Severing Spurious Correlations with Data Pruning

ICLR 2025, Spotlight

Varun Mulchandani

vmmulcha@ncsu.edu

Jung-Eun Kim

jung-eun.kim@ncsu.edu

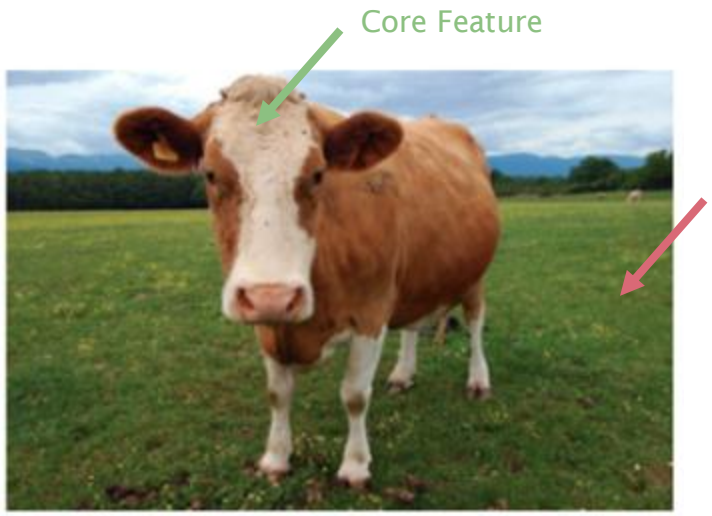
Cow Training Images

Camel Training Images

(Arjovsky et al. 2019)

Spurious Correlations

- During train time



W * Green Background = Cow

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

Spurious Correlations

- During test time

W * Green Background = Cow ✗

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

State of Existing Solutions

- Existing study settings where:
  - Strength(spurious signal) >> Strength(core, invariant signal).
  - E.g.: 97% of samples of a class have the spurious feature.
  - Easy to detect sample-wise presence - Identifiable.

- But we study both: Unidentifiable and Identifiable settings.

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

CelebA Gender Classification Train Set Setup

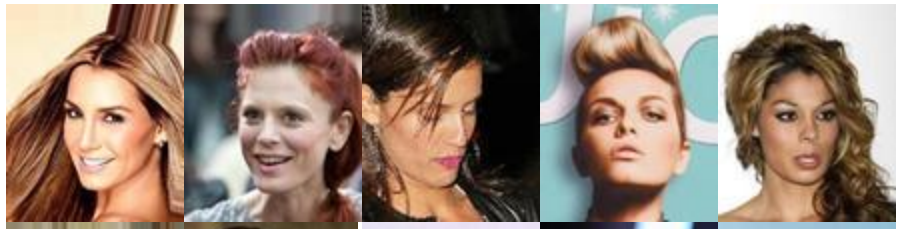
Male

Female

With glasses

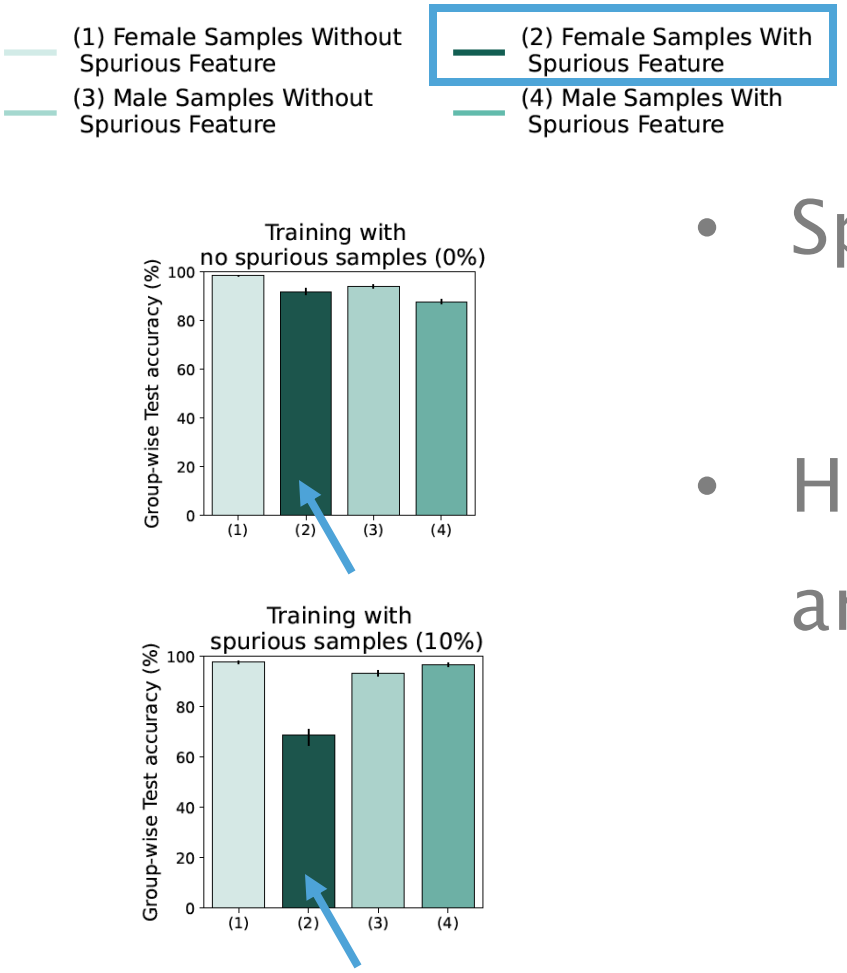
No glasses

No glasses



"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

When Spurious Signals are Weak



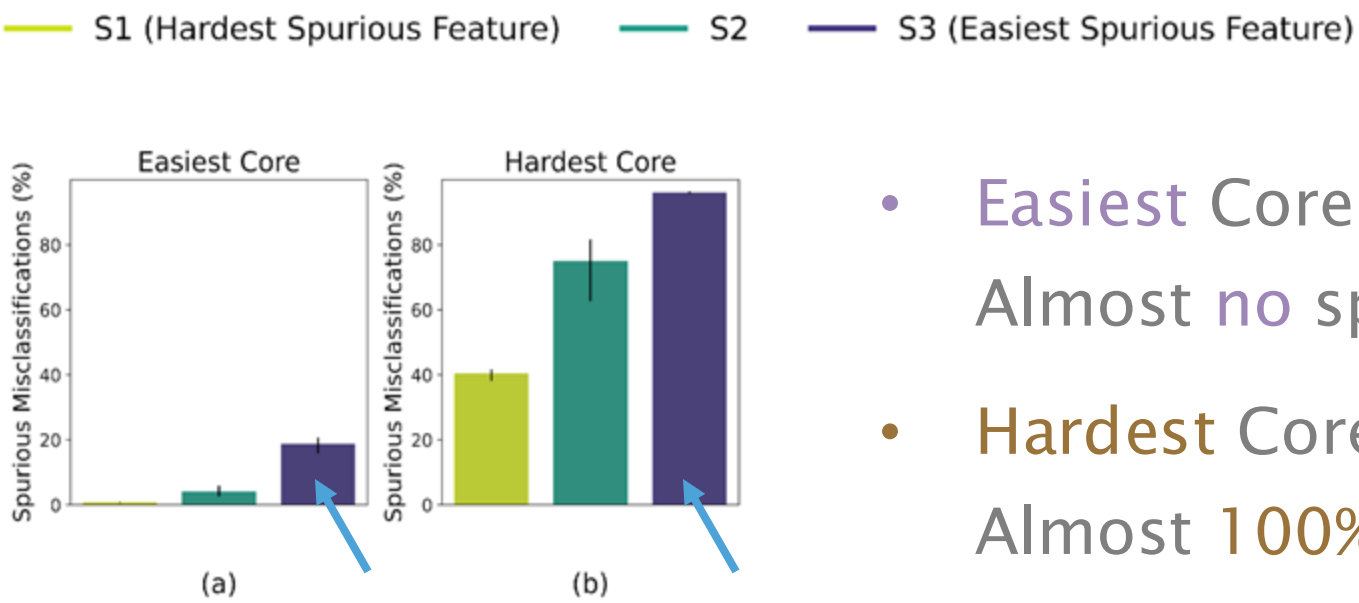
- Spurious Correlations are still relied upon.
- However, we show that spurious features are unidentifiable.

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

How to Overcome Spurious Correlations in Unidentifiable Settings?

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

Sample-Wise Contribution to Spurious Correlations

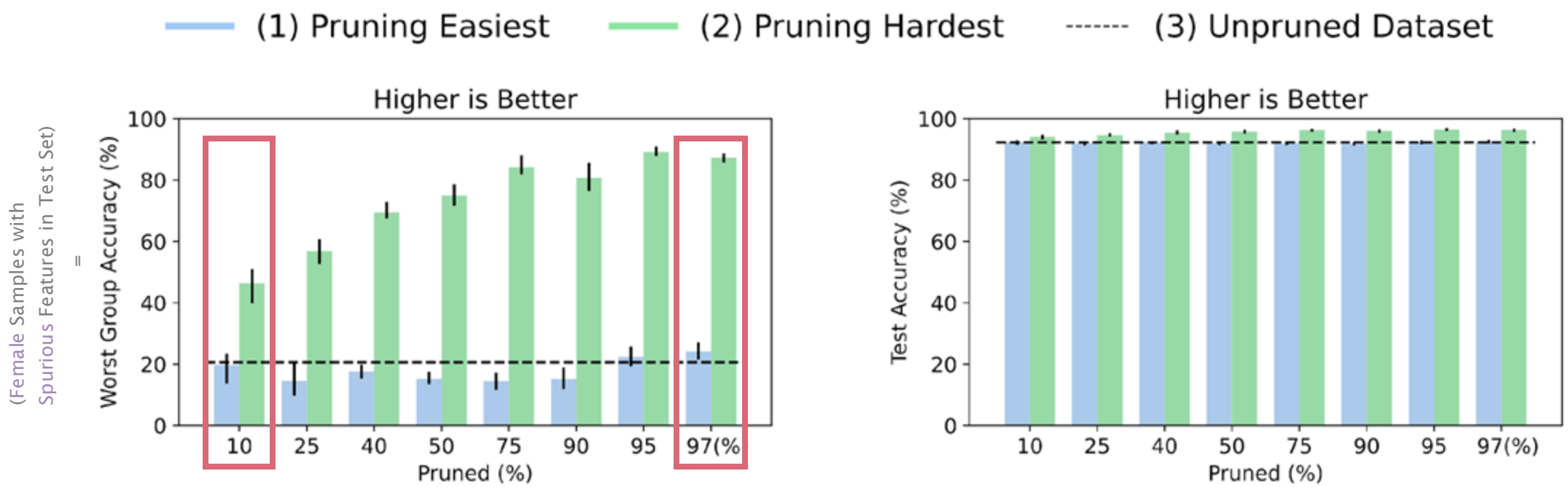


- Easiest Core + Spurious Feature:
  Almost no spurious misclassifications.
- Hardest Core + Spurious Feature:
  Almost 100% spurious misclassifications.

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

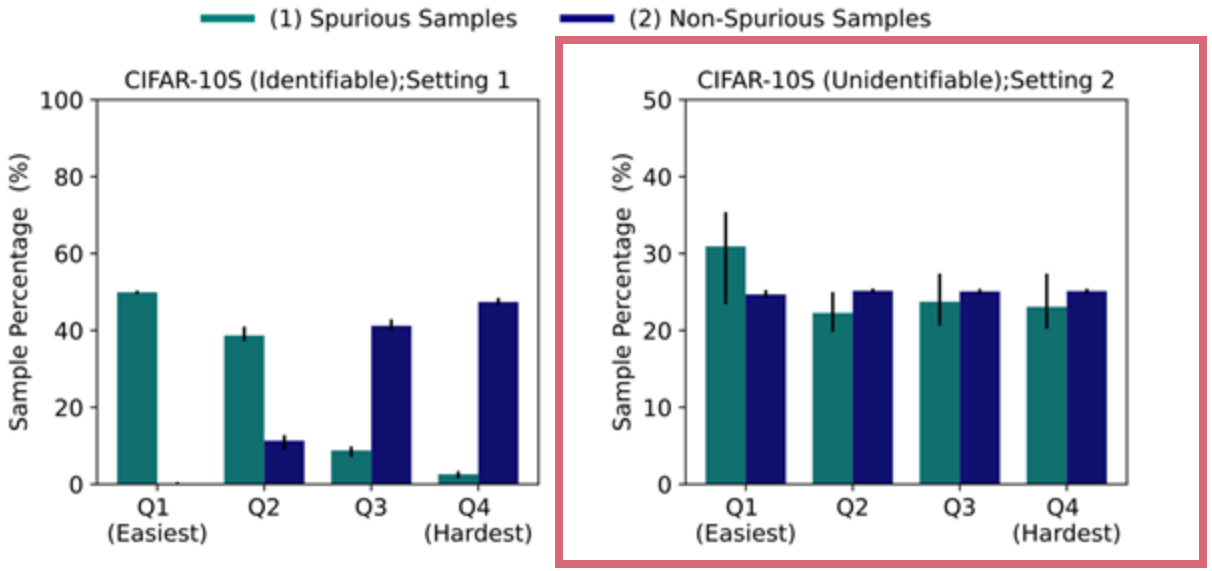
Samples with Hard Core + Spurious Features are Primary Contributors to SC Reliance

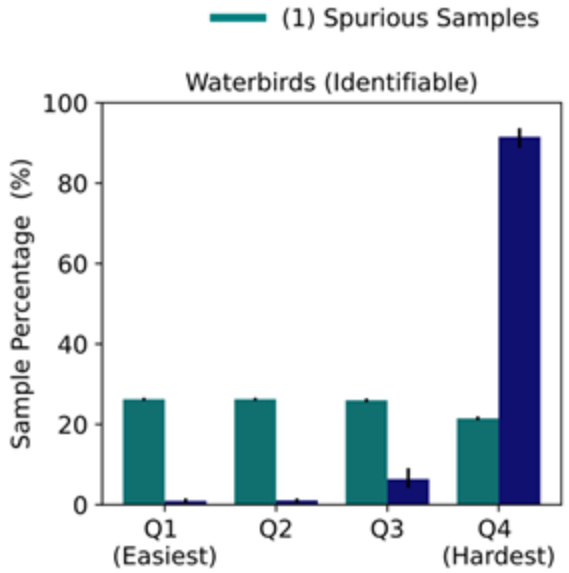
CelebA Gender Classification

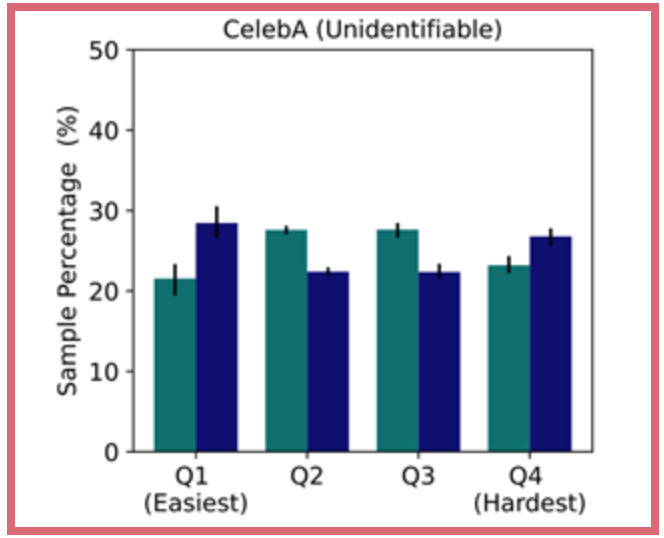


Pruned (%) represents number of samples with spurious features removed from train data.

Identifiable vs. Unidentifiable Settings







"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

Severing Spurious Correlations with Data Pruning

- In an unidentifiable setting,

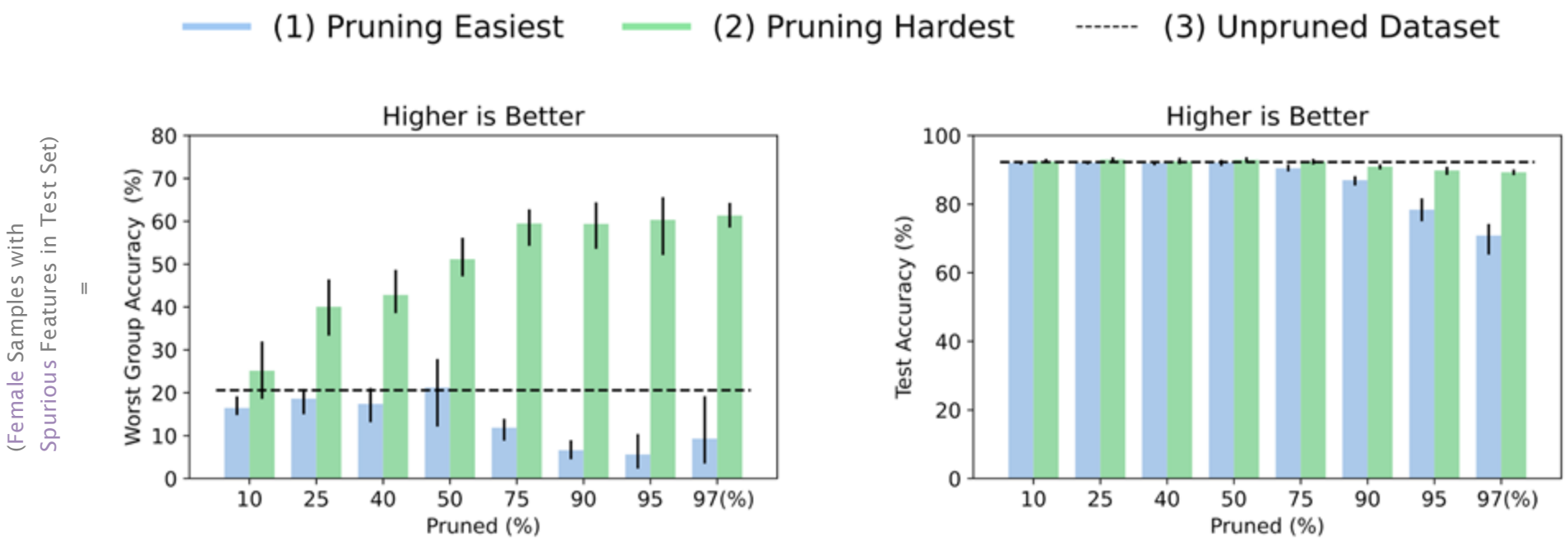
  you can just prune the hardest samples.
- In an identifiable setting,

  just prune the hardest samples w/ spurious features.

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

Severing Spurious Correlations with Data Pruning

Unidentifiable setting – CelebA Gender Classification



"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

Severing Spurious Correlations with Data Pruning

Identifiable setting

Method	Waterbirds (%)		MultiNLI (%)		Group Labels	
	Worst%	Mean%	Worst %	Mean%	Train	Val
ERM	74.81 (0.7)	98.10 (0.1)	65.9 (0.3)	82.8 (0.1)	✗	✗
CnC (Zhang et al., 2022)	88.5 (0.3)	90.9 (0.1)	-	-	✗	✓
JTT (Liu et al., 2021)	86.7	93.3	72.6	78.6	✗	✓
gDRO (Sagawa et al., 2020a)	86.0	93.2	**77.7**	81.4	✓	✓
DFR^{Tr} (Kirichenko et al., 2023)	90.2 (0.8)	97.0 (0.3)	71.5 (0.6)	82.5 (0.2)	✓	✓
PDE (Deng et al., 2023)	90.3 (0.3)	92.4 (0.8)	-	-	✓	✓
Ours	**90.93 (0.58)**	92.48 (0.72)	75.88 (1.62)	81.07 (0.25)	✓	✓

"Severing Spurious Correlations with Data Pruning"
Mulchandani & Kim, ICLR 2025

Our Core Insight

This paper discovers that spurious correlations are learned from a very small fraction of the samples containing spurious features.

They can be removed from the dataset even if one cannot determine/infer what spurious features/correlations are present in the dataset, to mitigate spurious correlations.

Thank you