SINGAPORE 2024

ICLR
International Conference On Learning Representations

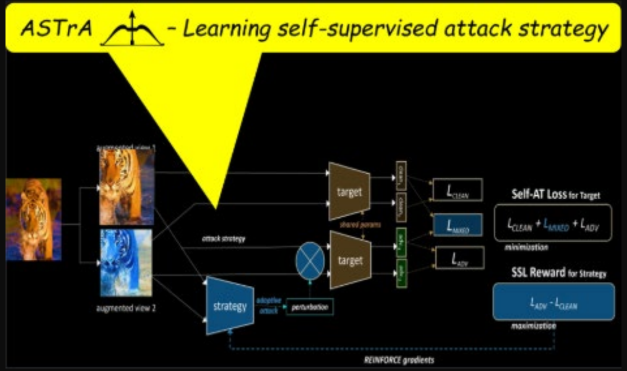
UCF
CENTER FOR RESEARCH IN COMPUTER VISION

LULEÅ TEKNISKA UNIVERSITET

# ASTrA: Adversarial Self-supervised Training with Adaptive-Attack

ICLR 2025 Singapore

https://prakashchhipa.github.io/projects/ASTrA



ASTrA - Learning self-supervised attack strategy

Prakash Chandra Chhipa[1]*, Gautam Vashishtha[2]*, Settur Jithamanyu[3]*, Rajkumar Saini[1], Mubarak Shah[4] , Marcus Liwicki[1]

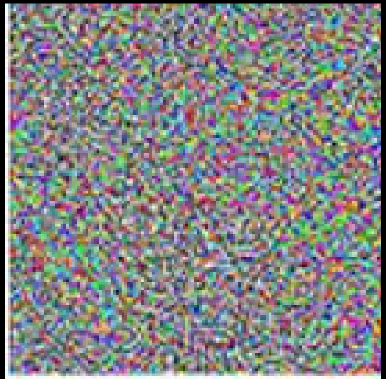
[1]Machine Learning Group, Lulea˚ Tekniska Universitet, Sweden
[2]Indian Institute of Technology, Gandinagar
[3]Indian Institute of Technology, Madras
[4]Center For Research in Computer Vision, University of Central Florida, USA

*equal contribution
presenter

# Self-supervised adversarial attacks – limitation

- Networks vulnerability to adversarial examples.

cat



δ=8/255.

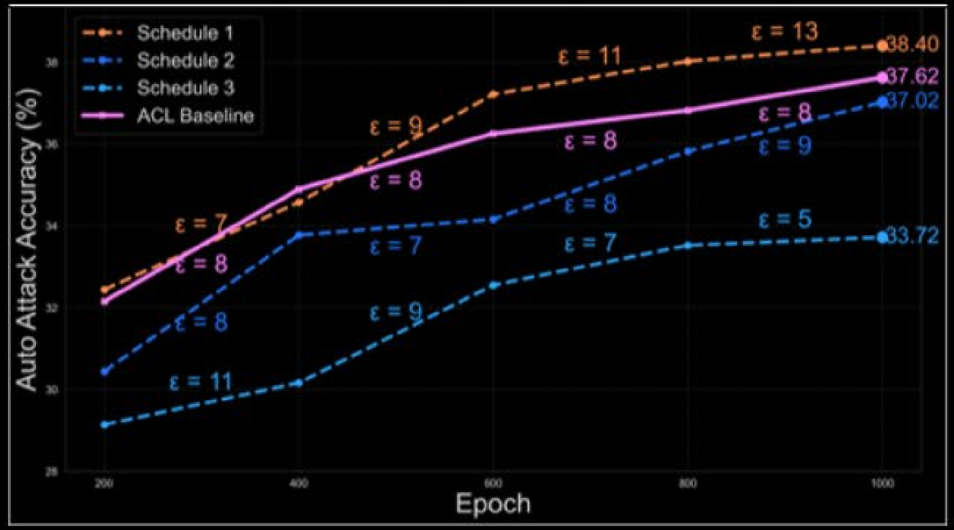
airliner

- Limitation: Hand-crafted adversarial attack strategy fail to adapt dynamically in Self-supervised adversarial training (Self-AT).
  - Does not align with model's learning dynamics
  - No correspondence between training examples and attack strategy parameters

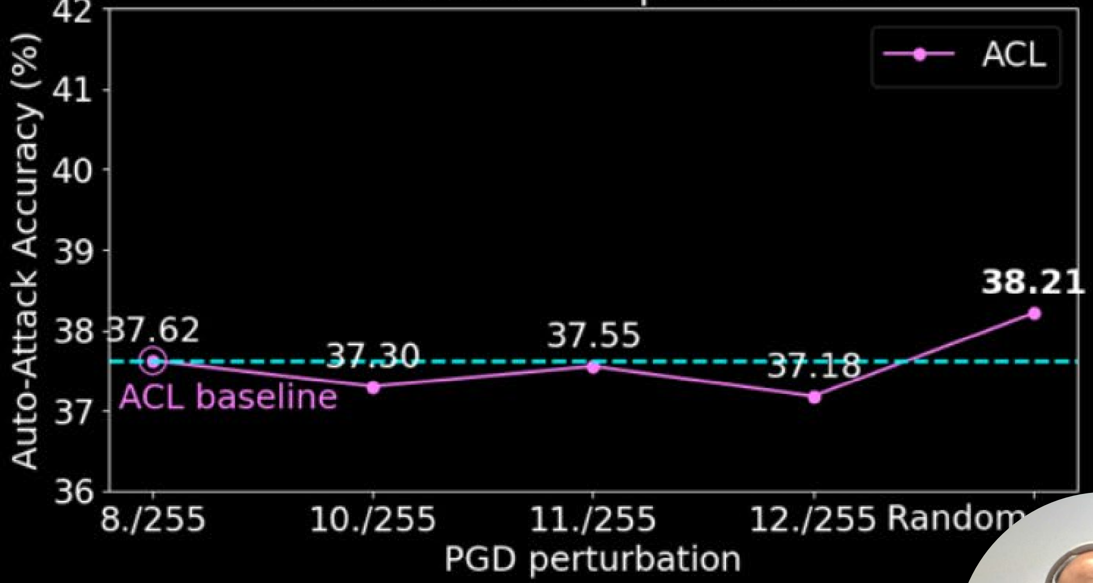
# Towards goal

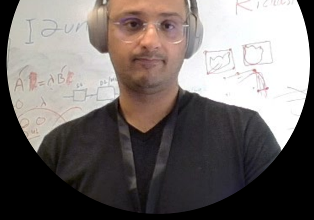
*ACL Method - Robust Pre-Training by Adversarial Contrastive Learning, NeurIPS 2020 (on CIFAR10)*

**Effect of Varying Attack Strategies during Different Training Stages**



Ablation on PGD perturbation

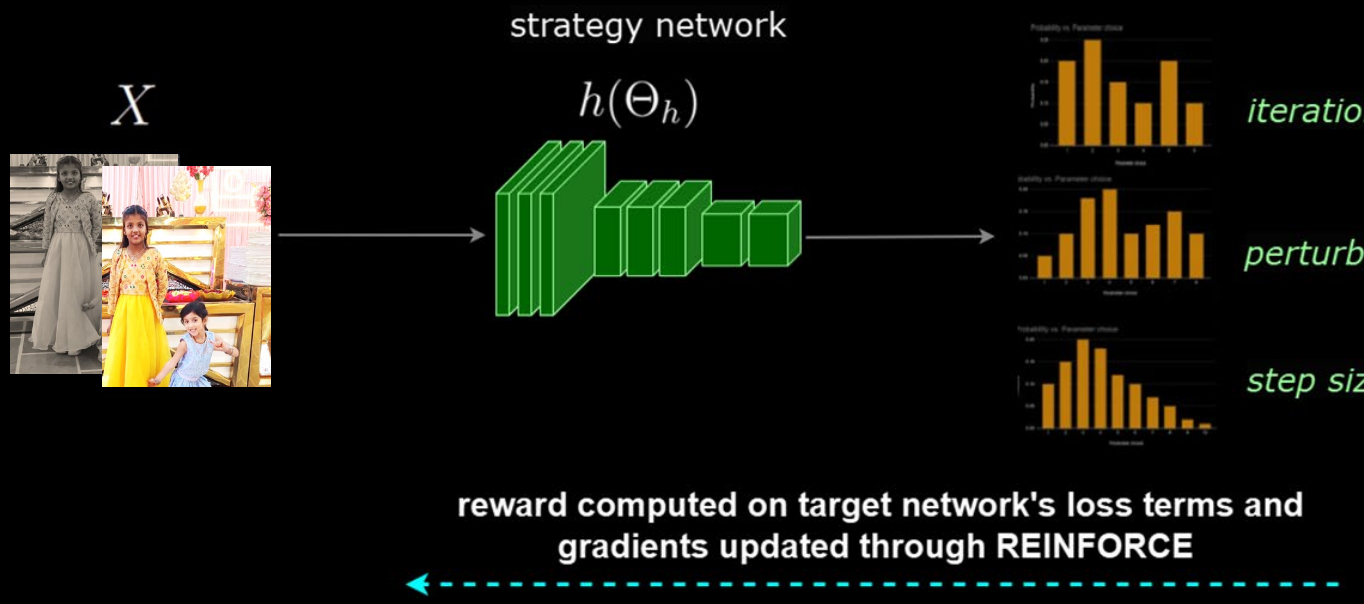




Develop adaptive, self-supervised adversarial attack strategy

# Learnable attacks in ASTrA

✓ Learnable strategy network autonomously finds optimal attacks.



*iteration (K) sampled from* ***distribution-1***

*perturbation ($\epsilon$) sampled from* ***distribution-2***

*step size (alpha) sampled from* ***distribution-3***

**reward computed on target network's loss terms and gradients updated through REINFORCE**

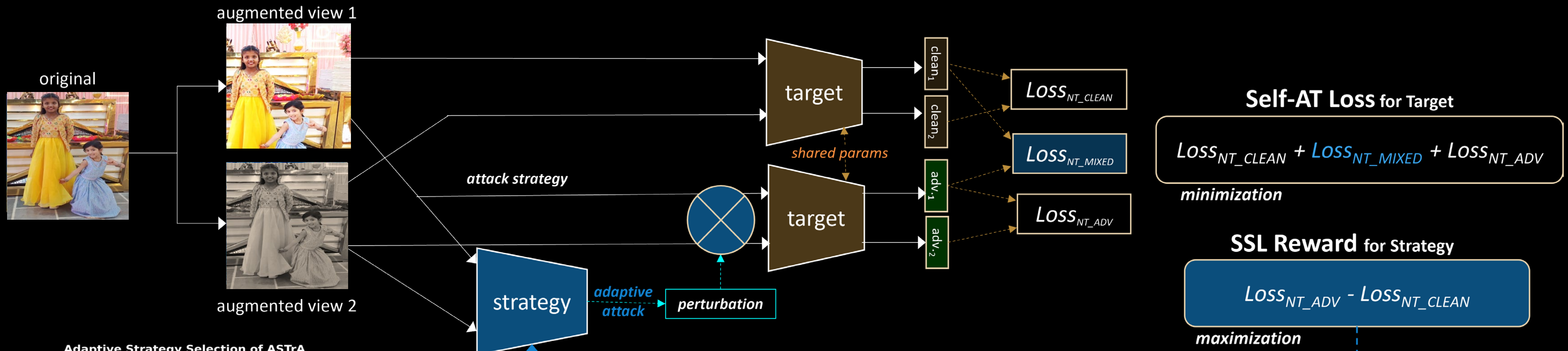
ICLR International Conference On Learning Representations
SINGAPORE 2024

UCF CENTER FOR RESEARCH IN COMPUTER VISION

LULEÅ TEKNISKA UNIVERSITET

# ASTrA framework

✓ Exploration-Exploitation using SSL contrastive reward and REINFORCE optimization



**Self-AT Loss** for Target

$Loss_{NT\_CLEAN} + Loss_{NT\_MIXED} + Loss_{NT\_ADV}$

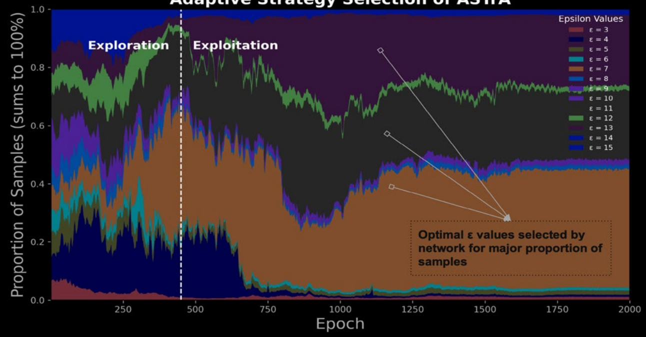
*minimization*

**SSL Reward** for Strategy

$Loss_{NT\_ADV} - Loss_{NT\_CLEAN}$

*maximization*

Adaptive Strategy Selection of ASTrA

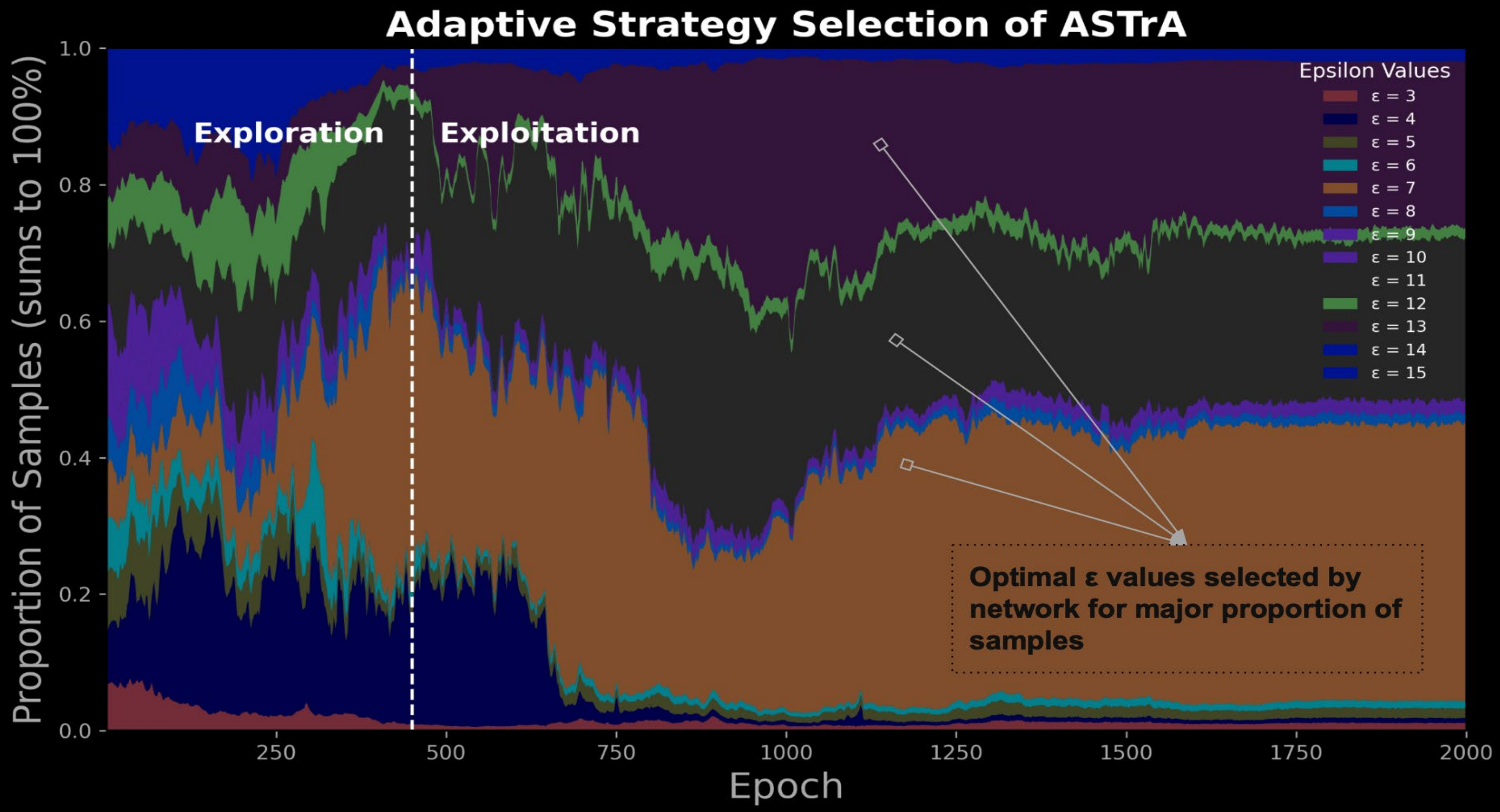


*ASTrA: Adversarial Self-supervised Training with Adaptive-Attack, ICLR 2025*

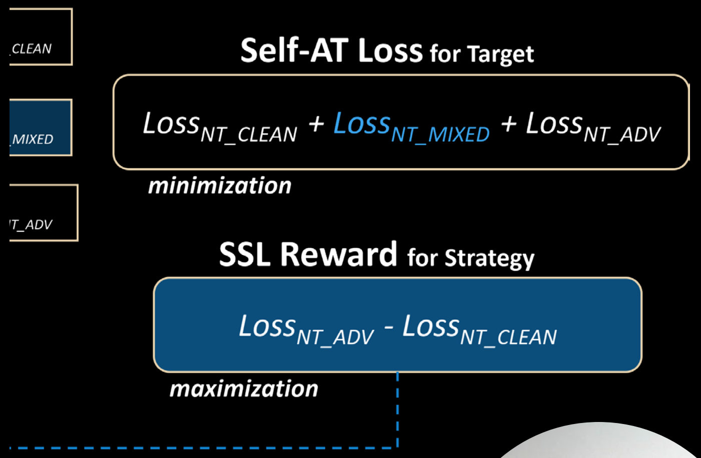
*https://prakashchhipa.github.io/projects/ASTrA*

# ASTrA framework

- ✓ Exploration-Exploitation using SSL contrastive reward and REINFORCE optimization



Optimal ε values selected by network for major proportion of samples



**Self-AT Loss** for Target

$Loss_{NT\_CLEAN} + Loss_{NT\_MIXED} + Loss_{NT\_ADV}$

*minimization*

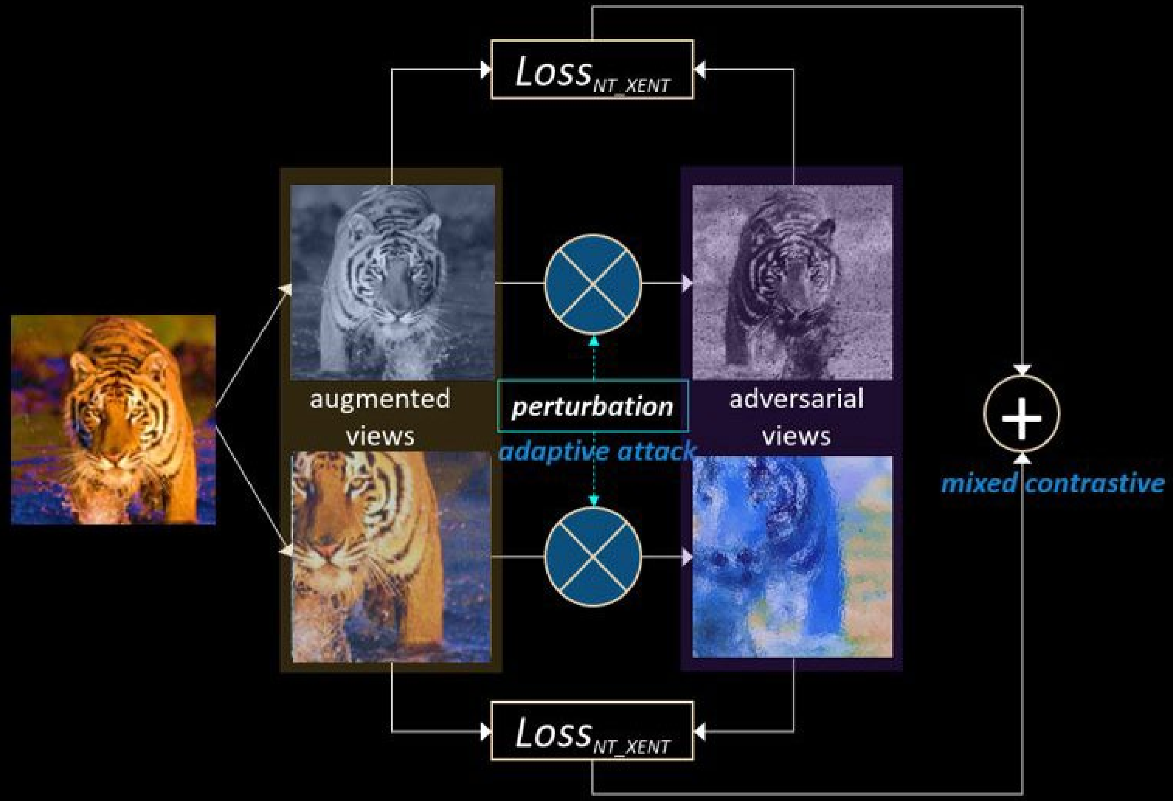
**SSL Reward** for Strategy

$Loss_{NT\_ADV} - Loss_{NT\_CLEAN}$

*maximization*

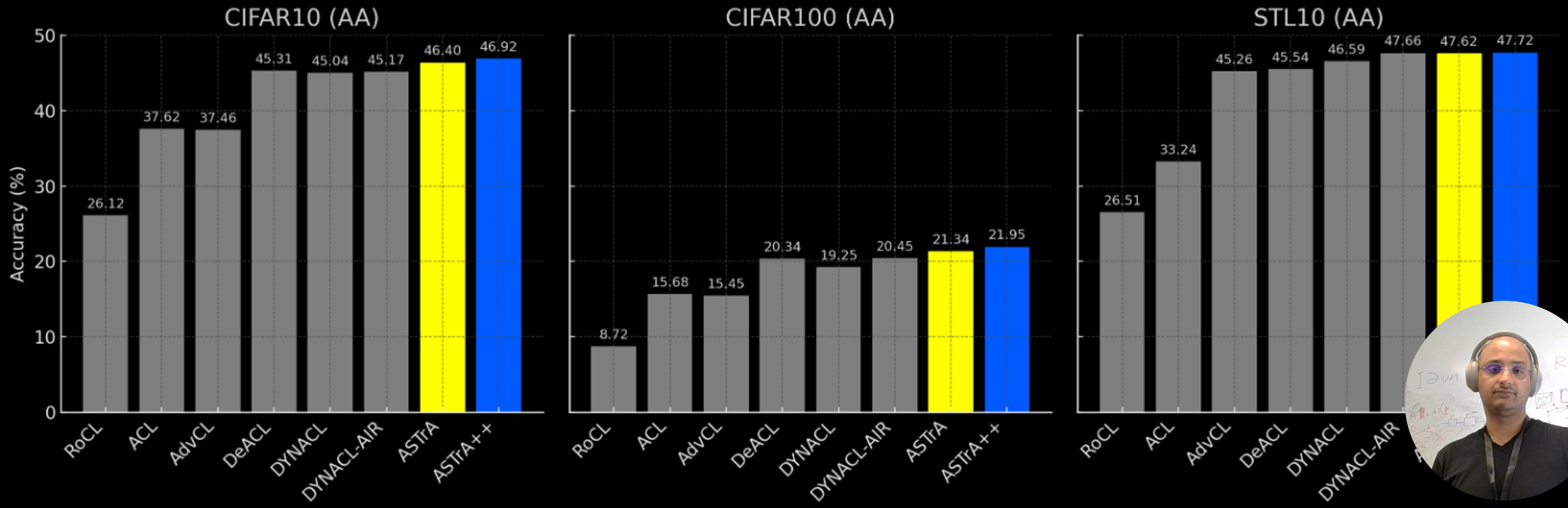
# Mixed contrastive loss

✓ Align representations using of clean view to corresponding (adaptively attacked) perturbed view.



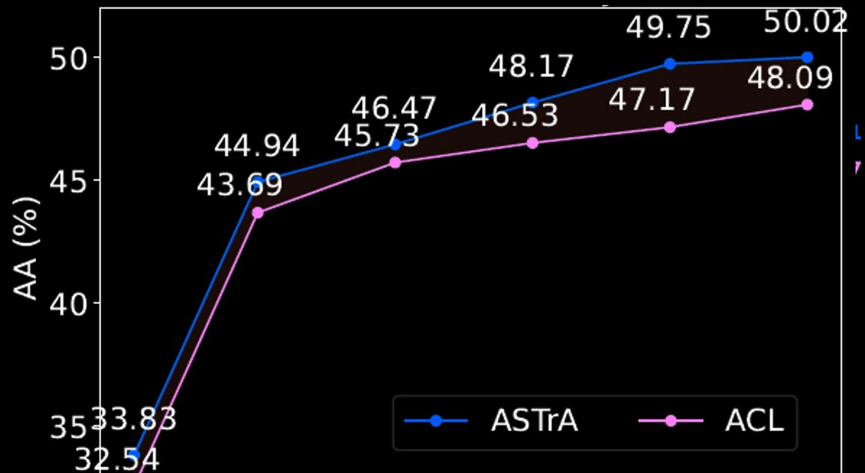
# Results *public benchmarks*

## Standard Linear Finetuning Performance – ASTrA vs. other Self-AT method

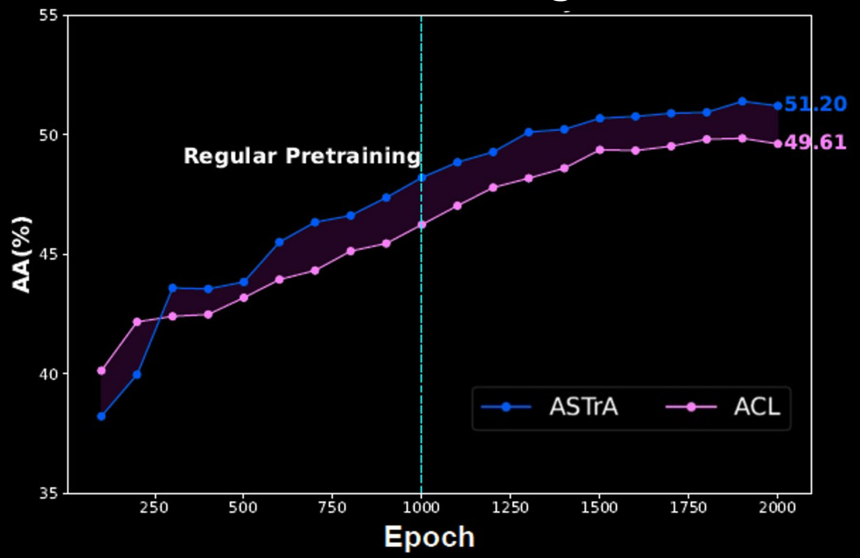


# Results *label efficiency, improved robustness and modularity*

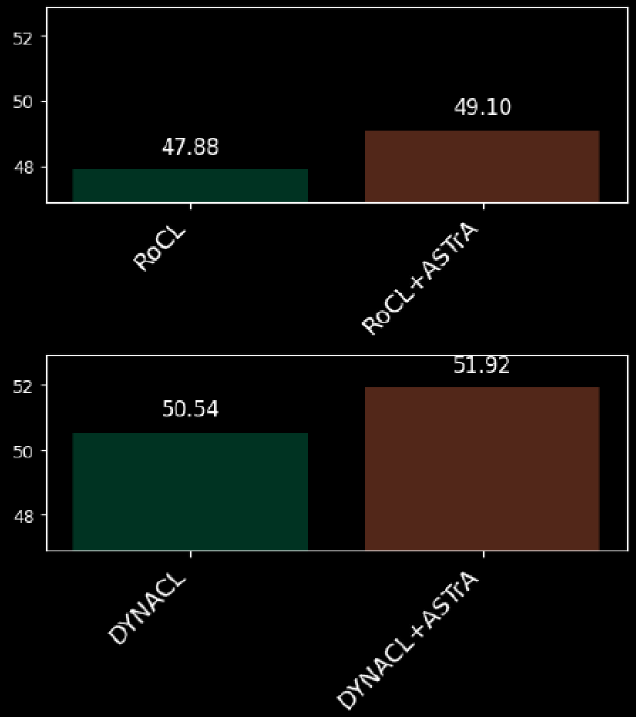
semi-supervised setting



robust overfitting



plug-and-play with self-ATs



# Contributions

✓ Self-supervised adaptive attack strategy – ASTrA framework

✓ *enables through exploration-exploitation on learning attack parameters*

✓ Mixed contrastive loss

✓ *improving distribution alignment*

# Achievements

✓ Improved robustness across benchmarks and evaluation protocols

✓ *STL, CIFARs, SLF, ALF, AFF*

✓Scalable and avoid robust overfitting

✓ *ImageNet100, longer pretraining*

✓Plug-and-play and modular

✓*Sefl-AT methods-DYNACL, RoCL*

# Thank you