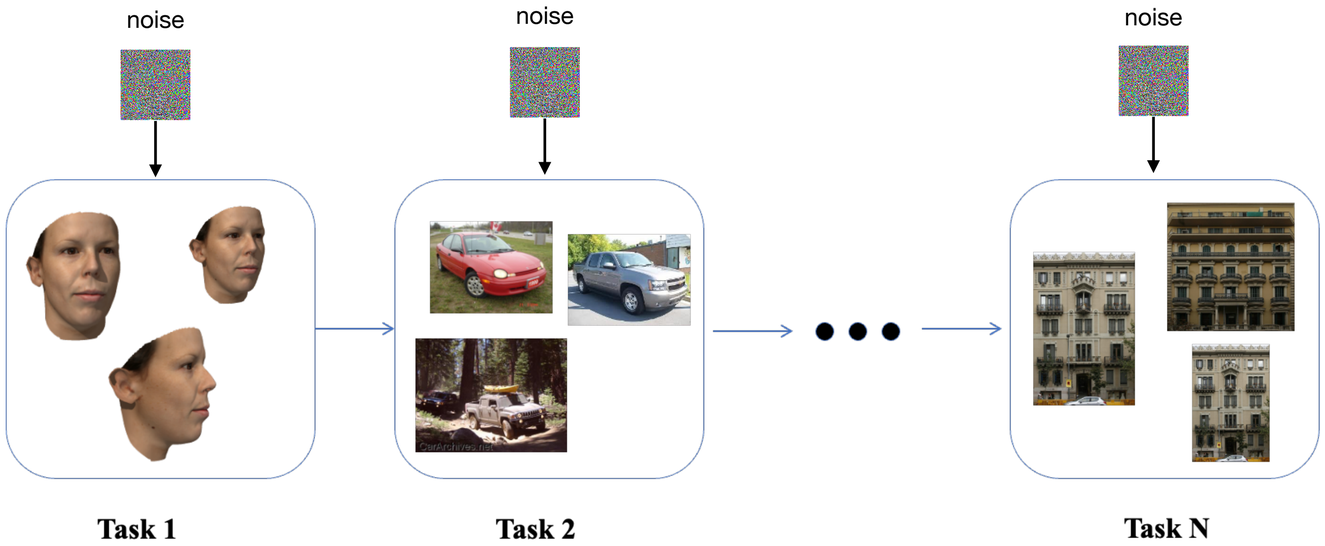
Dynamic Siamese Expansion Framework for Improving Robustness in Online Continual Learning

Fei Ye[1], Yulong Zhao[1], Qihe Liu[1], Junlin Chen[1], Adrian G. Bors[2], Jingling Sun[1], Rongyao Hu[1], Shijie Zhou[1]

[1]University of Electronic Science and Technology of China
[2]University of York, UK

Background

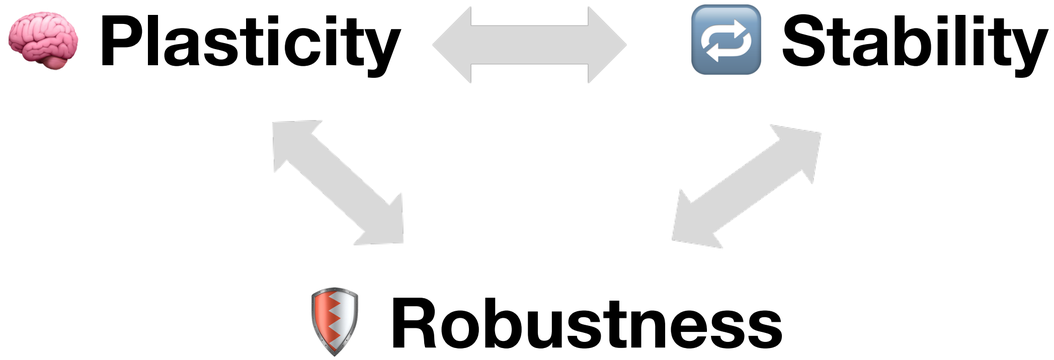


Continual Learning with
Noise/Adversarial Perturbation

Why Robust Continual Learning?

- Continual learning (CL) aims to learn sequential tasks without catastrophic forgetting.
- Real-world data → noisy / adversarial (illumination, weather, camera noise).
- Existing CL methods ≈ forgetting mitigation only → poor robustness.

Motivation

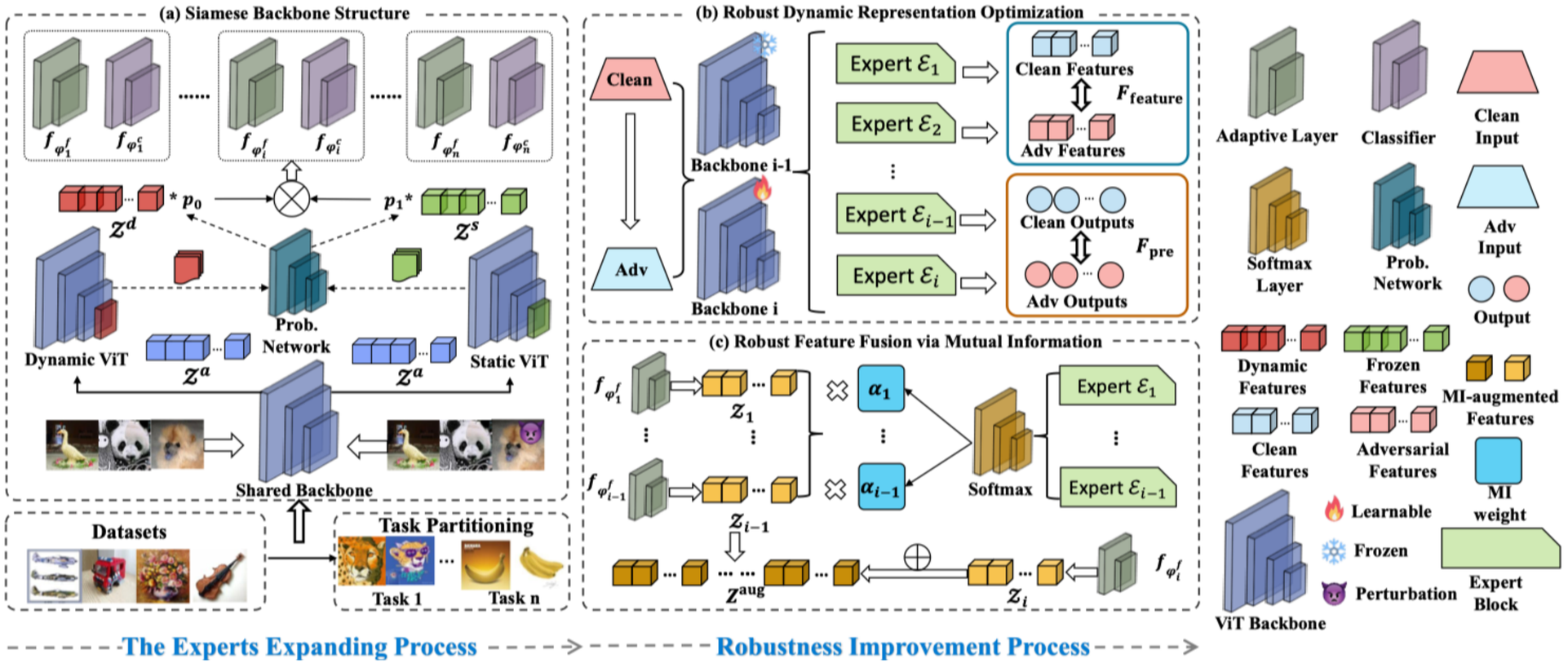


The Challenge of Online Continual Adversarial Defense (OCAD)

- OCAD = online CL + adversarial attacks.
- Requires balance:
  - -Plasticity (new task adaptation)
  - -Stability (old knowledge retention)
  - -Robustness (resistance to perturbation)

Our goal: a unified model that adapts to new tasks without forgetting and stays resistant to adversarial perturbations.

Framework: DSEM



DSEF builds experts dynamically on a Siamese ViT backbone and integrates robust representation optimization with mutual-information fusion.

Method I: Robust Dynamic Representation Optimization(RDRO)

(1) Prediction Consistency Term

$$F_{\text{pre}} = \sum_{i=1}^{j-1} \left\{ F_{\text{mse}}\left(F_{\varphi_i^c}\left(F_{\varphi_i^f}\left(F_{\theta^d}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x})) \right) \right) \right), \right.$$
$$\left. F_{\varphi_i^c}\left(F_{\varphi_i^f}\left(F_{\theta^s}(F_{\theta^a}(\mathbf{x})) \otimes F_{\theta^s}(F_{\theta^a}(\mathbf{x})) \right) \right) \right\},$$

(2) Adversarial Robustness Term

$$F'_{\text{pre}} = \min_{\theta_d} \left\{ F_{\text{pre}}(x) \sum_{i=1}^{j-1} \max_{\|\mathbf{x}'-\mathbf{x}\| \le \epsilon} F_{\text{ce}}\left(y, F_p(\mathbf{x}', E_j)\right) \right\}$$

where $x' = x + \nabla_x F_{ce}\left(F_p(x, E_j), y\right)$ represents adversarial samples generated within perturbation radius ϵ.

(3) Adversarial Robustness Term

$$F_{\text{feature}} = \min_{\theta_d} \frac{1}{j-1} \sum_{i=1}^{j-1} \left(\mathscr{L}_M(P(Z_i), P(\hat{Z}_i)), \mathscr{L}_M(P(Z_i'), P(\hat{Z}_i')) \right)$$

where $P(Z)$ and $P(\hat{Z})$ denote Borel probability measures for feature distributions before and after dynamic updates.

(4) Final RDRO Objective

The complete RDRO optimization combines predictive and feature constraints:

$$F_{\text{RDRO}} = F_{\text{feature}} + F'_{\text{pre}}$$

RDRO aligns both predictions and latent features of clean and adversarial samples, ensuring the dynamic backbone updates without forgetting.

Method II: Mutual Information-Based Robust Feature Fusion (MBRFF)

(1) Mutual Information Computation

For each previous expert E_i, the mutual information between its prediction Y_i and current task output Y is defined as:

$$I(Y_i; Y) = \sum_{y_i \in Y_i} \sum_{y \in Y} P(Y_i, Y)(y_i, y) \log \frac{P(Y_i, Y)(y_i, y)}{P(Y_i)(y_i)P(Y)(y)}$$

where $P(Y_i, Y)$ is the joint probability distribution between historical and current predictions.

(2) Adaptive Weight

$$\alpha_i = \frac{e^{I(Y_i;Y)}}{\sum e^{I(Y_c;Y)}}$$

α_i uses mutual information to weight historical experts by their relevance to the new task.

(3) Robust Feature Aggregation

Using these adaptive weights, historical representations are fused into an augmented feature:

$$Z_{\text{aug}} = \sum_{i=1}^{j-1} \alpha_i F^{\phi_{f_j}}\left(F^{\gamma_i}(x)[0]F_{\theta_d}(F_{\theta_a}(x)) \otimes F^{\gamma_i}(x)[1]F_{\theta_s}(F_{\theta_a}(x))\right)$$

MBRFF measures how relevant each past expert is to the current task through mutual information, then fuses their features adaptively.

Experiments Results

Standard results

Split CIFAR-10								
Methods	Refresh	Refresh (Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	**92.47%**	91.76%	92.46%	91.49%	91.42%	91.70%	49.80%	90.72%
FGSM	55.84%	58.09%	55.51%	60.78%	56.13%	42.39%	18.64%	**82.36%**
PGD	05.32%	06.43%	05.79%	07.29%	05.29%	06.23%	03.92%	**79.79%**
PGDL2	65.87%	68.64%	64.42%	69.58%	64.28%	52.17%	22.34%	**82.43%**
BIM	48.69%	47.96%	50.60%	48.79%	47.63%	48.47%	16.65%	**87.43%**
CW	00.39%	00.34%	00.39%	00.19%	00.27%	00.76%	00.29%	**82.43%**
AutoAttack	03.17%	04.79%	02.06%	02.77%	02.14%	03.87%	00.76%	**90.90%**
Average	38.82%	39.71%	40.12%	41.77%	38.16%	35.08%	16.05%	**85.15%**

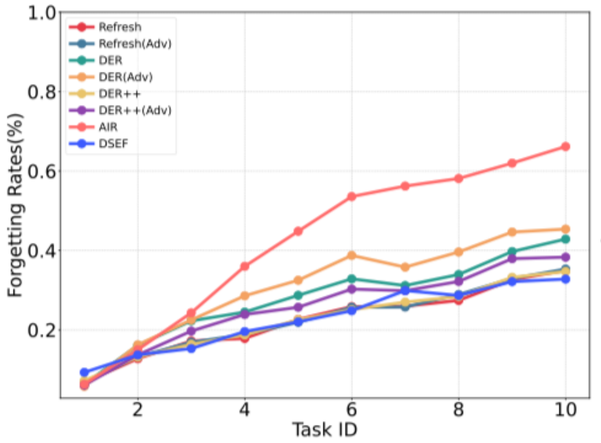
Split CIFAR-100								
Methods	Refresh	Refresh(Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	62.24%	61.37%	52.79%	48.67%	57.74%	57.49%	23.79%	**68.17%**
FGSM	25.89%	27.42%	21.49%	21.09%	22.95%	18.74%	09.43%	**51.71%**
PGD	03.29%	04.94%	03.76%	05.27%	04.16%	04.32%	01.46%	**44.79%**
PGDL2	32.96%	34.47%	27.68%	25.74%	30.73%	24.49%	12.93%	**53.42%**
BIM	25.47%	24.17%	22.49%	21.36%	22.98%	23.18%	10.27%	**60.79%**
CW	00.58%	00.29%	00.59%	00.94%	00.56%	00.79%	00.31%	**52.68%**
AutoAttack	02.34%	03.28%	02.44%	03.73%	02.84%	03.07%	00.89%	**66.52%**
Average	21.82%	22.27%	18.74%	18.11%	20.28%	18.86%	08.44%	**56.86%**

Complex results

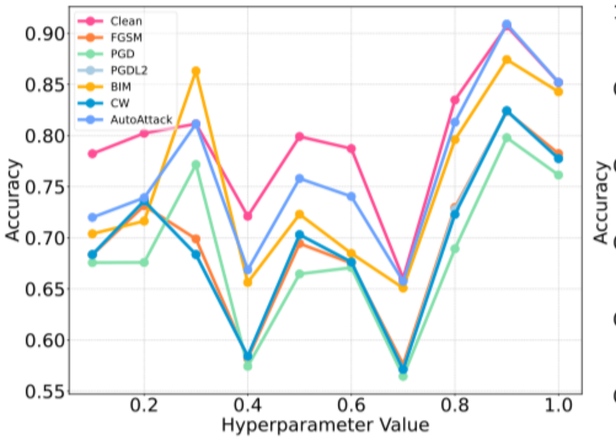
Split CUB200								
Methods	Refresh	Refresh (Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	**67.62%**	61.43%	58.75%	46.55%	65.03%	53.47%	29.37%	58.47%
FGSM	26.84%	**28.52%**	21.84%	19.83%	25.18%	20.82%	10.48%	26.79%
PGD	00.46%	00.96%	00.47%	00.56%	00.42%	00.52%	00.21%	**19.83%**
PGDL2	37.85%	**39.76%**	32.17%	27.37%	35.28%	28.94%	15.58%	27.48%
BIM	22.17%	18.74%	18.16%	15.83%	21.74%	17.12%	08.33%	**34.85%**
CW	05.16%	03.56%	04.32%	05.72%	04.76%	04.25%	04.15%	**26.23%**
AutoAttack	00.21%	00.15%	00.21%	00.29%	00.17%	01.65%	09.74%	**40.74%**
Average	22.90%	21.87%	19.41%	16.59%	21.79%	18.11%	11.12%	**33.48%**

Split TinyImageNet								
Methods	Refresh	Refresh(Adv)	DER	DER(Adv)	DER++	DER++ (Adv)	AIR	DSEF
Clean	63.28%	62.36%	54.32%	52.62%	**63.36%**	60.26%	30.27%	60.21%
FGSM	25.84%	25.17%	21.68%	18.97%	26.42%	18.47%	11.75%	**45.34%**
PGD	02.38%	03.12%	01.97%	02.65%	02.46%	02.13%	00.82%	**40.13%**
PGDL2	33.78%	34.58%	31.34%	28.18%	33.48%	25.17%	13.47%	**47.78%**
BIM	22.46%	22.84%	18.57%	19.67%	23.52%	19.42%	10.94%	**55.73%**
CW	00.64%	00.42%	00.65%	00.57%	00.69%	00.67%	00.34%	**47.77%**
AutoAttack	01.12%	01.25%	00.82%	01.14%	00.94%	00.86%	00.34%	**60.80%**
Average	21.35%	21.39%	18.47%	17.68%	21.55%	18.14%	09.70%	**51.10%**

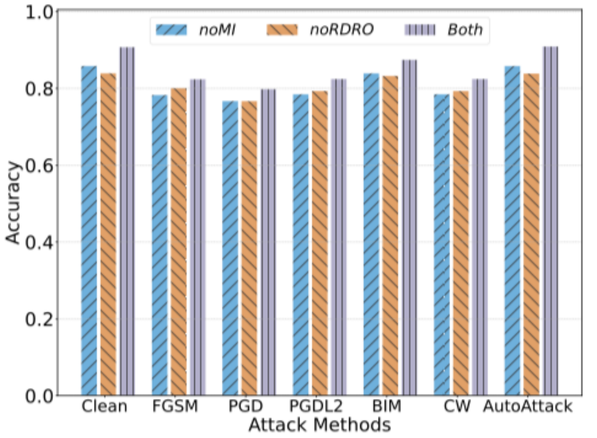
Visualization Results



(a) The forgetting curve.

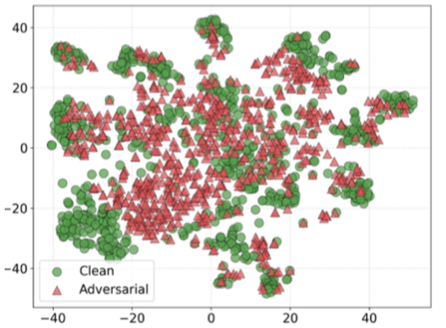


(b) The hyperparameter λ analysis.

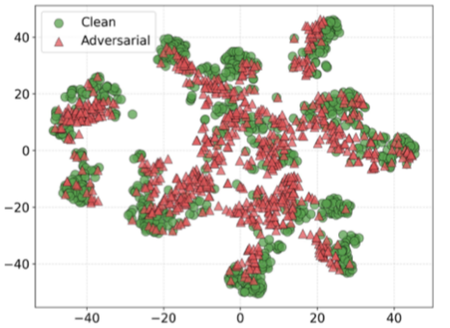


(c) Different configurations.

t-SNE



(a) TinyImageNet



(b) CIFAR-100

Conclusion & Future Work

✅ Proposed DSEF = Siamese backbone + RDRO + MBRFF.

✅ Achieves state-of-the-art robustness in OCAD setting.

🚀 Future: evaluate on new attack types & extend to multi-modal continual learning.

Thank you !