JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Learning Debiased Representations via Conditional Attribute Interpolation

Yi-Kai Zhang, Qi-Wei Wang, De-Chuan Zhan, Han-Jia Ye✉

Biased Representation

(a) Orange lifeboat

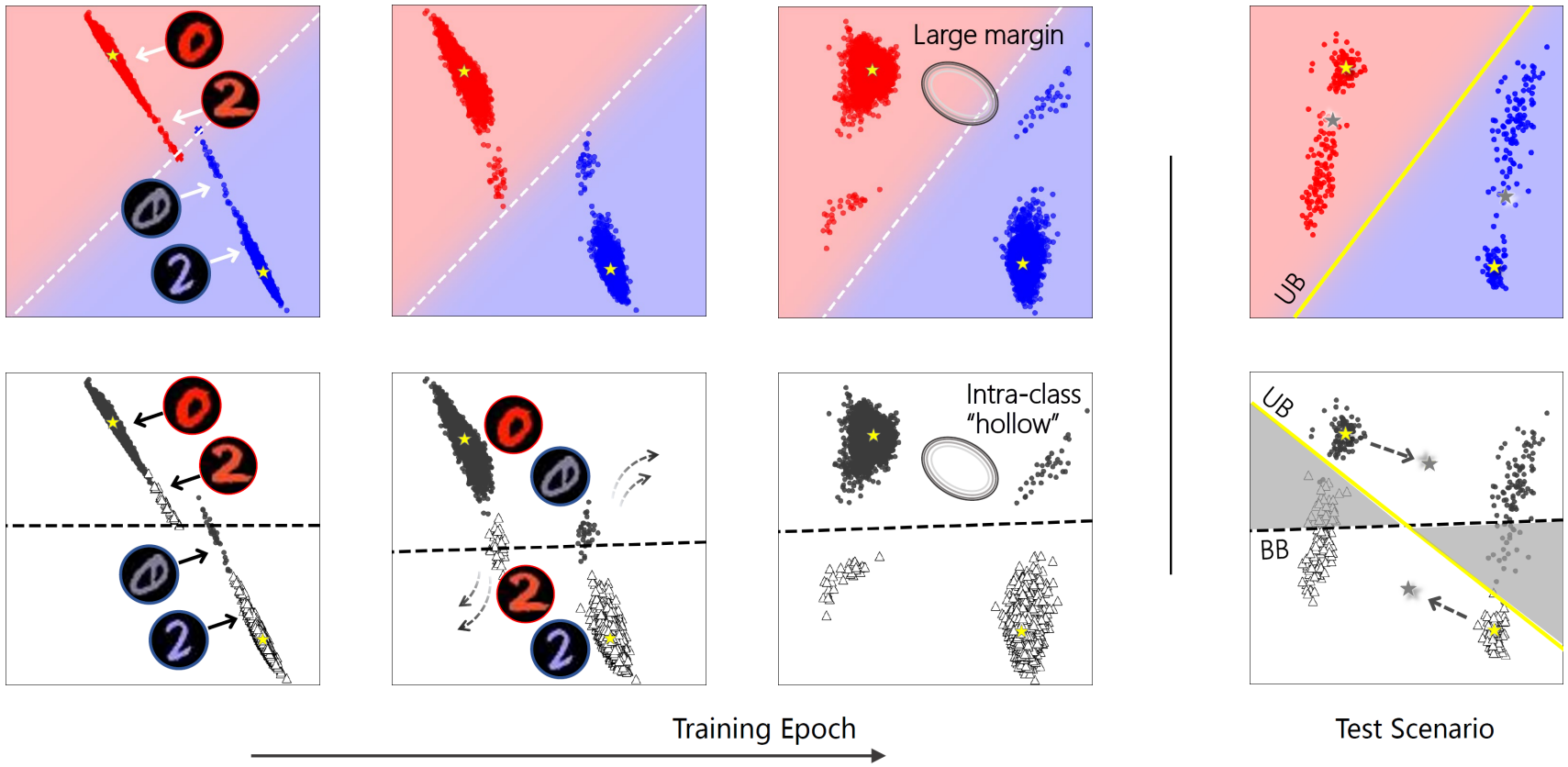
97.8%	lifeboat
0.5%	beacon
0.4%	container ship

(b) Orange cyclists

57.0%	lifeboat
13.8%	bicycle-built-for-two
9.0%	toyshop

- Classification of a standard **ImageNet pre-trained** `ResNet-50` of
  - (a) an ***orange*** **lifeboat** for training (with ***color*** and ***shape*** attributes) √
  - (b) an ***orange*** **cyclist** for test (***aligned*** with color, ***conflicting*** with shape) ×

Biased Training Dynamics



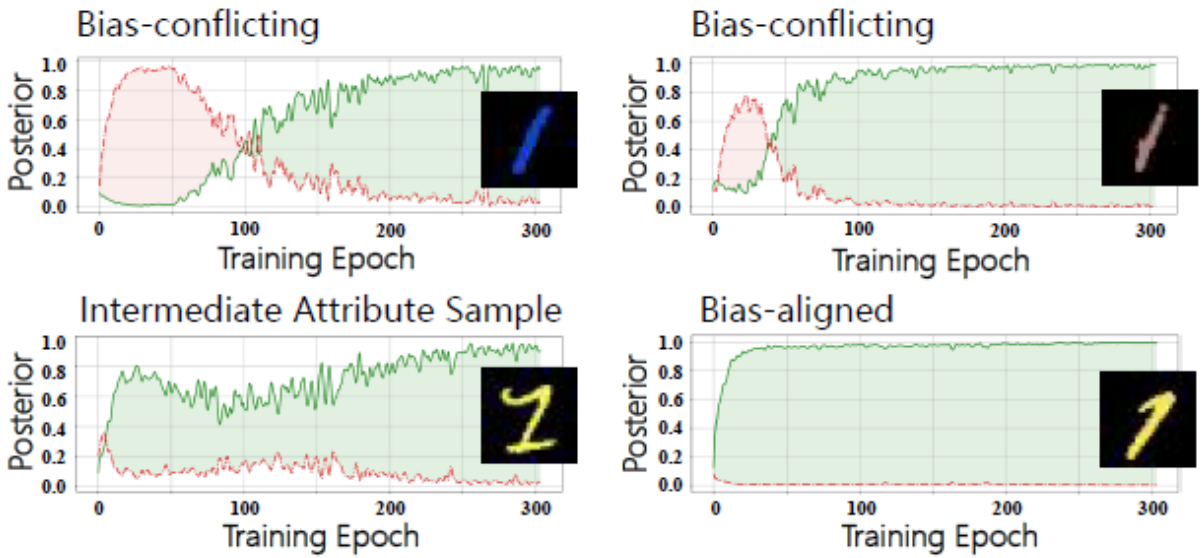
- ***Color*** attribute classifier for ● in red and ● in blue, ***Shape*** attribute classifier for 0 Ø in ● and 2 2 in △

  (a) The target attribute **shape** is learned later in a "***lazy***" manner.

  (b) The easier-to-learn **color** leaves a large margin and triggers the shape intra-class "***hollow***".

  (c) With the "hollow" **conditioned on** a particular shape, the true class representation is **deviated** toward color.

Find Intermediate Attribute Samples with χ-pattern



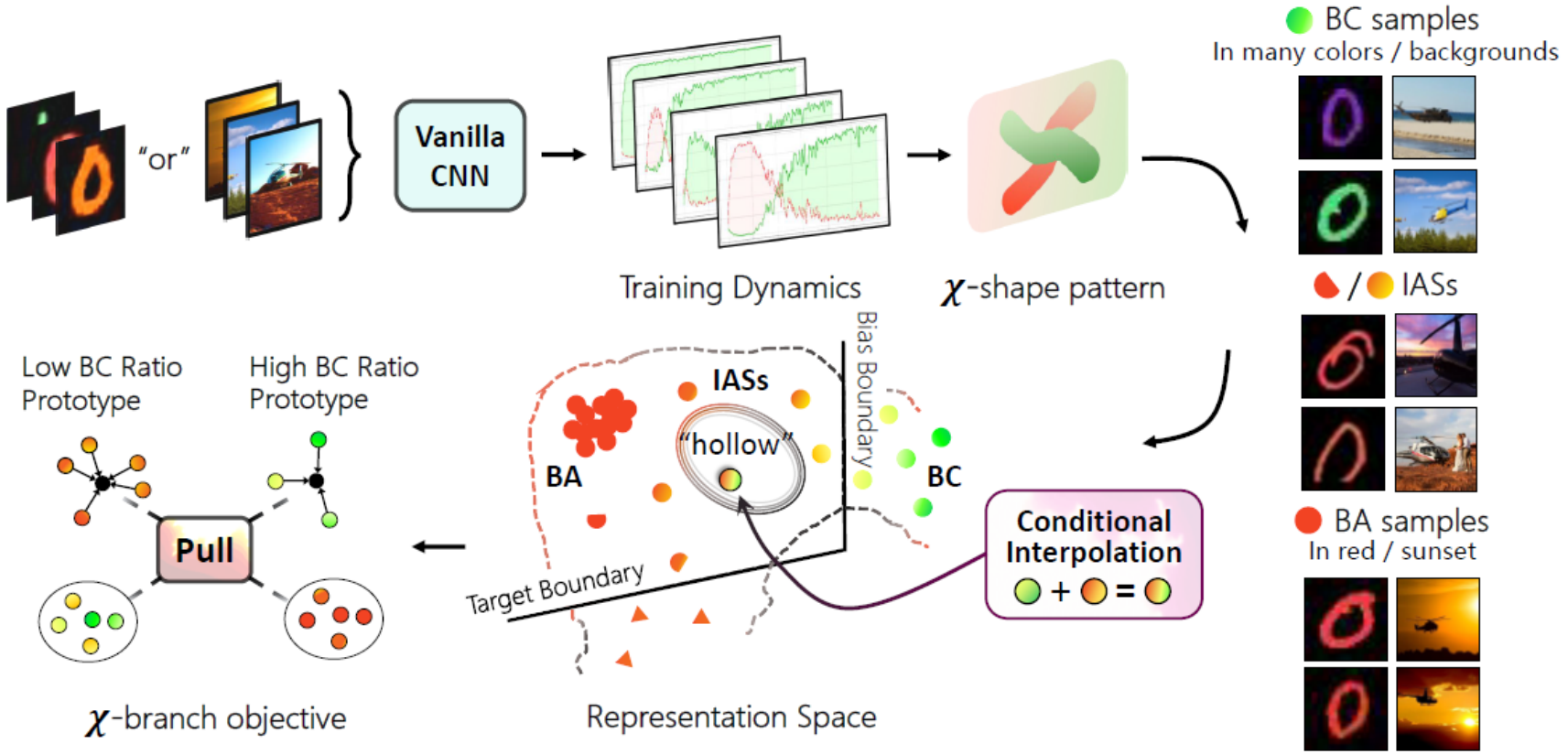
$$\mathcal{L}_{\mathrm{CE}}(\mathbf{x}_i) = \begin{pmatrix} \mathcal{L}^{gt}_{\mathrm{CE}}(\mathbf{x}_i) = \left\{-\log \Pr^t(y_i \mid \mathbf{x}_i)\right\}_{t=1}^{T} \\ \mathcal{L}^{b}_{\mathrm{CE}}(\mathbf{x}_i) = \left\{-\log \Pr^t(b_i \mid \mathbf{x}_i)\right\}_{t=1}^{T} \end{pmatrix},$$

$$\chi_{\mathrm{pattern}} = \begin{pmatrix} \mathrm{p}^{gt} = \left\{e^{-A_1 t}\right\}_{t=1}^{T} \\ \mathrm{p}^{b} = \left\{e^{A_2 t}\right\}_{t=1}^{T} \end{pmatrix},$$

- The change of posterior over the ***GT-class*** as well as the ***bias class***

$$\begin{aligned} \mathbf{s}(\mathbf{x}_i) &= \langle \mathcal{L}_{\mathrm{CE}}(\mathbf{x}_i), \chi_{\mathrm{shape}} \rangle \\ &= \langle \mathcal{L}^{gt}_{\mathrm{CE}}(\mathbf{x}_i), \mathrm{p}^{gt} \rangle + \langle \mathcal{L}^{b}_{\mathrm{CE}}(\mathbf{x}_i), \mathrm{p}^{b} \rangle \\ &= \sum_{t=1}^{T} \left(-e^{-A_1 t} \cdot \log \Pr\left(h_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i \mid \mathbf{x}_i\right)\right. \\ &\quad \left. -e^{A_2 t} \cdot \log \Pr\left(h_{\boldsymbol{\theta}}(\mathbf{x}_i) = b_i \mid \mathbf{x}_i\right)\right). \end{aligned}$$

Learning from χ-structured objective



$$\mathbf{p}_{\boldsymbol{\gamma},c} = \frac{1}{K} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{B}_{\boldsymbol{\gamma}}} f_{\boldsymbol{\phi}}(\mathbf{x}_i) \cdot \mathbb{I}[y_i = c], \quad \Pr(y_i \mid \mathbf{x}_i) = \frac{\exp\left(-\mathrm{d}\left(f_{\boldsymbol{\phi}}(\mathbf{x}_i), \mathbf{p}_{\boldsymbol{\gamma},y_i}\right)/\tau\right)}{\sum_{c \in [C]} \exp\left(-\mathrm{d}\left(f_{\boldsymbol{\phi}}(\mathbf{x}_i), \mathbf{p}_{\boldsymbol{\gamma},c}\right)/\tau\right)}$$

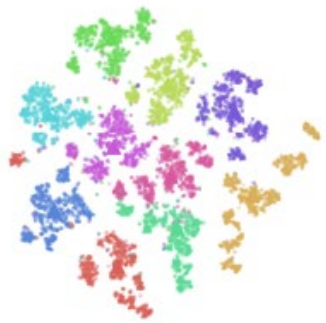
Benchmark Results of χ^2-model

- State-of-the-art on Colored `MNIST`, Corrupted `CIFAR-10`, Biased `CelebA` and Biased `NICO`.

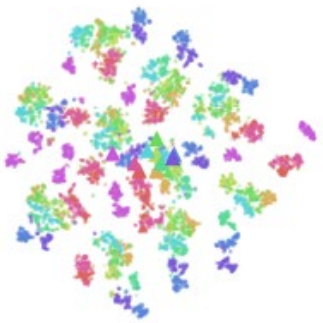
Dataset		Colored MNIST				Corrupted CIFAR-10			
Ratio (%)		99.9	99.5	99.0	95.0	99.9	99.5	99.0	95.0
Vanilla	○	28.58	59.29	74.42	87.13	26.91	30.16	37.71	41.60
+ p [44]	○	31.01	64.82	76.84	87.86	26.55	29.48	38.07	42.30
RUBi [7]	◑	27.82	70.80	86.58	96.77	33.70	34.70	34.59	47.23
ReBias [4]	◑	27.71	72.89	85.95	96.87	33.65	34.40	35.82	47.45
End [46]	●	28.19	81.81	88.10	96.99	31.30	33.83	34.02	38.77
DI [50]	●	33.18	80.63	86.28	**98.36**	32.09	33.37	37.65	51.27
LfF [36]	○	30.24	68.90	76.69	96.81	29.89	33.68	35.28	45.38
LFA [27]	○	22.31	64.13	81.83	95.45	32.49	35.74	39.63	47.25
χ-pattern + p	●	60.33	64.15	**93.53**	98.30	35.33	**39.31**	**41.32**	**53.37**
χ^2-model (Ours)	○	**66.91**	**88.73**	92.15	97.87	**35.67**	37.61	40.74	49.04

Data	Biased CelebA			NICO
Source	BA	BC	All	All
LfF [36]	73.69	**70.41**	72.05	34.44
DFA [27]	94.01	58.98	76.50	33.10
χ^2-model	**97.66**	60.79	**79.23**	**36.99**

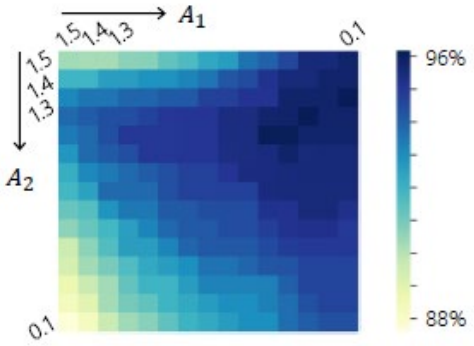
Measure	Acc. ↑	98%-σ ↓	AP ↑
Entropy [20]	78.33	632	83.52
Confidence [28]	80.33	590	85.61
Loss [36]	94.39	418	98.22
Pleiss *et al.* [39]	82.67	686	89.24
Zhao *et al.* [55]	90.33	451	96.04
χ-pattern (Ours)	**95.84**	**372**	**98.44**



(a) t-SNE on a_t

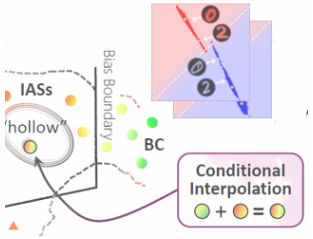


(b) t-SNE on a_b



(c) ablation study of A_1, A_2

Feel free to discuss with us~

- 

χ^2-model: a practical method to learn debiased representation
- Code is available at github.com/ZhangYikaii/chi-square

"Thanks!"

Paper

Code

Yi-Kai Zhang's Homepage
lamda.nju.edu.cn/zhangyk

Qi-Wei Wang's Homepage
lamda.nju.edu.cn/wangqiwei