20±1.60
DFA	×	65.22±4.41	89.66±1.09	29.75±0.71	51.13±1.28	63.90±0.30
LC	×	71.25±3.17	91.16±0.97	34.56±0.69	54.55±1.26	69.67±1.40
GERNE (ours)	×	**77.25±0.17**	90.98±0.13	**39.90± 0.48**	**56.53±0.32**	**76.80±1.21**

Experiment 2: GERNE with data augmentation for fair comparison:

- **Metrics:** Worst-Group Accuracy.

Methods	Group Info	Waterbirds	CelebA	Civil-Comments
Group DRO	✓	78.60±0.30	89.00±0.20	70.60±0.30
ReSample	✓	77.70±0.30	87.40±0.20	73.30±0.20
DFR	✓	**91.00±0.10**	90.40±0.05	69.60±0.10
LISA	✓	88.70±0.20	86.50±0.40	73.70±0.10
GERNE (ours)	✓	90.20±0.08	**91.98±0.05**	**74.65±0.07**
ERM	×	69.10±1.20	57.60±0.30	63.20±0.40
ReWeight	×	72.50±0.10	81.50±0.30	69.90±0.20
DFR	×	89.00±0.05	86.30±0.10	63.90±0.10
CnC	×	88.50±0.10	**88.80± 0.30**	68.90±0.70
GERNE (ours)	×	**90.21±0.15**	86.28±0.05	**71.00±0.10**

Discussion

- GERNE debiases models by **extrapolating gradients** from batches with different level of spurious correlations. By tuning the β, it can optimize either **Group-Balanced or Worst-Group Accuracy**.
- Results show that GERNE effectively mitigates spurious correlations, outperforming state-of-the-art methods on diverse vision and NLP benchmarks.

Acknowledgments: This study is funded by the Carl Zeiss Foundation within the project Sensorized Surgery, Germany (P2022-06-004).
Contact: {ihab.asaad, maha.shadaydeh, joachim.denzler}@uni-jena.de

Carl Zeiss Stiftung

SensX Sensorisierte Chirurgie