Learning Expandable and Adaptable Representations for Continual Learning

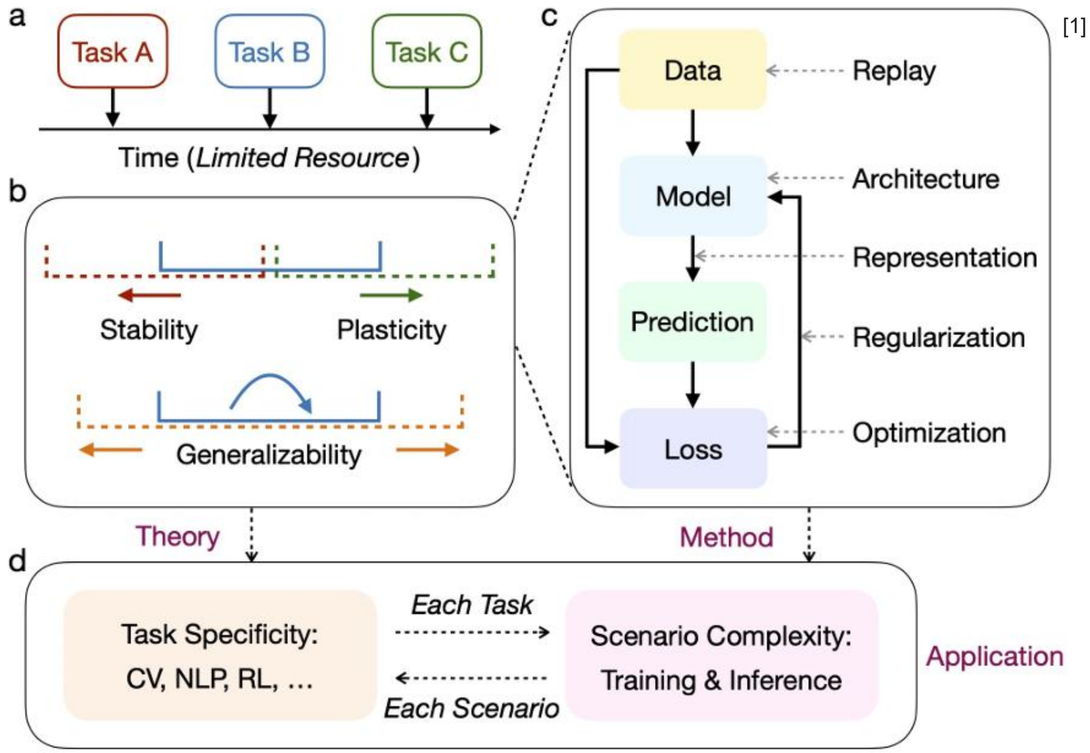
Ruilong Yu[1], Mingyan Liu[2], Fei Ye[1], Adrian G. Bors[3], Rongyao Hu[1], Jingling Sun[1] and Shijie Zhou[1]

[1]University of Electronic Science and Technology of China, China

[2]Harbin Institute of Technology, Shenzhen, China

[3]University of York, UK

Background



[1]

- **Continual learning (CL)** aims to learn sequential tasks without catastrophic forgetting, balancing between Stability and Plasticity.

- **Class-incremental Learning (CIL)** focuses on sequentially learning new classes while maintaining performance on previously learned ones.

- **Domain-incremental learning (DIL)** aims to adapt a model to new data distributions or domains without forgetting previously acquired knowledge.

[1] Wang L, Zhang X, Su H, et al. A comprehensive survey of continual learning: Theory, method and application[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(8): 5362-5383.

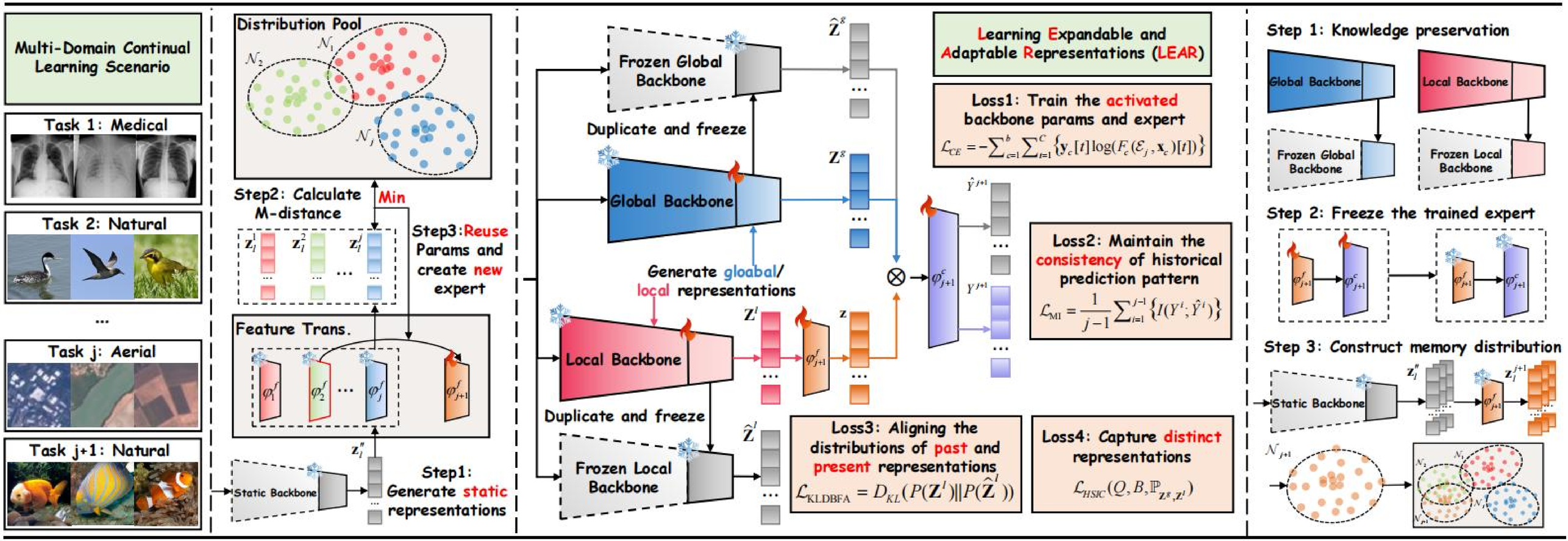
Motivation

- **Current Research Limitations:** Most continual learning studies focus on simplified scenarios confined to a single data domain, which fails to reflect real-world applications where models encounter multiple, evolving data domains.

- **Multi-domain Continual Learning (MDCL) Challenge:** When models learn across domains with significant distribution shifts (e.g., medical → aerial → natural images), they experience severe catastrophic forgetting while struggling to maintain performance across all previously learned domains.

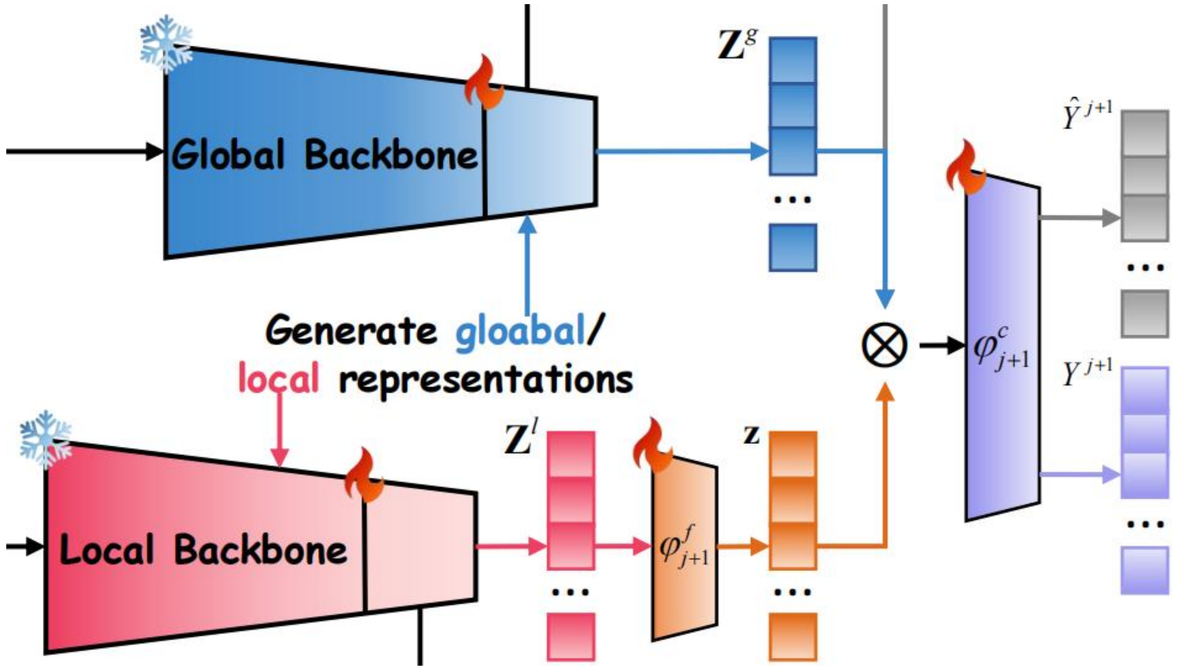
- **Research Objective**: Develop a framework that simultaneously optimizes for three critical properties in MDCL scenarios:
  - Plasticity: Rapid adaptation to novel domains
  - Stability: Preservation of previously learned knowledge
  - Efficiency: Computational and parameter efficiency when learning similar domains

Framework:LEAR



The data samples from new tasks are processed through a **collaborative backbone structure** to learn task-shared, task-specific and backbone distinct representations via the proposed **MIBPA, KLDBFA and HSICBCO**, respectively. **ESM** constructs memory distributions and selects relevant experts for network expansion and test evaluation.

Collaborative Backbone



LEAR introduces a collaborative backbone structure with dynamic expert expansion:

- Collaborative Backbone comprises **global** and **local** backbones designed to capture complementary representations:
  - Global Backbone: Incrementally updated to learn **task-shared** representations.
  - Local Backbone: Adapts to **task-specific** characteristics while preserving domain knowledge.
  - LEAR **concatenated** global-local representations for prediction.
  - For each new task, LEAR dynamically creates a **lightweight expert** consisting of feature transformation module and Linear classifier.

Mutual Information-Based Prediction Alignment (MIBPA)

$$\begin{aligned}\mathbf{Y}^i &= \left\{\mathbf{y}_c \,|\, \mathbf{y}_c = F_{\varphi_i^c}(F_{\theta^g}(\mathbf{x}_c) \oplus F_{\varphi_i^f}(F_{\theta^l}(\mathbf{x}_c))), c = 1, \cdots, b\right\}, \\ \hat{\mathbf{Y}}^i &= \left\{\mathbf{y}_c \,|\, \mathbf{y}_c = F_{\varphi_i^c}(F_{\hat{\theta}^g}(\mathbf{x}_c) \oplus F_{\varphi_i^f}(F_{\theta^l}(\mathbf{x}_c))), c = 1, \cdots, b\right\},\end{aligned} \tag{3}$$

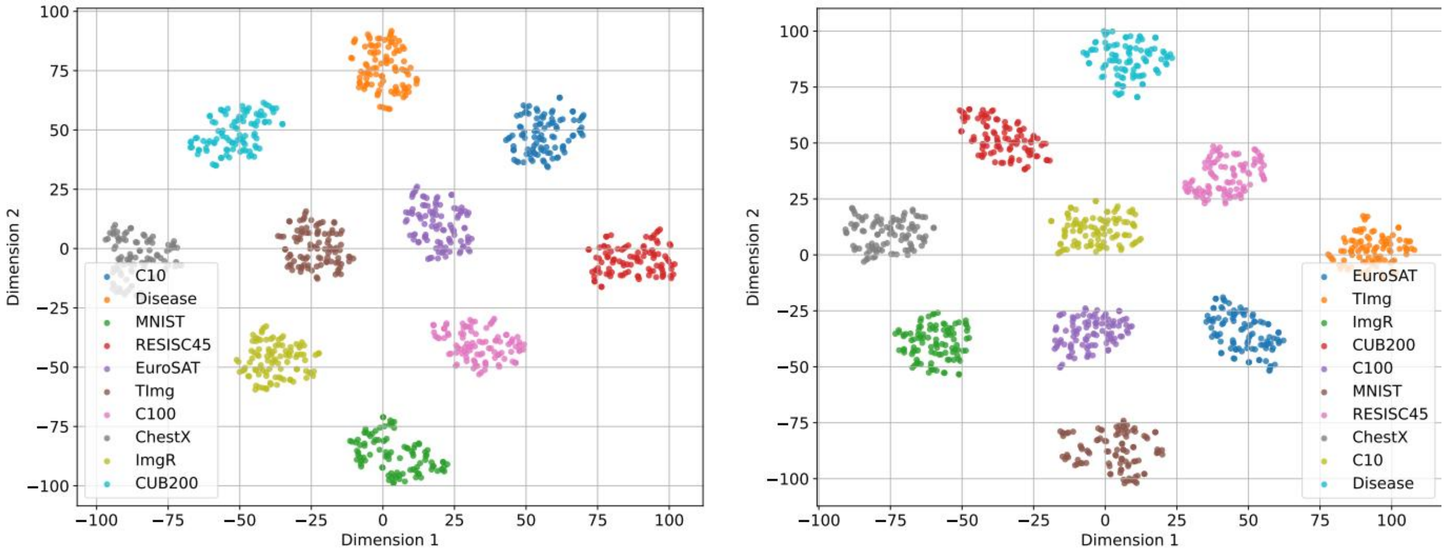
$$I(Y^i; \hat{Y}^i) = \sum\nolimits_{\hat{\mathbf{y}}^i \in \hat{Y}^i} \left\{ \sum\nolimits_{\mathbf{y}^i \in Y^i} \left\{ P(Y^i, \hat{Y}^i)(\mathbf{y}^i, \hat{\mathbf{y}}^i) \log \frac{P(Y^i, \hat{Y}^i)(\mathbf{y}^i, \hat{\mathbf{y}}^i)}{p(Y^i)(\mathbf{y}^i) p(\hat{Y})(\hat{\mathbf{y}}^i)} \right\} \right\}, \tag{4}$$

$$\mathcal{L}_{\mathrm{MI}} = \frac{1}{j-1} \sum\nolimits_{i=1}^{j-1} \left\{ I(Y^i; \hat{Y}^i) \right\}. \tag{5}$$

We introduce a novel regularization approach to **maintain prediction consistency** while updating the global backbone:

- Create parameter-shared auxiliary model by **duplicating and freezing** the global backbone's final layers.
- Generate two distinct prediction sets for **each historical expert** using current and frozen global backbones.
- **Minimize mutual information** between these prediction sets to preserve prediction patterns.

KL Divergence-Based Feature Alignment (KLDBFA)



(a) Feature visualization of CDM with T-SNE. (b) Feature visualization of ETI with T-SNE.

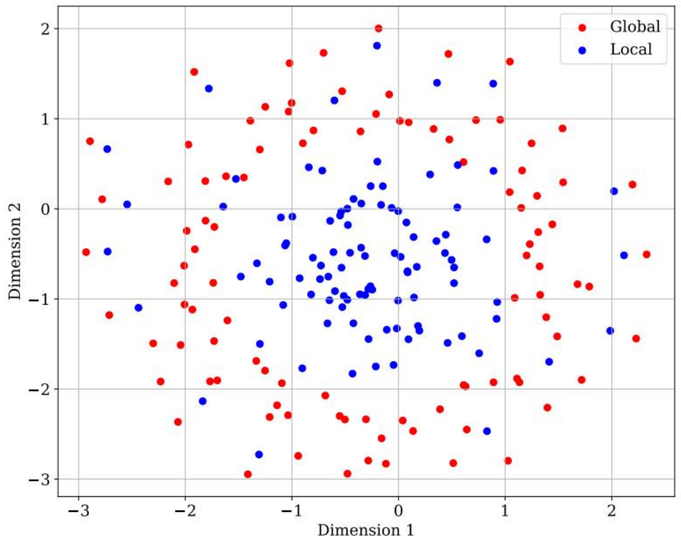
- **Rationale for using KL**: Modern representation evaluation metrics (e.g., FID) operate under Gaussian distribution assumptions in high-dimensional spaces. KL divergence offers directional constraint properties and computational efficiency.
- **Method**: We model feature distributions as Gaussians and apply KL to align current and historical feature distributions.

$$\mathbf{Z}^l = \{\mathbf{z}_c \,|\, \mathbf{z}_c = F_{\theta^l}(\mathbf{x}_c), c = 1, \cdots, b\}, \hat{\mathbf{Z}}^l = \{\mathbf{z}_c \,|\, \mathbf{z}_c = F_{\hat{\theta}^l}(\mathbf{x}_c), c = 1, \cdots, b\}. \tag{6}$$

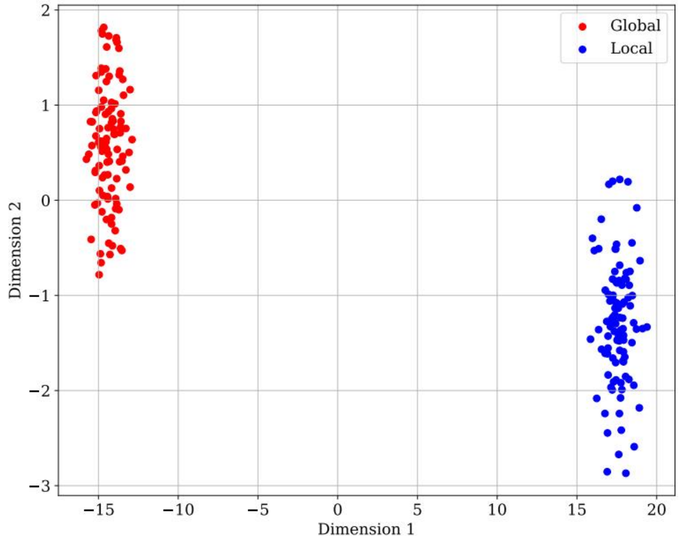
$$D_{KL}(P(\mathbf{Z}^l)||P(\hat{\mathbf{Z}}^l)) = \frac{1}{2}\left[\log\left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right) - d + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1)\right]$$

$$\mathcal{L}_{\mathrm{KLDBFA}} = D_{KL}(P(\mathbf{Z}^l) \,||\, P(\hat{\mathbf{Z}}^l)), \tag{7}$$

HSIC-Based Collaborative Optimization (HSICBCO)



(a) The features without HSICBCO.



(b) The features with HSICBCO.

To maximize **complementary learning** between global and local backbones, we propose **HSICBCO** to encourage global and local backbones to **capture distinct semantic information** by minimizing their statistical dependence.

$$C_{\mathbf{z}_g\mathbf{z}_l} = \mathbb{E}_{\mathbf{z}_g\mathbf{z}_l}\left\{ \left(f_Q(\mathbf{z}_g) - \mathbb{E}_{\mathbf{z}_g}\left[f_Q(\mathbf{z}_g)\right]\right) \otimes \left(f_B(\mathbf{z}_l) - \mathbb{E}_{\mathbf{z}_l}\left[f_B(\mathbf{z}_l)\right]\right) \right\}, \tag{8}$$

$$\begin{aligned} \mathcal{L}_{HSIC}(Q, B, \mathbb{P}_{\mathbf{z}^g,\mathbf{z}^l}) &= \left\| C_{\mathbf{z}_g,\mathbf{z}_l} \right\|_{HS}^2 = \mathbb{E}_{\mathbf{z}_g,\mathbf{z}'_g,\mathbf{z}_l,\mathbf{z}'_l}[k(\mathbf{z}_g,\mathbf{z}'_g)l(\mathbf{z}_l,\mathbf{z}'_l)] \\ &+ \mathbb{E}_{\mathbf{z}_g,\mathbf{z}'_g}[k(\mathbf{z}_g,\mathbf{z}'_g)]\mathbb{E}_{\mathbf{z}_l,\mathbf{z}'_l}[l(\mathbf{z}_l,\mathbf{z}'_l)] - 2\mathbb{E}_{\mathbf{z}_g,\mathbf{z}_l}[\mathbb{E}_{\mathbf{z}'_g}[k(\mathbf{z}_g,\mathbf{z}'_g)]\mathbb{E}_{\mathbf{z}'_l}[l(\mathbf{z}_l,\mathbf{z}'_l)]], \end{aligned} \tag{9}$$

Expert Selection Mechanism (ESM)

- **Memory Distribution Construction**: For each expert, ESM construct multivariate Gaussian distribution and store only **statistical information (mean and covariance)** rather than raw parameters.

$$\boldsymbol{\mu}_j = \frac{1}{m}\sum_{k=1}^{m}\{\mathbf{z}_k\}, \quad \boldsymbol{\Sigma}_j = \frac{1}{m-1}\sum_{k=1}^{m}\left\{(\mathbf{z}_k - \boldsymbol{\mu}_j)(\mathbf{z}_k - \boldsymbol{\mu}_j)^\top\right\}. \tag{11}$$

- **Expert Selection Process**: For new task, ESM compute Mahalanobis distance between incoming features and stored distributions and select expert with minimum average distance.

$$c^* = \underset{c=1,\cdots,j}{\operatorname{argmin}}\left\{\frac{1}{m'}\sum_{l=1}^{m'}\sqrt{(\mathbf{z}_l^c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\mathbf{z}_l^c - \boldsymbol{\mu}_c)}\right\}, \tag{12}$$

- **Parameter Transfer Mechanism**: Initialize new expert using parameters from selected expert for efficient learning across similar domains.
- **Advantages**: Accelerates convergence for similar domains, enables task-agnostic inference without task identity, and reduces parameter redundancy through selective transfer.

Experimental Results

Table 1: The classification accuracy (%) of all testing datasets after learning the **CDM** task sequence.

Methods	C10	Disease	MNIST	RESISC45	EuroSAT	TImg	C100	ChestX	ImgR	CUB200	Avg
DER++(Re)	22.78	34.00	11.33	8.24	18.18	7.66	31.79	25.99	54.04	78.23	29.22
CLS-ER	22.58	27.22	12.50	14.32	24.17	14.14	34.23	25.99	52.46	75.78	30.34
RanPAC	87.17	96.76	87.45	84.20	92.53	71.46	51.44	39.63	43.30	56.18	71.01
MoE	92.74	32.44	91.87	54.02	41.85	6.12	78.98	19.60	**78.43**	77.24	57.33
L2P	20.66	11.15	14.08	4.50	13.65	11.67	31.76	21.52	58.26	81.30	26.85
DAP	8.83	2.78	18.94	3.36	18.19	3.06	11.29	16.34	60.68	80.37	22.38
D-Prompt	25.99	9.08	16.57	4.13	7.30	22.73	38.36	17.33	59.06	82.44	28.30
C-Prompt	13.57	2.33	9.10	1.90	14.18	0.68	3.40	14.35	4.14	60.55	12.42
Ours	**95.44**	**98.46**	**96.59**	**92.04**	**95.00**	**81.24**	**85.10**	**45.95**	70.47	**85.80**	**84.61**

Table 2: The classification accuracy (%) of all testing datasets after learning the **ETI** task sequence.

Methods	EuroSAT	TImg	ImgR	CUB200	C100	MNIST	RESISC45	ChestX	C10	Disease	Avg
DER++(Re)	51.82	36.25	2.32	6.20	23.08	65.97	40.45	27.41	82.69	97.77	43.40
CLS-ER	45.76	29.33	16.19	33.08	30.74	67.10	46.44	30.26	80.29	97.62	47.68
RanPAC	92.64	70.87	43.75	56.13	51.83	88.08	83.51	40.34	86.74	96.88	71.08
MoE	43.01	0.94	55.48	25.16	74.22	97.67	84.57	33.38	**96.88**	**99.90**	61.12
L2P	10.88	1.45	0.97	4.04	14.55	12.17	22.48	9.16	89.69	98.62	26.40
DAP	9.93	1.07	1.84	17.65	15.74	15.44	22.19	16.34	86.20	98.44	28.48
D-Prompt	10.61	1.31	1.44	3.09	17.03	23.83	25.03	14.77	93.42	99.33	28.99
C-Prompt	11.66	0.67	0.62	0.33	1.72	13.04	6.14	16.62	21.74	96.21	16.88
Ours	**95.89**	**81.25**	**69.57**	**84.12**	**85.30**	**98.56**	**92.92**	**45.45**	96.60	99.30	**84.90**

Key Observations:

- LEAR consistently outperforms all baselines across **challenging domains**.
- Significant performance gains in task sequences with **severe distribution shifts.**
- Demonstrates **robustness** to domain ordering with consistent performance across all sequences.

Ablation Results

Table 4: Impact of individual and combined components on model performance in ETI.

Methods	EuroSAT	TImg	ImgR	CUB200	C100	MNIST	RESISC45	ChestX	C10	Disease	Avg
CB	19.59	8.37	11.81	12.29	36.92	49.88	61.14	30.04	92.79	99.00	42.18
SBE	85.38	70.49	55.42	73.45	77.05	90.05	85.93	37.73	93.65	99.13	76.83
CBE	92.38	76.49	62.42	78.45	82.05	95.05	90.93	38.73	94.17	99.04	80.97
CBE+MI	95.08	80.93	68.19	82.49	84.68	97.15	91.29	39.56	96.18	98.28	83.38
CBE+KL	92.18	78.02	66.13	78.97	82.50	92.24	90.54	44.58	95.81	99.13	82.01
CBE+HSIC	93.65	76.97	64.37	78.56	82.35	91.88	91.29	45.53	95.49	99.06	81.92
CBE+MI+KL	95.35	**81.61**	68.30	83.56	**85.54**	97.43	92.28	**46.44**	96.22	99.01	84.57
CBE+MI+HSIC	95.87	80.85	68.74	83.15	85.52	98.03	92.08	42.43	96.33	98.91	84.19
CBE+KL+HSIC	93.24	78.22	66.45	80.49	82.55	92.21	91.39	45.26	95.59	99.08	82.45
LEAR	**95.89**	81.25	**69.57**	**84.12**	85.30	**98.56**	**92.92**	45.45	**96.60**	**99.30**	**84.90**
LEAR w/o ESM	3.72	1.25	2.36	83.95	8.24	4.91	14.63	1.05	17.85	99.15	23.71

This table validates our method's key designs:

- The dual-backbone architecture and dynamic expert expansion are both crucial for performance;
- All regularization terms (MI/KL/HSIC) contribute to improvements;
- The Expert Selection Module (ESM) is essential, as random selection causes a significant performance drop.

Conclusion

- We propose LEAR, a novel framework for **Multi-Domain Continual Learning** that simultaneously addresses **stability**, **plasticity** and **efficiency**.
- Built on a **collaborative backbone** structure, we introduce **MIBPA** and **KLDBFA** to maintain historical prediction consistency and task-specific feature alignment during model updates, while **HSICBCO** ensures disentangled and complementary representations.
- **ESM** dynamically selects relevant experts for efficient network expansion and task-agnostic prediction.
- The empirical results demonstrate the effectiveness of the proposed approach.
- Propose a novel expert merging technology with self-distillation for effective model compression in the future.

Thank you!