Specifying What You Know or Not for Multi-Label Class-Incremental Learning

Aoting Zhang[1,3], Dongbao Yang[1,3*], Chang Liu[5], Xiaopeng Hong[4*], Yu Zhou[2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[3]School of Cyber Security, University of Chinese Academy of Sciences
[4]Harbin Institute of Technology
[5]Tsinghua University
{zhangaoting, yangdongbao}@iie.ac.cn, liuchang2022@tsinghua.edu.cn
hongxiaopeng@ieee.org, yzhou@nankai.edu.cn

Abstract

Existing class incremental learning is mainly designed for single-label classification task, which is ill-equipped for multi-label scenarios due to the inherent contradiction of learning objectives for samples with incomplete labels. We argue that the main challenge to overcome this contradiction in multi-label class-incremental learning (MLCIL) lies in the model's inability to clearly distinguish between known and unknown knowledge. This ambiguity hinders the model's ability to retain historical knowledge, master current classes, and prepare for future learning simultaneously. In this paper, we target at specifying what is known or not to accommodate Historical, Current, and Prospective knowledge for MLCIL and propose a novel framework termed as HCP. Specifically, (i) we clarify the known classes by dynamic feature purification and recall enhancement with distribution prior, enhancing the precision and retention of known information. (ii) We design prospective knowledge mining to probe the unknown, preparing the model for future learning. Extensive experiments validate that our method effectively alleviates catastrophic forgetting in MLCIL, surpassing the previous state-of-the-art by 3.3% on average accuracy for MS-COCO B0-C10 setting without replay buffers.

Introduction

To know what it is that you know, and to know what it is that you do not know–that is understanding.

—The Analects

Class incremental learning (CIL) (Aljundi et al. 2018; Douillard et al. 2022) is developed to continuously identify new classes while preserving old knowledge. Numerous studies endeavor to address the problem of catastrophic forgetting in CIL caused by the absence of old data. These studies are generally tailored to single-label class-incremental learning (SLCIL), assuming that each image only contains one single class. However, real-world images often feature multiple labels (*e.g.*, a street scene depicts cars, buses, persons, etc.). To this end, multi-label class incremental learning (MLCIL) has caught progressive attention (Dong et al. 2023), aiming to correctly classify an image into multiple classes that may be introduced across different sessions.

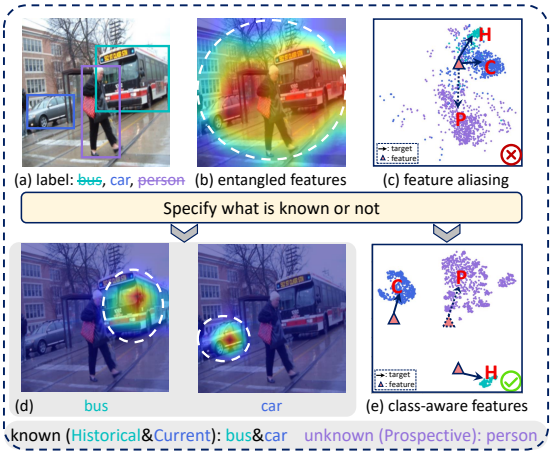


Figure 1: The contradiction of learning objectives in MLCIL arises from the model's inability to distinguish known and unknown knowledge. Current model fails to effectively recall prior known knowledge due to (a) the absence of historical labels, while (b) unknown classes' attention is inadvertently overlapped with known classes, and known classes are also entangled, resulting in (c) feature aliasing and contradictory learning objectives. By specifying what is known or not, (d) fine-grained class-aware features are focused, leading to (e) enhanced inter-class discriminability, alleviating the contradiction.

Different from SLCIL, MLCIL typically involves images that simultaneously contain objects from historical, current and prospective classes since the foreground class definition evolves between incremental sessions. Annotations are available only for the classes learned at the current session (car in Figure 1), leaving past (bus) and future (person) classes unlabelled. Directly adopting anti-forgetting techniques in SLCIL, such as knowledge distillation and exemplar replay, yields poor results, as they fail to tackle the contradiction of learning objectives caused by incomplete labels inherent to MLCIL. Specifically, the learning objective contradiction arises from the inconsistency among three critical

*Corresponding Author.

learning targets–preserving previously acquired knowledge, excelling in current classes, and preparing for future learning. Recent studies emphasize pseudo-labeling past classes and maintaining multiple knowledge to preserve previously acquired knowledge. For example, KRT (Dong et al. 2023) proposes a knowledge restoration and transfer framework to address the issues of known-class label absence. Despite the substantial progress, these methods ignore another aspect, namely the interference of future classes that the model does not know at the current session. Without "knowing" future classes, the model inadvertently activates these features into known class representations. As shown in Figure 1 (b), although the current model does not "know" person, it still pays high attention to the corresponding region and entangles the representations of person with known classes (car and bus). Meanwhile, due to the insufficient "knowing" ability, that is, fine-grained class-aware representation ability, past and current classes are also entangled in co-occurrence scenarios. Entangled features of historical, current and prospective classes blur the knowledge boundaries between tasks, resulting in the contradiction of learning objectives for aliased features. It not only impacts the learning for future classes but also degrades the recall of known knowledge due to the lack of target supervision, aggravating catastrophic forgetting.

In this work, we aim to specify what is known or not to accommodate Historical, Current, and Prospective knowledge for MLCIL and propose a novel framework termed as HCP. For clarifying known knowledge, HCP first proposes a dynamic feature purification module to capture fine-grained class-aware features, preventing feature aliasing across sessions. Specifically, each class is assigned a class embedding, activating relevant features within the image based on attention mechanism, and classes can be flexibly expanded by adding new embeddings. As shown in Figure 1 (d), our method focuses the attention of the bus embedding entirely on the bus itself and eliminates other noises. HCP then effectively recalls old known knowledge through pseudo-labeling with distribution prior, which alleviates the problem of large forgetting differences between classes. For probing unknown knowledge, we mine knowledge from images that encompass historical, current, and prospective classes to develop features pertinent to future classes. This prospective strategy helps the model learn a richer feature set and clearly defines the boundaries of current classes, thereby enhancing the discriminability of current features and preparing the model for future learning.

To summarize, our major contributions are as follows:

- We reveal the challenge of learning objectives contradiction in MLCIL, and propose a new framework named HCP, aiming to specify what is known or not to accommodate historical, current, and prospective knowledge.
- For clarifying the known, we develop dynamic feature purification that focuses on fine-grained class-aware features of known classes. In addition, we design recall enhancement with distribution prior to effectively preserve old known knowledge.
- For probing the unknown, we mine knowledge to develop features pertinent to future classes, boosting the model's discriminative capacity and preparing for future learning.
- Experiments on various settings demonstrate that our method achieves state-of-the-art performance and effectively mitigates catastrophic forgetting in MLCIL.

Related Work

Single-Label Incremental Learning aims to integrate new concepts without forgetting previously learned (Zhu et al. 2025). Current mainstream methods are typically divided into three categories. *Regularization-based methods* design a loss function to penalize changes in the weights or activations during learning new tasks (Schwarz et al. 2018; Huang et al. 2024b; Yang et al. 2022b). *Rehearsal-based methods* involve retaining a subset of previously encountered samples and merging them with new data for training (Bang et al. 2021; Chaudhry et al. 2018b; Huang et al. 2024a). Although they show impressive results, relying on a memory buffer raises concerns about the privacy of stored images and increases storage space. *Architectural-based methods* modify the network architecture by adding sub-networks or experts when new tasks arrive while keeping the previous network frozen (Douillard et al. 2022; Wang et al. 2022).

Multi-Label Classification has been a challenging problem compared with single-label classification. The straightforward approach is to treat each category independently and formulate it as multiple binary classification. However, it ignores the label correlation and spatial dependency between objects, which is important for multi-label. Therefore, several works (Wang et al. 2016; Chen et al. 2018) use the recurrent neural network (RNN) to capture label correlation, which face difficulty in parameter optimization. Others apply Graph Convolutional Network (GCN) (Zhou et al. 2020) to model label relationships (You et al. 2020; Chen et al. 2019), which capture spurious correlations when the label statistics are insufficient. Considering the binary cross entropy loss often suffers from the positive-negative imbalance issue, asymmetric loss (ASL) (Guo et al. 2019) is designed to dynamically down-weights and hard-thresholds easy negative examples. Recently, some approaches (Zhu et al. 2022; Li et al. 2023; Zhu et al. 2023) utilize transformer to model label correlation and improve multi-label prediction.

Multi-Label Class-Incremental Learning has gained widespread attention with the rapid development of class incremental learning and multi-label classification. PRS (Kim, Jeong, and Kim 2020) proposes a new sampling strategy for replay to alleviate the impact of imbalanced class distribution in buffers. OCDM (Liang and Li 2022) leverages greedy algorithm to update memory quickly and efficiently. AGCN (Du et al. 2023) utilizes a GCN network to build stable relationships between labels. KRT proposes a knowledge restore and transfer framework to solve the label absence of old classes. Although KRT achieves SOTA performance, the feature aliasing problem is still yet to be resolved, and the efficiency of KRT is impacted since it needs to go through the decoder multiple times to obtain task-level representation of each session. In this work, in addition to preserving previously acquired knowledge, we focus on accommodating his-

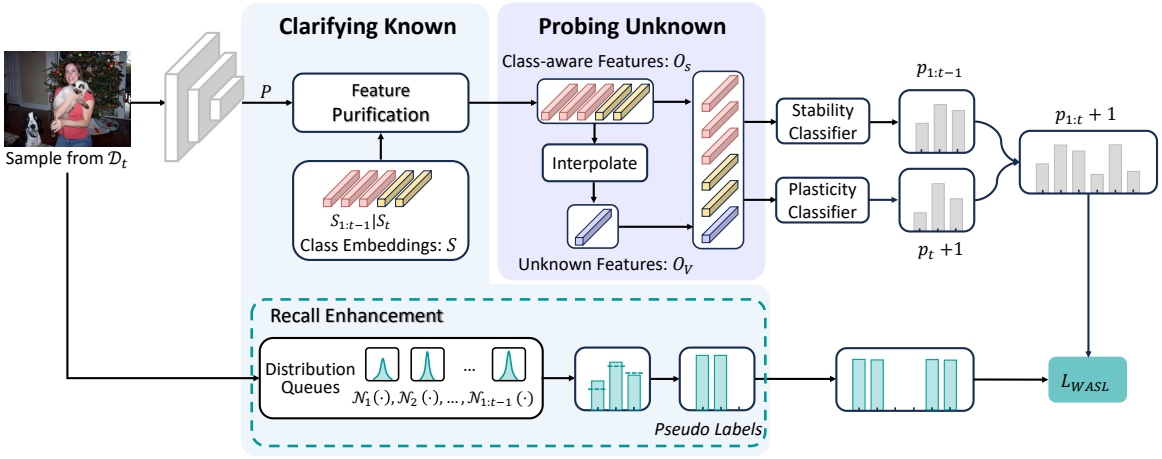


Figure 2: Framework of HCP, which leverages Clarifying Known and Probing Unknown to accommodate historical, current, and prospective knowledge. For clarifying known knowledge, we design dynamic Feature Purification to capture fine-grained class-aware features O_s to avoid feature aliasing across sessions, and Recall Enhancement with distribution prior to effectively retain historical known knowledge. For probing unknown knowledge, we interpolate known features as prospective class to help enrich the feature set, enhancing the discriminability of known features and facilitate future learning.

torical, current and prospective knowledge simultaneously, overcoming the problem of learning target contradiction.

Proposed Method

Problem Formulation

The objective of MLCIL is to build a unified model that can recognize all encountered classes. We assume total T sessions to simulate the continuous process. Given the dataset $D = \{(x, y)\}$, where x is the image with corresponding ground-truth label y, we split the dataset into T subsets according to the lexicographical order of category names, with $\{D_1, \ldots, D_T\}$ and their label sets $\{C_1, \ldots, C_T\}$. All label sets are mutually exclusive, *i.e.* $\forall m, n (m \neq n), C_m \cap C_n = \emptyset$. At session t, the model is trained on only D_t, with label space C_t. Different from SLCIL where each image has only one label, ground-truth y in the multi-label scenario may contain classes from old sessions $C_{1:t-1}$ and future sessions $C_{t+1:T}$., so we preserve only the labels belonging to C_t and discard others. During testing, the model is evaluated to recognize all seen classes $C_{1:t} = C_1 \cup \cdots \cup C_t$.

Overview

Our proposed framework HCP is illustrated in Figure 2. The key idea of HCP is to specify what is known or not at the current incremental session to accommodate historical, current, and prospective knowledge, thereby resolving learning target contradiction. Specifically, to clarify the known knowledge, the HCP framework initially introduces a dynamic feature purification module, where each class embedding focuses on fine-grained class-aware features without covering multiple classes, avoiding feature aliasing across sessions. It can flexibly introduce new classes in incremental learning by continuously adding new class embeddings. Moreover, we enhance the recall of historical knowledge by effectively utilizing priors of previous models, alleviating the problem of large forgetting differences between classes. To probe the unknown knowledge, we interpolate class features as a prospective class, which pushes all other non-target class features away from the generated component. As a result, the features of known classes are optimized to be more compact, facilitating future learning.

Clarifying Known Knowledge

In this section, we introduce feature purification module and how it adapts to multi-label incremental task. We then analyse the confidence forgetting between classes and further propose recall enhancement with distribution prior.

Feature Purification. To avoid feature aliasing between sessions caused by noise and impurity from non-target features, we propose feature purification module to extract fine-grained class-aware features from entangled multi-class global features. Compared with KRT, it ensures unique representations for each class and predicts historical and current classes in parallel.

Given a sample from D_t, global features are first extracted and then reshaped to patch tokens $P \in \mathbb{R}^{L \times d}$, where $L = h \cdot w$ and h, w, d represent the height, width and dimension. To aggregate the object information and extract fine-grained class features, each class is assigned a learnable embedding and we get a sequence of class embeddings $S \in \mathbb{R}^{M \times d}$, where $M = |C_{1:t}|$ is the number of known classes at session t. Feature purification module which consists of L multi-head self-attention blocks, takes class embeddings S and feature tokens P as input to generate purified class features $O_S \in \mathbb{R}^{M \times d}$ and enhanced patch features $O_P \in \mathbb{R}^{M \times d}$ (we omit the mini-batch):

$$(Q, K, V) = (W_q, W_k, W_v)[P, S], \tag{1}$$

$$O = W_o \text{softmax}\left(\frac{QK^T}{\sqrt{d/h}}\right) V + b_o, \tag{2}$$

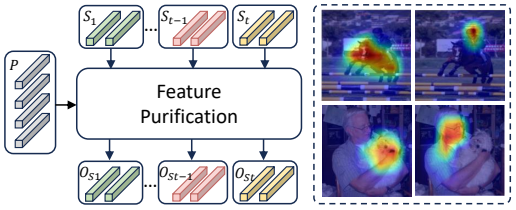


Figure 3: Illustration of feature purification. Each session appends new class embeddings S_t for new class features O_{St}.

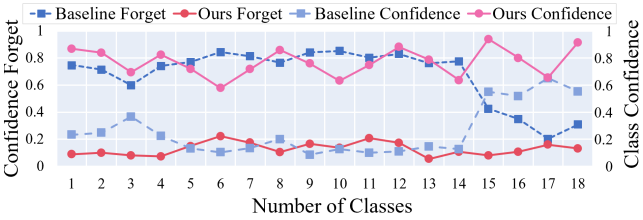


Figure 4: Confidence forgetting varies greatly among classes, making it difficult to effectively recall known knowledge by a unified and static pseudo-label threshold.

where $O = [O_P, O_S]$ and h is the number of attention heads. Based on the attention mechanism, each class embedding not only pays attention to the spatial information of global feature tokens, but also captures contextual relationships with other class embeddings. Then class features O_S are fed to the classifier to obtain the output logits p.

As shown in Figure 3, it can flexibly adapt to the learning of new sessions by appending new class embeddings, which is more efficient where old and new classes can be predicted in parallel. Following previous work (Yan, Xie, and He 2021), we utilize stability classifier to predict old logits $p_{1:t-1}$ and plasticity classifier to get new classification logits p_t, which are merged into a complete logits $p_{1:t} = [p_{1:t-1}, p_t]$ for classification. During learning new classes, we freeze old embeddings $S_{1:t-1} \in \mathbb{R}^{|C_{1:t-1}| \times d}$ and stability classifier to maintain the old knowledge while new class embeddings $S_t \in \mathbb{R}^{|C_t| \times d}$ and plasticity classifier can adapt to new data. Before entering the next session $t+1$, the plasticity classifier and stability classifier are combined to form a new stability classifier, and then a new plasticity classifier is created for the new session.

Recall Enhancement. Since the historical model has obtained effective information of old classes, we leverage the class probabilities $P = [p_1, \cdots, p_{|1:C_{t-1}|}]$ predicted by the old model to recall historical knowledge (Yang et al. 2022a, 2023a,b). The missing past labels $\hat{y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{|1:C_{t-1}|}]$ can be obtained: $\hat{y}_k = 1$ if $p_k \geq \varepsilon$ otherwise 0, where ε controls the quality of pseudo-labels. However, category concepts present different learning difficulties for the model. Some classes have clear and easily distinguishable features therefore less-forgetting, while others may be subtle therefore more-forgetting. Treating all classes uniformly brings numerous noises, leading to the dilemma of either mislabeling false positive labels or neglecting true positive labels.

To illustrate this problem, we perform a quantitative analysis of the forgetting of class confidence which reflects how certain the model is about predictions. After performing incremental learning ($B10 - C2$ protocol, explained in Experiments) on VOC datasets, we calculate the classification confidence distribution of each old class $p_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, where μ_k is the mean confidence and σ_k denotes variance:

$$\mu_k = \frac{1}{|D_k|} \sum_{i=1}^{|D_{t-1}|} p_{ik} \cdot \mathbb{1}(y_{ik} = 1), \tag{3}$$

$$\sigma_k^2 = \frac{1}{|D_k|} \sum_{i=1}^{|D_{t-1}|} \left(p_{ik} \cdot \mathbb{1}(y_{ik} = 1) - \mu_k\right)^2, \tag{4}$$

where $|D_k|$ is the number of samples where the true label for class k is present. p_{ik} is the predicted logit for class k of the i-th sample. $\mathbb{1}$ is the indicator function (1 when class k is present, and 0 otherwise). Then, following the accuracy forgetting in SLCIL (Chaudhry et al. 2018a), we define *confidence forgetting* of each class as the difference between the maximum confidence gained throughout the past learning sessions and the confidence after finishing the current learning session:

$$F_k = \max_{t \in \{1, \cdots, T-1\}} (\mu_{t,k} - \mu_{T,k}), \tag{5}$$

where $\mu_{t,k}$ is the mean confidence of class k on task t when the model complete learning task t. As shown in Figure 4, the classification confidence of baseline suffers from severe forgetting up to 0.85, and the degree of forgetting varies greatly between classes, which makes it difficult to determine an optimal threshold for all classes. Here, we introduce a class-specific strategy to formulate different thresholds for each class based on the confidence distribution learned by the previous model. According to the statistical principle, 3σ rule can guarantee the diversity of pseudo labels. For the representative examples of old classes, we define thresholds as mean confidence $\varepsilon_k = \mu_k$ and get the class-specific pseudo labels for training in session t:

$$\hat{y}_k = \begin{cases} 1, & p_k \geq \varepsilon_k, \\ 0, & p_k < \varepsilon_k. \end{cases} \tag{6}$$

Since class distributions drift with sessions, we update distribution queues after completing each session to mitigate performance degradation caused by error accumulation.

Probing Unknown Knowledge

Previous works (Song et al. 2023; Zhou et al. 2022) have validated that learning extra classes at each session can reserve embedding space for future classes and enhance the model's forward compatibility. A straightforward way is to introduce real classes from other datasets as auxiliary, which is impractical due to access issues and potential domain discrepancies that may affect the current learning session. In multi-label scenarios, where images can encompass both known and latent unknown classes, we propose to mine knowledge in existing data to synthesize simulated features of future classes, thereby enriching the feature set. This prospective strategy pushes all features of the known classes presented

in the current sample away from synthetic features, leading to a more compact and optimized representation of known classes, which better facilitates future learning.

For an image in training data D_t, it has ground-truth label $Y = [y_1, \cdots, y_M] \in \mathbb{R}^M$ where $M = |C_{1:t}|$, $y_i = 1$ means i-th class exists in the image. Based on the dynamic feature purification, we obtain all class features $O_S \in \mathbb{R}^{|C_{1:t}| \times d}$, including features of known classes present and absent in current image, denoted as $O_{S1} = \{o_i | y_i = 1, o_i \in O_S\} \in \mathbb{R}^{M_1 \times d}$ and $O_{S2} = \{o_i | y_i = 0, o_i \in O_S\} \in \mathbb{R}^{M_2 \times d}$, respectively. The attention of present classes are localized to the fine-grained regions under label supervision, while attentions of other absent class features are freely distributed in other foreground regions. We leverage the implicit information in O_{S2} to expand the model's feature set. Specifically, we randomly interpolate these absent class features to synthesize extra features $O_V \in \mathbb{R}^{1 \times d}$ as an unknown class:

$$O_V = \sum_{i=1}^{M_2} \bar{\lambda}_i \cdot o_i, \bar{\lambda}_i = \lambda_i / \textstyle\sum_{j=1}^{M_2} \lambda_j, \tag{7}$$

where λ_i is randomly sampled from Beta distribution. Unknown features O_V are then fed into the classifier jointly with real class features O_S for classification results $P \in \mathbb{R}^{(M+1)}$. To this end, the original M-class problem is extended to a $(M+1)$-class. The label of unknown features is also a binary label to indicate whether it is real or synthetic. Combined with the stability and plasticity classifiers, we attach unknown classification to the plasticity classifier, which predicts $|C_t| + 1$ classes. We must admit that the attention of absent classes may be paid to present classes due to inter-class similarity. However, the random synthetic features can simulate the distribution of future classes, and treating it as an additional class improves the model's ability to identify subtle differences, enhancing the compact representation of real classes and reserving space for future learning.

Loss Function

In long-term MLCIL tasks, pseudo-labels for previous classes rise with incremental sessions, leading to a disproportionate ratio of old to new class labels. Consequently, the model's pace of learning new classes tends to decelerate, which is particularly pronounced during the early training session. To sharpen the model's focus on new classes, we increase the weight of the classification loss for the new classes and adopt asymmetric loss (ASL). Given an image, we obtain its class probabilities $P = [p_1, \cdots, p_K] \in \mathbb{R}^K$, where there are $K = |\mathcal{C}_{1:t}| + 1$ classes including unknown class. The weighted ASL loss is as follows:

$$L_{WASL} = \frac{1}{K} \sum_{k=1}^{K} w_k \cdot \begin{cases} (1-p_k)^{\gamma+} \log(p_k), & y_k = 1, \\ (p_k)^{\gamma-} \log(1-p_k), & y_k = 0, \end{cases} \tag{8}$$

where w_k is set to $\sqrt{\frac{|C_{1:t}|}{|C_t|}}$ for new classes and otherwise to 1. y_k is the binary label to indicate whether label k is present or not. $\gamma+$ and $\gamma-$ are used to manipulate the impact of positive and negative samples.

Experiments

Experimental Setups

Datasets and Protocols. HCP is evaluated on MS-COCO 2014 (Lin et al. 2014) and PASCAL VOC 2007 (Everingham 2007) datasets. MS-COCO contains 82,081 training images and 40,137 test images, which covers 80 common objects with an average of 2.9 labels per image. PASCAL VOC contains 5,011 images in the train-val set, and 4,952 images in the test set. It covers 20 common objects, with an average of 1.6 labels per image. Similar to CIL works, we define different MLCIL protocols by a unified notation $Bi-Cj$, where i denotes the number of classes learned in the base session and j is the number of classes to be learned in each subsequent incremental session. We perform $B40-C10$ and $B0-C10$ protocols on MS-COCO dataset and $B10-C2$ and $B0-C4$ protocols on VOC 2007. We compare our method HCP with several baselines, representative SLCIL methods and state-of-the-art SLCIL methods with and without replay buffers.

Evaluation Metrics. Similar to CIL, we adopt two widely used metrics for evaluation: average accuracy (Avg. Acc) and last accuracy (Last Acc). Following KRT, we use the mean average precision (mAP) to evaluate all the categories that have been learned in each session and report the average mAP (the average of the mAP of all sessions) and the last mAP (final session mAP). Two more metrics are all reported for a comprehensive multi-label performance evaluation, $i.e.$, the per-class F1 measure (CF1) and overall F1-measure (OF1) alongside the last accuracy.

Implementation Details. For fair comparison, we follow the experimental setting in KRT and use TResNetM (Ridnik et al. 2021) pre-trained on ImageNet-21k (Deng et al. 2009) as the backbone. We train the model with a batch size of 64 for 20 epochs, using Adam (Kingma and Ba 2014) optimizer and OneCycleLR scheduler with a weight decay of 1e-4. In the base session, we set the learning rate to 4e-5. In the following sessions, it adjusts to 1e-4 for MS-COCO and 4e-5 for VOC. In dynamic feature purification module, we set 3 attention blocks for VOC and 1 for MS-COCO. Our codes are available at https://github.com/InfLoop111/HCP.

Comparison Results

Results on MS-COCO. As shown in Table 1, fine-tuning (FT) and SLCIL methods like PODNet suffer from severe forgetting in MLCIL tasks, with last accuracy 16.9% and 25.6% respectively, while our method achieves 71.2% on B0-C10 when the buffer size is set to 0. Similarly, multi-label online incremental learning methods do not perform well, and our method outperforms PRS and OCDM by a large margin. Compared with SOTA MLCIL methods, our method still maintains a leading position, with improvements of up to 3.8% in average accuracy over KRT (buffer size=1000) and a greater increase across all metrics compared to AGCN. Consistency improvements in B40-C10 setting underline the robustness of our HCP. It is notable that even without replay, our method exceeds all others with memory buffers in both average accuracy and last accuracy. Figure 5(a) and (b) show the performance curves as the number of classes increases. As incremental session progresses,

Method	Source Task	Buffer Size	MS-COCO B0-C10				MS-COCO B40-C10			
			Avg. Acc	Last Acc			Avg. Acc	Last Acc		
			mAP (%)	CF1	OF1	mAP (%)	mAP (%)	CF1	OF1	mAP (%)
Upper-bound	Baseline	-	-	76.4	79.4	81.8	-	76.4	79.4	81.8
FT	Baseline	0	38.3	6.1	13.4	16.9	35.1	6.0	13.6	17.0
PODNet	SLCIL		43.7	7.2	14.1	25.6	44.3	6.8	13.9	24.7
AGCN	MLCIL		72.4	53.9	56.6	61.4	73.9	58.7	59.9	69.1
KRT	MLCIL		74.6	55.6	56.5	65.9	77.8	64.4	63.4	74.0
HCP	MLCIL		**77.9**	**60.4**	**65.3**	**71.2**	**78.9**	**64.9**	**68.6**	**75.3**
TPCIL	SLCIL	5/class	63.8	20.1	21.6	50.8	63.1	25.3	25.1	53.1
PODNet	SLCIL		65.7	13.6	17.3	53.4	65.4	24.2	23.4	57.8
DER++	SLCIL		68.1	33.3	36.7	54.6	69.6	41.9	43.7	59.0
AGCN	MLCIL		72.9	56.7	58.5	63.6	74.5	59.8	61.3	69.7
KRT	MLCIL		75.8	60.0	61.0	68.3	78.0	66.0	65.9	74.3
HCP	MLCIL		**79.4**	**70.3**	**72.9**	**74.5**	**79.4**	**71.5**	**74.1**	**76.7**
iCaRL	SLCIL	20/class	59.7	19.3	22.8	43.8	65.6	22.1	25.5	55.7
BiC	SLCIL		65.0	31.0	38.1	51.1	65.5	38.1	40.7	55.9
TPCIL	SLCIL		69.4	51.7	52.8	60.6	72.4	60.4	62.6	66.5
PODNet	SLCIL		70.0	45.2	48.7	58.8	71.0	46.6	42.1	64.2
DER++	SLCIL		72.7	45.2	48.7	63.1	73.6	51.5	53.5	66.3
AGCN	MLCIL		73.2	59.5	60.3	66.0	75.2	64.1	65.2	71.7
KRT	MLCIL		76.5	63.9	64.7	70.2	78.3	67.9	68.9	75.2
HCP	MLCIL		**79.6**	**70.4**	**73.0**	**74.6**	**79.6**	**71.9**	**74.5**	**77.2**
PRS	MLOIL	1000	48.8	8.5	14.7	27.9	50.8	9.3	15.1	33.2
OCDM	MLOIL		49.5	8.6	14.9	28.5	51.3	9.5	15.5	34.0
AGCN	MLCIL		73.0	59.4	65.9	59.0	75.0	63.1	64.8	71.1
KRT	MLCIL		75.7	61.6	63.6	69.3	78.3	67.5	68.5	75.1
HCP	MLCIL		**79.5**	**70.2**	**72.8**	**74.4**	**79.5**	**71.8**	**74.4**	**76.7**

Table 1: Performance on MS-COCO, with comparison methods categorized by different source tasks. Buffer size 0 indicates no rehearsal is required, rendering many SOTA SLCIL approaches inapplicable. Best results among rows are highlighted in **bold**.

Method	Buffer Size	VOC B0-C4		VOC B10-C2	
		Avg. Acc	Last Acc	Avg. Acc	Last Acc
Upper bound	-	-	93.6	-	93.6
FT		82.1	62.9	72.9	43.0
iCarL	2/class	87.2	72.4	79.0	66.7
BiC		86.8	72.2	81.7	69.7
TPCIL		87.6	77.3	80.7	70.8
PODNet		88.1	76.6	81.2	71.4
DER++		87.9	76.1	82.3	70.6
KRT		90.7	83.4	87.7	80.5
HCP		**93.5**	**89.2**	**92.1**	**86.3**
HCP	0	92.9	87.9	90.1	81.9

Table 2: Comparison results on PASCAL VOC dataset.

FP	RE	PU	Sessions						Avg. Acc
			1	2	3	4	5	6	
			97.58	89.10	86.30	61.31	56.16	46.76	72.87
✓			97.64	90.60	87.26	70.73	66.29	66.29	82.09
✓	✓		97.84	94.17	90.72	83.44	76.39	73.45	86.00
✓		✓	97.57	93.11	90.31	84.77	75.39	71.64	85.47
✓	✓	✓	97.80	94.70	90.84	89.86	85.26	81.85	**90.05**

Table 3: Ablation study of each component. The module names are abbreviated as follows: FP-Feature Purification, RE-Recall Enhancement, PU-Probing Unknown.

our method exhibits stronger superiority, which demonstrates our effectiveness in long-term incremental scenarios.

Results on PASCAL VOC. Table 2 shows consistent improvements on VOC. For B0-C4 protocol, when buffer size is set to 0, our method outperforms KRT with 2/class replay in buffers by 2.2% in Avg. Acc and 4.5% in Last Acc respectively, which demonstrates the superiority of our method in scenarios with limited data access. HCP further achieves the best average mAP value of 93.5% with extra buffers. For B10-C2 setting, HCP also exceeds other competitive methods by up to 4.4% and 5.8% and reaches the upper bound performance. Comparative curves in Figure 5(c) and (d) exhibit a widening gap between ours and other methods.

Ablation Study

Effectiveness of Component. All ablation experiments are conducted on VOC B10-C2 setting. Tab. 3 reports results of Last Acc, Avg. Acc. Fine-tuning the old model with new data serves as baseline. After adding feature purification (FP), the model aggregates fine-grained features from multi-

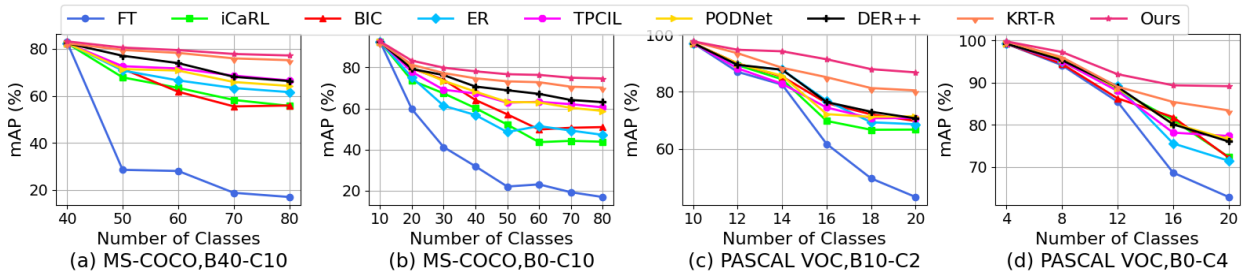
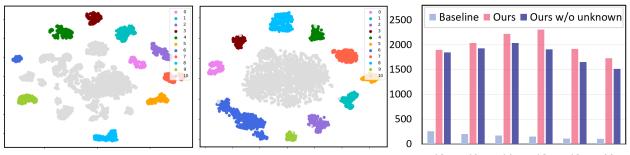


Figure 5: Performance curves (mAP%) on MS-COCO and PASCAL VOC datasets under different protocols.

Recall Strategy	Sessions						Avg. Acc
	1	2	3	4	5	6	
$\varepsilon = 0.8$	97.85	92.79	90.40	78.22	70.01	68.94	83.08
$\varepsilon = 0.9$	97.87	93.30	90.81	82.10	75.69	70.82	85.08
Top-2	97.87	92.72	90.27	80.87	74.10	68.94	84.24
RE	97.80	94.70	90.84	89.86	85.26	81.85	**90.05**

Table 4: Ablation of recall strategy, where the first two rows use a unified threshold, the third row represents top-K filtering, and RE is our Recall Enhancement.



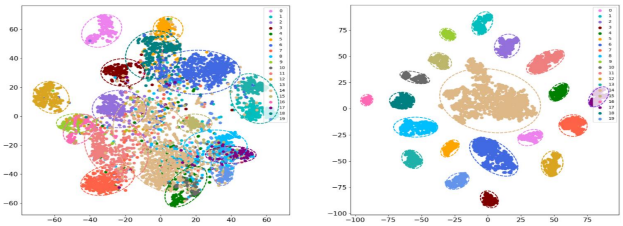
(a) Real features (b) Generated features (c) C-H Indexes

Figure 6: The distribution of (a) real future features can be simulated by (b) generated unknown features (gray in color). (c) Calinski-Harabasz (C-H) Indexes of class features.

object images, and we get a 9.22% Avg. gain. Recall Enhancement (RE) effectively recalls old known knowledge and alleviates forgetting difference among classes, which obtains a 13.13% Avg. gain. For probing unknown knowledge, mining knowledge to generate features as unknown class enhances discriminability between real classes, boosting the performance of Avg. Acc by 12.60%. Three modules jointly achieve the best, verifying the effectiveness of specifying what is known or not to accommodate historical, current and prospective knowledge.

Ablation of Recall Strategy. We compare different pseudo label selection strategies to recall past knowledge, including a unified threshold ε, top-K filtering and enhancement with distribution prior. Table 4 shows that results are sensitive to the threshold. Our recall enhancement (RE), which considers the forgetting difference between classes and automatically determines the optimal threshold for each class, provides supervision with higher quality for old classes.

Analysis of Feature Aliasing. In Figure 7, we visualize the feature distributions of the baseline and our method after incremental learning on VOC dataset. It can be seen that the baseline method exhibits severe inter-class confusion and



(a) Baseline (b) Ours

Figure 7: t-SNE visualization after incremental learning.

experiences catastrophic forgetting, while our method distinctly separates all classes without feature aliasing.

Analysis of Generated Unknown Features. Figure 6 (a) (b) compares the real future features and generated unknown features (gray in color), which both promote compact representation of current features. Besides, our interpolated features are very similar to the distribution of real future ones. This observation verifies the effectiveness of our probing unknown knowledge, which can provide valuable foresight for future learning. For quantitative illustration, we report the Calinski-Harabasz Index of feature representations in Figure 6 (c), where a higher index indicates better separation between classes and compactness within classes. The index of ours at each session is significantly higher than the baseline. Moreover, without probing the unknown, the index drops a lot in the later sessions, which illustrates the advantage of virtual features for preparing the model for future learning.

Conclusion

In this paper, we present a novel method HCP for multi-label class-incremental learning, which specifies what is known or not at current learning session to accommodate historical, current and prospective knowledge. To clarify the known knowledge, feature purification is proposed to capture class-aware features from entangled global features, preventing feature aliasing within and between sessions. We analyze the confidence forgetting and further design recall enhancement to effectively retain historical known knowledge. To probe the unknown, we interpolate class features as prospective class to enhance the discriminative capacity and prepare for future learning. This provides a fresh insight into multi-label CIL problems. Comparisons with previous methods and our ablation study demonstrate the superiority of our overall design and the importance of each component in HCP.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant NO 62406318, 62376266, 62076195, 62376070), and by the Key Research Program of Frontier Sciences, CAS (Grant NO ZDBS-LY-7024).

References

Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, 139–154.

Bang, J.; Kim, H.; Yoo, Y.; Ha, J.-W.; and Choi, J. 2021. Rainbow Memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8218–8227.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018a. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision*, 532–547.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2018b. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Chen, S.-F.; Chen, Y.-C.; Yeh, C.-K.; and Wang, Y.-C. 2018. Order-free rnn with visual attention for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Dong, S.; Luo, H.; He, Y.; Wei, X.; Cheng, J.; and Gong, Y. 2023. Knowledge restore and transfer for multi-label class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18711–18720.

Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9285–9295.

Du, K.; Lyu, F.; Li, L.; Hu, F.; Feng, W.; Xu, F.; Xi, X.; and Cheng, H. 2023. Multi-label continual learning using augmented graph convolutional network. *IEEE Transactions on Multimedia*.

Everingham, M. 2007. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html.*

Guo, H.; Zheng, K.; Fan, X.; Yu, H.; and Wang, S. 2019. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 729–739.

Huang, L.; An, Z.; Zeng, Y.; Xu, Y.; et al. 2024a. KFC: Knowledge reconstruction and feedback consolidation enable efficient and effective continual generative learning. In *The Second Tiny Papers Track at ICLR 2024*.

Huang, L.; Zeng, Y.; Yang, C.; An, Z.; Diao, B.; and Xu, Y. 2024b. eTag: Class-incremental learning via embedding distillation and task-oriented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12591–12599.

Kim, C. D.; Jeong, J.; and Kim, G. 2020. Imbalanced continual learning with partitioning reservoir sampling. In *Proceedings of the European Conference on Computer Vision*, 411–428.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, M.; Wang, D.; Liu, X.; Zeng, Z.; Lu, R.; Chen, B.; and Zhou, M. 2023. PatchCT: Aligning patch set and label set with conditional transport for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15348–15358.

Liang, Y.-S.; and Li, W.-J. 2022. Optimizing class distribution in memory for multi-label online continual learning. *arXiv preprint arXiv:2209.11469*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Ridnik, T.; Lawen, H.; Noy, A.; Ben Baruch, E.; Sharir, G.; and Friedman, I. 2021. TResNet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, 1400–1409.

Schwarz, J.; Czarnecki, W.; Luketina, J.; Grabska-Barwinska, A.; Teh, Y. W.; Pascanu, R.; and Hadsell, R. 2018. Progress & Compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, 4528–4537.

Song, Z.; Zhao, Y.; Shi, Y.; Peng, P.; Yuan, L.; and Tian, Y. 2023. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24183–24192.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. CNN-RNN: A unified framework for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2285–2294.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Yan, S.; Xie, J.; and He, X. 2021. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3014–3023.

Yang, D.; Zhou, Y.; Hong, X.; Zhang, A.; and Wang, W. 2023a. One-shot replay: Boosting incremental object detection via retrospecting one object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3127–3135.

Yang, D.; Zhou, Y.; Hong, X.; Zhang, A.; Wei, X.; Zeng, L.; Qiao, Z.; and Wang, W. 2023b. Pseudo object replay and mining for incremental object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 153–162.

Yang, D.; Zhou, Y.; Shi, W.; Wu, D.; and Wang, W. 2022a. RD-IOD: Two-level residual-distillation-based triple-network for incremental object detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(1): 1–23.

Yang, D.; Zhou, Y.; Zhang, A.; Sun, X.; Wu, D.; Wang, W.; and Ye, Q. 2022b. Multi-view correlation distillation for incremental object detection. *Pattern Recognition*, 131: 108863.

You, R.; Guo, Z.; Cui, L.; Long, X.; Bao, Y.; and Wen, S. 2020. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12709–12716.

Zhou, D.-W.; Wang, F.-Y.; Ye, H.-J.; Ma, L.; Pu, S.; and Zhan, D.-C. 2022. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9046–9056.

Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1: 57–81.

Zhu, X.; Cao, J.; Ge, J.; Liu, W.; and Liu, B. 2022. Two-stream transformer for multi-label image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3598–3607.

Zhu, X.; Liu, J.; Liu, W.; Ge, J.; Liu, B.; and Cao, J. 2023. Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1473–1482.

Zhu, Z.; Hong, X.; Ma, Z.; Zhuang, W.; Ma, Y.; Dai, Y.; and Wang, Y. 2025. Reshaping the Online Data Buffering and Organizing Mechanism for Continual Test-Time Adaptation. In *European Conference on Computer Vision*, 415–433.