VIETNAM NATIONAL UNIVERSITY,
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Pham Thi Quynh Trang

Mitigating Catastrophic Forgetting in Continual Learning via Memory Replay, Knowledge Transfer, and Adaptive Architecture Optimization

DOCTOR OF PHILOSOPHY IN INFORMATION SYSTEMS DISSERTATION

Major: Information Systems
Code: 9480104

Hanoi - 2026

VIETNAM NATIONAL UNIVERSITY,
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Pham Thi Quynh Trang

Mitigating Catastrophic Forgetting in Continual Learning via Memory Replay, Knowledge Transfer, and Adaptive Architecture Optimization

SUMMARY OF DOCTOR DISSERTATION IN INFORMATION SYSTEM

Major: Information Systems
Code: 9480104

Supervisor: Assoc. Prof. Nguyen Tri Thanh
Co-supervisor: Dr. Le Duc Trong

Hanoi - 2026

Abstract

Deep neural networks perform well when training data is fully available before deployment. However, in practice, data arrives continuously and changes over time, making it impractical to retrain models from scratch whenever new data appears. **Continual learning** addresses this by enabling models to learn from sequential tasks without forgetting prior knowledge, a phenomenon known as **catastrophic forgetting**.

Existing approaches fall into three families, each with its own limitations: *replay-based* methods (storing past data for rehearsal, raising privacy concerns), *knowledge transfer* methods (distilling knowledge into new models, constrained by fixed parameter capacity), and *architecture-based* methods (expanding the network per task, with parameter costs growing with sequence length).

This dissertation proposes solutions across three progressive stages:

- **Stage 1** *(memory available, but raw data cannot be stored)*: Prototypes approximate past class distributions, combined with an ELI module to prevent representational drift across tasks.
- **Stage 2** *(no memory available)*: CL-plugin enables selective knowledge transfer via distillation; ZeroRFF reformulates learning as a closed-form regression requiring neither gradient updates nor stored samples.
- **Stage 3** *(long task sequences with diverse domains)*: Four adaptive architecture methods are proposed, most notably SO-LoRA, which shares a universal orthogonal basis across all tasks and restricts each task to learning sparse compositions over it, keeping the adapter footprint constant regardless of sequence length.

Experiments across multiple benchmarks demonstrate state-of-the-art performance on continual relation extraction, aspect sentiment classification, and class-incremental image recognition.

PREAMBLE

Research Motivations

Static machine learning trains models on fixed datasets and deploys them without further adaptation. While widely adopted, this paradigm suffers from four interconnected limitations: inability to adapt to changing data distributions, poor generalization to unseen data, the lack of any continual update mechanism, and vulnerability to concept drift. **Continual Learning (CL)** addresses these shortcomings by enabling models to incrementally acquire new knowledge while retaining previously learned information, reducing the need for periodic full retraining and making adaptive AI systems viable in real-world, non-stationary environments. Figure 1 contrasts the two paradigms, and Figure 2 illustrates the rapid growth of CL research from 243 publications in 2018 to over 1,800 in 2024.

Figure 1: (a) Static machine learning trains on all data once. (b) Adaptive machine learning learns continually from new data as it arrives.

Despite substantial progress, three critical gaps motivate this dissertation. The first gap concerns **data privacy**: the most effective continual learning approach replay requires storing raw samples from past tasks, which is impractical or prohibited in sensitive domains such as healthcare and finance. The second gap concerns **memory**

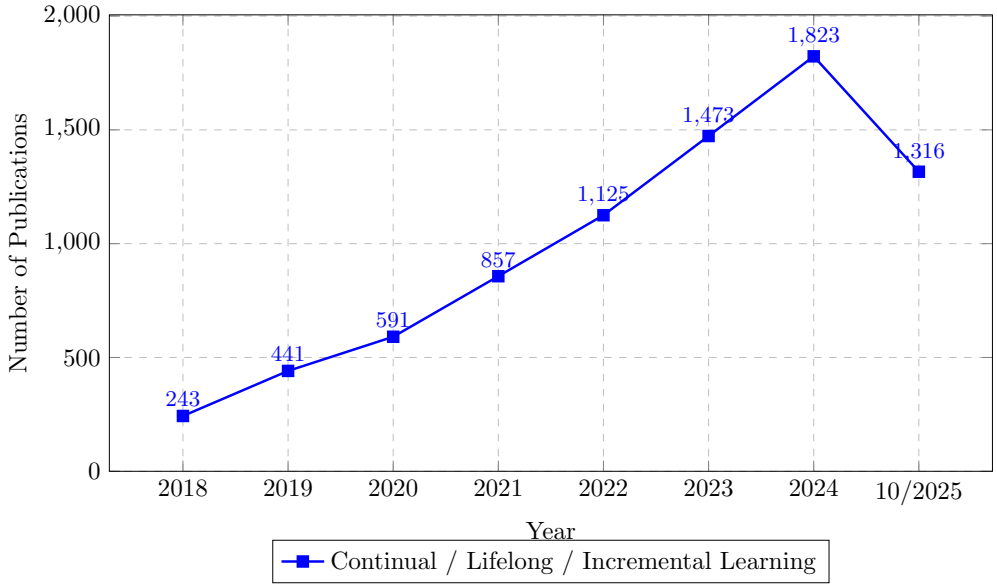


Figure 2: Annual publication trends for *continual learning*, *lifelong learning*, and *incremental learning* in DBLP (2018–2025).

dependency: even privacy-conscious prototype approximations become unreliable as the feature space grows complex, and any residual memory buffer may be infeasible under strict computational budgets. The third gap concerns **model capacity**: when tasks are highly diverse or sequences grow long, a fixed parameter budget forces a trade-off between stability and plasticity that neither regularization nor distillation can fully resolve.

These three gaps define the progressive logic of the dissertation. Chapter 2 answers the first gap by replacing stored real samples with Gaussian-perturbed class prototypes, complemented by an energy-based latent alignment module. Chapter 3 answers the second gap through exemplar-free knowledge transfer: a CL-plugin with contrastive distillation for the task-incremental setting, and ZeroRFF, which applies random Fourier features over a frozen backbone with closed-form analytic solutions for class-incremental learning. Chapter 4 answers the third gap through adaptive architecture strategies a multi-head framework with Online Prototype Equilibrium and Adaptive Prototypical Feedback, DAKTA integrating LoRA with a Kolmogorov-Arnold classifier, and SO-LoRA enforcing subspace separation via a fixed orthogonal basis and group sparsity.

The three research questions that structure the investigation are: (Q1) how can prototype augmentation mitigate forgetting while preserving privacy; (Q2) to what extent can selective knowledge transfer reduce information redundancy without exemplar storage; and (Q3) how can dynamic parameter allocation simultaneously prevent forgetting, limit computational overhead, and enable cross-task knowledge sharing. The correspond-

ing objectives are to develop and validate novel methods in each of these categories, with contributions published across seven peer-reviewed venues. Figure 3 illustrates the dissertation's overall structure.

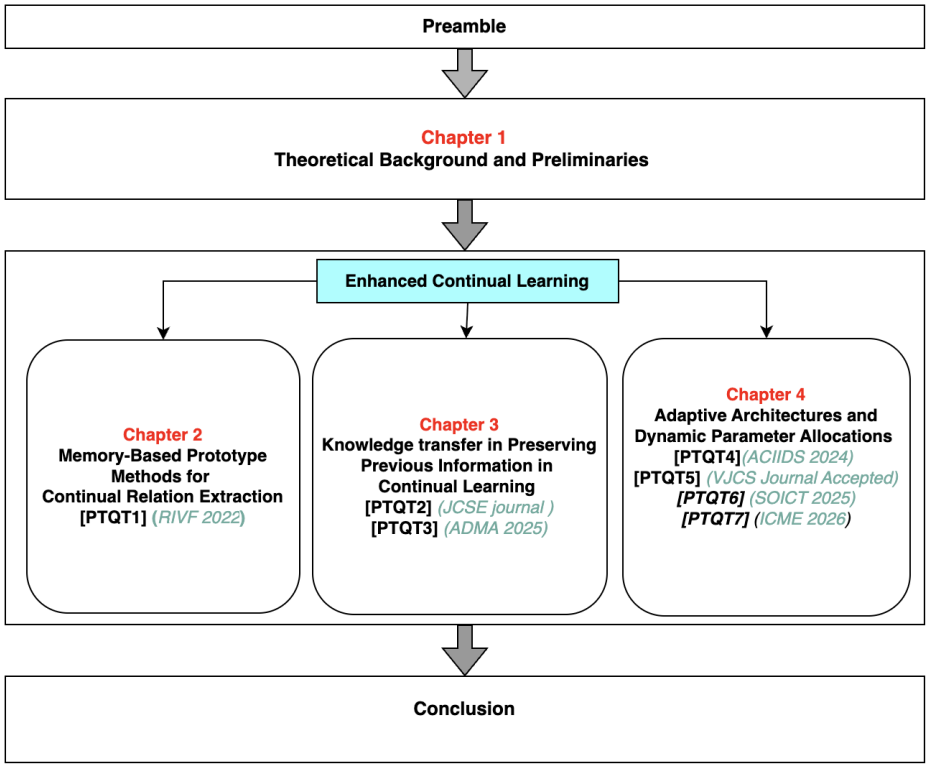


Figure 3: Dissertation outline and mapping between chapters and publications.

Chapter 1

Theoretical Background and Preliminaries

Continual Learning (CL), also referred to as lifelong learning or incremental learning, is a machine learning paradigm in which a model is trained sequentially on a series of tasks, aiming to perform as if all data had been observed simultaneously. Through the lens of deep learning, CL forms the foundation for intelligent systems that adapt to changing environments, mirroring the human capacity to continuously acquire, update, and utilize knowledge. The core obstacle in this paradigm is **catastrophic forgetting**: when a model adapts to a new task, updates to the shared parameters degrade its performance on previously learned tasks. This conflict between retaining old knowledge and acquiring new information is formally known as the **stability–plasticity dilemma**, and it drives the main research challenges of the field.

Two foundational definitions shape the theoretical landscape of CL. Thrun (1996) first proposed that a system completing N tasks should leverage the accumulated knowledge to benefit the $(N+1)$-th task, without retaining the raw data of earlier ones. Chen and Liu (2018) later extended this formulation to emphasize three essential properties: continuous learning across a sequence of tasks, maintenance of a Knowledge Base (KB) that accumulates and preserves learned information, and the ability to exploit past knowledge to improve future performance. Unlike the earlier definition, Chen and Liu explicitly formalise the preservation of all accumulated information as a fundamental goal, rather than merely optimizing for the most recent task.

Beyond catastrophic forgetting, CL must contend with several interconnected chal-

lenges. **Knowledge transfer** requires careful consideration of what knowledge to share, when to transfer it, and how to implement the transfer without introducing negative interference. **Task interference** arises when representations learned for one task impair performance on others, demanding a delicate balance between shared and task-specific parameters. **Limited memory capacity** constrains the amount of past information that can be retained, while **distribution drift** forces models to adapt to gradual or abrupt changes in data statistics. Finally, **sample efficiency** requires that new skills be acquired from limited examples, rather than through extensive retraining.

Depending on how task identity and label spaces are structured, eight canonical scenarios have been identified. **Instance-Incremental Learning (IIL)** involves a single task whose samples arrive in a sequential stream. **Domain-Incremental Learning (DIL)** preserves the label space while shifting the input distribution across domains, and task identity is not required at test time. **Task-Incremental Learning (TIL)** introduces disjoint label sets with the task identity explicitly available at inference, making it the simplest setting. **Class-Incremental Learning (CIL)**, the most studied scenario, shares the same disjoint-label structure but withholds the task identity at test time, requiring unified classification across all observed classes. **Task-Free (TFCL)** and **Online (OCL)** settings further remove task boundaries or restrict training to a single pass over each sample. The **Blurred Boundary (BBCL)** scenario models gradual transitions with overlapping class distributions, while **Continual Pre-training (CPT)** targets the sequential updating of large foundation models with new domain-specific corpora.

Five families of methods have been developed to address these challenges. **Regularization-based** approaches augment the loss with a penalty that slows learning for parameters deemed important to previous tasks, as in Elastic Weight Consolidation (EWC), Synaptic Intelligence (SI), and Learning without Forgetting (LwF). **Replay-based** methods store or synthesise past samples that are revisited during subsequent training (e.g., iCaRL, Dark Experience Replay, Gradient Episodic Memory), at the cost of memory overhead and potential privacy violations. **Optimization-based** approaches directly modify the training procedure through gradient projection or meta-learning strategies. **Representation-based** methods focus on learning transferable, drift-resistant feature spaces via contrastive or self-supervised pre-training, and prototype maintenance. Finally, **Architecture-based** methods allocate task-specific modules within a shared backbone, as in Progressive Neural Networks, PackNet, and adapter-based designs, allowing task-specific adaptation without excessive growth in total parameter count.

Performance in CL is assessed across three complementary dimensions. Overall effectiveness is measured by Average Accuracy (AA) and Average Incremental Accuracy (AIA). Memory stability is captured by the Forgetting Measure (FM) and Backward Transfer (BWT), where a negative BWT indicates that earlier task performance has degraded. Learning plasticity is evaluated through the Intransigence Measure (IM) and Forward Transfer (FWT), which quantify the model's ability to learn new tasks and benefit from previously acquired representations, respectively

Summary

A critical review of the existing literature reveals three open research gaps that motivate the contributions of this dissertation. First, replay-based methods that store real training data raise significant *privacy concerns*, calling for privacy-preserving alternatives through synthetic data generation. Second, current distillation-based knowledge transfer mechanisms are *inefficient*, motivating selective distillation strategies that leverage the representational power of pre-trained models. Third, architecture-based methods typically *lack cross-task knowledge sharing*, pointing toward dynamic architecture designs that balance task-specific adaptation with shared knowledge exploitation. These three gaps define the core research directions pursued in the subsequent chapters of this dissertation.

Chapter 2

Memory-Based Replay Methods for Continual Relation Extraction

Relation Extraction (RE) is the task of identifying and classifying semantic relations between entity pairs in text, and it serves as a core subtask of Information Extraction in applications such as question answering, sentiment analysis, and structured search. Conventional RE models operate over a fixed set of predefined relations, which limits their applicability in real-world scenarios where new relation types continuously emerge. **Continual Relation Extraction (CRE)** addresses this limitation by training models sequentially on new relation sets while preventing catastrophic forgetting of previously learned relations. Among the main families of solutions regularization-based, dynamic architecture, and memory-based methods memory-based approaches consistently achieve the strongest performance in NLP tasks by maintaining an episodic memory bank of past samples for rehearsal.

The most relevant prior work is **Consistent Representation Learning (CRL)**, which combines supervised contrastive learning with knowledge distillation to preserve stable relation embeddings across sequential tasks. Despite strong performance, CRL exhibits three key limitations: it depends entirely on storing real data, making it vulnerable to memory constraints and privacy concerns; its per-relation memory budget shrinks as more tasks are observed, risking overfitting on sparse exemplars; and it performs poorly on imbalanced datasets such as TACRED, where dominant classes suppress minority

relation learning during contrastive replay.

To address these limitations, this chapter proposes **ProtoAug**, which integrates prototype augmentation into the CRL framework to reduce reliance on stored real data. The method involves four stages: (1) a BERT-based encoder extracts relation representations optimised with supervised contrastive loss; (2) k-means clustering selects the most representative sample per relation for memory storage; (3) virtual samples are generated by perturbing relation prototypes with Gaussian noise, enriching the memory bank without storing additional real data; and (4) both real and synthetic samples are used for contrastive replay and knowledge distillation to preserve discriminability of previously learned relations. Figure 2.3 illustrates the overall architecture.

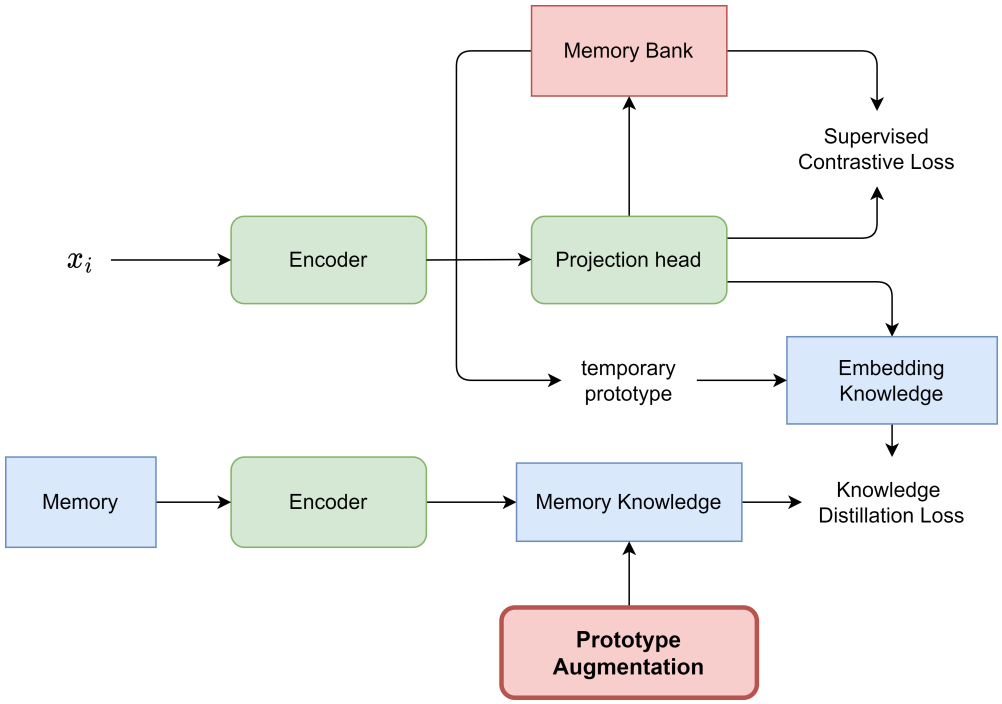


Figure 2.3: Architecture of the consistent representation module with prototype augmentation.

A further extension introduces an **Energy-based Latent space Alignment (ELI)** module to address representational drift across tasks. After training each task, an Energy-Based Model (EBM) is trained to assign low energy to latent representations from the previous model snapshot and high energy to those from the current model, using Langevin dynamics for approximate sampling. At inference time, the ELI algorithm iteratively moves each test sample's latent vector along the energy gradient toward the low-energy region, realigning shifted representations back toward the previous task's feature manifold without additional data storage. Figure 2.4 shows the full architecture incorporating both

prototype augmentation and ELI.

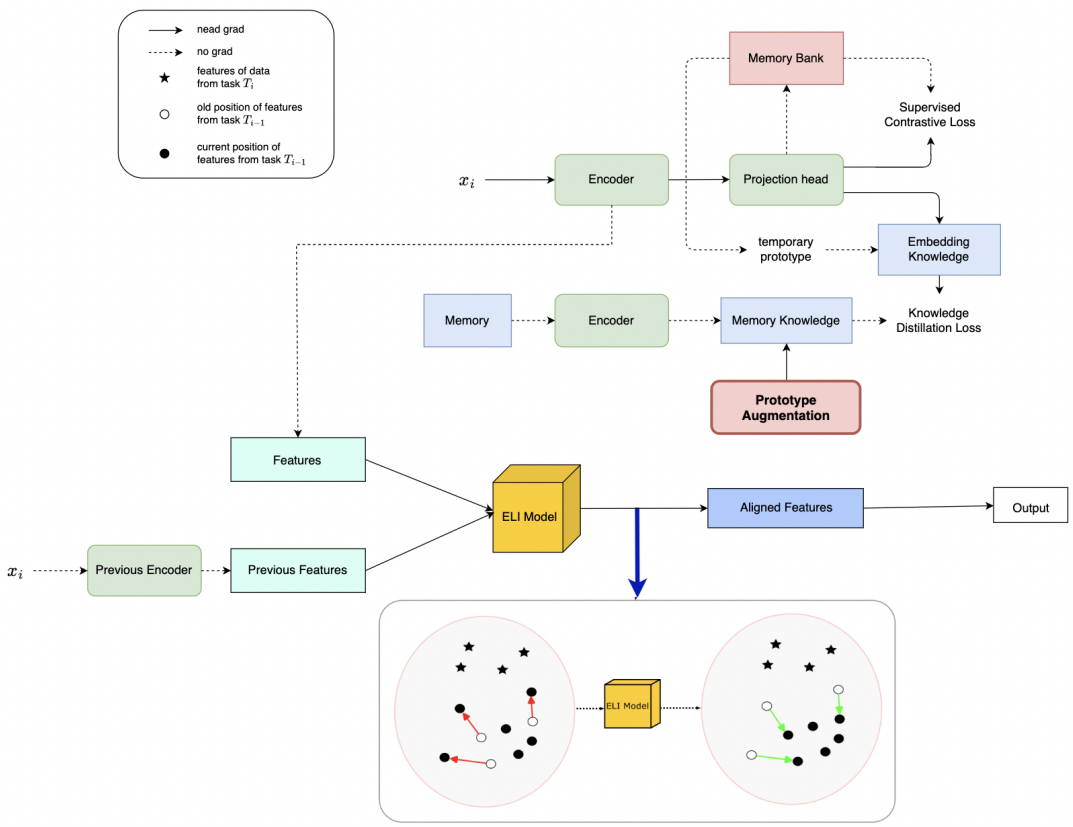


Figure 2.4: Overall architecture of the proposed continual relation extraction model combining prototype augmentation and energy-based latent alignment (ELI).

Experiments on the FewRel and TACRED benchmarks under the task-incremental setting across ten sequential tasks confirm the effectiveness of the proposed approach. Tables 2.1 and 2.2 report the accuracy of each method at every task step. On FewRel, the proposed model outperforms CRL on eight out of ten tasks by 1–2% and surpasses all other baselines across the full task sequence. On TACRED a more challenging, imbalanced dataset the gains are even more pronounced, with the model exceeding CRL by up to 4% on Task 4 and consistently outperforming EMAR+BERT with margins ranging from 1.1% to 11.1%. The improvements become more pronounced as the number of tasks increases, demonstrating long-term stability and superior mitigation of catastrophic forgetting.

Summary

In summary, this chapter demonstrates that replacing real stored samples with prototype-augmented synthetic data is not only sufficient but beneficial for continual relation extraction. The combination of k-means-based sample selection, Gaussian

Table 2.1: Accuracy (%) of each model across all tasks on the **FewRel** dataset. Bold = best; underline = second best.

Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
EA-EMR	89.0	69.0	59.1	54.2	47.8	46.1	43.1	40.7	38.6	35.2
EMAR	88.5	73.2	66.6	63.8	55.8	54.3	52.9	50.9	48.8	46.3
CML	91.2	74.8	68.2	58.2	53.7	50.4	47.8	44.4	43.1	39.7
EMAR+BERT	**98.8**	89.1	89.5	85.7	83.6	84.8	79.3	80.0	77.1	73.8
RP-CRE	97.9	92.7	91.6	89.2	88.4	86.8	85.1	84.1	82.2	81.5
CRL	98.2	94.6	92.5	90.5	89.4	87.9	86.9	85.6	84.5	83.1
Ours	98.2	**94.7**	**93.9**	**92.2**	**90.1**	**89.6**	**88.6**	**87.1**	**86.1**	**84.6**

prototype augmentation, and energy-based latent alignment provides a privacy-preserving, memory-efficient alternative to existing rehearsal methods, setting the stage for the subsequent chapter, which investigates knowledge distillation approaches that eliminate the need for any stored data entirely.

Table 2.2: Accuracy (%) of each model across all tasks on the **TACRED** dataset. Bold = best; underline = second best.

Model	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
EA-EMR	47.5	40.1	38.3	29.9	24.0	27.3	26.9	25.8	22.9	19.8
EMAR	73.6	57.0	48.3	42.3	37.7	34.0	32.6	30.0	27.6	25.1
CML	57.2	51.4	41.3	39.3	35.9	28.9	27.3	26.9	24.8	23.4
EMAR+BERT	96.6	85.7	81.0	78.6	73.9	72.3	71.7	72.2	72.6	71.0
RP-CRE	97.6	90.6	86.1	82.4	79.8	77.2	75.1	73.7	72.4	72.4
CRL	<u>97.7</u>	<u>93.2</u>	<u>89.8</u>	<u>84.7</u>	<u>84.1</u>	<u>81.3</u>	<u>80.2</u>	<u>79.1</u>	<u>79.0</u>	<u>78.0</u>
Ours	**97.8**	**95.2**	**90.6**	**88.2**	**84.6**	**81.7**	**82.8**	**80.5**	**80.0**	**79.6**

Chapter 3

Knowledge Transfer in Preserving Previous Information in Continual Learning

This chapter investigates knowledge transfer as a mechanism for mitigating catastrophic forgetting in continual learning, without relying on the storage of real exemplars from previous tasks. Two complementary strategies are developed: the first transfers knowledge through contrastive ensemble distillation within a **Task-Incremental Learning (TIL)** framework applied to Aspect Sentiment Classification (ASC); the second retains knowledge entirely within a frozen pre-trained backbone through closed-form analytic learning under a **Class-Incremental Learning (CIL) /** Generalized CIL (GCIL) setting. Together, they cover two qualitatively different routes to stability gradient-based distillation and gradient-free analytic computation both achieving exemplar-free continual learning.

Part 1: Knowledge Distillation with Contrastive Learning (TIL).

The first approach addresses the stability-plasticity trade-off by combining architectural isolation with contrastive learning. A **Continual Learning Plugin (CL-plugin)** is injected into BERT at two locations, replacing the standard adapter. The CL-plugin contains two sub-modules: a **Knowledge Sharing Sub-Module (KSM)**, implemented as capsule layers with dynamic routing to cluster shared knowledge across tasks, and a **Task-Specific Sub-Module (TSM)**, implemented as fully-connected layers with neuron-level soft masks that protect task-critical parameters from gradient updates during subsequent

tasks. Figure 3.3 illustrates the overall architecture, and Figure 3.4 details the CL-plugin structure.

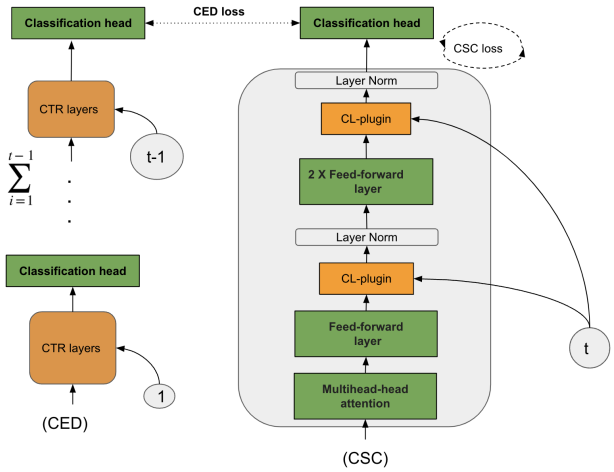


Figure 3.3: Overall architecture of the proposed CL-plugin model.

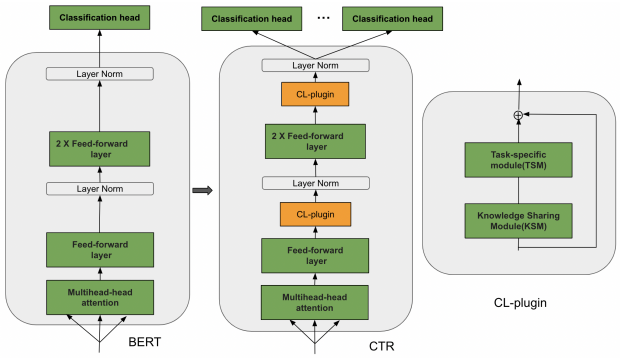


Figure 3.4: General architecture of BERT (left), the CTR system (middle), and the CL-plugin (right) integrated into BERT.

Two contrastive losses augment the standard cross-entropy objective. The **Contrastive Ensemble Distillation (CED) loss** distils knowledge from all previous task models into the current task by treating the logit pair of the same input from an old and the current model as a positive pair, and all other pairs as negatives. Summing over all $t-1$ previous teacher models yields the full CED loss. The **Contrastive Supervised Learning of Current Task (CSC) loss** maximises intra-class similarity and minimises inter-class similarity within the current task's batch, operating on the masked TSM output before the classification head. The three objectives are combined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CED}} + \mathcal{L}_{\text{CSC}}.$$

Experiments are conducted on 19 ASC tasks drawn from four benchmark datasets (HL5Domains, Liu3Domains, Ding9Domains, SemEval14), summarised in Table 3.1.

Results averaged over five random task orders are reported in Table 3.2. The proposed model achieves an Average Accuracy of 88.50% and the highest Macro-F1 of 82.35% among all evaluated continual learning methods in the Adapter-BERT family. CTR, the closest competitor, scores 88.21% accuracy and 81.19% Macro-F1 under the same randomised ordering, falling below both metrics. LAMOL reports a higher accuracy (88.91%) but relies on a GPT-2 generative backbone for pseudo-rehearsal, which incurs substantially higher computational cost and represents a fundamentally different strategy.

Table 3.1: Number of training/validation/test sentences per task across all four datasets.

Data source	**Task / domain**	**Train**	**Valid.**	**Test**
Liu3Domains	Speaker	352	44	44
	Router	245	31	31
	Computer	283	35	36
HL5Domains	Nokia6610	271	34	34
	Nikon4300	162	20	21
	Creative	677	85	85
	CanonG3	228	29	29
	ApexAD	343	43	43
Ding9Domains	CanonD500	118	15	15
	Canon100	175	22	22
	Diaper	191	24	24
	Hitachi	212	26	27
	Ipod	153	19	20
	Linksys	176	22	23
	MicroMP3	484	61	61
	Nokia6600	362	45	46
	Norton	194	24	25
SemEval14	Restaurant	3452	150	1120
	Laptop	2163	150	638

The ablation study in Table 3.3 confirms that every component contributes positively: removing the CL-plugin alone degrades accuracy by 2.67% and Macro-F1 by

2.60%, while removing both CED and CSC causes a 2.02% drop in accuracy and a 3.03% drop in Macro-F1.

Part 2: Exemplar-Free Knowledge Retention via Analytic Learning (GCIL).

The second approach eliminates gradient-based training altogether. Motivated by the limitations of Generalized Analytic Continual Learning (GACL), which applies a linear classifier directly on raw feature vectors and struggles with non-linearly separable distributions, this work proposes **ZeroRFF** a method that integrates **Random Fourier Features (RFF)** into the analytic continual learning framework. Figure 3.8 shows the overall architecture.

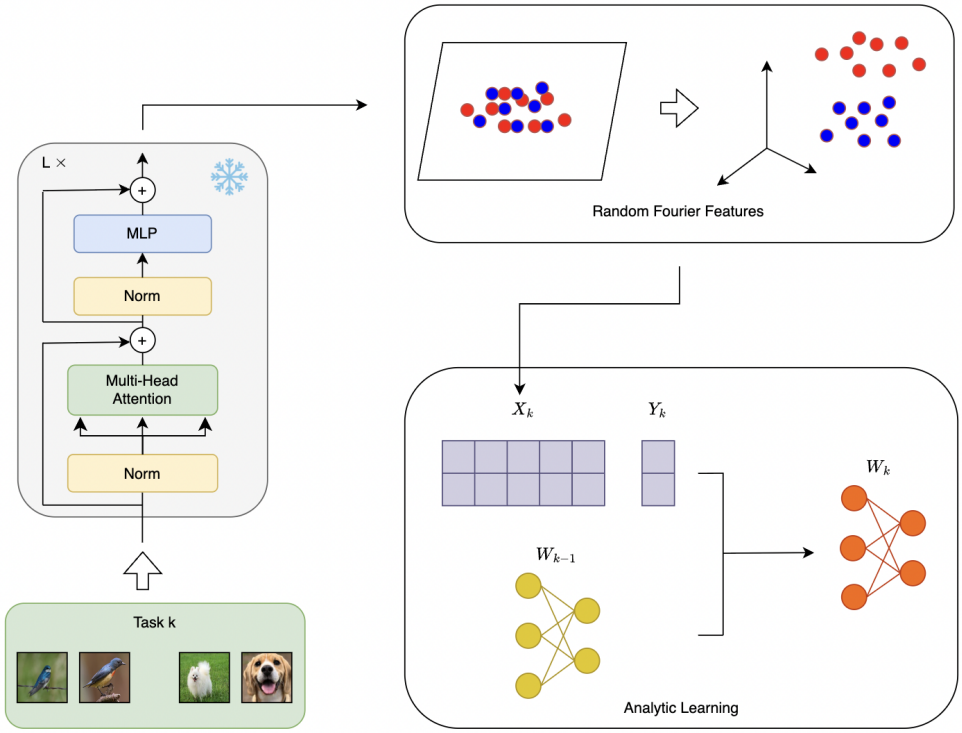


Figure 3.8: Overall architecture of ZeroRFF for Generalized Class-Incremental Learning.

ZeroRFF operates in four sequential phases per task. First, a frozen DeiT-S/16 backbone (pre-trained on ImageNet) extracts high-level feature representations $X_k^{(E)}$ from the input images, ensuring consistent representations across all tasks without any parameter updates. Second, the extracted features are transformed by the RFF module into a D=5000-dimensional space via $z(x) = \sqrt{2/D}\cos(\omega^\top x + \beta)$, where $\omega \sim \mathcal{N}(0, \sigma^{-2}I)$ and $\beta \sim \text{Uniform}(0, 2\pi)$, approximating the implicit feature space of an RBF kernel and rendering the data approximately linearly separable. Third, the analytic learning component computes closed-form optimal parameters by recursively updating a cor-

relation matrix R_k via the Sherman-Morrison formula and deriving the weight matrix $\hat{W}_{\text{FCN}}^{(k)}$ analytically, with no exemplar storage and no iterative optimisation. Fourth, a task management loop increments the task counter and loads the next dataset until all K tasks have been processed, after which the final weight matrix is returned for inference. Algorithm 1 summarises the full procedure.

Algorithm 1 ZeroRFF Training

Require: Tasks $\mathcal{D}_1^{\text{train}}, \ldots, \mathcal{D}_K^{\text{train}}$; frozen backbone weights Θ_{backbone}
Ensure: $\hat{W}_{\text{FCN}}^{(K)}$

Init: $R_0 \leftarrow \gamma I, \quad \hat{W}_{\text{FCN}}^{(0)} \leftarrow 0$
for $k = 1$ **to** K **do**
  $X_k^{(E)} \leftarrow f_{\text{backbone}}(X_k^{\text{train}}, \Theta_{\text{backbone}})$
  $X_k^{(B)} \leftarrow \text{rff}(X_k^{(E)})$
  $R_k \leftarrow R_{k-1} - R_{k-1} X_k^{(B)\top} \left(I + X_k^{(B)} R_{k-1} X_k^{(B)\top}\right)^{-1} X_k^{(B)} R_{k-1}$
  $\hat{W}^{(k)} \leftarrow \hat{W}^{(k-1)} - R_{k-1} X_k^{(B)\top} X_k^{(B)} \hat{W}^{(k-1)} + R_k X_k^{(B)\top} Y_k^{\text{train}}$
end for
return $\hat{W}_{\text{FCN}}^{(K)}$

Experiments on CIFAR-100 under the Si-Blurry GCIL setting (disjoint class ratio 50%, blurry sample ratio 10%, five seeds) are reported in Table 3.4. ZeroRFF achieves A_{AUC}=60.48, A_{AVG}=58.55, and A_{Last}=72.44, improving over the best memory-free baseline GACL by approximately 2% on all three metrics. Notably, ZeroRFF also surpasses most memory-based methods using a buffer of 500 samples, and approaches the performance of the memory-based MVP-R (memory size 2000) while requiring no stored data whatsoever.

Summary

In summary, this chapter presents two novel exemplar-free knowledge transfer strategies that occupy distinct positions on the continual learning design space. The CL-plugin approach is best suited to scenarios with rich inter-task semantic overlap, where contrastive distillation across task-specific model snapshots adds meaningful signal. ZeroRFF is preferable when a strong frozen backbone is available and training efficiency or data privacy is paramount, offering closed-form guarantees and constant memory footprint regardless of the number of tasks. Both methods advance the goal of preserving previously acquired knowledge without the data storage and privacy burdens associated with exemplar replay.

Table 3.2: Average Accuracy (Acc.) and Macro-F1 (MF1) over five random sequences of 19 tasks. Bold = best in category.

Scenario	**Backbone**	**Model**	**Acc.(%)**	**MF1(%)**
Non-CL (SDL)	BERT	MTL	91.91	88.11
	BERT	SDL	85.84	76.35
	BERT (Frozen)	SDL	78.14	58.13
	Adapter-BERT	SDL	85.96	78.07
	W2V	SDL	77.01	51.89
Continual Learning	BERT (Frozen)	NFH	85.51	76.64
	Adapter-BERT	NFH	85.51	76.64
	BERT (Frozen)	EWC	86.37	74.52
		OWM	87.02	79.31
		HAT	86.74	78.16
		A-GEM	86.06	78.44
		DER++	84.27	75.08
		KAN	85.49	77.38
		SRK	84.76	78.52
	Adapter-BERT	HAT	86.14	78.52
		BCL	88.29	81.40
		LAMOL	**88.91**	80.59
		CTR	89.47	83.62
		CTR*	88.21	81.19
		Ours	88.50	**82.35**

Table 3.3: Ablation study results.

Model variant	**Acc.(%)**	**MF1(%)**
Ours (full model)	**88.50**	**82.35**
Without CL-plugin	85.83	79.75
Without CSC + CED	86.48	79.32
Without CSC	87.52	79.93
Without CED	87.98	80.10

Table 3.4: Experimental results on CIFAR-100 under the Si-Blurry setting (%). Bold = best memory-free method.

Memory	**Method**	A_{AUC}	A_{AVG}	A_{Last}
2000	EWC++	53.31 ± 1.70	50.95 ± 1.50	52.55 ± 0.71
	ER	56.17 ± 1.84	53.80 ± 1.46	55.60 ± 0.69
	RM	53.22 ± 1.82	52.99 ± 1.65	55.25 ± 0.61
	MVP-R	60.62 ± 1.03	57.58 ± 0.56	64.30 ± 0.29
500	EWC++	48.31 ± 1.81	44.56 ± 0.40	40.52 ± 0.83
	ER	51.59 ± 1.91	48.03 ± 0.80	44.09 ± 0.80
	RM	41.07 ± 1.30	38.10 ± 0.59	32.66 ± 0.34
	MVP-R	56.20 ± 1.44	53.61 ± 0.04	55.35 ± 0.43
0 (memory-free)	LwF	40.71 ± 1.21	38.49 ± 0.56	27.03 ± 2.92
	L2P	42.68 ± 2.70	39.89 ± 0.45	28.59 ± 3.34
	DualPrompt	41.34 ± 2.59	38.59 ± 0.82	22.74 ± 3.40
	MVP	45.07 ± 1.24	44.93 ± 0.54	39.94 ± 0.47
	SLDA	53.00 ± 3.85	50.09 ± 2.77	61.79 ± 3.81
	GACL	57.99 ± 2.46	56.24 ± 3.12	70.31 ± 0.06
	ZeroRFF (Ours)	$\mathbf{60.48 \pm 4.33}$	$\mathbf{58.55 \pm 6.05}$	$\mathbf{72.44 \pm 0.13}$

Chapter 4

Adaptive Architectures and Dynamic Parameter Allocations

Architecture-based continual learning addresses catastrophic forgetting by dynamically allocating dedicated model capacity for each new task, creating isolated or partially shared parameter spaces that minimize interference between sequential training phases. Unlike regularization methods that constrain updates or replay methods that store past data, these approaches modify the model structure itself. This chapter presents three complementary contributions under the **Class-Incremental Learning (CIL)** scenario, organized across two methodological lineages: a ResNet-18 backbone trained from scratch for online CIL benchmarks, and a pre-trained ViT-B/16 backbone for parameter-efficient fine-tuning (PEFT) settings.

Part 1: Multi-Head Approach with Online Prototype Equilibrium, Adaptive Prototypical Feedback, and Energy-Based Latent Alignment.

The first contribution augments the Online Prototype Learning (OnPro) framework to address two limitations: OnPro uses a single shared classification head that conflates within-task prediction (WP) and task-ID prediction (TP), and treats all class pairs uniformly regardless of their susceptibility to confusion. The proposed method, illustrated in Figure 4.2, introduces two complementary mechanisms. **Online Prototype Equilibrium (OPE)** maintains discriminative representations by simultaneously pulling together online prototypes of the same class and pushing apart those of different classes, balancing new-class learning with preservation of previously seen classes through:

$$\mathcal{L}_{\text{OPE}} = \mathcal{L}_{\text{pro}}^{\text{new}}(\mathcal{P}, \hat{\mathcal{P}}) + \mathcal{L}_{\text{pro}}^{\text{seen}}(\mathcal{P}^b, \hat{\mathcal{P}}^b).$$

Adaptive Prototypical Feedback (APF) identifies class pairs prone to misclassification by computing distances between stored prototypes via a symmetric Gaussian kernel, then oversamples confusing pairs during memory replay to sharpen their decision boundaries. The two-stage sampling strategy selects $\alpha \cdot m$ samples proportionally to confusion probability and randomly fills the remainder to preserve global balance.

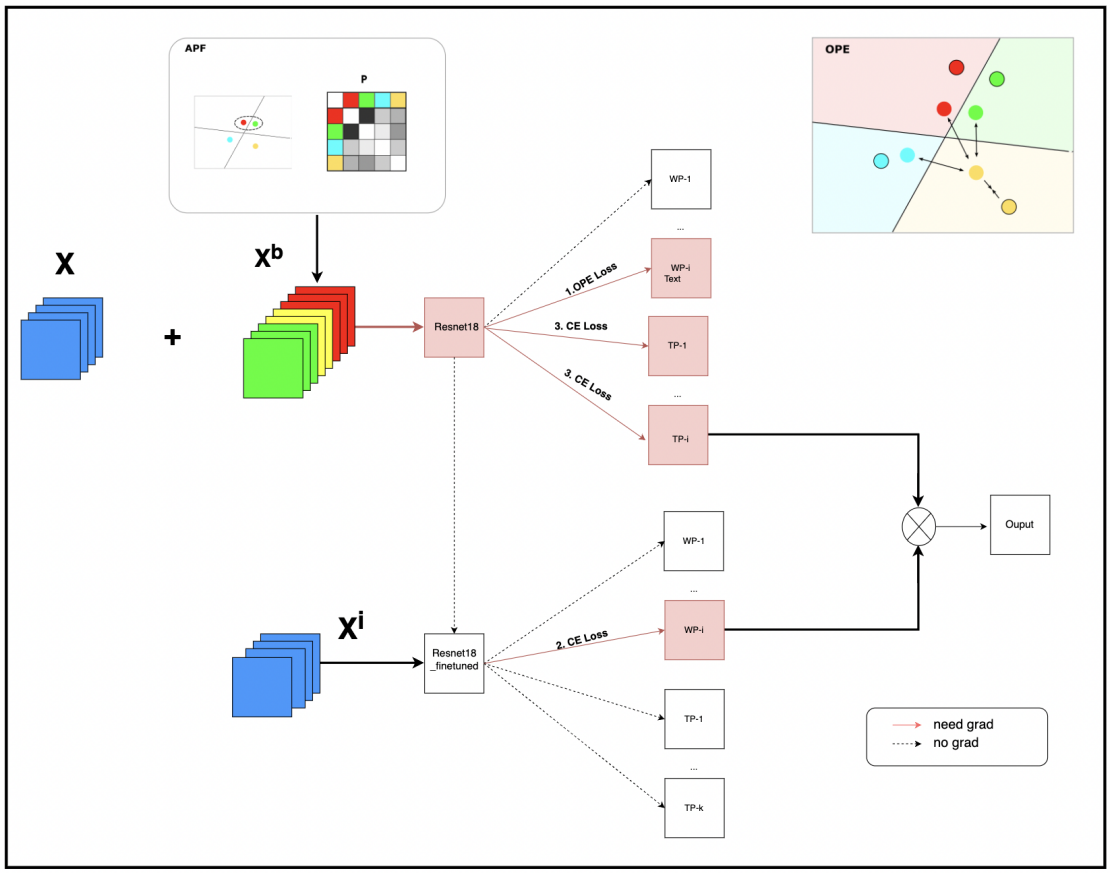


Figure 4.2: Overall architecture of the multi-head model with OPE, APF, and dual classification heads (WP and OOD).

The classification layer is restructured into two specialized heads: an **OOD head** that estimates task probability by distinguishing in-distribution samples of the current task from out-of-distribution samples in the memory buffer, and a **WP head** for fine-grained intra-task prediction. The final CIL prediction is decomposed as $P(X_{k,j}|x) = P(X_{k,j}|x,k)\,P(X_k|x)$, combining the outputs of both heads. Training proceeds in three steps: (1) joint training of the feature extractor and OOD head; (2) fine-tuning of the WP head with the feature extractor frozen; (3) fine-tuning of all previous OOD heads on replay data to reduce early-task bias.

An **Energy-based Latent Alignment (ELI)** module is further integrated to counteract representational drift. After each task, an Energy-Based Model (EBM) is trained

to assign low energy to latent representations from the previous encoder snapshot and high energy to those from the current encoder, using Langevin dynamics for sampling. At inference, ELI iteratively shifts each test sample's latent vector along the energy gradient back toward the previous task's feature manifold, correcting representational shift without any additional data storage.

Tables 4.1 and 4.2 report average accuracy and forgetting on CIFAR-10 and CIFAR-100. The proposed method (ModifiedOnPro) consistently outperforms OnPro in accuracy across all buffer sizes by up to 11.2 points on CIFAR-10 ($\mathcal{M}$=100, from 52.4% to 63.6%) and up to 6.4 points on CIFAR-100 ($\mathcal{M}$=500, from 20.7% to 27.1%). On the CCH5000 medical imaging benchmark (Table 4.5), the method achieves the best forgetting score of 2.9 in the two-classes-per-task setting, a 27.5% improvement over LO2LN.

Table 4.1: Average accuracy (%) with different memory buffer sizes $\mathcal{M}$. Bold = best.

	CIFAR-10			**CIFAR-100**		
Method	$\mathcal{M}$=100	$\mathcal{M}$=200	$\mathcal{M}$=500	$\mathcal{M}$=500	$\mathcal{M}$=1000	$\mathcal{M}$=2000
iCaRL	31.0	33.9	42.0	12.8	16.5	17.6
DER++	31.5	39.7	50.9	16.0	21.4	23.9
SCR	40.2	48.5	59.1	19.3	26.5	32.7
OCM	47.5	59.6	70.1	19.7	27.5	34.4
OnPro	52.4	63.0	67.7	20.7	27.2	34.9
Ours	**63.6**	**69.3**	**72.8**	**27.1**	**30.3**	**36.3**

Part 2: DAKTA Directional Kolmogorov-Arnold Classifier for Task Arithmetic.

The second contribution, DAKTA (Figure 4.6), builds upon the Incremental Task Arithmetic with LoRA (ITA-LoRA) framework and addresses its key limitation: a standard linear classification head that applies uniform decision boundaries across all tasks. DAKTA replaces the linear head with a **Kolmogorov-Arnold Classifier (KAC)**, which computes class activations as weighted sums of learnable Gaussian Radial Basis Functions (RBFs), providing locally sensitive, channel-selective boundaries that significantly reduce inter-task interference. A **Directional Vector Fusion** mechanism further combines RBF-based logits with a cosine similarity term:

$$l_c(x) = l_c^{\text{init}}(x) + \gamma_c \cdot \cos(x, W_c),$$

Table 4.2: Average forgetting (%) with different memory buffer sizes $\mathcal{M}$. Lower is better.

	CIFAR-10			CIFAR-100		
Method	$\mathcal{M}$=100	$\mathcal{M}$=200	$\mathcal{M}$=500	$\mathcal{M}$=500	$\mathcal{M}$=1000	$\mathcal{M}$=2000
iCaRL	52.7	49.3	38.3	16.5	11.2	10.4
DER++	57.8	46.7	33.6	41.0	34.8	33.2
SCR	43.2	35.5	24.1	29.3	20.4	11.5
OCM	35.5	23.9	13.5	18.3	15.2	10.8
OnPro	24.2	16.5	13.7	17.8	14.9	**10.0**
Ours	**23.2**	18.3	**13.5**	19.2	**14.5**	11.2

capturing both local similarity patterns and global directional alignment for more robust class discrimination. **Second-order Fisher Information Matrix (FIM) regularization** is applied to both the ViT backbone and the KAC head, penalizing deviations of important parameters from their pre-trained values:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{reg}}^{\text{backbone}} + \mathcal{L}_{\text{reg}}^{\text{cls}}.$$

After training each task, the parameter changes are stored as task vectors and composed at inference by averaging, inheriting the stability guarantees of the task arithmetic framework.

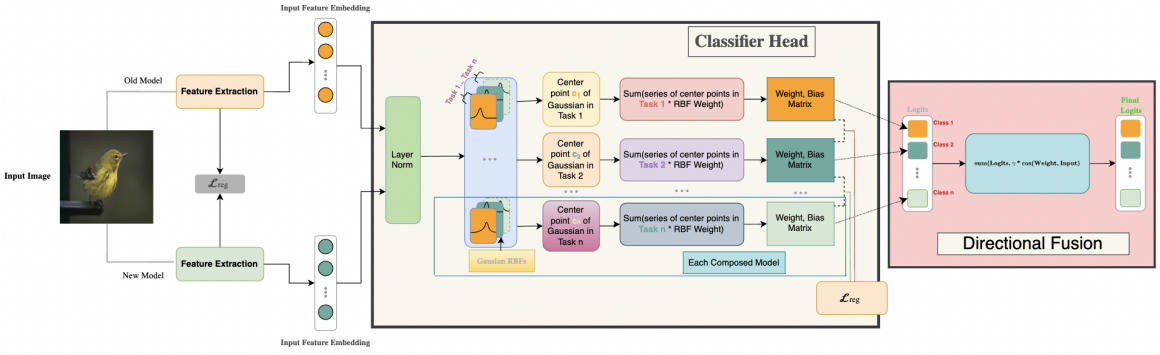


Figure 4.6: Pipeline of the DAKTA framework integrating ViT+LoRA backbone, Kolmogorov-Arnold Classifier (KAC), Incremental Task Arithmetic (ITA), and Fisher-based regularization.

Table 4.6 reports final accuracy across six benchmarks. DAKTA achieves the best

Table 4.5: Comparative results on CCH5000 (medical imaging) with one and two classes per task. Bold = best.

Method	1 class/task		2 classes/task	
	Acc	Fgt	Acc	Fgt
iCaRL	91.1	9.0	93.0	6.8
UCIR	92.0	5.5	93.9	4.4
PODNet	89.2	6.0	92.0	5.2
DER	91.0	5.6	93.0	6.4
LO2LN	**94.5**	3.9	**94.6**	4.0
OnPro	89.5	**3.6**	92.8	3.8
Ours	91.6	4.6	**94.6**	**2.9**

results on five out of six datasets, surpassing ITA-LoRA by up to 2.04 points (CIFAR-100) and by 2.04 points on CUB-200, a fine-grained recognition benchmark where the KAC's local sensitivity matters most. The forgetting comparison further confirms DAKTA's superiority: it achieves the lowest forgetting score on all six benchmarks, reducing forgetting by up to 0.94 points relative to ITA-LoRA on CIFAR-100. The ablation study (Table 4.7) validates that both the KAC head and the FIM regularization are necessary, with the full model consistently outperforming variants that remove either component.

Table 4.6: Final accuracy (%) comparison with SOTA methods. Bold = best.

Model	**C-100**	**CUB**	**Caltech**	**MIT**	**RESISC**	**CropDis.**
CODA	86.48	78.54	90.57	77.73	69.50	74.65
SEED	83.39	85.35	90.04	86.34	74.81	92.77
InfLoRA	87.17	79.14	90.53	79.14	79.92	89.05
DER++	83.02	76.10	86.11	68.88	67.23	**98.77**
ITA-LoRA	89.96	85.55	92.65	86.60	82.00	95.85
ITA-$(IA)^3$	90.66	85.67	92.67	84.74	83.73	95.41
DAKTA (Ours)	**91.21**	**87.59**	**93.61**	**87.97**	**85.82**	97.75

Table 4.7: Ablation study on second-order regularization and KAC in DAKTA. Bold = best.

Variant	C-100	CUB	Cal.	MIT	RES.	Crop.
w/o Reg.	85.24	80.57	89.10	82.25	79.38	90.84
w/o KAC	89.91	84.12	92.35	85.24	82.25	94.82
Full Model	**91.21**	**87.59**	**93.61**	**87.97**	**85.82**	**97.75**

Part 3: SO-LoRA Sparse Orthogonal LoRA for Continual Learning.

The third contribution, SO-LoRA (Figure 4.8), addresses the fundamental tension between per-task adapter isolation (stable but $\mathcal{O}(T)$ memory) and shared adapters (constant memory but prone to interference). SO-LoRA reconceptualizes the problem geometrically: it maintains a fixed orthonormal basis $A \in \mathbb{R}^{d\times r}$ (initialized once via QR decomposition and frozen) and restricts each task to learn a sparse coefficient matrix $B^{(t)} \in \mathbb{R}^{r\times k}$:

$$W^{(t)} = W_0 + AB^{(t)}.$$

Since $A^\top A = I_r$, interference between tasks i and j is provably bounded by $|\langle \Delta W_i, \Delta W_j \rangle_F| \leq \|\Delta B_i\|_{2,1}\|\Delta B_j\|_{2,1}$, and is directly controlled by minimizing the $\ell_{2,1}$ (group lasso) norm on $B^{(t)}$, which pushes each task to activate only a sparse subset of basis directions. A **gradient-aware soft mask** $M \in \mathbb{R}^r$ further protects historically important rank dimensions: after each task, row-gradient norms are aggregated via an exponential moving average, and during the next task's training, gradients for highly protected but currently irrelevant rows are suppressed by the scaling factor $S_j = 1 - M_{t-1,j}(1 - \phi_j)$.

Tables 4.9 and 4.10 summarise the main results. On ImageNet-R, SO-LoRA ranks first at all task granularities (5/10/20 tasks), with only a small accuracy drop from 79.78% to 78.15% as the sequence grows from 5 to 20 tasks, whereas O-LoRA degrades sharply from 72.76% to 60.98%. On ImageNet-A (severe domain shift), SO-LoRA achieves 60.27% final accuracy, nearly 5 points above SD-LoRA, highlighting the importance of the gradient-aware mask when current data depart from the pre-training distribution. On CUB-200 (fine-grained recognition), SO-LoRA attains the highest accuracy (81.58%), confirming that sparse orthogonal compositions capture subtle cues without amplifying interference. Figure 4.9 visualizes the accuracy trajectory across sequence lengths.

The ablation study (Table 4.11) confirms that the gradient-aware mask provides the

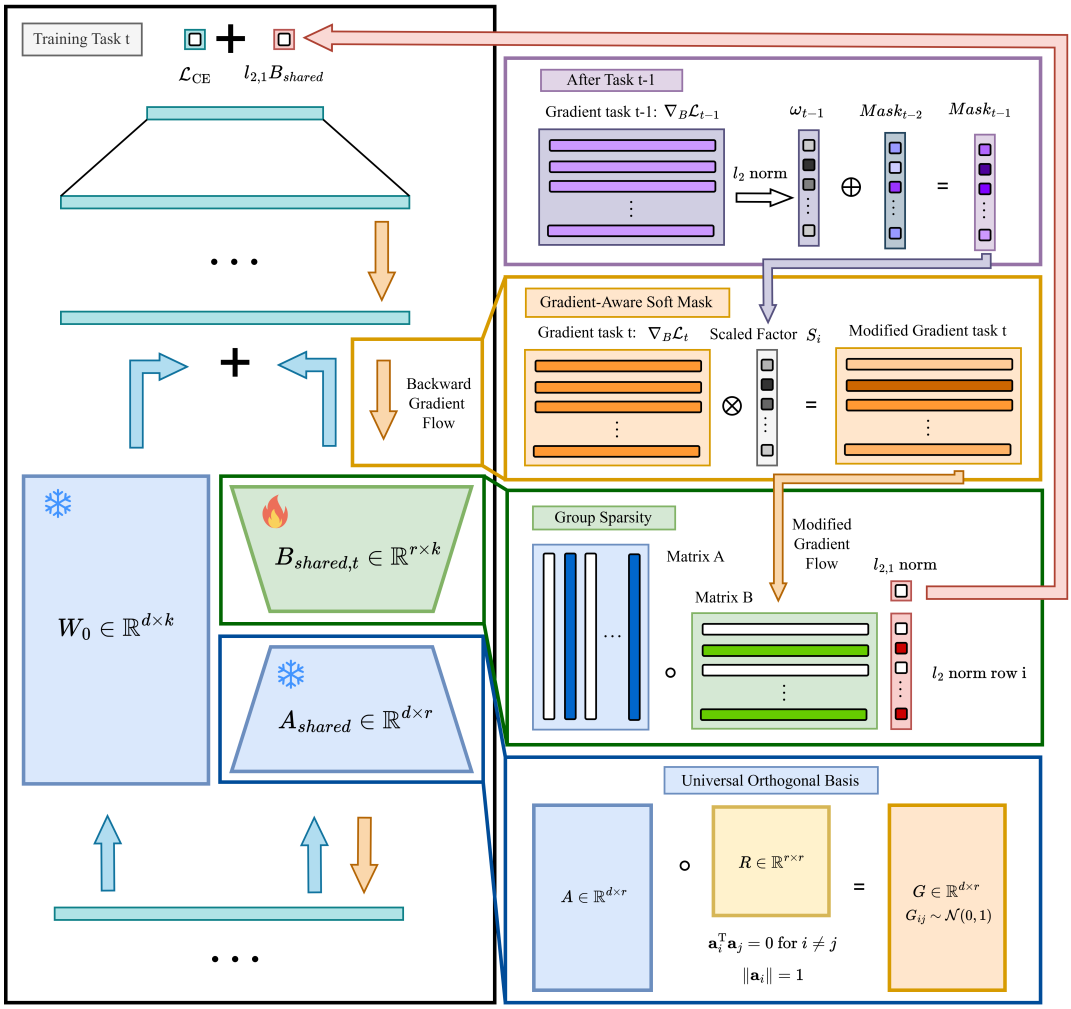


Figure 4.8: SO-LoRA represents adaptation as sparse coefficients $B^{(t)}$ over a frozen orthonormal basis A. Group sparsity limits overlap; the gradient-aware mask M protects historically important dimensions.

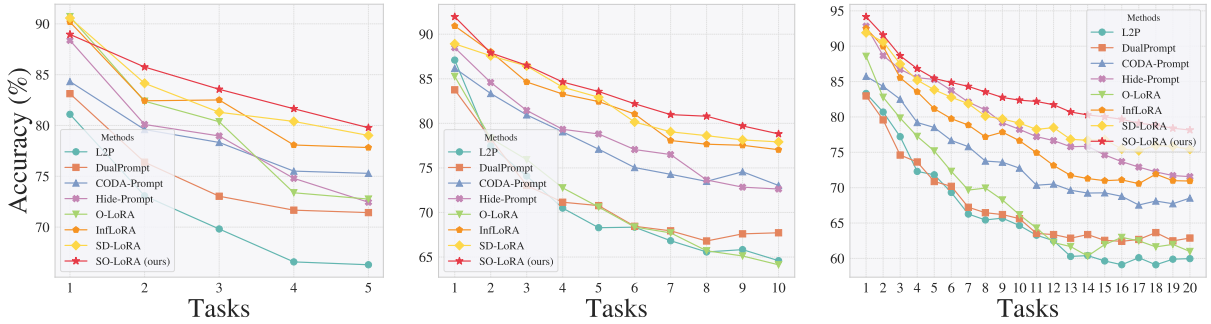
most consistent gains, especially under domain shift (ImageNet-A) and long horizons (20 tasks on ImageNet-R), while group sparsity adds complementary benefits particularly in out-of-domain settings.

Summary

In summary, this chapter presents three complementary architecture-based strategies that collectively address the key limitations of existing continual learning methods. The multi-head OPE/APF/ELI approach improves both feature representation and classification for online CIL by disentangling within-task and task-ID prediction and correcting latent drift. DAKTA replaces the conventional linear head with a locally sensitive Kolmogorov-Arnold classifier, extending Fisher-based regularization to the classification layer and achieving superior accuracy and lower forgetting across diverse fine-grained

Table 4.9: Performance on ImageNet-R under different task granularities. AAA = Average Anytime Accuracy. Bold = best.

Method	IN-R (5)		IN-R (10)		IN-R (20)	
	AAA↑	Acc↑	AAA↑	Acc↑	AAA↑	Acc↑
CODA	78.68	75.29	77.11	73.03	73.71	68.53
HiDe	77.42	72.45	76.84	72.63	75.88	71.56
O-LoRA	76.71	72.76	71.30	64.13	68.04	60.98
InfLoRA	81.63	77.83	80.24	77.04	76.78	70.94
SD-LoRA	82.85	79.02	81.05	77.92	80.62	75.43
SO-LoRA	**83.94**	**79.78**	**83.71**	**78.83**	**83.18**	**78.15**



(5 tasks) (10 tasks) (20 tasks)

Figure 4.9: Average accuracy on ImageNet-R across sequence lengths (5, 10, 20 tasks).

benchmarks. SO-LoRA provides a theoretically grounded, parameter-constant solution that geometrically isolates task representations within a fixed orthogonal basis and dynamically protects critical dimensions via a gradient-aware mask, demonstrating strong scalability across long task sequences and severe distribution shifts.

Table 4.10: Performance on ImageNet-A, CIFAR-100, and CUB-200 (10 tasks). Bold = best.

Method	IN-A		C100		CUB	
	AAA↑	Acc↑	AAA↑	Acc↑	AAA↑	Acc↑
CODA	56.30	45.94	91.37	86.82	82.36	72.41
HiDe	55.72	44.67	91.17	86.53	81.71	74.15
O-LoRA	54.17	44.46	89.96	82.51	79.66	70.12
InfLoRA	61.63	51.02	91.44	86.09	81.43	71.13
SD-LoRA	63.24	55.35	**91.71**	**87.28**	85.15	77.38
SO-LoRA	**67.89**	**60.27**	90.88	86.61	**87.71**	**81.58**

Table 4.11: Ablation on 10-task benchmarks. Impact of group sparsity and gradient-aware mask. Bold = best.

Variant	ImageNet-A		ImageNet-R (10)	
	AAA	Acc	AAA	Acc
SO-LoRA (full)	**67.89**	**60.27**	**83.71**	**78.83**
w/o sparsity	66.86	59.30	83.52	78.77
w/o mask	66.31	58.15	83.40	78.37

Conclusion and Future Work

This dissertation addresses catastrophic forgetting in continual learning through three progressive stages.

Stage 1 introduces privacy-preserving prototype replay: past class distributions are approximated using prototype augmentation with Gaussian noise, while an Energy-based Latent Alignment (ELI) module corrects representational drift after each task. Experiments on FewRel and TACRED confirm that this approach matches or surpasses methods that store real data.

Stage 2 eliminates stored data entirely. CL-plugin achieves state-of-the-art Macro-F1 on a 19-task aspect sentiment benchmark through selective knowledge distillation. ZeroRFF extends this to vision tasks by computing closed-form analytic classifiers over frozen backbone features, requiring no exemplar storage. Together, these methods confirm that exemplar-free continual learning is viable across both NLP and vision domains.

Stage 3 addresses capacity constraints that limit replay and distillation approaches. Four methods are proposed: (i) OPE/APF/ELI disentangles within-task and task-identity prediction, yielding the best forgetting score on CCH5000; (ii) DAKTA replaces the linear classifier with a Kolmogorov-Arnold Classifier with Fisher Information Matrix regularization, achieving consistent accuracy gains across six benchmarks; (iii) SO-LoRA fixes a universal orthogonal adapter basis and restricts each task to sparse coefficient compositions, maintaining a constant parameter footprint regardless of sequence length.

Taken together, the three stages form a context-sensitive answer to catastrophic forgetting: prototype replay when limited storage is permissible, knowledge distillation when data retention is infeasible, and adaptive architectures when task diversity makes a fixed parameter budget insufficient.

Limitations include the assumption of known task boundaries, lack of evaluation on very long task sequences, and coverage limited to text and image classification. Future

directions include a context-adaptive meta-framework that automatically selects among replay, distillation, and architecture expansion, as well as extensions to large language models and multi-modal settings.

List of Publications

[P1] **Quynh-Trang Pham Thi**, Anh-Cuong Pham, Ngoc-Huyen Ngo, and Duc-Trong Le. "Memory-Based Method using Prototype Augmentation for Continual Relation Extraction." In 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-6. IEEE, 2022. Scopus.

[P2] Thanh-Hai Dang, **Quynh-Trang Pham Thi**, Duc-Trong Le, and Tri-Thanh Nguyen. Contrastive Learning for Boosting Knowledge Transfer in Task-Incremental Continual Learning of Aspect Sentiment Classification Tasks. VNU Journal of Science: Computer Science and Communication Engineering 40, no. 1 (2024).

[P3] Duc-Hung Nguyen, Tri-Thanh Nguyen, Thanh-Hai Dang and **Quynh-Trang Pham Thi**. ZeroRFF: A Random Fourier Features and Analytic Learning Method for Generalized Class Incremental Learning. The 21st International Conference on Advanced Data Mining and Applications 2025. Scopus Rank B.

[P4] **Quynh-Trang Pham Thi**, Duc-Hung Nguyen, Thanh Hai Dang, Duc-Trong Le, Tri-Thanh Nguyen, and Quang-Thuy Ha. "Towards Robust Continual Learning: A Multi-Head Approach with Online Prototype Equilibrium and Adaptive Prototypical Feedback." In Asian Conference on Intelligent Information and Database Systems, pp. 275-286. Singapore: Springer Nature Singapore, 2024. Scopus Rank B.

[P5] **Quynh-Trang Pham Thi**, Duc-Hung Nguyen, Anh-Duc Nguyen-Tran, Hung-Dung Nguyen, Tri-Thanh Nguyen, Quang-Thuy Ha and Thanh Hai Dang. "Addressing Catastrophic Forgetting in Continual Learning using Multihead Approach and Latent Alignment via Energy-Based Model." Vietnam journal of computer science. (Accepted). WoS Q3.

[P6] **Quynh-Trang Pham Thi**, Van-Truong Le, Thanh-Huyen Pham, Thanh-Nga Hoang Thi and Duc-Trong Le. DAKTA: Directional Kolmogorov-Arnold Classifier for Task Arithmetic in Continual Learning. SOICT 2025. Scopus.

[P7] **Quynh-Trang Pham Thi**, Quang-Dat Mac, Thanh-Ha Nguyen and Duc-Trong Le. SO-LoRA: Sparse Orthogonal LoRA for Parameter-Efficient Continual Learning. In 2026 IEEE International Conference on Multimedia and Expo (ICME 2026). Scopus Rank A.

The list contains seven publications.