Towards Effective and Efficient Self-Supervised Robust Pre-Training

Xilie Xu
PhD Student at School of Computing, NUS

26th Aug 2023

AIGC'23 Forum 4: "Foundation Model and Fine-Tuning Strategy"

Chaired by Dr. Jingfeng Zhang

Outline

- Backgrounds
  - Adversarial attack and defense
  - Robust pre-training
- How to make self-supervised robust pre-training
  - More efficient
  - More effective
- Future directions

Outline

- Backgrounds
  - Adversarial attack and defense
  - Robust pre-training
- How to make self-supervised robust pre-training
  - More efficient
  - More effective
- Future directions

Gap between AI development and deployment

Develop AI-based applications
in an idealized environment

Image from https://blog.si-log.net/transport-by-sea-by-land-or-by-air-the-differences-and-similarities

Deploy AI-base applications
in the wild

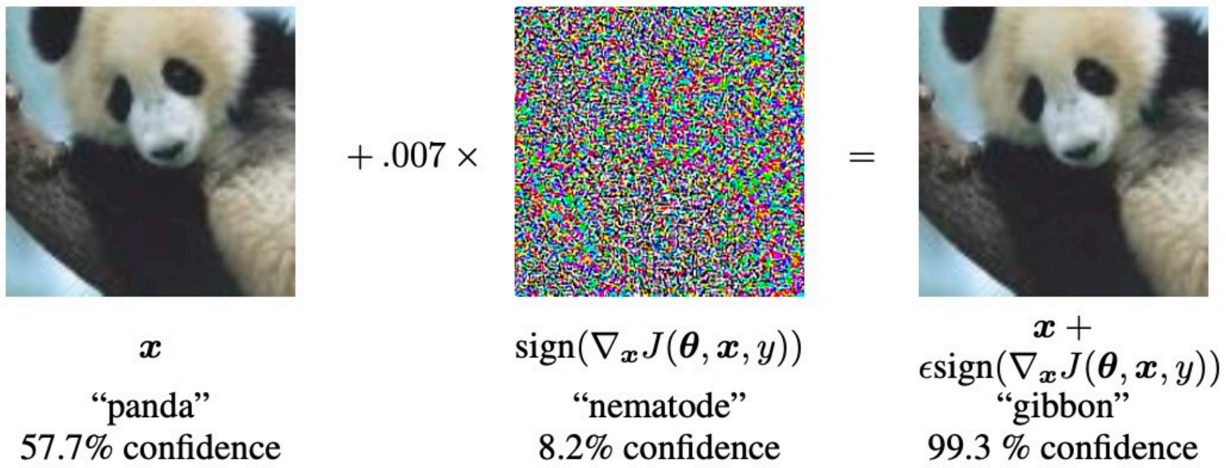
Image from https://www.primeins.com/insurance-news/how-to-protect-your-boat-from-a-tropical-storm-or-hurricane

Threats ("storm")

- Poisoning attack
- Backdoor attack
- Distribution shift
- **Adversarial attack**

Adversarial attacks

Objective: Make the model misclassify the adversarial data.



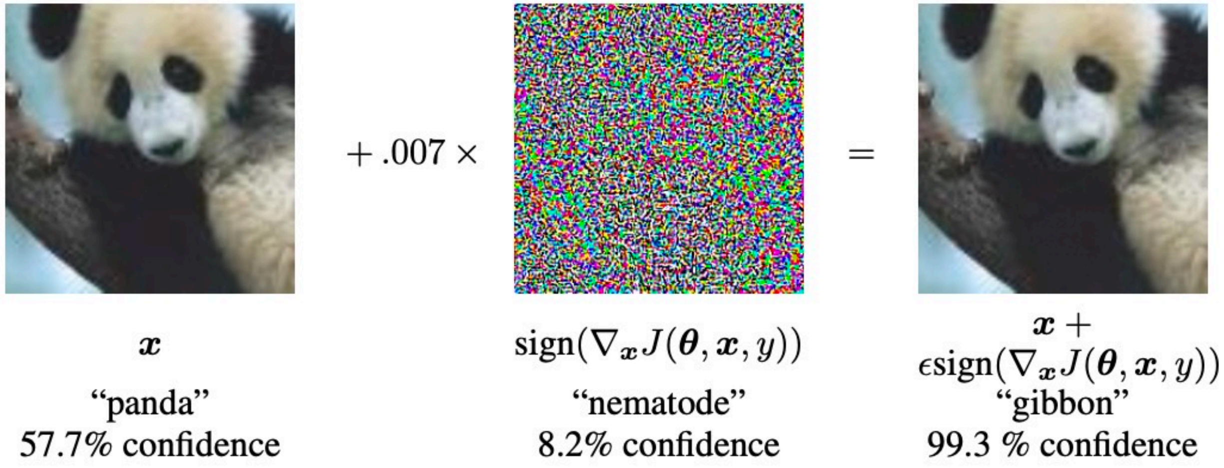
[Goodfellow et al. 2014]

Natural data x + Imperceptible adversarial perturbation = Adversarial data $\tilde{x}$

Adversarial attacks

Objective: Make the model misclassify the adversarial data.

$$\tilde{x} = argmax_{\tilde{x} \in \mathcal{B}_{\epsilon}[x]} \ell(f(\tilde{x}), y)$$



[Goodfellow et al. 2014]

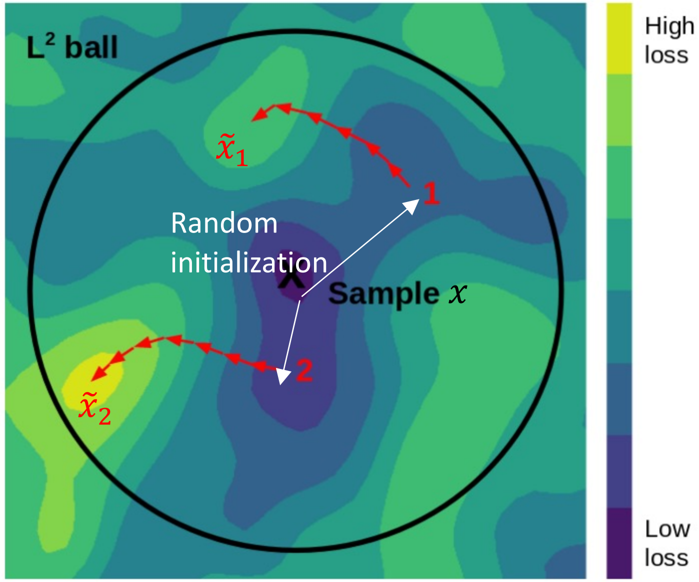


Image modified from https://towardsdatascience.com/know-your-enemy-7f7c5038bdf3

Natural data x + Imperceptible adversarial perturbation = Adversarial data $\tilde{x}$

Projected gradient descent (PGD)

[Madry et al. ICLR 2018]

Supervised adversarial training (SAT)

- Minimax formulation of SAT

$$\underbrace{min_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\ell(f(\tilde{x}_i),y_i)}_{\text{outer minimization}},\text{ where }\underbrace{\tilde{x}_i=argmax_{\tilde{x}_i\in\mathcal{B}_\epsilon[x_i]}\ell(f(\tilde{x}_i),y_i)}_{\text{inner maximization}}$$

[Madry et al. ICLR 2018]

- Realization

Alternatively conduct steps (1) and (2):

(1) generate adversarial data maximizing the loss;

(2) minimize loss on the generated adversarial data w.r.t. model parameters.

SAT

- Minimax formulation of SAT

$$\underbrace{min_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\ell(f(\tilde{x}_i),y_i)}_{\text{outer minimization}},\text{ where }\underbrace{\tilde{x}_i=argmax_{\tilde{x}_i\in\mathcal{B}_\epsilon[x_i]}\ell(f(\tilde{x}_i),y_i)}_{\text{inner maximization}}$$

[Madry et al. ICLR 2018]

- Realization
- Drawback: SAT requires a large amount of labelled data (for each task).

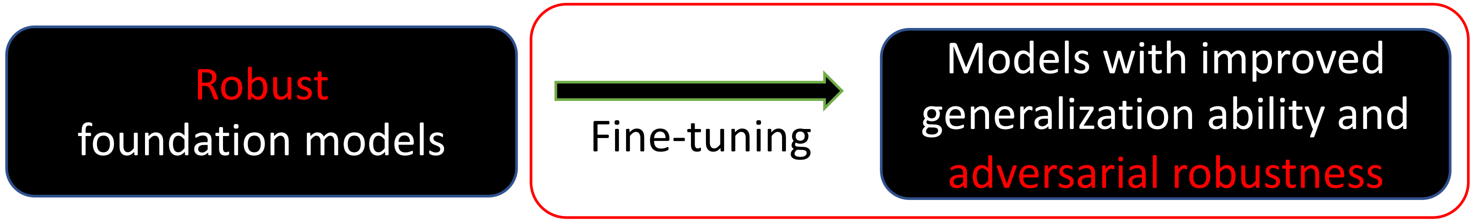
SAT

- Minimax formulation of SAT

$$\underbrace{min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(\tilde{x}_i), y_i)}_{\text{outer minimization}}, \text{ where } \tilde{x}_i = \underbrace{argmax_{\tilde{x}_i \in \mathcal{B}_\epsilon[x_i]} \ell(f(\tilde{x}_i), y_i)}_{\text{inner maximization}}$$

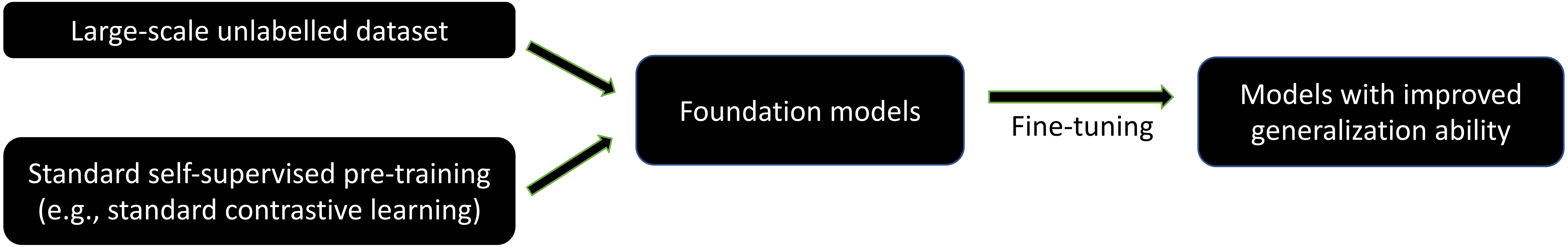
outer minimization [Madry et al. ICLR 2018] inner maximization

- Realization
- Drawback: SAT requires a large amount of labelled data (for each task).

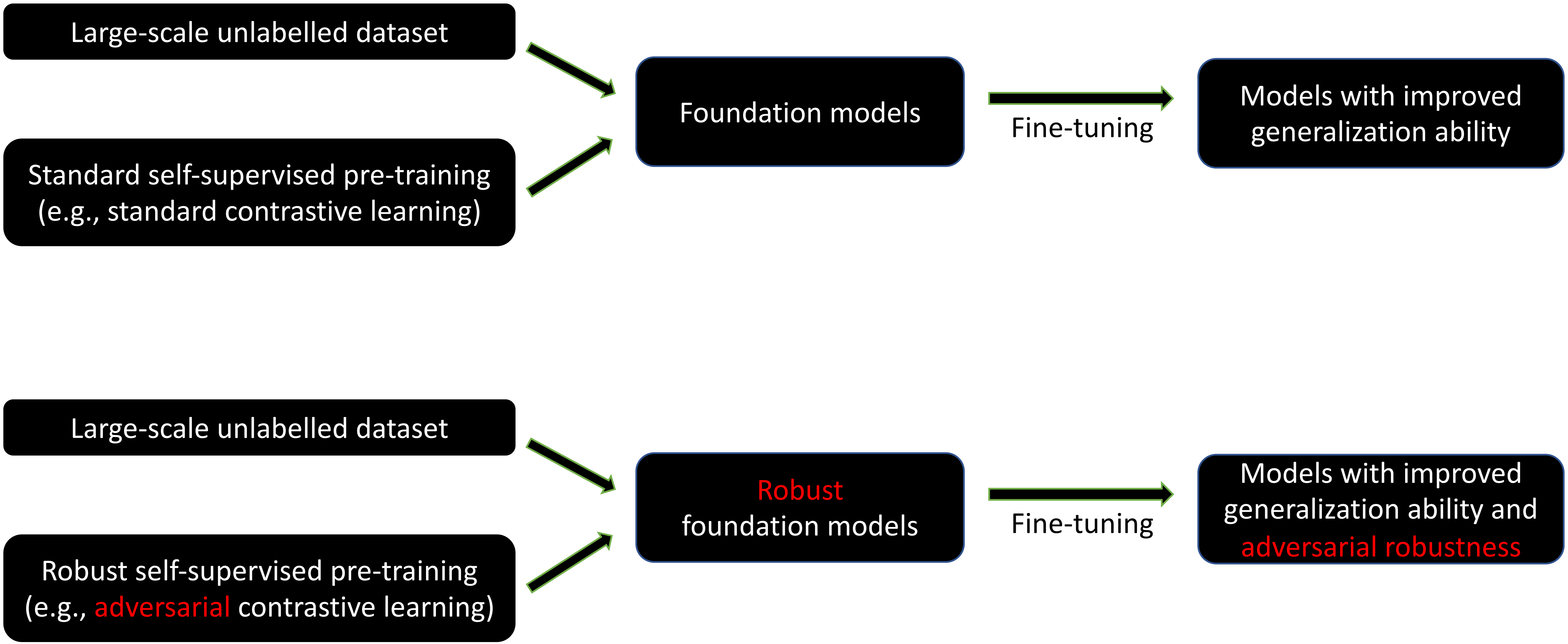


Small labelled datasets

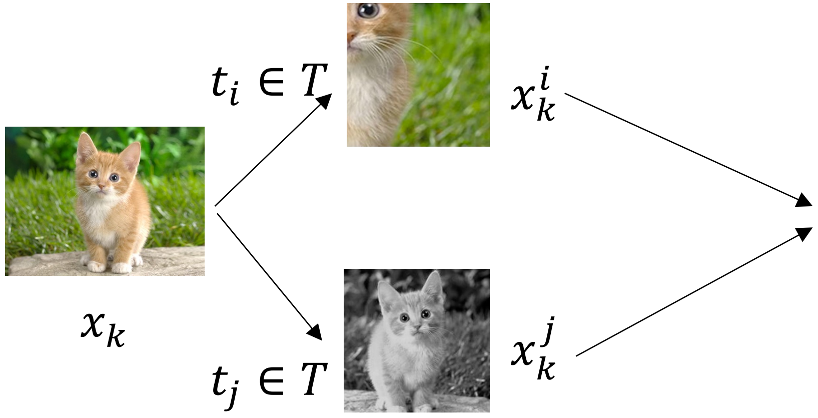
Robust pre-training



Robust pre-training



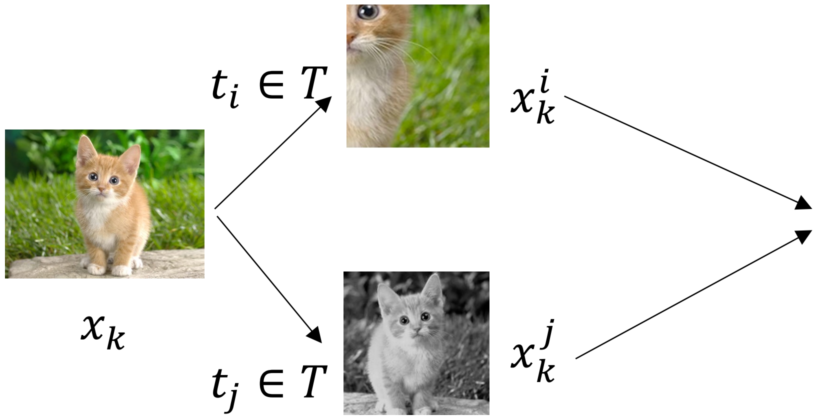
Adversarial contrastive learning (ACL)



Standard contrastive learning (e.g., SimCLR)

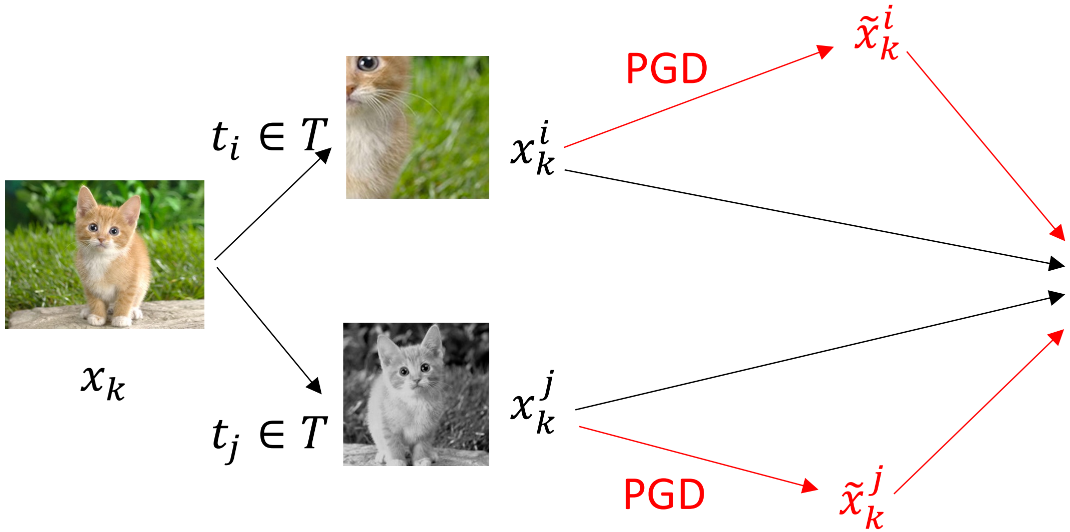
$$\ell_{\mathrm{CL}}(x_k^i, x_k^j; \theta) = -\sum_{u \in \{i,j\}} \log \frac{e^{\mathrm{sim}\left(f_\theta(x_k^i), f_\theta(x_k^j)\right)/t}}{\sum_{x \in B^i \cup B^j \setminus \{x_k^u\}} e^{\mathrm{sim}\left(f_\theta(x_k^u), f_\theta(x)\right)/t}},$$

ACL



Standard contrastive learning (e.g., SimCLR)

$$\ell_{\mathrm{CL}}(x_k^i, x_k^j;\theta) = -\sum_{u\in\{i,j\}} \log \frac{e^{\mathrm{sim}\left(f_\theta(x_k^i),\, f_\theta(x_k^j)\right)/t}}{\sum_{x\in B^i\cup B^j\setminus\{x_k^u\}} e^{\mathrm{sim}\left(f_\theta(x_k^u),\, f_\theta(x)\right)/t}},$$



The objective function of ACL

$$\ell_{\mathrm{ACL}}(x_k;\theta) = (1+\omega)\cdot \boxed{\ell_{\mathrm{CL}}(\tilde{x}_k^i, \tilde{x}_k^j;\theta)} + (1-\omega)\cdot \ell_{\mathrm{CL}}(x_k^i, x_k^j;\theta),$$

$$\text{where}\quad \tilde{x}_k^i, \tilde{x}_k^j = \mathop{\arg\max}_{\substack{\tilde{x}_k^i \in \mathcal{B}_\epsilon[x_k^i] \\ \tilde{x}_k^j \in \mathcal{B}_\epsilon[x_k^j]}} \ell_{\mathrm{CL}}(\tilde{x}_k^i, \tilde{x}_k^j;\theta),$$

Outline

- Backgrounds
  - Adversarial robustness
  - Robust pre-training
- How to make self-supervised robust pre-training
  - More efficient
  - More effective
- Future directions

Efficient ACL
via Robustness-aware Coreset Selection (RCS)

- Why do we need to speed up ACL?
  - ACL is extremely time-consuming.

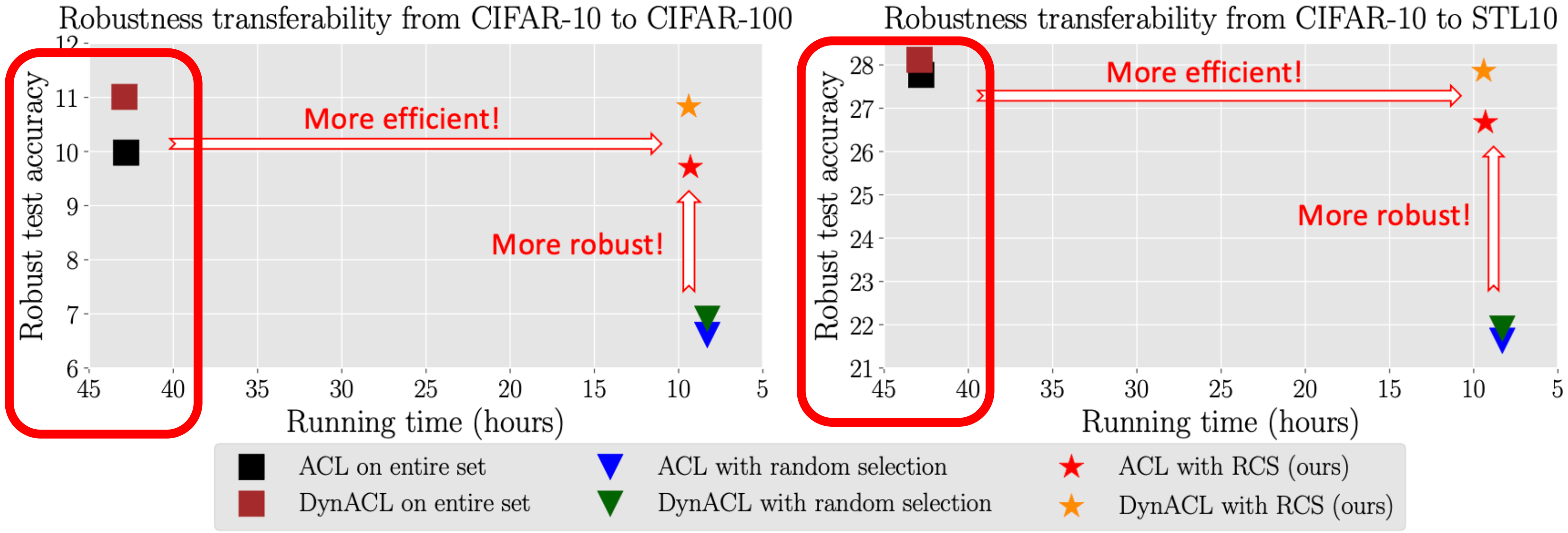


Figure 1: We learn a representation using CIFAR-10 [4] dataset (without requiring labels) via ACL [12] and DynACL [15]. Then, we evaluate the representation's robustness transferability to CIFAR-100 [4] and STL10 [22] (using labels during finetuning) via standard linear finetuning. We demonstrate the running time of robust pre-training w.r.t. different coreset selection (CS) strategies and report the robust test accuracy under AutoAttack [17]. Experimental details are in Appendix B.4.

Efficient ACL via RCS

- Why do we need to speed up ACL?
  - ACL is extremely time-consuming.
  - ACL has not been applied to ImageNet-1K yet due to computational prohibition.

Figure 4: Robustness evaluations on the CIFAR-10 (left three panels) and CIFAR-100 (right three panels) task. The number after the dash line denotes subset fraction $k \in \{0.05, 0.1, 0.2\}$.

Table 1: Robustness transferability from ImageNet-1K to CIFAR-10.

Pre-training	Runing time (hours)	SLF		ALF		AFF	
		SA (%)	RA (%)	SA (%)	RA (%)	SA (%)	RA (%)
Standard CL	147.4	84.36±0.17	0.01±0.01	10.00±0.00	10.00±0.00	86.63±0.12	49.71±0.16
ACL on entire set	650.2	-	-	-	-	-	-
ACL with Random	94.3	68.75±0.06	15.89±0.06	59.57±0.28	27.14±0.19	84.75±0.18	50.12±0.21
ACL with RCS	111.8	**70.02**±0.12	**22.45**±0.13	**63.94**±0.21	**31.13**±0.17	**85.23**±0.23	**52.21**±0.14

Table 2: Robustness transferability from ImageNet-1K to CIFAR-100.

Pre-training	Runing time (hours)	SLF		ALF		AFF	
		SA (%)	RA (%)	SA (%)	RA (%)	SA (%)	RA (%)
Standard CL	147.4	57.34±0.23	0.01±0.01	9.32±0.01	0.06±0.01	61.33±0.12	25.11±0.15
ACL on entire set	650.2	-	-	-	-	-	-
ACL with Random	94.3	38.53±0.15	10.50±0.13	28.44±0.23	11.93±0.21	59.63±0.33	25.46±0.26
ACL with RCS	111.8	**40.28**±0.17	**14.55**±0.10	**33.15**±0.26	**14.89**±0.16	**60.25**±0.18	**28.24**±0.13

Efficient ACL via RCS: Methodology

- Idea: Find an informative training subset
  - Decreasing the number of training samples
  - Preserving the robust representations

Efficient ACL via RCS: Methodology

- Idea: Find an informative training subset
- Intuitive solution: selects training data from the entire set whose gradients are most beneficial to maintaining adversarial robustness.

Efficient ACL via RCS: Methodology

- Idea: Find an informative training subset
- Intuitive solution: selects training data from the entire set whose gradients are most beneficial to maintaining adversarial robustness.
- Representational divergence (RD)
  - The smaller the RD is, the representations are of less sensitivity to adversarial perturbations, thus being more robust.

$$\ell_{\mathrm{RD}}(x;\theta) = d(g \circ f_\theta(\tilde{x}), g \circ f_\theta(x)) \quad \text{s.t.} \quad \tilde{x} = \arg\max_{x' \in \mathcal{B}_\epsilon[x]} d(g \circ f_\theta(x'), g \circ f_\theta(x))$$

Efficient ACL via RCS: Methodology

- Idea: Find an informative training subset
- Intuitive solution: selects training data from the entire set whose gradients are most beneficial to maintaining adversarial robustness.
- Representational divergence (RD)
- Objective function of RCS

Unlabeled validation set

$$S^* = \arg\min_{S \subseteq X, |S|/|X| \leq k} \mathcal{L}_{\mathrm{RD}}(U; \arg\min_{\theta} \mathcal{L}_{\mathrm{ACL}}(S; \theta))$$

Coreset

Subset fraction

Representational divergence (RD)

$$\ell_{\mathrm{RD}}(x; \theta) = d(g \circ f_\theta(\tilde{x}), g \circ f_\theta(x)) \quad \text{s.t.} \quad \tilde{x} = \arg\max_{x' \in \mathcal{B}_\epsilon[x]} d(g \circ f_\theta(x'), g \circ f_\theta(x))$$

Efficient ACL via RCS: Methodology

- Solve the objective function of RCS
  - Transformation of RCS

$$S^* = \underset{S \subseteq X, |S|/|X| \leq k}{\arg\min} \mathcal{L}_{\mathrm{RD}}(U; \arg\min_{\theta} \mathcal{L}_{\mathrm{ACL}}(S; \theta))$$

One-step gradient approximation

$$S^* = \underset{S \subseteq X, |S|/|X| \leq k}{\arg\min} \mathcal{L}_{\mathrm{RD}}(U; \theta - \eta \nabla_{\theta} \mathcal{L}_{\mathrm{ACL}}(S; \theta))$$

Transform into a problem of maximizing a set function subject to a cardinality constraint

$$S^* = \underset{S \subseteq X, |S|/|X| = k}{\arg\max} G_{\theta}(S)$$

$$G_{\theta}(S \subseteq X) \triangleq -\mathcal{L}_{\mathrm{RD}}(U; \theta - \eta \nabla_{\theta} \mathcal{L}_{\mathrm{ACL}}(S; \theta))$$

Efficient ACL via RCS: Methodology

- Solve the objective function of RCS
  - Transformation of RCS $S^* = \arg\max_{S \subseteq X, |S|/|X|=k} G_\theta(S)$ $G_\theta(S \subseteq X) \triangleq -\mathcal{L}_{\mathrm{RD}}(U; \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{ACL}}(S; \theta))$
  - Greedy search for solving a proxy set problem

$$\hat{S}^* = \arg\max_{S \subseteq X, |S|/|X|=k} \hat{G}_\theta(S)$$

Theorem 1. *We define a proxy set function* $\hat{G}_\theta(S) \triangleq G_\theta(S) + |S|\sigma$, *where* $\sigma = 1 + \nu_1 + \nu_2 L_2 + \eta M L_2(L_1 + \eta k N(L_1 L_4 + L_2 L_3))$, $\nu_1 \to 0^+$, *and* $\nu_2 > 0$ *are positive constants. Given Assumption 1,* $\hat{G}_\theta(S)$ *is monotone and* γ*-weakly submodular where* $\gamma > \gamma^* = \frac{1}{2\sigma - 1}$.

Efficient ACL via RCS: Methodology

- Solve the objective function of RCS
  - Transformation of RCS
  - Greedy search for solving a proxy set problem
  - Guaranteed lower bound of the original problem by solving the proxy set problem

$$S^* = \arg\max_{S \subseteq X, |S|/|X| = k} G_\theta(S) \quad G_\theta(S \subseteq X) \triangleq -\mathcal{L}_{\mathrm{RD}}(U; \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{ACL}}(S; \theta))$$

$$\hat{S}^* = \arg\max_{S \subseteq X, |S|/|X| = k} \hat{G}_\theta(S)$$

Theorem 2. *Given a fixed parameter θ, we denote the optimal solution of Eq. (5) as $G_\theta^* = \sup_{S \subseteq X, |S|/|X| = k} G_\theta(S)$. Then, $\hat{S}^*$ in Eq. (6) found via greedy search satisfies*

$$G_\theta(\hat{S}^*) \geq G_\theta^* - (G_\theta^* + kN\sigma) \cdot e^{-\gamma^*}.$$

Efficient ACL via RCS: Methodology

- Solve the objective function of RCS
  - Transformation of RCS $S^* = \underset{S \subseteq X, |S|/|X|=k}{\arg\max} G_\theta(S)$ $G_\theta(S \subseteq X) \triangleq -\mathcal{L}_{\mathrm{RD}}(U; \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{ACL}}(S; \theta))$
  - Greedy search for solving a proxy set problem $\hat{S}^* = \underset{S \subseteq X, |S|/|X|=k}{\arg\max} \hat{G}_\theta(S)$
  - Guaranteed lower bound of the original problem by solving the proxy set problem
  - Algorithm

Algorithm 1 Robustness-aware Coreset Selection (RCS)

Input: Unlabeled training set X, unlabeled validation set U, batch size β, model $g \circ f_\theta$, learning rate for RCS η, subset fraction $k \in (0, 1]$
Output: Coreset S
Initialize $S \leftarrow \emptyset$
Split entire set into minibatches $X = \{B_m\}_{m=1}^{\lceil |X|/\beta \rceil}$
for each minibatch $B_m \in X$ **do**
  Compute gradient $q_m \leftarrow \nabla_\theta \mathcal{L}_{\mathrm{ACL}}(B_m; \theta)$
end for
// Conduct greedy search via batch-wise selection
for $1, \dots, \lfloor k|X|/\beta \rfloor$ **do**
  Compute gradient $q_U \leftarrow \nabla_\theta \mathcal{L}_{\mathrm{RD}}(U; \theta)$
  Initialize $best\_gain = -\infty$
  for each minibatch $B_m \in X$ **do**
    Compute marginal gain $\hat{G}(B_m|S) \leftarrow \eta q_U^\top q_m$
    if $\hat{G}(B_m|S) > best\_gain$ **then**
      Update $s \leftarrow m$, $best\_gain \leftarrow \hat{G}(B_m|S)$
    end if
  end for
  Update $S \leftarrow S \cup B_s$, $X \leftarrow X \setminus B_s$
  Update $\theta \leftarrow \theta - \eta q_s$
end for

Algorithm 2 Efficient ACL via RCS

Input: Unlabeled training set X, unlabeled validation set U, total training epochs E, learning rate η', batch size β, warmup epoch ω, epoch interval for executing RCS λ, subset fraction k, learning rate for RCS η
Output: Adversarially pre-trained feature extractor f_θ
Initialize parameters of model $g \circ f_\theta$
Initialize training set $S \leftarrow X$
for $e = 0$ **to** $E - 1$ **do**
  if $e\%\lambda == 0$ **and** $e \geq \omega$ **then**
    $S \leftarrow \mathrm{RCS}(X, U, \beta, g \circ f_\theta, \eta, k)$ //by Algorithm 1
  end if
  for batch $m = 1, \dots, \lceil |S|/\beta \rceil$ **do**
    Sample a minibatch B_m from S
    Update $\theta \leftarrow \theta - \eta' \nabla_\theta \mathcal{L}_{\mathrm{ACL}}(B_m; \theta)$
  end for
end for

Efficient ACL via RCS: Empirical results

- Our proposed RCS is
  - more efficient (higher speed-up ratio)
  - more effective (higher test accuracy)

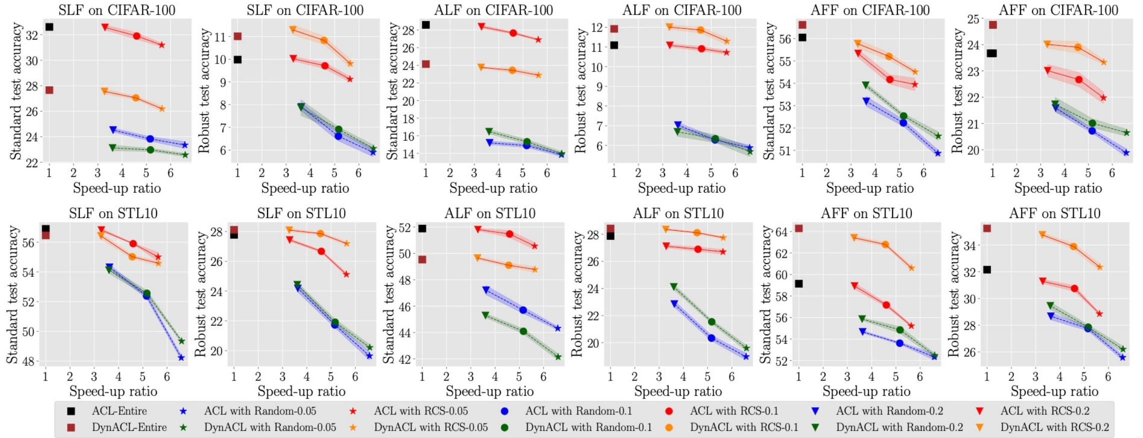


Figure 2: Robustness transferability from CIFAR-10 to CIFAR-100 (upper row) and STL10 (bottom row). The number after the dash line denotes subset fraction $k \in \{0.05, 0.1, 0.2\}$.

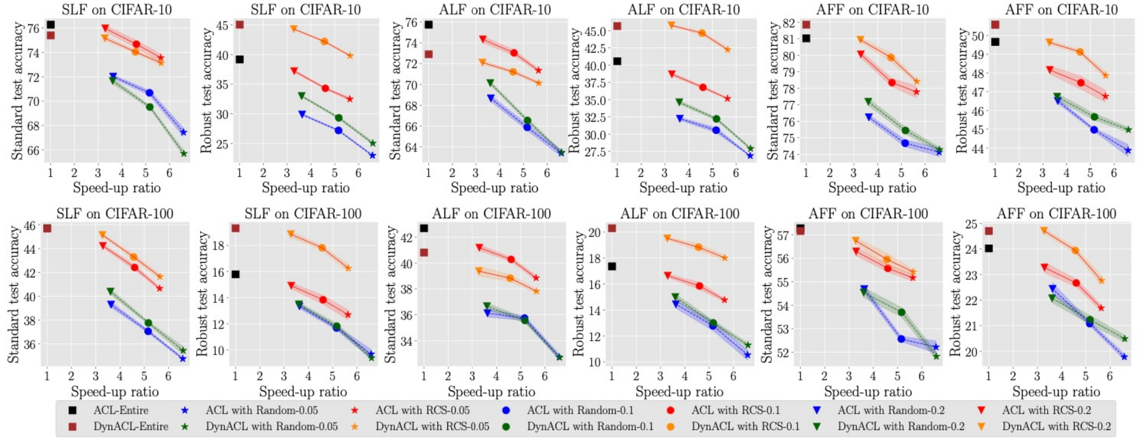


Figure 4: Robustness evaluations on the CIFAR-10 (left three panels) and CIFAR-100 (right three panels) task. The number after the dash line denotes subset fraction $k \in \{0.05, 0.1, 0.2\}$.

The upper-right (ours) is better!

Efficient ACL via RCS: Empirical results

- For the first time to conduct ACL on ImageNet-1K using WRN-28-10

Figure 4: Robustness evaluations on the CIFAR-10 (left three panels) and CIFAR-100 (right three panels) task. The number after the dash line denotes subset fraction $k \in \{0.05, 0.1, 0.2\}$.

Table 1: Robustness transferability from ImageNet-1K to CIFAR-10.

Pre-training	Runing time (hours)	SLF		ALF		AFF	
		SA (%)	RA (%)	SA (%)	RA (%)	SA (%)	RA (%)
Standard CL	147.4	84.36±0.17	0.01±0.01	10.00±0.00	10.00±0.00	86.63±0.12	49.71±0.16
ACL on entire set	650.2	-	-	-	-	-	-
ACL with Random	94.3	68.75±0.06	15.89±0.06	59.57±0.28	27.14±0.19	84.75±0.18	50.12±0.21
ACL with RCS	111.8	**70.02**±0.12	**22.45**±0.13	**63.94**±0.21	**31.13**±0.17	**85.23**±0.23	**52.21**±0.14

Table 2: Robustness transferability from ImageNet-1K to CIFAR-100.

Pre-training	Runing time (hours)	SLF		ALF		AFF	
		SA (%)	RA (%)	SA (%)	RA (%)	SA (%)	RA (%)
Standard CL	147.4	57.34±0.23	0.01±0.01	9.32±0.01	0.06±0.01	61.33±0.12	25.11±0.15
ACL on entire set	650.2	-	-	-	-	-	-
ACL with Random	94.3	38.53±0.15	10.50±0.13	28.44±0.23	11.93±0.21	59.63±0.33	25.46±0.26
ACL with RCS	111.8	**40.28**±0.17	**14.55**±0.10	**33.15**±0.26	**14.89**±0.16	**60.25**±0.18	**28.24**±0.13

Efficient ACL via RCS: Empirical results

- RCS for speeding up SAT on ImageNet-1K (supervised setting)
  - Maintaining standard transferability
  - Maintaining robustness transferability

Table 18: Standard transferability [43] of adversarially pre-trained ResNet-50 from ImageNet-1K to CIFAR-10 and CIFAR-100, respectively. We report the standard test accuracy (%) via standard linear finetuning (SLF) and standard full finetuning (SFF). The number after the dash line denotes subset fraction $k \in \{0.05, 0.1, 0.2\}$.

Pre-training	Runing time (hours)	CIFAR-10		CIFAR-100	
		SLF	SFF	SLF	SFF
Standard training [43] on entire set	-	78.84	97.41	57.09	84.21
SAT [43] on entire set	286.1	93.53	98.09	77.29	86.99
Fast-AT [20] on entire set	10.4	90.91	97.54	73.35	83.33
SAT with Random-0.05	38.7	85.72	95.27	69.29	82.34
SAT with RCS-0.05	**48.2**	**92.68**	**97.65**	**75.35**	**84.71**
SAT with Random-0.1	45.8	87.14	95.60	71.23	83.62
SAT with RCS-0.1	**55.4**	**92.92**	**97.82**	**75.41**	**85.22**
SAT with Random-0.2	70.3	87.69	96.10	72.05	84.14
SAT with RCS-0.2	**79.8**	**93.48**	**98.06**	**76.39**	**85.44**

Table 16: Robustness transferability of adversarially pre-trained WRN-28-10 from ImageNet-1K to CIFAR-10. Here, "RA" stands for robust test accuracy under PGD-20 attacks following the setting of Hendrycks et al. [51]. The number after the dash line denotes subset fraction $k \in \{0.05, 0.1, 0.2\}$.

Pre-training	Runing time (hours)	ALF		AFF	
		SA (%)	RA (%)	SA (%)	RA (%)
Standard training on entire set	66.7	10.12	10.04	84.73	51.91
SAT [51] on entire set	341.7	85.90	50.89	89.35	59.68
SAT with Random-0.05	53.6	69.59	31.58	85.55	53.53
SAT with RCS-0.05	**68.6**	**79.72**	**44.44**	**87.99**	**56.87**
SAT with Random-0.1	70.2	73.28	33.86	86.78	54.95
SAT with RCS-0.1	**81.9**	**81.92**	**45.10**	**88.87**	**57.69**
SAT with Random-0.2	103.4	75.46	39.62	86.64	56.46
SAT with RCS-0.2	**121.9**	**83.94**	**46.88**	**89.54**	**58.13**

Efficient ACL via RCS: Conclusions

- We proposed robustness-aware coreset selection (RCS) that can
  - speed up (supervised and self-supervised) robust pre-training
  - maintain (standard and robustness) transferability

Outline

- Backgrounds
  - Adversarial attack and defense
  - Robust pre-training
- How to make self-supervised robust pre-training
  - More efficient
  - More effective
- Future directions

Effective ACL
via adversarial invariant regularization (AIR)

- Motivation
  - The style-independence property of learned representations, which eliminates the effects of nuisance style factors in standard contrastive learning (SCL), has been shown to significantly improve the transferability of representations.

Algorithm	Shorthand	Paper	KNN accuracy
Bootstrap Your Own Latent: A new approach to self-supervised Learning	BYOL	arXiv	80.09
Representation Learning via Invariant Causal Mechanisms	ReLIC	arXiv	79.26
A Simple Framework for Contrastive Learning of Visual Representations	SimCLR	arXiv	77.79
Unsupervised Learning of Visual Features by Contrasting Cluster Assignments	SwAV	arXiv	72.11
Momentum Contrast for Unsupervised Visual Representation Learning	MoCo	arXiv	63.14
Barlow Twins: Self-Supervised Learning via Redundancy Reduction	Barlow	arXiv	56.81

[Mitrovic et al., ICLR 2021]

Performance evaluated on CIFAR-10

Image from https://github.com/NightShade99/Self-Supervised-Vision

Effective ACL via AIR

- Motivation
  - The style-independence property of learned representations, which eliminates the effects of nuisance style factors in standard contrastive learning (SCL), has been shown to improve the transferability of representations.

It is unclear how the style-independence property benefits ACL-learned robust representations.

Effective ACL via AIR : Methodology

- ACL in the view of causality

$$\ell_{\mathrm{ACL}}(x_k;\theta) = (1+\omega)\cdot \ell_{\mathrm{CL}}(\tilde{x}_k^i, \tilde{x}_k^j;\theta) + (1-\omega)\cdot \ell_{\mathrm{CL}}(x_k^i, x_k^j;\theta),$$

$$\text{where} \quad \tilde{x}_k^i, \tilde{x}_k^j = \mathop{\arg\max}_{\substack{\tilde{x}_k^i \in \mathcal{B}_\epsilon[x_k^i] \\ \tilde{x}_k^j \in \mathcal{B}_\epsilon[x_k^j]}} \ell_{\mathrm{CL}}(\tilde{x}_k^i, \tilde{x}_k^j;\theta),$$

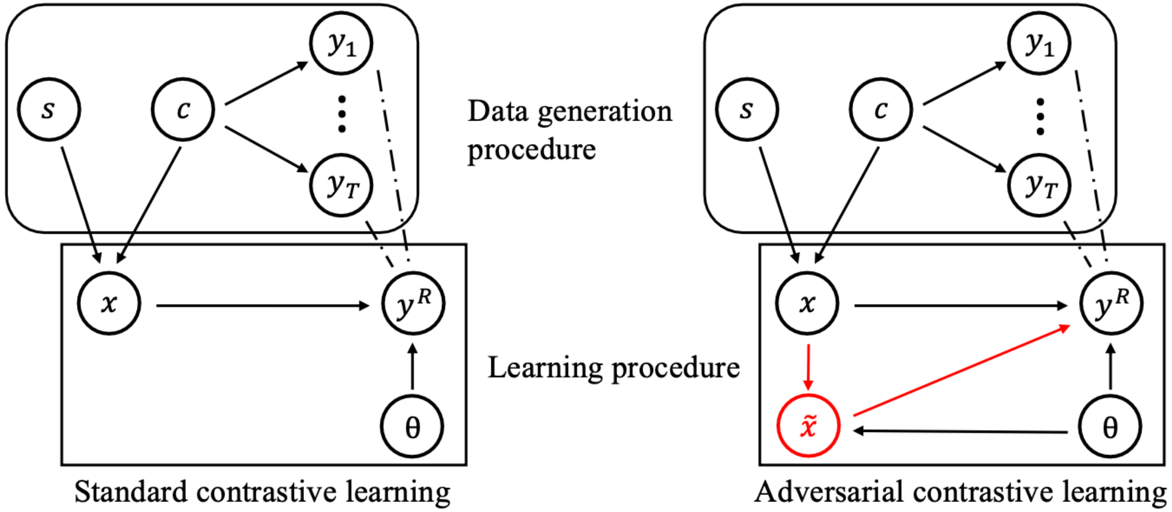


Figure 1: Causal graph of standard contrastive learning [35] (left panel) and adversarial contrastive learning (right panel). x is unlabeled data, s is style variable, c is content variable, $\tilde{x}$ is the generated adversarial data, and θ is the parameter of representation. The dashdotted lines denote that the proxy label $y^R \in \mathcal{Y}^R$ is a refinement of the target label $y_t \in \mathcal{Y} = \{y_i\}_{i=1}^T$. All other arrows are causal.

Effective ACL via AIR : Methodology

- ACL in the view of causality

$$\ell_{\mathrm{ACL}}(x_k;\theta) = (1+\omega)\cdot \ell_{\mathrm{CL}}(\tilde{x}_k^i, \tilde{x}_k^j;\theta) + (1-\omega)\cdot \ell_{\mathrm{CL}}(x_k^i, x_k^j;\theta),$$

$$\text{where} \quad \tilde{x}_k^i, \tilde{x}_k^j = \mathop{\arg\max}_{\substack{\tilde{x}_k^i \in \mathcal{B}_\epsilon[x_k^i] \\ \tilde{x}_k^j \in \mathcal{B}_\epsilon[x_k^j]}} \ell_{\mathrm{CL}}(\tilde{x}_k^i, \tilde{x}_k^j;\theta),$$

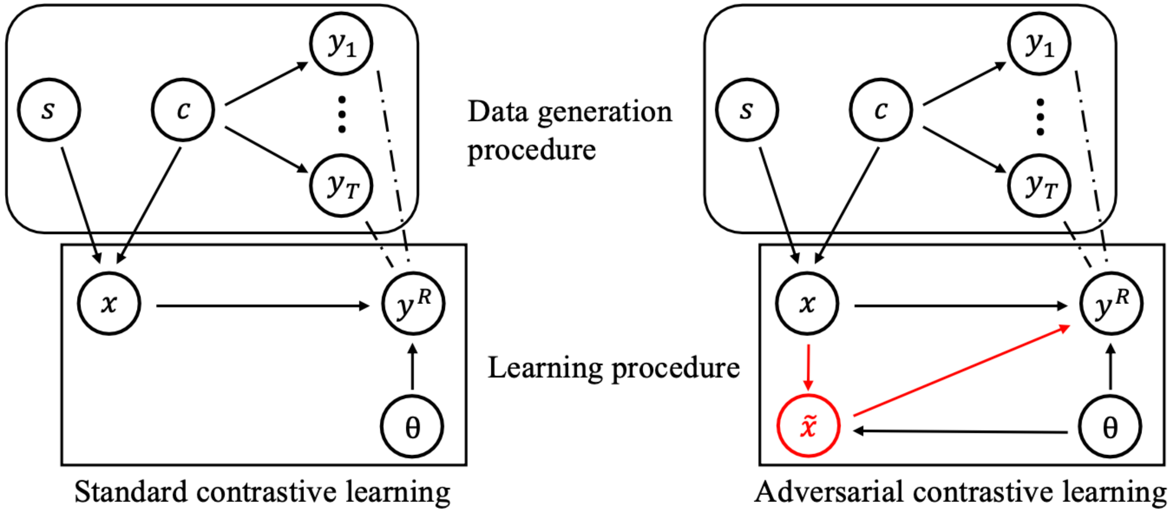
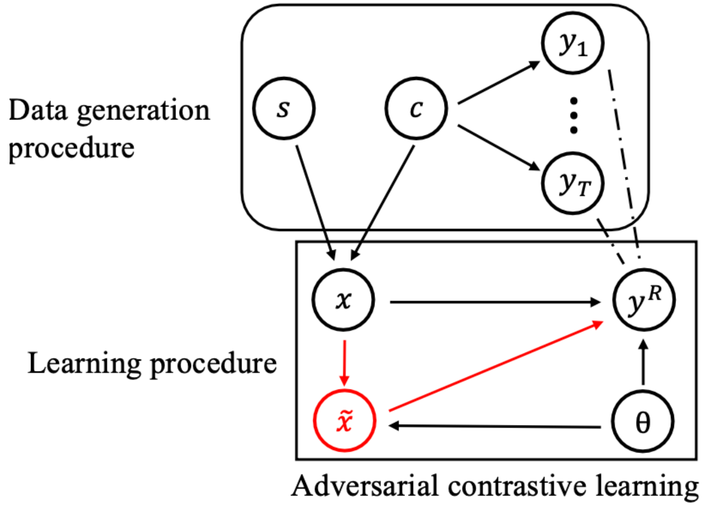


Figure 1: Causal graph of standard contrastive learning [35] (left panel) and adversarial contrastive learning (right panel). x is unlabeled data, s is style variable, c is content variable, $\tilde{x}$ is the generated adversarial data, and θ is the parameter of representation. The dashdotted lines denote that the proxy label $y^R \in \mathcal{Y}^R$ is a refinement of the target label $y_t \in \mathcal{Y} = \{y_i\}_{i=1}^T$. All other arrows are causal.

The rationality of the causal graph

Theorem 1. *The learning objective of the proxy task used in ACL which is to maximize the conditional probability both $p(y^R|x)$ and $p(y^R|\tilde{x})$ is equivalent to the learning objective of ACL [26] which is to minimize the sum of standard and adversarial contrastive losses.*

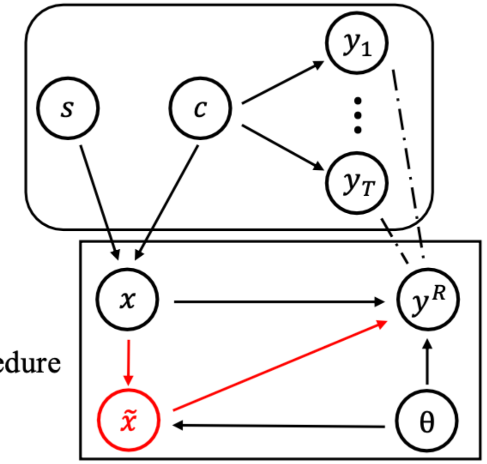
Effective ACL via AIR: Methodology



- Adversarial invariant regularization (AIR)
  - The conditional probability learned via ACL

$$p(y^R|x) = p(y^R|\tilde{x})p(\tilde{x}|x)$$

Effective ACL via AIR: Methodology

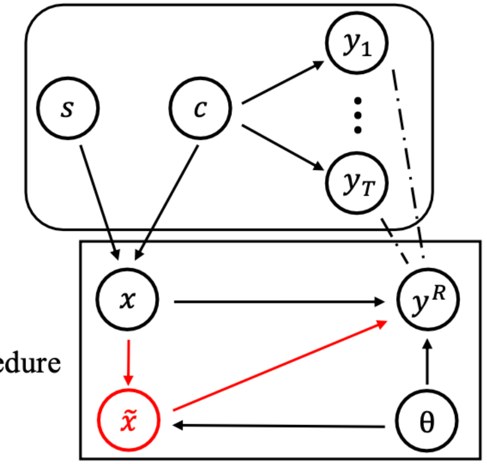


- Adversarial invariant regularization (AIR)
  - The conditional probability learned via ACL $p(y^R|x) = p(y^R|\tilde{x})p(\tilde{x}|x)$
  - Style-independent criterion

$$p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) = p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x) \quad \forall \tau_i, \tau_j \in \mathcal{T},$$

$$p^{do(\tau_u)}(y^R|\tilde{x}) = \frac{e^{\mathrm{sim}(f_\theta(x), f_\theta(\tilde{x}^u))/t}}{\sum_{x_k \in B} e^{\mathrm{sim}\left(f_\theta(x_k), f_\theta(\tilde{x}_k^u)\right)/t}}, \quad p^{do(\tau_u)}(\tilde{x}|x) = \frac{e^{\mathrm{sim}(f_\theta(\tilde{x}^u), f_\theta(x^u))/t}}{\sum_{x_k \in B} e^{\mathrm{sim}\left(f_\theta(\tilde{x}_k^u), f_\theta(x_k^u)\right)/t}}$$

Effective ACL via AIR: Methodology

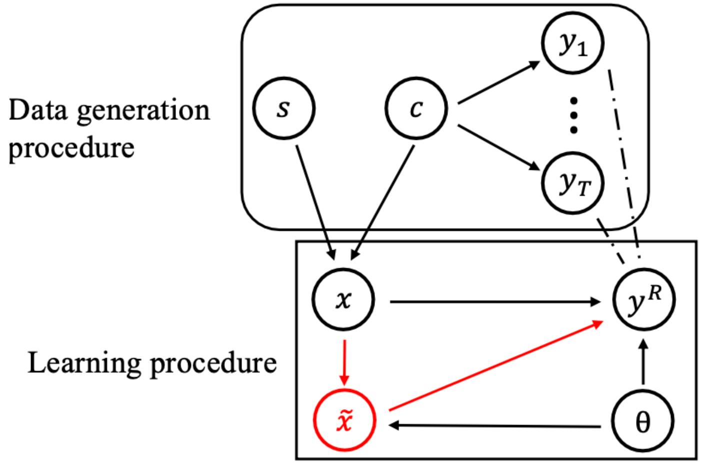


- Adversarial invariant regularization (AIR)
  - The conditional probability learned via ACL $p(y^R|x) = p(y^R|\tilde{x})p(\tilde{x}|x)$
  - Style-independent criterion $p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) = p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x) \quad \forall \tau_i, \tau_j \in \mathcal{T},$
  - Loss function of AIR

$$\mathcal{L}_{\text{AIR}}(B;\theta) = \text{KL}\left(p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) \| p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x); B\right)$$

$$p^{do(\tau_u)}(y^R|\tilde{x}) = \frac{e^{\text{sim}(f_\theta(x), f_\theta(\tilde{x}^u))/t}}{\sum_{x_k \in B} e^{\text{sim}\left(f_\theta(x_k), f_\theta(\tilde{x}_k^u)\right)/t}}, \quad p^{do(\tau_u)}(\tilde{x}|x) = \frac{e^{\text{sim}(f_\theta(\tilde{x}^u), f_\theta(x^u))/t}}{\sum_{x_k \in B} e^{\text{sim}\left(f_\theta(\tilde{x}_k^u), f_\theta(x_k^u)\right)/t}}$$

Effective ACL via AIR: Methodology



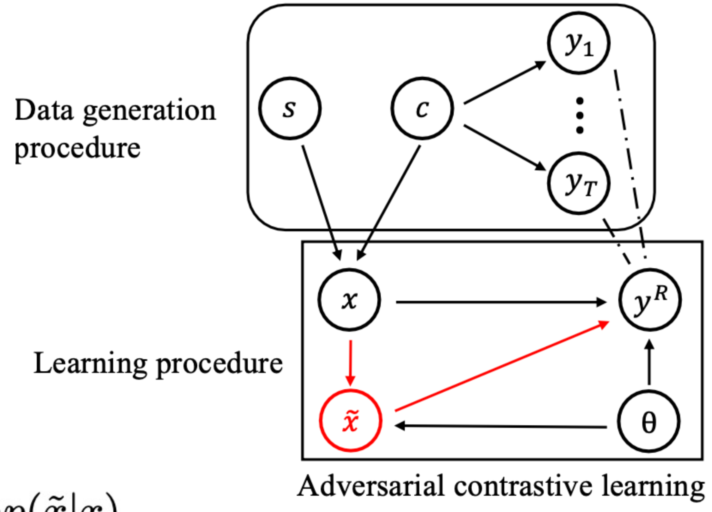
Adversarial contrastive learning

- Adversarial invariant regularization (AIR)
  - The conditional probability learned via ACL $p(y^R|x) = p(y^R|\tilde{x})p(\tilde{x}|x)$
  - Style-independent criterion $p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) = p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x) \quad \forall \tau_i, \tau_j \in \mathcal{T},$
  - Loss function of AIR $\mathcal{L}_{\mathrm{AIR}}(B;\theta) = \mathrm{KL}\left(p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) \| p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x); B\right)$
  - Standard invariant regularization (SIR): a special case of AIR

$$\mathcal{L}_{\mathrm{SIR}}(B;\theta) = \mathrm{KL}\left(p^{do(\tau_i)}(y^R|x) \| p^{do(\tau_i)}(y^R|x); B\right),$$

$$\text{where} \quad p^{do(\tau_u)}(y^R|x) = \frac{e^{\mathrm{sim}(f_\theta(x), f_\theta(x^u))/t}}{\sum_{x_k \in B} e^{\mathrm{sim}\left(f_\theta(x_k), f_\theta(x_k^u)\right)/t}} \quad \forall u \in \{i, j\}$$

Effective ACL via AIR: Methodology



- Adversarial invariant regularization (AIR)
  - The conditional probability learned via ACL $p(y^R|x) = p(y^R|\tilde{x})p(\tilde{x}|x)$
  - Style-independent criterion $p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) = p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x) \quad \forall \tau_i, \tau_j \in \mathcal{T},$
  - Loss function of AIR $\mathcal{L}_{\text{AIR}}(B;\theta) = \text{KL}\left(p^{do(\tau_i)}(y^R|\tilde{x})p^{do(\tau_i)}(\tilde{x}|x) \| p^{do(\tau_j)}(y^R|\tilde{x})p^{do(\tau_j)}(\tilde{x}|x); B\right)$
  - SIR: a special case of AIR $\mathcal{L}_{\text{SIR}}(B;\theta) = \text{KL}\left(p^{do(\tau_i)}(y^R|x) \| p^{do(\tau_i)}(y^R|x); B\right)$
- Our proposed invariant regularization (IR)

$$\arg\min_{\theta} \sum_{x \in U} \ell_{\text{ACL}}(x;\theta) + \underbrace{\lambda_1 \cdot \mathcal{L}_{\text{SIR}}(U;\theta) + \lambda_2 \cdot \mathcal{L}_{\text{AIR}}(U;\theta)}_{\textit{invariant regularization}},$$

Effective ACL via AIR: Theoretical analysis

- Theoretical justification of the effectiveness
  - The style-independence property is generalizable to the downstream tasks

Proposition 4. *Let $\mathcal{Y} = \{y_t\}_{t=1}^{T}$ be a label set of a downstream classification task, $\mathcal{Y}^R$ be a refinement of $\mathcal{Y}$, and $\tilde{x}_t$ be the adversarial data generated on the downstream task. Assuming that $\tilde{x} \in \mathcal{B}_\epsilon[x]$ and $\tilde{x}_t \in \mathcal{B}_\epsilon[x]$, we have the following results:*

$$p^{do(\tau_i)}(y^R|\tilde{x}) = p^{do(\tau_j)}(y^R|\tilde{x}) \Longrightarrow p^{do(\tau_i)}(y_t|\tilde{x}_t) = p^{do(\tau_j)}(y_t|\tilde{x}_t) \quad \forall \tau_i, \tau_j \in \mathcal{T},$$

$$p^{do(\tau_i)}(\tilde{x}|x) = p^{do(\tau_j)}(\tilde{x}|x) \Longrightarrow p^{do(\tau_i)}(\tilde{x}_t|x) = p^{do(\tau_j)}(\tilde{x}_t|x) \quad \forall \tau_i, \tau_j \in \mathcal{T}.$$

Effective ACL via AIR: Theoretical analysis

- Theoretical justification of the effectiveness
  - The style-independence property is generalizable to the downstream tasks

Proposition 4. *Let $\mathcal{Y} = \{y_t\}_{t=1}^{T}$ be a label set of a downstream classification task, $\mathcal{Y}^R$ be a refinement of $\mathcal{Y}$, and $\tilde{x}_t$ be the adversarial data generated on the downstream task. Assuming that $\tilde{x} \in \mathcal{B}_\epsilon[x]$ and $\tilde{x}_t \in \mathcal{B}_\epsilon[x]$, we have the following results:*

$$p^{do(\tau_i)}(y^R|\tilde{x}) = p^{do(\tau_j)}(y^R|\tilde{x}) \Longrightarrow p^{do(\tau_i)}(y_t|\tilde{x}_t) = p^{do(\tau_j)}(y_t|\tilde{x}_t) \quad \forall \tau_i, \tau_j \in \mathcal{T},$$
$$p^{do(\tau_i)}(\tilde{x}|x) = p^{do(\tau_j)}(\tilde{x}|x) \Longrightarrow p^{do(\tau_i)}(\tilde{x}_t|x) = p^{do(\tau_j)}(\tilde{x}_t|x) \quad \forall \tau_i, \tau_j \in \mathcal{T}.$$

  - We can treat adversarial attacks and common corruptions as style factors
  - IR regulates the representations to be invariant of style factors

Effective ACL via AIR: Experimental results

- Performance evaluated on various tasks

Table 1: Robustness evaluations via SLF across various tasks.

Pre-training	λ_1	λ_2	CIFAR-10 AA (%)	CIFAR-10 SA (%)	CIFAR-100 AA (%)	CIFAR-100 SA (%)	STL10 AA (%)	STL10 SA (%)
ACL [26]	0.0	0.0	37.39±0.06	78.27±0.09	15.78±0.05	45.70±0.09	35.80±0.06	67.90±0.09
ACL with SIR [35]	0.5	0.0	37.51±0.04	78.97±0.08	15.76±0.06	47.16±0.11	36.36±0.09	68.09±0.13
ACL with AIR	0.0	0.5	38.70±0.09	79.96±0.05	16.03±0.12	49.60±0.15	36.86±0.08	68.61±0.10
ACL with IR	0.5	0.5	**38.89**±0.06	**80.03**±0.07	**16.14**±0.07	**49.75**±0.10	**36.94**±0.06	**68.91**±0.07
DynACL [19]	0.0	0.0	45.05±0.04	75.39±0.05	19.31±0.06	45.67±0.09	46.49±0.05	69.59±0.08
DynACL with SIR [35]	0.5	0.0	44.70±0.03	76.45±0.06	19.67±0.09	46.13±0.10	46.56±0.08	70.41±0.09
DynACL with AIR	0.0	0.5	45.23±0.08	78.01±0.11	20.37±0.08	46.77±0.11	47.62±0.07	71.98±0.12
DynACL with IR	0.5	0.5	**45.27**±0.04	**78.08**±0.06	**20.45**±0.07	**46.84**±0.12	**47.66**±0.06	**72.30**±0.10

- Performance evaluated via various fine-tuning methods

Table 2: Robustness benchmark on the CIFAR-10 task evaluated via SLF, ALF, and AFF.

Pre-training	λ_1	λ_2	SLF AA (%)	SLF SA (%)	ALF AA (%)	ALF SA (%)	AFF AA (%)	AFF SA (%)
ACL [26]	0.0	0.0	37.39±0.06	78.27±0.09	40.61±0.07	75.56±0.09	49.42±0.07	82.14±0.18
ACL with SIR [35]	0.5	0.0	37.51±0.04	78.97±0.08	40.30±0.08	76.49±0.05	50.36±0.07	82.62±0.08
ACL with AIR	0.0	0.5	38.70±0.09	79.96±0.05	41.09±0.06	77.99±0.12	50.32±0.09	82.67±0.09
ACL with IR	0.5	0.5	**38.89**±0.06	**80.03**±0.07	**41.39**±0.08	**78.29**±0.10	**50.44**±0.04	**82.71**±0.06
DynACL [33]	0.0	0.0	45.05±0.04	75.39±0.05	45.65±0.05	72.90±0.08	50.52±0.05	81.86±0.11
DynACL with SIR [35]	0.5	0.0	44.70±0.03	76.45±0.06	45.42±0.10	74.78±0.14	50.58±0.07	81.66±0.18
DynACL with AIR	0.0	0.5	45.23±0.08	78.01±0.11	46.12±0.09	77.01±0.12	50.66±0.05	82.62±0.10
DynACL with IR	0.5	0.5	**45.27**±0.04	**78.08**±0.06	**46.14**±0.07	**77.42**±0.10	**50.68**±0.08	**82.74**±0.11

- Robustness under common corruption

Table 3: Test accuracy (%) evaluated on CIFAR-10-C (corruption severity ranges from 1 to 5) of CIFAR-10 pre-trained models after SLF and AFF, respectively. Standard deviation is in Table 20.

Pre-training	λ_1	λ_2	SLF CS-1	SLF CS-2	SLF CS-3	SLF CS-4	SLF CS-5	AFF CS-1	AFF CS-2	AFF CS-3	AFF CS-4	AFF CS-5
ACL [26]	0.0	0.0	76.57	74.73	71.78	67.75	62.78	79.15	76.01	72.54	69.47	65.27
ACL with SIR [35]	0.5	0.0	77.31	75.46	72.21	68.14	63.27	79.05	76.29	72.73	69.43	65.29
ACL with AIR	0.0	0.5	78.30	76.34	73.27	69.10	64.24	79.24	76.54	72.81	69.64	65.32
ACL with IR	0.5	0.5	**78.55**	**76.67**	**73.33**	**69.12**	**64.28**	**79.49**	**76.86**	**72.95**	**69.73**	**65.37**
DynACL [33]	0.0	0.0	73.92	71.69	69.01	66.22	62.51	79.77	76.44	72.95	69.74	65.60
DynACL with SIR [35]	0.5	0.0	75.81	72.88	69.31	66.24	62.20	80.59	77.31	73.67	70.39	66.05
DynACL with AIR	0.0	0.5	76.33	73.46	69.97	67.19	63.13	80.93	77.71	74.11	70.81	66.58
DynACL with IR	0.5	0.5	**76.62**	**73.62**	**70.16**	**67.37**	**63.29**	**80.98**	**77.87**	**74.31**	**70.96**	**66.75**

Effective ACL via AIR: Conclusions

- We proposed an invariant regularization that can
  - regulate (both standard and robust) representations to be style-independent
  - improve both generalization ability and robustness transferability against adversarial attacks and common corruptions

Thank you for your attention!

- Summary
  - More efficient robust pre-training via robustness-aware coreset selection
  - More effective robust pre-training via adversarial invariant regularization

- Future directions
  - The application of robust foundation models in computer vision tasks
    - Segmentation
    - Point cloud classification
    - Human-object interaction detection
    - …
  - The potential of robust self-supervised pre-training in building robust language foundation models