Learning Multi-Source and Robust Representations for Continual Learning

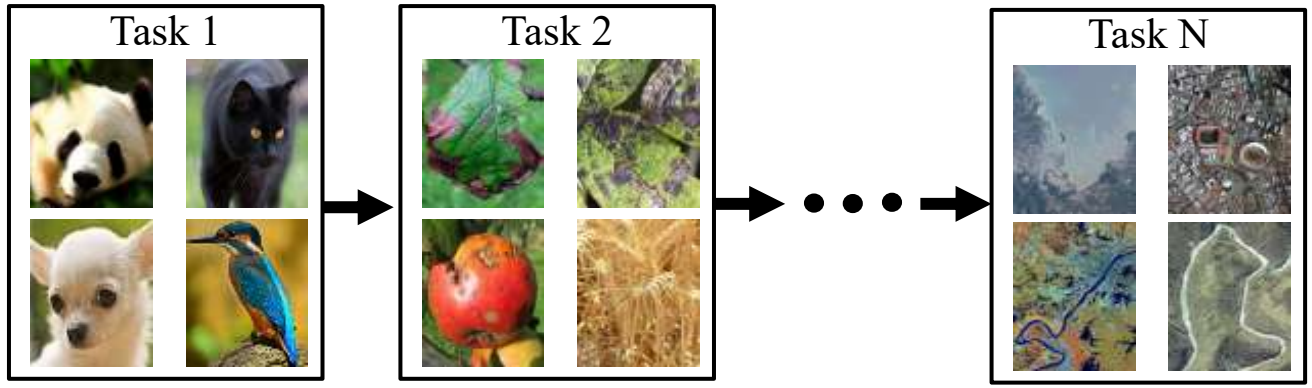
Fei Ye[1], Yongcheng Zhong[1], Qihe Liu[1], Adrian G. Bors[2], Jingling Sun[1], Rongyao Hu[1], Shijie Zhou[1]

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

Background

Continual Learning (CL)

- Goal: Enable models to learn continuously from a sequence of tasks.
- Key challenge: Catastrophic forgetting — new knowledge overwrites old ones.
- Two essential abilities:
  Plasticity – adapt quickly to new tasks.
  Stability – preserve past knowledge.



Existing Solutions

Category	Key Idea	Limitation
Rehearsal-based	Store a small memory buffer of old samples	Limited memory → poor scalability
Dynamic expansion	Add new sub-networks for new tasks	Leads to parameter growth
Regularization-based	Constrain parameter updates	Over-regularization → reduced plasticity

Motivation

Limitation of Pretrained Backbones(CL)

- Recent CL methods use pretrained ViTs or CNNs to improve stability.
- However:
    - Relying on a **single pretrained backbone** restricts adaptability.
    - Fixed representations often fail to generalize to new domains.
    - Updating too many layers → instability; freezing too many → rigidity.

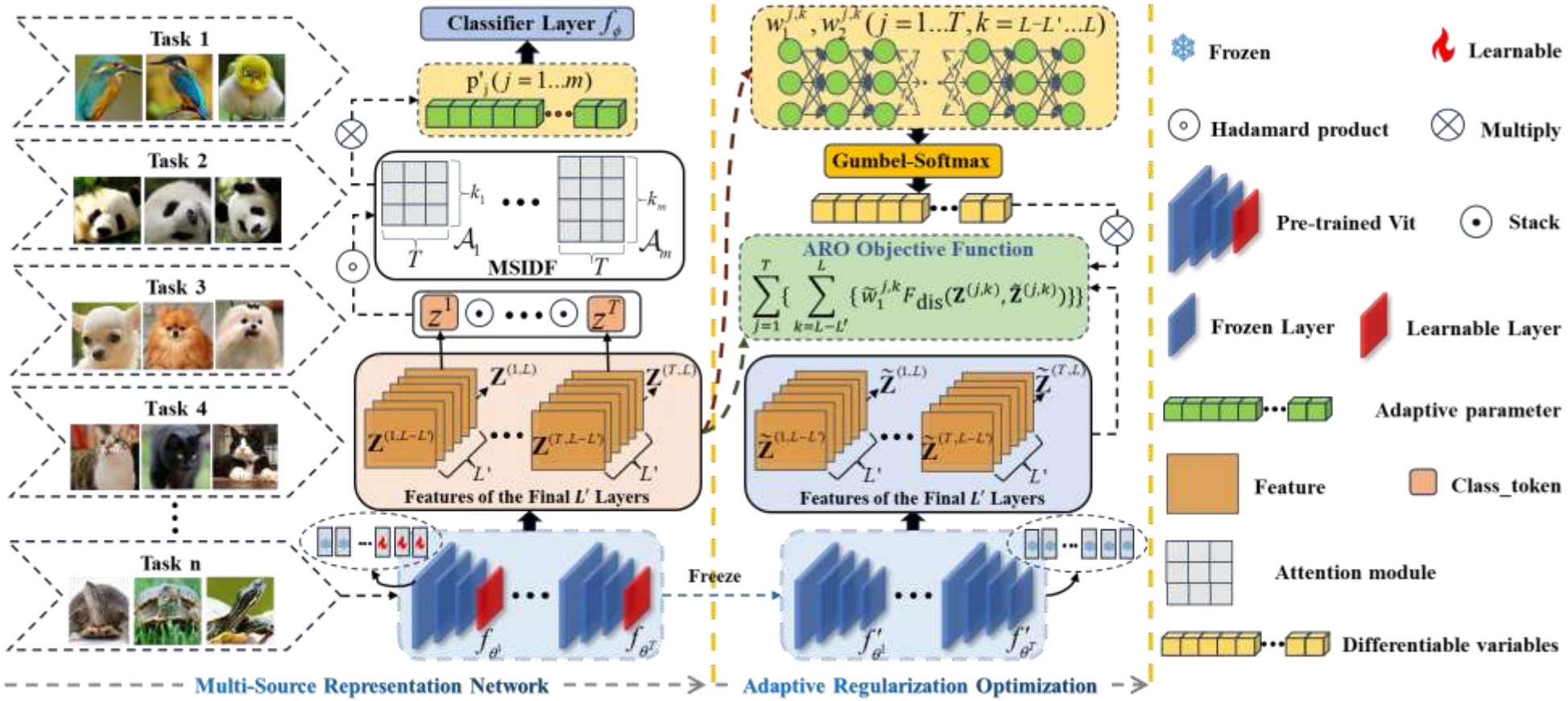
Our Motivation

Can we **leverage multiple pretrained models** to build a **robust and adaptive representation space**, achieving both **stability** and **plasticity** without architectural expansion?

LMSRR Framework

- A novel framework for learning multi-source and robust representations in continual learning.
- Core idea:
    - Fuse features from **multiple ViT backbones**
    - Adapt representations through **dynamic and adaptive optimization**

Framework: LMSRR



LMSRR dynamically fuses multi-source features and adaptively regulates representation learning, achieving an optimal **balance between stability and plasticity** in continual learning.

Method I: Multi-Scale Interaction and Dynamic Fusion

(1) Stack features:

$$\tilde{\mathbf{z}}_s = f_{\theta^1}(\mathbf{x}_s) \bullet \cdots \bullet f_{\theta^T}(\mathbf{x}_s)$$

(2) Multi-scale attention:

- Apply *m* learnable attention modules $\mathcal{A}_j$ with window size k_j.
- Each module highlights relevant cross-backbone patterns via W^j.

(3) Adaptive fusion:

- Learn weights $\boldsymbol{p_j \rightarrow softmax \rightarrow p'_j = \frac{exp(p_j)}{\sum_{c=1}^{m} exp(p_c)}}$
- Final representation:

$$\mathbf{Z_s} = \sum_{j=1}^{m} p'_j \cdot \mathbf{Z_s^j}$$

Advantages:

- Output dimension fixed (independent of *T*).
- Automatically focuses on most informative backbone combinations per input.

Goal of MSIDF: Fuse multi-source ViT features without redundancy, with adaptivity.

Method II : Multi-Level Representation Optimization (MLRO)

Mechanism:

- Keep a frozen copy of each ViT from previous task $(f'_{\theta j})$
- For current task, extract features from last L' layers of both:
  - Active network: $\mathbf{Z}^{(\mathbf{j},\mathbf{k})}$
  - Frozen network: $\widetilde{\boldsymbol{Z}}^{(\boldsymbol{j},\boldsymbol{k})}$
- Minimize L2 distance between them:

$$\mathcal{L}_{\mathrm{MLRO}} = \sum_{j=1}^{T} \sum_{k=L-L'}^{L} ||\mathbf{Z}^{(\mathbf{j},\mathbf{k})} - \widetilde{\boldsymbol{Z}}^{(\boldsymbol{j},\boldsymbol{k})}||_2$$

Goal of MLRO: Prevent representation drift during fine-tuning → preserve stability.

Method III: Adaptive Regularization Optimization (ARO)

Problem: MLRO applies uniform regularization → may over-constrain some layers.

Solution: Introduce learnable gate per layer to control regularization strength.

Key Idea:

For each layer k in backbone j, learn switch $\left(w_1^{j,k}, w_2^{j,k}\right)$

Use Gumbel-Softmax to get differentiable weight $\widetilde{w}_1^{\{\mathrm{j,k}\}}$

$$w_1^{j,k} = \frac{\exp((\log(w_1^{j,k}) + g_1)/\tau)}{\sum_{t=1}^{2}\{\exp((\log(w_t^{j,k}) + g_t)/\tau)\}},$$

Modify loss:

$$L_{ARO} = \sum_{j=1}^{T}\left\{\sum_{k=L-L'}^{L}\left\{w_1^{j,k} F_{\mathrm{dis}}(\mathbf{Z}^{(j,k)}, \mathbf{Z}^{(j,k)})\right\}\right\}$$

Why it works:

If a layer is critical for new task :

$\widetilde{w}_1^{\mathrm{j,k}} \approx 0 \rightarrow$ less constraint.

If a layer encodes old knowledge :

$\widetilde{w}_1^{\mathrm{j,k}} \approx 1 \rightarrow$ strong constraint.

Experiments Results

Standard datasets results

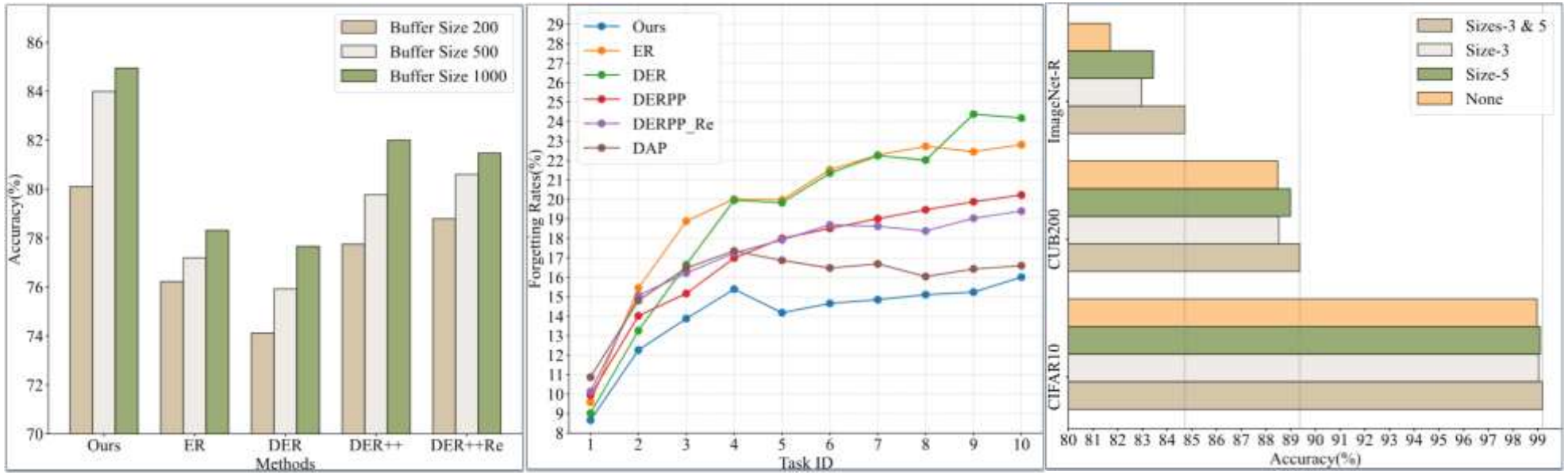
Buffer	Method	CIFAR-10		Tiny ImageNet		R-MNIST
		Average	Last	Average	Last	Domain-IL
-	EWC [51]	68.29±3.92	**97.07±0.74**	19.20±0.31	75.15±3.18	77.35±5.77
	SI [63]	68.05±5.91	94.18±0.88	36.32±0.13	65.80±3.25	71.91±5.83
	LwF [37]	63.29±2.35	96.75±0.35	15.85±0.58	77.95±3.60	-
	PNN [50]	95.13±0.72	96.63±0.10	67.84±0.29	69.03±1.01	-
	DAP [27]	**97.13±2.06**	96.05±3.39	**92.49±0.60**	**94.95±1.20**	**88.58±2.53**
200	ER [49]	91.19±0.94	97.50±0.35	38.17±2.00	79.40±0.28	85.01±1.90
	GEM [39]	90.44±0.94	96.60±0.35	-	-	80.80±1.15
	A-GEM [12]	83.88±1.49	97.90±0.07	22.77±0.03	78.65±3.32	81.91±0.76
	iCaRL [48]	88.99±2.13	97.07±0.32	28.19±1.47	47.45±0.78	-
	FDR [7]	91.01±0.68	97.78±0.24	40.36±0.68	81.40±0.70	85.22±3.35
	GSS [3]	88.80±2.89	97.42±0.24	-	-	79.50±0.41
	HAL [11]	82.51±3.20	94.60±0.14	-	-	84.02±0.98
	DER [8]	91.40±0.92	97.80±0.28	40.22±0.67	79.15±0.21	90.04±2.61
	DER++ [8]	91.92±0.60	97.72±0.38	40.87±1.16	78.35±0.49	90.43±1.87
	DER++(re) [56]	92.01±3.03	97.65±3.03	47.61±8.87	81.40±1.41	91.64±2.26
	Ours	**98.85±0.05**	**99.35±0.21**	**92.08±0.31**	**96.00±0.01**	**94.20±1.24**
500	ER [49]	93.61±0.27	97.15±0.28	48.64±0.46	80.80±1.69	88.91±1.44
	GEM [39]	92.16±0.69	96.63±0.17	-	-	81.15±1.98
	A-GEM [12]	89.48±1.45	97.40±0.78	25.33±0.49	81.00±0.42	80.31±6.29
	iCaRL [48]	88.22±2.62	96.57±0.10	31.55±3.27	50.65±1.20	-
	FDR [7]	93.29±0.59	97.32±0.24	49.88±0.71	81.10±0.56	89.67±1.63
	GSS [3]	91.02±1.57	96.97±0.24	-	-	81.58±0.58
	HAL [11]	84.54±2.36	94.22±0.60	-	-	85.00±0.96
	DER [8]	93.40±0.39	97.90±0.28	51.78±0.88	79.30±1.13	92.24±1.12
	DER++ [8]	93.88±0.50	98.10±0.01	51.91±0.68	76.20±5.23	92.77±1.05
	DER++(re) [56]	93.06±0.38	97.75±0.38	54.06±0.79	79.65±1.34	93.28±0.75
	Ours	**99.15±0.05**	**99.48±0.04**	**92.75±0.32**	**96.23±0.40**	**96.97±1.58**
1000	ER [49]	95.34±0.16	97.67±0.67	55.92±0.90	80.30±0.82	90.42±1.07
	GEM [39]	93.67±0.32	97.37±0.17	-	-	81.15±1.98
	A-GEM [12]	85.61±2.01	97.45±0.42	24.29±1.28	79.65±2.19	81.30±5.33
	iCaRL [48]	91.40±1.06	96.85±0.35	63.87±0.25	54.00±2.82	-
	FDR [7]	94.02±0.64	97.60±0.56	56.05±0.71	80.25±0.49	91.68±1.01
	GSS [3]	91.79±2.16	96.10±1.70	-	-	82.25±2.42
	HAL [11]	87.33±1.46	92.27±3.21	-	-	89.33±2.01
	DER [8]	92.33±0.61	97.72±0.07	56.62±1.13	78.50±0.42	93.13±0.28
	DER++ [8]	94.99±0.26	97.94±0.08	58.05±0.52	79.95±0.35	93.82±0.39
	DER++(re) [56]	93.66±1.00	97.40±0.01	61.91±1.15	80.45±3.18	93.37±0.58
	Ours	**99.21±0.06**	**99.43±0.03**	**93.24±0.24**	**96.10±0.57**	**97.05±0.04**

Experiments Results

Complex datasets results

Method	CIFAR-100		CUB-200		Imagenet-R		Cars196	
	Average	Last	Average	Last	Average	Last	Average	Last
ER [49]	73.37±0.43	93.35±1.34	30.57±4.81	35.57±14.86	24.85±4.06	45.85±0.01	30.52±4.4	54.32±5.07
A-GEM [12]	48.06±0.57	92.80±0.32	13.22±0.31	42.18±0.01	16.87±2.65	47.56±12.31	8.07±0.15	16.45±7.41
FDR [7]	76.29±1.44	93.60±1.34	23.94±0.07	45.58±0.19	15.74±3.69	42.14±10.75	31.41±1.30	58.36±1.17
GSS [3]	57.50±1.93	92.80±2.98	27.04±0.28	42.01±0.08	17.83±0.88	33.44±6.75	34.67±2.27	56.80±4.15
DER [8]	74.93±1.06	93.25±0.35	26.19±2.07	51.79±1.08	18.26±1.67	25.26±0.47	39.75±0.36	68.02±5.20
DER++ [8]	75.64±0.60	92.60±0.14	33.40±1.48	49.83±1.63	22.87±5.83	43.10±10.51	35.39±3.38	60.56±8.45
DER++refresh [56]	77.71±0.85	93.40±1.13	35.77±3.20	50.85±0.47	23.74±3.03	31.00±0.01	33.94±2.46	60.29±4.73
CoFiMA [41]	94.21±0.47	96.13±0.59	**90.66±0.76**	92.54±0.28	83.76±0.53	85.86±0.58	87.28±0.54	90.33±0.45
DAP [27]	90.11±0.33	92.30±2.12	71.83±1.44	72.23±2.85	83.22±1.25	84.61±2.85	39.79±1.85	65.35±2.21
L2P [57]	95.36±0.12	96.80±0.14	86.30±0.21	90.81±0.24	**86.01±0.30**	87.50±0.90	79.55±0.86	84.45±0.12
Ours	**95.76±0.08**	**98.70±0.37**	88.91±0.64	**94.31±0.12**	84.35±0.52	**88.43±0.15**	**90.14±0.06**	**95.32±0.39**

Ablation Study



(a) Same backbone comparison. **(b) The forgetting curve.** **(c) Configuration comparison.**

(a) Comparison of performance of various models with varying buffer sizes on ImageNet-R, where each model uses the same backbone. (b) Comparison of forgetting curves of the proposed approach with other benchmark methods on ImageNet-R. (c) Performance variations of the proposed MSIDF method under different configurations.

Conclusion

✅ We propose LMSRR, a novel continual learning framework that orchestrates multiple pre-trained ViTs to learn robust, adaptive representations.

✅ Our method introduces three key innovations:

- MSIDF: Dynamically fuses multi-source features via learnable attention and adaptive weighting.
- MLRO: Preserves stability by minimizing representation drift across tasks.
- ARO: Relieves over-regularization through layer-wise adaptive gating.

✅ LMSRR achieves state-of-the-art performance without expanding the model or using task-specific parameters.