INTER-DOMAIN INTRA-CLASS KNOWLEDGE COMPLEMENT FOR LONG-TAIL DOMAIN INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain Incremental Learning (DIL) aims to learn from a continuous stream of domain-specific data. However, existing DIL approaches predominantly assume class-balanced distributions, overlooking realistic scenarios where data typically follow a long-tailed distribution. To bridge this gap, we investigate the more practical Long-Tailed Domain Incremental Learning (LDIL) problem, which introduces two major challenges: severe catastrophic forgetting of head classes and underrepresentation of tail classes. To address these issues, we propose a novel Inter-**D**omain Intra-**C**lass Knowledge **Comp**lement (DC-Comp) approach that mines discriminative information from previously seen domains to alleviate distributional undersampling in tail classes of new domains and to mitigate the forgetting of head-class knowledge. Specifically, a Discriminativity Modeling Network associated with an Intra-class Balanced Discriminativity Mining mechanism is developed to capture class-specific local features in long-tailed data. In addition, an Inter-domain Discriminativity Reserving and Complementing scheme is introduced to retain these discriminative features and further leverage them to generate pseudo-samples during new domain training, effectively mitigating both distribution imbalance and catastrophic forgetting. Comprehensive experiments conducted under two practical LDIL settings, unified and random cross-domain distributions, demonstrate that DC-Comp achieves state-of-the-art performance. Our code will be released.

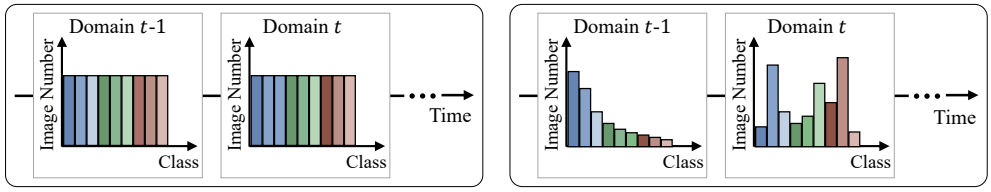
1 INTRODUCTION

Domain Incremental Learning (DIL) focuses on learning from a sequential stream of domain-specific data without access to domain identifiers during inference Wang et al. (2022a); Li et al. (2024c). The major challenge in DIL is catastrophic forgetting, where previously acquired knowledge deteriorates as new domains are learned Liu et al. (2024); Xu et al. (2024c). Most existing DIL methods assume class-balanced distributions, where each class contains an equal number of samples Xu et al. (2025b), as illustrated in fig. 1(a). However, in real-world applications such as autonomous driving, object data naturally exhibit a long-tailed distribution that continuously evolves because object frequencies vary across weather conditions and environments, as shown in fig. 1(b). This more realistic setting, termed Long-Tailed Domain Incremental Learning (LDIL), introduces distinct challenges: amplified forgetting of head-class knowledge and severe underlearning of tail classes caused by sample scarcity, as illustrated in fig. 2(a)(i).

A straightforward solution to LDIL is to integrate long-tail learning strategies with DIL methods Liu et al. (2024). Conventional long-tail learning methods typically rely on data resampling Li et al. (2021); Chen et al. (2022); Shi et al. (2023); Du et al. (2024) or inter-class distribution transfer Li et al. (2023); Hong et al. (2021); Chen et al. (2023); Li et al. (2022). As shown in fig. 2(a)(ii), data resampling aims to balance tail-class samples through repeated sampling and augmentation. However, the limited number of tail-class instances provides restricted diversity relative to the underlying distribution, hindering effective feature learning. Conversely, inter-class distribution transfer (fig. 2(a)(iii)) attempts to align head- and tail-class distributions but often introduces semantic drift due to class heterogeneity, leading to incorrect feature associations. Therefore, existing methods achieve only limited test performance in LDIL, as shown in fig. 2(b)(i)–(iii). Furthermore, recent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

advances in long-tail class-incremental learning Wang et al. (2024b); Liu et al. (2022) mainly focus on class-incremental settings without considering domain shifts, thereby limiting their applicability to scenarios involving domain variation.



(a) Data of Conventional DIL **(b) Data of DIL with Long-tail Distribution**

Figure 1: Comparison of data distribution in conventional domain incremental learning (DIL) and long-tail DIL (LDIL).

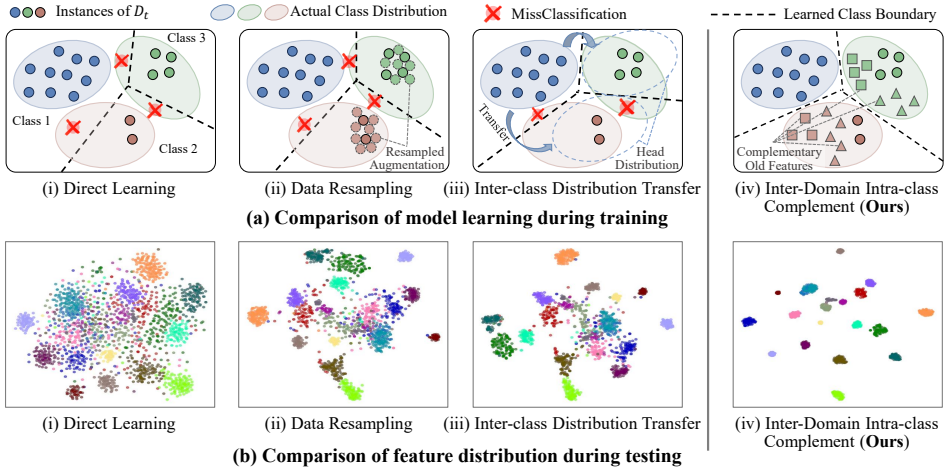


Figure 2: Comparison of existing techniques and our method. (a) Illustration of each approach's influence on training. (b) Visualization of the testing feature distributions.

To address the above challenges, we propose a novel Inter-Domain Intra-Class Knowledge Complement (DC-Comp) approach. As illustrated in fig. 2(a)(iv), DC-Comp leverages discriminative information from historical domains to complement the data distribution of the new domain, thereby alleviating distribution undersampling in underrepresented classes. Specifically, a Discriminativity Modeling Network is introduced to extract local features containing class-discriminative cues. This network is optimized through an Intra-class Balanced Discriminativity Mining mechanism, which employs a multi-round probabilistic sampling strategy to mitigate instance-level imbalance. Furthermore, an Inter-domain Discriminativity Reserving and Complementing scheme is designed to store information-equivalent semantic features with minimal memory overhead. Then, based on these preserved semantics, pseudo-samples are generated to exploit complementary visual information for alleviating data imbalance while simultaneously mitigating catastrophic forgetting through implicit cross-domain alignment. To comprehensively evaluate DC-Comp, two novel LDIL settings, unified and random cross-domain distributions, are constructed to represent consistent and varying inter-class structures, respectively. Extensive experiments on multiple cross-domain benchmarks demonstrate that DC-Comp consistently achieves state-of-the-art performance against advanced DIL and long-tail learning approaches.

To summarize, our contributions are threefold: (1) We present pioneering research on the practical and important LDIL problem and propose a novel DC-Comp framework to jointly address severe underrepresentation and catastrophic forgetting in LDIL scenarios. (2) DC-Comp introduces a discriminative information mining and complement mechanism that employs probabilistic modeling and sampling to adaptively extract and preserve class-discriminative features from historical domains. These features are further utilized as complementary knowledge for learning new domains and mitigating forgetting. (3) Extensive experiments demonstrate that DC-Comp achieves substantial performance gains over existing methods in LDIL scenarios.

2 Related Work

2.1 Domain Incremental Learning

Domain Incremental Learning (DIL) suffers from catastrophic forgetting Mirza et al. (2022); Lamers et al. (2023). Existing DIL methods typically adopt data replay or parameter regularization to alleviate forgetting. However, replay-based approaches require storing historical exemplars, which hinders data privacy Jeeveswaran et al. (2024); Aljundi et al. (2019); Rolnick et al. (2019); Zheng et al. (2024). Regularization-based methods constrain model parameters to preserve old knowledge Bonato et al. (2024); Li et al. (2024c), but such constraints often limit the model's adaptability to new domains Xu et al. (2024b). Recently, prompt-based learning has emerged as a promising paradigm in DIL Wang et al. (2022a); Liu et al. (2024); Lu et al. (2024); Yang et al. (2024), effectively mitigating forgetting by isolating domain-specific knowledge. Despite their effectiveness, most existing approaches assume class-balanced data distributions Liu et al. (2024); Wang et al. (2022c;b); Feng et al. (2026), overlooking the inherently long-tailed nature of real-world data. This oversight leads to insufficient learning for tail classes and limited generalization to realistic conditions.

2.2 Long-tail Learning

Long-tail learning arises from the observation that real-world datasets often exhibit highly imbalanced class distributions Kang et al. (2020); Wang et al. (2020); Mahajan et al. (2018); Buda et al. (2018). Existing methods can be broadly divided into three types: (1) Class-rebalancing methods alleviate sample imbalance through resampling or loss reweighting to emphasize tail classes Hong et al. (2021); Chen et al. (2023); Cao et al. (2019). Although these techniques partially improve tail-class recognition, they are prone to overfitting due to limited sample diversity. (2) Information-transfer methods aim to bridge the representation gap by transferring discriminative knowledge from head to tail classes Du et al. (2023). Some methods mix head and tail samples to enrich tail-class features Du et al. (2023), while others align feature distributions across classes Shi et al. (2023). However, semantic discrepancies between head and tail classes often result in feature distortion, degrading classifier robustness Xu et al. (2024a). (3) Classifier-retraining methods decouple representation learning from classifier calibration. Some methods balance class-wise logit magnitudes through label over-smoothing Sun et al. (2025b), whereas others directly impose monotonic constraints on class-wise classifier norms Zhang et al. (2026). Although effective in static long-tailed recognition, these methods assume sufficiently learned and stationary representations; consequently, classifier-level calibration alone cannot compensate for domain-specific tail information lost during sequential learning. While these approaches have shown promise in static settings, none explicitly addresses the coupled effects of class imbalance, sequential domain shifts, and catastrophic forgetting. This gap motivates the design of our method, which explicitly models and exploits class-specific discriminative information along domain transitions.

2.3 Long-tail Incremental Learning

Long-Tail Incremental Learning (LIL) extends the principles of long-tail learning to incremental scenarios, where imbalanced data arrive continuously Wang et al. (2024b); Liu et al. (2022). Current LIL techniques are predominantly studied under class-incremental settings, where novel classes are introduced sequentially. These methods attempt to transfer knowledge from head to tail classes to mitigate imbalance Wang et al. (2024b). However, when adapted to domain-incremental contexts, they typically treat instances of the same class across domains as independent, failing to exploit shared semantic information Wang et al. (2022a; 2023). In contrast, our proposed DC-Comp method is designed not only to mitigate catastrophic forgetting but also to enhance tail-class learning by leveraging cross-domain discriminative information. By preserving and transferring class-specific semantics across domains, DC-Comp effectively addresses the limitations of existing LIL and DIL methods, establishing a new paradigm for Long-Tail Domain Incremental Learning (LDIL).

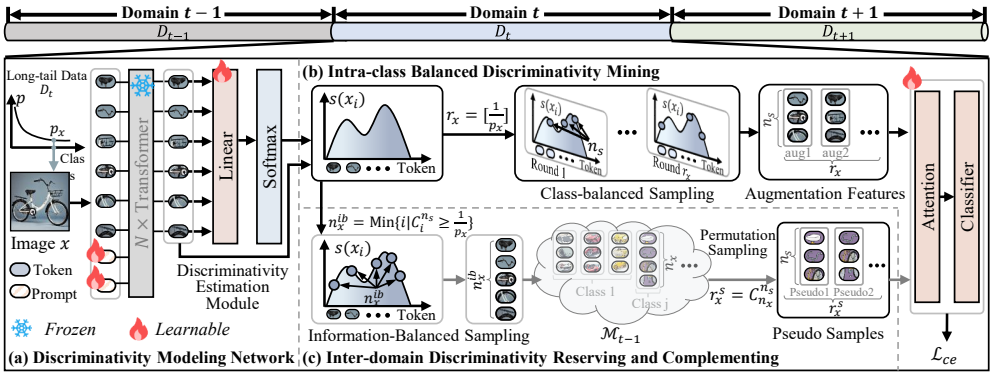


Figure 3: Overview of our DC-Comp approach. The Discriminativity Modeling Network (DMN) extracts features from each input sample and generates a discriminativity score for each local feature. The Intra-class Balanced Discriminativity Mining module (IBDM) improves discriminativity-score prediction through differentiable sampling. The Inter-domain Discriminativity Reserving and Complementing scheme (IDRC) selects and preserves discriminative features from historical domains and generates pseudo-samples to alleviate information imbalance across classes.

3 Proposed Method

3.1 Problem Formulation

Long-tailed domain-incremental learning (LDIL) introduces a sequence of T datasets $\mathcal{D} = \{D_t\}_{t=1}^{T}$. All datasets share a common label space $\mathcal{C}$ and exhibit class-imbalanced distributions, *e.g.*, Pareto distributions Shao et al. (2024); Wang et al. (2024a); Zhao et al. (2024). In this paradigm, only the current domain D_t is accessible during training, while previous datasets $\{D_t\}_{t=1}^{t-1}$ are inaccessible Liu et al. (2024). To quantify class imbalance, we define the relative class-balance ratio for a given image x as:

$$p_x = \frac{p(y_x)}{\max_{i \in \mathcal{C}} p(y_i)}, \tag{1}$$

where y_x is the label of x and $p(y_i)$ denotes the proportion of instances labeled as y_i in the current domain D_t. This metric highlights the disparity in representation among different classes, guiding our discriminative sampling strategy. To evaluate model performance, each domain is associated with a test set D_t^{te}.

3.2 Discriminativity Modeling Network

Inspired by recent prompt-learning-based DIL methods Liu et al. (2024); Wang et al. (2022a), we build our approach upon a prompt-learning mechanism. Specifically, a pretrained Vision Transformer (ViT) model $\mathbf{\Theta}$ is employed as the backbone, with its parameters kept frozen. When data from a new domain D_t are obtained, a learnable prompt set $\mathcal{P}_t$ is instantiated as follows:

$$\mathcal{P}_t = \{(\mathbf{p}_t^1, \mathbf{k}_t^1), (\mathbf{p}_t^2, \mathbf{k}_t^2), \ldots, (\mathbf{p}_t^{N_\mathbf{P}}, \mathbf{k}_t^{N_\mathbf{P}})\}, \tag{2}$$

where $\mathbf{p}_t^i \in \mathbb{R}^{L_\mathbf{P} \times D}$ is a prompt embedding and $\mathbf{k}_t^i \in \mathbb{R}^D$ serves as a key for cosine-similarity-based matching. Here, $L_\mathbf{P}$ denotes the prompt length, and D is the input dimension of the ViT transformer layers. Given an image x, the matched prompt is retrieved by

$$(\mathbf{p}^x, \mathbf{k}^x) = \arg \max_{(\mathbf{p}_t^i, \mathbf{k}_t^i) \in \mathcal{P}_t} \big(\gamma(q(x), \mathbf{k}_i)\big), \tag{3}$$

where $q(x)$ is a query function, typically implemented using a pretrained ViT to extract the image feature, and γ is the cosine similarity function. The input image x is then tokenized Ronen et al. (2023) into a sequence representation $\boldsymbol{h}_x \in \mathbb{R}^{L_h \times d}$, where L_h is the sequence length. The matched prompt $\boldsymbol{p}^x$ and $\boldsymbol{h}_x$ are then fed into the transformer layers. The resulting output token sequence $\boldsymbol{h}_x^*$ is represented as

$$\boldsymbol{h}_x^* = \mathbf{\Theta}([\boldsymbol{h}_x, \boldsymbol{p}^x]). \tag{4}$$

Discriminativity Estimation Module: Existing methods typically exploit a single [CLS] token Liu et al. (2024) to generate a class prediction, which requires implicit aggregation of discriminative information and has limited capacity when training instances of specific classes are scarce. We instead use a lightweight module to identify discriminative tokens within the input x. Specifically, the Discriminativity Estimation Module (DEM), parameterized by $\boldsymbol{\psi}$, contains a linear layer and a softmax layer. Since each token corresponds to an image patch, the DEM process can be represented as

$$\{s(x_1), s(x_2), \cdots, s(x_{L_h})\} = \boldsymbol{\psi}(\boldsymbol{h}_x^*), \tag{5}$$

where x_i represents the image patch associated with token $[\boldsymbol{h}_x^*]_i$ and $s(x_i)$ is a patch-level discriminativity score.

3.3 Intra-class Balanced Discriminativity Mining

To improve discriminative-information recognition under the imbalanced data distribution in LDIL, we introduce a class-balanced multi-round sampling algorithm, as shown in Fig. 3(b). Specifically, given the token sequence $\boldsymbol{h}_x^*$, discriminativity scores $s(x_i)$, and relative class-balance ratio p_x, we sample from $\boldsymbol{h}_x^*$ over $r_x = \lfloor \frac{1}{p_x} \rfloor$ rounds, selecting n_s tokens per round according to $s(x_i)$.

Consequently, we obtain a set of r_x augmentation groups $\mathcal{G}_x = \{aug_j\}_{j=1}^{r_x}$, where aug_j contains n_s tokens sampled from $\boldsymbol{h}_x^*$ and n_s is a hyperparameter. Each augmentation feature group aug_j is treated as an augmented sample of x and represented by a matrix $X_j \in \mathbb{R}^{n_s \times d}$. We then use an attention layer to process each augmentation feature group aug_j:

$$\boldsymbol{f}_j = \text{softmax}(X_j \boldsymbol{W}_a + b_a)^\top X_j, \tag{6}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{d \times 1}$ and $b_a \in \mathbb{R}$ are learnable weights and bias, respectively. $\boldsymbol{f}_j \in \mathbb{R}^d$ is the generated feature that aggregates the semantic information of the n_s sampled tokens. A classifier parameterized by $\boldsymbol{W}_c \in \mathbb{R}^{d \times C}$ then produces the class prediction, and the cross-entropy (CE) loss is calculated as follows:

$$\mathcal{L}_{ce} = CE(\boldsymbol{f}_j \boldsymbol{W}_c, y_x). \tag{7}$$

3.4 Inter-domain Discriminativity Reserving and Complementing

Discriminative Pseudo-sample Generation: When $t > 1$, as shown in Fig. 3(c), a token memory $\mathcal{M}_{t-1}$ is introduced to store historical tokens, which provide complementary information for learning the new domain. Specifically, given a class y_x and n_x tokens from a historical instance in $\mathcal{M}_{t-1}$, we sample token combinations to generate different pseudo-samples. The number of possible pseudo-samples is

$$r_x^s = C_{n_x}^{n_s}, \tag{8}$$

where n_x is the number of tokens preserved from a historical instance and n_s is the number selected for each pseudo-sample as introduced in Section 3.3. The total number of possible pseudo-samples for each class i can be as large as

$$N_{r,i} = \sum_{j=1}^{N_i} C_{n_x^j}^{n_s}, \tag{9}$$

where N_i is the number of instances with label y_x in $\mathcal{M}_{t-1}$, and n_x^j is the number of tokens in an instance. Following Eq. 6 and Eq. 7, the generated pseudo-samples are also fed into the attention layer and classifier. The abundant discriminative information in the pseudo-samples not only provides complementary cues that alleviate information imbalance across domains but also promotes semantic alignment between features from different domains, significantly mitigating forgetting caused by domain gaps. Moreover, because the pseudo-samples bypass the backbone and are fed directly into the lightweight classification module, they introduce limited training overhead.

Information-Balanced Sampling: After training on domain t, we sample discriminative tokens from $\boldsymbol{h}_x^*$ and add them to the token memory to provide valuable information for later domains. Sampling the same number of tokens for all instances can lead to information imbalance in the preserved tokens because head classes typically dominate the stored information. To address this issue, we introduce an information-balanced sampling algorithm in which each instance is assigned

a different token-sampling number according to the relative class-balance ratio p_x. Specifically, the instance-specific information-balanced token-sampling number is obtained by

$$n_x^{ib} = \min\{i \in \mathbb{N} : i \geq n_s,\ C_i^{n_s} \geq \frac{1}{p_x}\}. \tag{10}$$

Because a small number of preserved tokens can generate many pseudo-samples according to Eq. 8, the potential number of pseudo-sample configurations is balanced across classes with negligible additional storage overhead. Adding the sampled tokens to the previous token memory $\mathcal{M}_{t-1}$ yields the updated token memory $\mathcal{M}_t$ and effectively preserves discriminative information from the current domain while retaining only sparse high-level semantics rather than raw data.

Training and Inference: When training on the initial domain, only the modules in Fig. 3(a) and (b) are adopted. When training on domain t ($t > 1$), the modules in Fig. 3(a)–(c) are used to optimize the model.

During inference, given an input image x, the inference prompt is matched against all historical prompt pools $\mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \mathcal{P}_T$ according to Eq. 3. Following Fig. 3(a), the matched prompt and x are fed into the pretrained ViT to obtain the token sequence $\boldsymbol{h}_x^*$, from which an aggregated image feature is obtained as

$$\boldsymbol{f}_x = \text{softmax}(\boldsymbol{h}_x^* \boldsymbol{W}_a + b_a)^\top \boldsymbol{h}_x^*, \tag{11}$$

where we empirically find that using all tokens in $\boldsymbol{h}_x^*$ yields better performance than sampling n_s tokens from $\boldsymbol{h}_x^*$. This is because aggregating all relevant tokens during inference improves the robustness of the generated features. The final prediction is obtained by

$$\hat{y} = \text{softmax}(\boldsymbol{f}_x \boldsymbol{W}_c). \tag{12}$$

4 Experiment

4.1 Experimental Setups

Datasets: Three multi-domain datasets are used in this paper: DomainNet Peng et al. (2019), ImageNet-R Hendrycks et al. (2021), and ImageNet-C Hendrycks & Dietterich (2019). DomainNet is a large-scale cross-domain dataset, comprising 345 classes with 586,575 images across six distinct stylistic domains: Real, Quickdraw, Sketch, Painting, Infograph, and Clipart. ImageNet-R contains 30,000 images from 200 classes, categorized into 15 stylistic domains (e.g., art, cartoon, and graffiti). We use a 200-class subset of ImageNet-C covering 15 types of image-quality degradation and environmental perturbation.

Long-tail Settings: Following ImageNet-LT Liu et al. (2019), we subsample the original datasets according to a Type II Pareto distribution with parameter $\alpha = 6$ to generate multi-domain long-tailed datasets Shao et al. (2024). The sampling process ensures that the relative class frequencies follow the Pareto distribution, with the sample count of each class determined by its predefined rank. We consider two long-tailed configurations:

(1) *Unified Long-Tailed Subset* (Unif): All domains share the same long-tailed class distribution. Head classes (high-sample-count classes) and tail classes (low-sample-count classes) are determined globally across all domains, preserving the original dataset's class-frequency hierarchy.

(2) *Randomized Long-Tailed Subset* (Rand): Each domain has a unique randomized long-tailed class distribution. Within each domain, classes with higher native sample counts are more likely to become head classes, while those with fewer samples are assigned as tail classes.

Evaluation Metrics: For both settings, the test sets remain unmodified and retain their original class distributions. After the model is trained on the t-th domain, average accuracy (**Avg-ACC**) is used to evaluate its knowledge-accumulation capacity and is calculated as

$$A_t = \frac{1}{t} \sum_{i=1}^{t} a_{t,i}, \tag{13}$$

where $a_{t,i}$ denotes the test accuracy on the i-th domain. Therefore, A_T indicates the overall performance of M_T across all seen domains.

Table 1: Performance comparison on different benchmarks. "Origin," "Unif," and "Rand" indicate the original benchmark, class-aware cross-domain unified-distribution setting, and class-independent cross-domain random-distribution setting, respectively.

Method	Pub.	DomainNet			ImageNet-R			ImageNet-C		
		Unif	Rand	Origin	Unif	Rand	Origin	Unif	Rand	Origin
Dual-Prompt	*ECCV 2022*	47.70	50.75	53.63	48.97	54.87	58.21	64.83	65.80	84.97
L2P-Prompt	*CVPR 2022*	47.10	50.03	51.79	47.57	52.99	57.74	61.24	62.19	75.04
S-Prompts	*NeurIPS 2022*	40.39	47.72	57.37	33.53	51.25	53.18	37.06	34.90	69.22
ESN	*AAAI 2023*	10.00	39.12	45.31	2.85	10.61	17.57	4.48	5.79	69.37
CODA-Prompt	*CVPR 2023*	48.25	51.93	54.51	47.41	51.74	55.22	60.75	61.27	72.83
CPrompt	*CVPR 2024*	42.73	47.28	52.47	42.46	57.37	59.54	42.34	57.83	69.97
InfLoRA	*CVPR 2024*	40.02	49.29	51.42	13.30	27.23	32.02	9.84	7.40	53.09
DGR	*CVPR 2024*	39.17	49.39	55.80	3.01	18.41	15.54	6.86	5.07	80.19
C-Prompt	*IJCV 2024*	40.29	49.25	58.68	43.78	57.35	62.55	48.23	55.41	81.16
MOS	*AAAI 2025*	36.71	44.96	51.01	15.64	32.28	41.30	23.09	20.40	63.69
DCE	*ICML 2025*	31.45	39.85	47.74	20.90	30.70	32.67	29.92	32.42	50.21
KA-Prompt	*ICML 2025*	47.38	54.76	<u>62.91</u>	53.68	60.94	66.51	52.7	58.53	85.43
HEDP	*ICML 2026*	35.70	37.58	44.57	8.01	44.71	55.96	7.86	8.28	59.82
Dual-Prompt										
+ Resampling	*JAIR 2002*	48.81	50.48	51.81	47.81	55.38	57.21	65.70	66.63	85.21
+ Focal Loss	*ICCV 2017*	47.62	49.84	51.71	48.50	55.46	56.35	65.94	66.60	84.61
+ CSA	*NeurIPS 2023*	47.95	48.85	50.12	45.81	45.30	47.59	<u>68.81</u>	<u>68.51</u>	80.84
+ ProCo	*PAMI 2024*	47.58	49.91	51.68	48.72	55.20	58.58	65.99	66.36	84.43
+ GBG	*AAAI 2024*	33.84	44.87	53.79	38.23	51.86	57.37	48.96	51.78	68.85
+ LOS	*ICLR 2025*	34.39	45.83	54.32	38.26	52.81	58.51	50.52	52.56	68.86
+ SAMN	*CVPR 2026*	39.43	46.23	53.11	32.20	44.56	43.98	46.24	52.60	65.00
KA-Prompt										
+ Resampling	*JAIR 2002*	50.80	<u>55.50</u>	61.29	50.60	57.54	61.09	57.59	58.06	<u>85.80</u>
+ Focal Loss	*ICCV 2017*	<u>50.86</u>	55.40	61.56	38.48	57.40	60.96	58.08	60.27	85.18
+ CSA	*NeurIPS 2023*	50.15	53.10	57.73	54.20	63.71	65.80	53.26	56.99	81.70
+ ProCo	*PAMI 2024*	41.78	50.64	60.00	54.76	64.57	66.85	52.68	59.92	80.90
+ GBG	*AAAI 2024*	47.68	54.91	62.73	<u>57.85</u>	<u>68.44</u>	<u>70.51</u>	53.77	60.04	85.22
+ LOS	*ICLR 2025*	36.70	51.75	62.30	37.02	61.31	58.64	16.57	36.01	83.78
+ SAMN	*CVPR 2026*	40.96	51.52	60.32	27.43	57.21	67.32	9.98	21.82	83.19
DC-Comp	*This Paper*	**53.78**	**56.87**	**63.04**	**62.77**	**69.80**	**71.46**	**78.74**	**80.32**	**90.45**
Gain ↑	-	+2.92	+1.37	+0.13	+4.92	+1.36	+0.95	+9.93	+11.81	+4.65

Implementation Details: We build our framework upon Dual-Prompt Wang et al. (2022b), a simple prompt-based approach. The Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used to train the model. The default batch size and learning rate for all benchmarks are set to 128 and 0.005, respectively. The default number of training epochs is 5. The training images are resized to 224×224. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

Compared Methods: We compare DC-Comp with the state-of-the-art incremental-learning methods Dual-Prompt Wang et al. (2022b), L2P-Prompt Wang et al. (2022c), S-Prompts Wang et al. (2022a), ESN Wang et al. (2023), CODA-Prompt Smith et al. (2023), CPrompt Gao et al. (2024), InfLoRA Liang & Li (2024), C-Prompt Liu et al. (2024), MOS Sun et al. (2025a), DCE Li et al. (2025), KA-Prompt Xu et al. (2025b), HEDP Feng et al. (2026), and the long-tail class-incremental learning method DGR He (2024). In addition, we combine the long-tail learning methods Resampling Japkowicz & Stephen (2002), Focal Loss Lin et al. (2017), CSA Shi et al. (2023), ProCo Du et al. (2024), GBG Li et al. (2024b), LOS Sun et al. (2025b), and SAMN Zhang et al. (2026) with the advanced DIL approaches Dual-Prompt and KA-Prompt for a more comprehensive comparison. The main comparisons use ViT-B/16 pretrained on ImageNet-21k as the backbone.

4.2 Comparison Results

The experimental comparisons on the original DomainNet, ImageNet-R, and ImageNet-C benchmarks and their long-tailed versions (Unif and Rand) are reported in Table 1. The best and second-best results are indicated by boldface and underlining, respectively. Among the recently introduced baselines, HEDP exhibits limited robustness to long-tailed domain streams, achieving 8.01% and 7.86% on ImageNet-R and ImageNet-C under Unif, respectively. Integrating LOS or SAMN with Dual-Prompt and KA-Prompt yields setting-dependent results: the strongest of these vari-

ants, KA-Prompt+LOS, reaches 61.31% on ImageNet-R under Rand, but remains 8.49 points below DC-Comp. In comparison, KA-Prompt+GBG provides the strongest long-tail plug-in baseline on ImageNet-R, reaching 57.85%, 68.44%, and 70.51% under the Unif, Rand, and Origin settings, respectively.

Under the Unif distribution: The Unif distribution represents a scenario in which the sample proportion of each class remains consistent across domains. As shown in table 1, the proposed DC-Comp surpasses the strongest baselines, including KA-Prompt+Focal Loss, KA-Prompt+GBG, and Dual-Prompt+CSA, achieving improvements of **2.92/4.92/9.93** on DomainNet, ImageNet-R, and ImageNet-C, respectively. These gains are primarily attributed to the Inter-domain Discriminativity Reserving and Complementing module, which effectively captures and utilizes historical discriminative information from tail classes, thereby facilitating more robust learning of new domains.

Under the Rand distribution: The Rand distribution represents a dynamic scenario in which class-wise sample proportions vary across domains. In this setting, most methods perform better than under the Unif setting because previously encoded domain knowledge, in which certain classes may be more prominent, helps mitigate the severe underlearning of tail classes. As shown in table 1, DC-Comp maintains its superiority over the strongest competing methods, namely KA-Prompt+Resampling on DomainNet, KA-Prompt+GBG on ImageNet-R, and Dual-Prompt+CSA on ImageNet-C, achieving improvements of **1.37/1.36/11.81** across the three benchmarks. These results demonstrate that the proposed information-preservation strategy is more effective than existing approaches for tail-class learning.

Under the Origin distribution: The Origin distribution refers to the setting in which the original datasets are used directly for training. DC-Comp surpasses the state-of-the-art KA-Prompt, KA-Prompt+GBG, and KA-Prompt+Resampling methods, achieving gains of **0.13/0.95/4.65** across the three benchmarks. These results demonstrate that DC-Comp remains effective even without long-tail subsampling. This robustness stems from the proposed discriminative-information-preservation mechanism, which provides complementary knowledge when learning new domains and is particularly beneficial when training samples are limited, as observed in ImageNet-R and ImageNet-C.

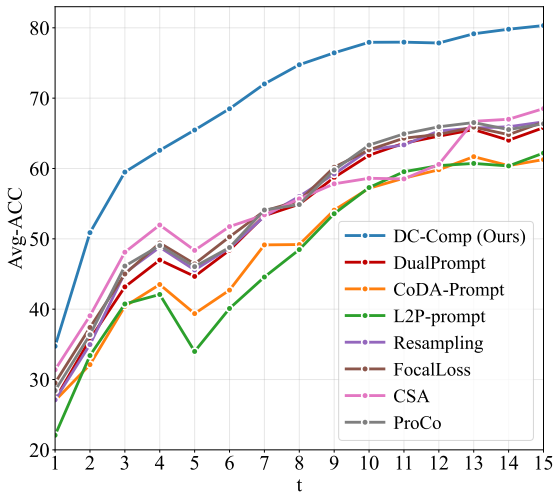


Figure 4: Performance trends compared with state-of-the-art methods.

Visualization of model learning trends: To further analyze the model's learning process, we visualize the average accuracy (Avg-ACC) across seen domains throughout the LDIL process in fig. 4, using ImageNet-C under the Rand long-tail setting. The results indicate that our method consistently outperforms existing approaches across all stages, with increasing gains over the CSA baseline (DIL + Long-tail). This advantage is attributed to our inter-domain intra-class knowledge-complement designs, which effectively mine and exploit discriminative information to facilitate knowledge acquisition and retention in LDIL scenarios.

4.3 Ablation Study and Additional Analysis

Analysis of the proposed modules. Since the DMN (Fig. 3(a)) and IBDM (Fig. 3(b)) modules are interdependent, they are collectively referred to as IBDM for clarity in this section. Figure 5 presents the ablation studies of IBDM and IDRC, conducted on ImageNet-C under the Rand long-tail setting. When the IBDM module is applied independently, it achieves a **4.71%** improvement over the baseline. This enhancement is primarily attributed to the class-balanced multi-round sampling algorithm, which mitigates the effects of class imbalance to some extent. When IDRC is further integrated,

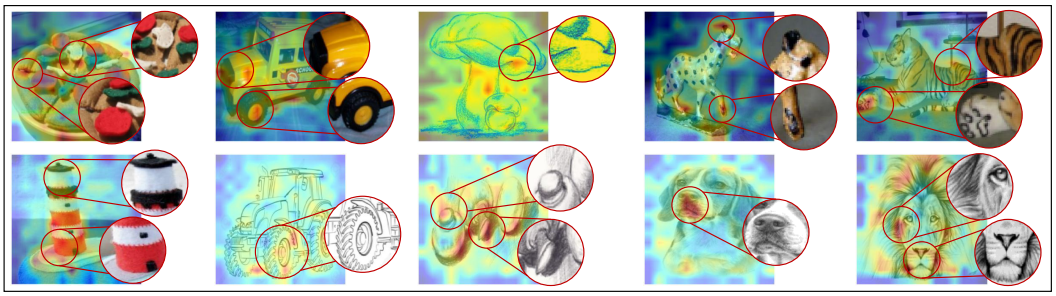


Figure 7: Visualization of learned discriminativity scores and recognized discriminative regions by our DC-Comp approach.

the performance gain increases to a remarkable **14.52%**, as the preserved discriminative historical information provides valuable complementary knowledge when learning from new-domain data.

Analysis of the hyperparameter: In fig. 6, the performance of DC-Comp is evaluated under different values of the hyperparameter n_s, which denotes the number of tokens sampled in IBDM. Model performance improves as n_s increases from 1 to 7 because more sampled tokens encode richer semantic information. However, when n_s exceeds 7, performance declines because the excessive number of local object regions introduces non-discriminative features that interfere with model learning. Based on these observations, n_s is set to 7 by default.

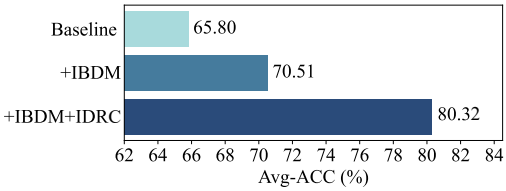


Figure 5: Ablation on components.

Visualization of learned discriminativity scores: Figure 7 visualizes the discriminativity scores predicted by DC-Comp and the corresponding discriminative regions across various instances. For both man-made and natural objects in cartoon and real-world domains, DMN effectively locates category-relevant feature regions. This result also demonstrates the effectiveness of our intra-class balanced discriminativity mining mechanism, which encourages the network to produce high-confidence predictions for discriminative regions.

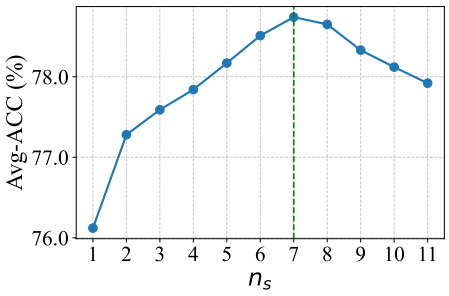


Figure 6: Ablation on the token sampling number n_s in IBDM.

Additional analysis: We provide the training algorithm, together with additional experiments on computational efficiency, token preservation strategies, compatibility with different DIL baselines, performance changes from balanced to imbalanced settings and Head/Tail accuracy analysis. Detailed results and discussions are provided in Appendices A, B, C, D, E, F, and G.

5 Conclusion and Limitation

In this paper, we pioneer the study of the practical long-tail domain incremental learning (LDIL) problem and propose DC-Comp, a novel approach that preserves discriminative semantics from past domains to alleviate distribution undersampling in tail classes without storing raw historical data. Our method incorporates a Discriminativity Modeling Network with an Intra-class Balanced Discriminativity Mining mechanism to efficiently mine key information from known domains according to the class distribution. Furthermore, an Inter-domain Discriminativity Reserving and Complementing scheme effectively retains lightweight, high-level discriminative features and generates pseudo-samples during training on new domains, mitigating both inter-class information imbalance and catastrophic forgetting. We also introduce two realistic LDIL settings to support future research. Extensive experiments validate the effectiveness of our approach.

Limitation: Our model relies on the semantic relevance across domains. When the discriminative information across domains is in conflict or rarely relevant, our performance would be limited.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and feedback on presentation. All research decisions, experimental designs, implementations, analyses, and reported results were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *NeurIPS*, 32, 2019.

Jacopo Bonato, Francesco Pelosin, Luigi Sabetta, and Alessandro Nicolosi. Mind: Multi-task incremental network distillation. In *AAAI*, volume 38, pp. 11105–11113, 2024.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019.

Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI*, volume 36, pp. 356–364, 2022.

Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *ICCV*, pp. 19277–19287, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE TPAMI*, 2024.

Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *CVPR*, pp. 15814–15823, 2023.

Yu Feng, Zhen Tian, Haoran Luo, Xie Yu, Diancheng Cheng, Haoyue Zheng, Shuai Lyu, Ping Zong, Lianyuan Li, Xin Ge, et al. Hedp: A hybrid energy-distance prompt-based framework for domain incremental learning. *arXiv preprint arXiv:2605.05776*, 2026.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *CVPR*, pp. 28463–28473, 2024.

Jiangpeng He. Gradient reweighting: Towards imbalanced class-incremental learning. In *CVPR*, pp. 16668–16677, 2024.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8340–8349, 2021.

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, pp. 6626–6636, 2021.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Kishaan Jeeveswaran, Elahe Arani, and Bahram Zonooz. Gradual divergence for seamless adaptation: A novel domain incremental learning method. In *ICML*, volume 235, pp. 21486–21501. PMLR, 2024.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Christiaan Lamers, René Vidal, Nabil Belbachir, Niki van Stein, Thomas Bäeck, and Paris Giampouras. Clustering-based domain-incremental learning. In *ICCV*, pp. 3384–3392, 2023.

Jian Li, Ziyao Meng, Daqian Shi, Rui Song, Xiaolei Diao, Jingwen Wang, and Hao Xu. Fcc: Feature clusters compression for long-tailed visual recognition. In *CVPR*, pp. 24080–24089, 2023.

Lan Li, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Addressing imbalanced domain-incremental learning through dual-balance collaborative experts. In *ICML*, 2025.

Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, pp. 6929–6938, 2022.

Qiwei Li, Kunlun Xu, Yuxin Peng, and Jiahuan Zhou. Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation. *IJCV*, pp. 1–16, 2024a.

Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, pp. 5212–5221, 2021.

Weiqi Li, Fan Lyu, Fanhua Shang, Liang Wan, and Wei Feng. Long-tailed learning as multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3190–3198, 2024b.

Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In *ECCV*, pp. 127–144. Springer, 2024c.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *CVPR*, pp. 23638–23647, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Xialei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Long-tailed class incremental learning. In *ECCV*, pp. 495–512. Springer, 2022.

Zichen Liu, Yuxin Peng, and Jiahuan Zhou. Compositional prompting for anti-forgetting in domain incremental learning. *IJCV*, pp. 1–18, 2024.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pp. 2537–2546, 2019.

Yue Lu, Shizhou Zhang, De Cheng, Yinghui Xing, Nannan Wang, Peng Wang, and Yanning Zhang. Visual prompt tuning in null space for continual learning. *NeurIPS*, 37:7878–7901, 2024.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, pp. 181–196, 2018.

M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *CVPR*, pp. 3001–3011, 2022.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pp. 1406–1415, 2019.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *NeurIPS*, 32, 2019.

Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *CVPR*, pp. 4613–4622, 2023.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. *NeurIPS*, 37:123007–123031, 2024.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? *NeurIPS*, 36:75669–75687, 2023.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *CVPR*, pp. 11909–11919, 2023.

Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. Mos: Model surgery for pre-trained model-based class-incremental learning. In *AAAI*, volume 39, pp. 20699–20707, 2025a.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, 2025b.

Haohui Wang, Weijie Guan, Chen Jianpeng, Zi Wang, and Dawei Zhou. Towards heterogeneous long-tailed learning: Benchmarking, metrics, and toolbox. *NeurIPS*, 37:73098–73123, 2024a.

Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, pp. 728–744. Springer, 2020.

Xi Wang, Xu Yang, Jie Yin, Kun Wei, and Cheng Deng. Long-tail class incremental learning via independent sub-prototype construction. In *CVPR*, pp. 28598–28607, 2024b.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *NeurIPS*, 35:5682–5695, 2022a.

Yabin Wang, Zhiheng Ma, Zhiwu Huang, Yaowei Wang, Zhou Su, and Xiaopeng Hong. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *AAAI*, volume 37, pp. 10209–10217, 2023.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pp. 631–648. Springer, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pp. 139–149, 2022c.

Kunlun Xu, Haozhuo Zhang, Yu Li, Yuxin Peng, and Jiahuan Zhou. Mitigate catastrophic remembering via continual knowledge purification for noisy lifelong person re-identification. In *ACM MM*, pp. 5790–5799, 2024a.

Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *CVPR*, pp. 16604–16613, 2024b.

Kunlun Xu, Xu Zou, and Jiahuan Zhou. Lstkc: Long short-term knowledge consolidation for lifelong person re-identification. In *AAAI*, volume 38, pp. 16202–16210, 2024c.

Kunlun Xu, Yibo Feng, Jiangmeng Li, Yongsheng Qi, and Jiahuan Zhou. C^2prompt: Class-aware client knowledge interaction for federated continual learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Kunlun Xu, Xu Zou, Gang Hua, and Jiahuan Zhou. Componential prompt-knowledge alignment for domain incremental learning. In *ICML*, 2025b.

Longrong Yang, Hanbin Zhao, Yunlong Yu, Xiaodong Zeng, and Xi Li. Rcs-prompt: Learning prompt to rearrange class space for prompt-based continual learning. In *ECCV*, pp. 1–20. Springer, 2024.

Shuo Zhang, Chenqi Li, and Tingting Zhu. Why not hyperparameter-friendly optimisation? a monotonic adaptive norm rescaling approach for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36947–36956, June 2026.

Zhe Zhao, HaiBin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. *NeurIPS*, 37:7493–7520, 2024.

Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Multi-layer rehearsal feature augmentation for class-incremental learning. In *ICML*, 2024.

A Training Overhead Analysis

To comprehensively evaluate computational efficiency, we report training-time overheads on ImageNet-R under the Rand long-tail setting in table 2, matching the corresponding column of the main comparison in table 1. All measurements were obtained from single-seed runs on one NVIDIA RTX 4090 GPU. DC-Comp requires only 243 seconds (**DMN+IBDM: 213 seconds**; **IDRC: 30 seconds**) while achieving the highest Avg-ACC of 69.80%. In comparison, Dual-Prompt requires 152 seconds but is **14.93** less accurate. Its computational cost also increases substantially when combined with existing long-tail techniques: +Resampling, +CSA, +ProCo, and +GBG require 226, 439, 378, and 2610 seconds, respectively. Likewise, KA-Prompt requires 1201 seconds, while its corresponding variants require 3960, 2510, 2300, and 1995 seconds. When integrated with Dual-Prompt, the recently introduced LOS and SAMN strategies require 1176 and 1196 seconds, respectively, which are **4.84**× and **4.92**× the training time of DC-Comp, while their Avg-ACC values remain 16.99 and 25.24 points lower. Their KA-Prompt variants require 1984 and 2033 seconds (**8.16**× and **8.37**× the cost of DC-Comp) and trail DC-Comp by 8.49 and 12.59 points, respectively. Overall, DC-Comp offers a favorable accuracy–efficiency trade-off because its lightweight information-preservation and feature-complement operations avoid the costly multi-stage optimization used by several competing methods.

Table 2: Training-time comparison on ImageNet-R under the Rand long-tail setting.

Method	**Pub.**	**Training Time (s)**	**Avg-ACC**
Dual-Prompt Wang et al. (2022b)	*ECCV 2022*	152	54.87
DGR He (2024)	*CVPR 2024*	26446	18.41
C-Prompt Liu et al. (2024)	*IJCV 2024*	660	57.35
MOS Sun et al. (2025a)	*AAAI 2025*	9313	32.28
KA-Prompt Xu et al. (2025b)	*ICML 2025*	1201	60.94
HEDP Feng et al. (2026)	*ICML 2026*	4298	44.71
Dual-Prompt			
+ Resampling Japkowicz & Stephen (2002)	*JAIR 2002*	226	55.38
+ CSA Shi et al. (2023)	*NeurIPS 2023*	439	45.30
+ ProCo Du et al. (2024)	*PAMI 2024*	378	55.20
+ GBG Li et al. (2024b)	*AAAI 2024*	2610	51.86
+ LOS Sun et al. (2025b)	*ICLR 2025*	1176	52.81
+ SAMN Zhang et al. (2026)	*CVPR 2026*	1196	44.56
KA-Prompt			
+ Resampling Japkowicz & Stephen (2002)	*JAIR 2002*	3960	57.54
+ CSA Shi et al. (2023)	*NeurIPS 2023*	2510	63.71
+ ProCo Du et al. (2024)	*PAMI 2024*	2300	64.57
+ GBG Li et al. (2024b)	*AAAI 2024*	1995	<u>68.44</u>
+ LOS Sun et al. (2025b)	*ICLR 2025*	1984	61.31
+ SAMN Zhang et al. (2026)	*CVPR 2026*	2033	57.21
DC-Comp	*This Paper*	243	**69.80**

B Token Preservation Strategies

In addition to sampling tokens based on the predicted discriminativity scores illustrated in Fig. 3(c), we investigate alternative strategies, including fixed-position sampling of [CLS] or patch tokens and random sampling, as presented in table 3. Specifically, *Fixed/[CLS]* and *Fixed/Patch* refer to preserving the [CLS] token and the central patch tokens of an image, respectively.

Among these strategies, *Fixed/[CLS]* generally yields lower performance than the other strategies because it primarily encodes global features of the entire image Dosovitskiy et al. (2021), in which domain bias tends to dominate the representation Li et al. (2024a), making it less discriminative for later domains. *Fixed/Patch* also shows limited effectiveness because fixed-position patches are not consistently the most discriminative for all classes and tend to be redundant across instances

of the same class. The *Random* strategy samples tokens uniformly at random. It outperforms both *Fixed/[CLS]* and *Fixed/Patch* by introducing more diverse local information from historical samples. However, it remains suboptimal compared with our method because randomly sampled patches are likely to include non-discriminative or background information. In contrast, our method uses the predicted discriminativity scores to sample patches, as depicted in Fig. 3(c) and visualized in Fig. 7, thereby preserving class-relevant information.

Table 3: Performance comparison of different token-preservation methods.

Method	DomainNet			ImageNet-C		
	Unif	Rand	Origin	Unif	Rand	Origin
Fixed/[CLS]	31.42(22.36↑)	42.03(14.84↑)	49.20(13.84↑)	64.35(14.39↑)	66.18(14.14↑)	77.88(12.57↑)
Fixed/Patch	44.71(9.07↑)	50.12(6.75↑)	58.64(4.40↑)	61.77(16.97↑)	70.20(10.12↑)	83.92(6.53↑)
Random	47.68(6.10↑)	51.21(5.66↑)	57.95(5.09↑)	68.98(9.76↑)	71.27(9.05↑)	86.22(4.23↑)
Ours	**53.78**	**56.87**	**63.04**	**78.74**	**80.32**	**90.45**

C Adapting to Different Baselines

DC-Comp demonstrates broad compatibility and effectively enhances various baseline models across diverse long-tail settings and multiple cross-domain benchmarks. Figures 8–10 compare the average accuracy (Avg-ACC) of different baseline models (CODA-Prompt, L2P-Prompt, C-Prompt, and Dual-Prompt) with that of their DC-Comp-enhanced variants on DomainNet, ImageNet-R, and ImageNet-C, respectively. Each figure contains three panels representing the unified (Unif), random (Rand), and original (Origin) settings.

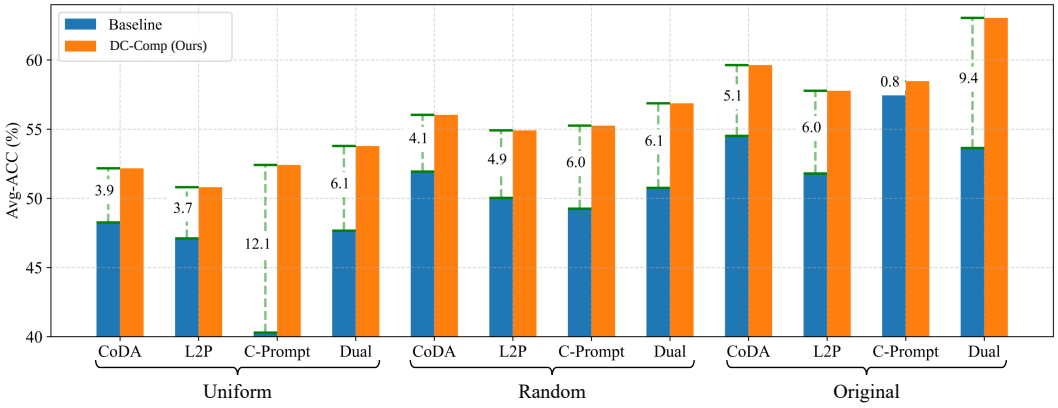


Figure 8: Adapting DC-Comp to different baselines on the **DomainNet** benchmark.

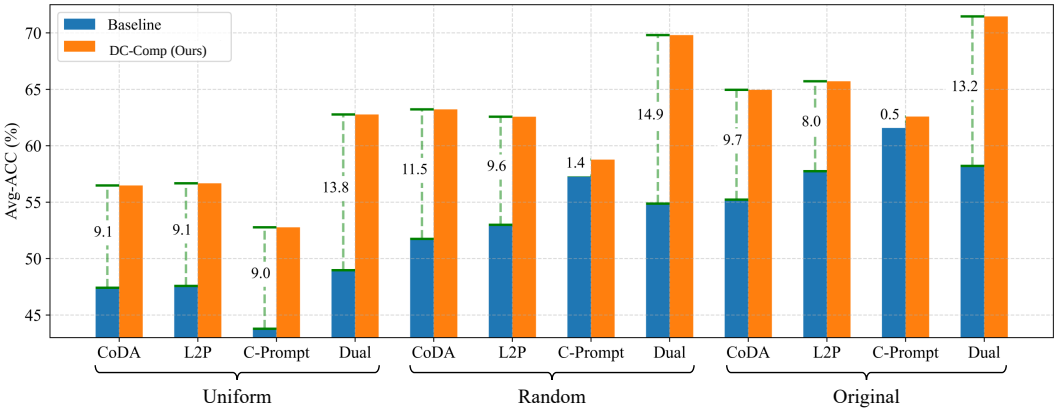


Figure 9: Adapting DC-Comp to different baselines on the **ImageNet-R** benchmark.

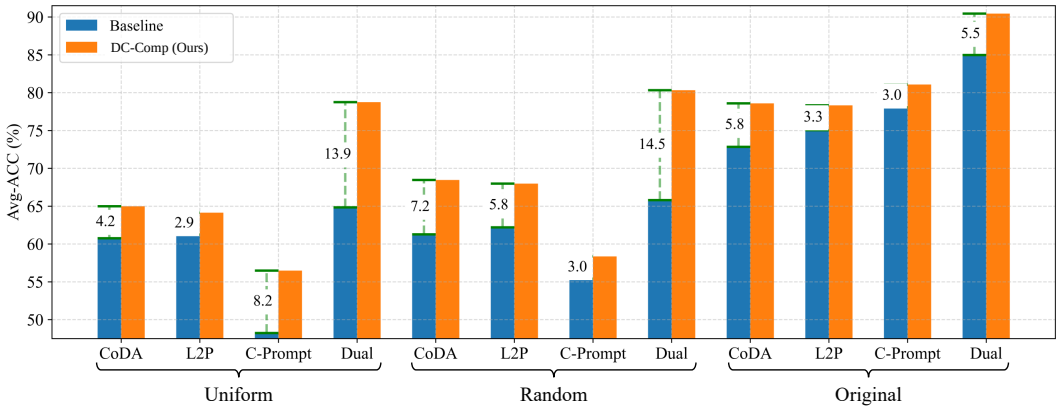


Figure 10: Adapting DC-Comp to different baselines on the **ImageNet-C** benchmark.

In the unified (Unif) long-tail setting, DC-Comp consistently improves all baselines. On DomainNet, it improves CODA-Prompt by 3.9 points, L2P-Prompt by 3.7 points, C-Prompt by 12.1 points, and Dual-Prompt by 6.1 points. Similarly, on ImageNet-R, the improvements range from 9.0 points (C-Prompt) to 13.8 points (Dual-Prompt), while on ImageNet-C, the gains range from 2.9 points (L2P-Prompt) to 13.9 points (Dual-Prompt). These results indicate that DC-Comp effectively mitigates the impact of unified long-tail distributions by enhancing the representation of minority classes.

In the random (Rand) long-tail setting, DC-Comp continues to deliver substantial improvements. On DomainNet, it improves CODA-Prompt by 4.1 points and C-Prompt by 6.0 points. On ImageNet-R, the method achieves an improvement of up to 11.5 points for CODA-Prompt, while on ImageNet-C, it improves CODA-Prompt by 7.2 points and Dual-Prompt by 14.5 points. These results demonstrate DC-Comp's adaptability to irregular class distributions, ensuring robust feature learning across heterogeneous domains.

For the original (Origin) datasets, DC-Comp still provides notable improvements. On DomainNet, it improves CODA-Prompt by 5.1 , L2P-Prompt by 6.0 points, C-Prompt by 0.8 points, and Dual-Prompt by 9.4 points. On ImageNet-R, the gains range from 0.6 points (C-Prompt) to 13.2 points (Dual-Prompt), while on ImageNet-C, the improvements include 3.3 points for L2P-Prompt and 5.5 points for Dual-Prompt. These results indicate that DC-Comp not only mitigates the impact of long-tail distributions but also strengthens overall model performance without long-tail subsampling, thereby enhancing domain-invariant feature learning.

The consistent performance improvements across all three datasets (DomainNet, ImageNet-R, and ImageNet-C) and data configurations (Unif, Rand, and Origin) underscore the robustness and adaptability of DC-Comp. Its ability to address both unified and random long-tail settings, as well as inherent cross-domain variations in the original datasets, highlights its capacity to learn discriminative and domain-invariant features effectively. These properties make DC-Comp applicable to visual-recognition tasks in which long-tail distributions and domain variations present simultaneous challenges.

D Performance Variation When Shifting from Balanced to Imbalanced Scenarios

To comprehensively investigate the influence of long-tail distributions relative to class-balanced scenarios, we introduce a sample-count-matched balanced (Balance) setting. This configuration avoids the confounding effect of reduced data volume in long-tailed subsets, thereby isolating the impact of class imbalance on performance degradation. Specifically, the Balance dataset maintains the same total data volume per domain as the Unif dataset while using a class-balanced distribution. As presented in Table 4, which compares model performance on the Balance and Unif subsets of DomainNet, ImageNet-R, and ImageNet-C, the long-tailed distribution still induces substantial performance degradation for most methods and datasets. C-P and KA-P denote integrations of a long-tail learning method with the C-Prompt and KA-Prompt baselines, respectively.

Specifically, under the Balance setting, the state-of-the-art CSA (KA-P) achieves 54.28 on DomainNet but drops sharply to 50.15 under Unif. Similarly, on ImageNet-R, the performance of the state-of-the-art ProCo (KA-P) declines from 57.76 (Balance) to 54.76 (Unif), resulting in a substantial drop of 3.00. On ImageNet-C, Dual-Prompt drops from 67.96 (Balance) to 64.83 (Unif). This trend is consistently observed across other methods. In contrast, DC-Comp exhibits superior robustness to long-tailed distributions, achieving the highest performance under both the Balance and Unif settings. Notably, DC-Comp exhibits the smallest drops (**2.17**, **0.58**, and **0.49** on DomainNet, ImageNet-R, and ImageNet-C, respectively) when shifting the data distribution from Balance to Unif, highlighting its effectiveness in mitigating long-tail challenges.

Table 4: Performance comparison under the balanced (Balance) and unified long-tailed (Unif) settings, with Drop indicating the performance difference.

Methods	**Pub.**	**DomainNet**			**ImageNet-R**			**ImageNet-C**		
		Balance↑	Unif↑	Drop↓	Balance↑	Unif↑	Drop↓	Balance↑	Unif↑	Drop↓
Dual-Prompt	*ECCV 2022*	50.80	47.70	<u>3.10</u>	51.51	48.97	2.54	<u>67.96</u>	<u>64.83</u>	<u>3.13</u>
CODA-Prompt	*CVPR 2023*	51.35	48.25	<u>3.10</u>	49.83	47.41	2.42	64.19	60.75	3.44
CPrompt	*CVPR 2024*	47.20	42.73	4.47	51.79	42.46	9.33	58.13	42.34	15.79
C-Prompt	*IJCV 2024*	47.68	40.29	7.39	48.74	43.78	4.96	56.85	48.23	8.62
KA-Prompt	*ICML 2025*	53.99	47.38	6.61	56.39	53.68	2.71	61.81	52.70	9.11
ProCo (C-P)	*PAMI 2024*	52.66	49.36	3.30	48.79	45.06	3.73	56.37	48.71	7.66
CSA (KA-P)	*NeurIPS 2023*	<u>54.28</u>	<u>50.15</u>	4.13	54.94	54.20	<u>0.74</u>	62.66	53.26	9.40
ProCo (KA-P)	*PAMI 2024*	47.94	41.78	6.16	<u>57.76</u>	<u>54.76</u>	3.00	61.18	52.68	8.50
DC-Comp	*This Paper*	**55.95**	**53.78**	**2.17**	**63.35**	**62.77**	**0.58**	**79.23**	**78.74**	**0.49**

E HEAD/TAIL ACCURACY ANALYSIS

To examine whether the overall gains of DC-Comp extend to different regions of the class-frequency spectrum, we report the final-stage Head and Tail accuracy in table 5. For every domain, the observed classes are ranked by their training frequencies and divided into three approximately equally sized groups. We compute macro-averaged accuracy over the classes with test samples in the highest- and lowest-frequency groups, and then average the resulting Head and Tail accuracy across all seen domains. This protocol avoids domination by frequent classes and applies the same group construction and aggregation rule to all methods.

Under the more persistent Unif imbalance, DC-Comp achieves the best Head and Tail accuracy on all three benchmarks. In particular, it improves Tail accuracy over the strongest competing method by **18.03**, **29.22**, and **24.23** on DomainNet, ImageNet-R, and ImageNet-C, respectively. These substantial gains indicate that DC-Comp does not obtain its overall advantage merely from dominant classes; instead, the preserved cross-domain discriminative semantics provide complementary evidence for chronically underrepresented classes. DC-Comp simultaneously improves Head accuracy by **6.92**, **24.50**, and **21.90** points, demonstrating that tail-class enhancement is achieved without sacrificing the retention of well-represented knowledge.

Under the dynamically changing Rand distribution, DC-Comp ranks first in five of the six Head/Tail comparisons and second in the remaining one. It yields particularly large Tail improvements on DomainNet and ImageNet-C (**4.56** and **23.55** points), while remaining competitive with KA-Prompt+LOS on ImageNet-R (58.93 versus 59.15). The consistently strong performance across both frequency groups supports the central design of DC-Comp: IDRC complements scarce current-domain evidence using discriminative semantics from earlier domains, while the information-preserving representation also limits the forgetting of previously learned head-class knowledge.

Table 5: Comparison of Head/Tail accuracy across different methods. "Unif" and "Rand" denote the class-aware cross-domain unified-distribution and class-independent cross-domain random-distribution settings, respectively. The best and second-best results under each setting are indicated by boldface and underlining.

Method	Pub.	DomainNet		ImageNet-R		ImageNet-C	
		Head	Tail	Head	Tail	Head	Tail
Unif setting							
Dual-Prompt + LOS	*ICLR 2025*	51.57	5.52	36.11	14.48	<u>63.48</u>	41.99
Dual-Prompt + SAMN	*CVPR 2026*	47.97	<u>18.18</u>	27.78	18.40	47.74	<u>43.86</u>
KA-Prompt + LOS	*ICLR 2025*	<u>56.51</u>	4.72	<u>43.67</u>	<u>23.56</u>	43.62	0.99
KA-Prompt + SAMN	*CVPR 2026*	52.73	13.64	33.44	17.41	10.14	8.93
DC-Comp	*This Paper*	**63.43**	**36.21**	**68.17**	**52.78**	**85.38**	**68.09**
Rand setting							
Dual-Prompt + LOS	*ICLR 2025*	52.67	36.28	53.65	16.30	<u>56.59</u>	50.29
Dual-Prompt + SAMN	*CVPR 2026*	50.60	39.62	44.88	13.60	54.71	<u>51.13</u>
KA-Prompt + LOS	*ICLR 2025*	<u>59.24</u>	41.95	<u>62.02</u>	**59.15**	42.06	32.34
KA-Prompt + SAMN	*CVPR 2026*	56.20	<u>45.82</u>	57.79	43.41	22.09	19.80
DC-Comp	*This Paper*	**63.06**	**50.38**	**71.32**	<u>58.93</u>	**79.81**	**74.68**

F Evaluation of Forgetting and Backward Transfer

We calculate the Forgetting Measure (FM) and Backward Transfer (BT) Xu et al. (2025a) for head and tail classes. The results show that DC-Comp outperforms the state-of-the-art Dual-Prompt+CSA, improving FM by **2.04/1.34 points** and BT by **1.20/5.11 points** for head/tail classes, respectively.

Table 6: Forgetting Measure (FM) and Backward Transfer (BT) on head and tail classes.

Method	Forgetting Measure ↓		Backward Transfer ↑	
	Head	Tail	Head	Tail
Dual-Prompt+CSA	3.25	2.11	1.12	5.12
Ours	**1.21**	**0.77**	**2.32**	**10.23**

G ALGORITHM OF THE PROPOSED METHOD

The detailed training process of DC-Comp is presented as pseudocode in Algorithm 1. When $t = 1$, we train only DMN+IBDM. When $t > 1$, we train each domain in the order DMN+IBDM $\rightarrow$ IDRC to improve training efficiency. Specifically, the model learns discriminative features using DMN and IBDM, after which discriminative tokens are sampled and added to the token memory. Finally, the tokens from the previous t domains are jointly used to optimize the attention layer and classifier.

Algorithm 1 Training process of Inter-Domain Intra-Class Knowledge Complement (DC-Comp).

Input: Current domain data D_t, token memory $\mathcal{M}_{t-1}$
Output: Trained model, token memory $\mathcal{M}_t$, prompt pool $\mathcal{P}_t$

for each epoch $e = 1$ to N_{epoch} **do**
  for each $(\boldsymbol{x}, y) \in D_t$ **do**

    # Discriminativity Modeling Network (DMN)
    # Discriminativity Estimation Module (DEM)
    Tokenize $\boldsymbol{x}$ into a sequence representation $\boldsymbol{h}_x \in \mathbb{R}^{L_h \times d}$;
    Retrieve the matched prompt $\boldsymbol{p}^x$;
    Extract the token sequence $\boldsymbol{h}_x^*$ using the pretrained ViT $\boldsymbol{\Theta}$ via Eq. (4);
    Compute discriminative scores $\{s(x_i)\}_{i=1}^{L_h} = \boldsymbol{\psi}(\boldsymbol{h}_x^*)$;

    # Intra-class Balanced Discriminativity Mining (IBDM)
    Calculate sampling rounds $r_x = \lfloor \frac{1}{p_x} \rfloor$;
    Generate $\mathcal{G}_x = \{aug_j\}_{j=1}^{r_x}$ via n_s-token sampling according to $\{s(x_i)\}_{i=1}^{L_h}$;
    Compute features $\boldsymbol{f}_j$ via Eq. (6);
    Update the model with $\mathcal{L}_{ce}$ in Eq. (7);
  end for
end for

 # Inter-domain Discriminativity Reserving and Complementing (IDRC)
for each $(\boldsymbol{x}, y) \in D_t$ **do**
  Calculate the balanced sampling size n_x^{ib} via Eq. (10);
  Sample n_x^{ib} discriminative tokens as $\boldsymbol{h}_x^{sel}$;
  Update $\mathcal{M}_t \leftarrow \mathcal{M}_{t-1} \cup \{\boldsymbol{h}_x^{sel}\}$;
end for

if $t > 1$ **then**
  for each class $c \in \mathcal{C}$ **do**
    for each instance token group $\boldsymbol{h}_i \in \mathcal{M}_{t-1}$ **do**
      Calculate the number of pseudo-samples r_x^s via Eq. (8);
      Generate pseudo-samples by sampling token combinations;
      Update the attention layer and classifier with pseudo-sample features via Eq. (11);
    end for
  end for
end if

Return the trained model, $\mathcal{M}_t$, and $\mathcal{P}_t$