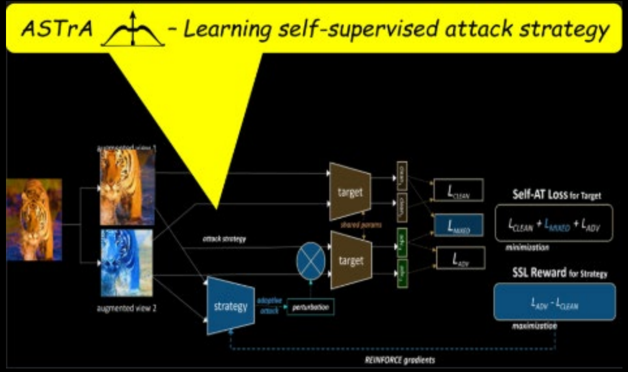
ASTrA: Adversarial Self-supervised Training with Adaptive-Attack

ICLR 2025 Singapore

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



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

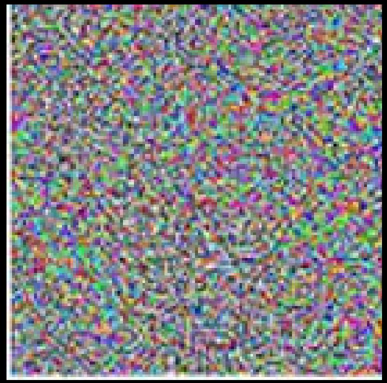
[1]Machine Learning Group, Lulea˚ Tekniska Universitet, Sweden
[2]Indian Institute of Technology, Gandinagar
[3]Indian Institute of Technology, Madras
[4]Center For Research in Computer Vision, University of Central Florida, USA

*equal contribution
presenter

Self-supervised adversarial attacks

- Networks vulnerability to adversarial examples.

cat

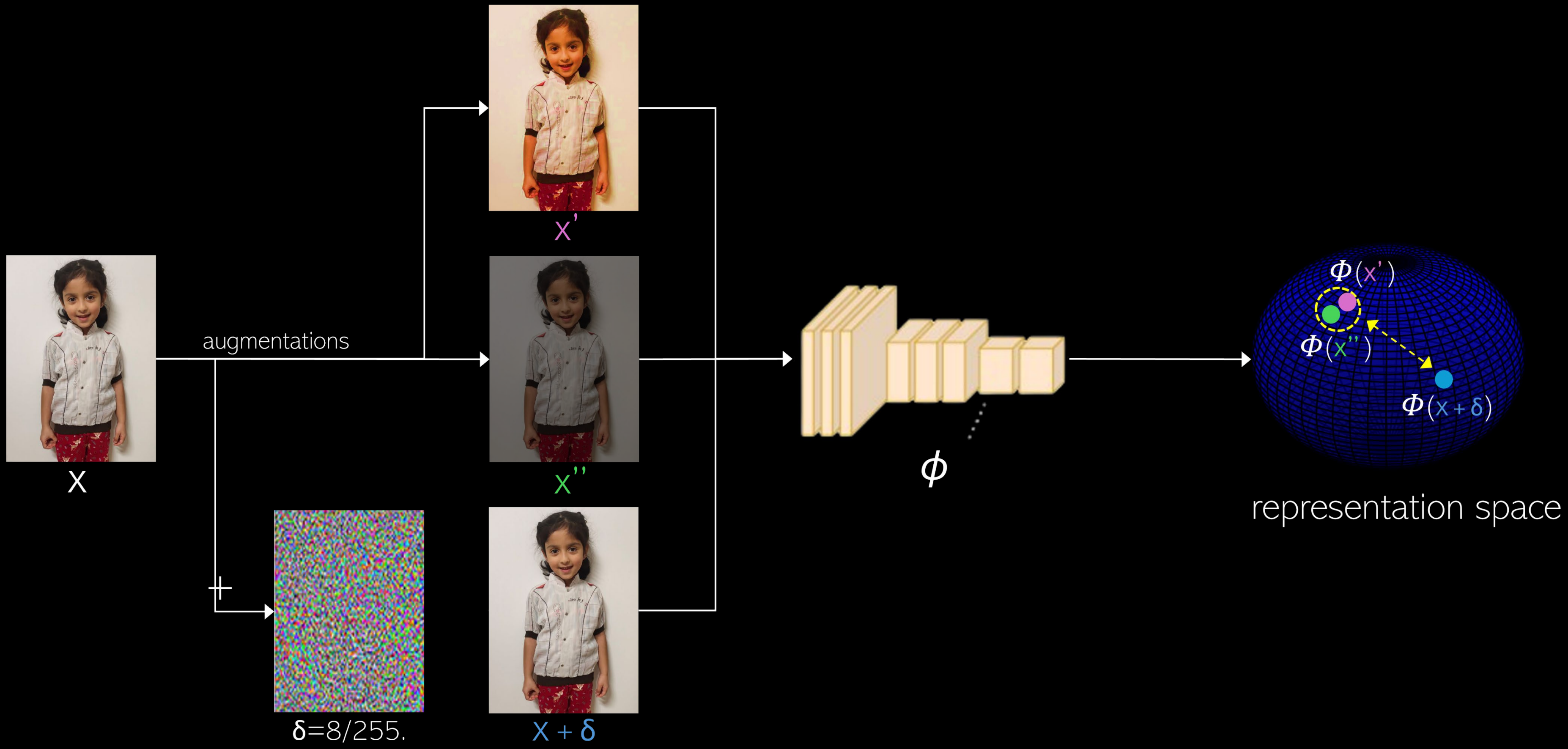


δ=8/255.

airliner

- Adversarial attacks are small, carefully crafted perturbations added to input data that cause a model to make incorrect predictions—while the altered input still looks unchanged to humans.

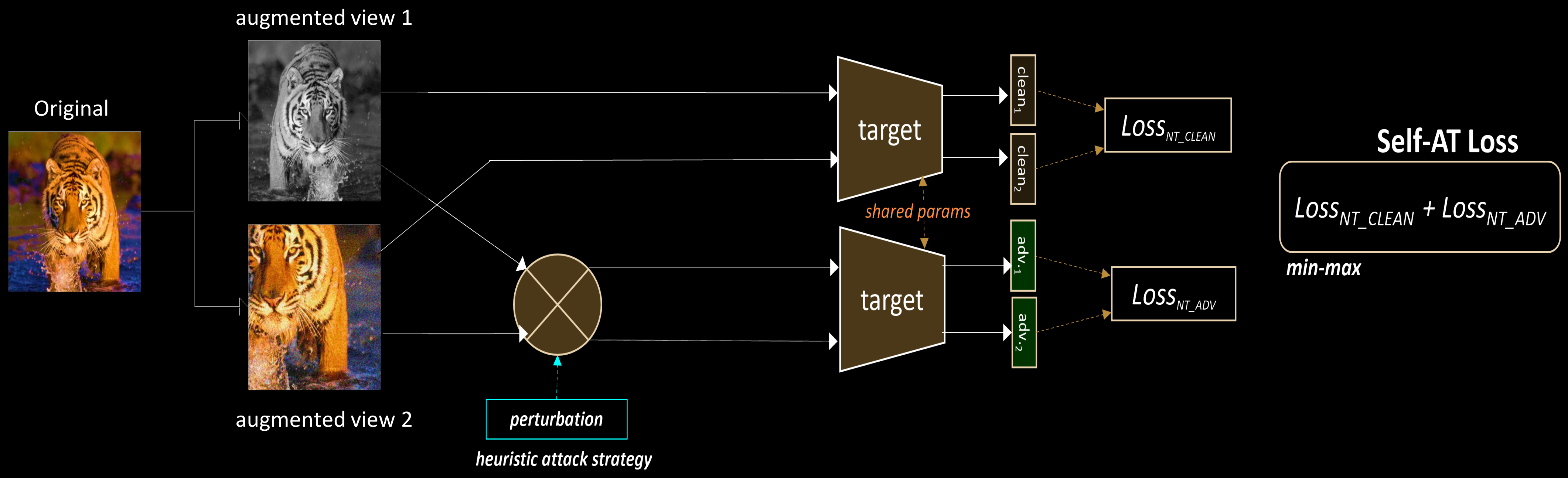
Representation learning perspective



Ian Goodfellow et al. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.
Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018.

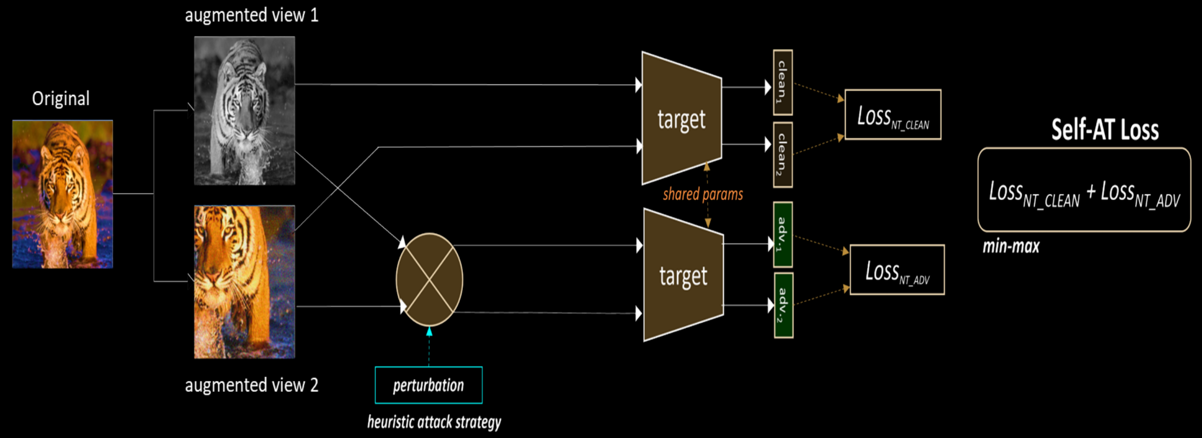
Existing Self-supervised adversarial training methods

- ACL (2020), RoCL (2020), AdvCL [2021], DeACL [2022], DYNACL [218] – adversarial contrastive learning



ACL: Ziyu Jiang et al. Robust pre-training by adversarial contrastive learning. Advances in neural information processing systems, 33:16199–16210, 2020.
RoCL: Minseon Kim et al. Adversarial self-supervised contrastive learning. Advances in neural information processing systems, 33:2983–2994, 2020.
AdvCL: Lijie Fan et al. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? Advances in neural information processing systems, 34:21480–21492, 2021.
DeACL: Chaoning Zhang et al. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In European Conference on Computer Vision, pages 725–742. Springer, 2022.
DYNACL: Rundong Luo et al. Rethinking the effect of data augmentation in adversarial contrastive learning. In International Conference on Learning Representations, 2023.

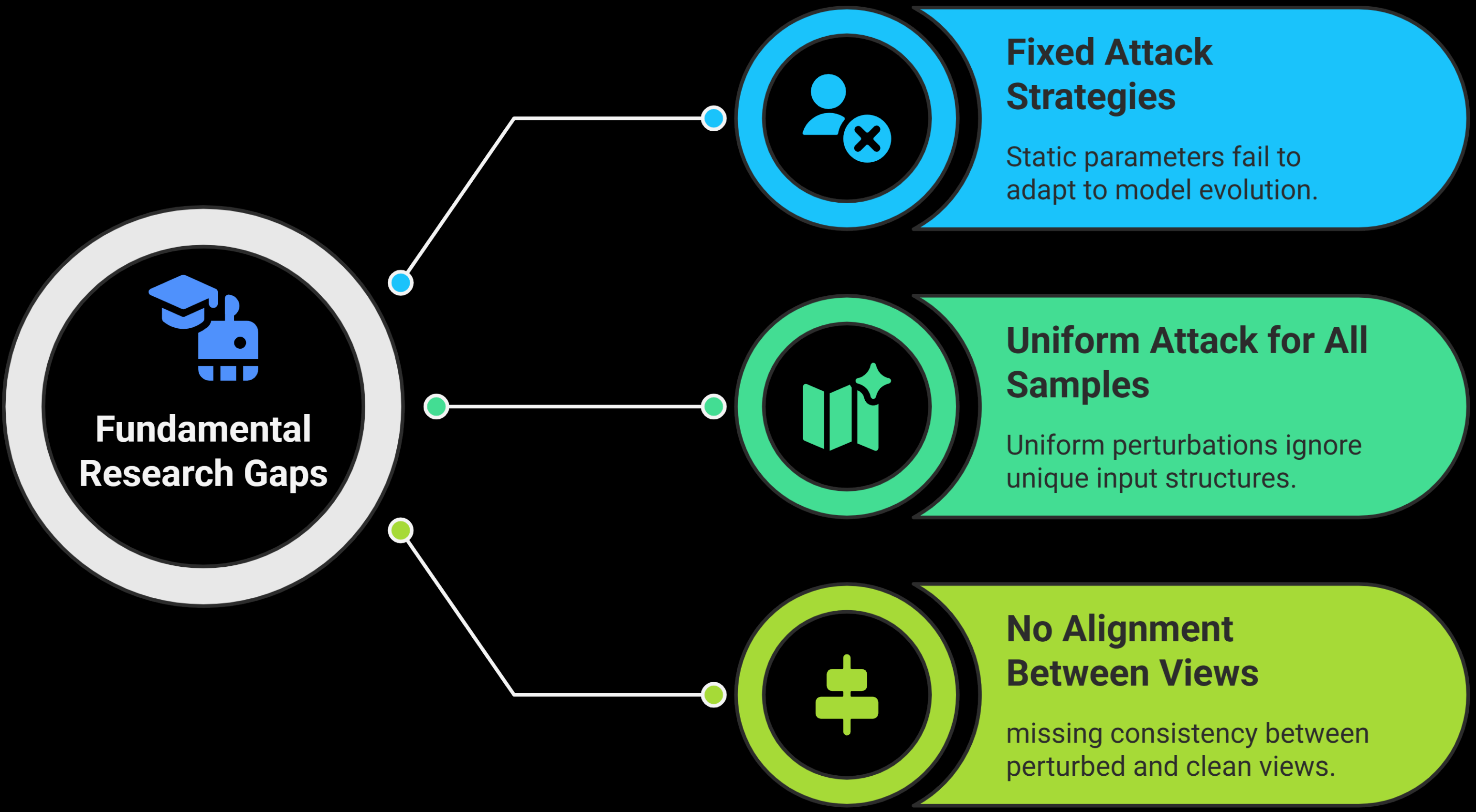
What they do?



- ACL [2020], RoCL [2020] – *baseline self-AT methods with added perturbed views in contrastive learning.*
- AdvCL [2021] – *added teacher-student knowledge distillation.*
- DeACL [2022] – *standard SSL to learn representations from unlabeled data and applying adversarial training (AT) using pseudo-labels from first stage encoder.*
- DYNACL [2023]- *dynamic rule-based augmentations (strong → weak) applied during self-AT training.*

ACL: Ziyu Jiang et al. Robust pre-training by adversarial contrastive learning. Advances in neural information processing systems, 33:16199–16210, 2020.
RoCL: Minseon Kim et al. Adversarial self-supervised contrastive learning. Advances in neural information processing systems, 33:2983–2994, 2020.
AdvCL: Lijie Fan et al. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? Advances in neural information processing systems, 34:21480–21492, 2021.
DeACL: Chaoning Zhang et al. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In European Conference on Computer Vision, pages 725–742. Springer, 2022.
DYNACL: Rundong Luo et al. Rethinking the effect of data augmentation in adversarial contrastive learning. In International Conference on Learning Representations, 2023.

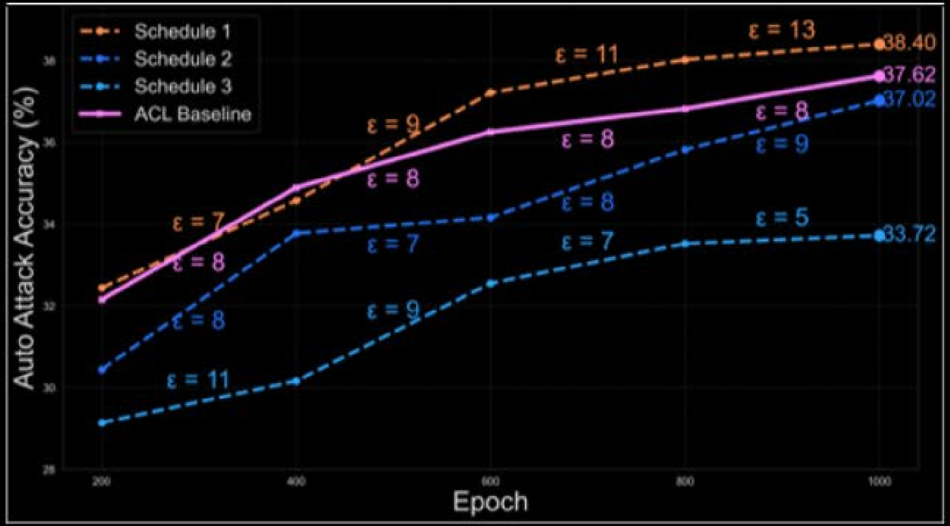
What don't they do?

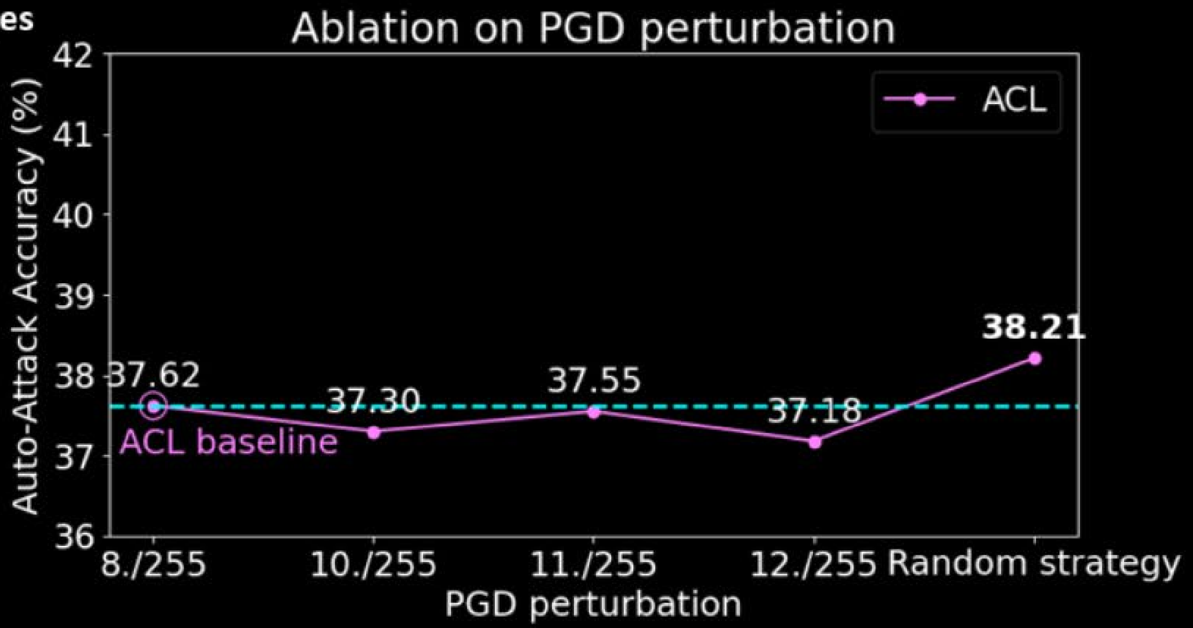


One-size-fits-all approach to adversarial attack strategy is fundamentally suboptimal and inefficient, as it disregards the evolving robustness of the model, ultimately limiting the effectiveness and scalability of adversarial training.

Towards goal – exploring rule based yet dynamic attack strategy

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

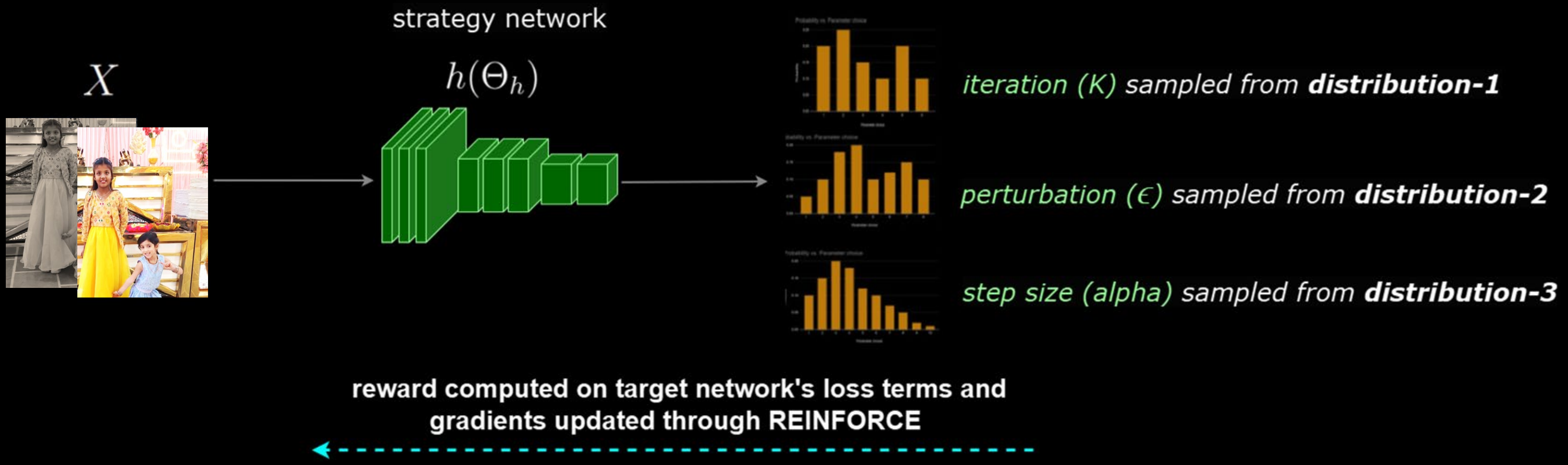




Main goal- develop adaptive, self-supervised adversarial attack strategy

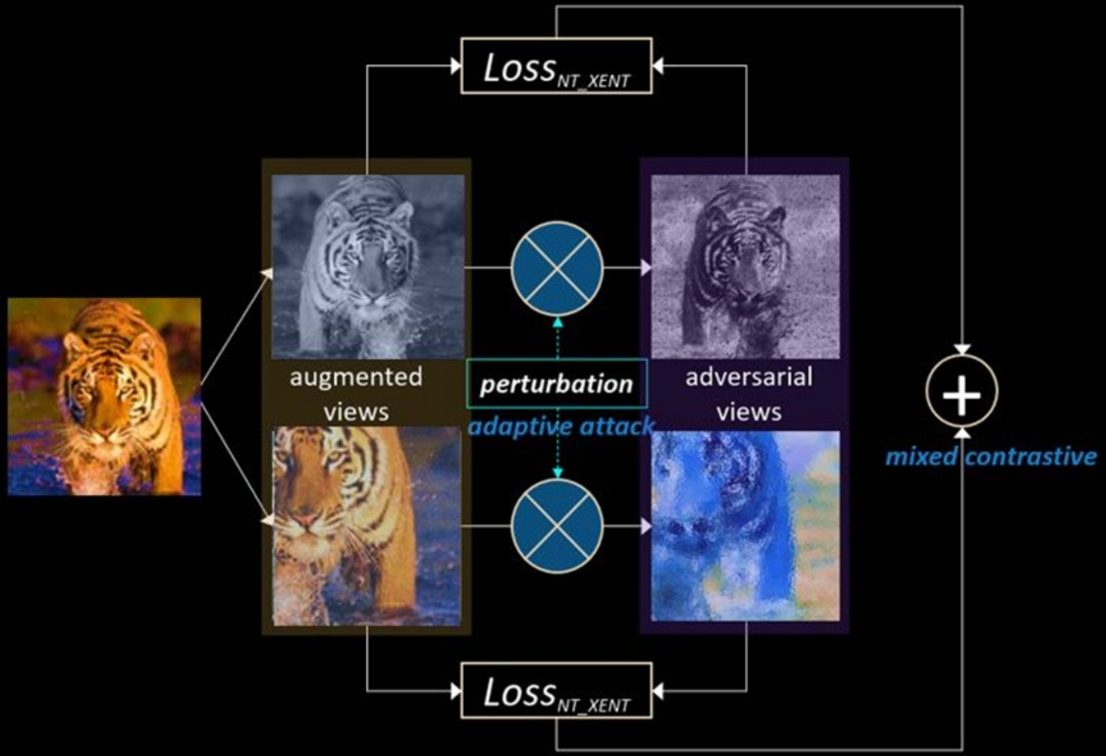
Learnable attacks in ASTrA – contribution 1

✓ Learnable strategy network autonomously finds optimal attacks.



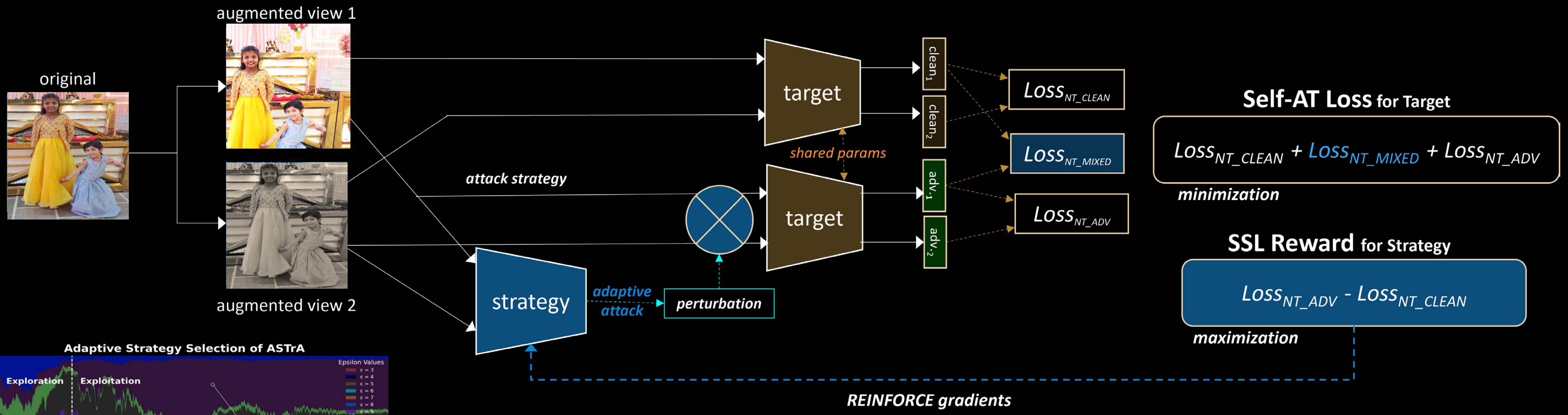
Mixed contrastive objective in ASTrA – contribution 2

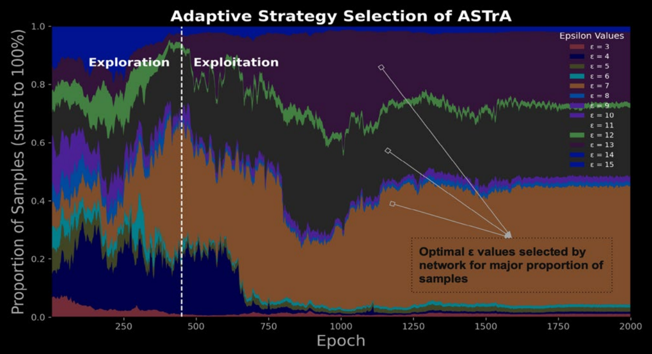
✓ Align representations using of clean view to corresponding (adaptively attacked) perturbed view.



ASTrA framework

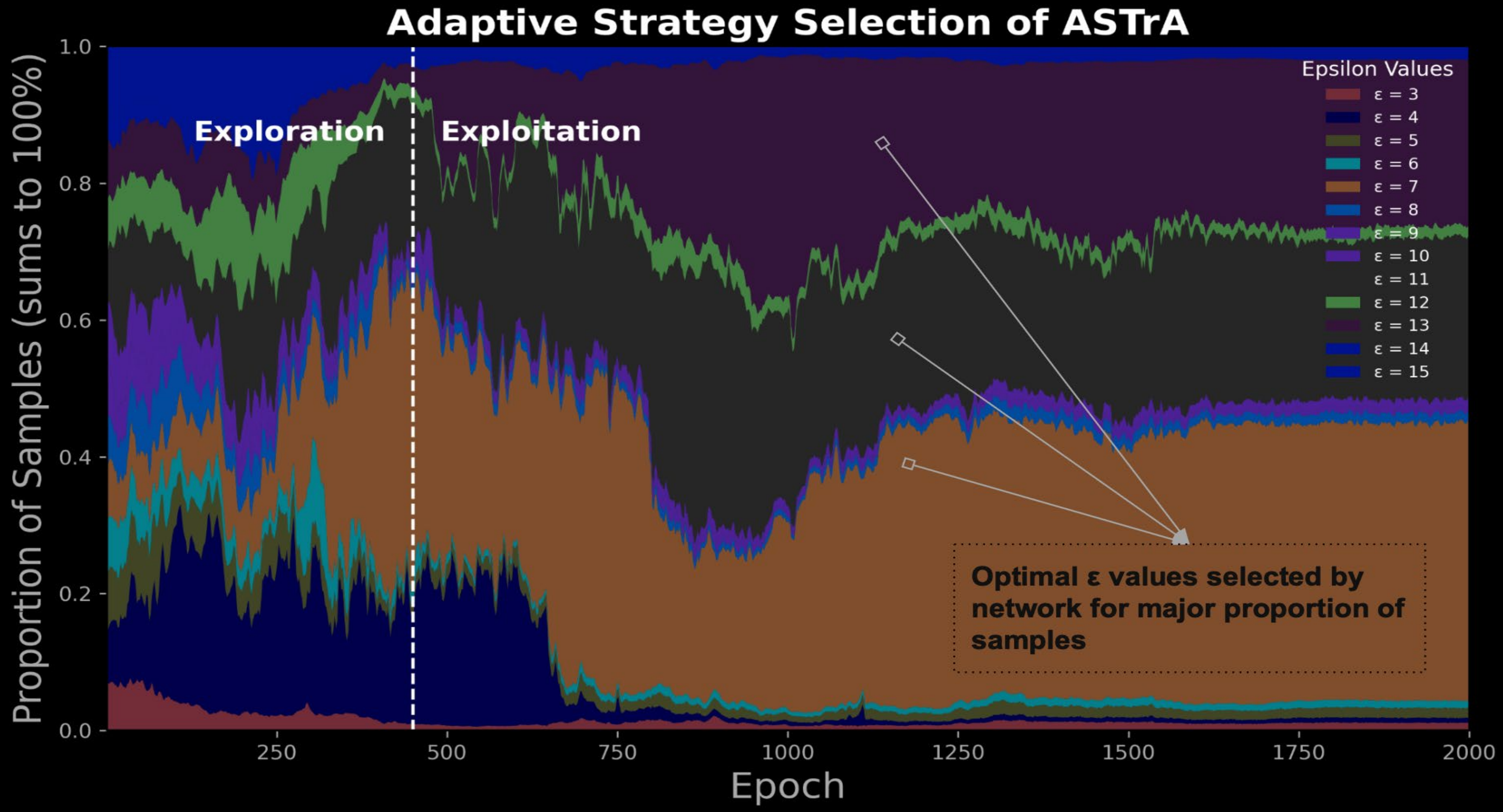
✓ Exploration-Exploitation using SSL contrastive reward and REINFORCE optimization





ASTrA framework

- ✓ Exploration-Exploitation lead to achieve optimal combination of attack paramters



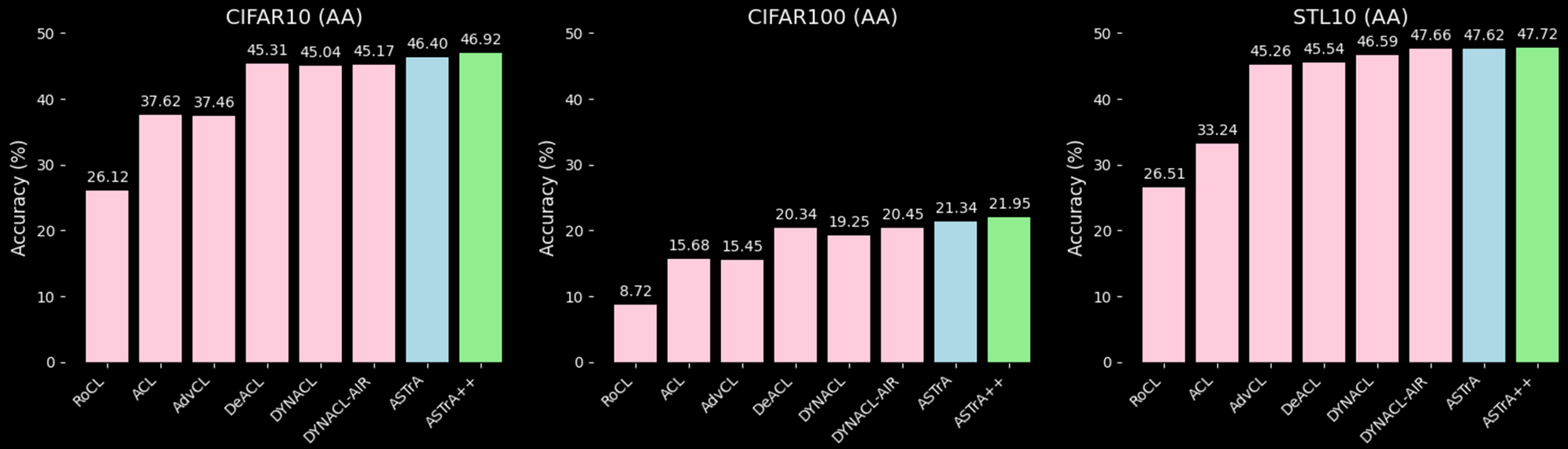
ICLR
International Conference On Learning Representations
SINGAPORE 2024

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

LULEÅ TEKNISKA UNIVERSITET

Results *public benchmarks*

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



AA: Auto Attack Accuracy, **RA**: Accuracy under PGD-20 as Robust Accuracy, and **SA**: Standard Accuracy.

ASTrA as plug-N-play *with other self-AT methods*

Standard linear-finetuning (SLF)

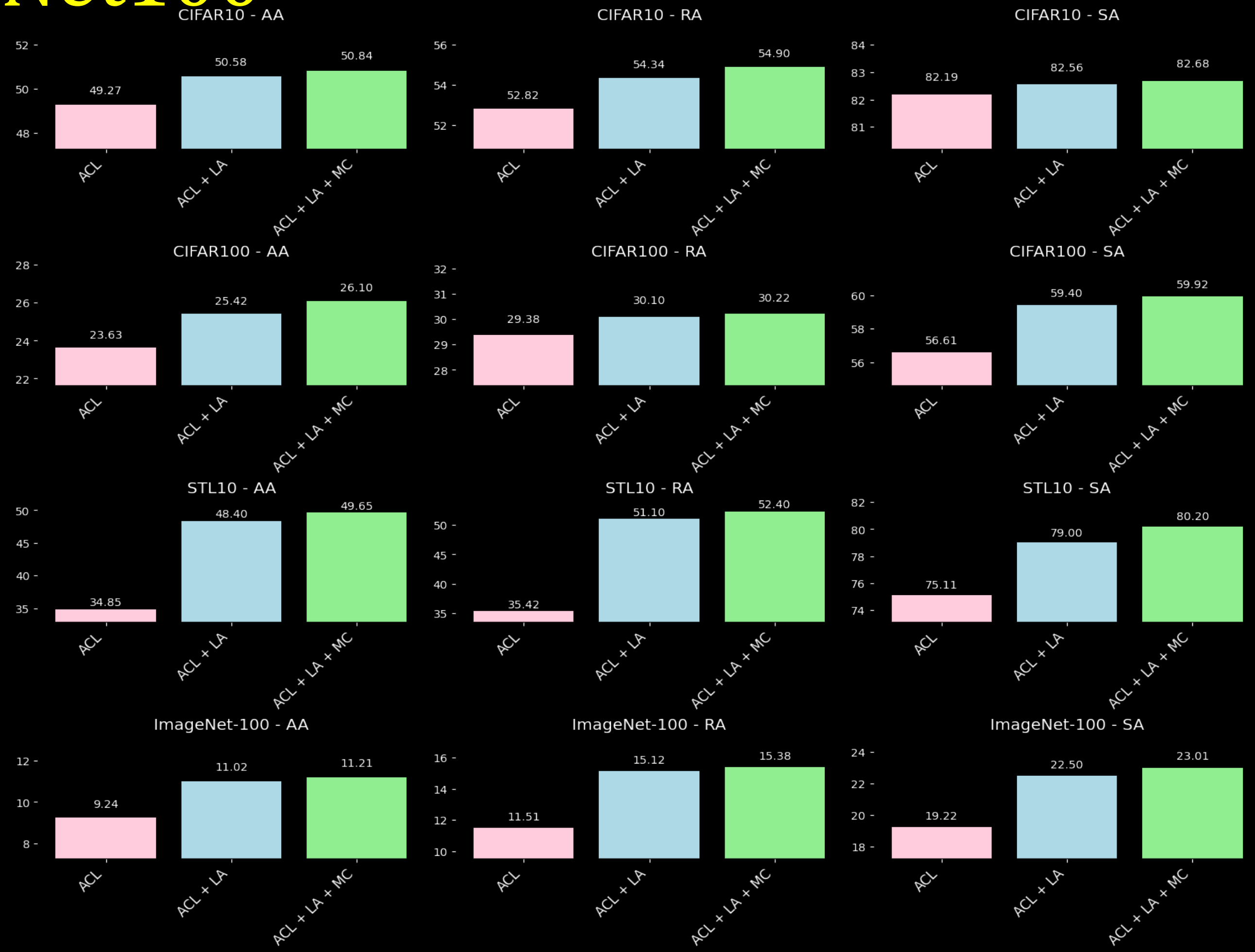
AA: Auto Attack Accuracy, **RA**: Accuracy under PGD-20 as Robust Accuracy, and **SA**: Standard Accuracy.

ICLR
International Conference On Learning Representations
SINGAPORE 2024

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

LULEÅ TEKNISKA UNIVERSITET

ASTrA - on ImageNet100



AA: Auto Attack Accuracy,
RA: Accuracy under PGD-20 as Robust Accuracy
SA: Standard Accuracy.

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