"A STUDY ON THE DISTRIBUTION OF SOCIAL BIASES IN SELF-SUPERVISED LEARNING VISUAL MODELS"

Kirill Sirotkin, Pablo Carballeira, Marcos Escudero-Viñolo

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

MOTIVATION

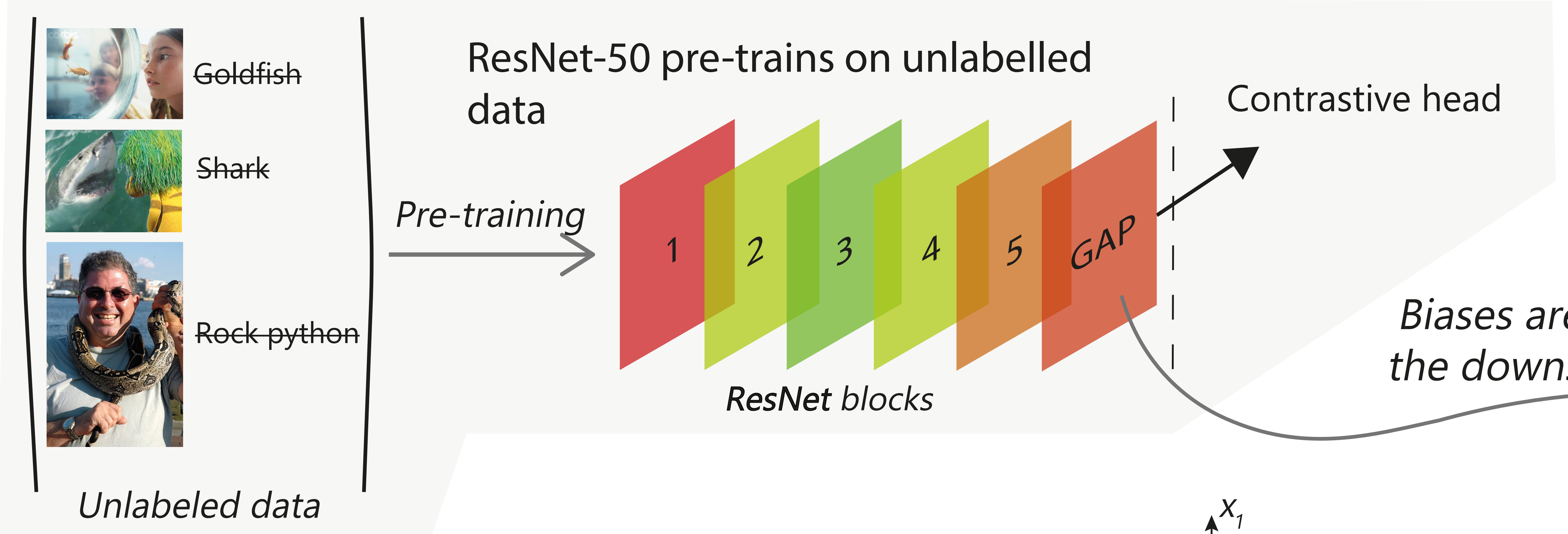
- Self-Supervised Learning (SSL) models can learn biases, **contrary** to common belief.
- It is unclear if some **types** of SSL learn **more** biases than others.

STUDIED SOCIAL BIASES INCLUDE

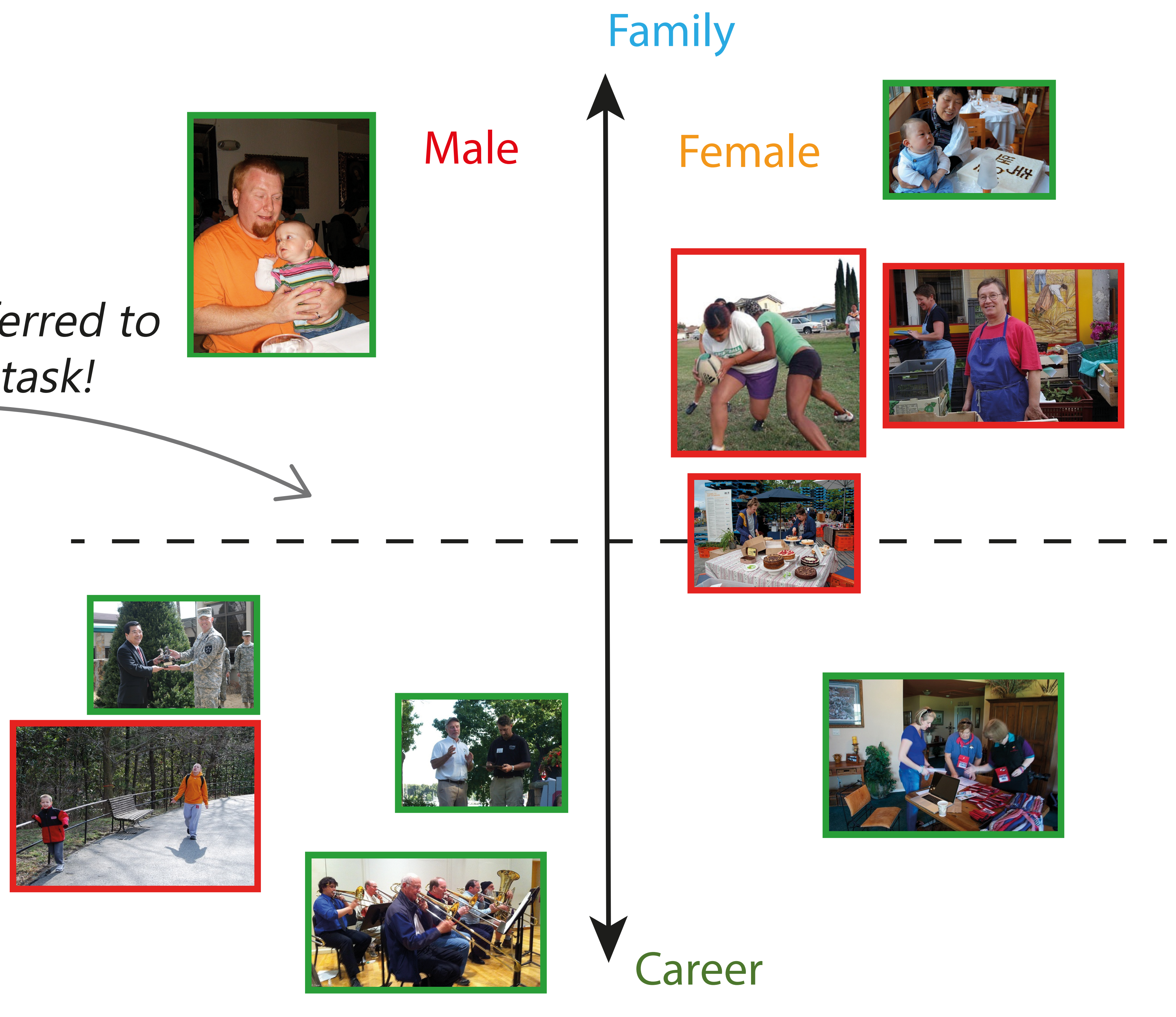
Concepts (X, Y)	Attributes (A, B)
Male and Female	Career and Family
Black and White	Weapon and Tool
Skinny and Overweight	Pleasant and Unpleasant

MAIN FINDINGS

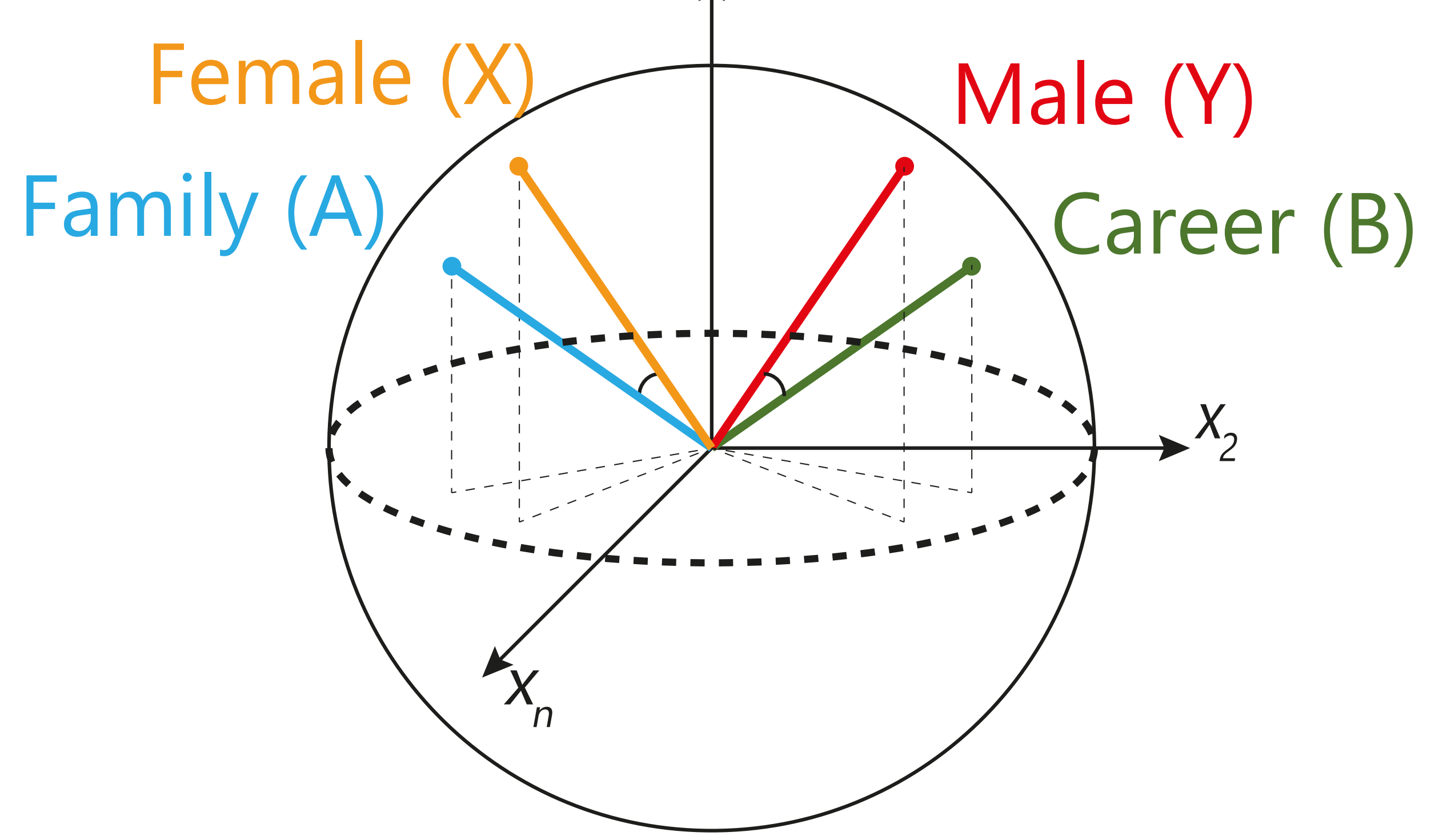
- **Contrastive** SSL models learn more biases than geometric or clustering ones.
- Most biases are learned in the **deep layers** of a network.
- SSL-learned biases **transfer** to the downstream task.
- There is a **trade-off** between the accuracy and a number of learned biases.



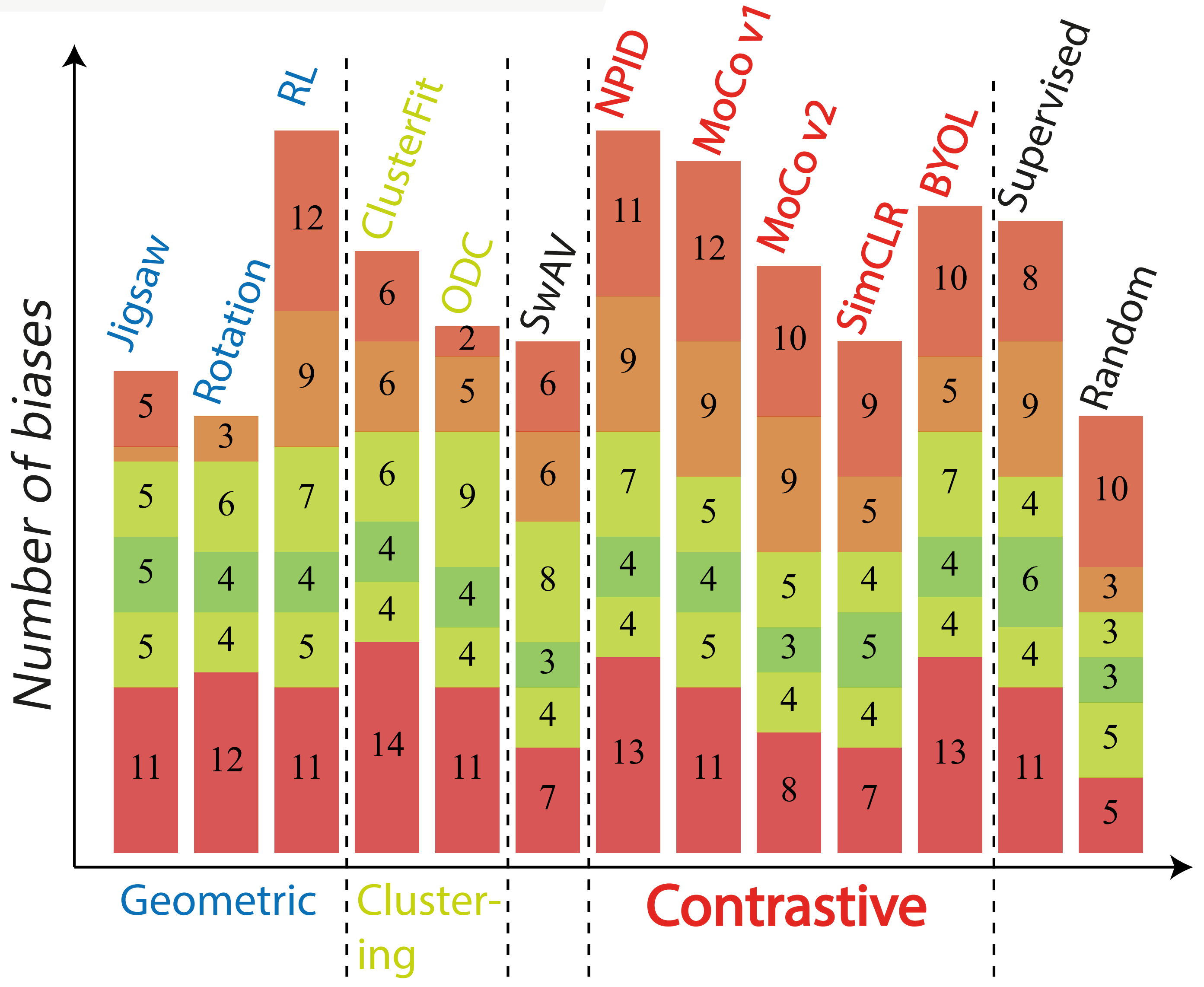
Biases are transferred to the downstream task!



Correct predictions are shown in green



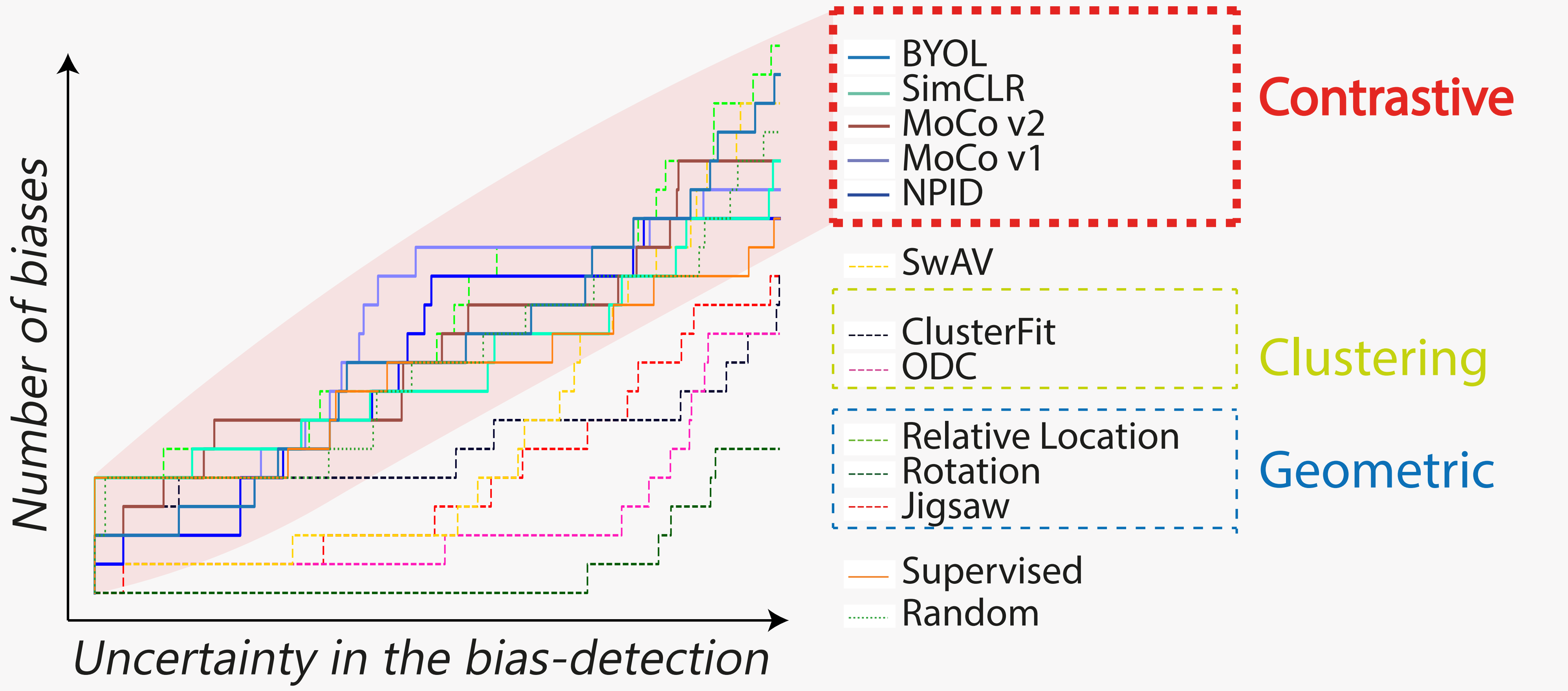
Images in the feature space



Bias-detection test

$$\sum_{x\in X} s(x, A, B) - \sum_{y\in Y} s(y, A, B)$$

$$\underset{a\in A}{\operatorname{mean}} \cos(x, a) - \underset{b\in B}{\operatorname{mean}} \cos(x, b) \qquad \underset{a\in A}{\operatorname{mean}} \cos(y, a) - \underset{b\in B}{\operatorname{mean}} \cos(y, b)$$



Universidad Autónoma de Madrid

Video Processing and Understanding Lab