

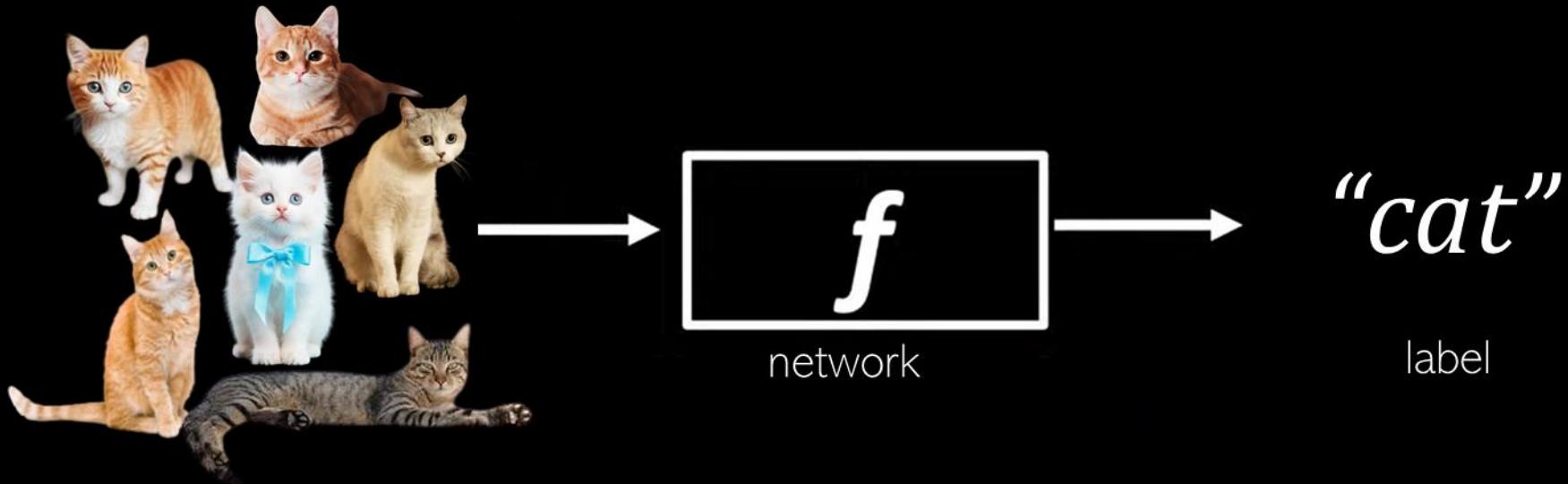
Towards Robust and Self-supervised Representation Learning

MBZUAI Research Talks

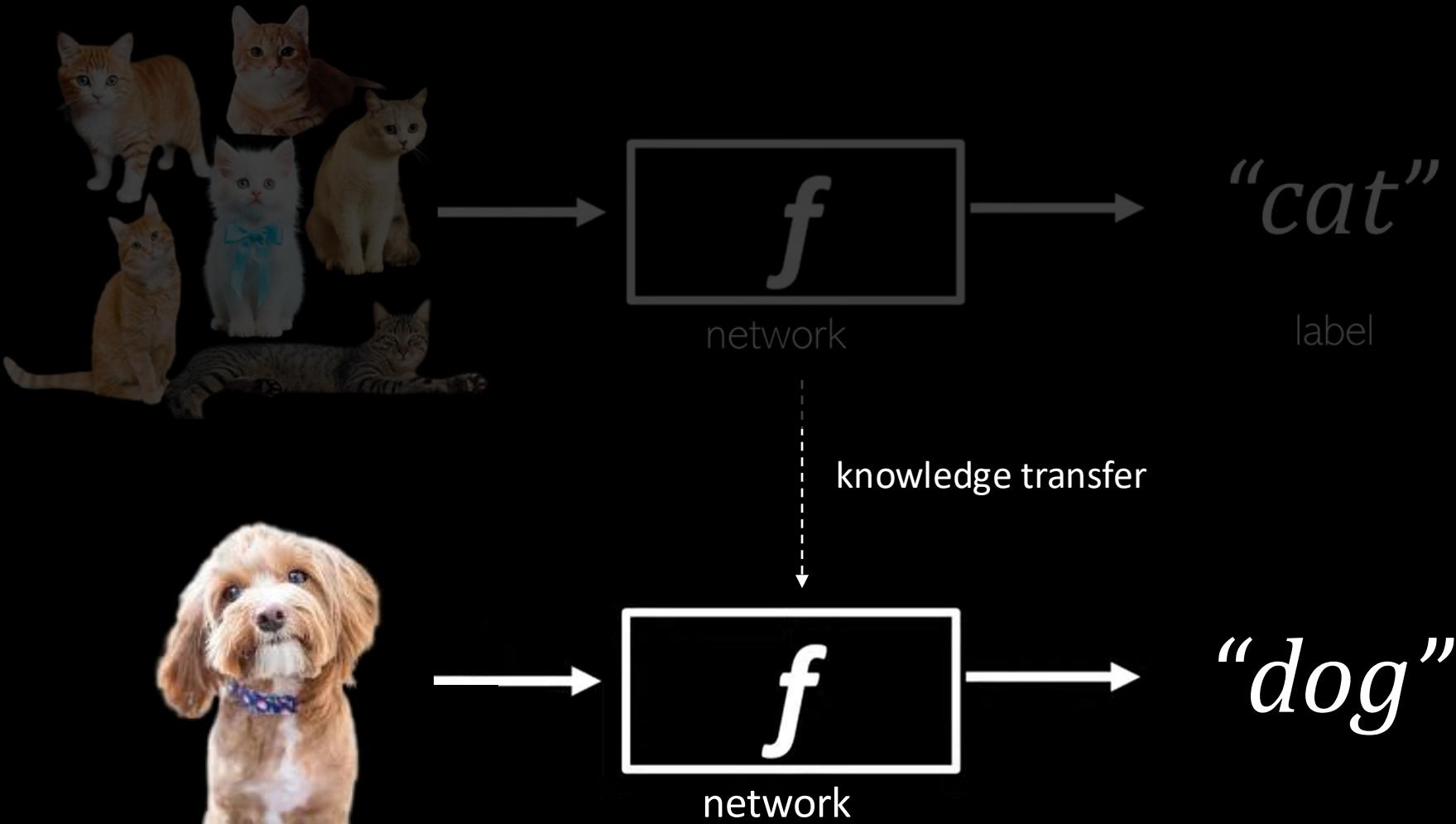
Prakash Chandra Chhipa and Rajkumar Saini
Machine Learning group



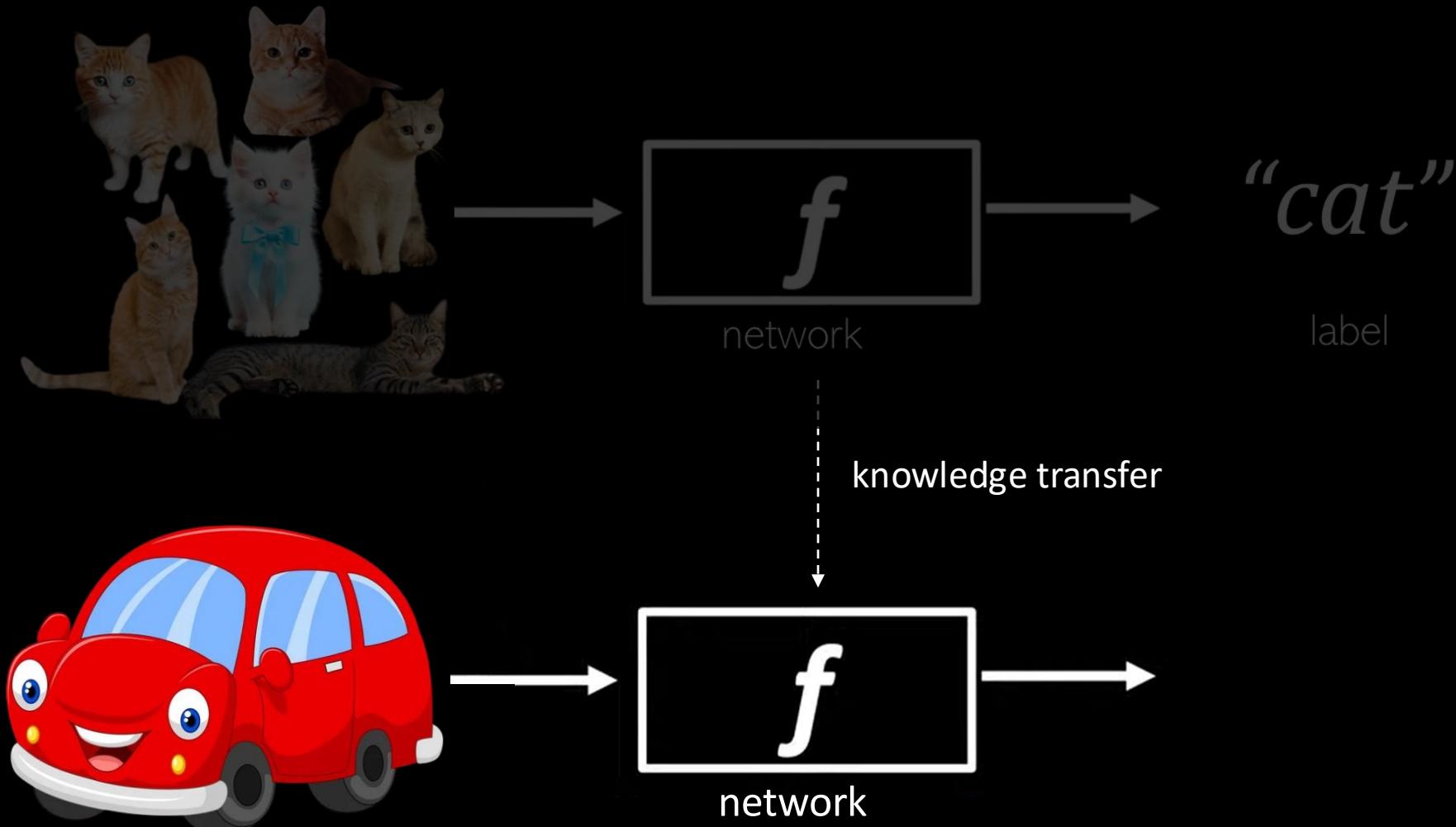
Supervised learning



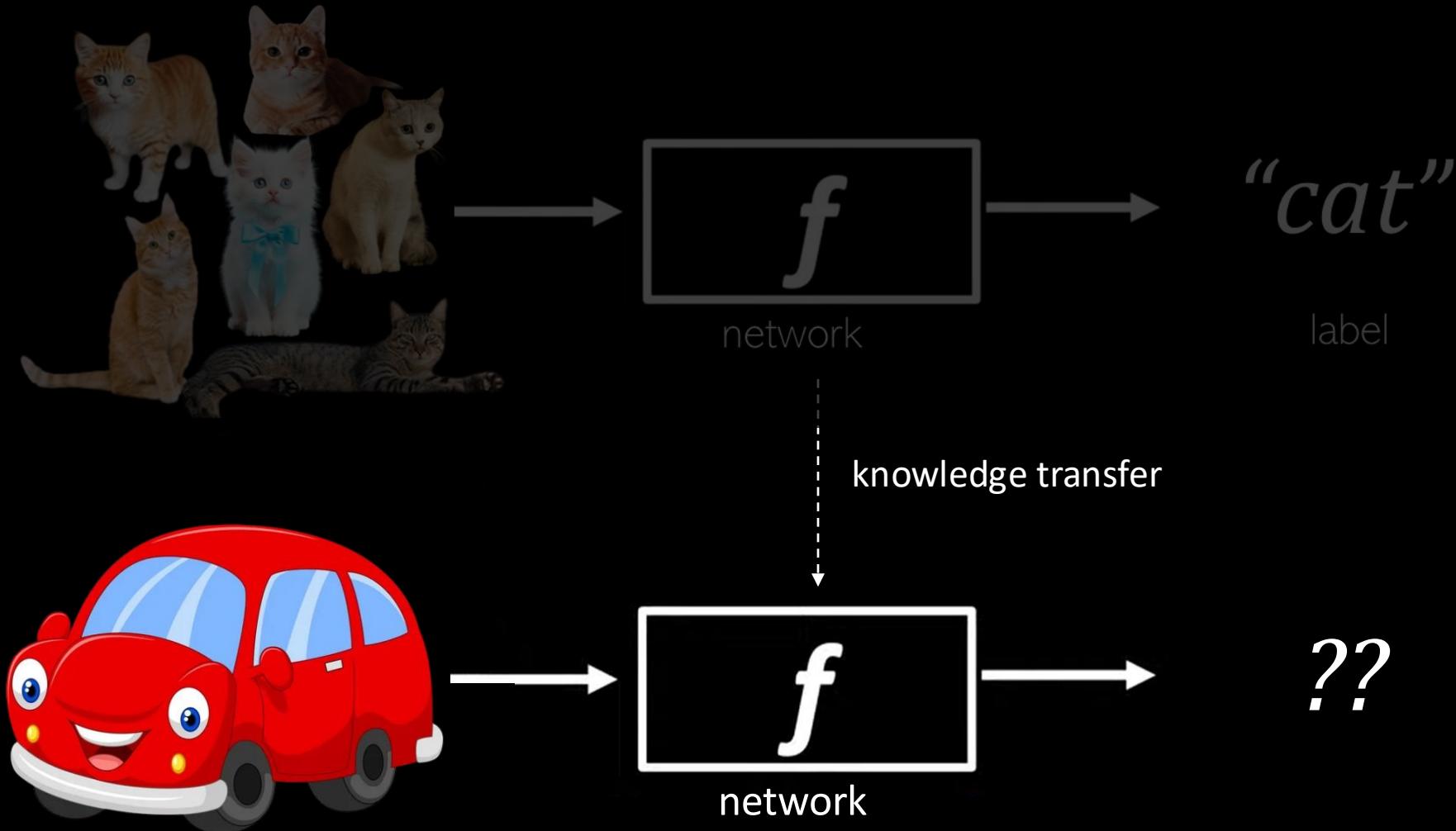
Supervised learning



Supervised learning



Supervised learning

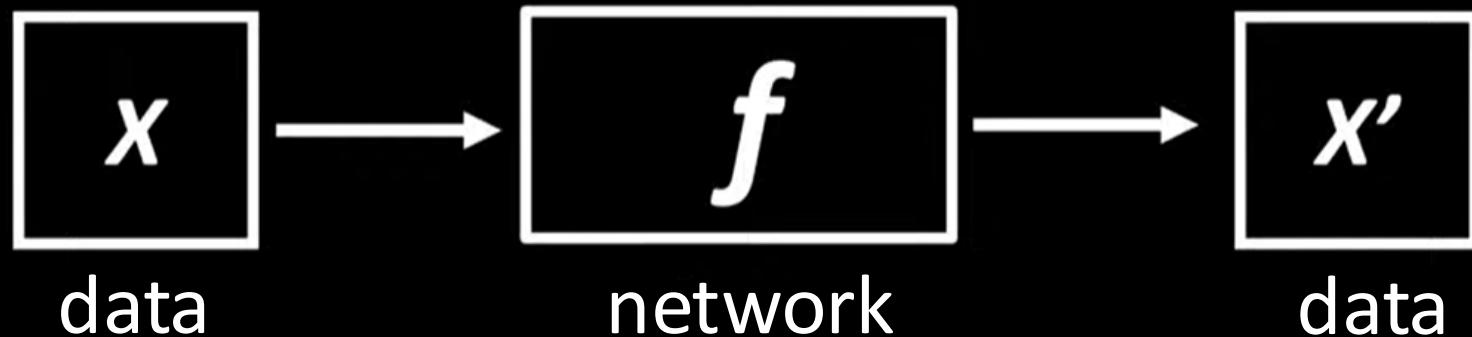


Poor generalizability

Not scalable

Self-supervised representation learning

Decade ago..



Learns representations by emphasizing inherent structures within data—without needing labels

Pseudo task (until 2020)

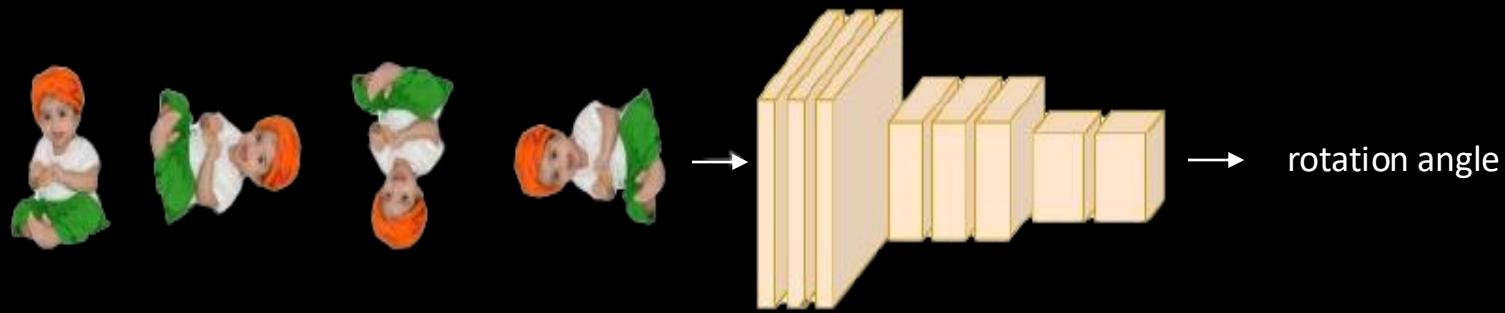
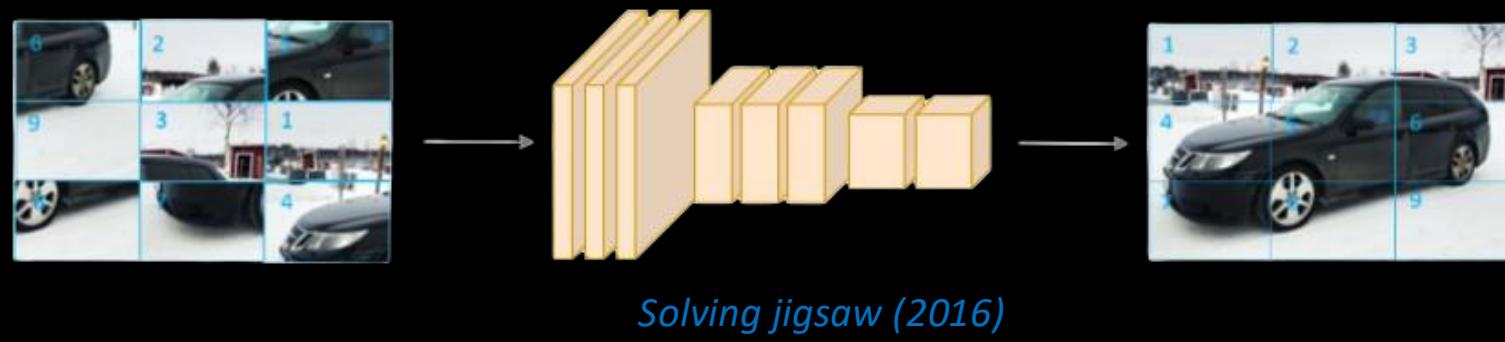
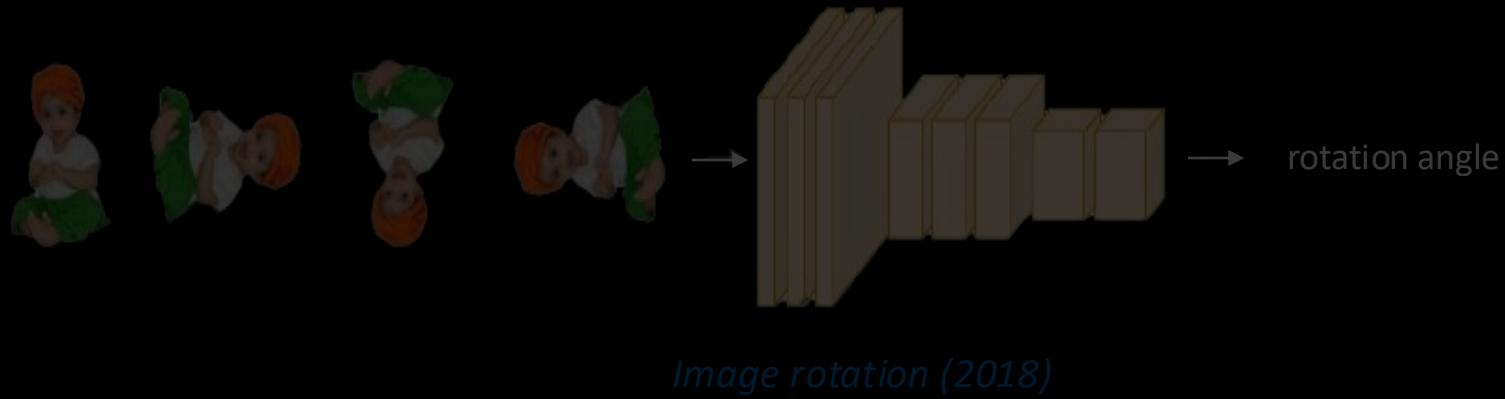
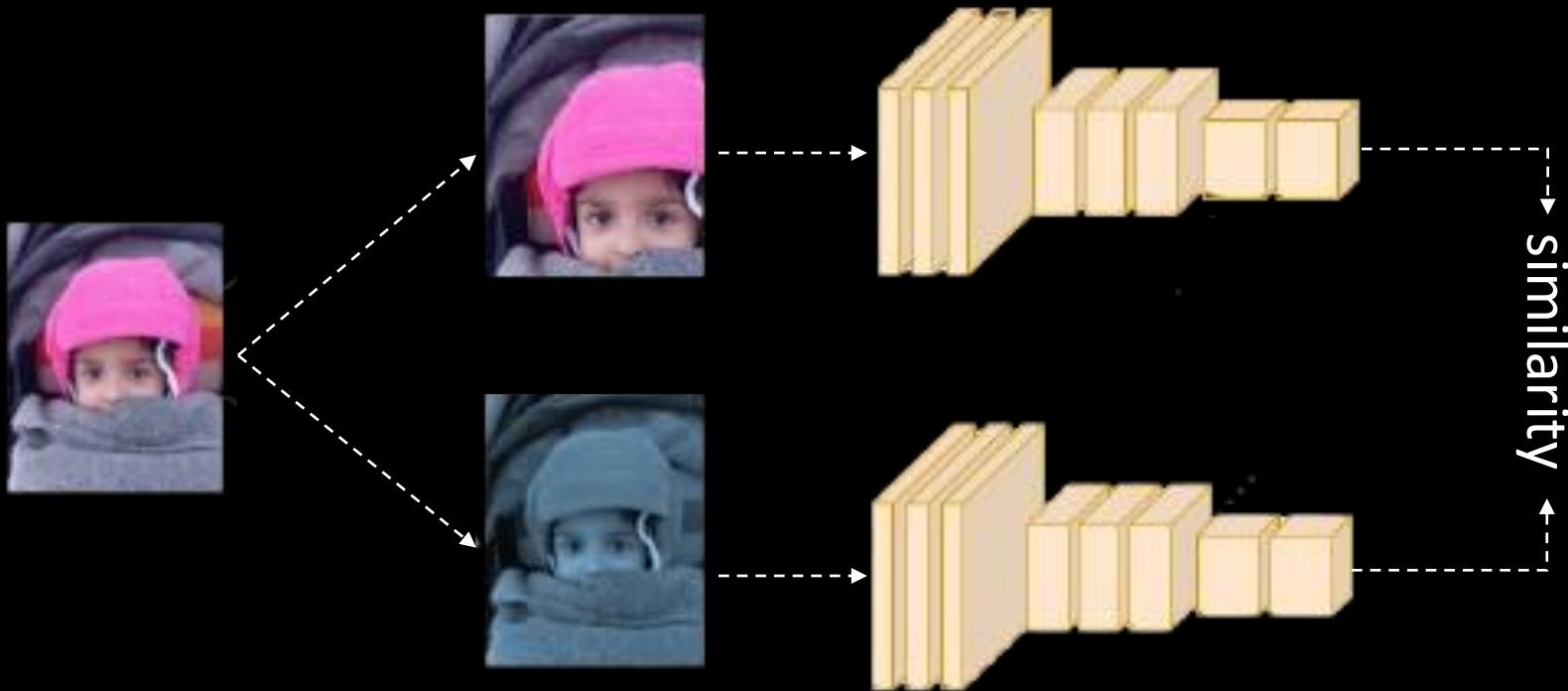


Image rotation (2018)

Pseudo task (until 2020)



Joint embedding (2020 onwards)



SimCLR (2020), MoCo (2020), SimSiam (2020), BYOL (2021), DINO (2021), Barlow Twin (2020), VICReg(2021)

Merits



near-supervised performance

on many large scale benchmarks - e.g. ImageNet



Cost Reduction

Lowers costs associated with data annotation on downstream tasks - detection, classification, etc.



Broad Adoption Scope

adopted across modalities - vision, language, audio, video and cross modalities.



What are the challenges?

Real-world scenarios

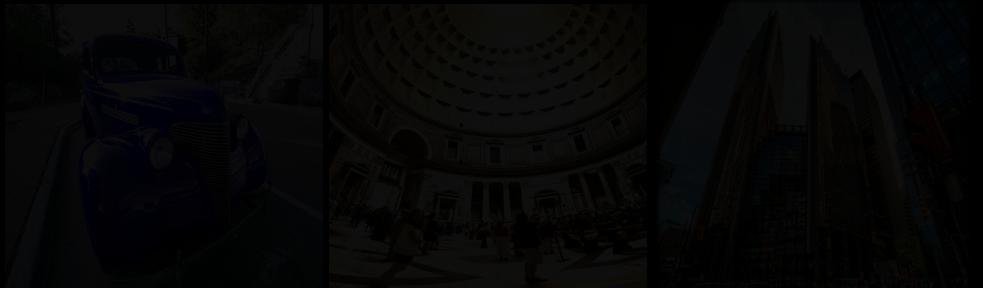
centered, clean



(distribution shifts)



sensors - spatial distortions



environments - pixel-level distortions (e.g., weather)

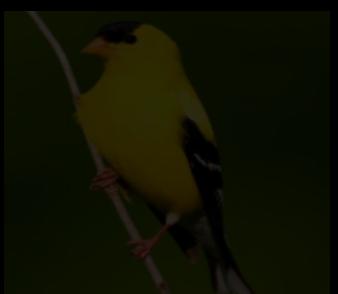


human - adversarial attacks



SENSORS - spatial distortions

centered, clean



(distribution shifts)



environments - pixel-level distortions (e.g., weather)



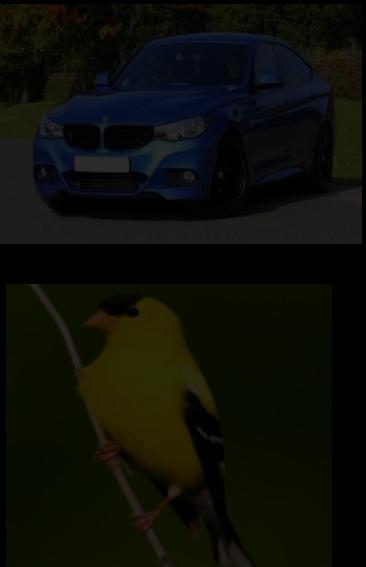
human - adversarial attacks



13

airliner

centered, clean



(distribution shifts)



environments - pixel-level corruptions



human - adversarial attacks

A diagram illustrating the creation of an adversarial example. It shows a dark gray square labeled "cat" followed by a plus sign (+), then a smaller square labeled " $\delta=8/255.$ ", followed by an equals sign (=), and finally a darker gray square labeled "airliner".

14

airliner

centered, clean

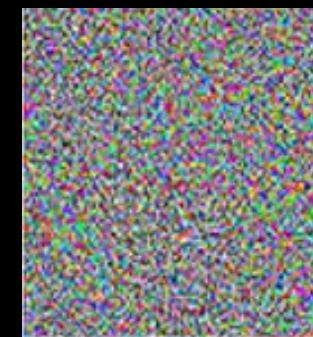


(distribution shifts)

human - adversarial attacks



+



=



sensors - spatial distortions



environments - pixel-level distortions (e.g., weather)

Real-world scenarios

sensors - spatial distortions



centered, clean



(distribution shifts)



environments - pixel-level distortions (e.g., weather)



human - adversarial attacks

$$\text{cat} + \delta=8/255. = \text{airliner}$$

Goal



Improve robustness in SSL

Goal



Improve **robustness** in SSL against real-world challenges



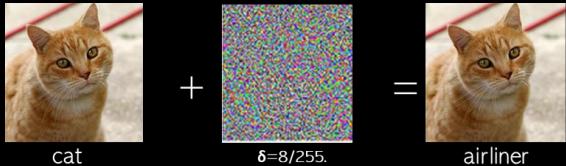
sensors - spatial distortions

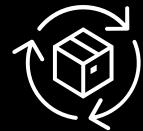


environments - pixel-level distortions (e.g., weather)

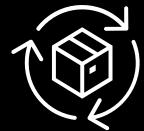


human - adversarial attacks



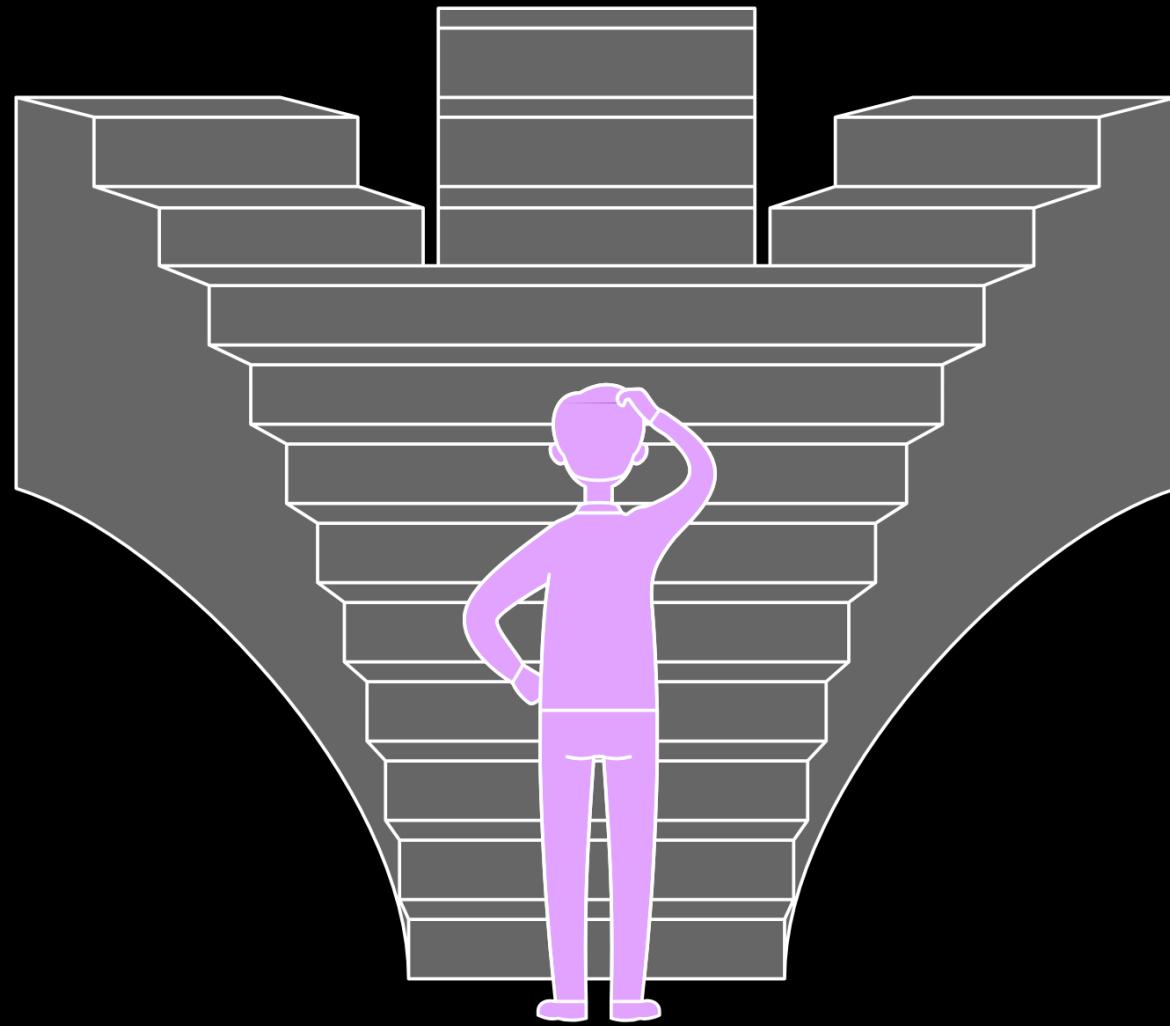
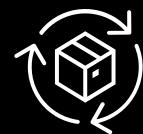


knowledge
distillation



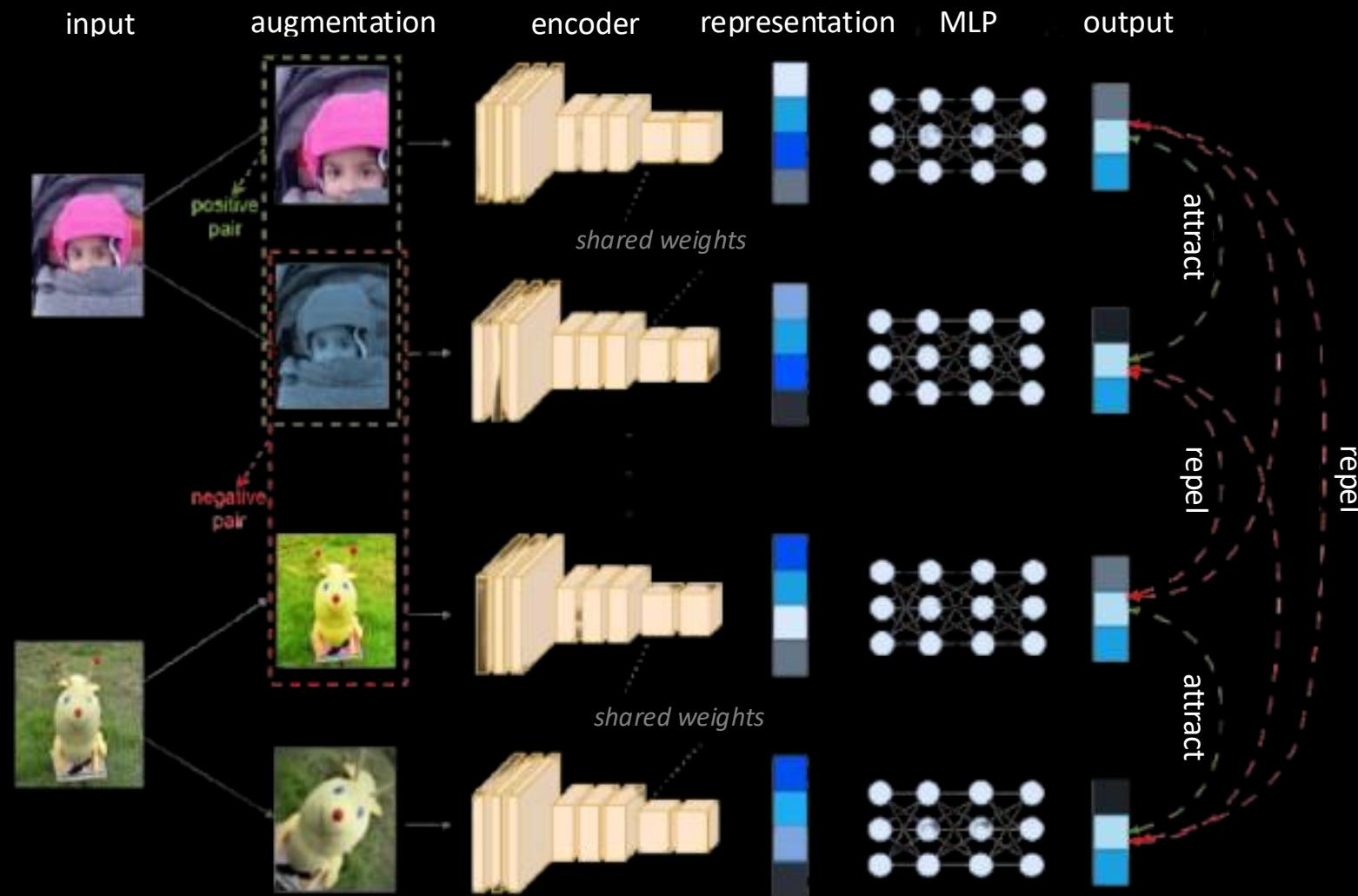
contrastive learning

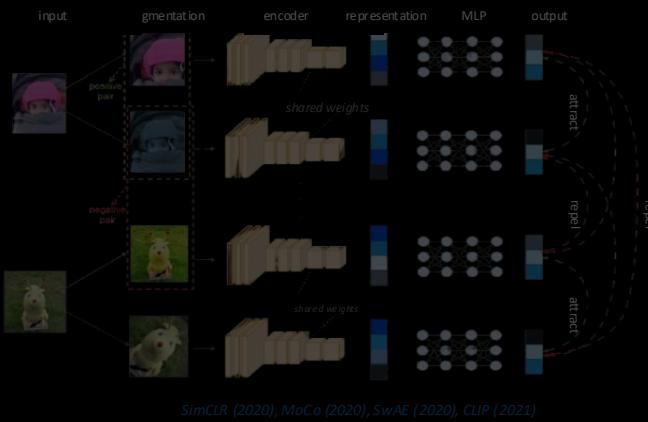
information maximization



Joint embedding and architectural methods (JEAM)

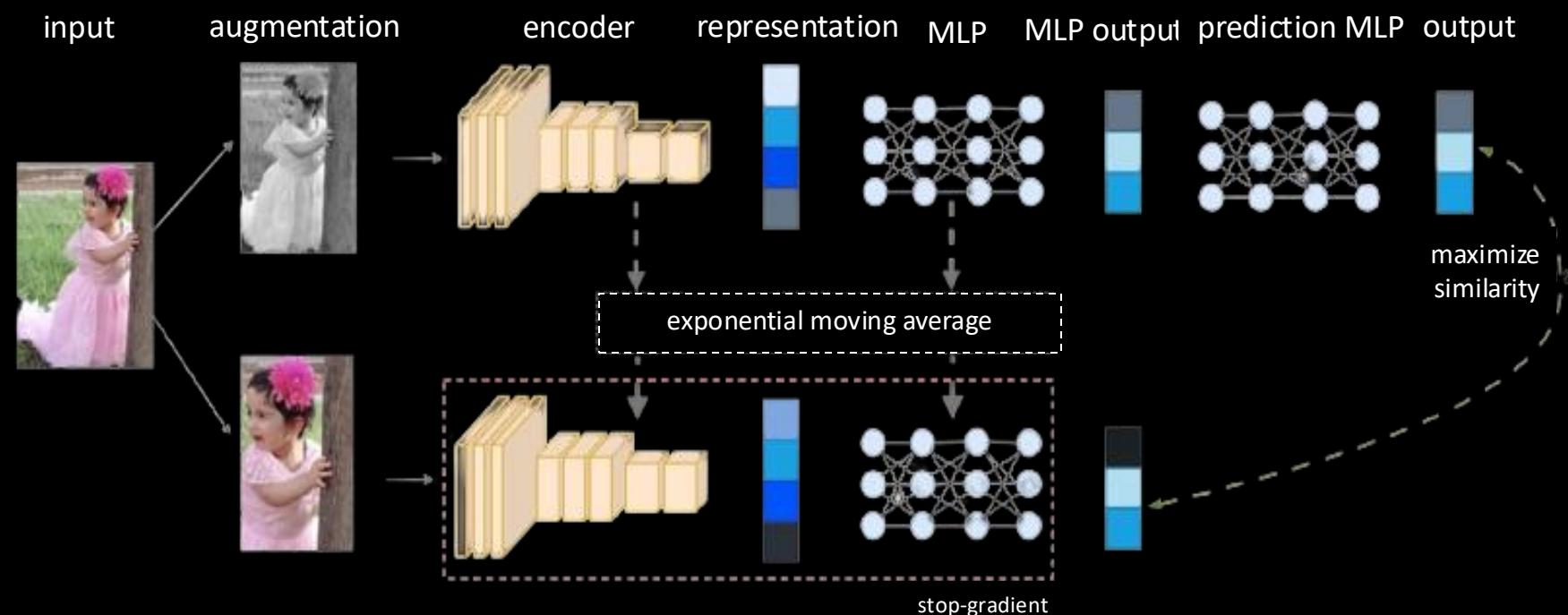
Contrastive learning



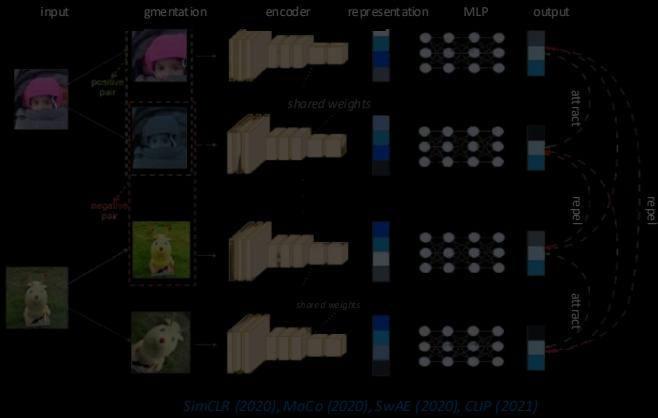


Contrastive learning

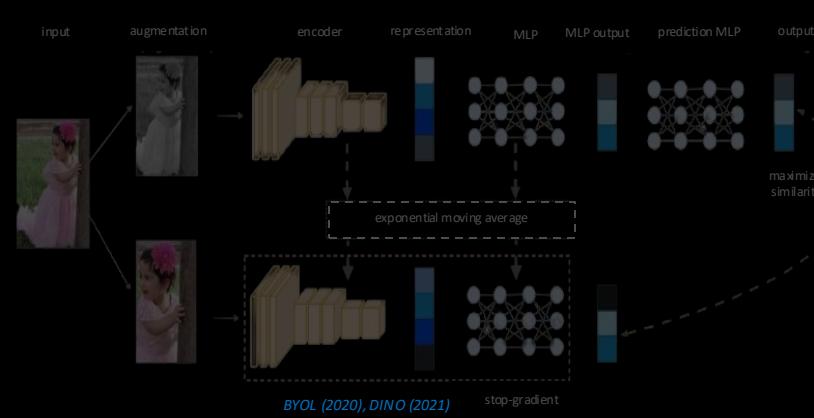
Knowledge distillation



BYOL (2020), DINO (2021)

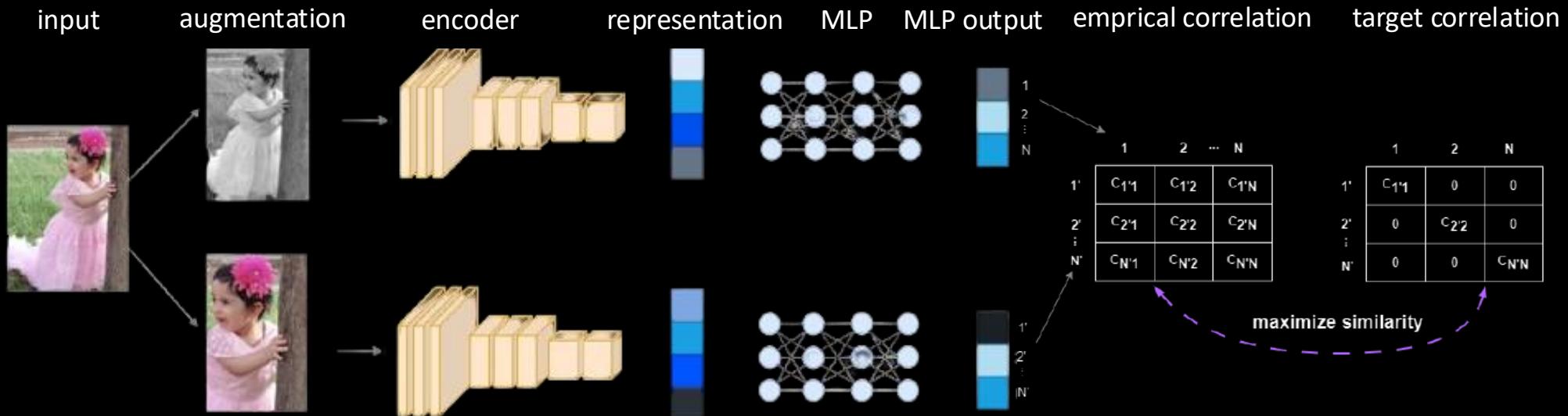


Contrastive learning



Knowledge distillation

Information maximization redundancy reduction

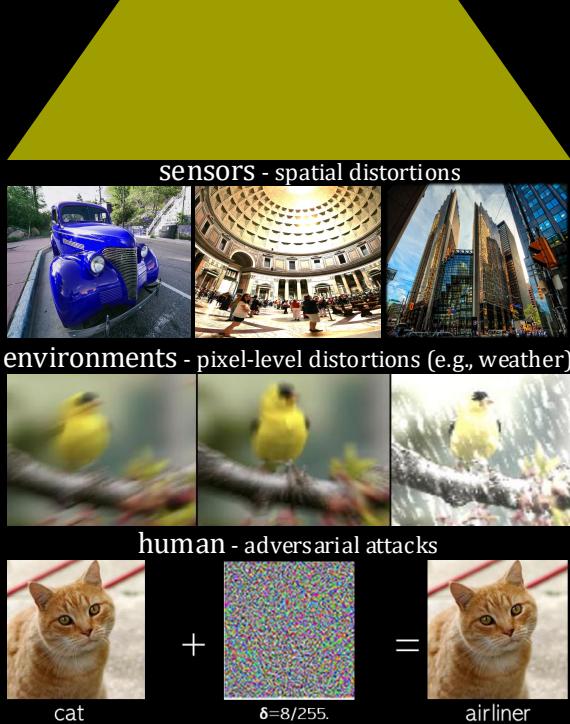


Barlow Twin (2020), VICReg(2021)

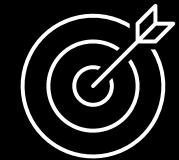
Goal



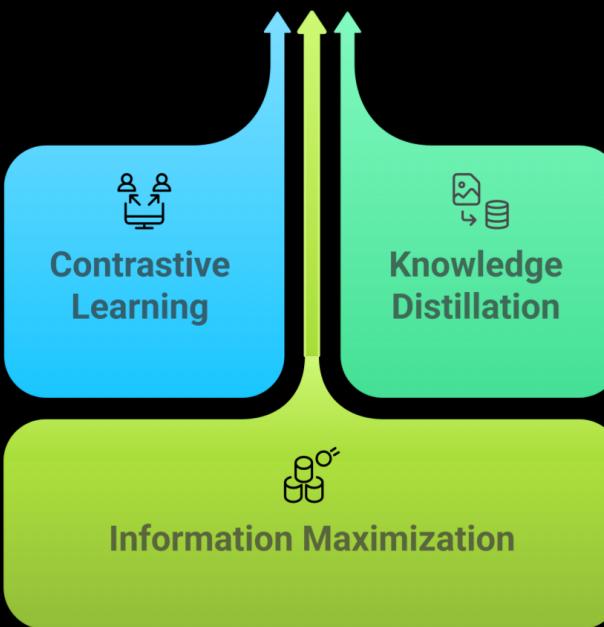
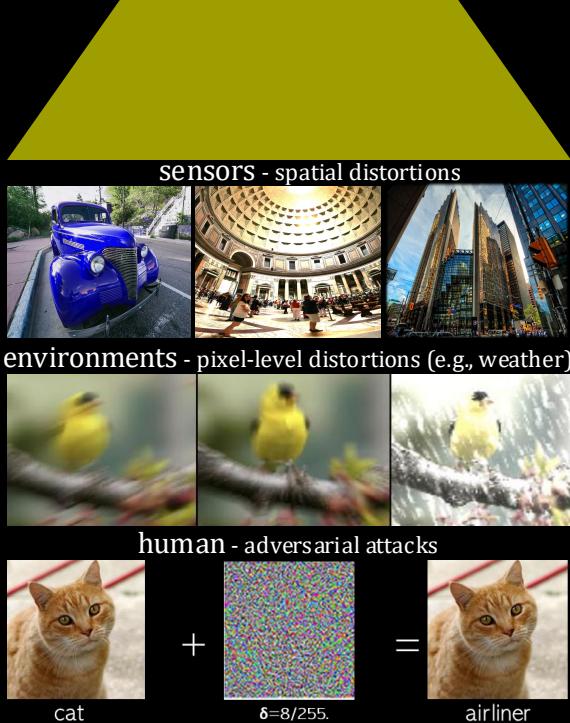
Improve **robustness** in SSL
against real-world challenges

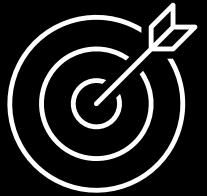


Refined Goal



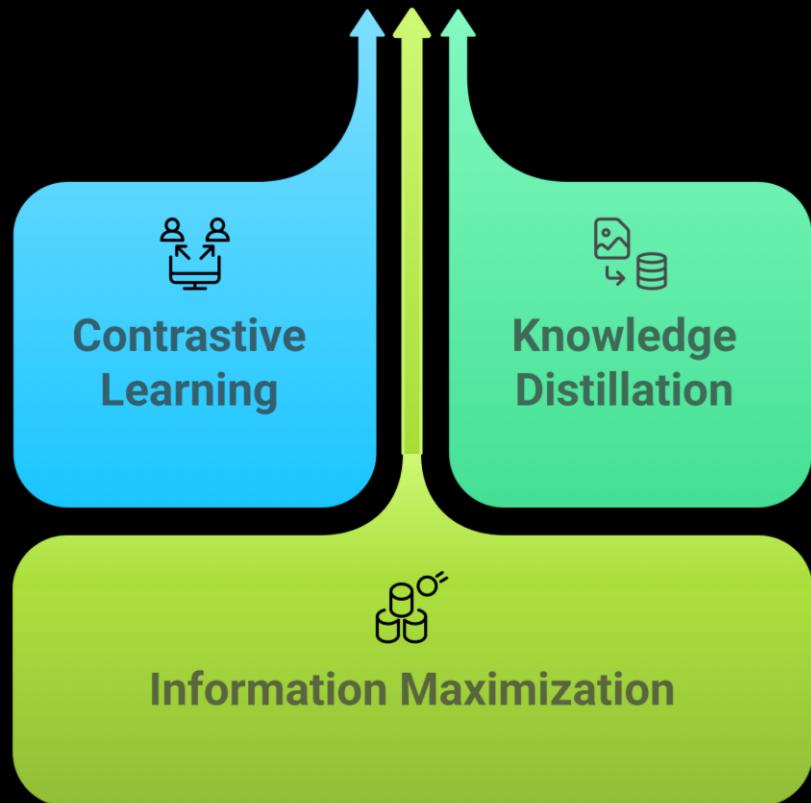
Improve robustness in SSL as plug-N-play
against real-world challenges





plug-N-play

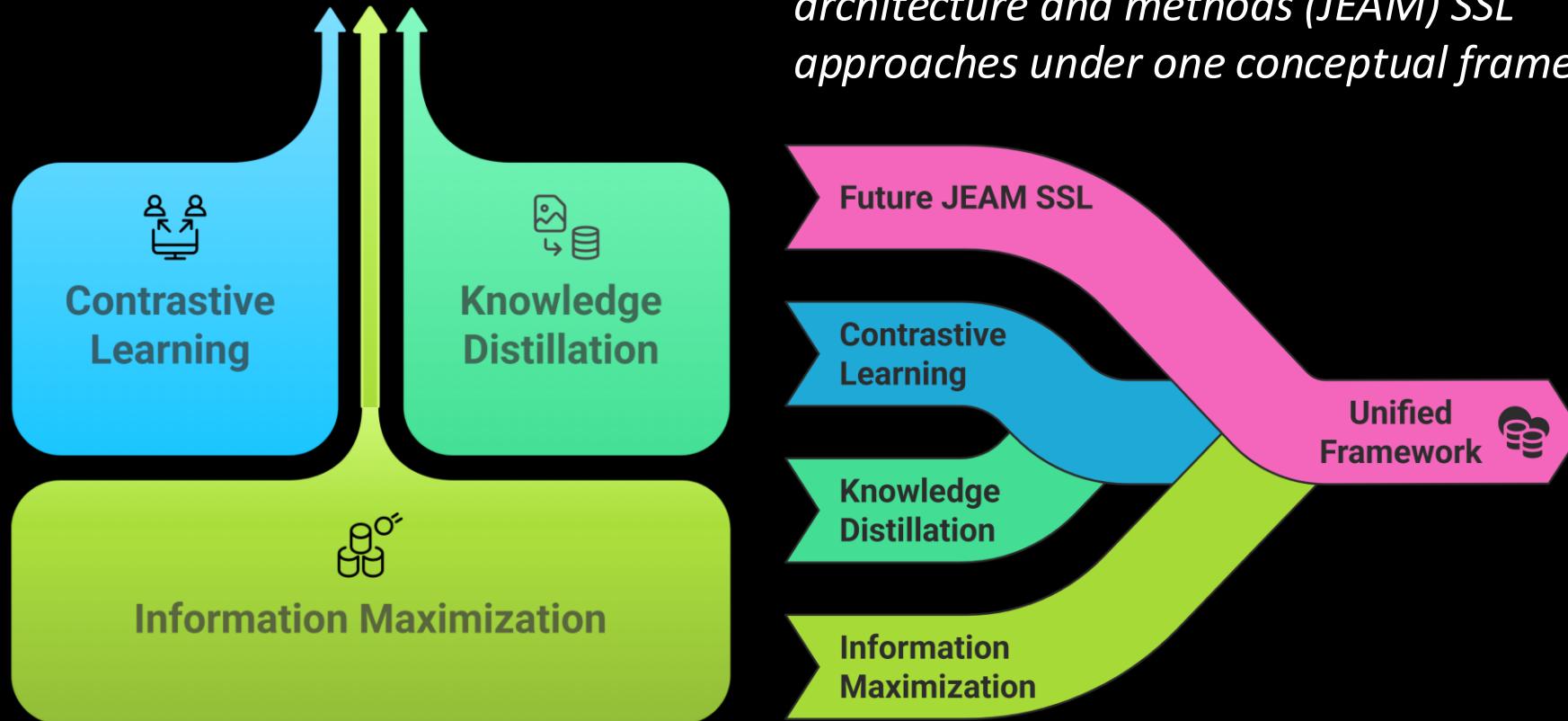
open to all JEAM SSL approaches

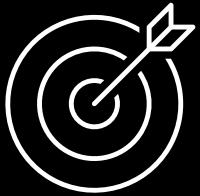




plug-N-play

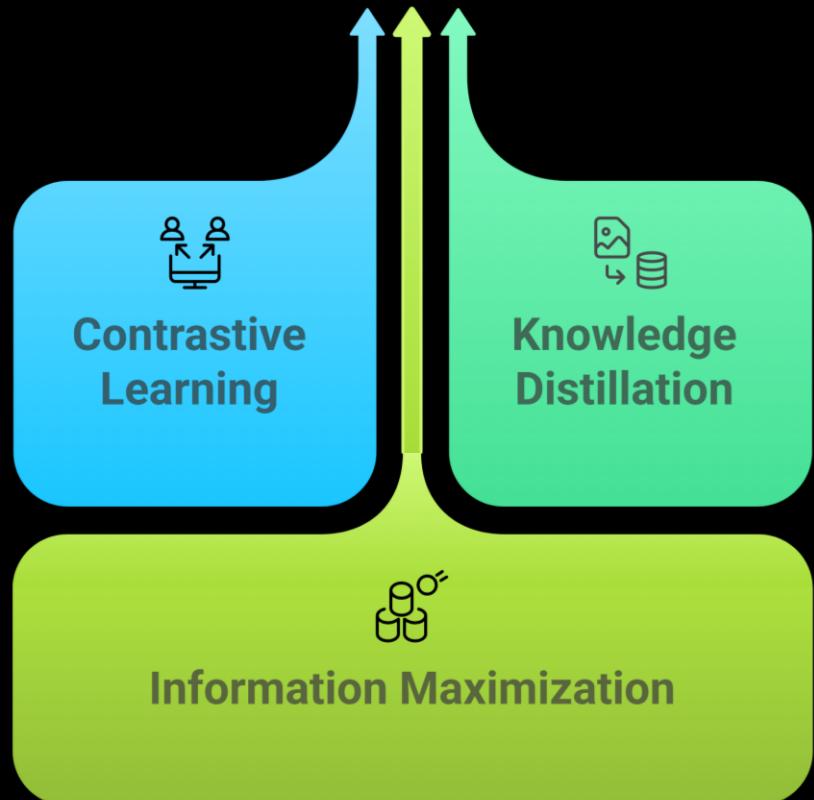
open to all JEAM SSL approaches



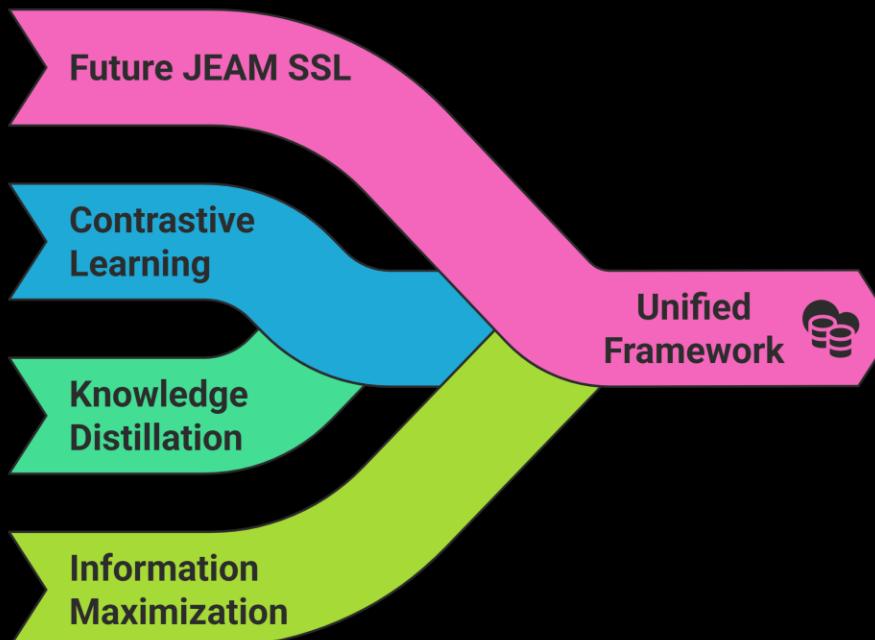


plug-N-play

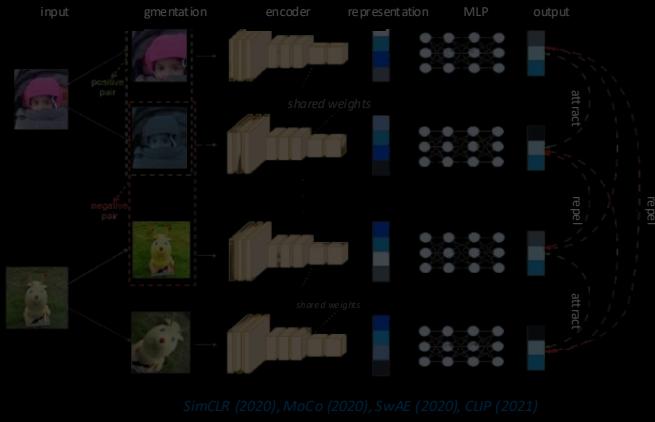
open to all JEAM SSL approaches



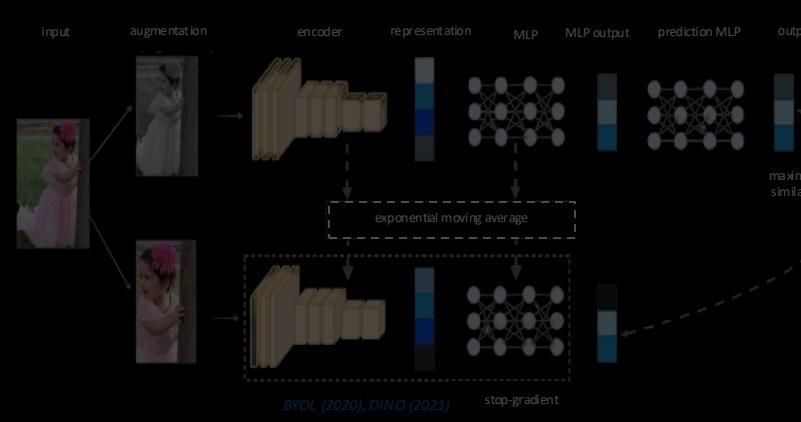
Foundational study of Joint embedding architecture and methods (JEAM) SSL approaches under one conceptual framework



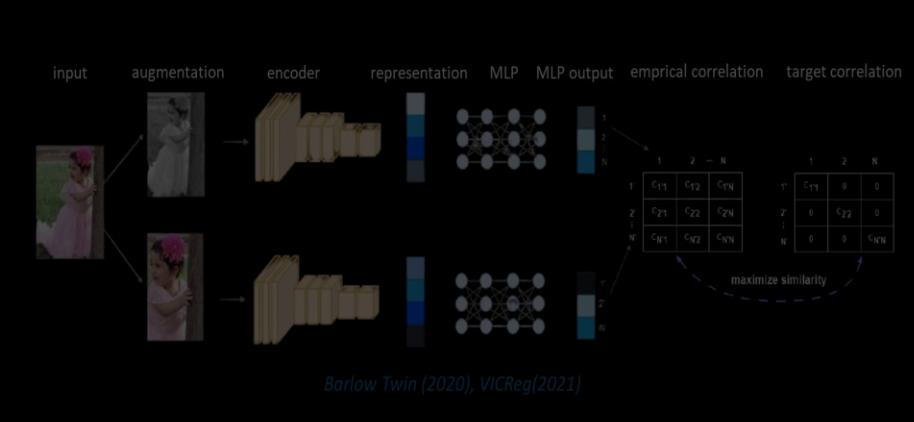
- **Harness commonalities**
Focus on shared aspects of SSL approaches.
- **Understand differences**
How one SSL approach is different than others, at which level.
- **Approach agnostic**
Aims to introduce robustness and domain-awareness to all.



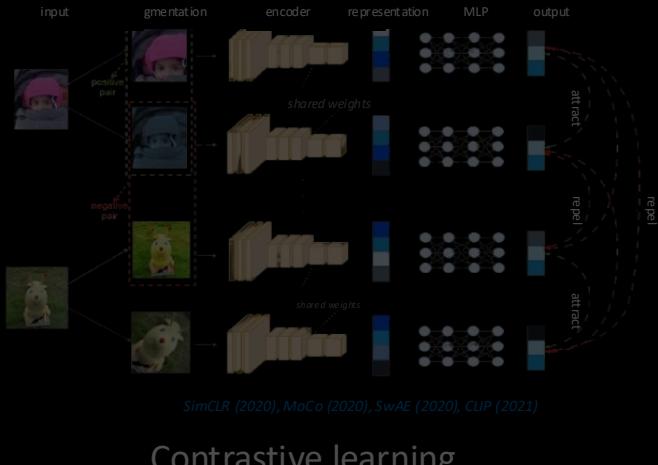
Contrastive learning



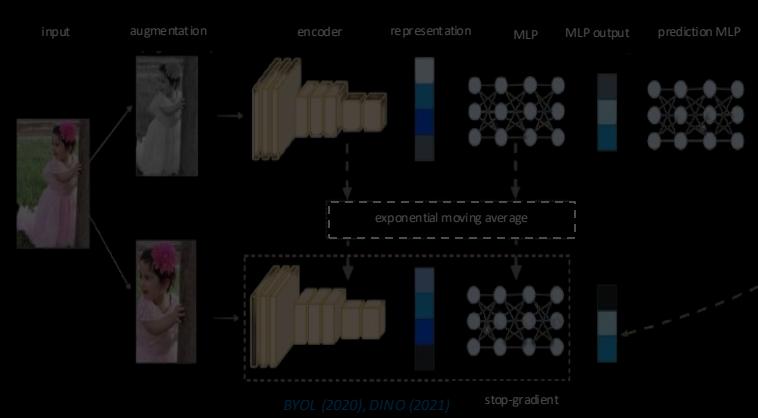
Knowledge distillation



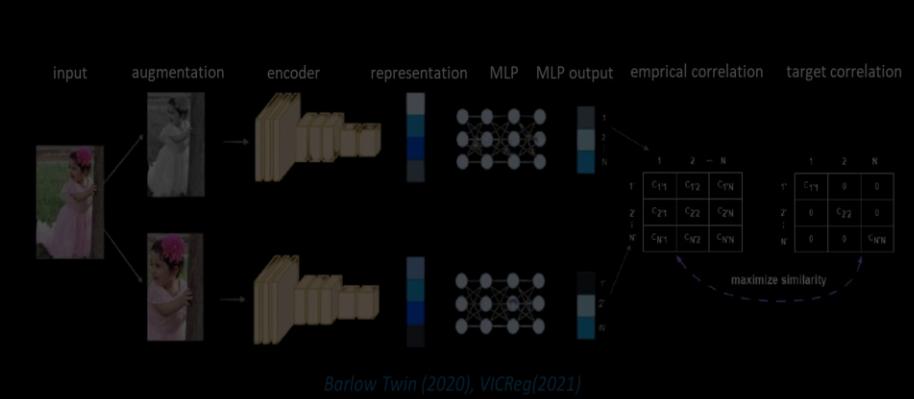
Information maximization



Contrastive learning



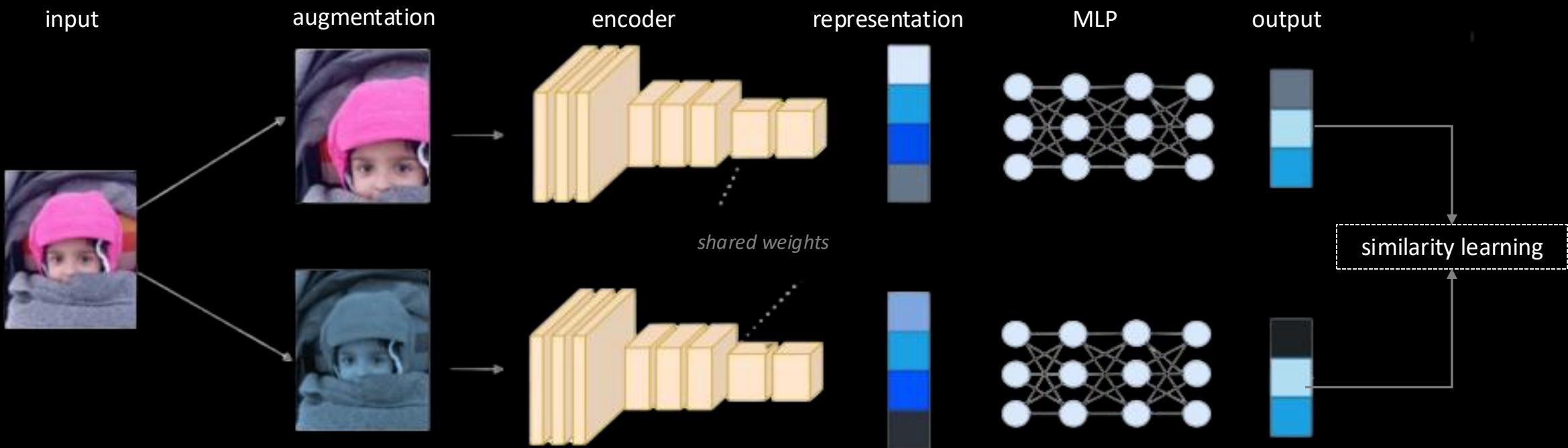
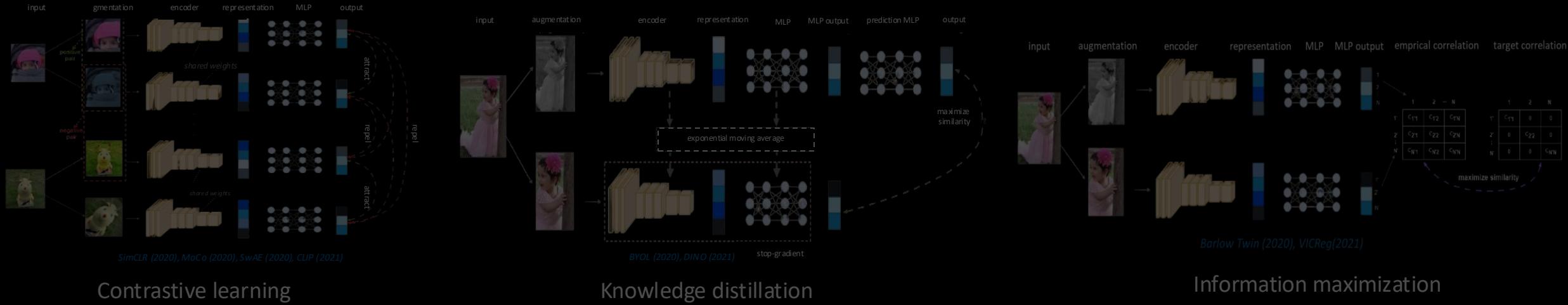
Knowledge distillation



Information maximization

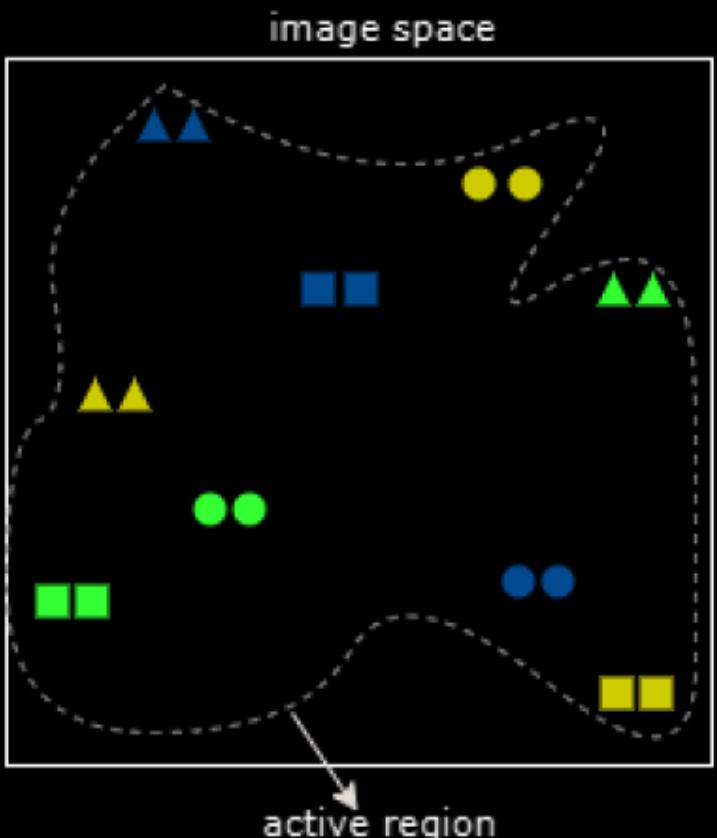
What is common in JEAM SSL approaches?

- ✓ input image → augmentations → generated view pairs
- ✓ Align representations of views in embedding space → employing similarity mechanism

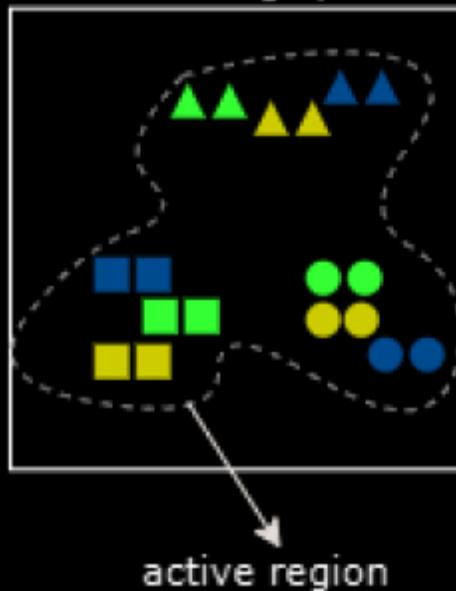


Graphical illustration

- augmented view
pair of same visual
concept
- ••
••
augmented view
pair of similar
visual concepts
- ••
••
augmented view
pair of different
visual concepts



embedding space

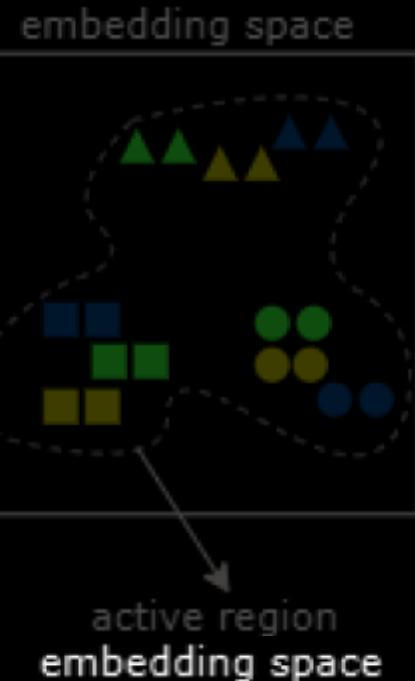
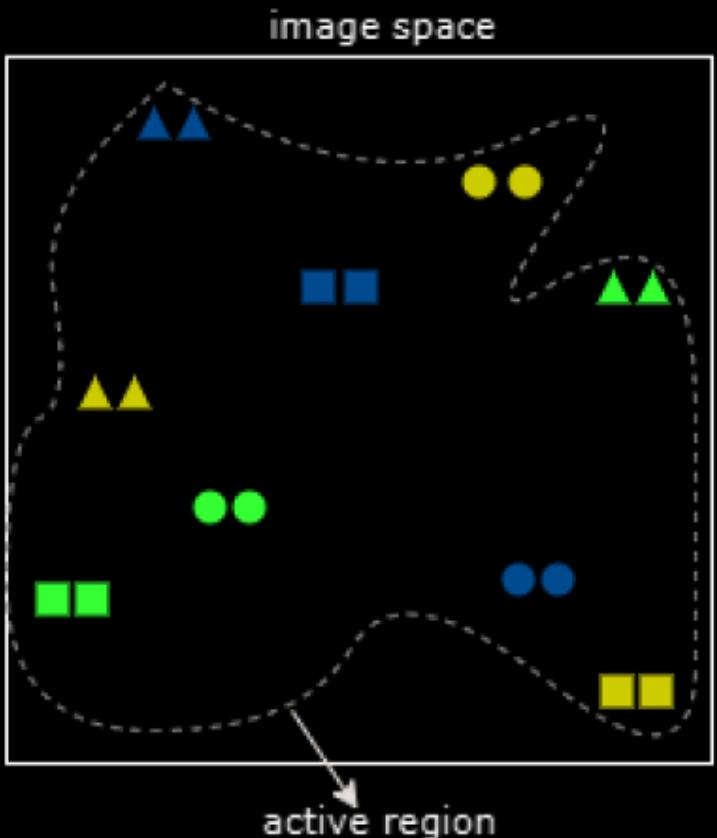


ideal

$$\lim_{\mathcal{L}_{ssl} \rightarrow 0} \Theta_{fg} = \Theta_{fg}^*$$

Graphical illustration

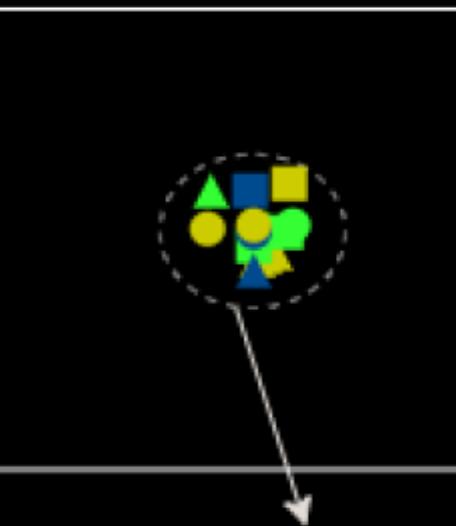
- augmented view pair of same visual concept
- augmented view pair of similar visual concepts
- augmented view pair of different visual concepts



ideal

$$\lim_{\mathcal{L}_{ssl} \rightarrow 0} \Theta_{fg} = \Theta_{fg}^*$$

trivial solution



joint learning parameters Θ_{fg} optimal joint parameters set Θ_{fg}^*

Analytical argument

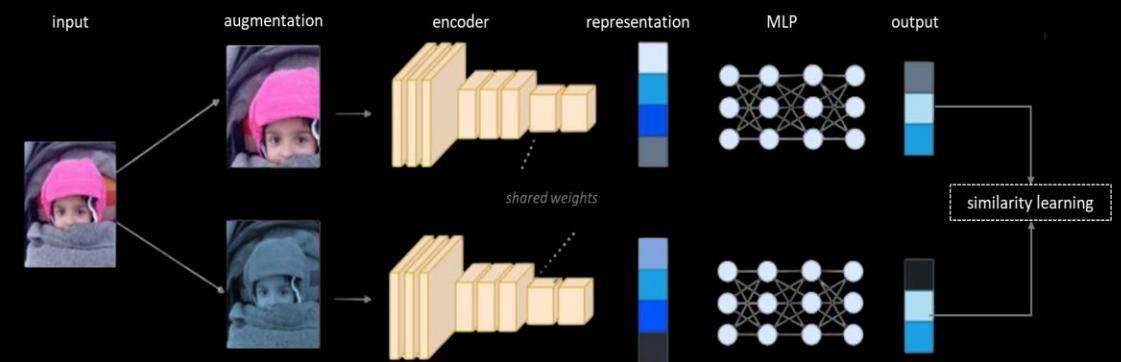
What collapsed representations for common SSL schema?

- Supervised learning: Unique labels force distinct representations, preventing collapse.
- Unsupervised learning: Reconstruction requires unique representations for each input.
- SSL schema: No distinct encoding requirement to reduce loss, allowing collapse into a single embedding.

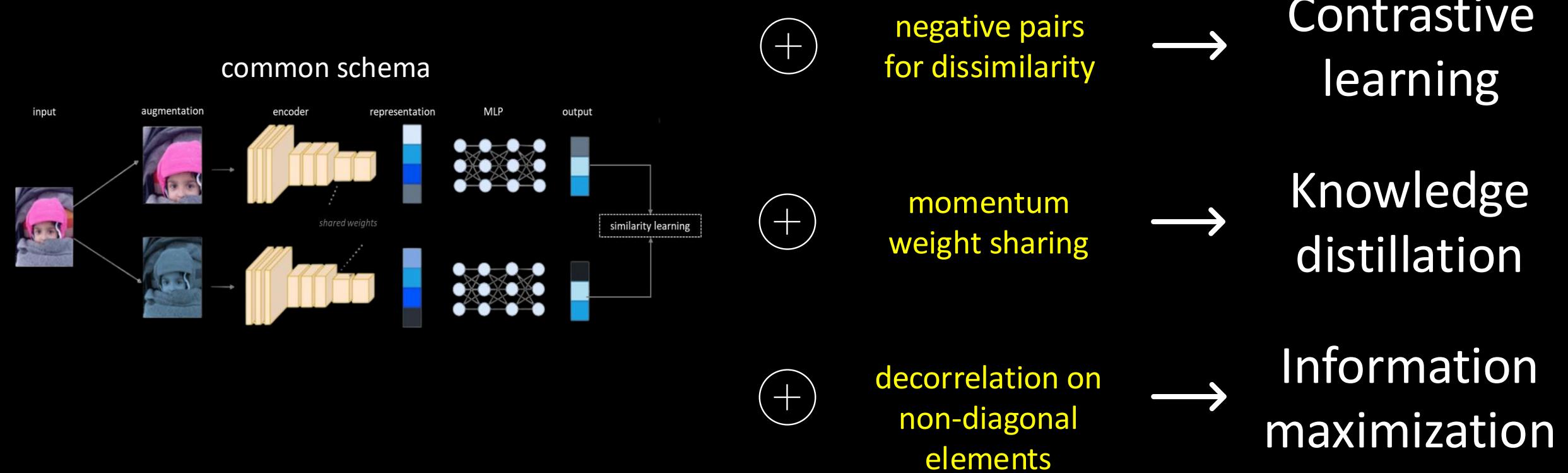
While supervised and unsupervised imposes constraints however SSL schema does not have explicit constraint.

Explicit constraints

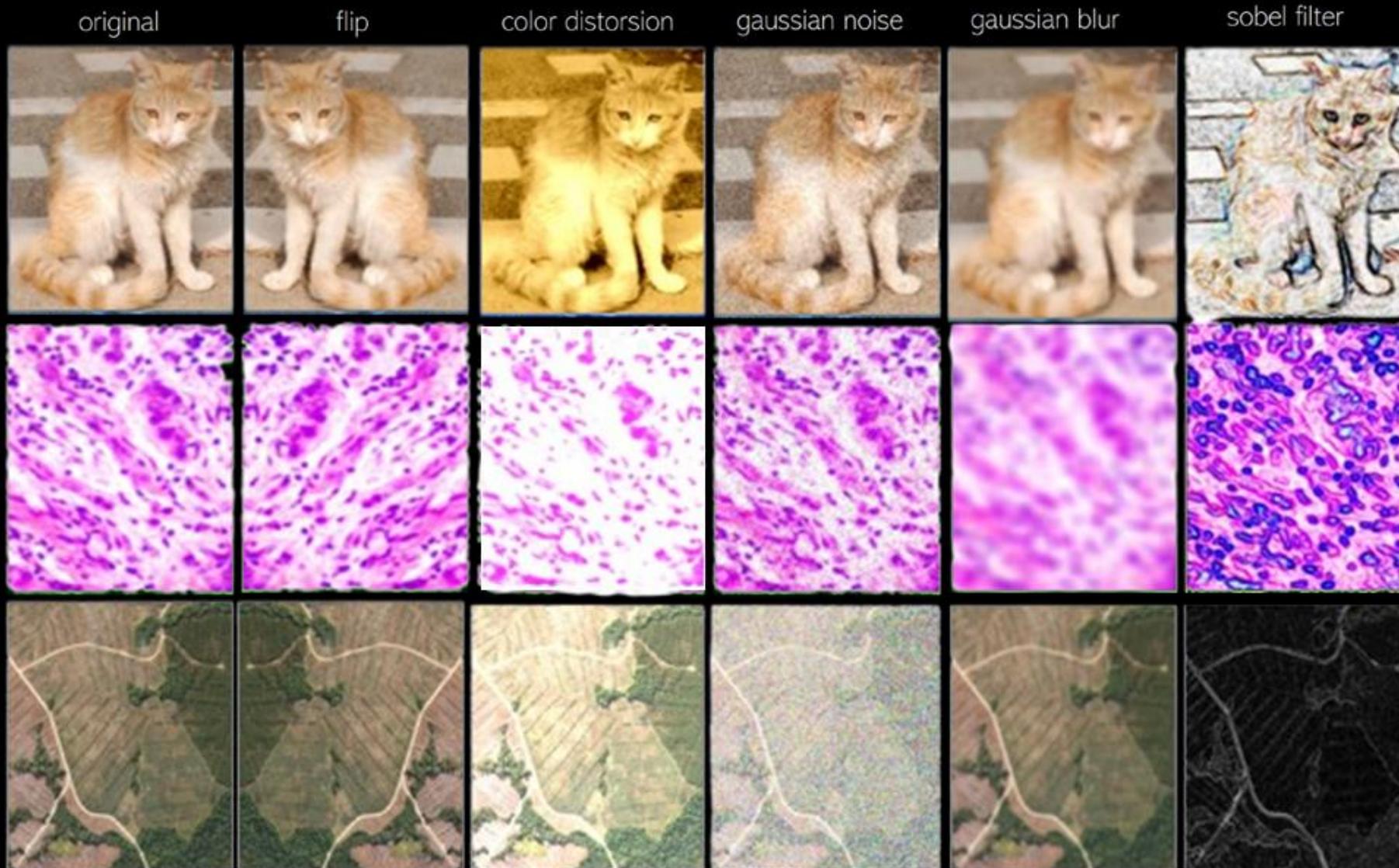
common schema



Explicit constraints

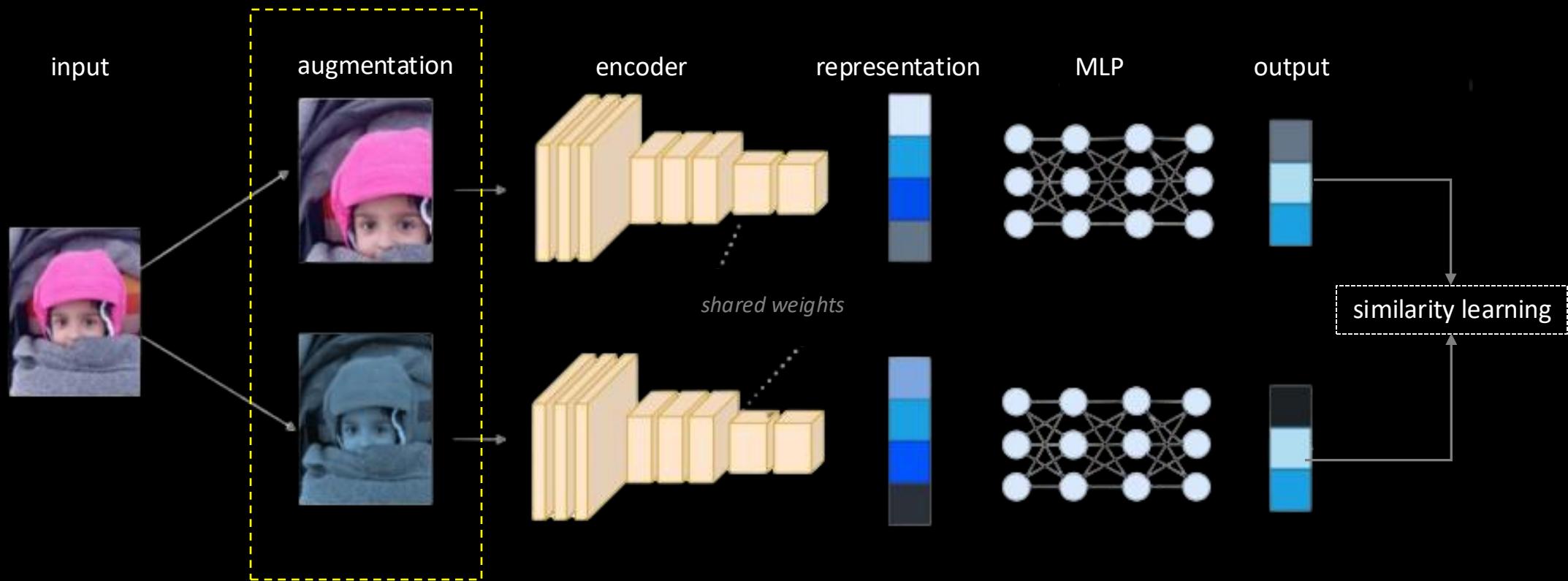


Domain sensitivity



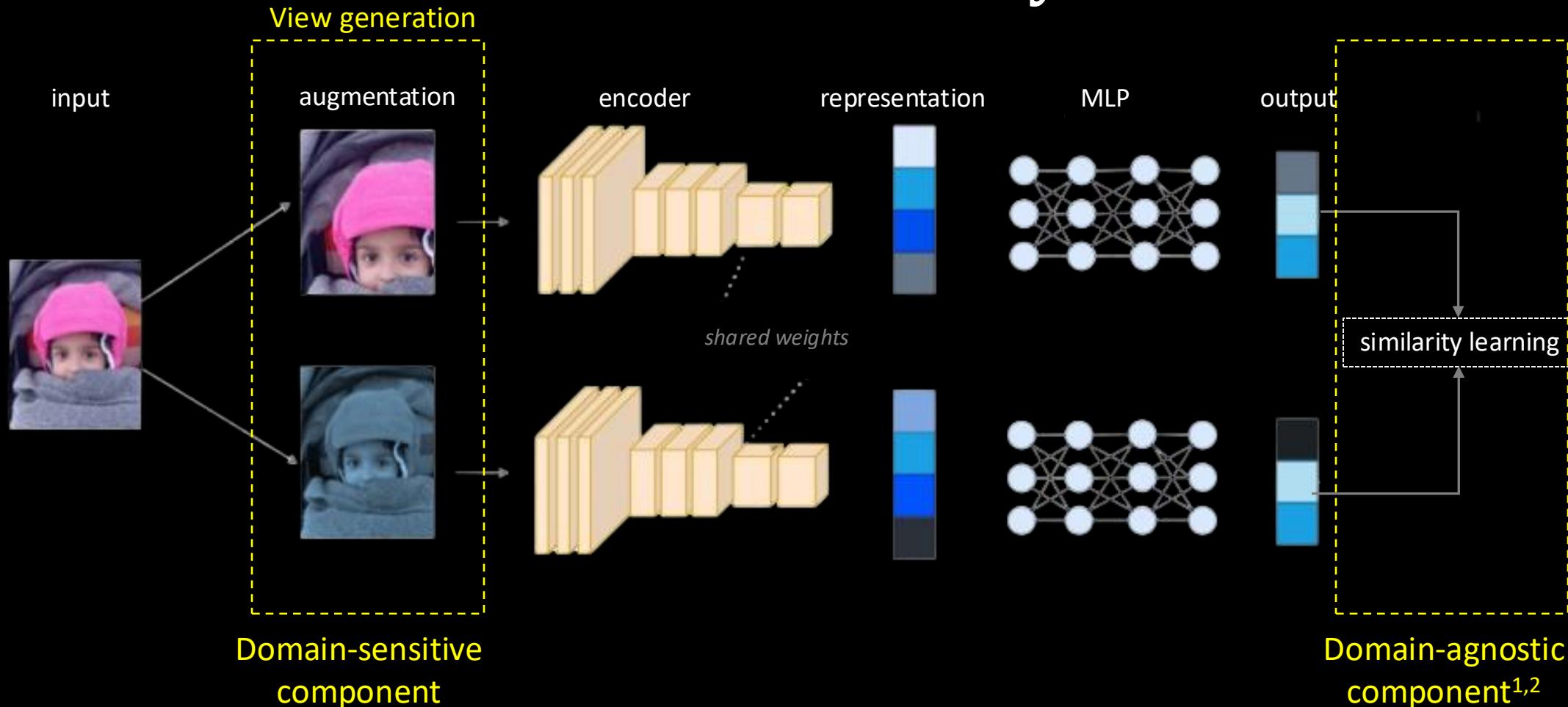
Augmentation serves the purpose for natural scenes but destroys the visual other domains.

Domain sensitivity



- *View generation in SSL methods are solely depends on **augmentations** for invariance in representation learning in SSL.*
- *Augmentation are designed for natural scenes.*

Domain sensitivity



- *View generation component in common schema of SSL is sensitive to domains because of applied augmentations.*
- *Being common across SSLs—suitable for improvements towards robustness and domain-awareness in plug-N-play manner.*
- *Loss objective considered to be domain-agnostic as operates on embedding space.*

¹Chopra, M., Chhipa, P. C., & Liwicki, M. (2023, June). Domain adaptable self-supervised representation learning on remote sensing satellite imagery. In 2023 International Joint Conference on Neural Networks (IJCNN)

²Gupta, E., Gupta, V., Chopra, M., Chhipa, P. C., & Liwicki, M. (2023, June). Learning self-supervised representations for label efficient cross-domain knowledge transfer on diabetic retinopathy fundus images. In 2023 International Joint Conference on Neural Networks (IJCNN)

How robust are state-of-the-art self-supervised
representation learning methods?

Global geometric transformations-perspective distortion (ImageNet-PD)

8 subsets, 4 without background and 4 with padded background

Original



Left View



Right View



Top View



Bottom View



Padded Background

Left View



Right View



Top View



Bottom View



Global geometric transformations-perspective distortion (ImageNet-PD)

8 subsets, 4 without background and 4 with padded background

Original



Left View



Right View



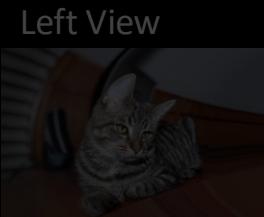
Top View



Bottom View



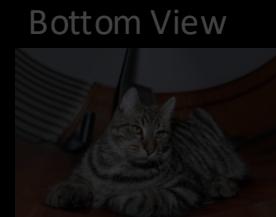
Padded Background



Left View



Right View



Top View



Photometric and localized structural corruptions (ImageNet-C)

15 subsets, 5 intensity level for each subsets

Gaussian Noise



Shot Noise



Impulse Noise



Defocus Blur



Motion Blur



Zoom Blur



Snow



Frost



Fog



Brightness



Contrast



Elastic



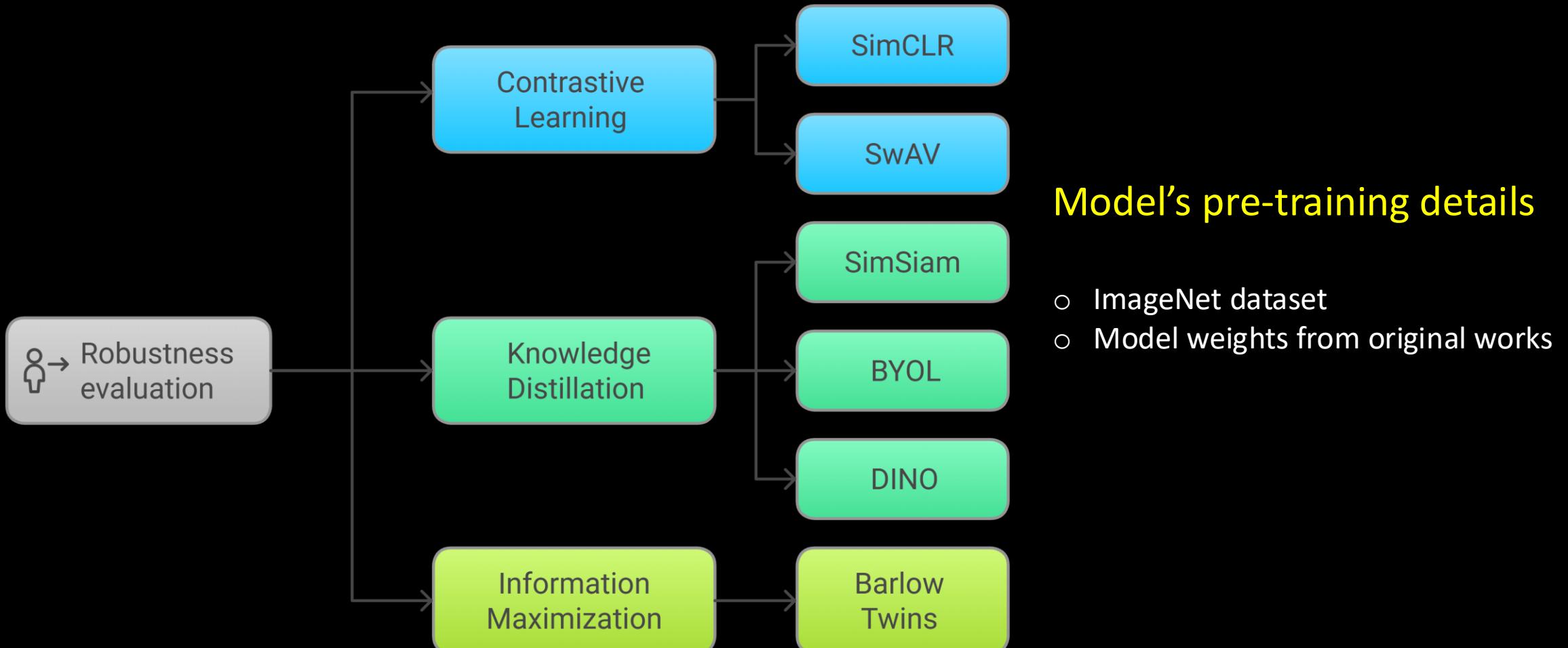
Pixelate



JPEG



Existing SSL's robustness



SimCLR: Ting Chen et al. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

SwAV: Mathilde Caron et al. Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems, 33:9912–9924, 2020.

SimSiam: Xinlei Chen et al. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15750–15758, 2021.

BYOL: Jean-Bastien Grill et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020.

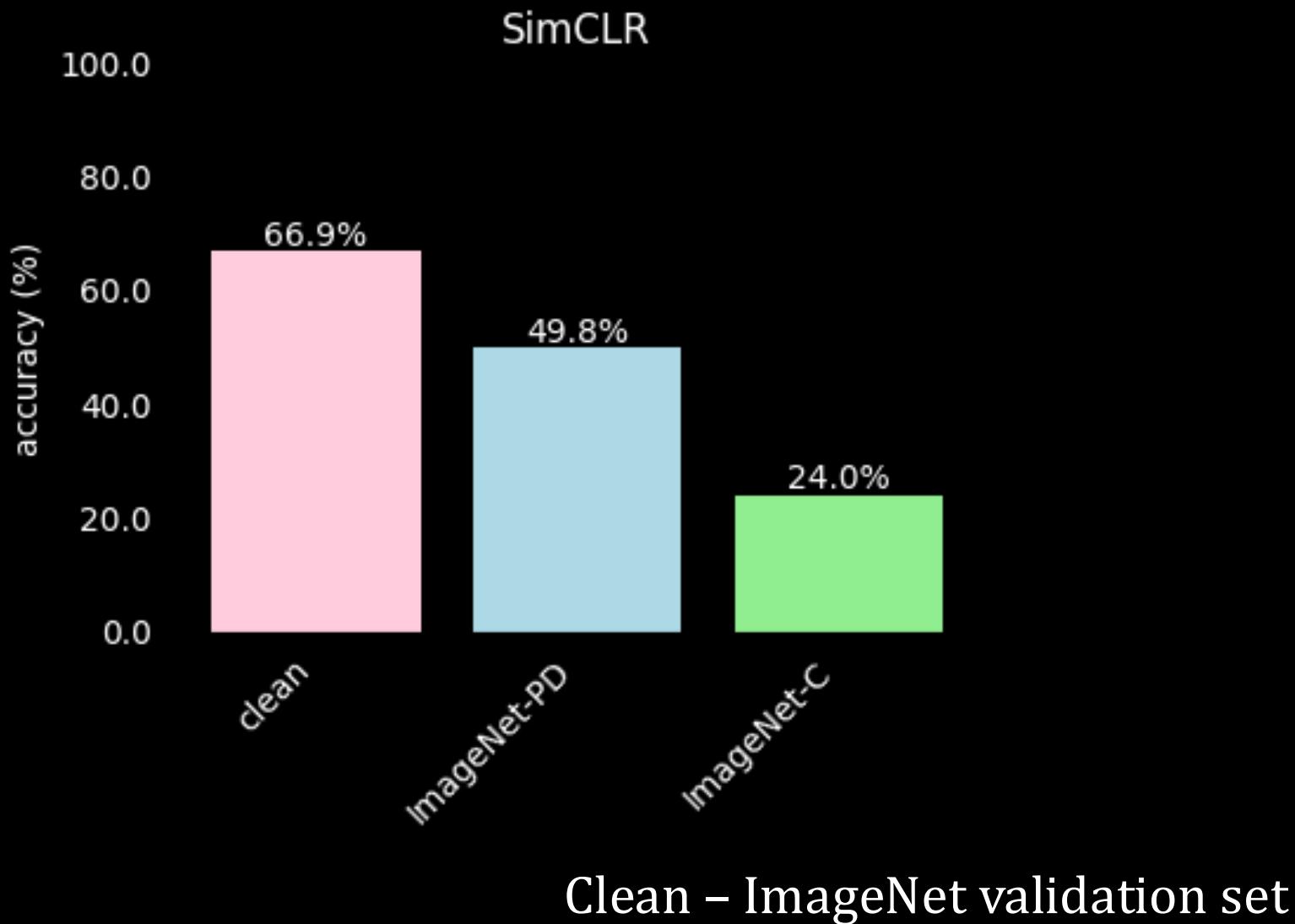
DINO: Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021.

Barlow Twins: Jure Zbontar et al. Barlow twins: Self-supervised learning via redundancy reduction. In International conference on machine learning, pages 12310–12320. PMLR, 2021.

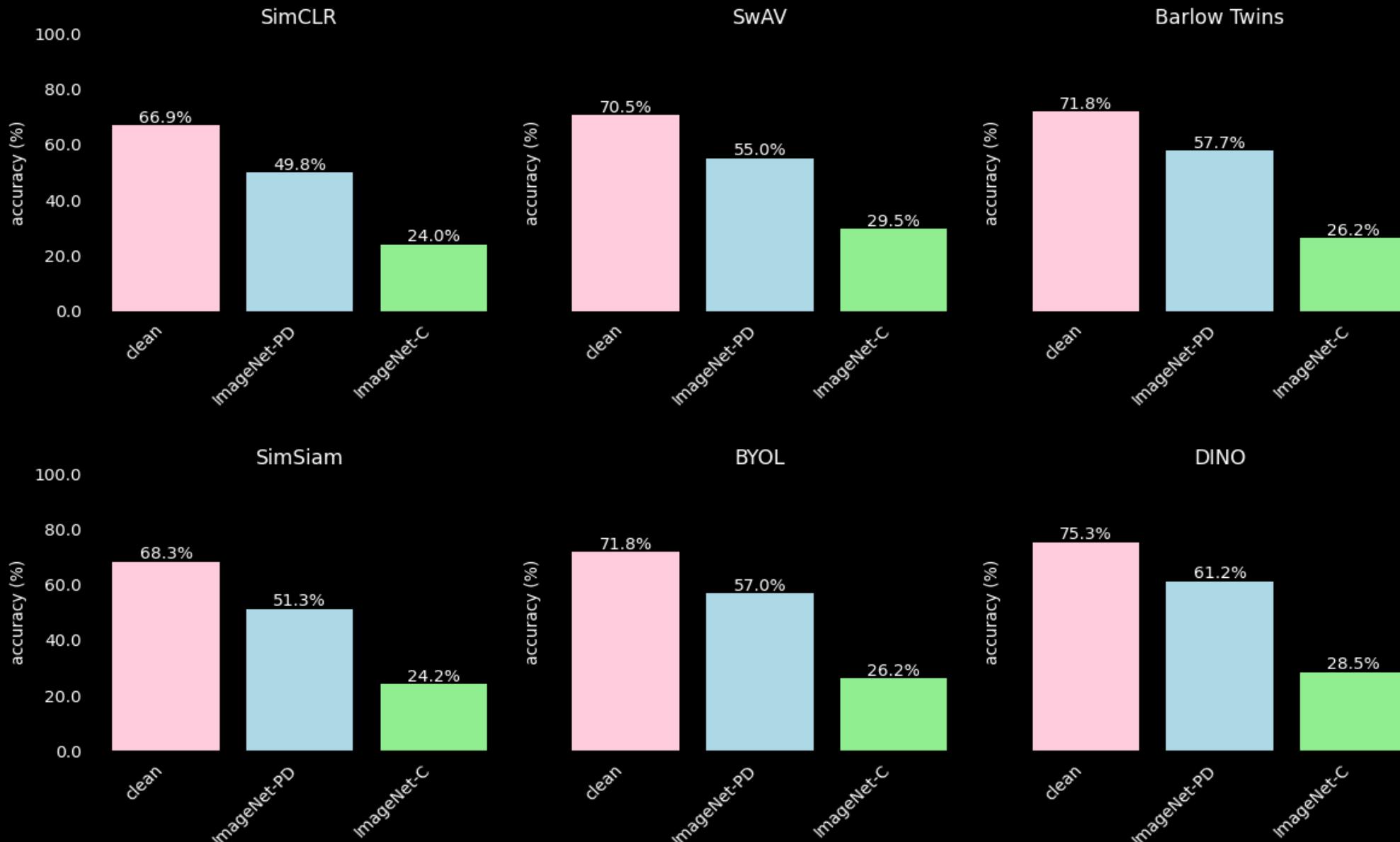
Prakash Chandra Chhipa et al. Can self-supervised representation learning methods withstand distribution shifts and corruptions?, International Conference on Computer Vision (ICCV) Workshops 2023.

Prakash Chandra Chhipa (2025). Towards Robust and Domain-aware Self-supervised Representation Learning (Doctoral dissertation, Luleå tekniska universitet).

Existing SSL's robustness



Existing SSL's robustness



Existing SSL's robustness

Findings

- **SSL models degrade under distribution shifts**—revealing fundamental vulnerabilities.
- **Real-world distortions** (perspective or pixel-level) pose major challenges for SSL.

Existing SSL's robustness

Findings

- **SSL models degrade under distribution shifts**—revealing fundamental vulnerabilities.
- **Real-world distortions** (perspective or pixel-level) pose major challenges for SSL.
- **ImageNet-C corruptions** often resemble adversarial noise, exposing gaps in adversarial robustness.
- **Moving forward:**
 - Robustness to perspective distortion
 - Adversarial robustness

Robustness against perspective distortion

Perspective distortion (PD)

Alters object appearance by changing shape, size, orientation, and spatial relationships in non-linear ways.



Perspective distortion (PD)

Alters object appearance by changing shape, size, orientation, and spatial relationships in non-linear ways.

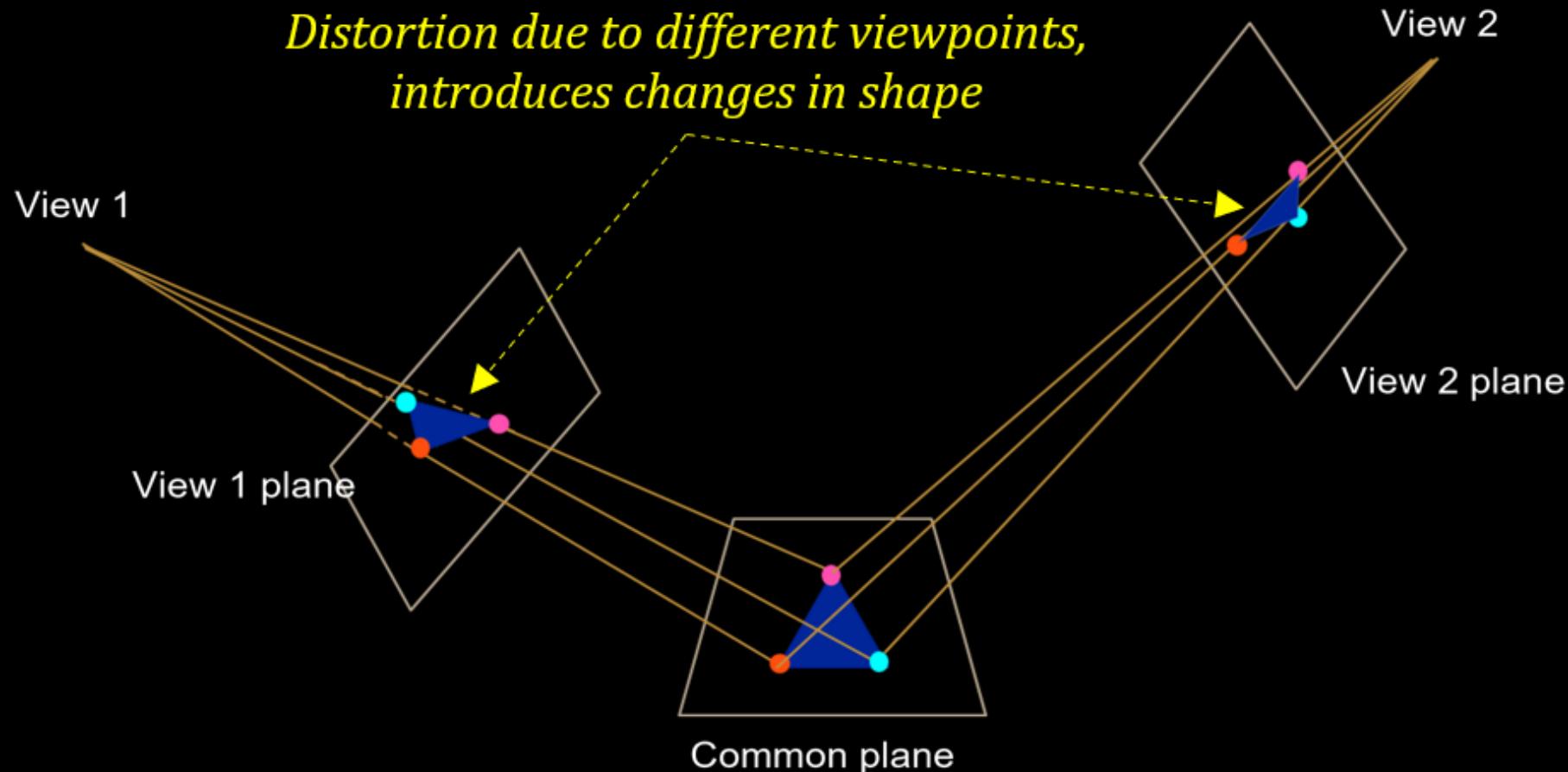


Perspective distortion (PD)

Alters object appearance by changing shape, size, orientation, and spatial relationships in non-linear ways.



Explanation- Why PD



Security cameras

Why it is critical to handle?

- Real-world images often deviate from ideal front-to-parallel views—e.g., security cameras, driving scenes, and mobile snapshots.



Crowds

Why it is critical to handle?

- Real-world images often deviate from ideal front-to-parallel views—e.g., security cameras, driving scenes, and mobile snapshots.
- Crowd and traffic scenes are often captured from elevated or oblique angles, distorting object geometry and scale.



Driving scenes

Why it is critical to handle?

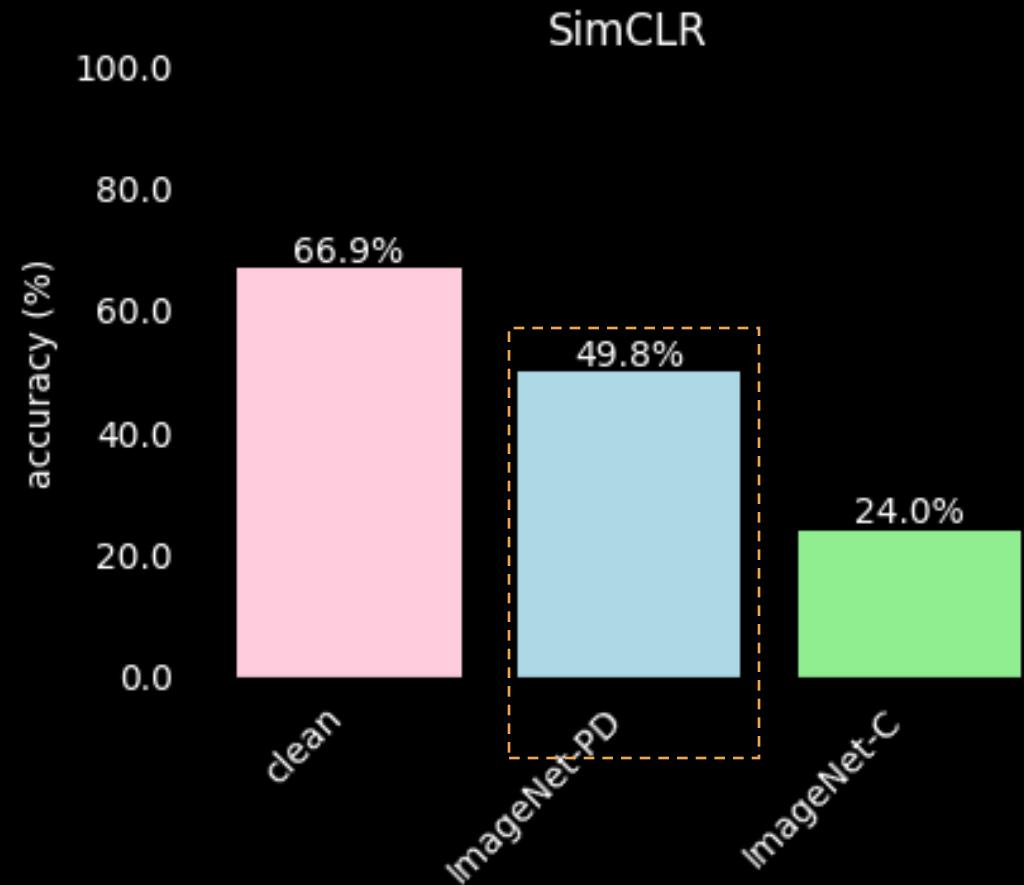
- Real-world images often deviate from ideal front-to-parallel views—e.g., security cameras, driving scenes, and mobile snapshots.
- Crowd and traffic scenes are often captured from elevated or oblique angles, distorting object geometry and scale.



Existing SSL models

Why it is critical to handle?

- **Real-world images often deviate from ideal front-to-parallel views**—e.g., security cameras, driving scenes, and mobile snapshots.
- **Crowd and traffic scenes** are often captured from elevated or oblique angles, distorting object geometry and scale.
- **Models trained on canonical views fail to generalize**—leading to poor performance in real-world tasks like detection, counting, or segmentation.
 - *SSL models poor robustness in previous section (RQ3)*



What makes perspective distortion even more challenging?

Affine Transformation Limitations

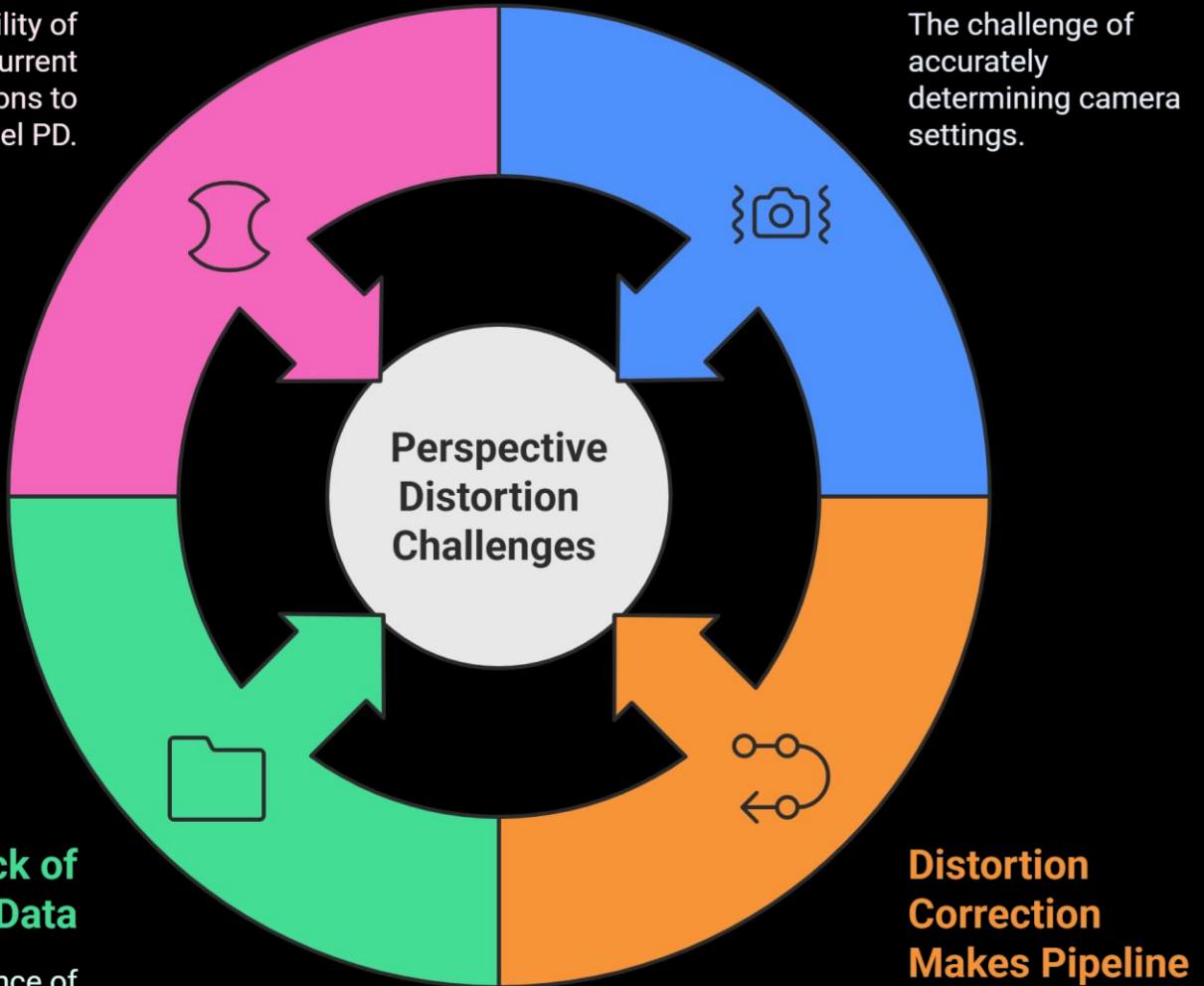
The inability of current transformations to model PD.

Lack of Training Data

The absence of comprehensive datasets for model training.

Camera Parameter Estimation

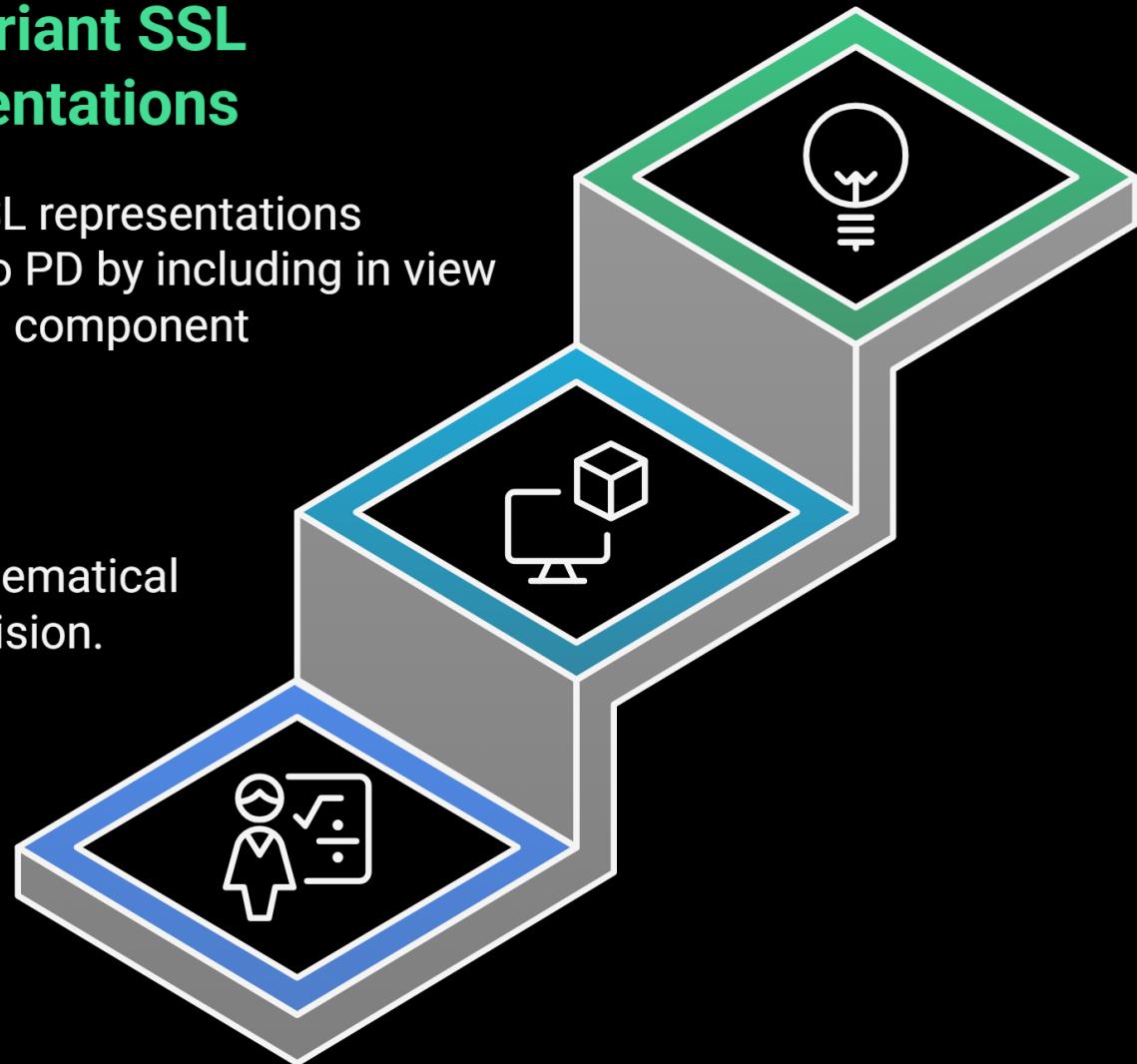
The challenge of accurately determining camera settings.



The inefficiency introduced by separate correction stages.

Mitigating perspective distortion

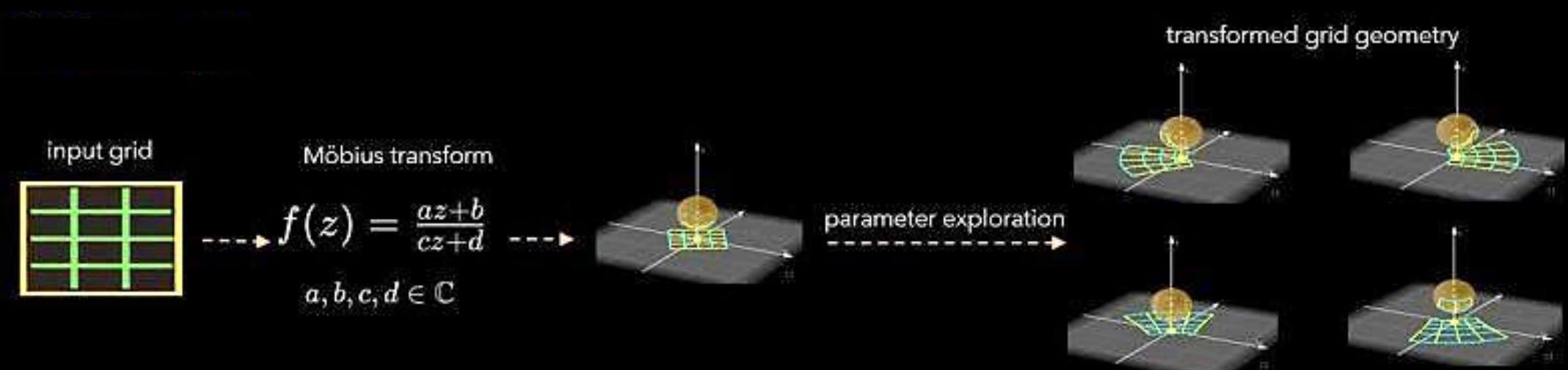
- 1 **Mathematical Modeling**
Synthesizing PD by modelling it mathematically by exploring complex geometry transformations.
- 2 **Computer Vision Transition**
Transitioning from mathematical modeling to computer vision.
- 3 **PD invariant SSL Representations**
Making SSL representations invariant to PD by including in view generation component



Towards making SSL more robust against PD

Möbius-based mitigating perspective distortion (MPD)

1. Modeling: Möbius transform → complex geometry, non-linear, conformal



Conformal transforms - Möbius

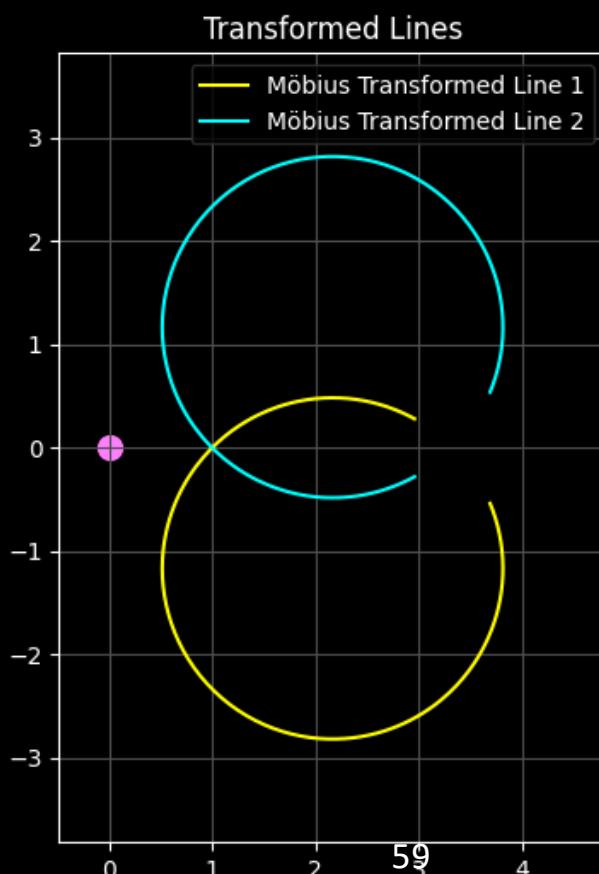
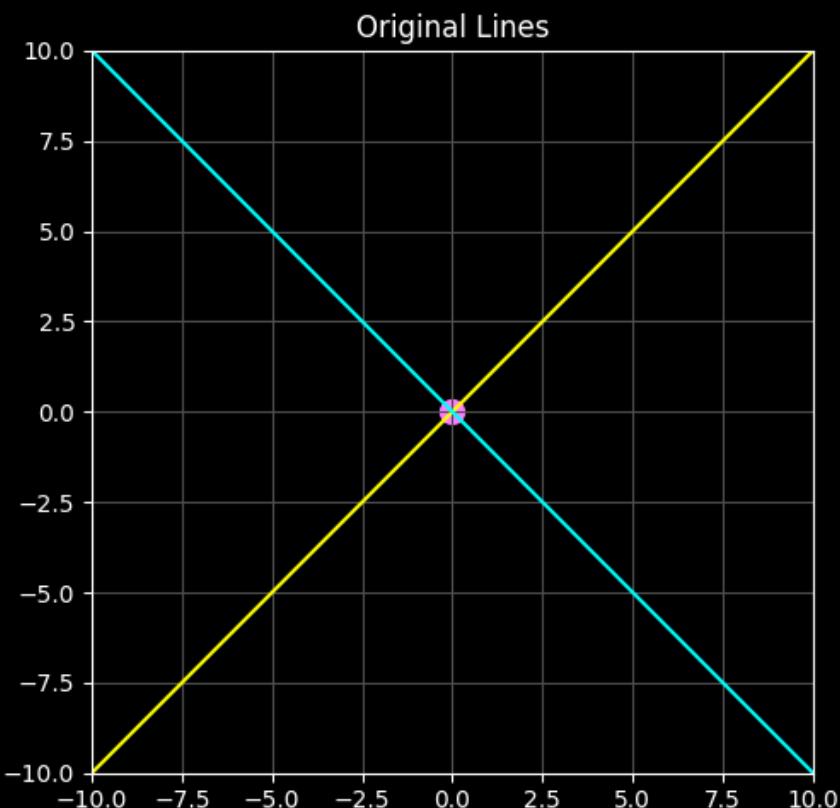
```
def mobius_transform(z, a, b, c, d):
    """Applies a Möbius transformation to a complex number z with given coefficients
    return (a * z + b) / (c * z + d)

# Define parameters for the Möbius transformation
a, b, c, d = 1, 1, 0.3, 1 # Keep these for now but adjust if needed

# Generate lines in the complex plane
x = np.linspace(-10, 10, 400)
line1 = x + 1j * x # 45 degree line
line2 = x - 1j * x # -45 degree line

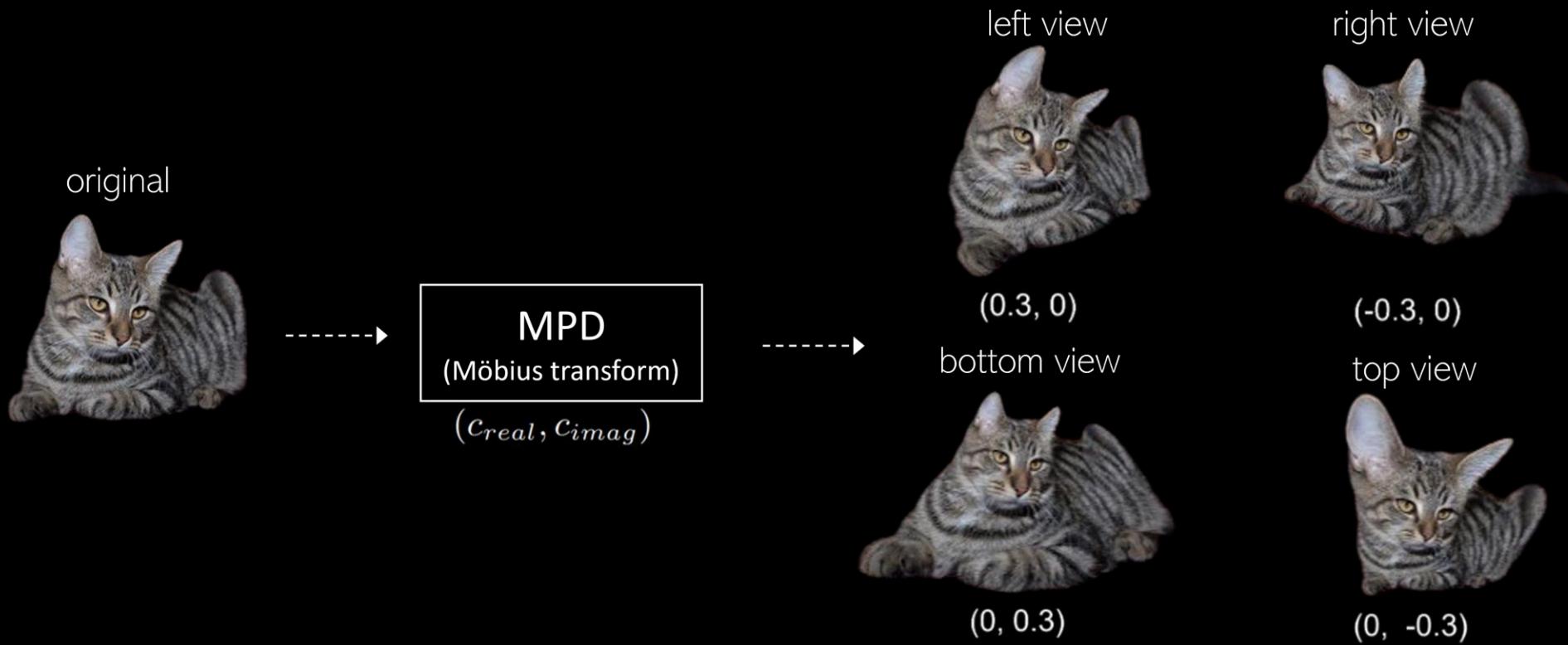
# Apply Möbius transformation to the lines
transformed_line1 = mobius_transform(line1, a, b, c, d)
transformed_line2 = mobius_transform(line2, a, b, c, d)

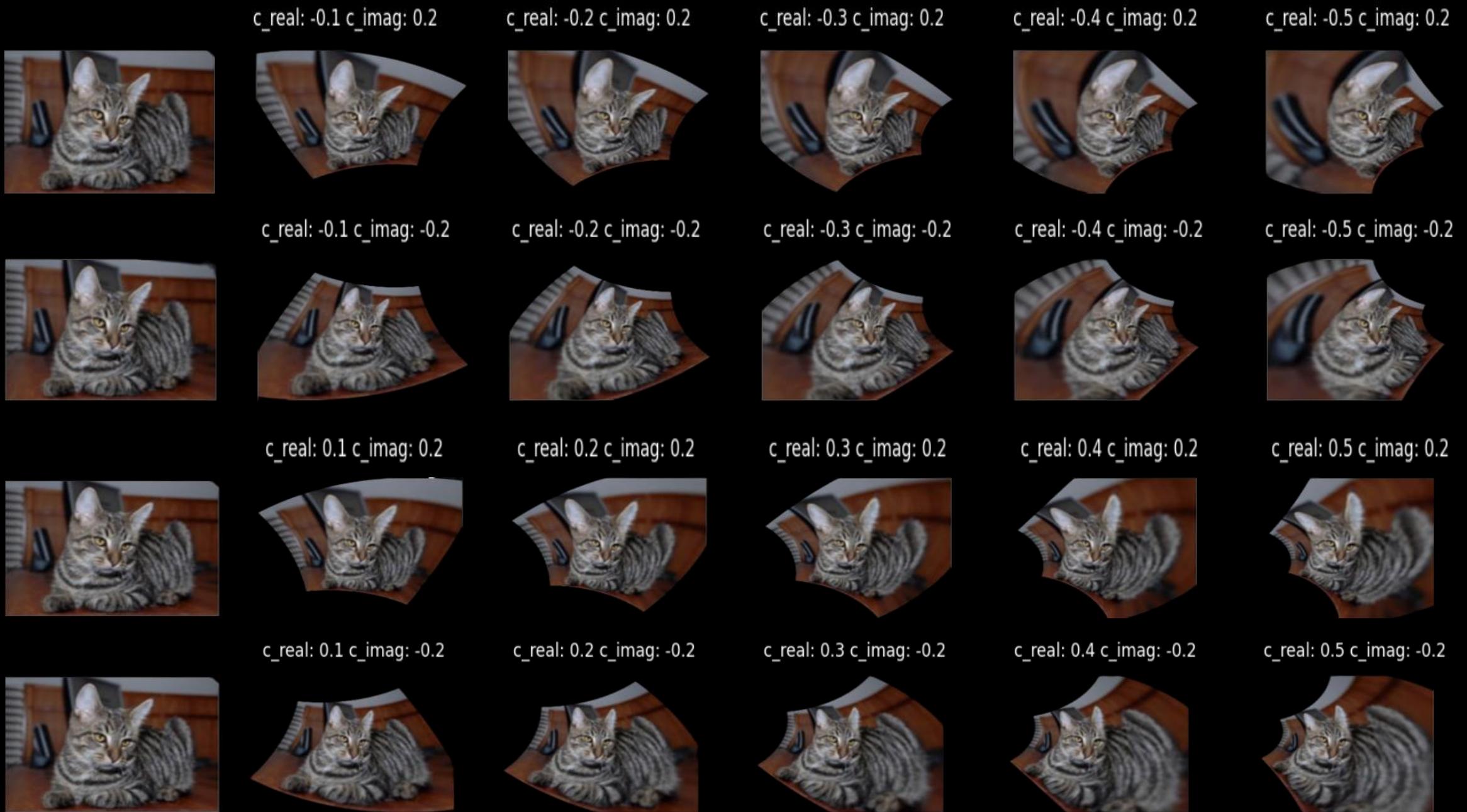
# Determine suitable axis limits for the transformed plot
transformed_x = np.concatenate((transformed_line1.real, transformed_line2.real))
transformed_y = np.concatenate((transformed_line1.imag, transformed_line2.imag))
```



Möbius-based mitigating perspective distortion (MPD)

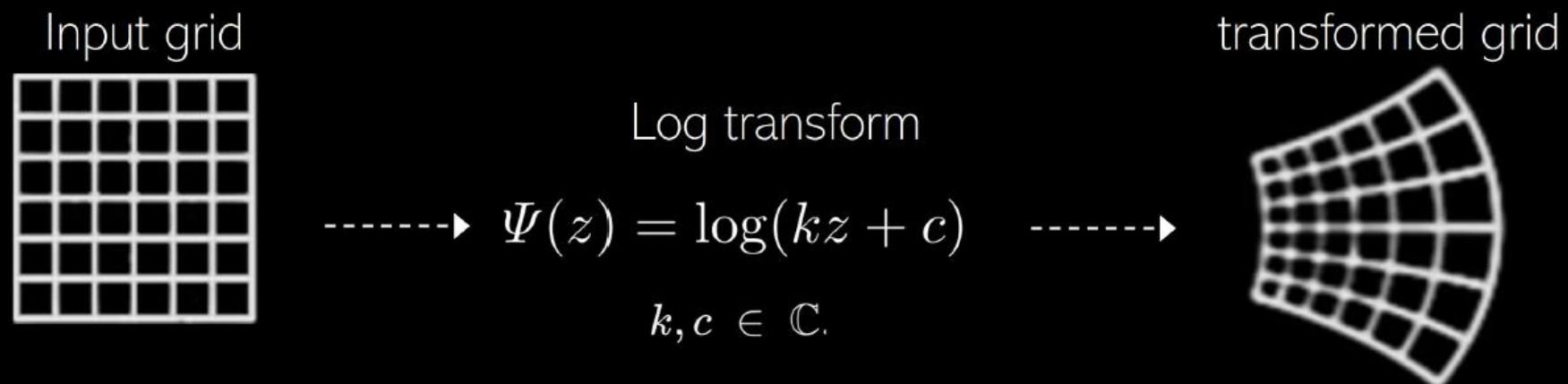
2. Geometry to Vision: pixel to complex space, transform, and back to pixel





Log Conformal Maps (LCM)

1. Modeling: log transform → complex geometry, non-linear, conformal



Conformal transforms - Log

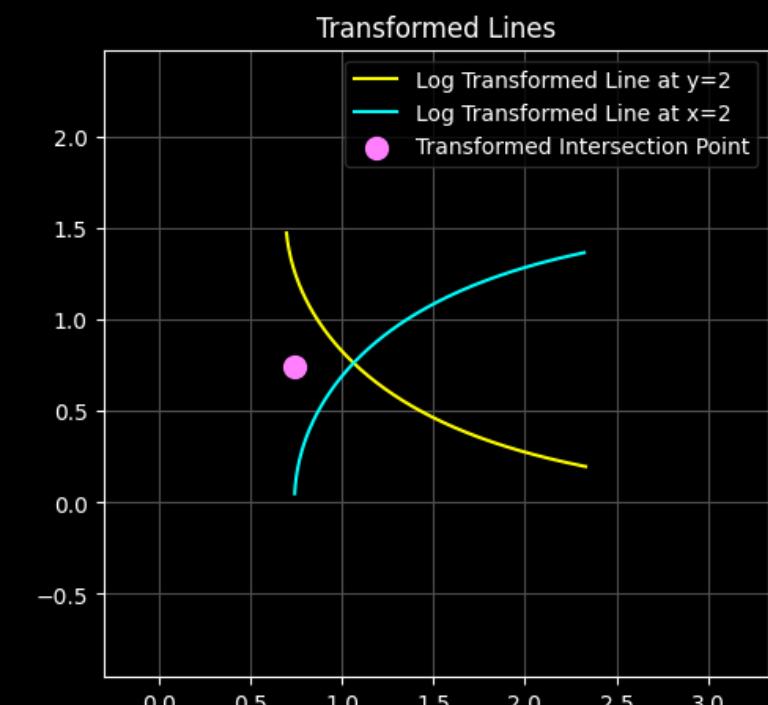
```
def log_conformal_map(z):
    """Applies a logarithmic conformal map to a complex number z."""
    # Avoid issues with zero by ensuring no zero values are passed to log
    return np.log(z + 0.1) # Adding a small constant to avoid log(0)

# Define a range avoiding zero, particularly negative or zero values which are
x = np.linspace(0.1, 10, 400)
line1 = x + 1j * 2 # Horizontal line slightly above the x-axis
line2 = 2 + 1j * x # Vertical line slightly right of the y-axis

# Apply Log Conformal Map to the lines
transformed_line1 = log_conformal_map(line1)
transformed_line2 = log_conformal_map(line2)

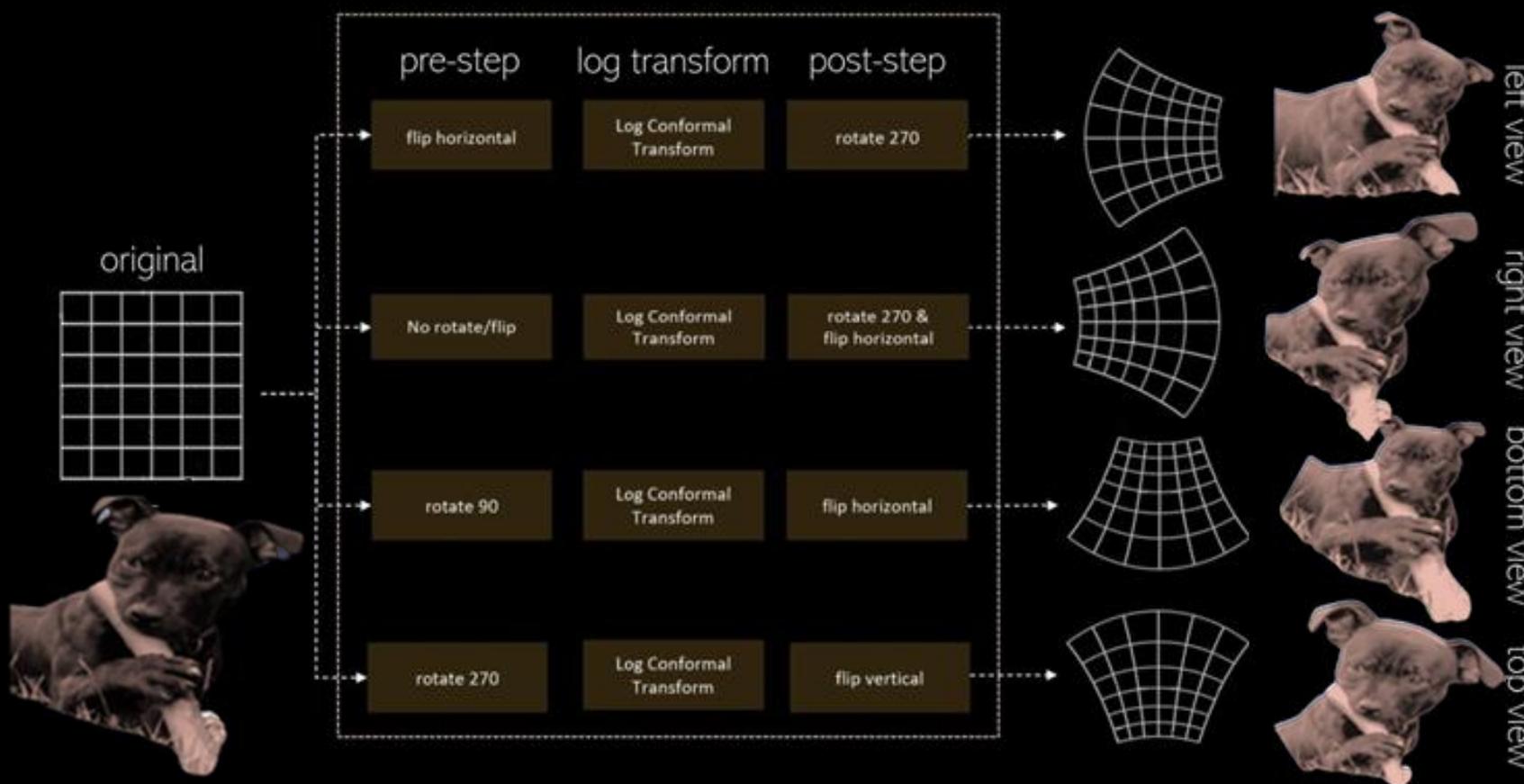
# Plotting the original and transformed lines
fig, ax = plt.subplots(1, 2, figsize=(12, 6), subplot_kw={'aspect': 'equal'})

# Original lines
ax[0].plot(line1.real, line1.imag, 'b-', label='Line at y=2')
ax[0].plot(line2.real, line2.imag, 'r-', label='Line at x=2')
ax[0].set_title('Original Lines')
```

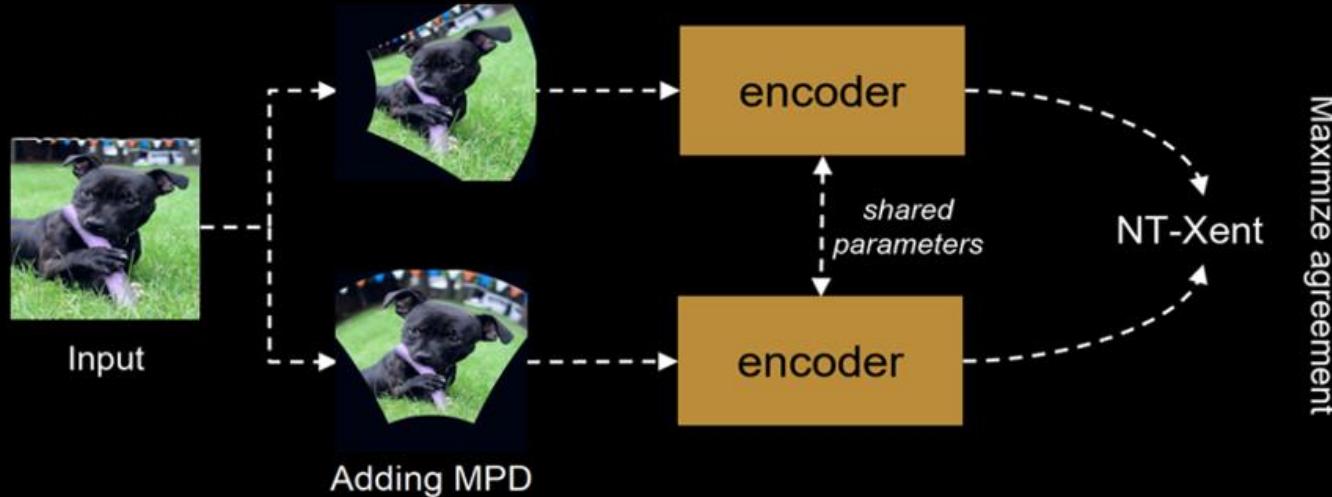


Log Conformal Maps (LCM)

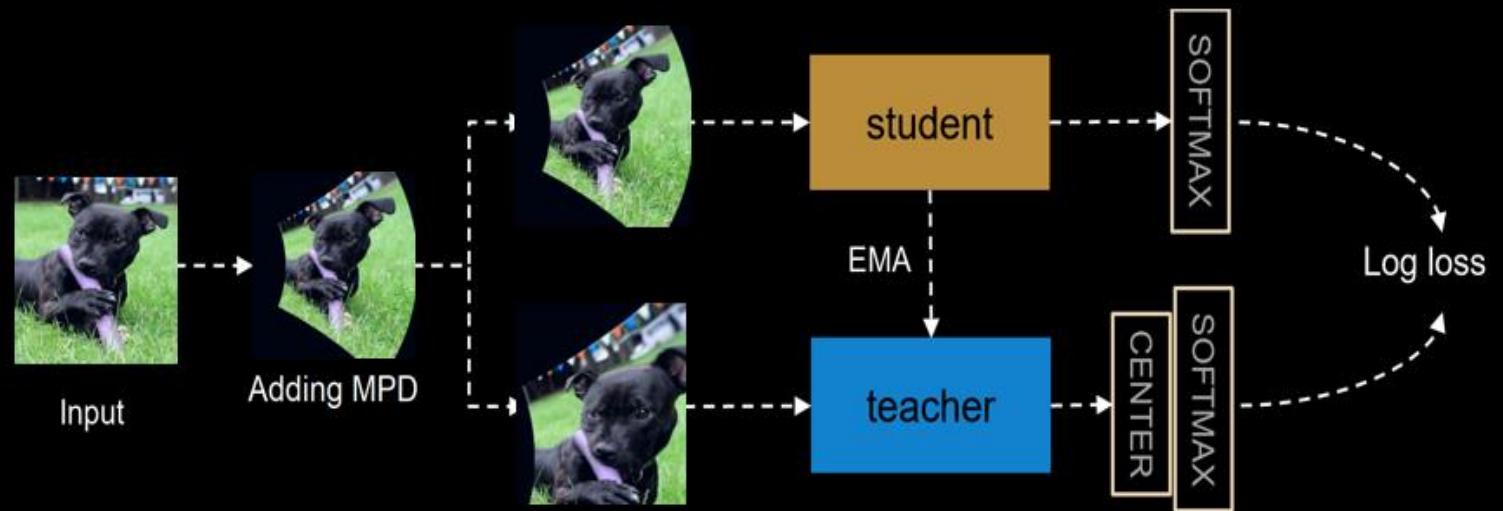
2. Geometry to Vision: pixel to complex space, transform, and back to pixel



SimCLR with MPD/LCM

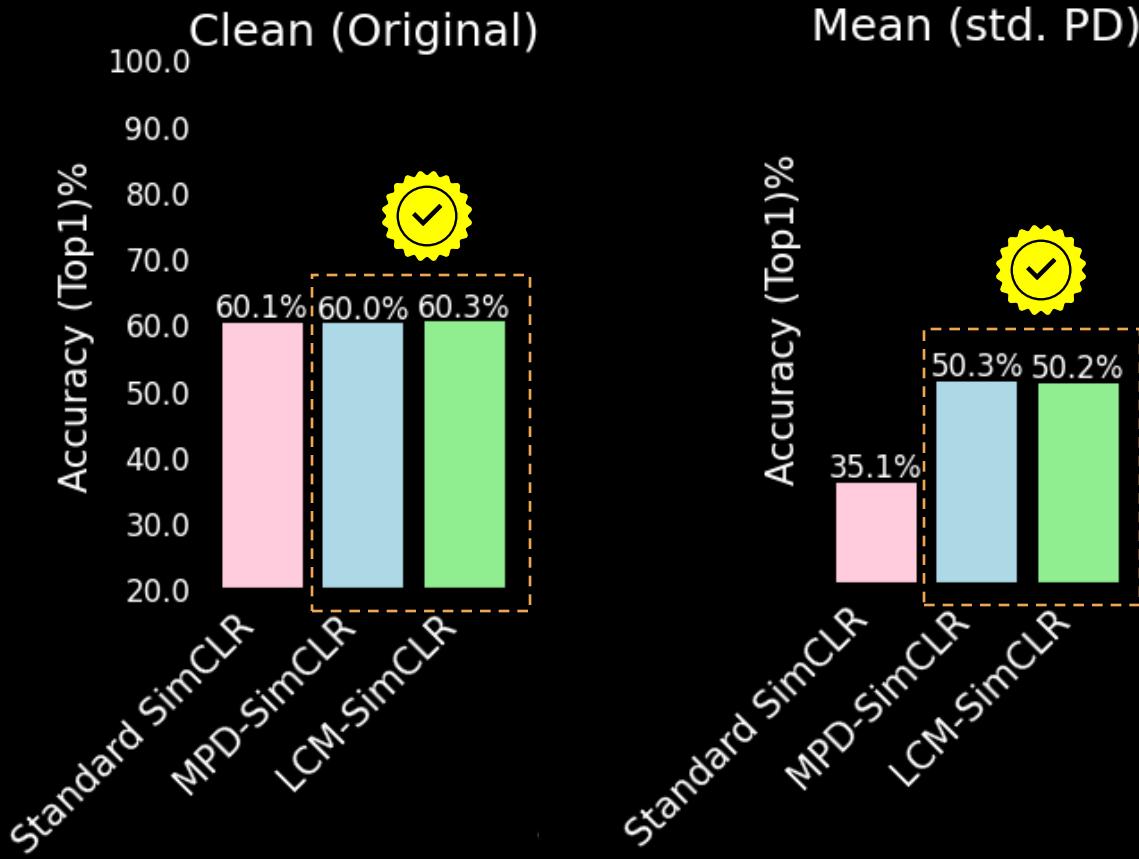


DINO with MPD/LCM

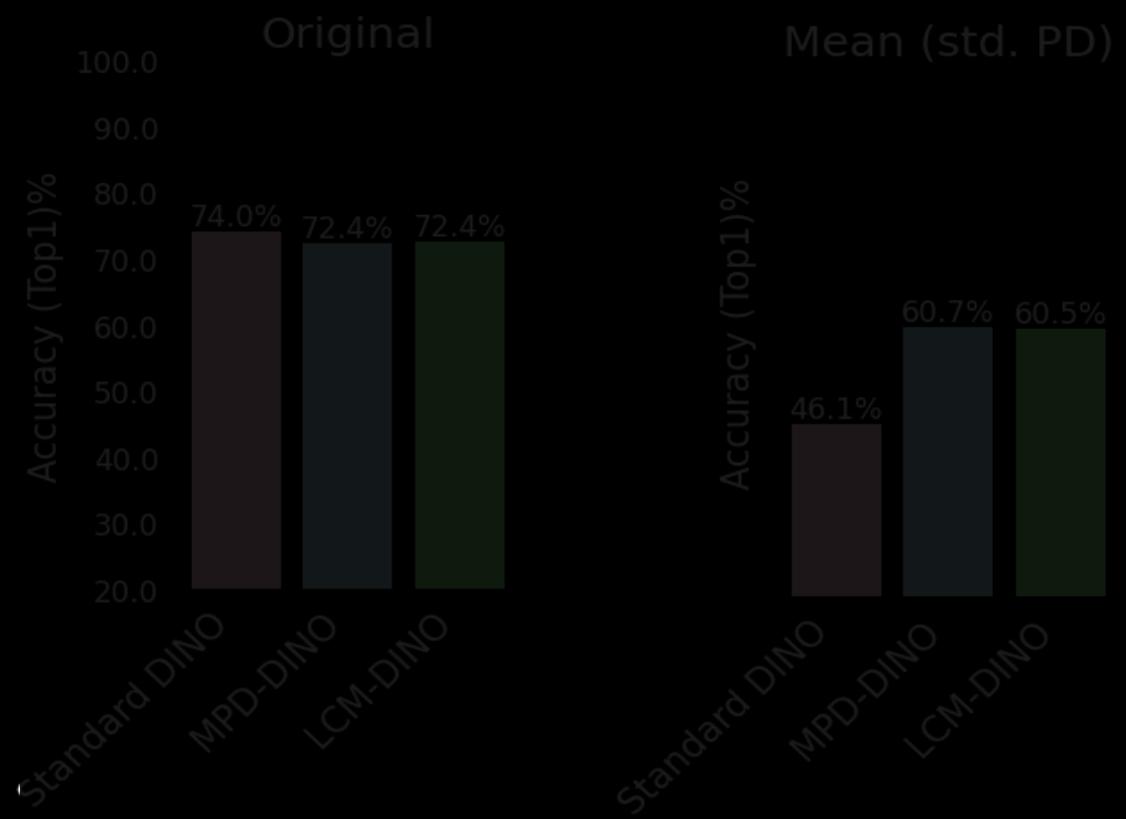


Robustness against PD – linear evaluation

SimCLR on ImageNet-PD



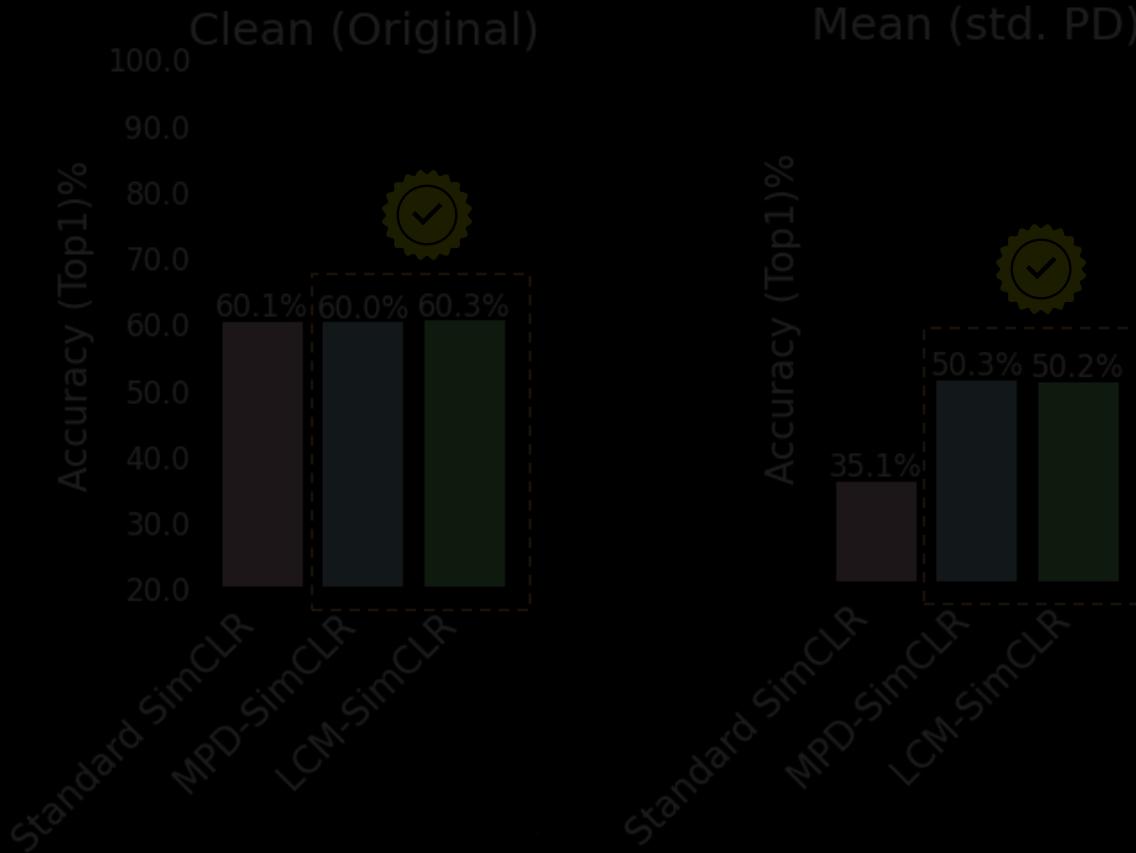
DINO on ImageNet-PD



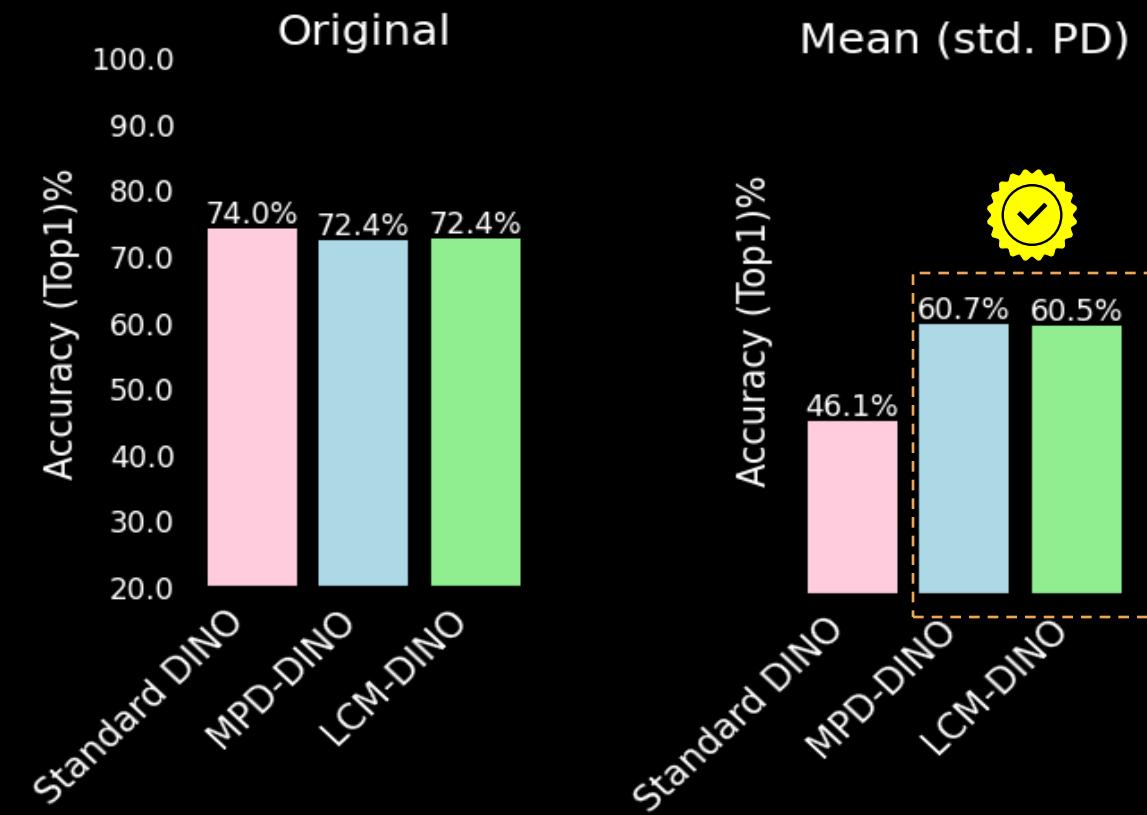
Pretraining following original protocol with batch size 512 on ImageNet dataset

Robustness against PD – linear evaluation

SimCLR on ImageNet-PD



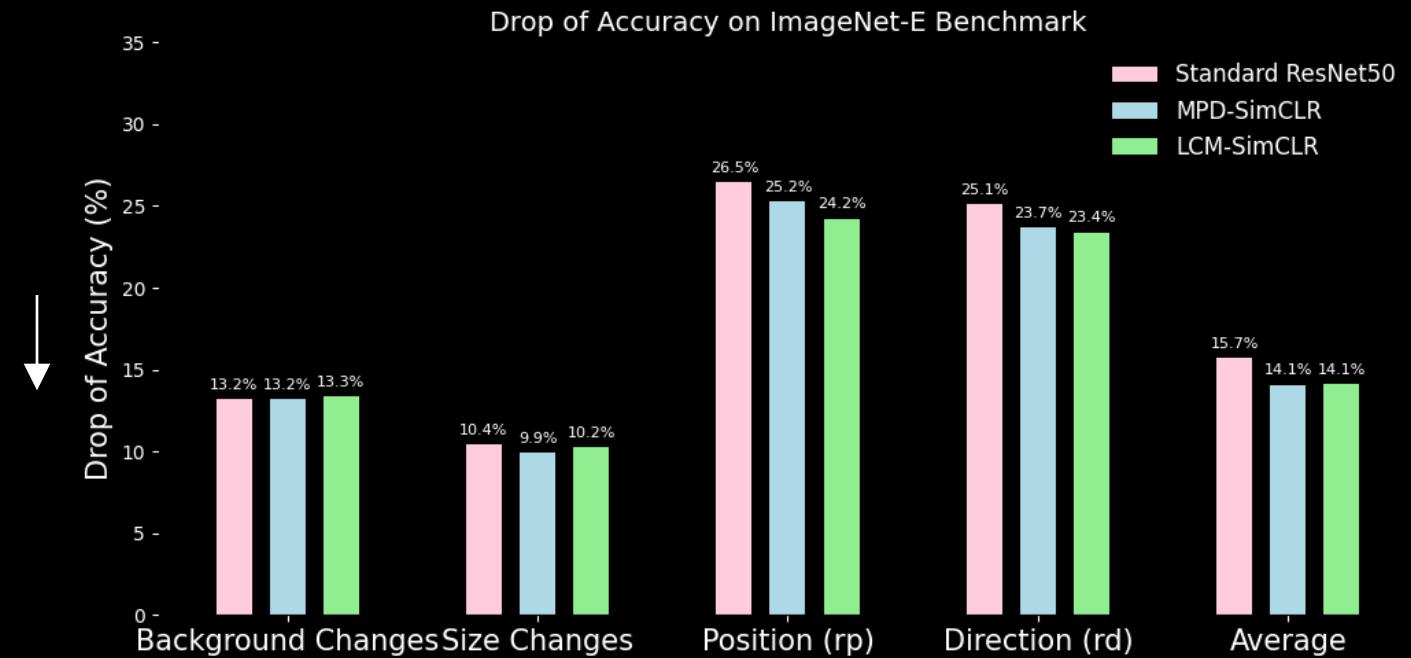
DINO on ImageNet-PD



Pretraining following original protocol with batch size 512 on ImageNet dataset

Robustness against PD – ImageNet-E

ImageNet-E examples – position, direction, and size subsets



Real-world applications with MPD & LCM

Crowd counting



shanghai part A

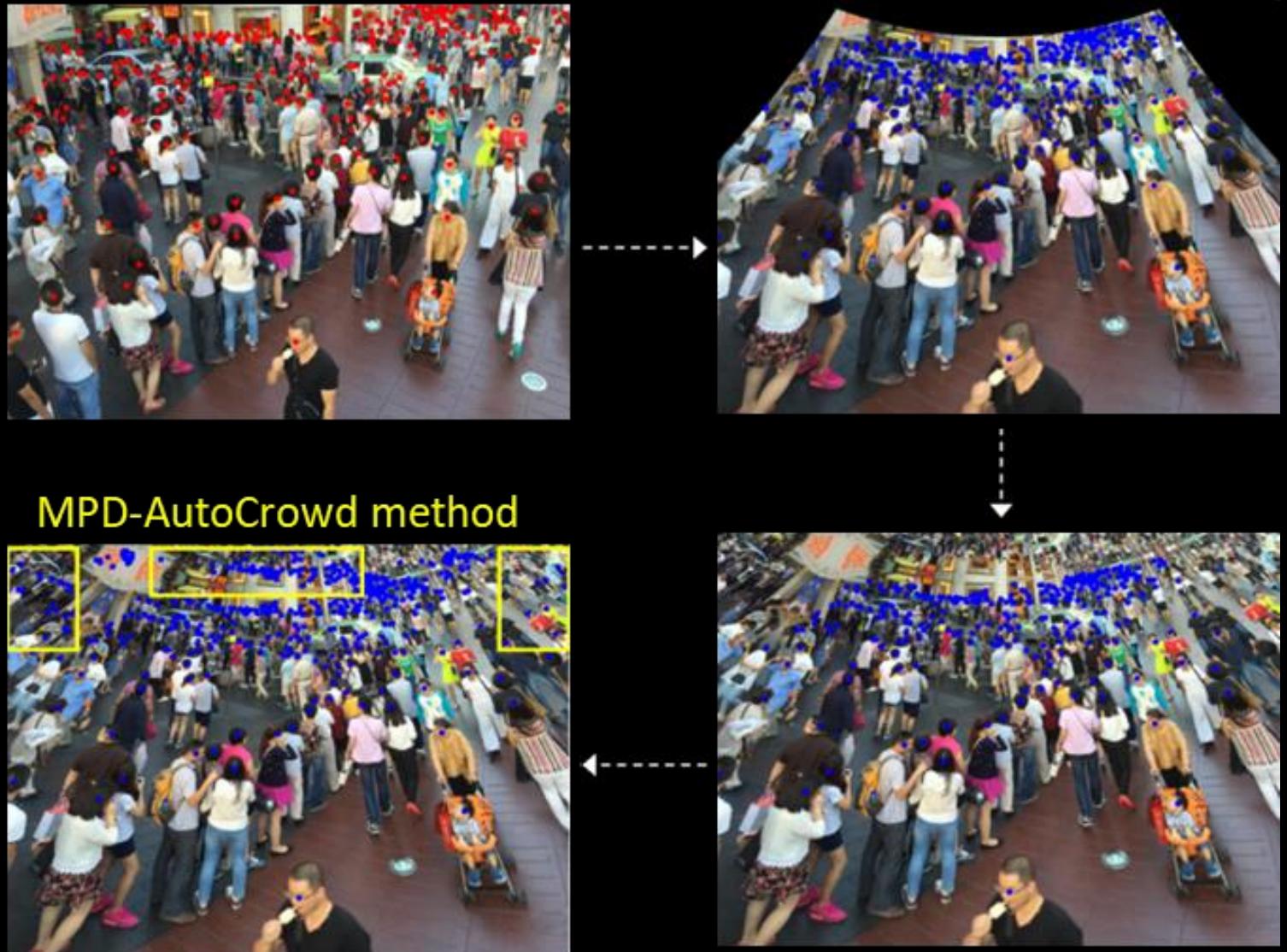


shanghai part B



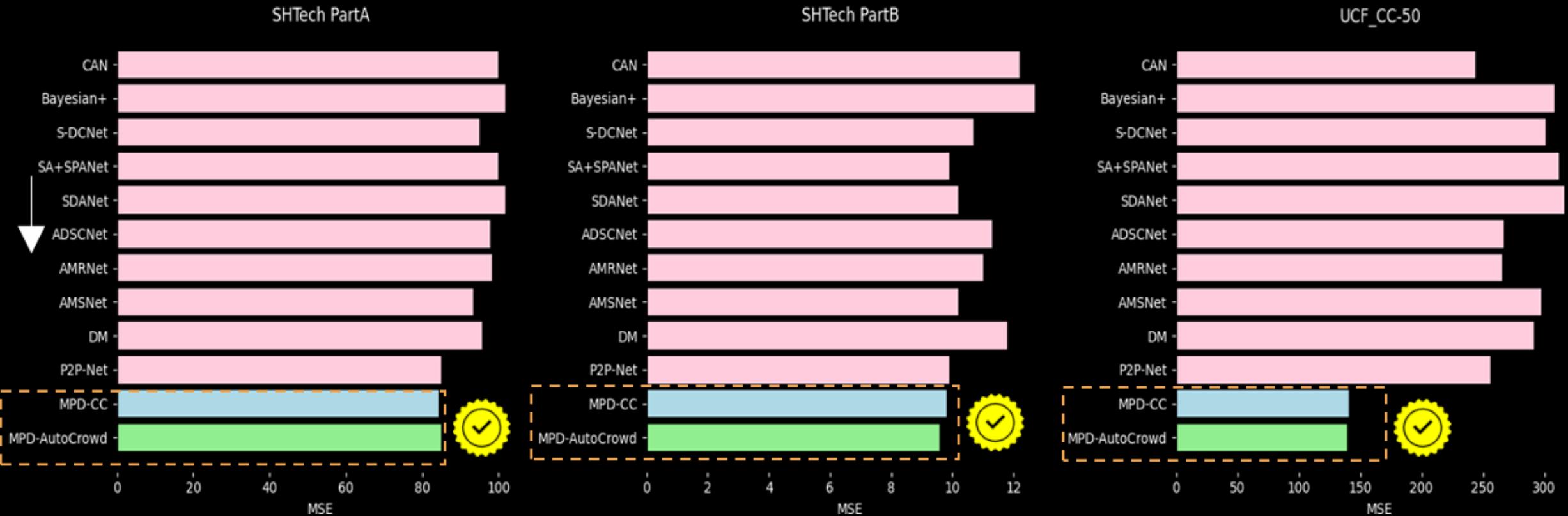
UCF CC 50

Proposed methods



These methods builds on P2P-Net (2021)

Comparative performance analysis

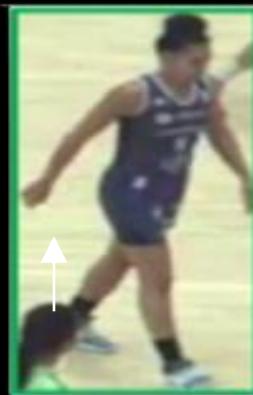
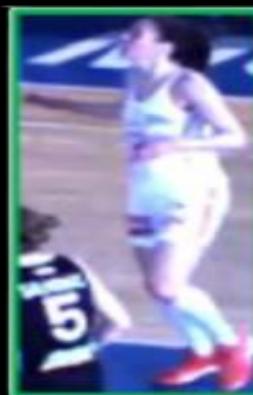


Person re-identification

Query

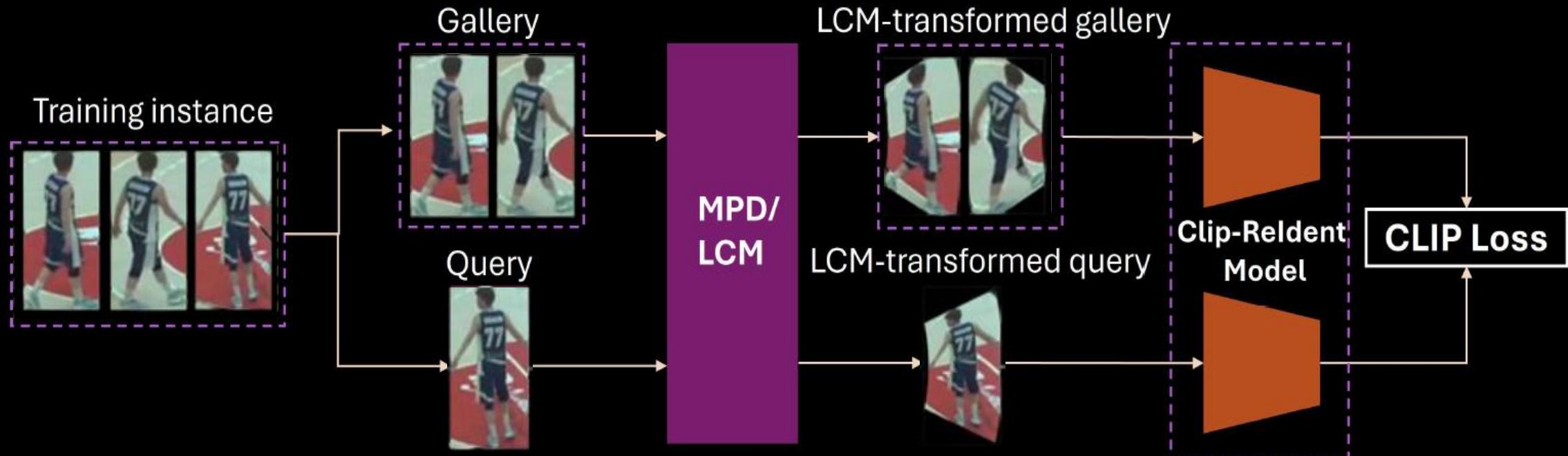


Gallery

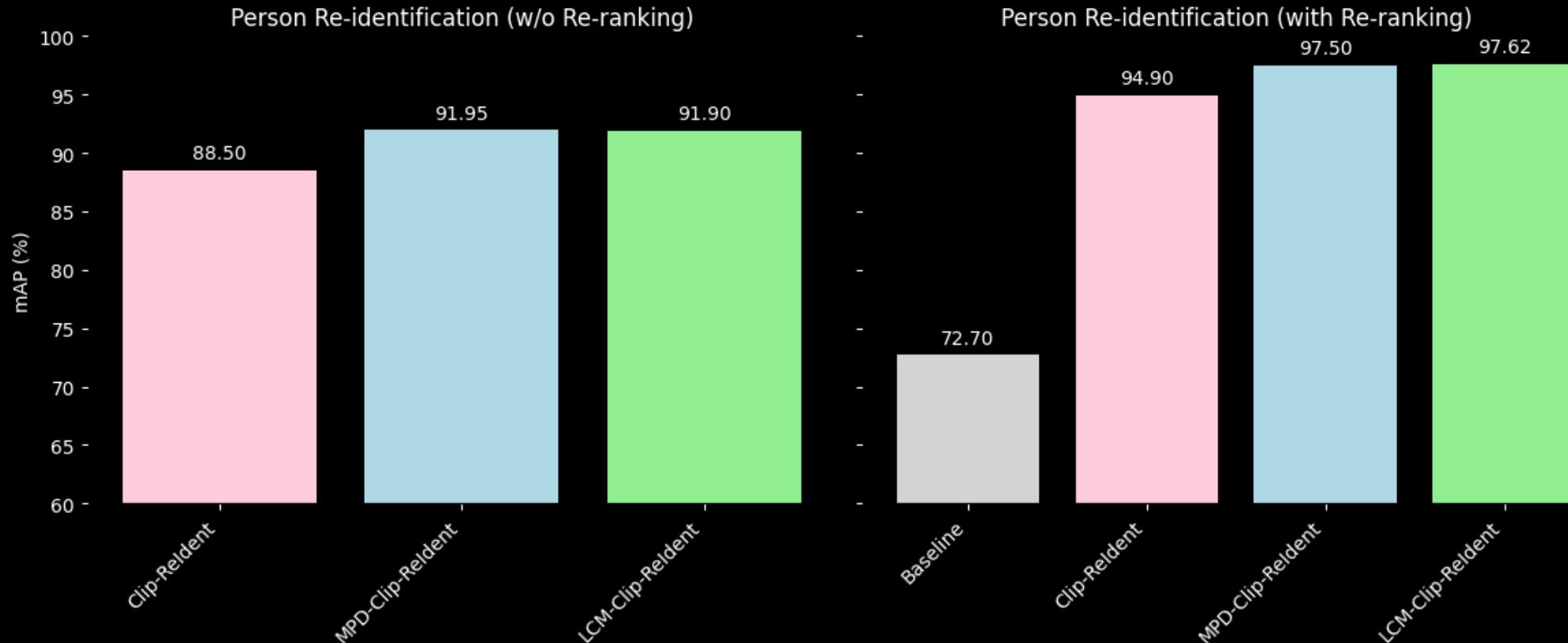


DeepSportRadar dataset

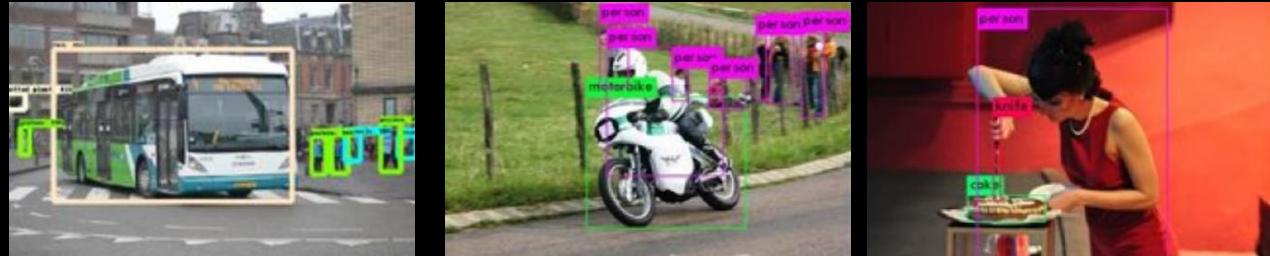
Person re-identification



Person re-identification



Object detection



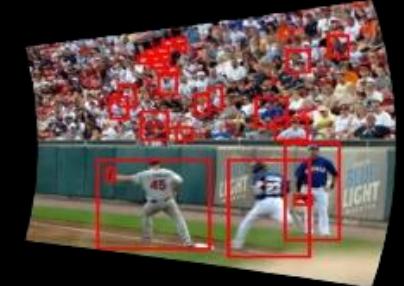
COCO dataset

MPD-OD: transforming boxes along with image

left view



right view



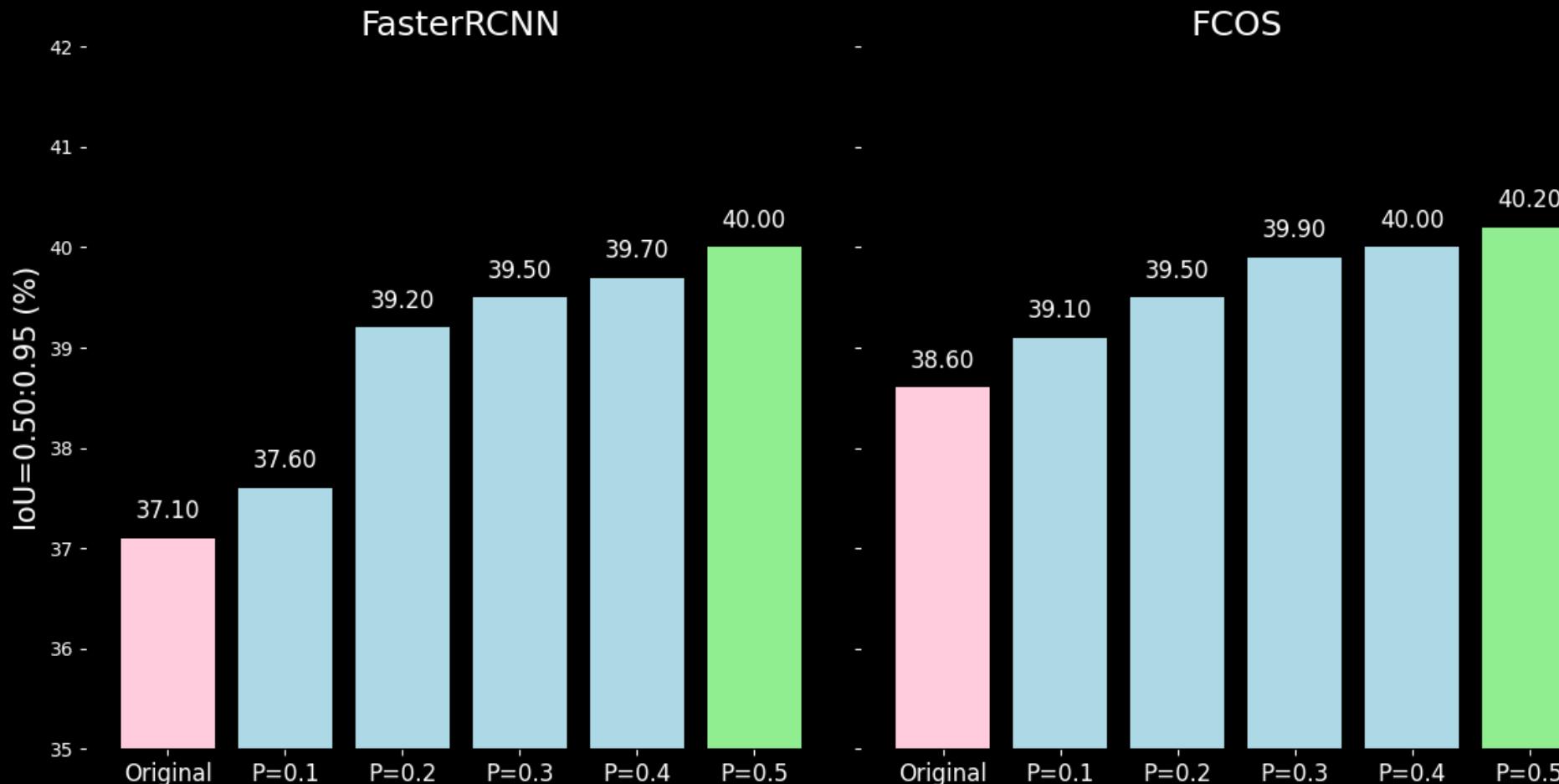
top view



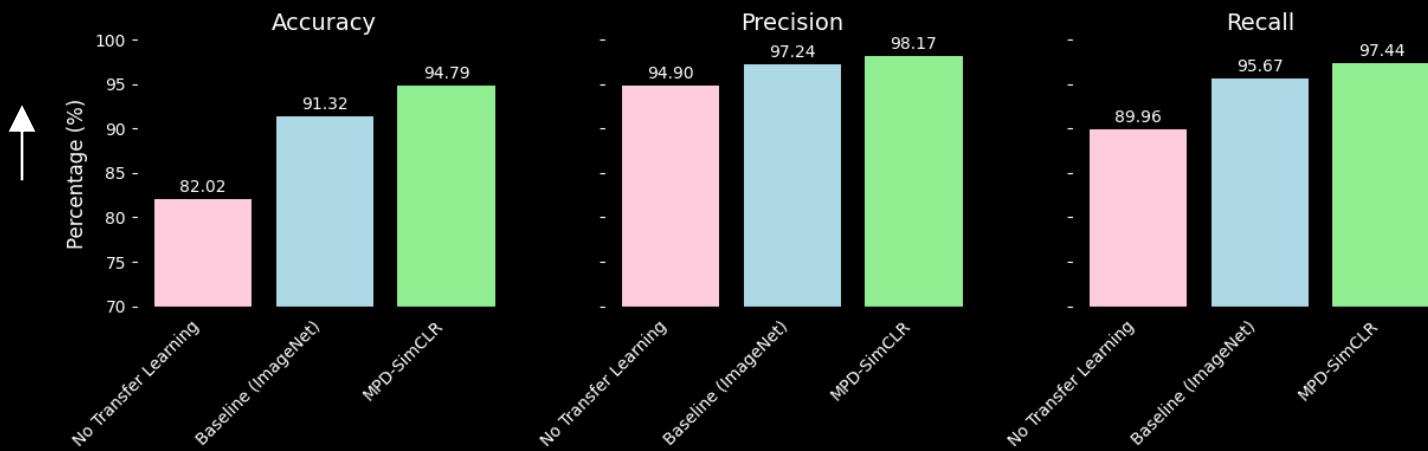
bottom
view



Object detection

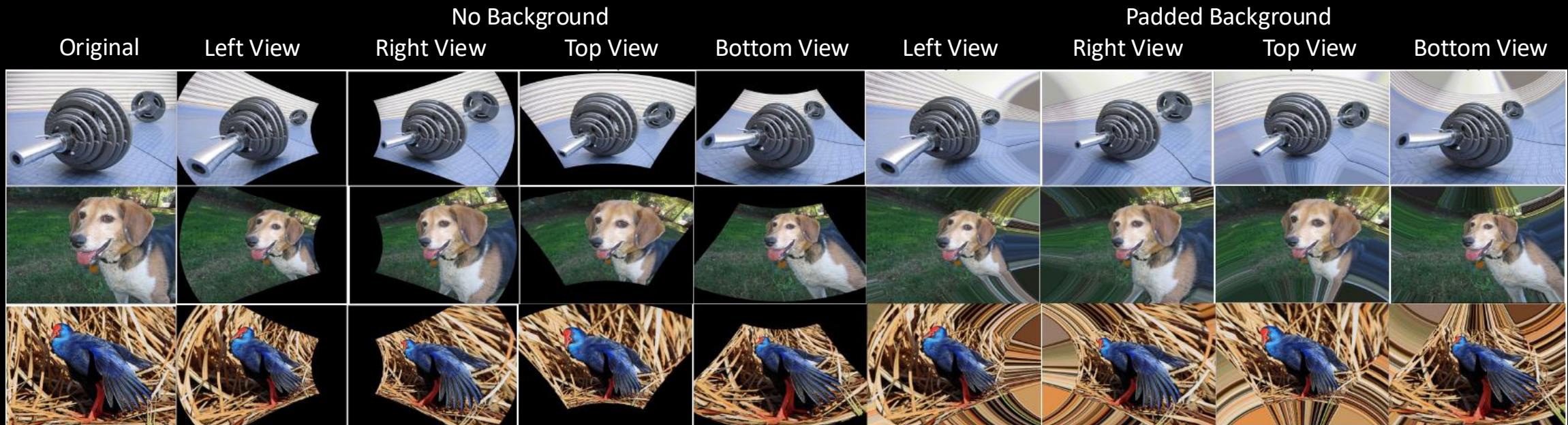


Transfer learning on fisheye view recognition



Public robustness benchmark: ImageNet-PD

ImageNet-PD



- ✓ ImageNet-PD is derived from the full validation set of ImageNet
- ✓ Eight subsets: Four orientation—left, right, top, bottom, and their padded background variants—each consisting of 50,000 images across 1,000 classes.

Are LLMs robust against perspective distortion?

ChatGPT 4 and perspective distortion (on human face)

ChatGPT 4 ▾



wearing same style clothes, is it same person or three different person?



Based on the images you've provided, it seems that these are three different individuals. Although there are similarities, particularly in the style of clothing and the general facial features suggesting a Middle-Eastern or South Asian descent, subtle differences in facial structure and hair suggest they are different people. For example, the facial hair styles, shapes of the head and faces, and the hairstyles vary among the images.



ChatGPT 4 and perspective distortion with colors effects - cat

ChatGPT 4 ▾



are they same cats or not?

The two images you've uploaded show different cats. The first cat has a more prominent gray color and a deeper, almost blue tint, while the second cat features a lighter gray with subtle brown tones. Additionally, the fur patterns differ slightly, and the facial features, such as the eyes and the shape of the ears, are not identical. These details suggest that they are not the same cat.

ChatGPT o3-mini-high and perspective distortion on cat

ChatGPT o3-mini-high



are these similar cats or same cat?

Reasoned about similarity of cats for 7 seconds

It's hard to be absolutely certain just from a photo, but these ginger tabbies look extremely similar in color, markings, and overall appearance—enough that they could be the same cat photographed in different poses. The coat pattern and facial features are consistent across all five images, suggesting they might be pictures of one cat rather than different individuals. However, without additional details (like very close comparisons of stripe patterns or information about how the photos were taken), we can't be 100% certain. They do appear very much alike.

ChatGPT 4o and perspective distortion and colors on two sets of cat

ChatGPT 4o



What do you understand about this visual? are they same cat or different ones?

This image shows **five different cats**, each with distinct features and appearances. Here's what stands out:

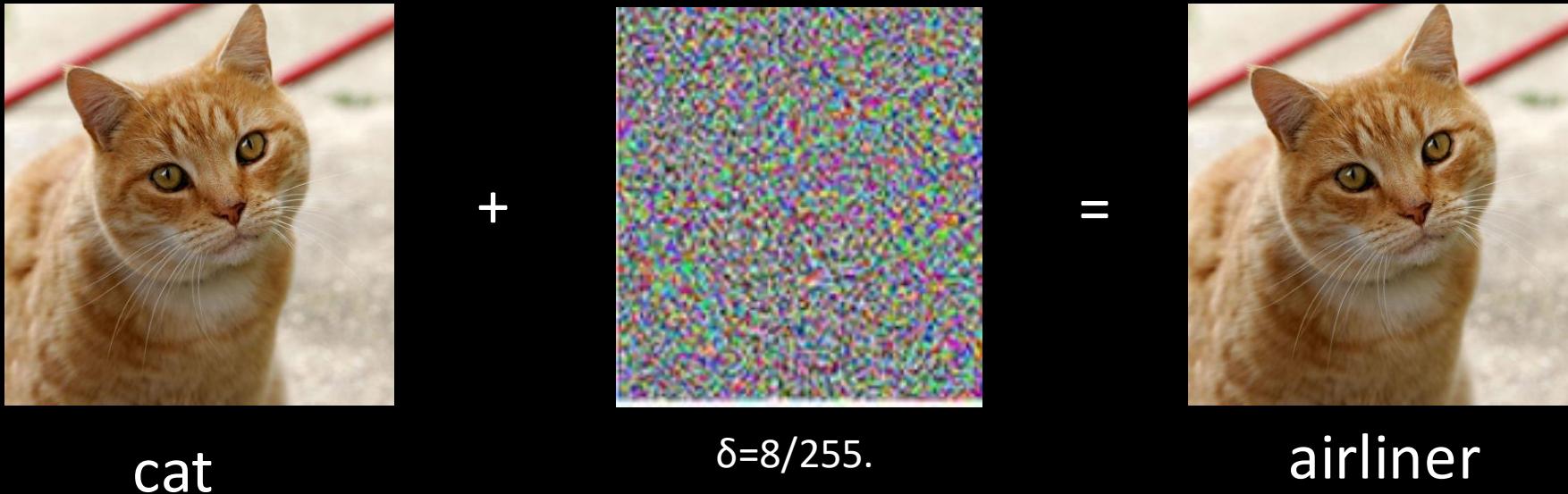
1. **Top-left cat** – Gray and white, almost grayscale in appearance.
2. **Top-center cat** – Unusual blue-toned fur with dark stripes. This coloration is not natural and seems artificially edited.
3. **Top-right cat** – Orange tabby, sitting in a relaxed upright pose.
4. **Bottom-left cat** – Another orange tabby, but smaller and with a more curious head tilt.
5. **Bottom-center cat** – Gray tabby, lying down with a different fur pattern than the others.

Even though two of the cats are orange tabbies and two are gray tabbies, their **poses, fur patterns, sizes, and color tones suggest they are different individual cats**. The blue-toned one especially appears digitally altered, reinforcing that this is a composite image, not a set of photos of the same cat in different styles.

Would you like help identifying if some are edited versions of the same image, or are you exploring something else with this visual?

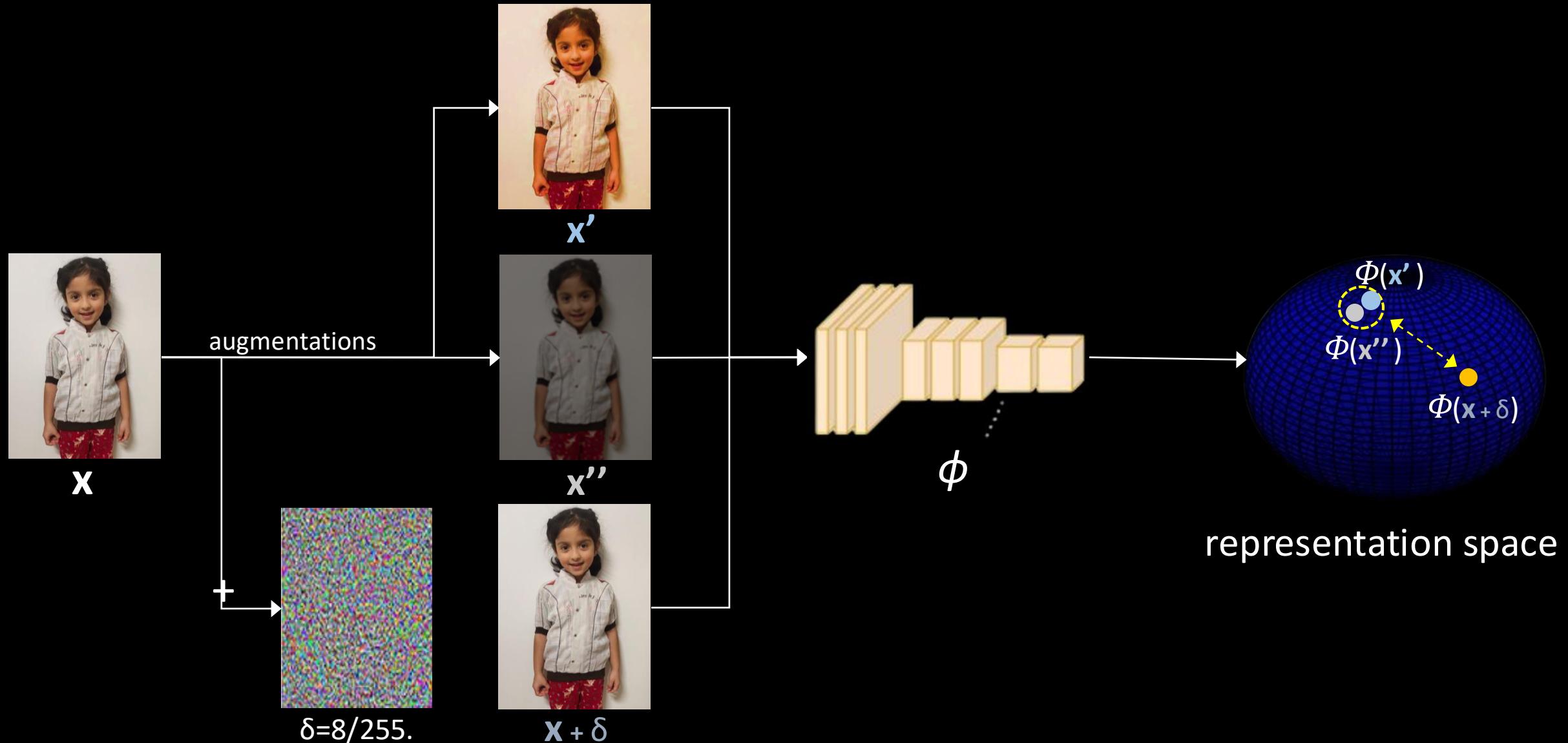
Robustness against adversarial attacks

Adversarial attacks

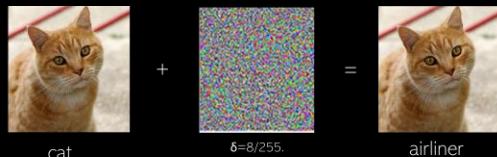


Adversarial attacks are small, carefully crafted perturbations added to input data that cause a model to make incorrect predictions

Representation learning perspective

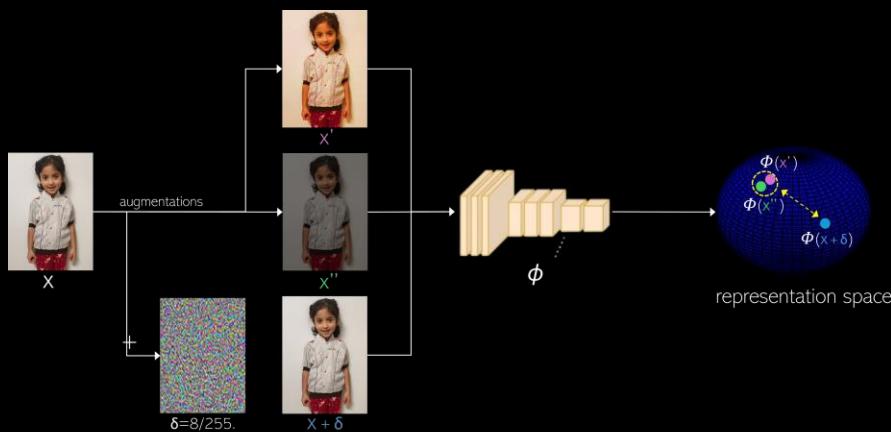


Projected gradient descent (PGD)



In adversarial attacks, small invisible changes can cause a model to wrong prediction.

Representation learning perspective

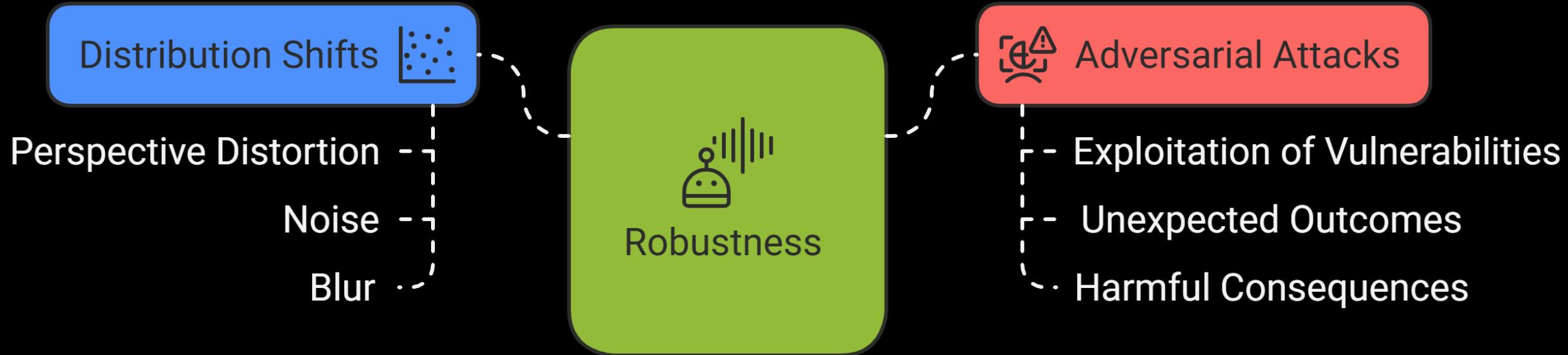


PGD is robust, gradient-based mechanism for approximating worst-case perturbations in ℓ_p -bounded neighborhoods.

- Iteratively adds perturbations that increase the model's loss (gradient ascent).
- Keeps perturbations bounded within a small radius (ϵ -ball) via projection.
- Finds worst-case input that the model is most vulnerable to—used to train for robustness.

Transition from robustness against distribution shifts to adversarial robustness

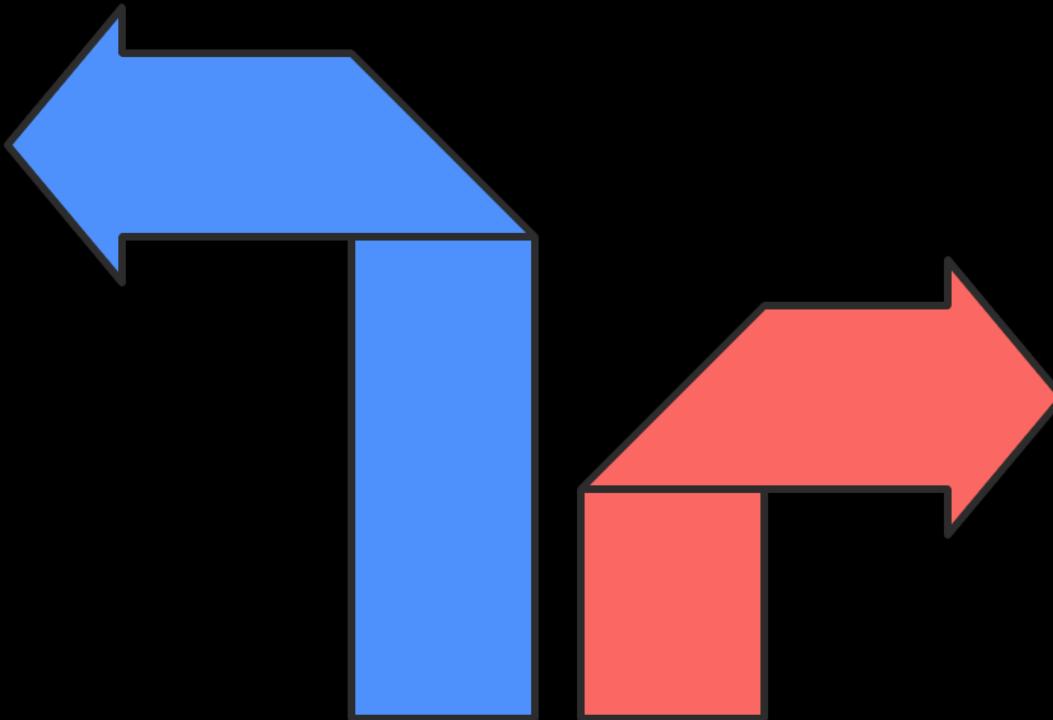
Nature of perturbation



Perturbation distribution

Distribution Shifts

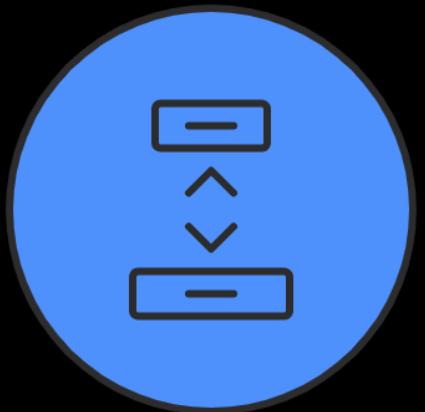
Perturbations follow a structured distribution, allowing for predictable modeling.



Adversarial Attacks

Perturbations are adaptive and unpredictable, requiring dynamic defense strategies.

Effect on representations



Distribution Shifts

Smooth representation changes



Adversarial Attacks

Disrupts representation space in target direction

Invariance mechanism

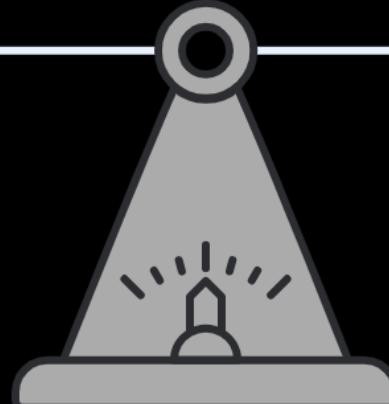
Augmentations
and
Transformations



Adversarial
Training
Methods



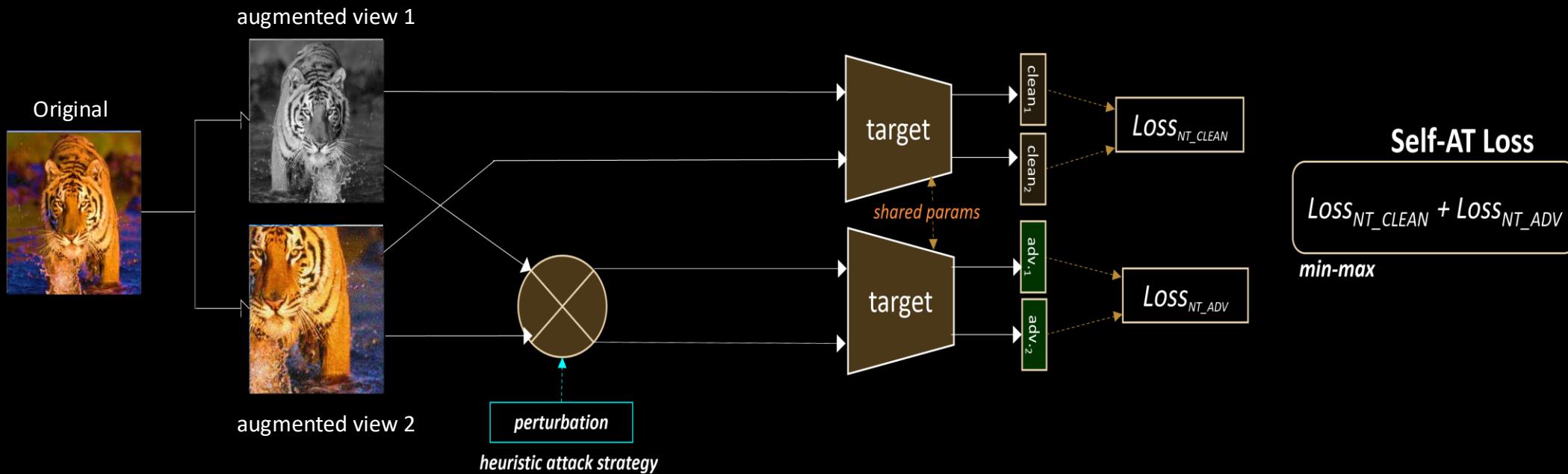
Distribution Shifts



Adversarial Attacks

Self-supervised adversarial training (Self-AT)

Existing self-supervised adversarial training methods—ACL (2020), RoCL (2020), AdvCL [2021], DeACL [2022], DYNACL [218]



PGD: Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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.

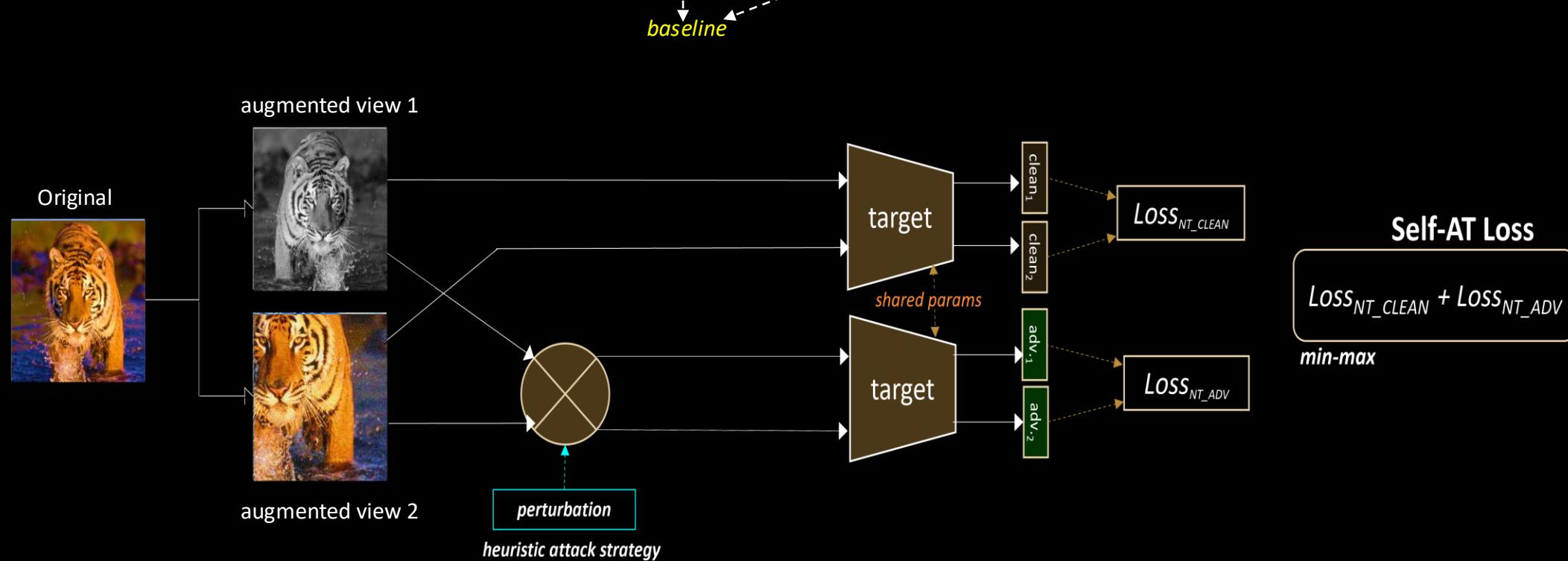
AdvCL: Lijie Fan et al. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in neural information processing systems*, 34:21480–21492, 2021.

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.

Self-supervised adversarial training (Self-AT)

Existing self-supervised adversarial training methods—ACL (2020), RoCL (2020), AdvCL [2021], DeACL [2022], DYNACL [218]



PGD: Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018.

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.

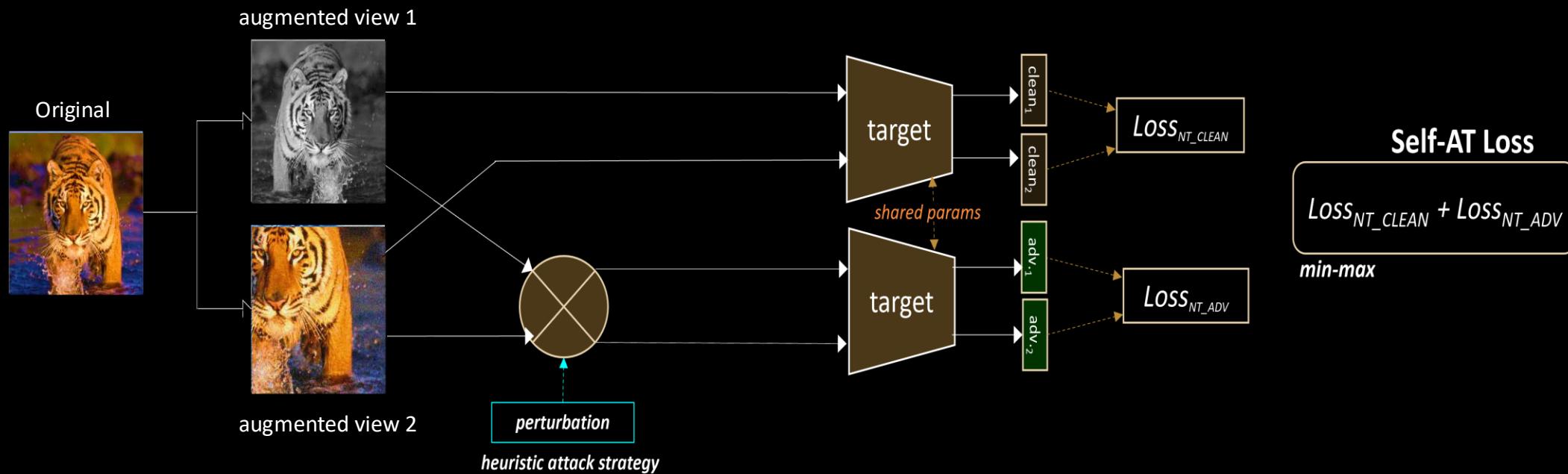
DeACL: Chaoning Zhang et al. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In European Conference on Computer Vision, pages 725–742. Springer, 2022.

DYNACL: Rundong Luo et al. Rethinking the effect of data augmentation in adversarial contrastive learning. In International Conference on Learning Representations, 2023.

Self-supervised adversarial training (Self-AT)

Existing self-supervised adversarial training methods—ACL (2020), RoCL (2020), AdvCL [2021], DeACL [2022], DYNACL [218]

baseline distillation approach



PGD: Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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. 97

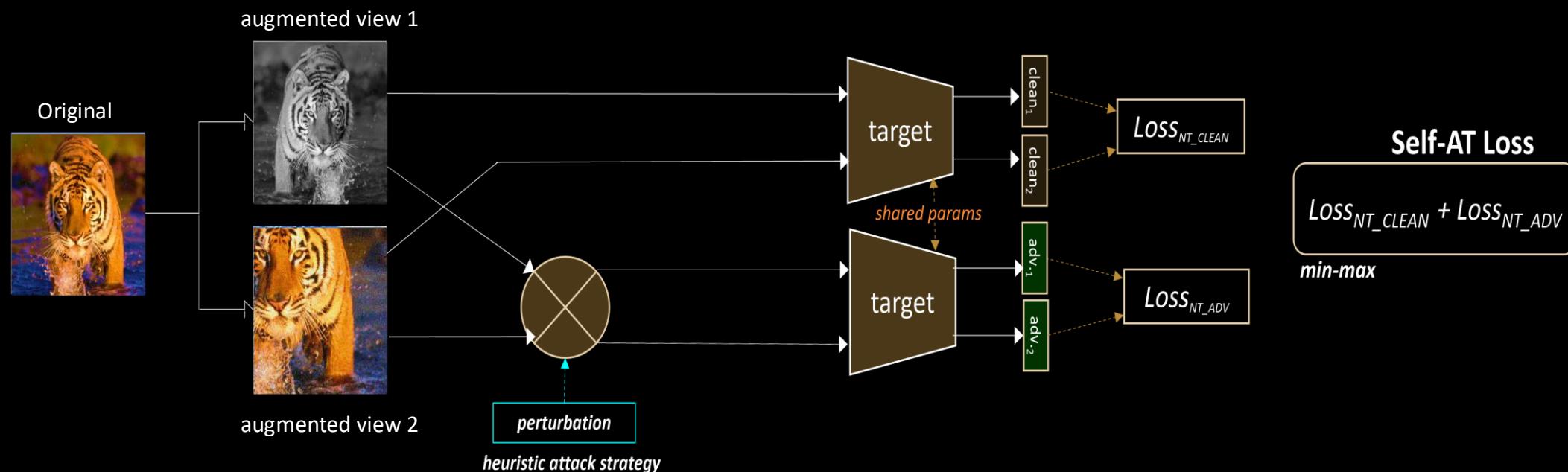
DeACL: Chaoning Zhang et al. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In *European Conference on Computer Vision*, pages 725–742. Springer, 2022.

DYNACL: Rundong Luo et al. Rethinking the effect of data augmentation in adversarial contrastive learning. In *International Conference on Learning Representations*, 2023.

Self-supervised adversarial training (Self-AT)

Existing self-supervised adversarial training methods—ACL (2020), RoCL (2020), AdvCL [2021], DeACL [2022], DYNACL [218]

decoupled approach – 3 stage
↓
baseline ↓
distillation approach



PGD: Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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.

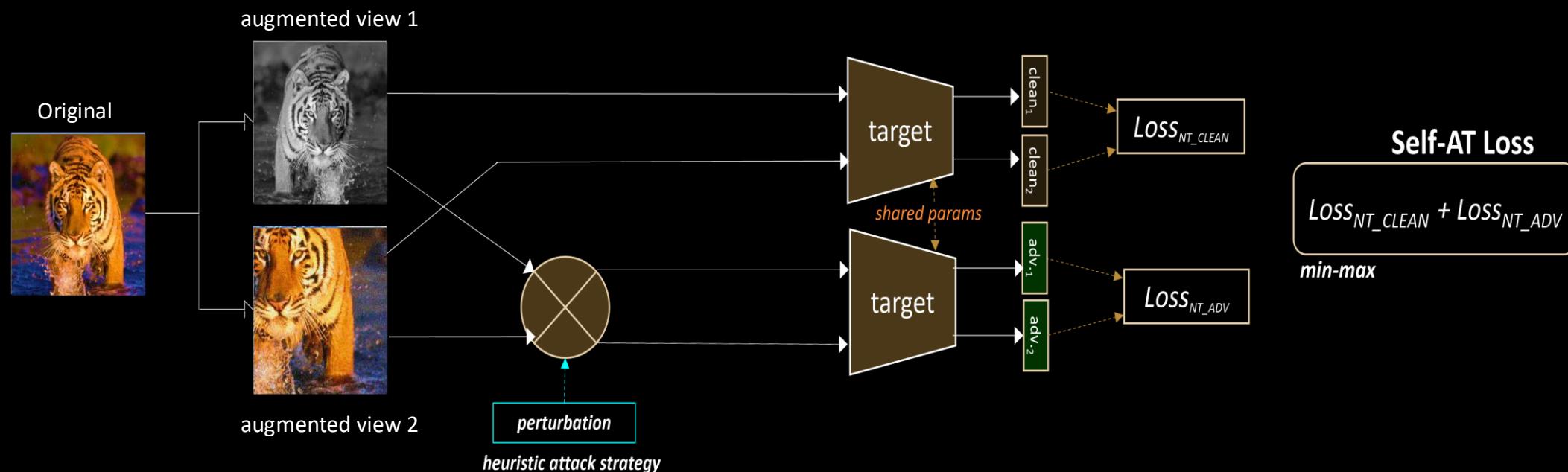
DeACL: Chaoning Zhang et al. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In *European Conference on Computer Vision*, pages 725–742. Springer, 2022.

DYNACL: Rundong Luo et al. Rethinking the effect of data augmentation in adversarial contrastive learning. In *International Conference on Learning Representations*, 2023.

Self-supervised adversarial training (Self-AT)

Existing self-supervised adversarial training methods—ACL (2020), RoCL (2020), AdvCL [2021], DeACL [2022], DYNACL [218]

decoupled approach – 3 stage
↓
baseline ↓
distillation approach ↓
dynamic augmentations



PGD: Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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.

Research gap



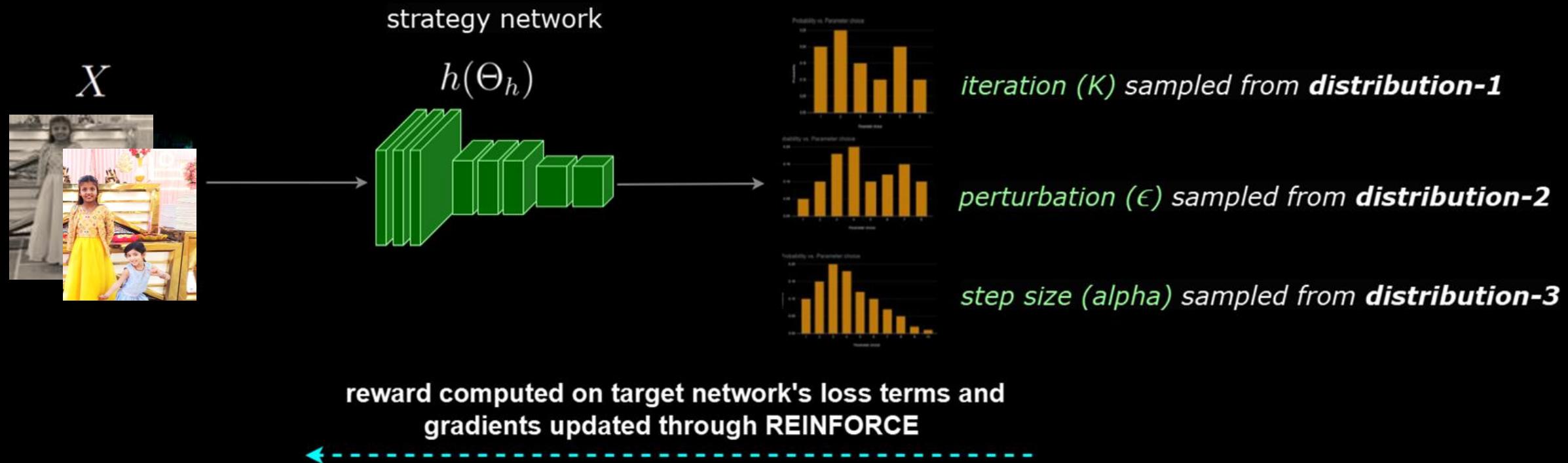
Research gap



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.

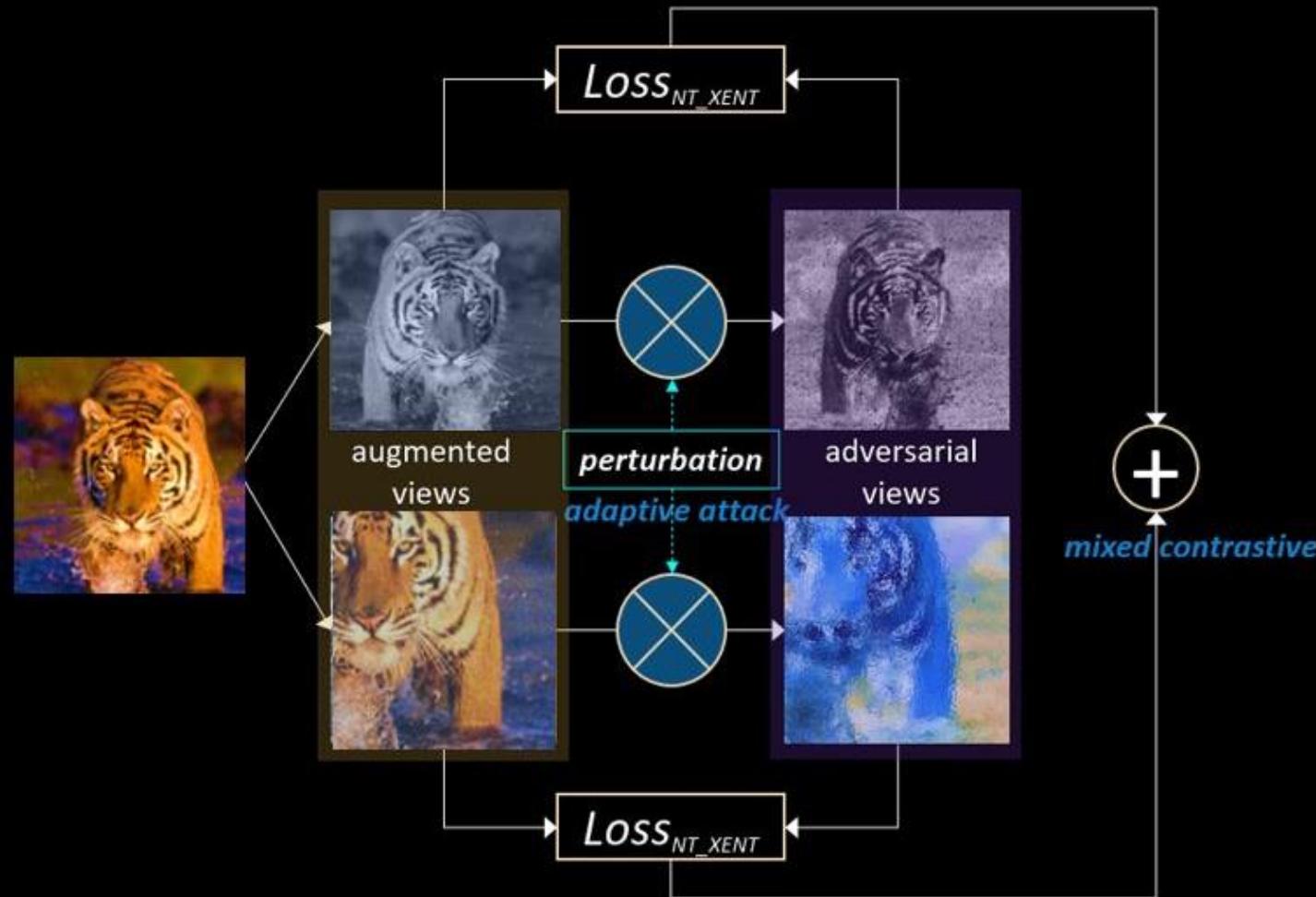
Learnable attacks

- ✓ Learnable strategy network autonomously finds optimal attacks.



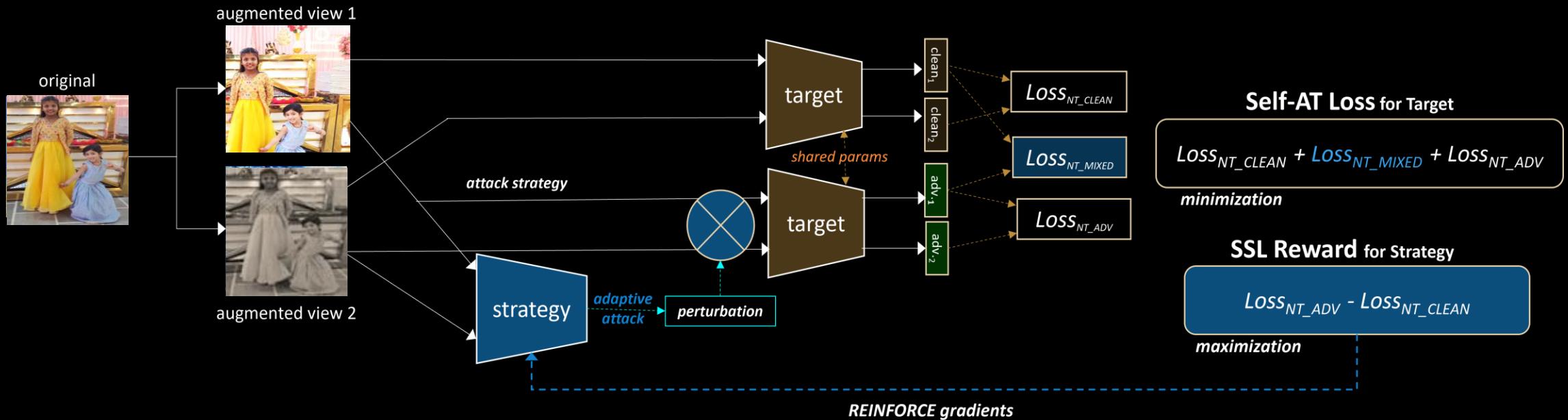
Mixed contrastive objective

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

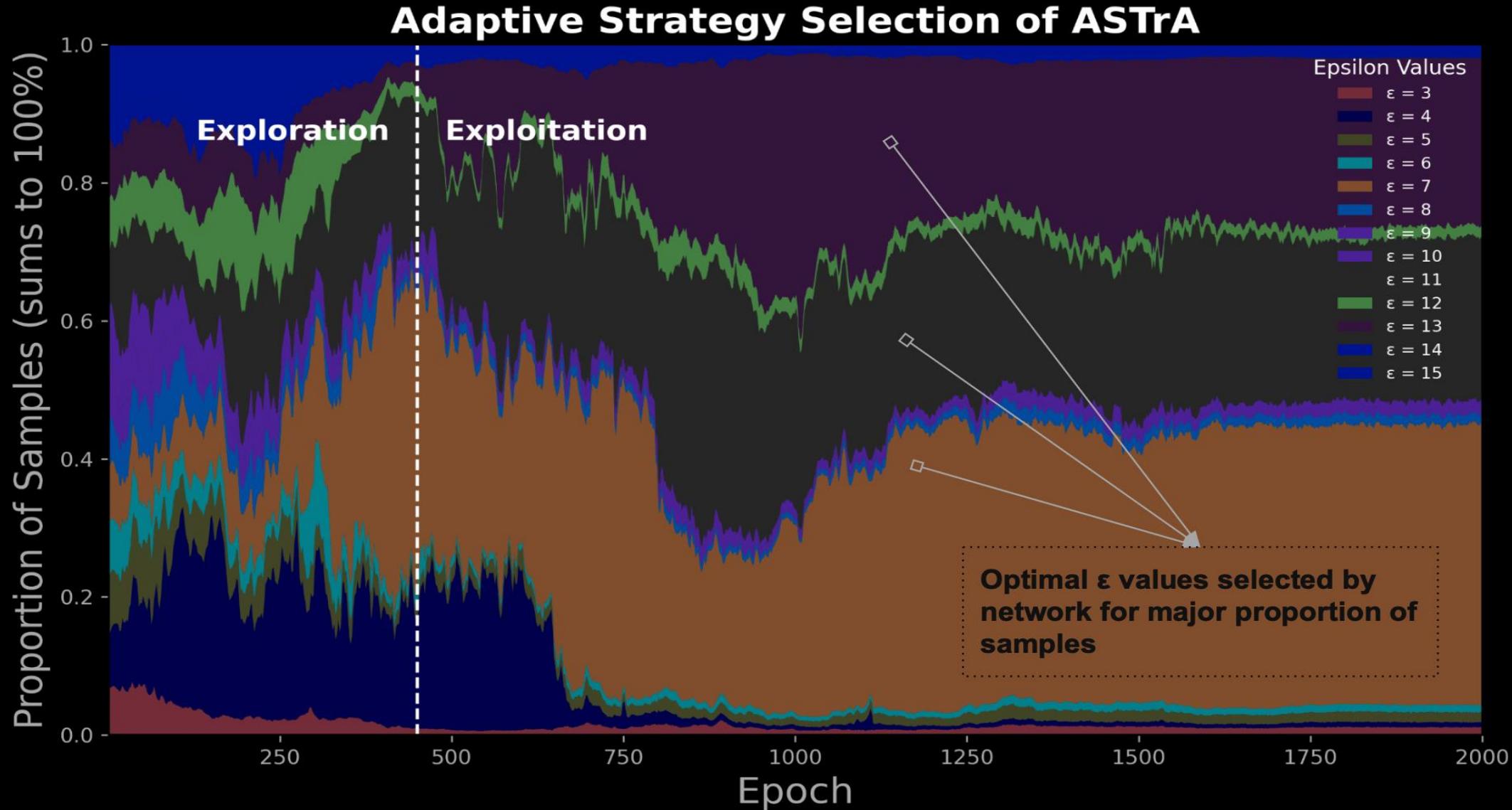


ASTrA – complete framework

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

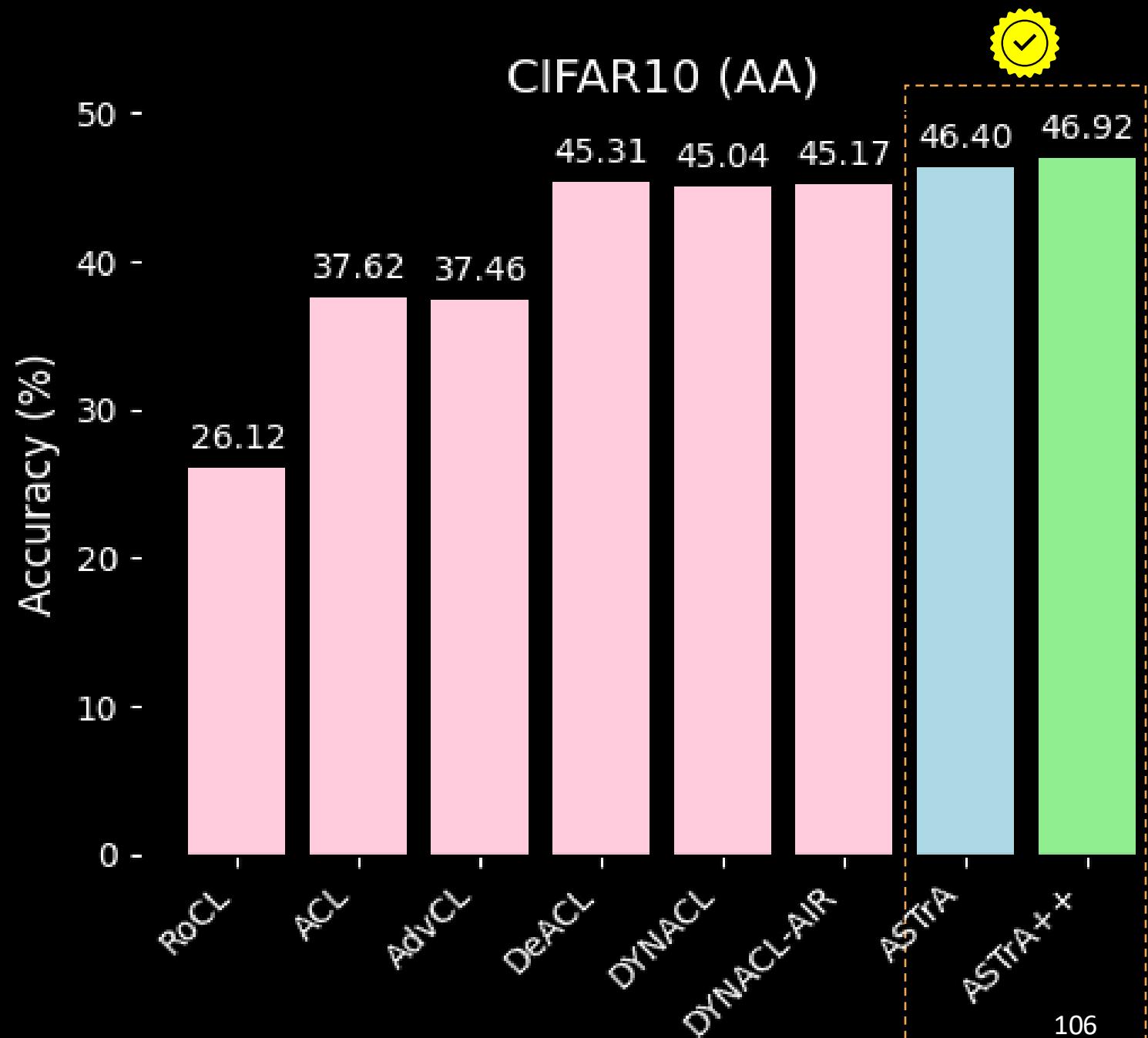


Exploration - exploitation



ASTrA – robustness performance

standard linear finetuning

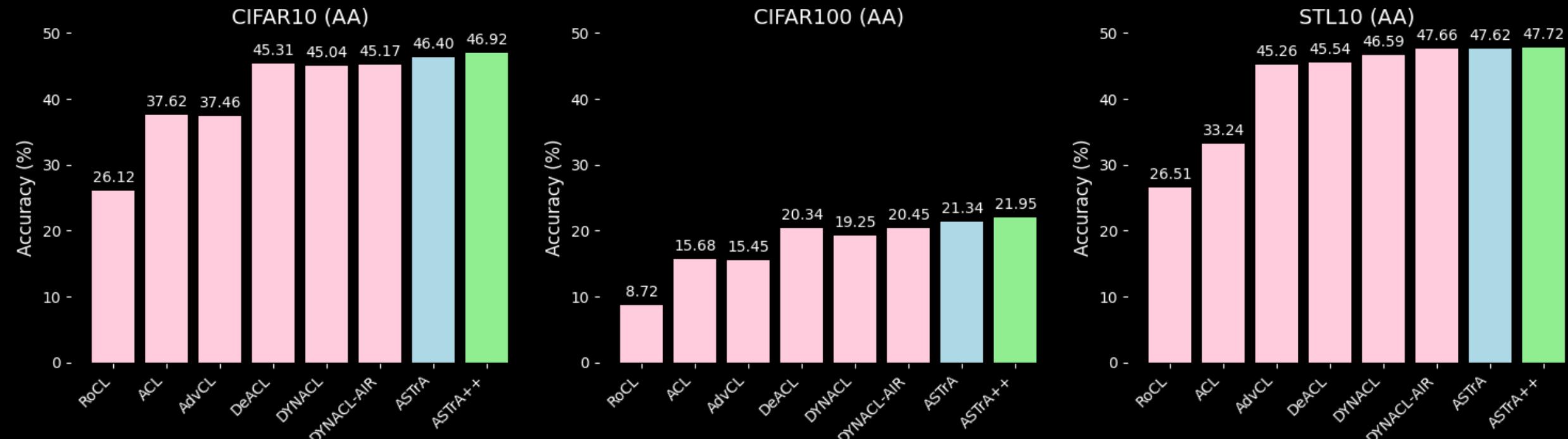


AA: Auto Attack Accuracy

Other results – CIFAR100, STL10, ImageNet100

ASTrA – robustness performance

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 – SLF on ImageNet100



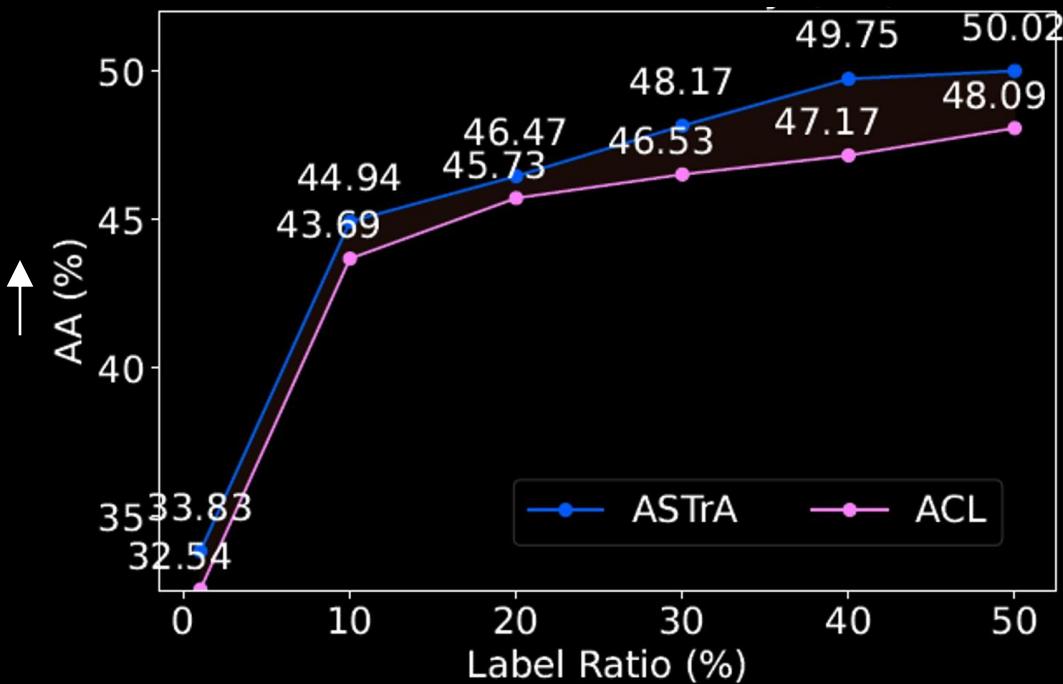
ASTrA as plug-N-play

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

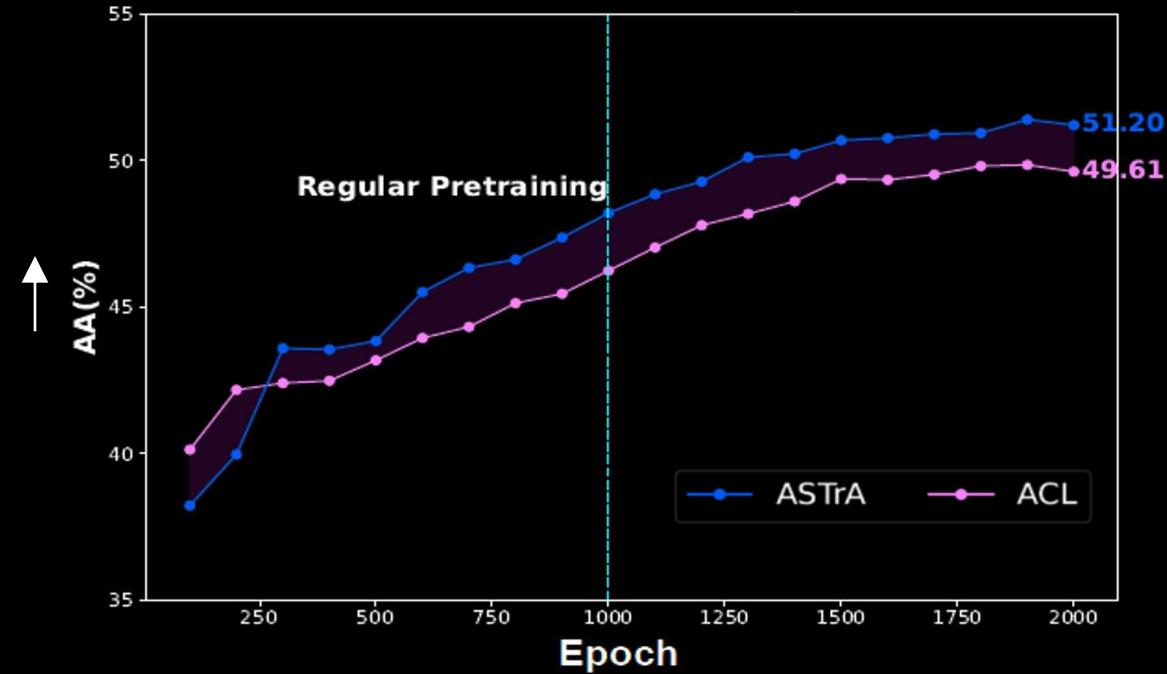


ASTrA - label efficiency & robust overfitting

semi-supervised setting



robust overfitting



AA: Auto Attack Accuracy

Outcomes

Unification of SSL approaches

Promotes potential improvements in approach agnostic and modular way.

Identified vulnerabilities of SSL models

Understanding lack of robustness of SSL models in real-world challenges.

Enhanced robustness against perspective distortion

Eliminated the need for camera parameters and real PD data, making SSL models robust without them.

Improved performance of real world applications

Addressed the consequences of perspective distortion for practical applications.

Enhanced adversarial robustness

Intorduced plug-n-play self-supervised adversarial traning framework with learning attack strategy.

Thank you everyone



<https://prakashchhipa.github.io/>