

# Contents and speakers

**Overview of trustworthiness** (Jindong Wang, 10min)

**Robust machine learning**  
(Jindong Wang, 40min)

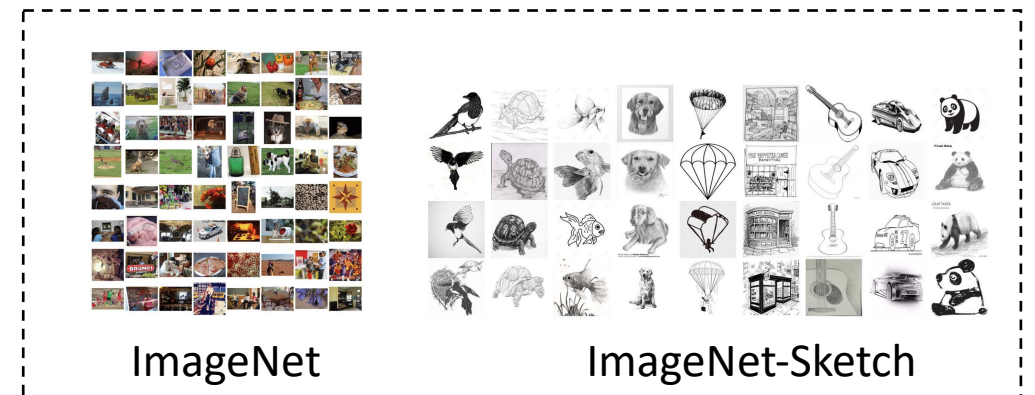
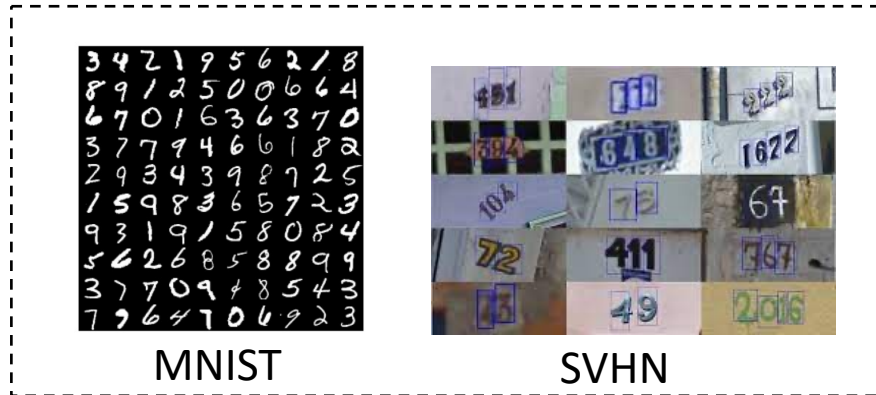
**Out-of-distribution generalization**  
(Haohan Wang, 40min)

**Interpretability**  
(Haohan Wang, 40min)

**Trustworthiness in the era of large models** (Jindong Wang, 40min)

# Robustness (domain generalization and more)

- Usually studied over benchmarks (that are constructed by the research community)



# Variants of Cross-Domain Robustness

- Different study scenarios defined over partition of the data

- Domain adaptation

- Using (unlabeled) data from the test domain
- (Ben-David 2007)

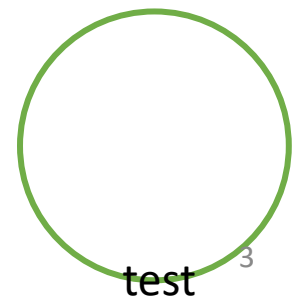
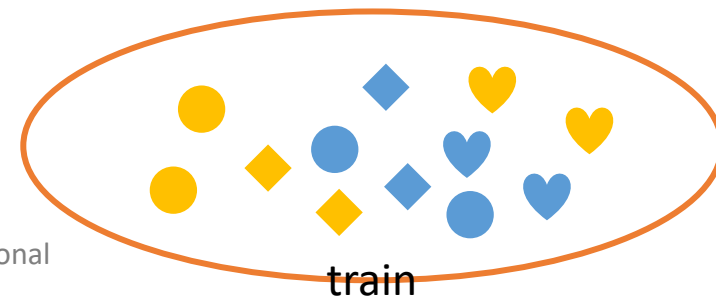
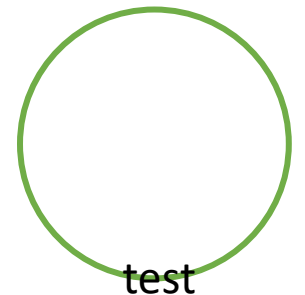
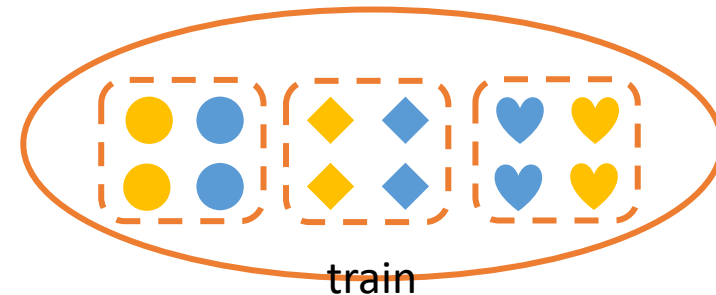
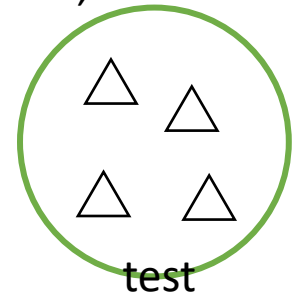
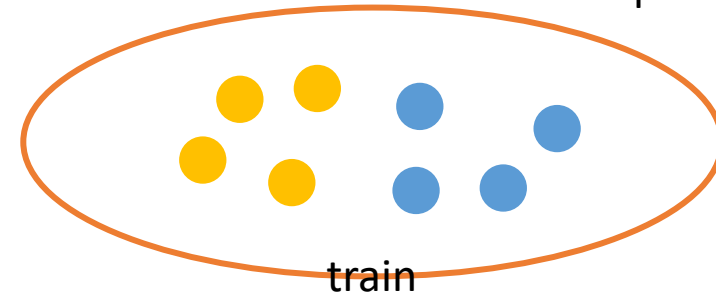
- Domain generalization

- Using partitions of data in the train domains
- (Muandet et al., 2013)

- Cross-domain generalization

- Not using any extra information
- (Wang et al, 2019)
- Used as the setup of this talk

shape-domain; color-label



# For two arbitrary domains/distributions?

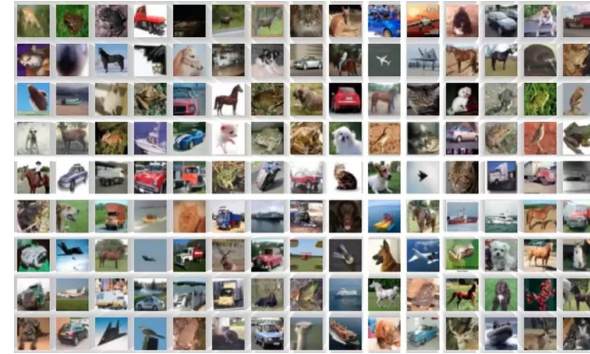
- Maybe not



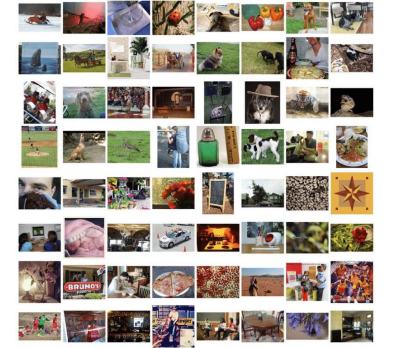
MNIST



FashionMNIST



CIFAR



ImageNet

So, there should be some regulations about what datasets can be used to study cross-domain robustness.

# Conventional Machine Learning Generalization

- On the theoretical end
  - Preliminary:
    - A standard generalization error bound of supervised machine learning

$$\varepsilon(\theta) \leq \hat{\varepsilon}(\theta) + \phi(\Theta, n, \delta)$$

Expected error during test

Empirical error during training

Other technical terms

- hypothesis space,
- number of samples
- probability of this bound

# Domain Adaptation: a study across **similar** but **different** domains

- On the theoretical end
  - “**Similar** but **different**” is barely rigorously defined
  - Domain Adaptation Bounds
    - (Ben-David et al., 2010) (Mansour et al., 2009) (Germain et al., 2016) (Zhang et al., 2019) (Dhouib et al., 2020)

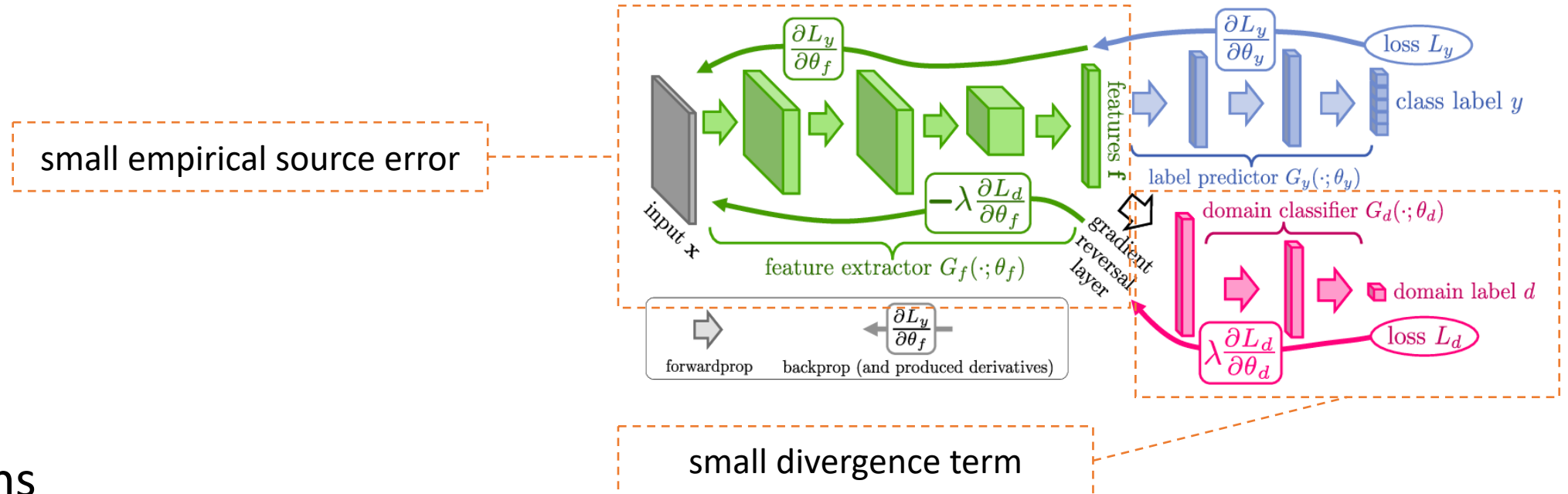
$$\epsilon_{\mathbf{P}_t}(\theta) \leq \hat{\epsilon}_{\mathbf{P}_s}(\theta) + \phi(\Theta, n, \delta) + D_{\Theta}(\mathbf{P}_s, \mathbf{P}_t) + \lambda$$

Divergence Between Distributions (estimable term)

Learnable nature of  
the problem  
(not estimable)

# Domain Adaptation Domain Adversarial Neural Networks (Ganin et al 2016)

- Design Rationale



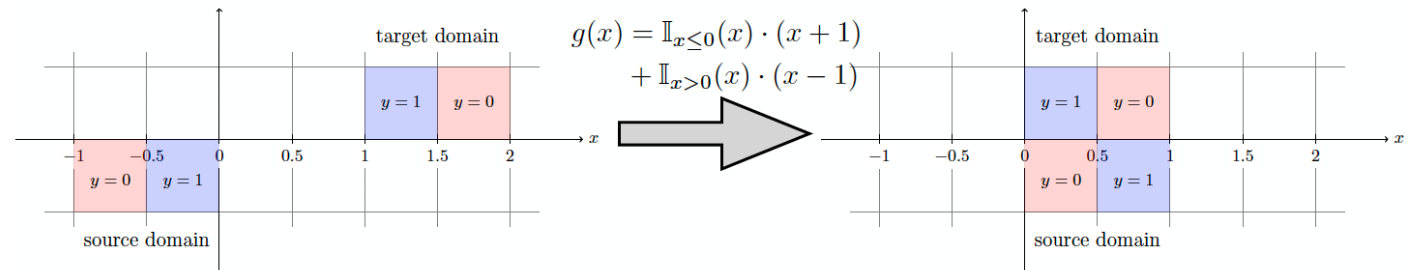
- Limitations

- methods fail to generalize in certain closely related source/target pairs, e.g., digit classification from **MNIST** to **SVHN** (Ganin et al 2016)
- Hypothesis of misaligned labelling function
  - (Zhao et al 2019) (Wu et al 2019)

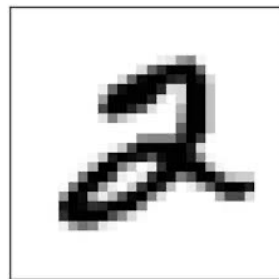
# Domain Adaptation

## the hypothesis of misaligned labelling function

- Domain adversarial neural network will not work if the labelling function shifts from training domain to test domain
  - (Zhao et al 2019)



- However, a human might prefer that samples across **similar** but **different** domains will have a shared labelling function
  - After all, there is a reason both of these digits are 2!



MNIST

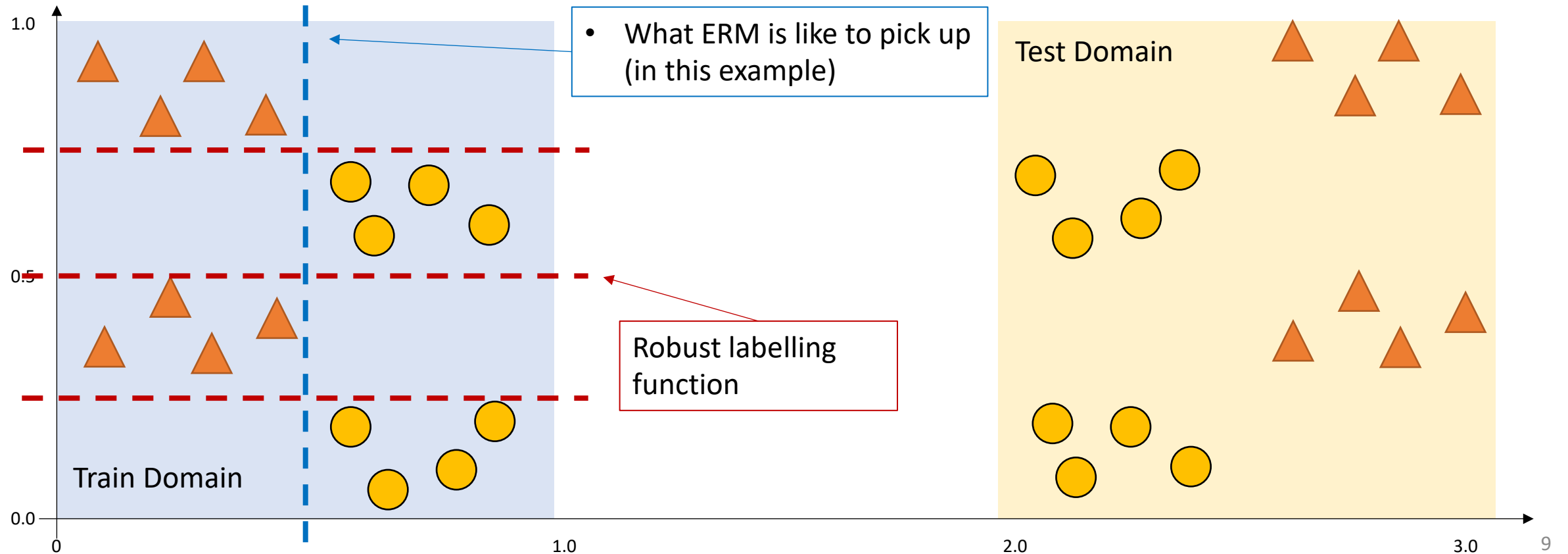


SVHN

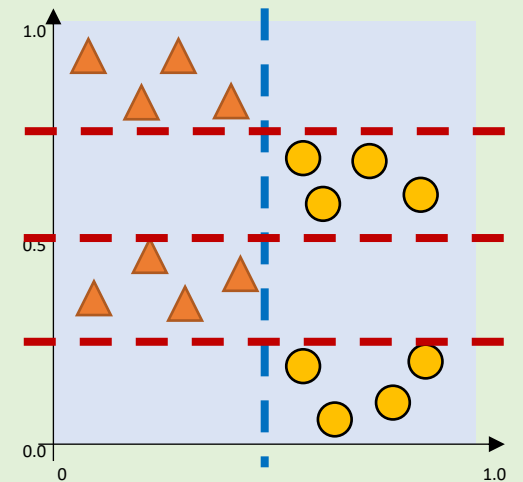
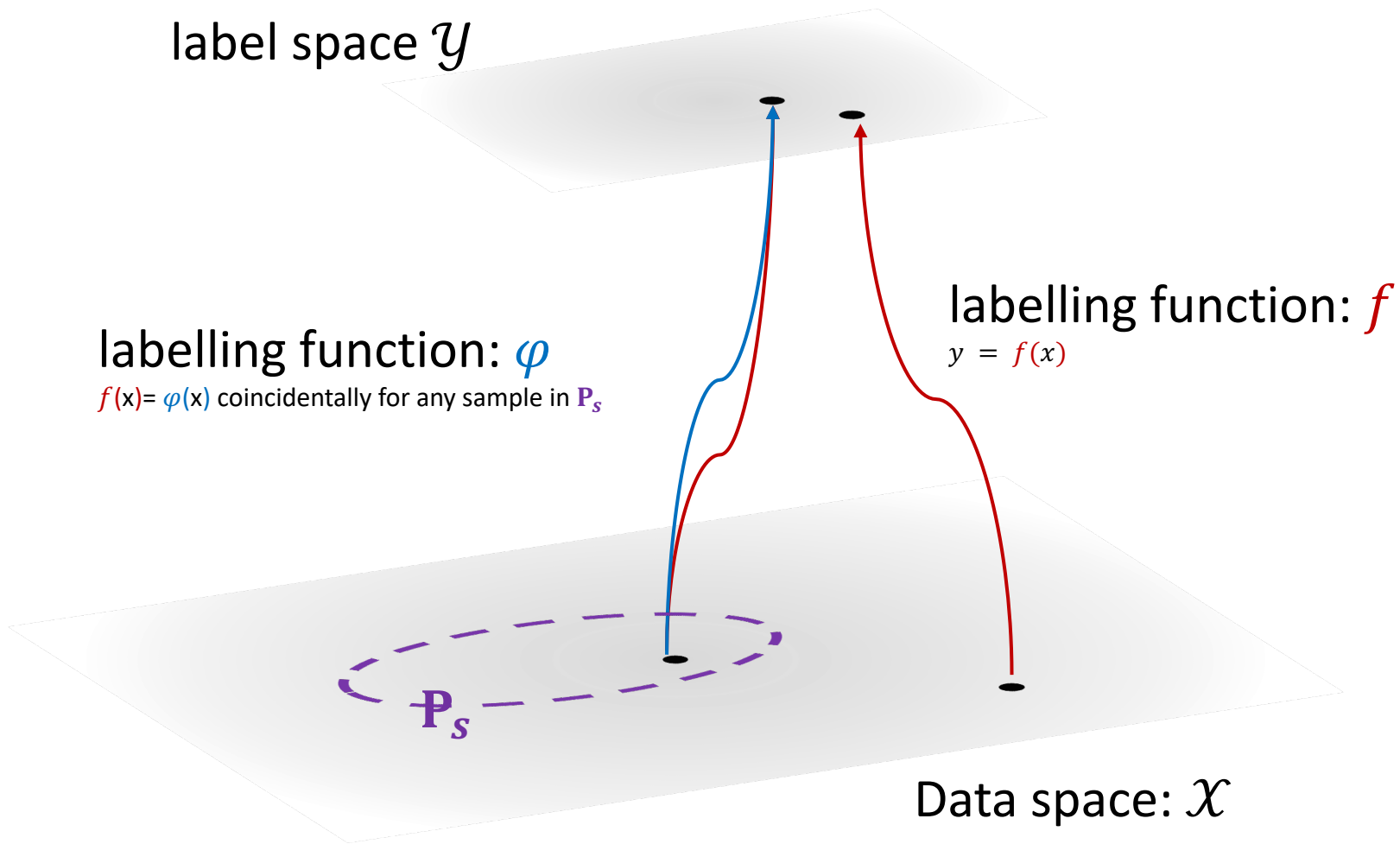


# Similar but Different: intuitively

- **Similar**: there is a shared labelling function
- **Different**: the training domain has an additional labelling function

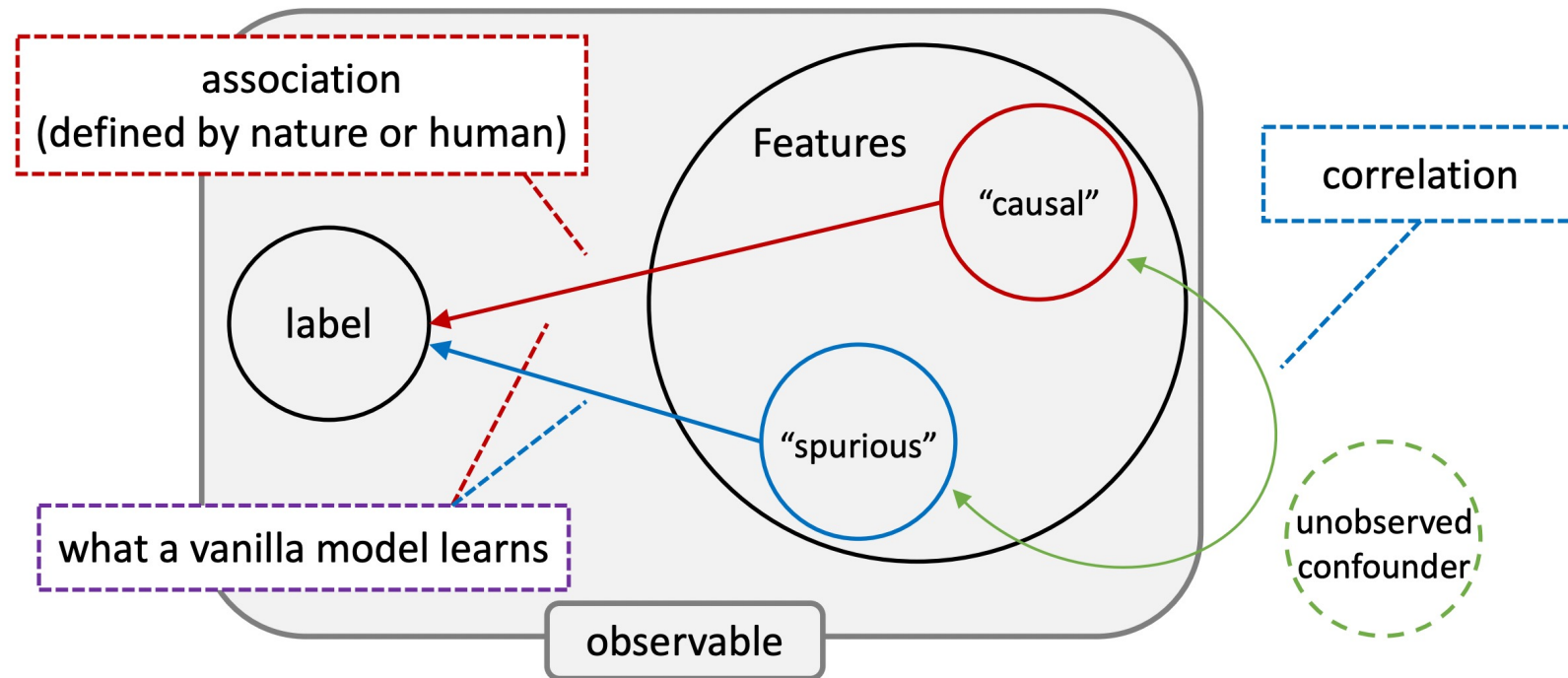


# Similar but Different: formally



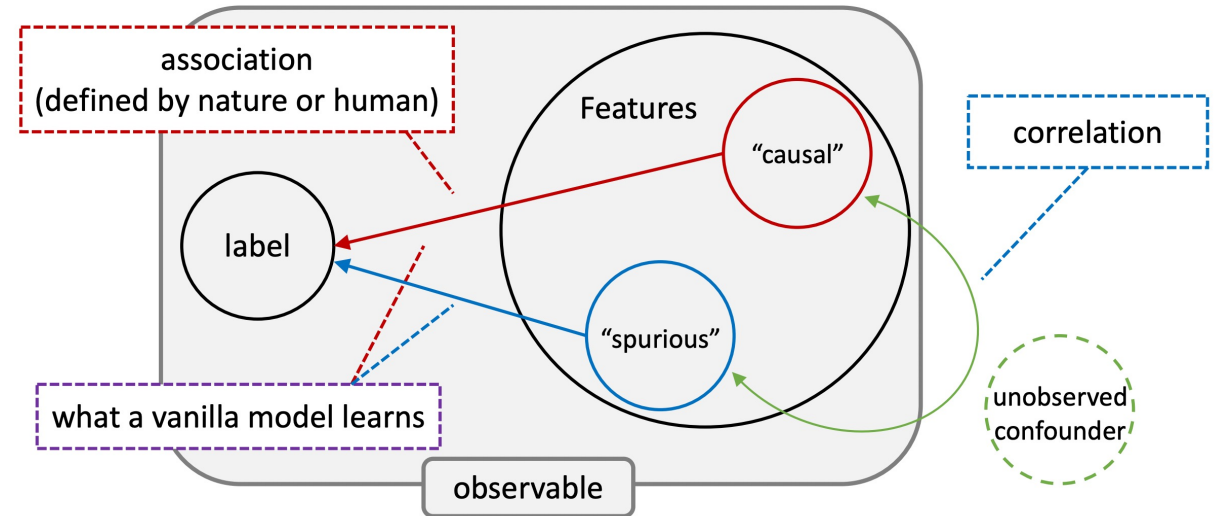
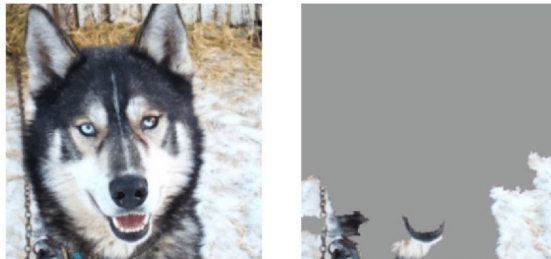
# Cross-domain Generalization

- Training models from one domain and test it in the other
  - These two domains are **similar** but **different**
    - **Similar**: there is a shared labelling function
    - **Different**: the training domain has an additional labelling function



# Cross-domain Generalization

- This understanding of data is cross the understanding of domain generalization
  - Domain adaptation
  - Domain generalization
  - Beyond Domain Generalization
  - Bias-in-Data



- (adversarial robustness)

# Generalization Error Bound of Robust ML

## Theorem (the Curse of Universal Approximation)

(Informal) Under multiple assumptions, with probability at least  $1 - \delta$ , we have

$$\varepsilon_{\mathbf{P}_t}(\theta) \leq \hat{\varepsilon}_{\mathbf{P}_s}(\theta) + \phi(\Theta, n, \delta) + c(\theta)$$

$$c(\theta) = \frac{1}{n} \sum_{(x,y) \in (X,Y)_{\mathbf{P}_s}} \mathbb{I}[\theta(x) \neq y] r(\theta, \mathcal{S}(\varphi, x))$$

the accuracy gain because  $\theta$  learns  $\varphi$

usually guided by additional domain knowledge

a robust model

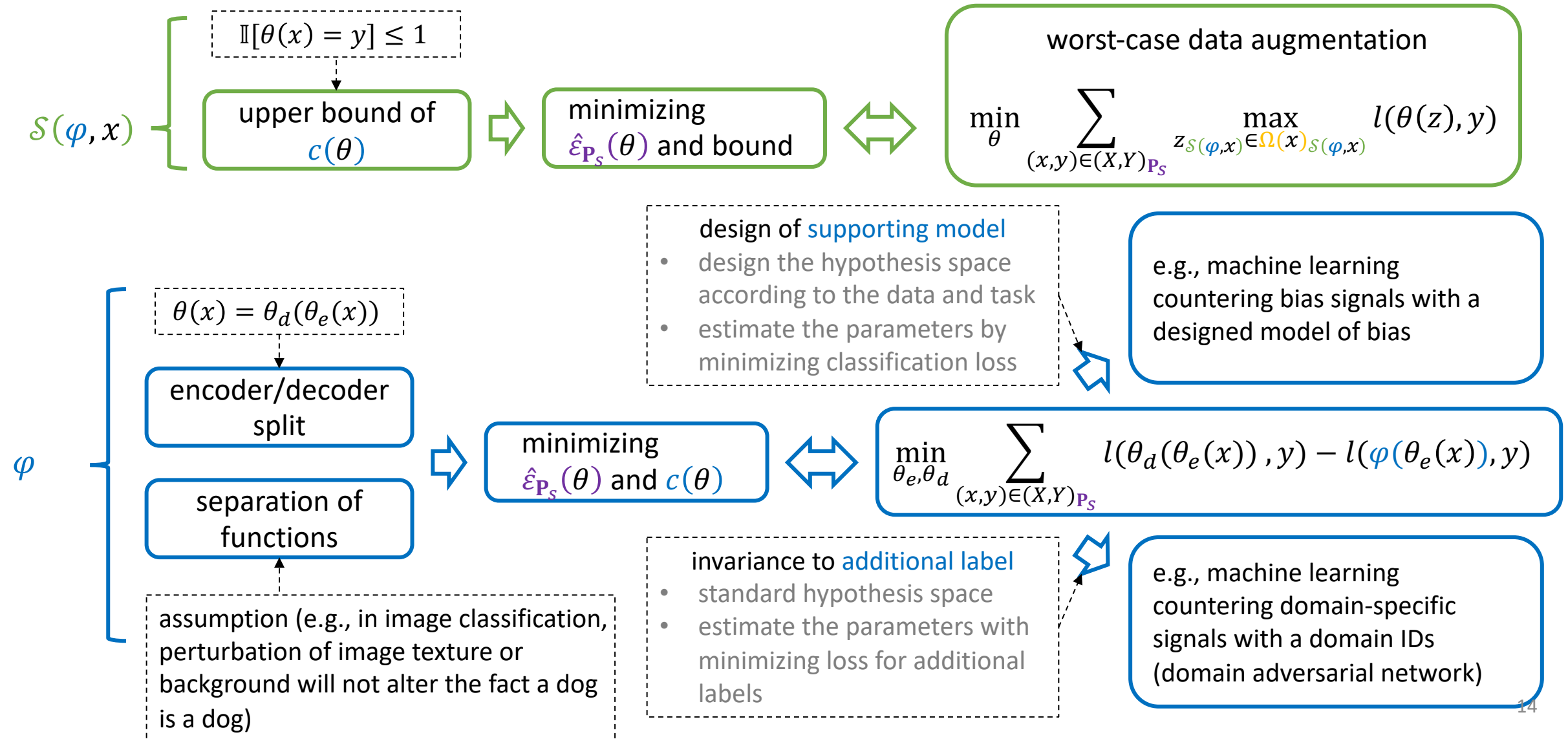
small  $\hat{\varepsilon}_{\mathbf{P}_s}(\theta)$

small  $c(\theta)$

$\mathcal{S}(\varphi, x)$  (i.e., the features the superficial function uses) is given

$\varphi$  is given

# Principled Understanding of Robust ML



# Principled Understanding of Robust ML

the Curse of Universal Approximation

$$\varepsilon_{P_t}(\theta) \leq \hat{\varepsilon}_{P_s}(\theta) + \phi(\Theta, n, \delta) + c(\theta)$$

robust model

small  $\hat{\varepsilon}_{P_s}(\theta)$

small  $c(\theta)$

1. Worst-case Data Augmentation (Adversarial Training)

- Regular data augmentation
- With alignment regularization (consistency loss)

2. Regularizing the hypothesis

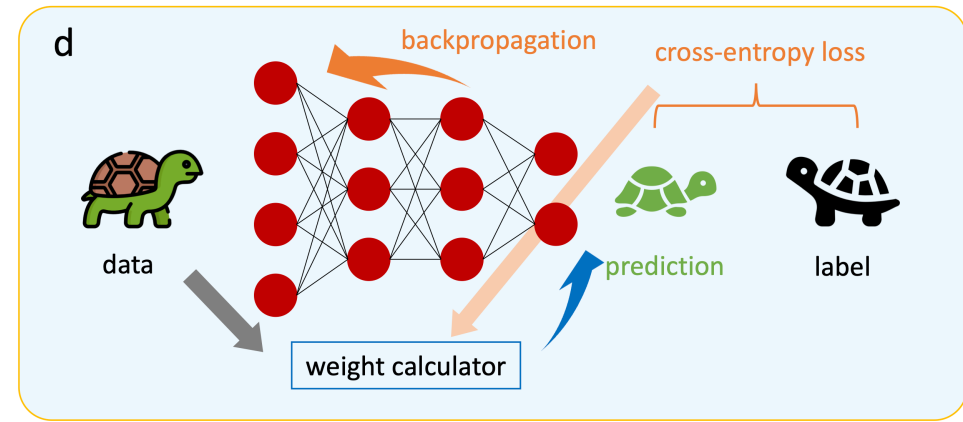
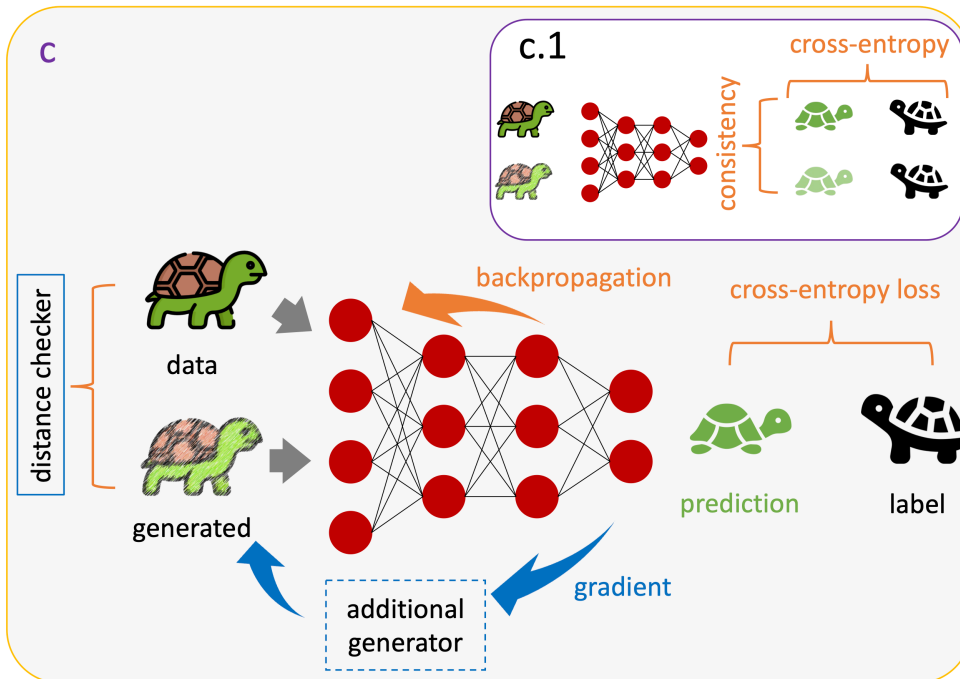
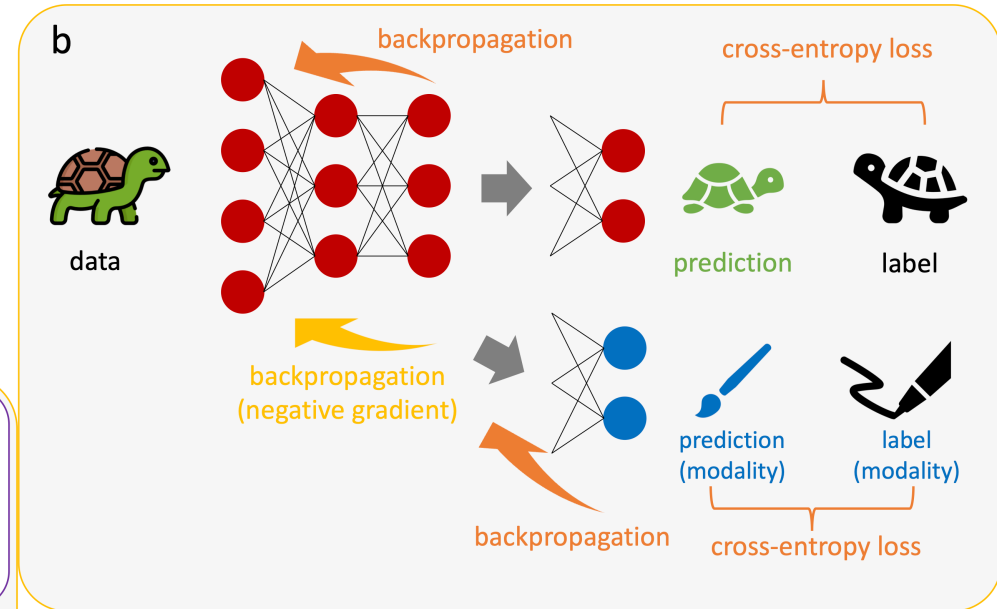
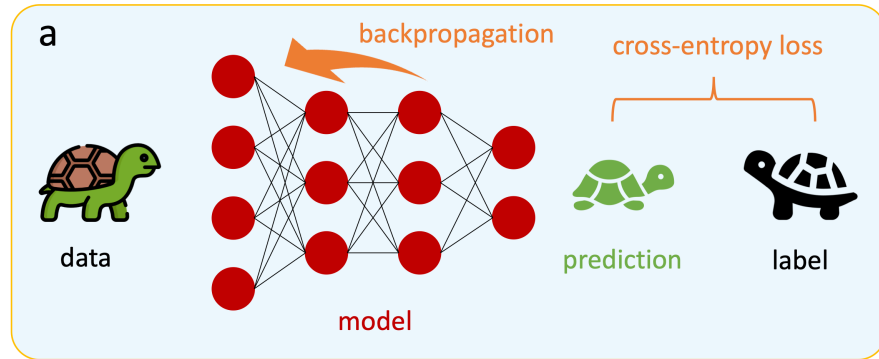
- With assumed function space of invariance
  - Learning by countering superficial/spurious features, debiasing
- With assumed label space of invariance
  - Learning embedding to fool an additional classifier, e.g., domain adversarial neural network

3. Worst-case Sample Reweighting (group-DRO Methods)

- Reweighting samples with learned functions

# Principled Understanding of Robust ML

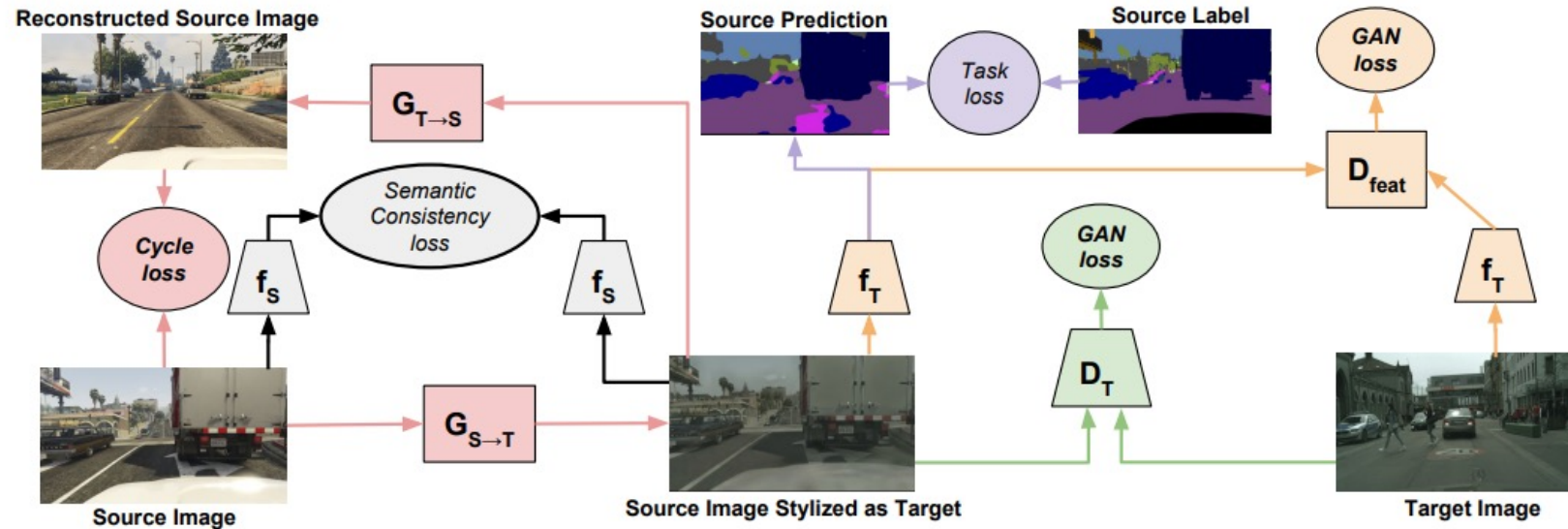
- Visual Summary





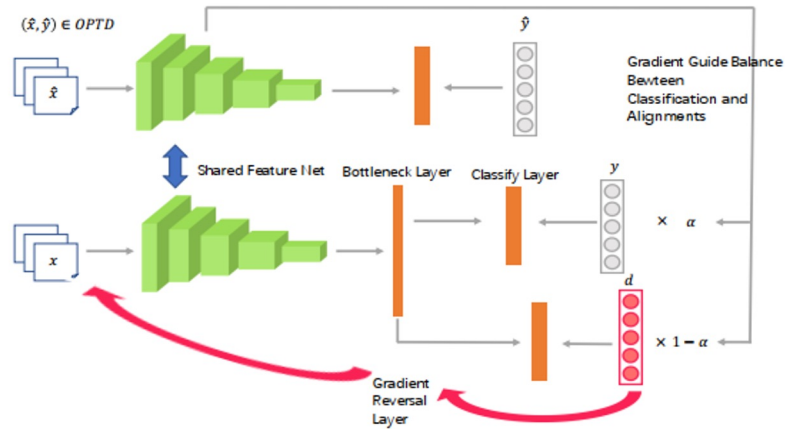
# Data Augmentation Based Methods

- Generating the data for the domain of interest
  - With domain information, e.g.,
    - Cycle-consistent adversarial adaptation



# Model selection and Optimization for OOD

- Mixup-guided optimization and selection: A mixup-guided solution
  - OPTD: an OOD dataset to balance classification loss and regularized items
  - VALD: an OOD dataset to select the best models



$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \\ \tilde{y} = y_i = y_j, \\ \text{where } d_i \neq d_j.$$

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \\ \tilde{y} = \lambda y_i + (1 - \lambda)y_j, \\ \text{where } d_i = d_j.$$

$$\omega^* = \arg \max_{\omega \in \Delta^{m-1}} (G\omega)^T (I(\ell_{optd} > 0)g_{optd} + I(\ell_{optd} = 0)G1/m), \\ \text{s.t. } (G\omega)^T g_j \geq I(J \neq \emptyset)(g_{optd}^T g_j), \forall j \in \bar{J} - J^*, \\ (G\omega)^T g_j \geq 0, \forall j \in J^*.$$

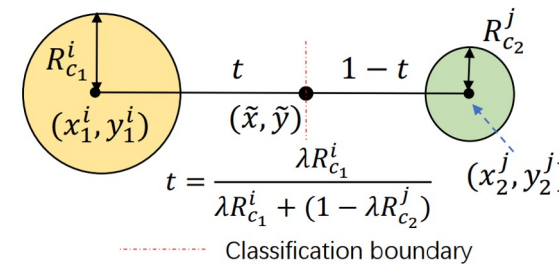
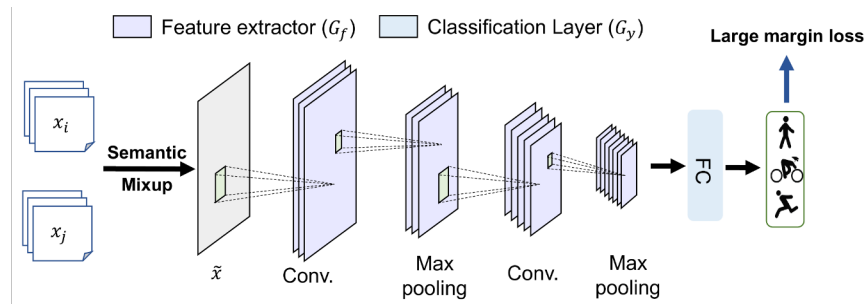
Methods	DSADS					USC-HAD					PAMAP2				
	0	1	2	3	AVG	0	1	2	3	AVG	0	1	2	3	AVG
ERM	89.69	81.45	81.05	78.20	82.60	80.33	59.88	74.15	73.93	72.07	87.28	73.10	49.03	78.76	72.04
ANDMask	85.35	73.07	85.04	82.06	81.38	79.51	61.53	76.32	65.52	70.72	88.22	79.11	53.35	83.22	75.97
GILE	79.67	75.00	77.00	67.00	74.65	78.67	63.00	77.00	61.67	70.08	83.33	68.67	44.00	76.67	68.25
DANN	87.54	81.27	78.42	83.03	82.57	81.33	64.02	72.91	66.37	71.16	88.93	75.60	47.35	86.78	74.66
<b>DANN+Ours</b>	<b>93.33</b>	<b>88.77</b>	<b>91.75</b>	<b>84.78</b>	<b>89.66</b>	<b>81.98</b>	<b>64.32</b>	<b>74.84</b>	<b>78.40</b>	<b>74.89</b>	<b>89.23</b>	<b>81.36</b>	<b>61.71</b>	<b>89.28</b>	<b>80.40</b>

- Lu et al. Towards Optimization and Model Selection for Domain Generalization: A Mixup-guided Solution. KDD workshop 2023. <https://arxiv.org/abs/2209.00652>

# Data augmentation: SDMix

- SDMix: Semantic-Discriminative Mixup

- Semantic-aware Mixup: overcome the semantic inconsistency brought by domain differences
- Enhancing Discrimination: introduce large margin loss to overcome discriminative slackness



$$\tilde{x} = \lambda x_1^i + (1 - \lambda) x_2^j,$$

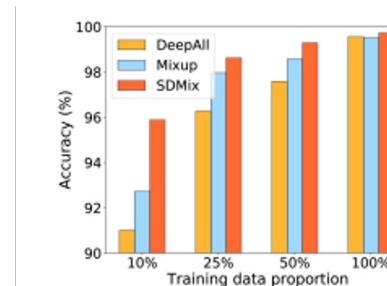
$$\tilde{y} = t y_1^i + (1 - t) y_2^j,$$

$$\lambda \sim \text{Beta}(\alpha, \alpha),$$

$$R_c^i = \max_{x \in \mathcal{D}_c^i} d(x, \mu_c^i),$$

$$R_c^t = \frac{\sum_{x \in \mathcal{D}_c^i} d(x, \mu_c^i)}{|\mathcal{D}_c^i|},$$

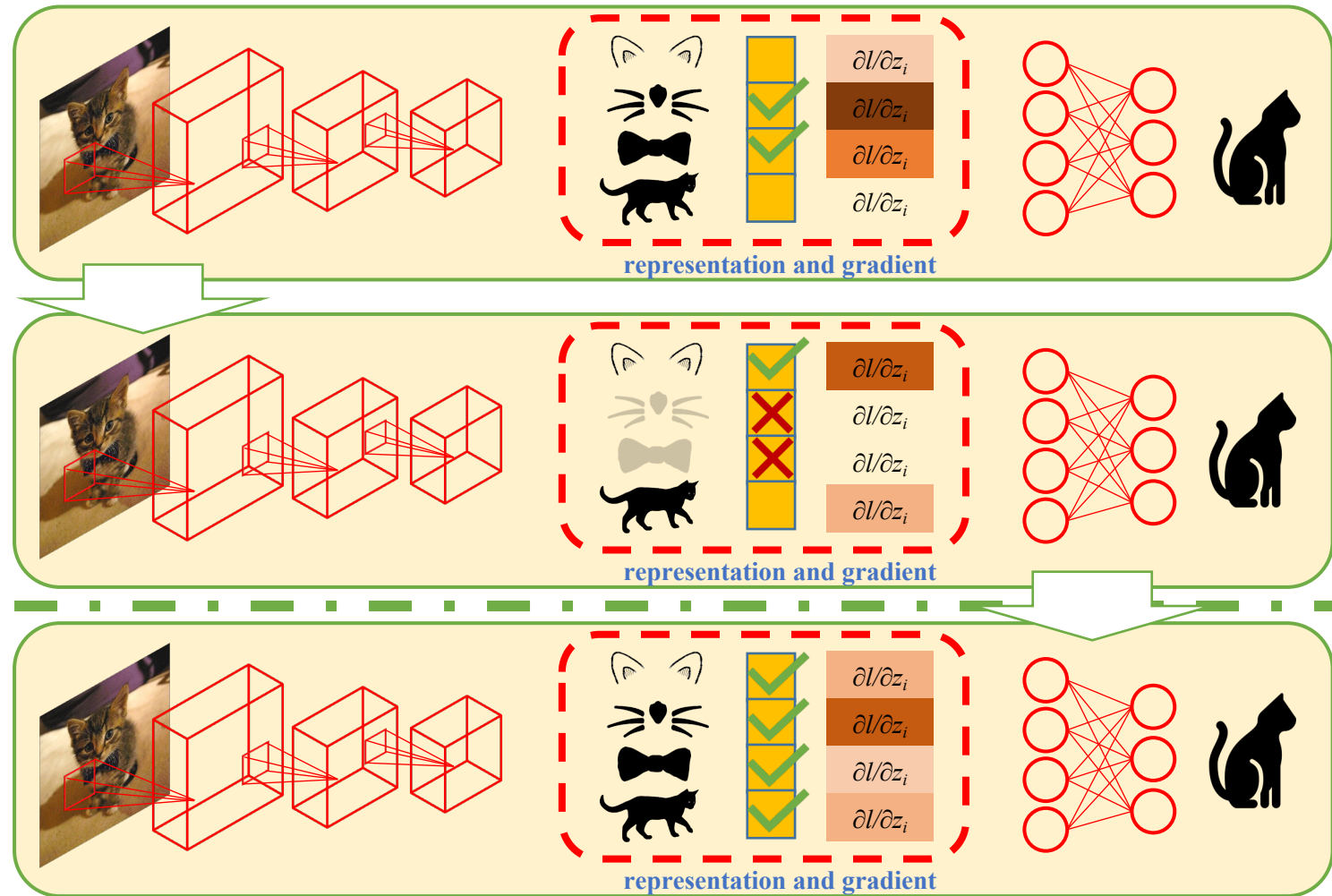
Source	Target	DeepAll	DANN	CORAL	ANDMask	GroupDRO	RSC	Mixup	SDMix
DSADS, USC, PAMAP2	UCI-HAR	<u>46.06</u>	39.10	44.44	43.22	33.20	45.28	40.24	<b>46.41</b>
USC, UCI-HAR, PAMAP2	DSADS	29.73	39.46	26.35	41.66	<u>51.41</u>	33.10	37.35	<b>52.66</b>
DSADS, USC, UCI-HAR	PAMAP2	43.84	36.61	32.93	40.17	33.80	<u>45.94</u>	23.12	<b>53.65</b>
DSADS, UCI-HAR, PAMAP2	USC	45.33	41.82	29.58	33.83	36.74	39.70	<u>47.39</u>	<b>53.54</b>
AVG	-	<u>41.24</u>	39.25	33.32	39.72	38.79	41.01	37.03	<b>51.57</b>



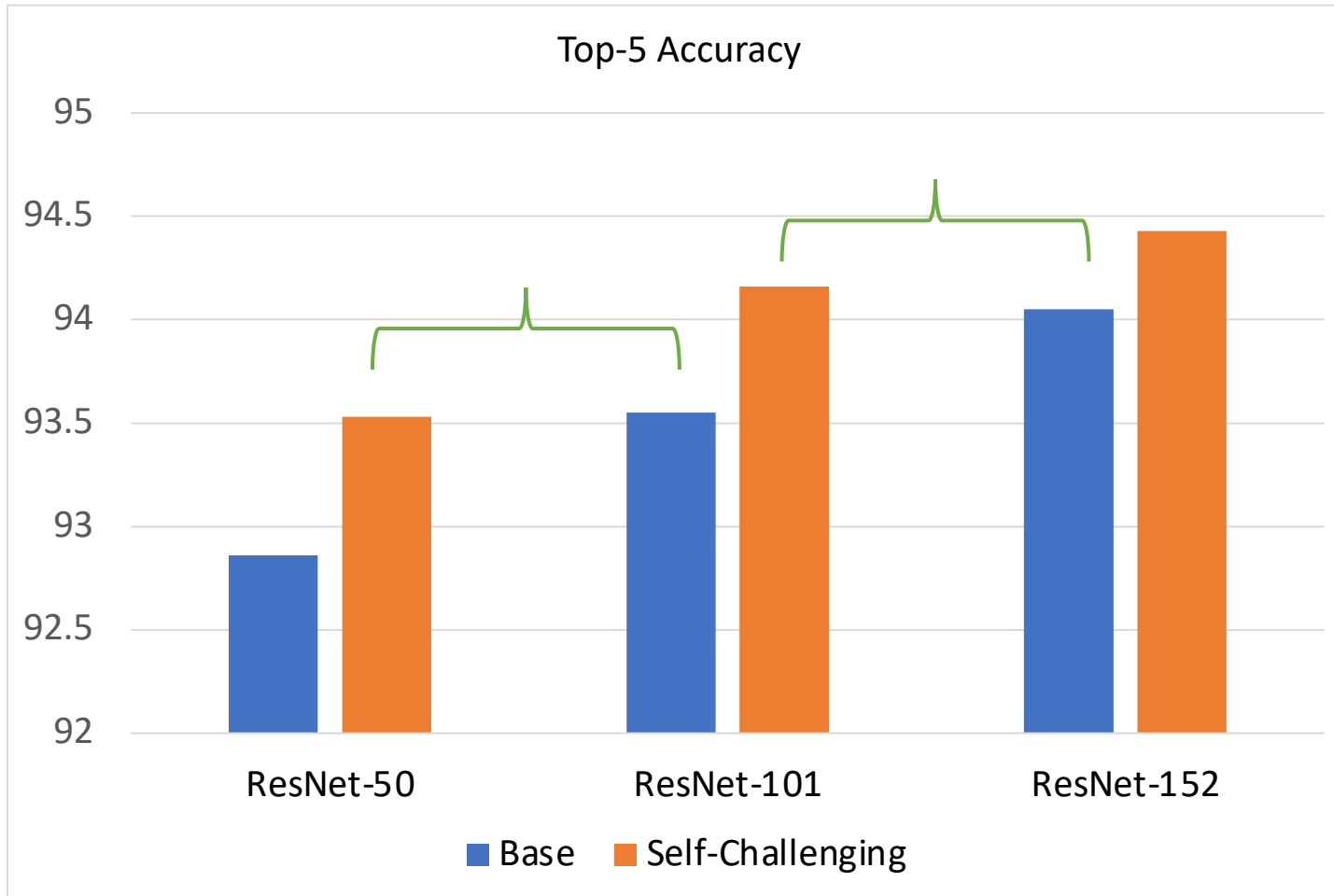
- Lu et al. Semantic-Discriminative Mixup for Generalizable Sensor-based Cross-domain Activity Recognition . IMWUT 2022. <https://arxiv.org/abs/2206.06629>

# Data Augmentation Based Methods

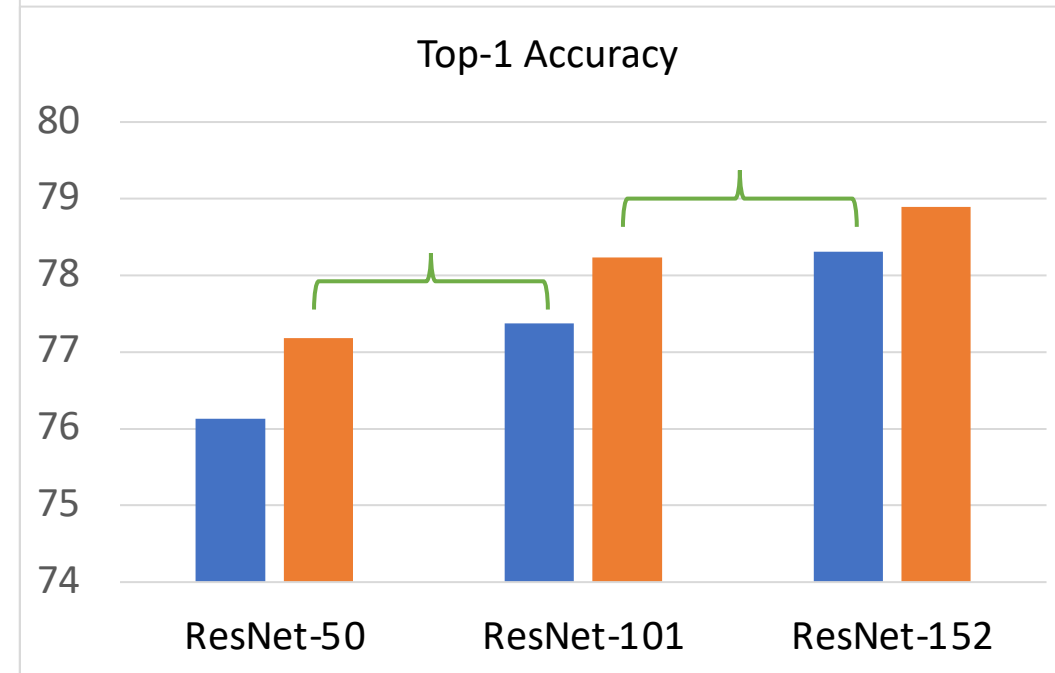
- Generating the data for the domain of interest
  - More general approach (without specific domain information)
  - We iteratively force the model to predict without the features it considers predictive in the previous iteration.



# Results: It Improves Accuracy on ImageNet Classification

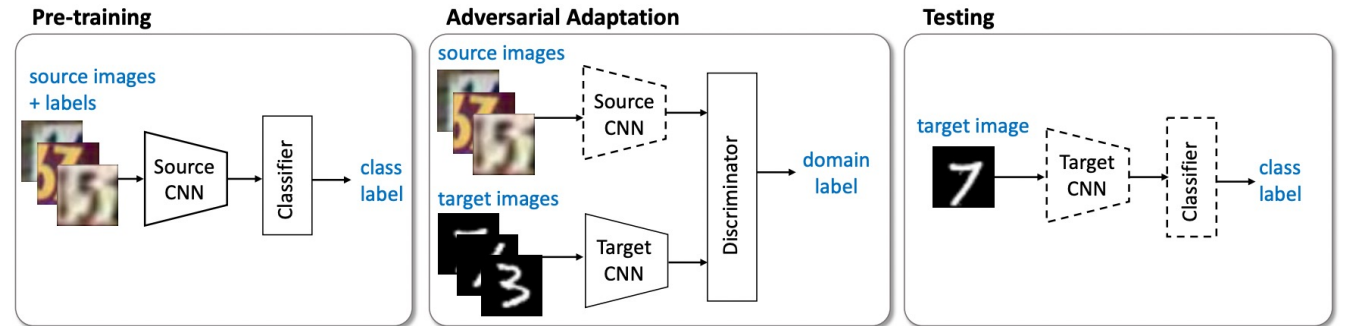
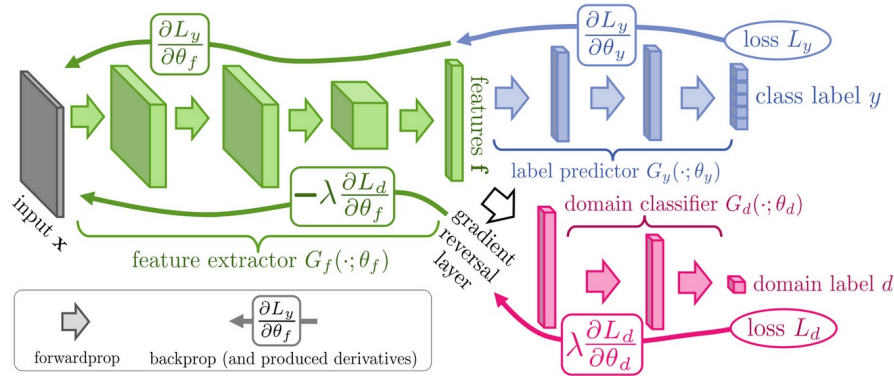


- Self-challenging can improve ImageNet classification accuracy with a margin that bridges the gap of model sizes.



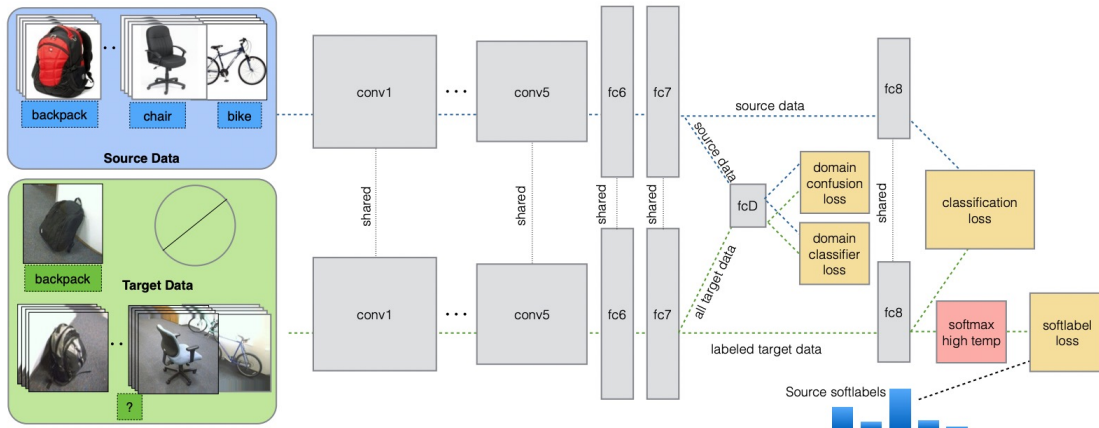
# Regularization Based Methods

- Regularizing for domain invariance

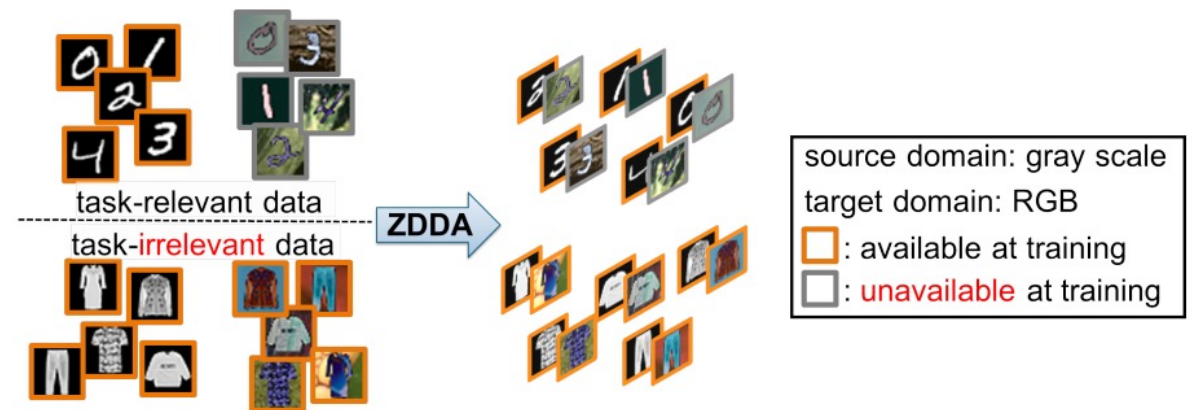


Adversarial Discriminative Domain Adaptation

Adversarial Discriminative Domain Adaptation  
(Split the feature extractor into two copies)



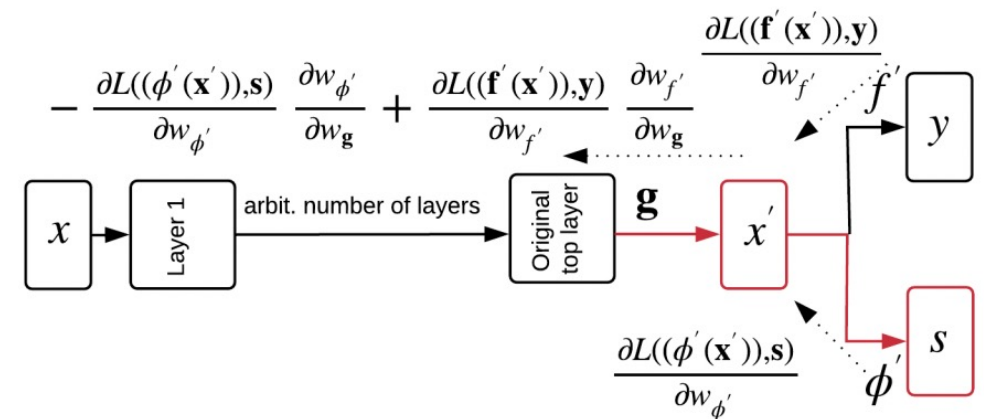
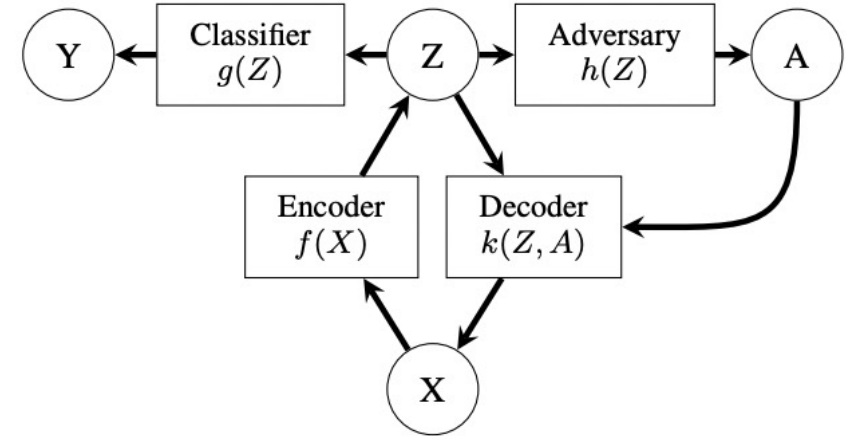
Simultaneous Deep Transfer Across Domains and Tasks  
(More information to align for domain classifier)



Zero-Shot Deep Domain Adaptation  
(Train domain classifier with additional data)

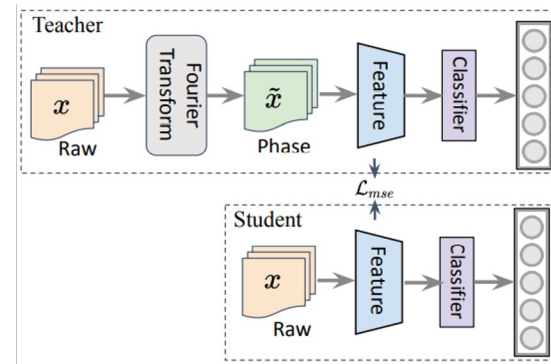
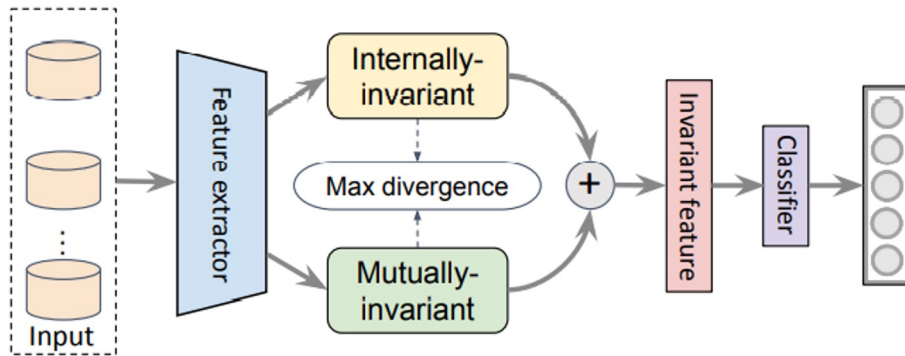
# Regularization Based Methods

- Extensions to studying machine learning fairness
  - Augmentation based methods
    - Learning Adversarially Fair and Transferable Representations
  - Regularization-based methods
    - One-network adversarial fairness.



# Algorithm: DIFEX for domain-invariant features

- DIFEX: Domain-Invariant Feature Exploration
  - What exactly are the domain-invariant features?
    - Internally-invariant features: captures the intrinsic knowledge of the signal
    - Mutually-invariant features: stays invariant w.r.t. other domain data

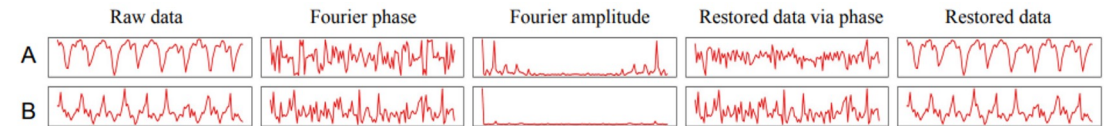


$$\mathcal{F}(\mathbf{x})(u, v) = \sum_{h=1}^{H-1} \sum_{w=0}^{W-1} \mathbf{x}(h, w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)}$$

$$\mathcal{P}(x)(u, v) = \arctan \left[ \frac{I(x)(u, v)}{R(x)(u, v)} \right]$$

$$\mathcal{L}_{align} = \frac{2}{N \times (N - 1)} \sum_{i \neq j} \|\mathbf{C}^i - \mathbf{C}^j\|_F^2$$

Source	Target	ERM	DANN	CORAL	Mixup	GroupDRO	RSC	ANDMask	SWAD	DIFEX
L,S,V	C	93.64	94.49	96.33	96.18	<b>97.81</b>	<u>96.89</u>	93.50	94.56	96.61
C,S,V	L	60.05	64.34	64.42	63.55	62.24	62.91	<u>64.83</u>	61.78	<b>67.21</b>
C,L,V	S	68.46	67.15	68.65	68.86	<u>69.23</u>	69.07	63.28	67.82	<b>74.31</b>
C,L,S	V	<u>74.02</u>	72.69	70.73	71.95	70.73	70.38	68.87	67.89	<b>75.24</b>
AVG	-	74.04	74.67	75.03	<u>75.13</u>	75.00	74.81	72.62	73.01	<b>78.34</b>

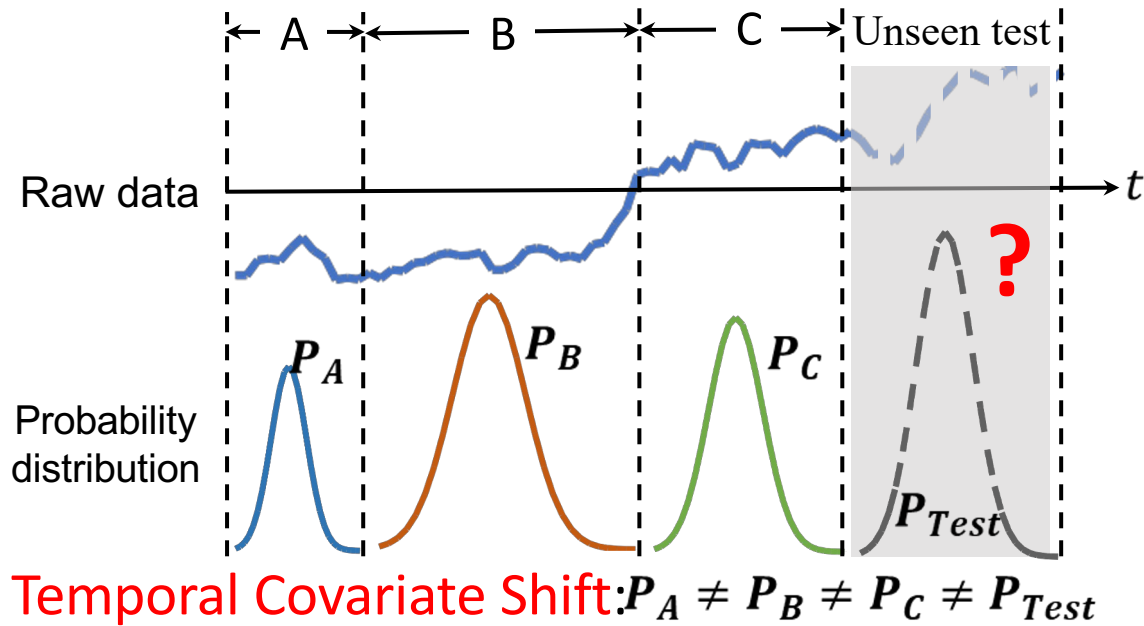


- Lu et al. Domain-invariant feature exploration for domain generalization. TMLR 2022. <https://arxiv.org/abs/2207.12020>



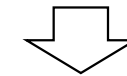
# New perspective: dynamic distributions

- Non-stationary time series
  - The statistical properties are changing over time
  - We formulate it as TCS: temporal covariate shift



Covariate shift:

$$P_{train}(x) \neq P_{test}(x), P_{train}(y|x) = P_{test}(y|x)$$



Temporal Covariate shift:

$$\mathcal{D} = \{D_1, D_2, \dots, D_K\}$$
$$P_{D_i}(x) \neq P_{D_j}(x), P_{D_i}(y|x) = P_{D_j}(y|x)$$

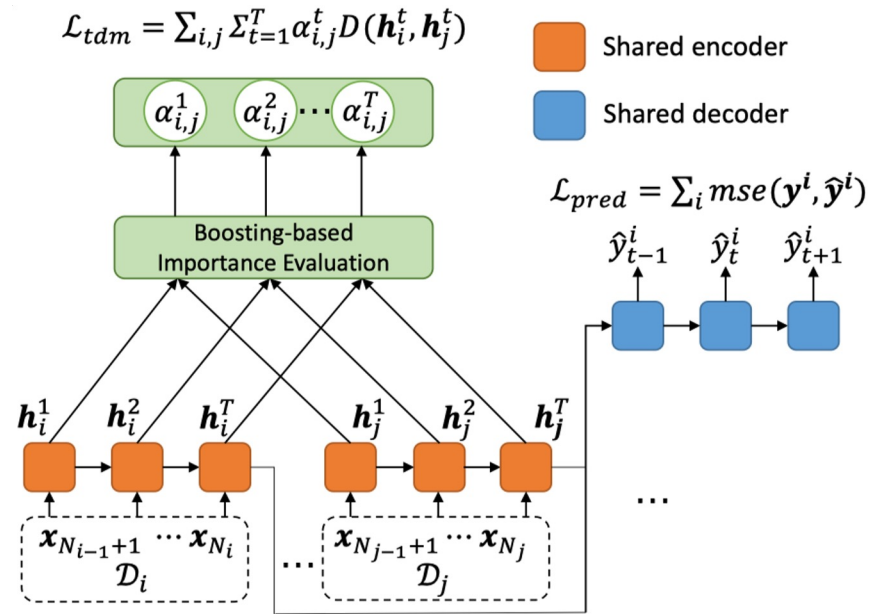
Construct worst-case distribution scenario

Match the big distribution gap

Good model

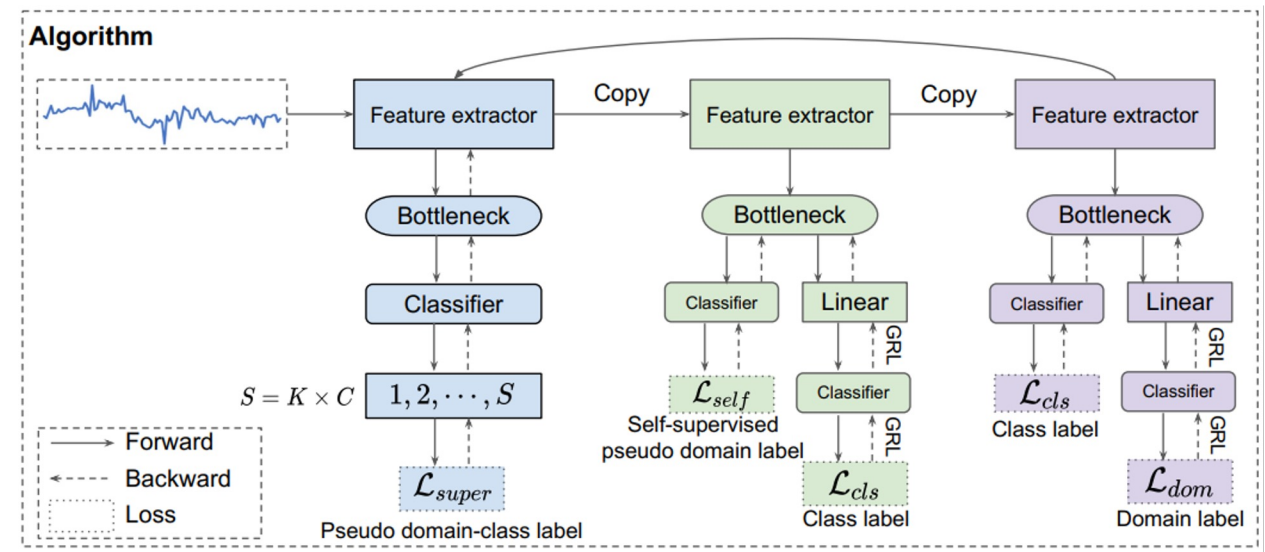
# Algorithms

- AdaRNN: Adaptive RNNs



$$\theta^*, \alpha^* = \arg \min_{\theta, \alpha} \mathcal{L}_{pred}(\theta) + \lambda \sum_{1 \leq i, j \leq K} \mathcal{L}_{tdm}(\mathbf{H}_i, \mathbf{H}_j; \alpha_{i,j}, \theta)$$

## • DIVERSIFY



- Du et al. AdaRNN: adaptive learning and forecasting for time series. CIKM 2021.
- Lu et al. DIVERSIFY to generalize: learning generalized representation for time series classification. ICLR 2023.

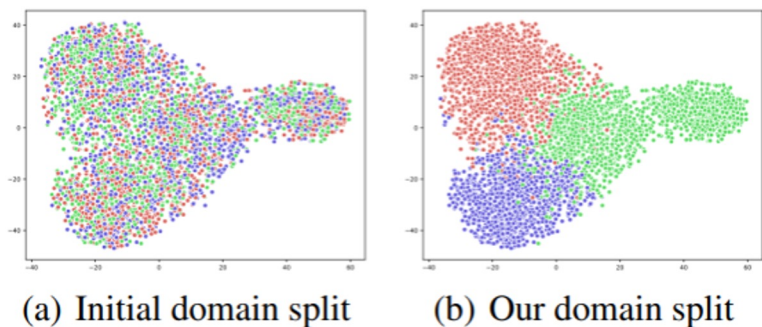
# Experimental results

- Forecasting

Weather forecasting and electric consumption

	Dongsi		Tiantan		Nongzhanguan		Dingling		$\Delta(\%)$	Electric Power
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE		
FBProphet [10]	0.1866	0.1403	0.1434	0.1119	0.1551	0.1221	0.0932	0.0736	-	0.080
ARIMA	0.1811	0.1356	0.1414	0.1082	0.1557	0.1156	0.0922	0.0709	-	-
GRU	0.0510	0.0380	0.0475	0.0348	0.0459	0.0330	0.0347	0.0244	0.00	0.093
MMD-RNN	0.0360	0.0267	0.0183	0.0133	0.0267	0.0197	0.0288	0.0168	-61.31	0.082
DANN-RNN	0.0356	0.0255	0.0214	0.0157	0.0274	0.0203	0.0291	0.0211	-59.97	0.080
LightGBM	0.0587	0.0390	0.0412	0.0289	0.0436	0.0319	0.0322	0.0210	-11.08	0.080
LSTNet [23]	0.0544	0.0651	0.0519	0.0651	0.0548	0.0696	0.0599	0.0705	-	0.080
Transformer [45]	0.0339	0.0220	0.0233	0.0164	0.0226	0.0181	0.0263	0.0163	-61.20	0.079
STRIFE [24]	0.0365	0.0216	0.0204	0.0148	0.0248	0.0154	0.0304	0.0139	-64.60	0.086
<b>AdaRNN</b>	<b>0.0295</b>	<b>0.0185</b>	<b>0.0164</b>	<b>0.0112</b>	<b>0.0196</b>	<b>0.0122</b>	<b>0.0233</b>	<b>0.0150</b>	<b>-73.57</b>	<b>0.077</b>

- Better domain split effects

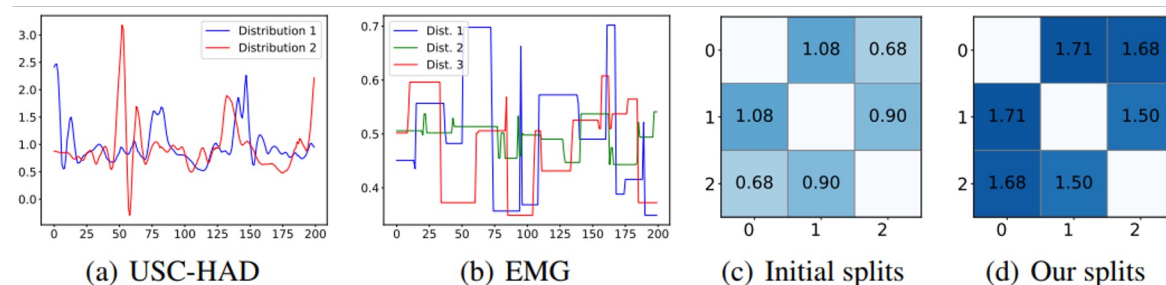


- Classification

Human activity recognition

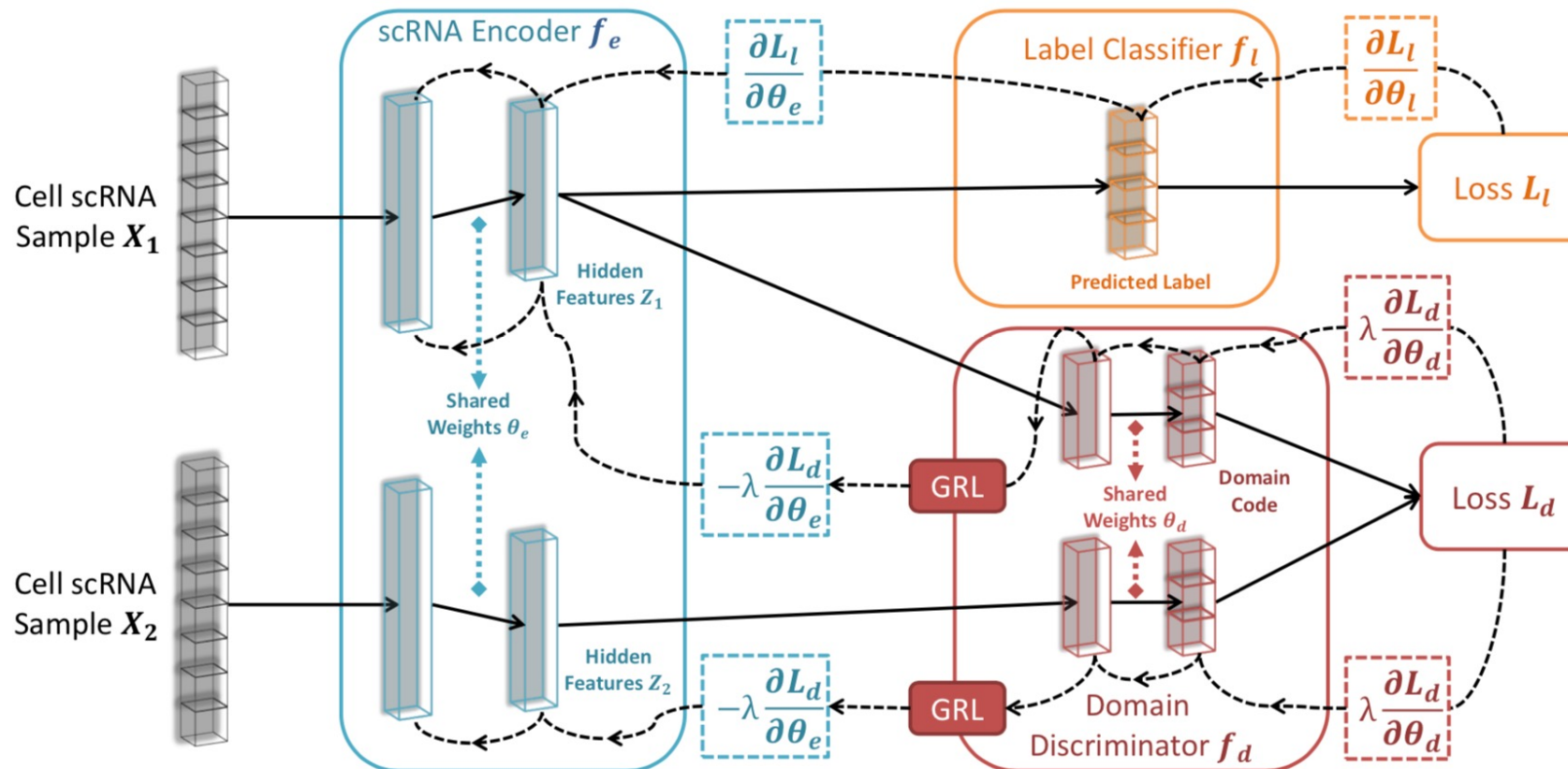
Target	Cross-position generalization						Cross-dataset generalization					One-Person-To-Another			
	0	1	2	3	4	AVG	0	1	2	3	AVG	DSADS	USC-HAD	PAMAP	AVG
ERM	41.5	26.7	35.8	21.4	27.3	30.6	26.4	29.6	44.4	32.9	33.3	51.3	46.2	53.1	50.2
DANN	45.4	25.3	38.1	28.9	25.1	32.6	29.7	45.3	46.1	43.8	41.2	-	-	-	-
CORAL	33.2	25.2	25.8	22.3	20.6	25.4	39.5	41.8	39.1	36.6	39.2	-	-	-	-
Mixup	<b>48.8</b>	<b>34.2</b>	37.5	29.5	29.9	36.0	37.3	<b>47.4</b>	40.2	23.1	37.0	62.7	46.3	58.6	55.8
GroupDRO	27.1	26.7	24.3	18.4	24.8	24.3	<b>51.4</b>	36.7	33.2	33.8	38.8	51.3	48.0	53.1	50.8
RSC	46.6	27.4	35.9	27.0	29.8	33.3	33.1	39.7	45.3	45.9	41.0	59.1	49.0	59.7	55.9
ANDMask	47.5	31.1	39.2	30.2	29.9	35.6	41.7	33.8	43.2	40.2	39.7	57.2	45.9	54.3	52.5
<b>DIVERSIFY</b>	47.7	32.9	<b>44.5</b>	<b>31.6</b>	<b>30.4</b>	<b>37.4</b>	48.7	46.9	<b>49.0</b>	<b>59.9</b>	<b>51.1</b>	<b>67.6</b>	<b>55.0</b>	<b>62.5</b>	<b>61.7</b>

- Really characterize the latent distributions inside a time series!



# Removing batch effects from scRNA data

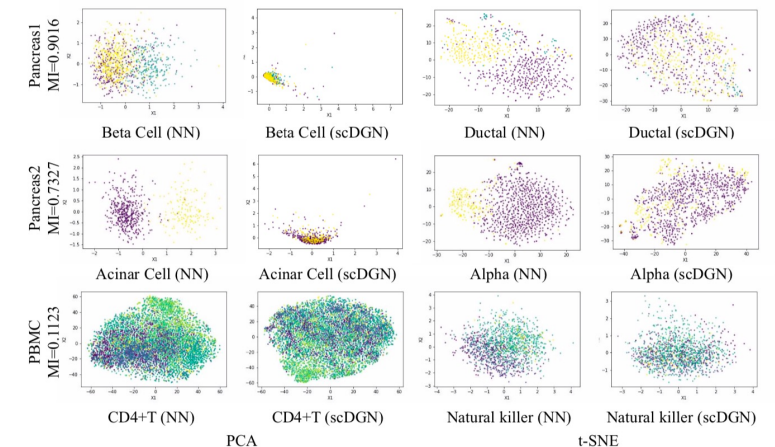
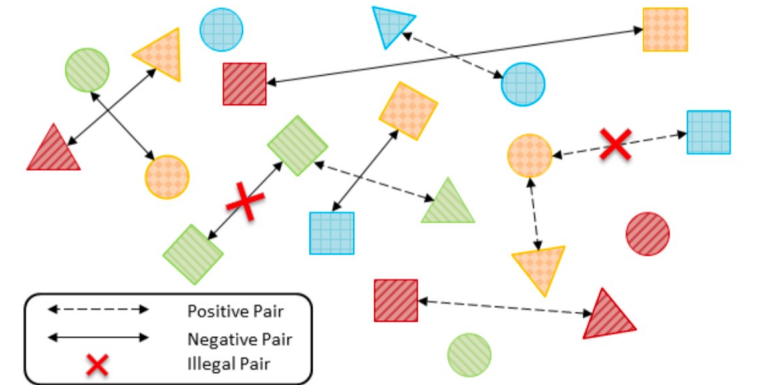
- Model



Ge, Songwei, et al. "Supervised adversarial alignment of single-cell RNA-seq data." Journal of Computational Biology 28.5 (2021): 501-513.

# Removing batch effects from scRNA data

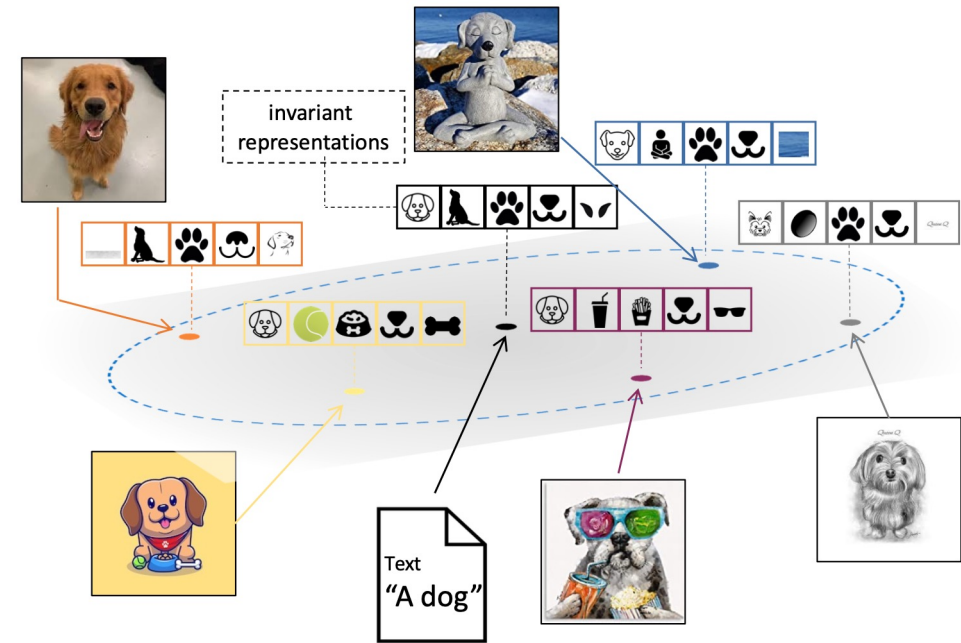
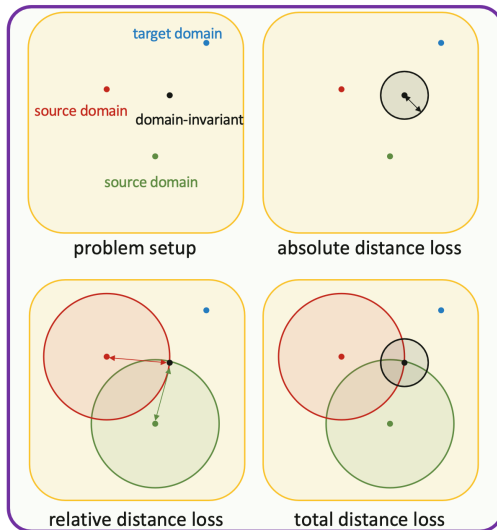
- Algorithm:
  - two types of sample pairs are considered
    - Samples from the same domain, with different cell types
    - Samples from different domains, with the same cell type
- Results:
  - High numerical results in classification of cell types
  - visualization



Ge, Songwei, et al. "Supervised adversarial alignment of single-cell RNA-seq data." *Journal of Computational Biology* 28.5 (2021): 501-513.

# A Sentence Speaks a Thousand Images

- Regularizing the representations invariant from different domains
  - Different loss functions for the distances between invariant (text) vs. images



Method	Backbone	Ens/MA	PACS	VLCS	OfficeHome	Terra	Ave
ERM [68]	ResNet18	No	81.5	73.2	63.3	43.6	65.4
Best SoTA competitor	ResNet18	No	83.4 [22]	74.1 [22]	63.8 [29]	44.5 [22]	66.5
ERM [16]	ResNet50	No	85.7	77.4	67.5	47.2	69.5
Best SOTA competitor	ResNet50	No	86.6 [48]	78.8 [52]	68.7 [52]	48.6 [38]	70.7
Ensemble [2]	ResNet50	Yes	87.6	78.5	70.8	49.2	71.5
SWAD [5]	ResNet50	Yes	88.1	79.1	70.6	50.0	71.9
EoA [2]	ResNet50	Yes	88.6	79.1	72.5	52.3	73.1
CLIP [77] (Teacher)	ViT B/16	No	96.1	82.3	82.3	50.2*	77.7
ERM + Hint	ResNet18	No	84.6	78.0	64.6	47.0	68.6
ERM + Hint + AD	ResNet18	No	85.1	78.5	65.6	48.2	69.4
ERM + Hint + RD	ResNet18	No	84.9	78.2	65.2	47.9	69.0
ERM + Hint + AD + RD (Our full method)	ResNet18	No	85.3	78.6	65.9	48.6	69.6
ERM + Hint	ResNet50	No	88.4	80.7	70.2	50.5	72.5
ERM + Hint + AD	ResNet50	No	89.0	81.5	71.3	52.2	73.5
ERM + Hint + RD	ResNet50	No	88.8	81.2	71.1	51.7	73.2
ERM + Hint + AD + RD (Our full method)	ResNet50	No	89.4	81.8	71.8	52.5	73.9
ERM + Hint + AD + RD + MT (Our full method)	ResNet50	Yes	90.2	82.4	72.6	54.0	74.8

# Robust Machine Learning

- So far, two branch of solutions:
  - Introducing invariance to the models (regularizing hypothesis space)
    - So that the model does not learn information different between source and target domains
  - Introducing new augmented data
    - Augment the data so that we will be able to train models with samples more like the test domain
- How about we just put these things together

# Existing Solutions

## (Worst-case) Data Augmentation

- (Geirhos et al., 2019)
- (Hermann and Kornblith, 2020)
- (Wang et al., 2022)
- (Hendrycks et al., 2019)
- (Mahabadi et al., 2020)
- (Shankar et al., 2018)
- (Huang et al., 2020)
- (Lee et al., 2021)
- (Huang et al., 2022)
- (Madry et al., 2018)

Usually requires specified knowledge about how to augment the data to introduce invariance

## Regularizing Hypothesis Space

- (Wang et al., 2019)
- (Bahng et al., 2019)
- (Wang et al., 2019b)
- (He et al., 2019)
- (Mahabadi et al., 2020)
- (Nam et al., 2020)
- (Ghifary et al., 2016)
- (Rozantsev et al., 2018)
- (Motiian et al., 2017)
- (Li et al., 2018)
- (Carlucci et al., 2018)

Usually requires specified knowledge from additional labels or functional invariance



# A Simple Heuristic (simply combining these two branches)

(Worst-case) Data Augmentation

Usually requires specified knowledge about how to augment the data to introduce invariance

Regularizing Hypothesis Space

Usually requires specified knowledge from additional labels or functional invariance

## Worst-case Data Augmentation with Regularization

- Generic data augmentation through frequency domain
  - Worst-case selection from choices of perturbation radii
- Additional classifier to regularize the learned embeddings
  - Augmentation offers labels of domains (original vs. augmented)

# Empirical Results

- Performances on 9-class ImageNet

	Vanilla	SN	LM	RUBi	ReBias	Mixup	Cutout	AugMix	WT	Reg	WR
Standard Acc.	90.80	88.40	67.90	90.50	91.90	92.50	91.20	92.90	92.50	93.10	<b>93.30</b>
Weighted Acc.	88.80	86.60	65.90	88.60	90.50	91.20	90.30	91.70	91.30	<b>92.20</b>	92.00
ImageNet-A	24.90	24.60	18.80	27.70	29.60	29.10	27.30	<b>31.50</b>	28.50	30.00	29.60
ImageNet-Sketch	41.10	40.50	36.80	42.30	41.80	40.60	38.70	41.40	43.00	42.50	<b>43.20</b>
average	61.40	60.03	47.35	62.28	63.45	63.35	61.88	64.38	63.83	64.45	<b>64.53</b>

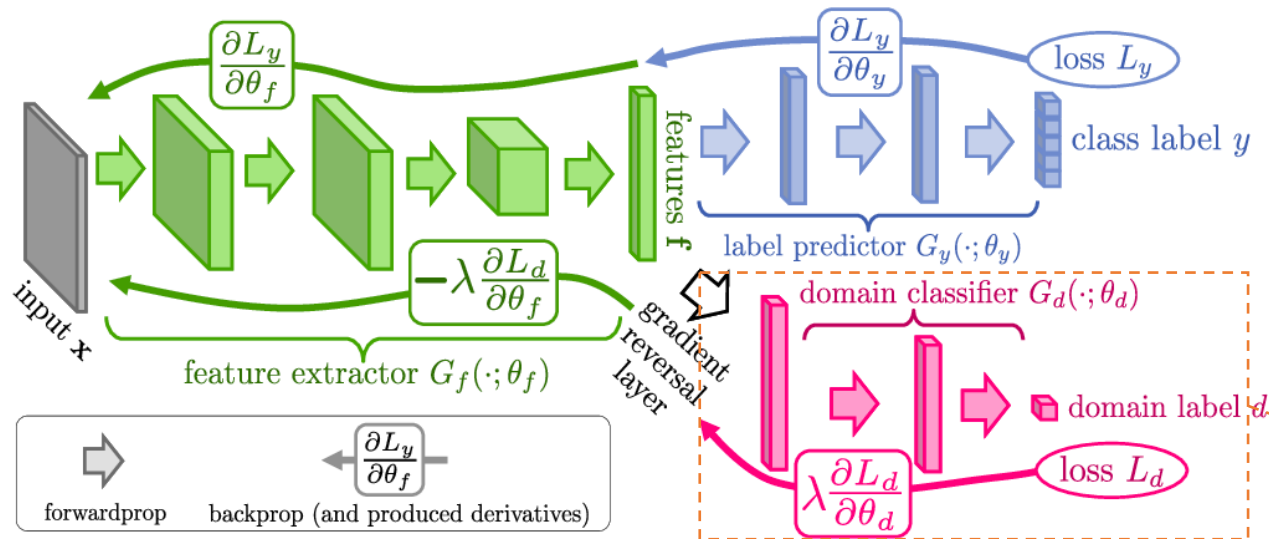
regularization

data augmentation

worst-case aug + regularization

(and ablation study)

# A Classifier Might Not Be Necessary



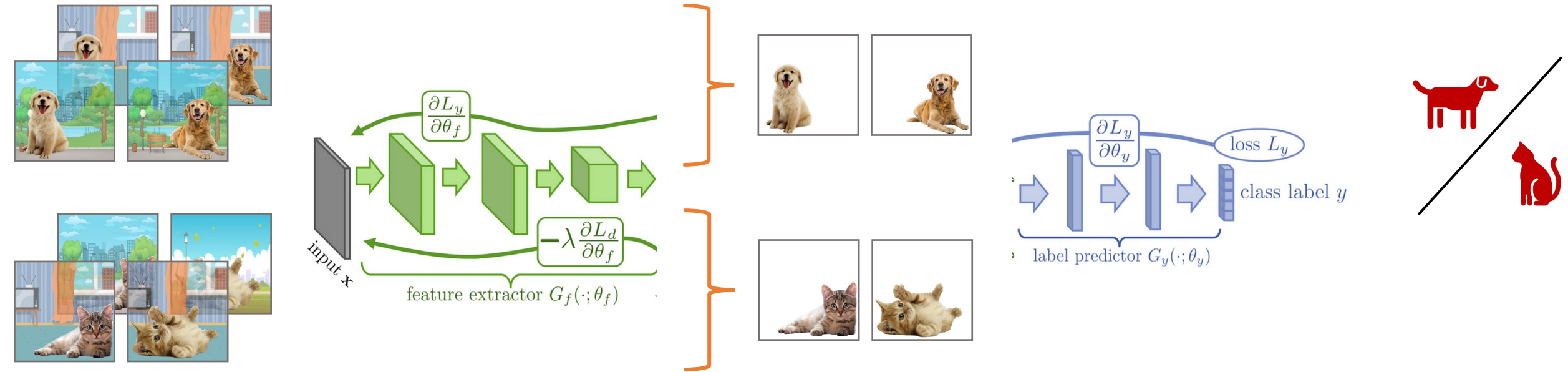
?

If we are doing data augmentation, do we really need the domain classifier there?

- This classifier (or its extensions) are there to push the embeddings invariant from two domains
- Not necessary if we have more efficient methods
  - Especially in the data augmentation settings

# A Classifier Might Not Be Necessary

- **Alignment regularization** pushes the model to learn the same representations from an image and its augmented counterpart.



# Alignment Regularization

- $\ell_2$  distance and cosine similarities
  - internal representations
  - speech recognition
  - (Liang et al., 2018)
- Squared  $\ell_2$  distance
  - logits
  - adversarial robust vision models
  - (Kannan et al., 2018)
- KL divergence
  - softmax outputs
  - adversarial robust vision models
  - (Zhang et al., 2019a)
- Jensen–Shannon divergence
  - embeddings
  - texture invariant image classification
  - (Hendrycks et al., 2020)
- and many others...

If there is a general method that can work well across applications, and enjoys some theoretical support?

# Our Solution: Squared L2 Norm as Alignment Regularization

- We recommend using **squared l2 norm** as regularization

## Empirically

We conduct a set of experiments and find out squared L2 norm is the best choice

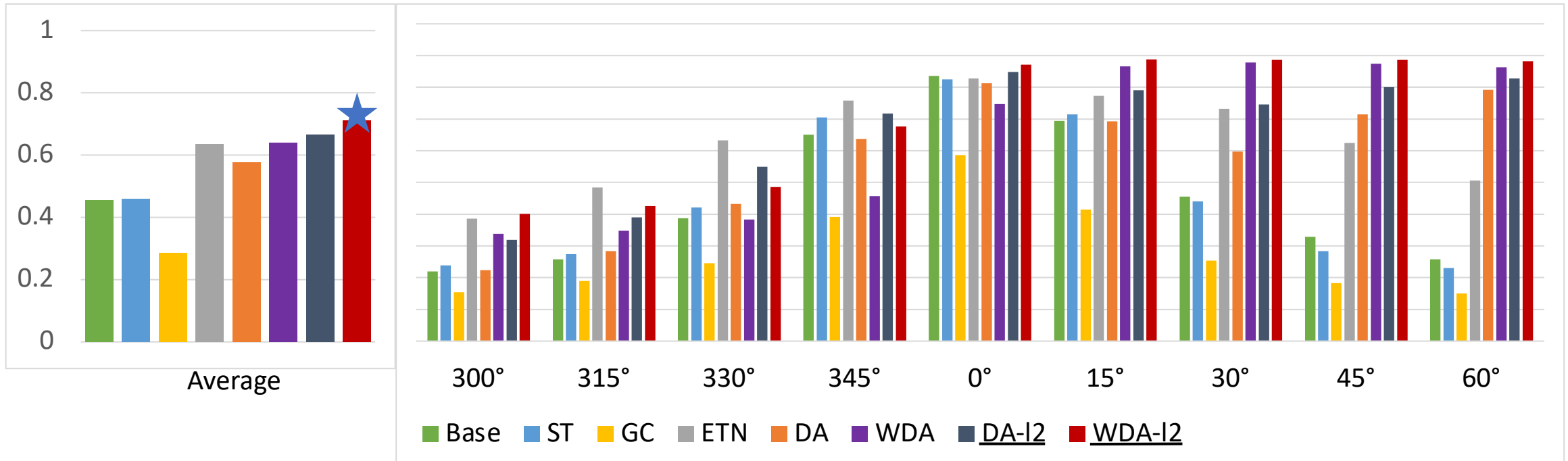
## Theoretically

We complement our empirical study with a formal proof

- We show a bounded worst-case error (robustness)
- We connect the regularization to the measure of invariance

# Comparison to the Top-performing Methods: Results





- Rotation-invariant Classification



**The simple method we identified can compete with top-performing methods specially designed for each task.**

# Re-weighting Based Methods

- Targeting a specific type of distributional shift
  - There are some minor samples in the training set
- Solutions:
  - Give majority samples lower weights
    - Group-DRO
  - Give majority samples even negative weights
    - VREx

		label: object	
		waterbird	landbird
spurious attribute: background	water background	 majority	 minority
	land background	 minority	 majority



# Group-DRO extensions

- Adversarially reweighted learning (ARL)
  - uses another model to identify samples

$$\lambda(\mathbf{x}) = 1 + |(\mathbf{X}, \mathbf{Y})| \cdot \frac{\phi(\mathbf{x})}{\sum_{(\mathbf{x}, y) \in (\mathbf{X}, \mathbf{Y})} \phi(\mathbf{x})}$$

- Learning from failures (LFF)
  - trains another model by amplifying its early-stage predictions

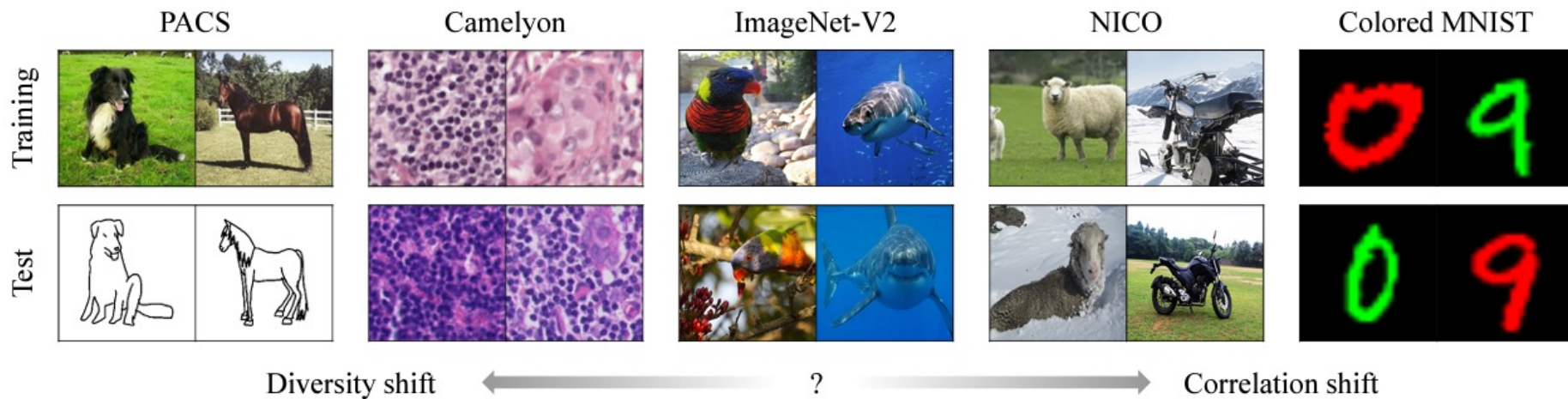
$$\lambda(\mathbf{x}) = \frac{\ell(\phi(\mathbf{x}), y)}{\ell(\phi(\mathbf{x}), y) + \ell(\theta(\mathbf{x}), y)}$$

# Is there a lost of domain generalization to search? (maybe, maybe not)

- In the search of lost domain generalization
  - (Gulrajani and Lopez-Paz 20')
  - The developed domain generalization methods cannot perform well once taking the averaged performances across a set of hyperparameters
    - ERM can is the best in comparison to the averaged performances
- A counterpoint
  - Extensions of ERM will probably be at least as good as ERM
    - Unless with an unrealistic choice of hyperparameters
      - Worst-case feature methods with too many features to be dropped
      - Worst-case sample methods with too many samples to be dropped (in comparison to batch size)

# Is there a lost of domain generalization to search? (probably not, again)

- A closer look at the datasets
  - (Ye *et al* 2020)
  - There are two kinds of diversity shift in the datasets
  - It's probably not the best to consider the shift as one thing
    - Diversity shift (the shift of support)
    - Correlation shift (the shift of density)



# Existing Solutions (and the previous best ones in each)

- If we split the datasets into two polarities
  - No method can do the best across these two settings

Algorithm	PACS	OfficeHome	TerraInc	Camelyon17	Average	Ranking score
RSC [38]	82.8 ± 0.4 <sup>↑</sup>	62.9 ± 0.4 <sup>↓</sup>	43.6 ± 0.5 <sup>↑</sup>	94.9 ± 0.2 <sup>↑</sup>	71.1	+2
MMD [48]	81.7 ± 0.2 <sup>↑</sup>	63.8 ± 0.1 <sup>↑</sup>	38.3 ± 0.4 <sup>↓</sup>	94.9 ± 0.4 <sup>↑</sup>	69.7	+2
SagNet [60]	81.6 ± 0.4 <sup>↑</sup>	62.7 ± 0.4 <sup>↓</sup>	42.3 ± 0.7	95.0 ± 0.2 <sup>↑</sup>	70.4	+1
ERM [90]	81.5 ± 0.0	63.3 ± 0.2	42.6 ± 0.9	94.7 ± 0.1	70.5	0
IGA [43]	80.9 ± 0.4 <sup>↓</sup>	63.6 ± 0.2 <sup>↑</sup>	41.3 ± 0.8 <sup>↓</sup>	95.1 ± 0.1 <sup>↑</sup>	70.2	0
CORAL [85]	81.6 ± 0.6 <sup>↑</sup>	63.8 ± 0.3 <sup>↑</sup>	38.3 ± 0.7 <sup>↓</sup>	94.2 ± 0.3 <sup>↓</sup>	69.5	0
IRM [9]	81.1 ± 0.3 <sup>↓</sup>	63.0 ± 0.2 <sup>↓</sup>	42.0 ± 1.8	95.0 ± 0.4 <sup>↑</sup>	70.3	-
VREx [44]	81.8 ± 0.1 <sup>↑</sup>	63.5 ± 0.1	40.7 ± 0.7 <sup>↓</sup>	94.1 ± 0.3 <sup>↓</sup>	70.0	-
GroupDRO [79]	80.4 ± 0.3 <sup>↓</sup>	63.2 ± 0.2	36.8 ± 1.1 <sup>↓</sup>	95.2 ± 0.2 <sup>↑</sup>	68.9	-
ERDG [105]	80.5 ± 0.5 <sup>↓</sup>	63.0 ± 0.4 <sup>↓</sup>	41.3 ± 1.2 <sup>↓</sup>	95.5 ± 0.2 <sup>↑</sup>	70.1	-
DANN [27]	81.1 ± 0.4 <sup>↓</sup>	62.9 ± 0.6 <sup>↓</sup>	39.5 ± 0.2 <sup>↓</sup>	94.9 ± 0.0 <sup>↑</sup>	69.6	-
MTL [16]	81.2 ± 0.4 <sup>↓</sup>	62.9 ± 0.2 <sup>↓</sup>	38.9 ± 0.6 <sup>↓</sup>	95.0 ± 0.1 <sup>↑</sup>	69.5	-
Mixup [101]	79.8 ± 0.6 <sup>↓</sup>	63.3 ± 0.5	39.8 ± 0.3 <sup>↓</sup>	94.6 ± 0.3	69.4	-
ANDMask [64]	79.5 ± 0.0 <sup>↓</sup>	62.0 ± 0.3 <sup>↓</sup>	39.8 ± 1.4 <sup>↓</sup>	95.3 ± 0.1 <sup>↑</sup>	69.2	-
ARM [103]	81.0 ± 0.4 <sup>↓</sup>	63.2 ± 0.2	39.4 ± 0.7 <sup>↓</sup>	93.5 ± 0.6 <sup>↓</sup>	69.3	-
MLDG [47]	73.0 ± 0.4 <sup>↓</sup>	52.4 ± 0.2 <sup>↓</sup>	27.4 ± 2.0 <sup>↓</sup>	91.2 ± 0.4 <sup>↓</sup>	61.0	-
<b>Average</b>	80.7	62.5	39.8	94.6	69.4	-

diversity shift dataset

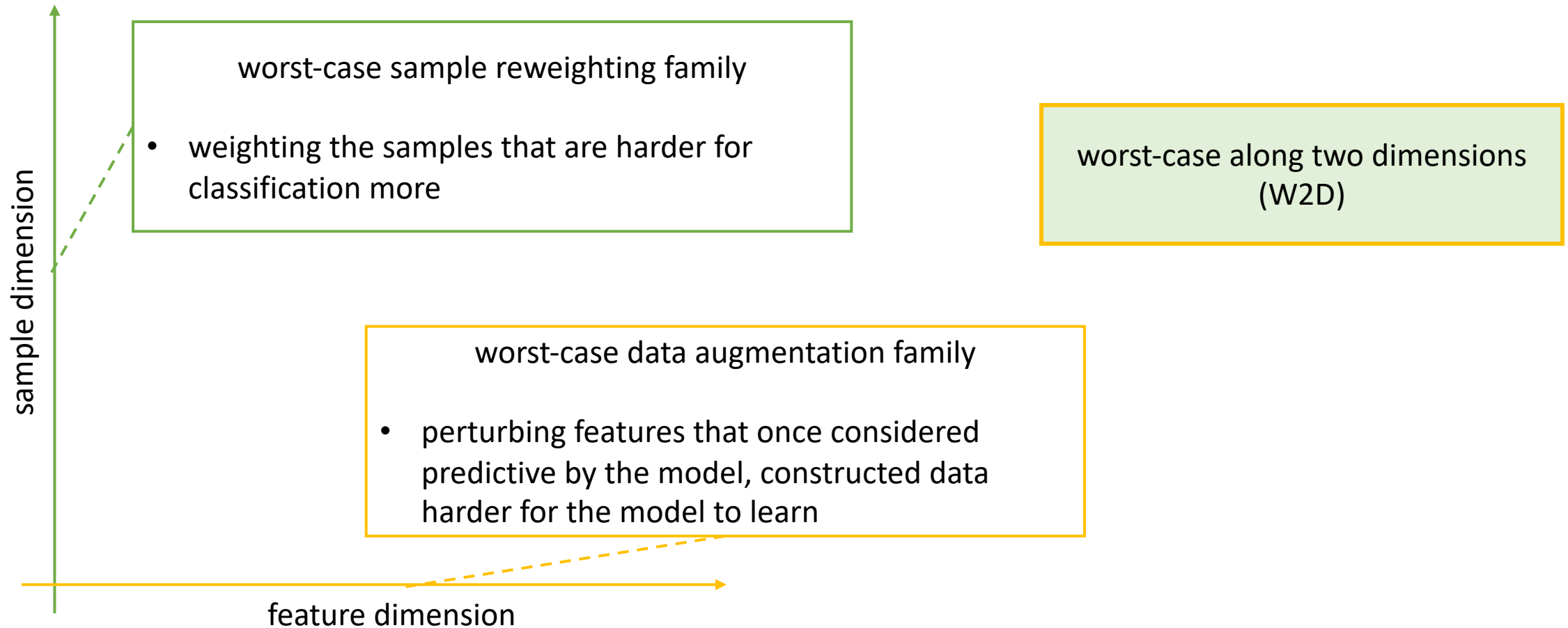
worst-case data augmentation family

Algorithm	Colored MNIST	CelebA	NICO	Average	Prev score	Ranking score
VREx [44]	56.3 ± 1.9 <sup>↑</sup>	87.3 ± 0.2	71.5 ± 2.3	71.7	-1	+1
GroupDRO [79]	32.5 ± 0.2 <sup>↑</sup>	87.5 ± 1.1	71.0 ± 0.4	63.7	-1	+1
ERM [90]	29.9 ± 0.9	87.2 ± 0.6	72.1 ± 1.6	63.1	0	0
IRM [9]	60.2 ± 2.4 <sup>↑</sup>	85.4 ± 1.2 <sup>↓</sup>	73.3 ± 2.1	73.0	-1	0
MTL [16]	29.3 ± 0.1	87.0 ± 0.7	70.6 ± 0.8	62.3	-2	0
ERDG [105]	31.6 ± 1.3 <sup>↑</sup>	84.5 ± 0.2 <sup>↓</sup>	72.7 ± 1.9	62.9	-2	0
ARM [103]	34.6 ± 1.8 <sup>↑</sup>	86.6 ± 0.7	67.3 ± 0.2 <sup>↓</sup>	62.8	-3	0
MMD [48]	50.7 ± 0.1 <sup>↑</sup>	86.0 ± 0.5 <sup>↓</sup>	68.9 ± 1.2 <sup>↓</sup>	68.5	+2	-1
RSC [38]	28.6 ± 1.5 <sup>↓</sup>	85.9 ± 0.2 <sup>↓</sup>	74.3 ± 1.9 <sup>↑</sup>	61.4	+2	-1
IGA [43]	29.7 ± 0.5	86.2 ± 0.7 <sup>↓</sup>	71.0 ± 0.1	62.3	0	-1
CORAL [85]	30.0 ± 0.5	86.3 ± 0.5 <sup>↓</sup>	70.8 ± 1.0	61.5	-1	-1
Mixup [101]	27.6 ± 1.8 <sup>↓</sup>	87.5 ± 0.5	72.5 ± 1.5	60.6	-2	-1
MLDG [47]	32.7 ± 1.1 <sup>↑</sup>	85.4 ± 1.3 <sup>↓</sup>	66.6 ± 2.4 <sup>↓</sup>	56.6	-4	-1
SagNet [60]	30.5 ± 0.7	85.8 ± 1.4 <sup>↓</sup>	69.8 ± 0.7 <sup>↓</sup>	62.0	+1	-2
ANDMask [64]	27.2 ± 1.4 <sup>↓</sup>	86.2 ± 0.2 <sup>↓</sup>	71.2 ± 0.8	61.5	-2	-2
DANN [27]	24.5 ± 0.8 <sup>↓</sup>	86.0 ± 0.4 <sup>↓</sup>	69.4 ± 1.7 <sup>↓</sup>	59.7	-2	-3
<b>Average</b>	34.5	86.4	70.8	63.7	-	-

worst-case sample reweighting family

correlation shift dataset

# The two dimensions of worst-case training



$$\varepsilon_{P_t}(\theta) \leq \hat{\varepsilon}_{P_s}(\theta) + \phi(\Theta, n, \delta) + c(\theta)$$

# The two dimensions of worst-case training

- Design Rationales
  - Through the integration of principled understanding
  - Also inspired by a psychological prior
- Method Details
  - Acceleration with hard samples
  - Consider all samples for performances upon convergence

---

**Algorithm 1:** W2D Algorithm

---

**Input:** data set  $(\mathbf{X}, \mathbf{Y})$ , percentage of samples used per batch  $\rho$ , percentage of whole batch patching  $\kappa$ , batch size  $\eta$ , maximum number of epochs  $T$ , and other RSC hyperparameters;

**Output:** Classifier  $f(\cdot; \theta)$ ;

randomly initialize the model  $\theta_0$ ;

calculate the number of iterations  $K = n/\eta$ ;

**while**  $t \leq (1 - \kappa)T$  **do**

**for** a batch of data  $(\mathbf{X}, \mathbf{Y})_k$  where  $k \leq K$  **do**

        forward pass to calculate the loss

$l(f(\mathbf{X}_i; \theta_{t,k-1}), \mathbf{Y}_i)$  of every sample in the batch;

        select the top  $\eta\rho$  samples with highest loss to construct  $(\mathbf{X}, \mathbf{Y})_{k,\rho}$ ;

        Train the model with  $(\mathbf{X}, \mathbf{Y})_{k,\rho}$  following (1).

**end**

**end**

**while**  $(1 - \kappa)T < t \leq T$  **do**

**for** a batch of data  $(\mathbf{X}, \mathbf{Y})_k$  where  $k \leq K$  **do**

        Train the model with  $(\mathbf{X}, \mathbf{Y})_k$  following (1).

**end**

**end**

---

# Results

Algorithm	PACS	OfficeHome	TerraInc	Camelyon	Average	Ranking score
<b>W2D</b>	83.4 ± 0.3	63.5 ± 0.1	44.5 ± 0.5	95.2 ± 0.3	71.7	+3
RSC [31]	82.8 ± 0.4	62.9 ± 0.4	43.6 ± 0.5	94.9 ± 0.2	71.1	+2
MMD [42]	81.7 ± 0.2	63.8 ± 0.1	38.3 ± 0.4	94.9 ± 0.4	69.7	+2
SagNet [49]	81.6 ± 0.4	62.7 ± 0.4	42.3 ± 0.7	95.0 ± 0.2	70.4	+1
ERM [62]	81.5 ± 0.0	63.3 ± 0.2	42.6 ± 0.9	94.7 ± 0.1	70.5	0
IGA [36]	80.9 ± 0.4	63.6 ± 0.2	41.3 ± 0.8	95.1 ± 0.1		
CORAL [59]	81.6 ± 0.6	63.8 ± 0.3	38.3 ± 0.7	94.2 ± 0.3		
IRM [2]	80.9 ± 0.4	63.6 ± 0.2	41.3 ± 0.8	95.1 ± 0.1		
VREx [38]	81.8 ± 0.4	63.5 ± 0.1	40.7 ± 0.7	94.1 ± 0.3		
GroupDRO [57]	80.4 ± 0.3	63.2 ± 0.2	36.8 ± 1.1	95.2 ± 0.2		
ERDG [79]	80.5 ± 0.5	63.0 ± 0.4	41.3 ± 1.2	95.5 ± 0.2		
DANN [16]	81.1 ± 0.4	62.9 ± 0.6	39.5 ± 0.2	94.9 ± 0.0		
MTL [8]	81.2 ± 0.4	62.9 ± 0.2	38.9 ± 0.6	95.0 ± 0.1		
Mixup [75]	79.8 ± 0.6	63.3 ± 0.5	39.8 ± 0.3	94.6 ± 0.3		
ANDMask [53]	79.5 ± 0.0	62.0 ± 0.3	39.8 ± 1.4	95.3 ± 0.1		
ARM [76]	81.0 ± 0.4	63.2 ± 0.2	39.4 ± 0.7	93.5 ± 0.6		
MLDG [41]	73.0 ± 0.4	52.4 ± 0.2	27.4 ± 2.0	91.2 ± 0.4		

diversity shift dataset

top position (tie) by inheriting the power of worst-case sample reweighting methods

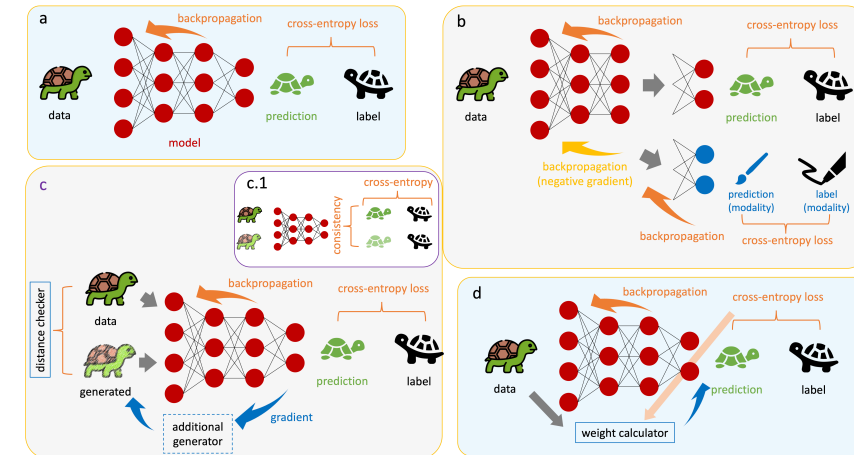
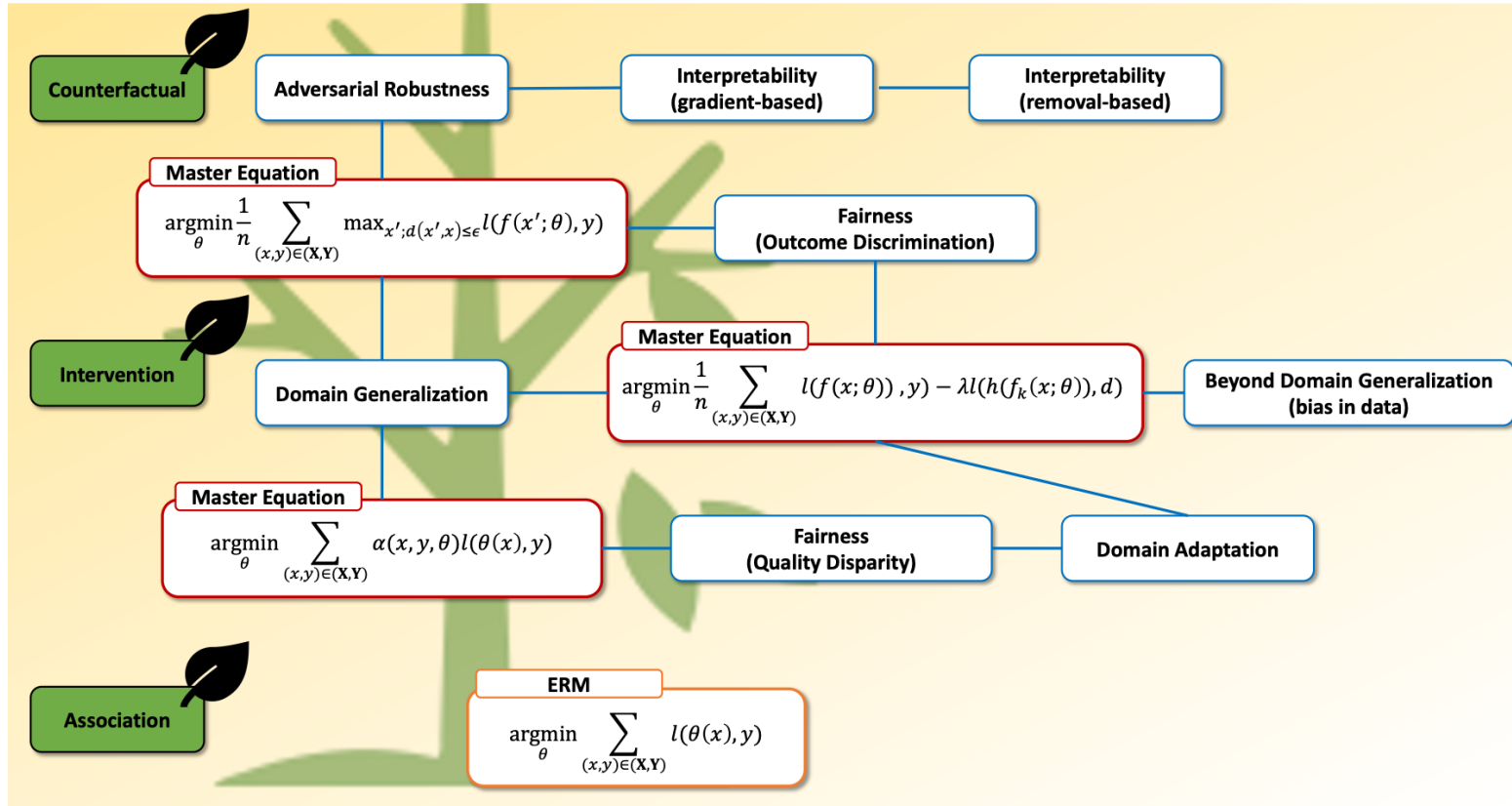
Algorithm	CMNIST	NICO	CelebA	Average	Prev score	Ranking score
VREx [38]	56.3 ± 1.9	71.5 ± 2.3	87.3 ± 0.2	71.7	-1	+1
GroupDRO [57]	32.5 ± 0.2	71.0 ± 0.4	87.5 ± 1.1	63.7	-1	+1
<b>W2D</b>	31.0 ± 0.3	71.9 ± 1.2	87.7 ± 0.4	63.5	+3	+1
ERM [62]	29.9 ± 0.9	72.1 ± 1.6	87.2 ± 0.6	63.1	0	0
IRM [2]	60.2 ± 2.4	73.3 ± 2.1	85.4 ± 1.2	73.0	-1	0
ERDG [79]	31.6 ± 1.3	72.7 ± 1.9	84.5 ± 0.2	62.9	-2	0
ARM [76]	34.6 ± 1.8	67.3 ± 0.2	86.6 ± 0.7	62.8	-3	0
MTL [8]	29.3 ± 0.1	70.6 ± 0.8	87.0 ± 0.7	62.3	-2	0
MMD [42]	50.7 ± 0.1	68.9 ± 1.2	86.0 ± 0.5	68.5	+2	-1
RSC [31]	27.6 ± 1.8	74.3 ± 1.9	85.9 ± 0.2	62.6	+2	-1
Mixup [75]	28.6 ± 1.5	72.5 ± 1.5	87.5 ± 0.5	62.5	-2	-1
CORAL [59]	30.0 ± 0.5	70.8 ± 1.0	86.3 ± 0.5	62.4	-1	-1
IGA [36]	29.7 ± 0.5	71.0 ± 0.1	86.2 ± 0.7	62.3	0	-1
MLDG [41]	32.7 ± 1.1	66.6 ± 2.4	85.4 ± 1.3	61.6	-4	-1
SagNet [49]	30.5 ± 0.7	69.8 ± 0.7	85.8 ± 1.4	62.0	+1	-2
ANDMask [53]	27.2 ± 1.4	71.2 ± 0.8	86.2 ± 0.2	61.5	-2	-2
DANN [16]	24.5 ± 0.8	69.4 ± 1.7	86.0 ± 0.4	60.0	-2	-3

correlation shift dataset

top position by inheriting the power of RSC

# Summary

- Master equations and principled solutions of learning robust models





# Implications

- A glimpse into large models
  - The connection between large model solutions to ERM

- Fine-tuning

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim P(X,Y)} l(f(x; \theta), y)$$

- Parameter-efficient fine-tuning

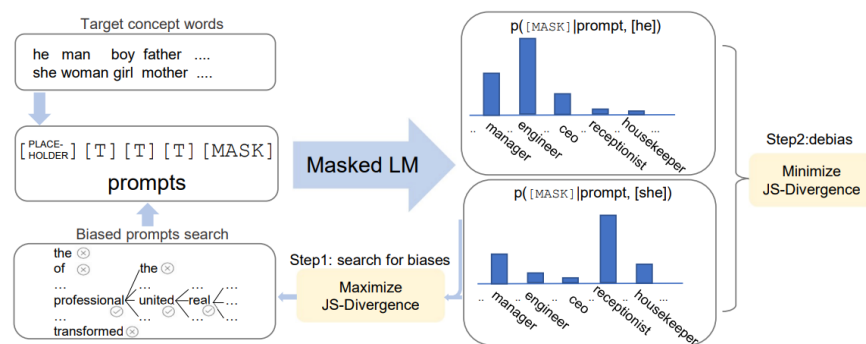
$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim P(X,Y)} l(f(x; [\Theta; \theta]), y)$$

- Prompting

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim P(X,Y)} l(f(h(x; \theta); \Theta), y)$$

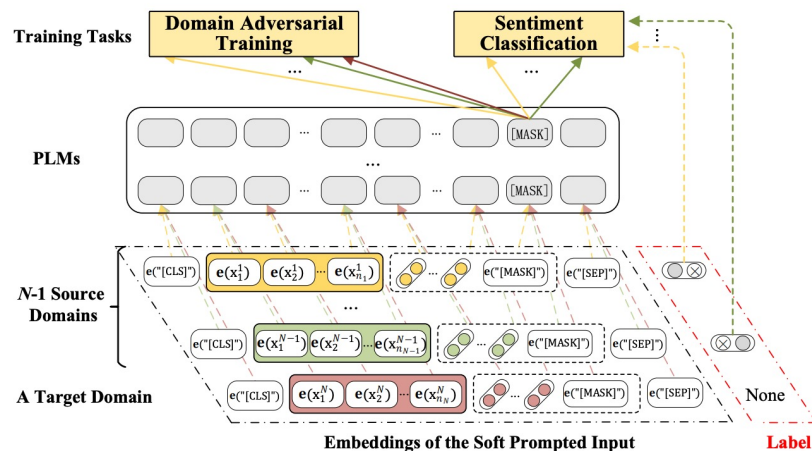
# Implications

- A glimpse into trustworthy solutions of large models
  - Data augmentation based



Guo, Yue, Yi Yang, and Ahmed Abbasi. "Auto-debias: Debiasing masked language models with automated biased prompts." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

- Regularization-based



Wu, Hui, and Xiaodong Shi. "Adversarial soft prompt tuning for cross-domain sentiment analysis." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

# For More Information

- Most of the contents are from our latest survey
  - Data-centric view of Trustworthy Machine Learning
    - in Robustness, Security, Interpretability, Fairness
  - Connection of methods via Pearl's Causal Hierarchy
  - Trustworthiness and techniques of large pretrained models
  - Opportunities for new methods in large models
  - Access: <http://trustai.one>
- Also checkout Jindong's first-ever survey on domain generalization
  - Wang, Jindong, et al. "Generalizing to unseen domains: A survey on domain generalization." *IEEE Transactions on Knowledge and Data Engineering* (2022).

Towards Trustworthy and Aligned Machine Learning:  
A Data-centric Survey with Causality Perspectives

Haoyang Liu<sup>†</sup>, Maheep Chaudhary<sup>†\*</sup>, and Haohan Wang

School of Information Sciences,  
University of Illinois Urbana-Champaign  
{hl57, haohanw}@illinois.edu, maheep001@e.ntu.edu.sg

<sup>†</sup> equal contribution