Supercharging Imbalanced Data Learning With Energy-based Contrastive Representation Transfer

NeurIPS 2021 Spotlight Presentation

Junya Chen*, Zidi Xiu*, Benjamin Goldstein, Ricardo Henao, Lawrence Carin, Chenyang Tao

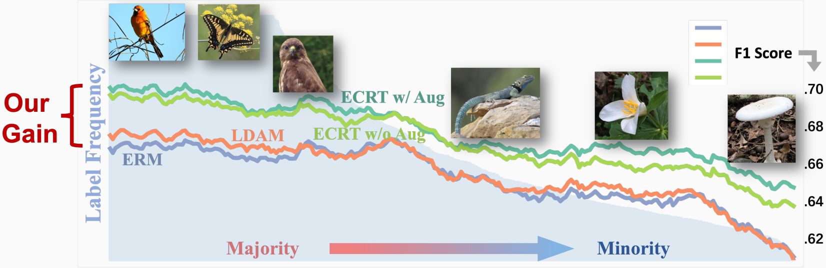
* Contributed equally

November 6, 2021

Duke University

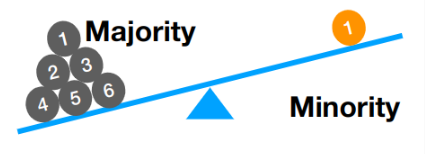
Contributions

- A **causal representation** encoder informed by an invariant generative mechanism
- Augmentation exploiting **feature independence** to enrich minority representation
- Novel **energy-based perspective** for to improve generalized contrastive learning
- Fresh understanding on how causal representation improves learning

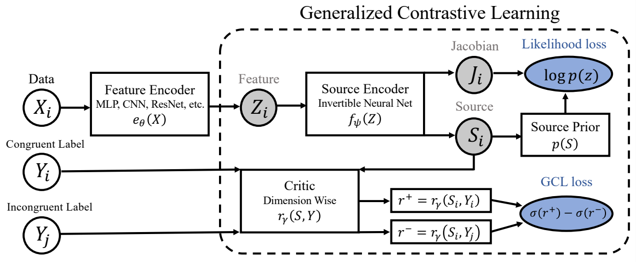


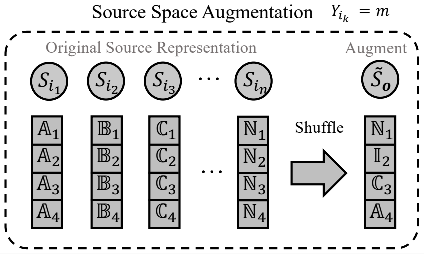
Presentation Outline

- Problem Setup: Class Imbalanced Learning
- Limitations of Current Solutions
- Causal Representation Transfer
- Minority Augmentation
- Experiments



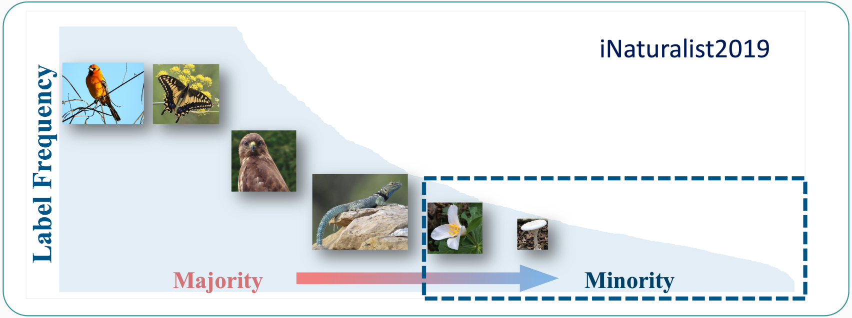
Our Solution: ECRT





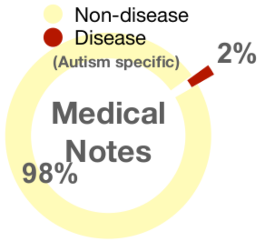
Imbalanced Learning is Ubiquitous in Real-world

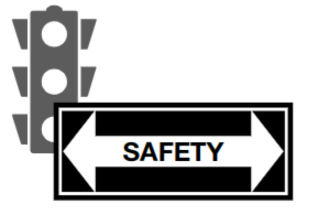
- Training samples are scarce for certain categories/tasks
  - Over 1000 classes
  - Skewed distribution in label frequency (3 ~ 500 training samples)



More Examples

- Fraud detection
- Spam filtering
- Rare disease
- Advertising click-throughs





Naive Classifier Is Not Suitable

- Imbalance induces bias towards the majority classes
- **But the minority classes are of primary interest**

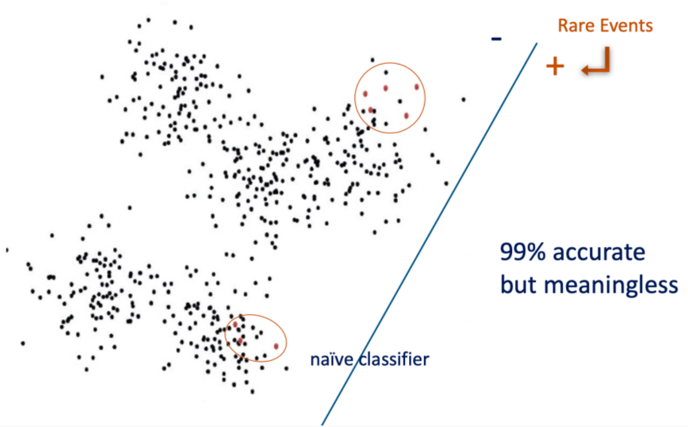
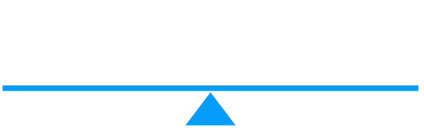


Image credit to Cynthia Rudin

Current Approaches to Class Imbalance

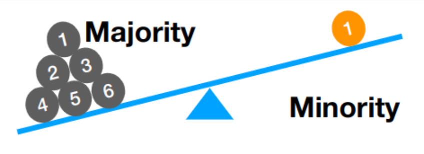
Two Popular Solutions

- ☐ Resampling
- ☐ Reweighting



Two Popular Solutions

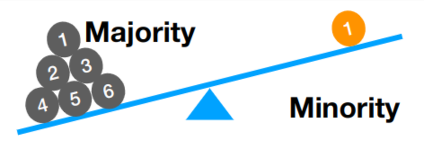
- Resampling
  - Alters training data sampling distributions
  - Upsampling minority, downsampling majority
- Reweighting



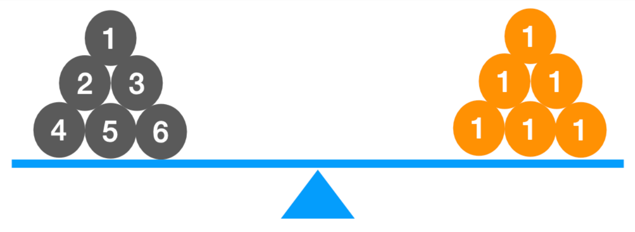
Imbalanced Data

Two Popular Solutions

- Resampling
  - Alters training data sampling distributions.
  - Upsampling minority, downsampling majority.
- Reweighting



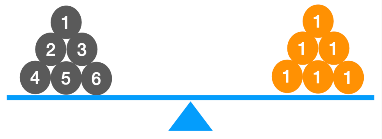
Imbalanced Data



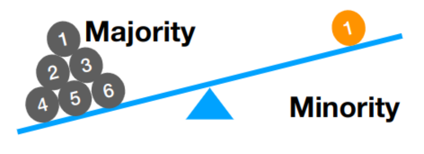
Resampling

Two Popular Solutions

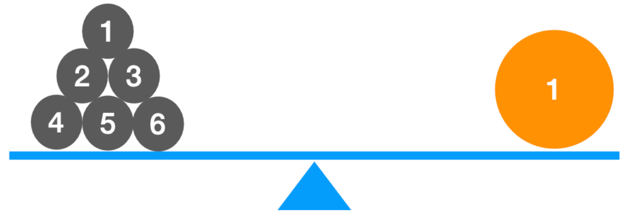
- ☐ Resampling
- ■ Reweighting
  - Amends the relative importance of each sample
  - Bases on the inverse of their frequency.



Resampling



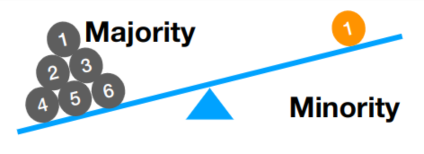
Imbalanced Data



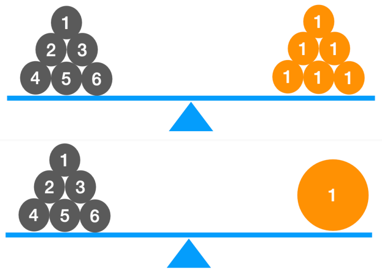
Reweighting

However, Simple Methods Might Not Be That Effective...

- Fails to maintain the original distribution $\Leftarrow$ **Lack of generalization**
- Creates artificial balance $\Leftarrow$ **Introduces biases**



Imbalanced Data

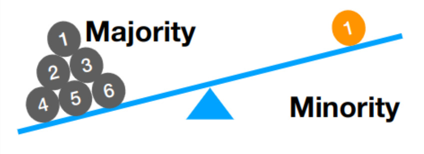


Resampling

Reweighting

Recent Developments: Adaptive Loss

- Focal[1]: modify the entropy loss
- LDAM[2]: modify the margin hinge loss



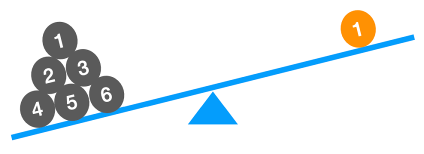
Imbalanced Data

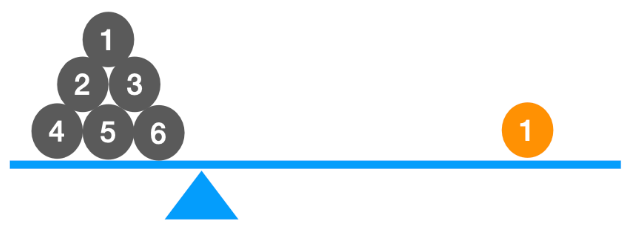
[1] Tsung-Yi Lin, et al. Focal loss for dense object detection. In ICCV, 2017.

[2] Kaidi Cao, et al. Learning imbalanced datasets with label-distribution-aware margin loss. In NeurIPS, 2019

Recent Developments: Adaptive Loss

- ☐ Focal
- ☐ LDAM
- ■ **Incur overfitting issues due to training bias and label noise.**

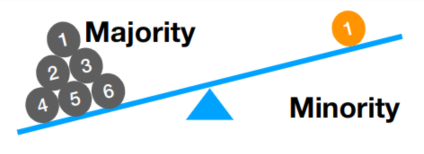




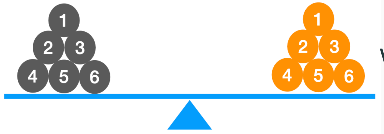
LDAM & Focal

Recent Developments: Augmentations

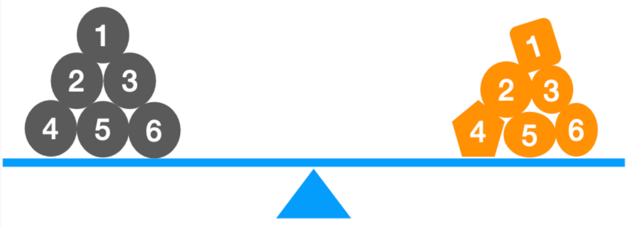
- Invariant transformations: rotation, scaling
- Generative augmentation procedures: Generalized adversarial learning (GAN)[1]
- **Might be biased by spurious relationships**



Imbalanced Data



What we hope

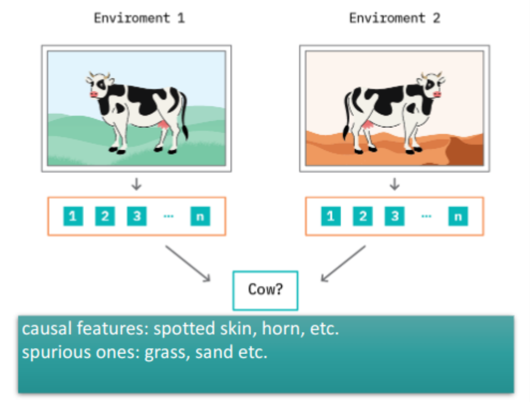


What we actually get

Why We Need Causal Transfer

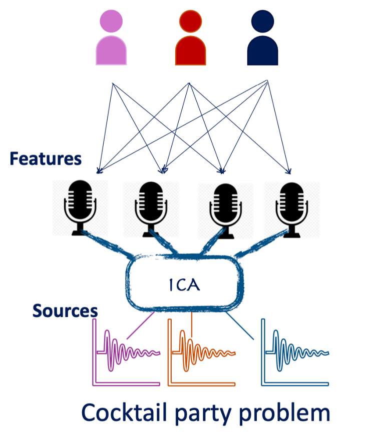
Motivation

Causal relations are invariant across environments



[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

So How Can We Find Such “Invariant” Features?

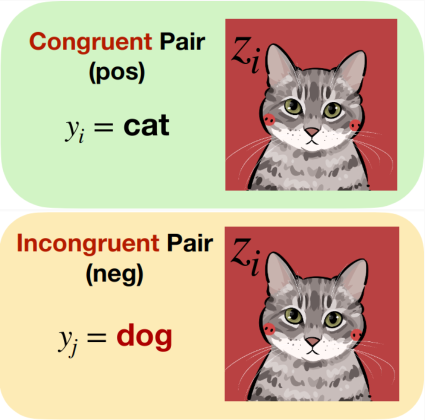


Mixing and demixing

Independent Component Analysis (ICA)

- **Decorrelates** features $\boldsymbol{Z}$ of the observed signal $\boldsymbol{X}$ into a source signal representation $\boldsymbol{S} = f_\psi(\boldsymbol{Z})$.
- Density $q(\boldsymbol{s}) = \prod_j q_j([\boldsymbol{s}]_j)$, where $[\boldsymbol{s}]_j$ is j-th **independent component (IC)** of $\boldsymbol{Z}([\boldsymbol{s}]_j$ is the j-th entry of vector $\boldsymbol{s}$.

So How Can We Find Such “Invariant” Features?

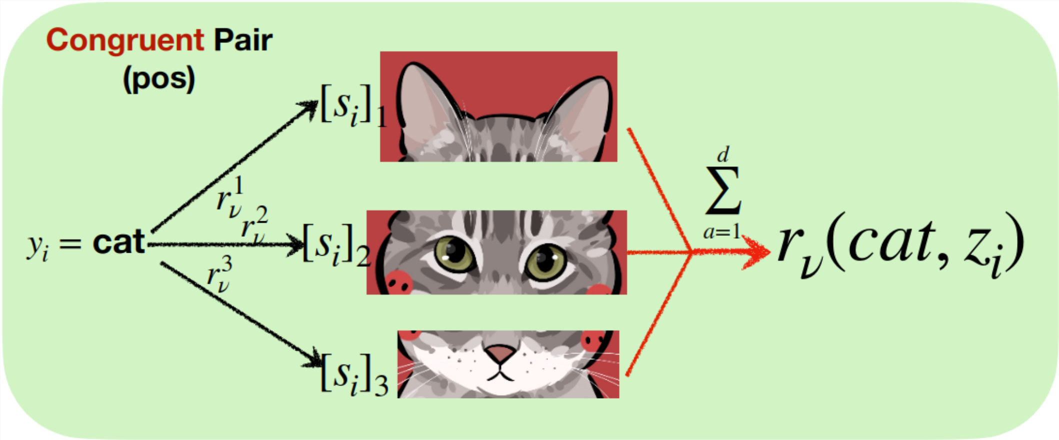


Nonlinear ICA

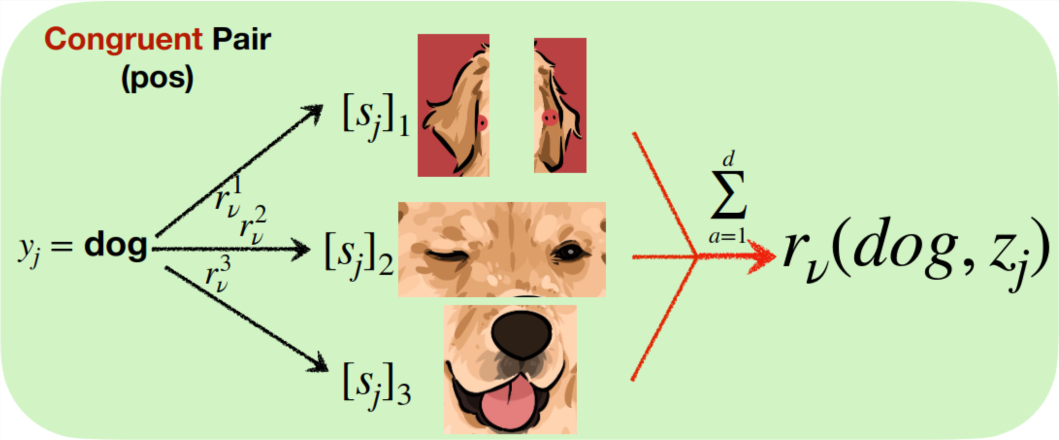
- Identification of Nonlinear ICA can be achieved, by requiring an **additional auxiliary label** y[1].
- NICA assumes that source signals are conditionally independent given y, then $f_\psi(\boldsymbol{z})$ can be identified using *generalized contrastive learning* **(GCL)**

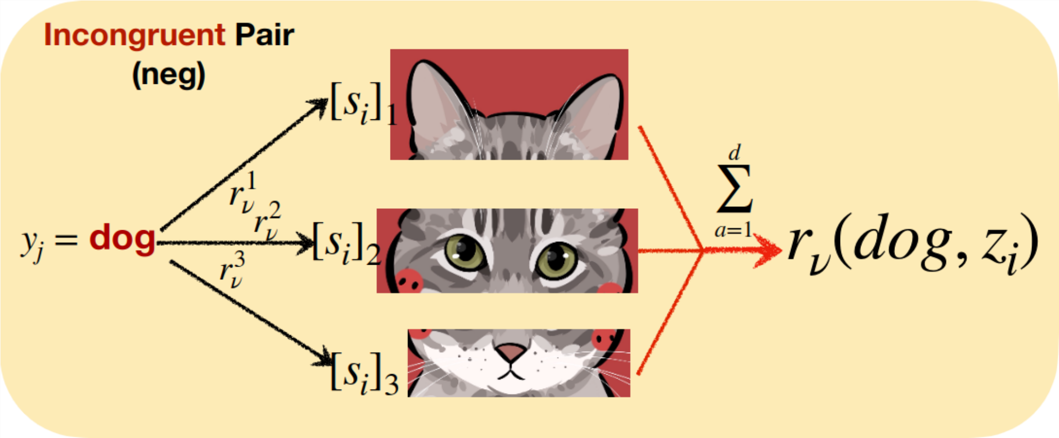
[1] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In AISTATS, pages 859–868, 2019.

Learn GCL with Our Special Guests

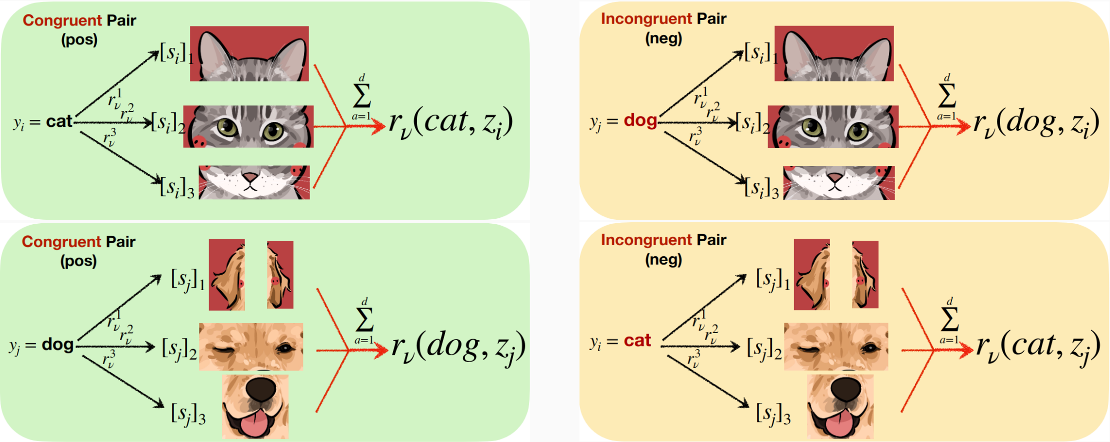


Learn GCL with Our Special Guests

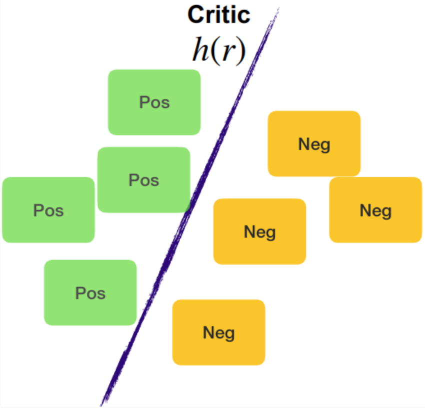



Incongruent Pair
(neg)
y_j = dog
r_ν^1
r_ν^2
r_ν^3
$[s_i]_1$
$[s_i]_2$
$[s_i]_3$
$\sum_{a=1}^{d}$
$r_\nu(dog, z_i)$

Learn GCL with Our Special Guests



Disentangling Representations with GCL



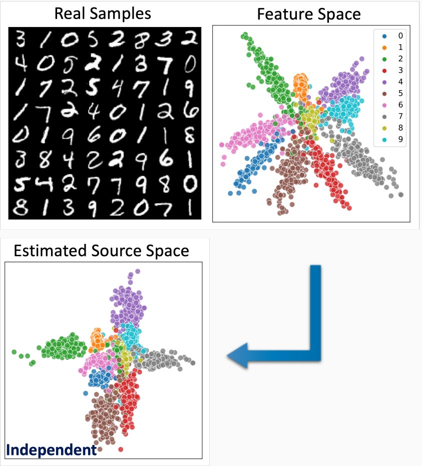
Generalized Contrastive Learning

- A generalized regression problem
- Critic function predicts whether **label y and representation z are correctly paired**, i.e., congruent

$$\arg\min_{f_\psi, r_\nu} \underbrace{\mathbb{E}_i[h(-r_\nu(y_i, \boldsymbol{z}_i))] + \mathbb{E}_{j\neq i}[h(r_\nu(y_j, \boldsymbol{z}_i))]}_{\mathcal{L}_{\text{GCL}}(f_\psi, r_\nu)}, \tag{1}$$

- Independent features achieved

Disentangling Representations with GCL



Generalized Contrastive Learning

- A generalized regression problem
- Critic function predicts whether **label y and representation z are correctly paired**, i.e., congruent

$$\arg\min_{f_\psi, r_\nu} \underbrace{\mathbb{E}_i[h(-r_\nu(y_i, \boldsymbol{z}_i))] + \mathbb{E}_{j\neq i}[h(r_\nu(y_j, \boldsymbol{z}_i))]}_{\mathcal{L}_{\mathrm{GCL}}(f_\psi, r_\nu)}, \tag{2}$$

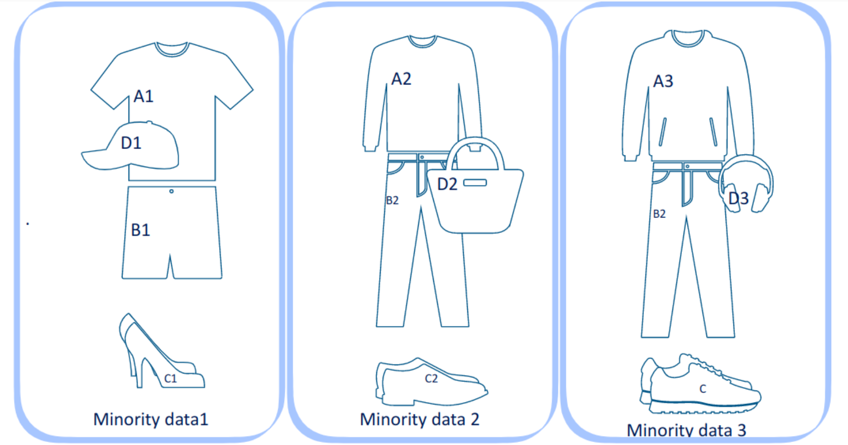
- Independent features achieved

We Can Greatly Improve Algorithm Efficiency for GCL

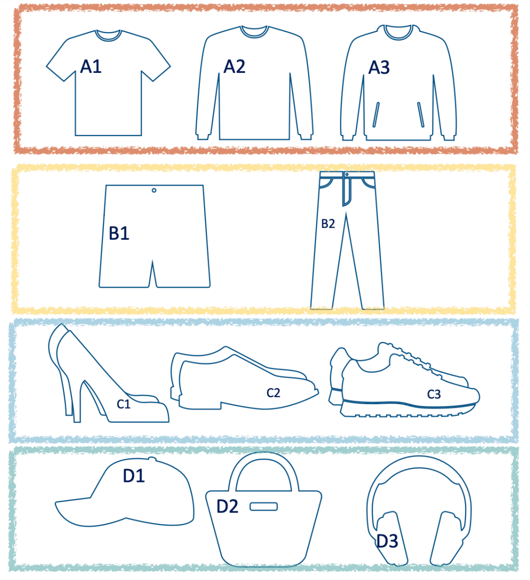
	Cat	Dog	Bird	Fish		Zebra
	Pos	Neg	Neg	Neg	...	Neg
	Neg	Pos	Neg	Neg	...	Neg

Energy-based GCL: Learning with group contrasts!

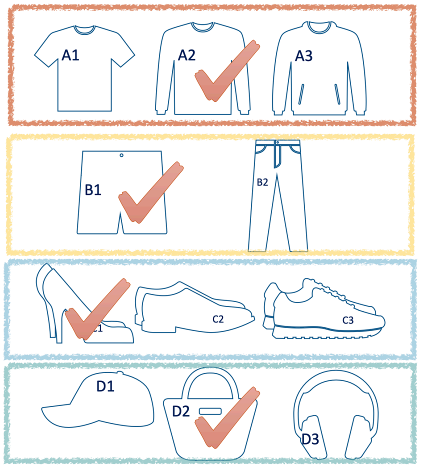
We Are Ready to Augment the Minority Groups



Organize the Closet by Independent Components

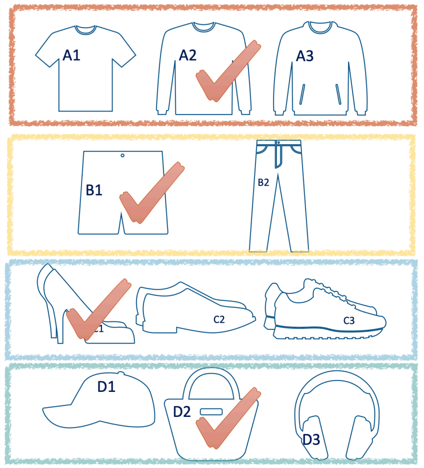


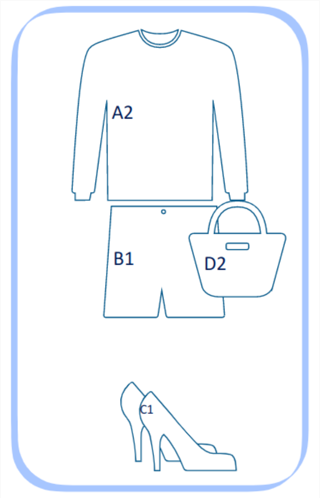
Even Better: a Decision-Free Closet!



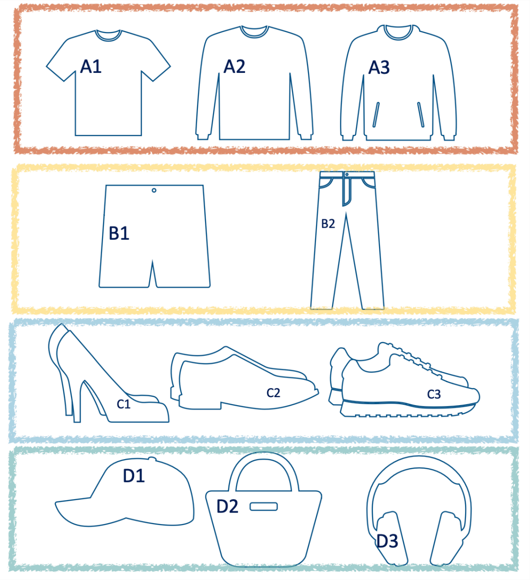
Randomly permute the coordinates within elements

Random Outfit Generator

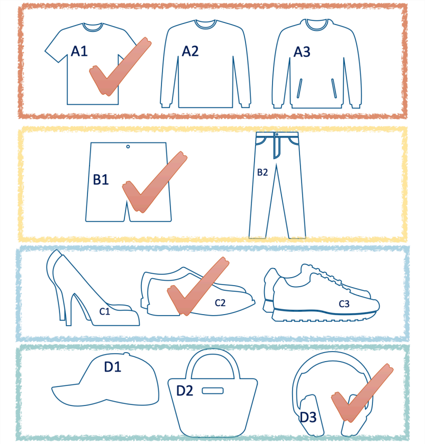




I Want More...

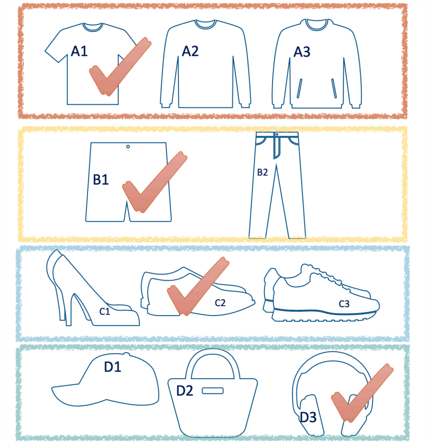


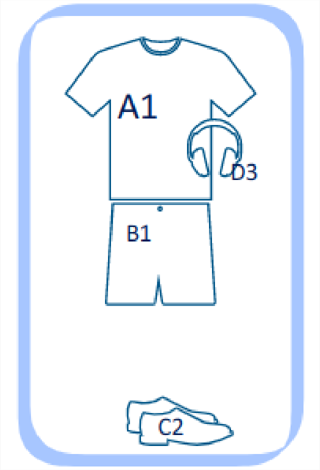
Try Again



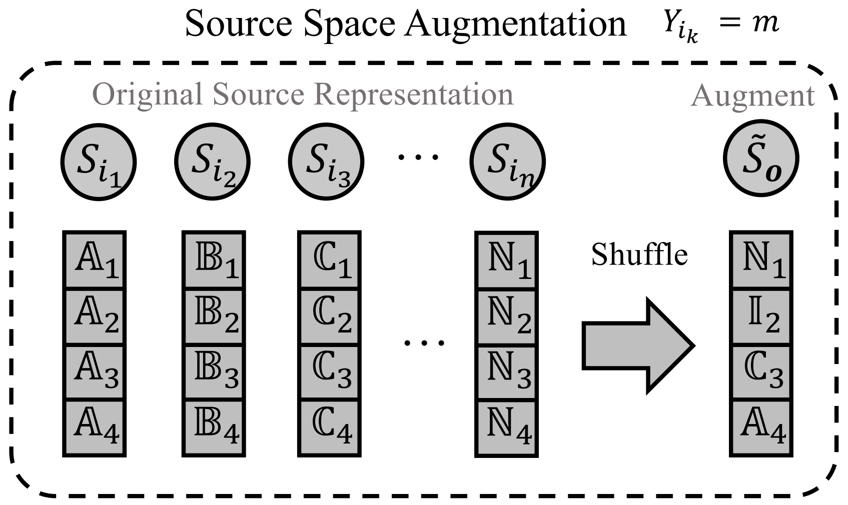
Randomly permute the coordinates within elements

Again and Again

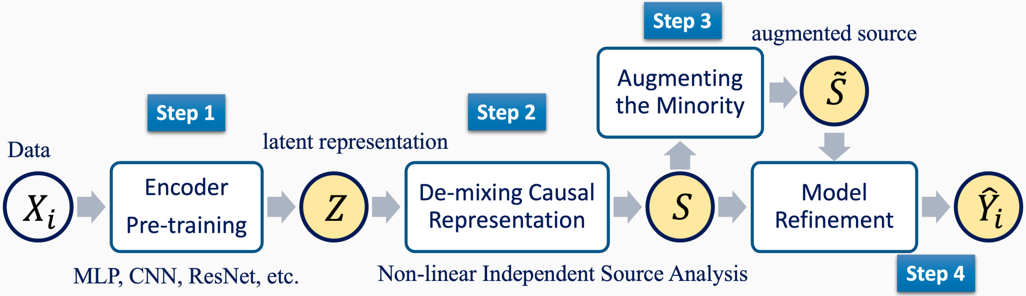




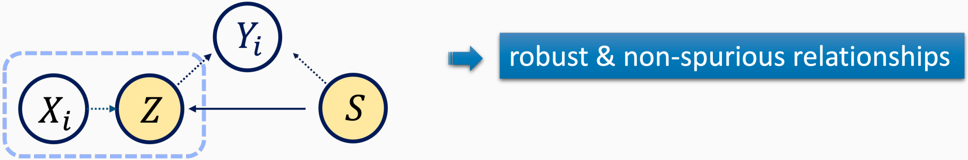
This is How Augmentation Works



ECRT Workflow Summary

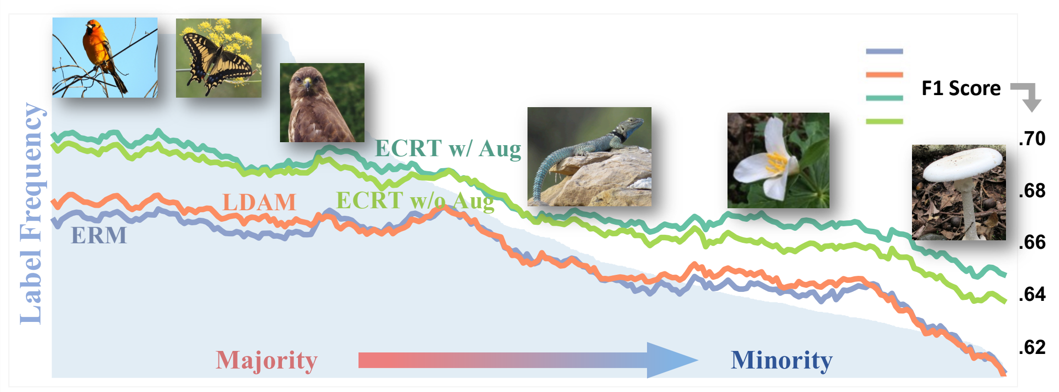


- Establishing causal invariance



- Enriching minority latent representations based on the feature independence

Experiments

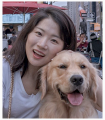


Experiments

Table 1: Comparison of performance on real-world datasets (↑ higher is better, ↓ lower is better).

	CIFAR100	iNaturalist	TinyImageNet	ArXiv
	Top-1↑	Top-1↑	Top-1↑	Acc↑
ERM	49.29	66.73	58.52	44.64
IW	43.97	67.63	60.50	46.02
GAN	47.64	67.40	60.69	45.42
VAT	46.47	67.06	59.69	45.82
FOCAL	43.32	66.63	58.27	46.01
LDAM	50.46	67.39	58.18	45.04
ECRT	**53.00**	**69.01**	**64.40**	**48.33**

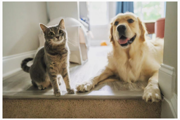
Our Team



Junya Chen[1,2] Zidi Xiu[1], Chenyang Tao[1] Ricardo Henao[1]
Benjamin Goldstein[1] Lawrence Carin[1,3]
[1] Duke; [2] Fudan; [3] KAUST
{firstname.lastname}@duke.edu

Guest Stars


Lychee and Goki as themselves