Confidence Self-Calibration for Multi-Label Class-Incremental Learning

Kaile Du[1], Yifan Zhou[1], Fan Lyu[2], Yuyang Li[1], Chen Lu[1], and Guangcan Liu[1]

[1] School of Automation, Southeast University, China
[2] NLPR, MAIS, CASIA, China

Abstract. The partial label challenge in Multi-Label Class-Incremental Learning (MLCIL) arises when only the new classes are labeled during training, while past and future labels remain unavailable. This issue leads to a proliferation of false-positive errors due to erroneously high confidence multi-label predictions, exacerbating catastrophic forgetting within the disjoint label space. In this paper, we aim to refine multi-label confidence calibration in MLCIL and propose a Confidence Self-Calibration (CSC) approach. Firstly, for label relationship calibration, we introduce a class-incremental graph convolutional network that bridges the isolated label spaces by constructing learnable, dynamically extended label relationship graph. Then, for confidence calibration, we present a max-entropy regularization for each multi-label increment, facilitating confidence self-calibration through the penalization of over-confident output distributions. Our approach attains new state-of-the-art results in MLCIL tasks on both MS-COCO and PASCAL VOC datasets, with the calibration of label confidences confirmed through our methodology. Our code is available at `https://github.com/Kaile-Du/CSC`.

Keywords: Multi-label class-incremental learning · Partial label · Max-entropy regularization · Class-incremental graph convolutional network

1 Introduction

Class-incremental learning (CIL) [14, 15, 27, 28, 31, 45, 54] refers to learning a sequence of new classes and being able to preserve the knowledge learned well. In CIL, the model learns only from samples associated with the incremental classes, leading to the challenge of catastrophic forgetting [32], where adaptation to new classes often results in a significant decline in performance on the old ones. Most CIL methods assume single-label annotations for inputs and employ techniques to mitigate single-label aware forgetting [1,20,29,43]. In contrast, there is limited exploration of multi-label class-incremental learning (MLCIL), which aims to continually adapt to the increasingly larger label space through non-forgetting learning. MLCIL focuses solely on learning the incremental classes, resulting in the absence of past and future labels, a phenomenon referred to as the *task-level partial label* [15] or category-incomplete [10] issue. A toy-like MLCIL depiction

---

Training label space C^1 Test label space C^1

task 1: f_{θ^1}

cat   cat

Training label space C^2 Test label space $C^{1:2}$

task 2: f_{θ^2}

dog   cat, dog   cat

Training label space C^3 Test label space $C^{1:3}$

task 3: f_{θ^3}

person   cat, dog, person   person

(a) Multi-label class-incremental learning   (b) The qualitative results

PODNet — dog / person — F-P

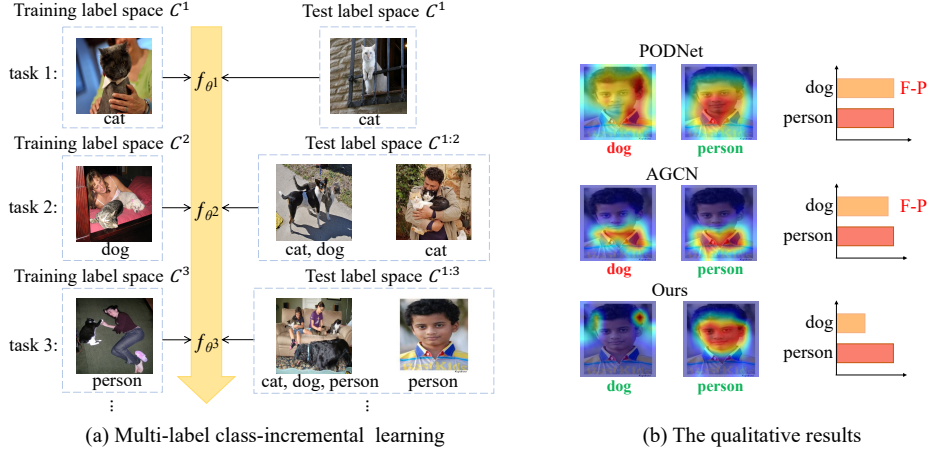AGCN — dog / person — F-P

Ours — dog / person

Fig. 1: (a) A simple depiction of multi-label class-incremental learning. There are a total of three labels, including "cat", "dog" and "person", which are sequentially learned and tested across three tasks. (b) After training on task 3, given a test image only with "person" processing by f_{θ^3}. Left: attention maps for the labels "dog" and "person" using the SOTA methods PODNet [12], AGCN [15] and ours. Right: The confident scores for the three methods. The false-positive (F-P) error occurs on the "dog" using PODNet and AGCN. The bottom shows the effectiveness of our method.

is shown in Fig. 1 (a). Given three labels in total, including "cat", "dog" and "person", which are sequentially learned and tested across three tasks. During task 2, only the current task class ("dog") is labeled at training time, while the classes from the previous ("cat") and future ("person") tasks remain unlabeled.

Partial label issue makes the new and old label spaces naturally disjoint. The previous efforts [10, 15] have overlooked the need to calibrate the MLCIL model, often resulting in an over-confident output distribution [35,38] with false-positive (F-P) errors [42]. As shown in Fig. 1 (b), even when presented with a test image containing only a person, SOTA methods like PODNet [12] and AGCN [15] still output high confidence for both old class "dog" and new class "person". This phenomenon indicates an F-P error and becomes rigorous with the enlargement of the label space of MLCIL. The erroneous over-confident multi-label predictions from poorly calibrated models will degrade performance on old classes, exacerbating catastrophic forgetting under the disjoint label space.

In this paper, we aim to calibrate MLCIL model and propose a Confidence Self-Calibration (CSC) approach under the partial label setting, comprising two components: (1) a class-incremental graph convolutional network (CI-GCN) structure for label relationship calibration and (2) a max-entropy regularization for confidence self-calibration. Firstly, while relationships between labels are beneficial in multi-label classification [8,50,51], with the partial label issue, it becomes challenging to construct cross-task relationships due to the missing labels from disjoint label space. To address this issue, we propose the CI-GCN

to calibrate the label relationships. The label relationship graph constructed by CI-GCN is continuously learned through gradient updates. Under the partial label challenge, CI-GCN not only preserves label relationships in the old label space but also consistently expands to establish relationships in the new label space. Secondly, miscalibrated confidences result in misclassifications in multi-label problems. To mitigate this issue, we propose a max-entropy regularization to penalize over-confident output distribution, effectively reducing the F-P rate to facilitate confidence self-calibration. Our main contributions include:

– A CI-GCN is designed for label relationship calibration, which bridges isolated label spaces by establishing cross-task label relationships. The dynamically extended label relationship graph is continually learned through gradient updates.
– A max-entropy regularization is proposed to facilitate the confidence self-calibration by penalizing over-confident output distribution, effectively reducing the false-positive errors.
– Extensive experimental results on MS-COCO and PASCAL VOC datasets demonstrate that the proposed method distinctly outperforms recent state-of-the-art approaches in MLCIL.

2   Related Work

Single-Label Class-Incremental Learning. Recent advances have been made in single-label class-incremental learning. For example, the regularization-based methods [19,21,33,52,54] incorporate a regularization term into the loss function to learn new classes without forgetting the old ones. EWC [19] utilizes a Fisher matrix to retain critical parameters from previous tasks. Then, oEWC [40] proposes an improvement to the calculation of parameter importance. The replay-based methods [3,9,24,26,30,36,39,41,49] overcome forgetting by storing partial data of learned classes in a small buffer and employing suitable sampling techniques. ER [39] trains with samples from old tasks in the buffer alongside new task samples. Building upon this, DER++ [1] integrates rehearsal with knowledge distillation and regularization, while iCaRL [36] prevents forgetting by selecting exemplars with the herding technique. And the architecture-based approaches, which decompose a model into task-sharing and task-specific components [13,48], have also been proposed. Moreover, L2P [47] introduces prompt-based learning [25] into class-incremental learning and achieves promising results based on a pre-trained ViT model [11].
Multi-Label Classification. Multi-label classification has garnered widespread attention. Relation-based methods [5–8,44,46] aim to leverage the label dependencies or region relations for image classification. Earlier approaches utilize RNNs [44] or LSTMs [46] to model the relationship between labels and semantics sequentially, while state-of-the-art methods leverage spatial information or graphical architectures. For example, Chen et al. [8] propose a multi-label classification model based on graph convolutional network to generate inter-dependent

object classifiers. KGGR [5] exploits prior knowledge to guide adaptive information propagation among categories to facilitate multi-label analysis, while Zhao et al. [51] construct complementary relationships by exploring structural and semantic correlation graphs. Our work is inspired by the GCN-based methods.

Multi-Label Class-Incremental Learning. With the ongoing research in the fields of class-incremental learning and multi-label classification, multi-label class-incremental learning has emerged as a new area of interest. OCDM [22] improves the replay-based method by formulating the memory update mechanism as an optimization problem, while PRS [17] proposes a novel sampling strategy for replay-based approaches to mitigate the impact of long-tail distribution in multi-label datasets. Dong et al. [10] propose a knowledge restoration and transfer framework to address the issues of label absence. AGCN [15] utilizes pseudo labels to construct fixed statistical label relationships and mitigate catastrophic forgetting. Note that KRT [10] is the current SOTA method. However, these methods have failed to recognize the significant occurrence of F-P errors caused by over-confident output distribution, thus lacking in model calibration.

3   Method

3.1   Problem Formulation

Following [10, 15], given T incremental learning tasks, with the corresponding training set denoted as $\{\mathcal{D}_{\text{trn}}^1, \cdots, \mathcal{D}_{\text{trn}}^T\}$ and the testing set as $\{\mathcal{D}_{\text{tst}}^1, \cdots, \mathcal{D}_{\text{tst}}^T\}$. For each incremental state t, the training dataset is defined as $\mathcal{D}_{\text{trn}}^t = \{(x_i^t, y_i^t)\}$, where x_i^t is the i-th training sample, and $y_i^t \in \mathcal{Y}_{\text{trn}}^t$ denotes the corresponding ground-truth labels. The training label space $\mathcal{Y}_{\text{trn}}^t = \mathcal{C}^t$, and $\mathcal{C}^t$ denotes the current class collection, and the past class collection is represented by $\mathcal{C}^{1:t-1} = \bigcup_{i=1}^{t-1} \mathcal{C}^i$. $\mathcal{C}^{1:t} = \mathcal{C}^{1:t-1} \cup \mathcal{C}^t$, and $\mathcal{C}^{1:t-1} \cap \mathcal{C}^t = \varnothing$. The testing label space $\mathcal{Y}_{\text{tst}}^t$ is equivalent to $\mathcal{C}^{1:t}$ for the t-th task, formulated as $\mathcal{Y}_{\text{tst}}^t = \mathcal{C}^{1:t}$. Under the task-level partial label condition, we intend to calibrate the model based on the connected label space. Our goal is to establish relationships between new and old task labels to connect the label space, and then calibrate the over-confident output distribution on this basis.

3.2   Overall Framework

We propose a Confidence Self-Calibration (CSC) framework under the partial label setting, comprising two major components: (1) A stacked **Class-Incremental Graph Convolutional Network** (CI-GCN) structure (Sec. 3.3), inspired by the GCN in traditional multi-label learning [8, 50, 51]. CI-GCN first employs a feature extractor to obtain class-agnostic feature maps, followed by the stacked structure to generate cross-task graph representations for calibrating label relationships. (2) A **Max-Entropy Regularization** (Sec. 3.4) applied to the old classes to calibrate multi-label confidence in MLCIL. When CSC outputs predictions, it combines class-related features with graph representations to classify all the seen labels so far under the partial label condition.
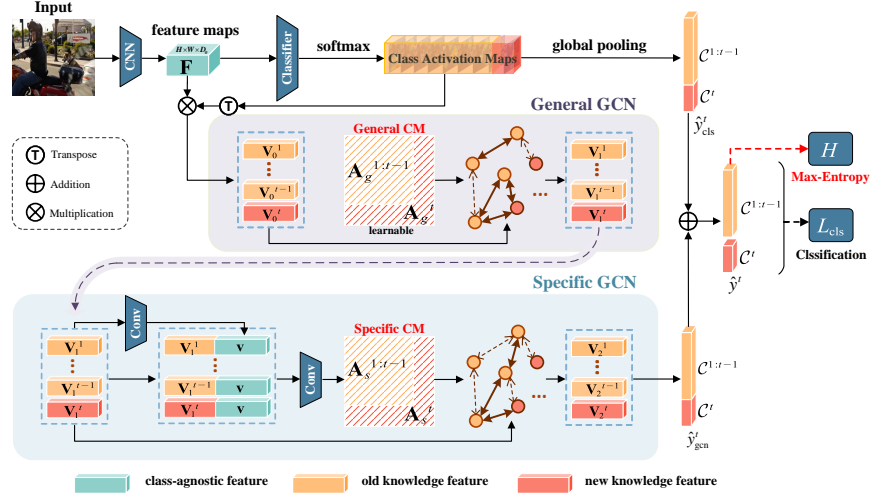
Fig. 2: Overview of confidence self-calibration framework. Given an image, we first employ a CNN backbone to extract the class-agnostic feature maps $\mathbf{F}$. Then, inspired by CAM [53], we use the class activation maps as masks to decouple $\mathbf{F}$ into class-aware label representations $\mathbf{V}_0$. Our model constructs general and specific graphs among $\mathbf{V}_0$ to generate $\mathbf{V}_2$, which aggregates rich cross-task label relationships. Finally, we combine graph-based representations $\hat{y}_{\mathrm{gcn}}^t$ with class-related outputs $\hat{y}_{\mathrm{cls}}^t$ for classification.

3.3   Class-Incremental Graph Convolutional Network

Utilizing graph structure to construct label relationships has been proven effective in GCN-based methods [7,8,50,51]. These methods [7,8,15,50,51] rely on the input of graph nodes, $i.e.$, multiple labels, and a correlation matrix (CM) to provide graph edges for label relationships. However, these approaches necessitate prior statistical counting of label co-occurrences across the entire label set to construct the CM. Given the partial label setting, access to past and future labels is restricted, posing a challenge in constructing cross-task CM.

As shown in Fig. 2, CI-GCN is a stacked structure that does not depend on prior information, comprising one layer of general GCN and one layer of specific GCN. A learnable general CM is input into the general GCN, capturing general class dependencies by learning new and old task knowledge within the model through gradient updates during training, thus allowing the general GCN to automatically construct cross-task label relationships without relying on statistical relationships. Note that the general CM is shared for all test images in the same task during testing. Subsequently, the specific GCN is a finer structure to provide image-dependent relationships, where the specific CM generated by each image may vary. The CMs in the two GCNs will continually and automatically expand as the number of learned classes increases. General CM and specific CM capture general and specific class dependencies, respectively, thereby aiding in complementing missing label information from the overall to the detailed level.

Constructing CMs. Before the first task, we randomly initialize the general CM $\mathbf{A}_g^1$, which will expand and be updated in the subsequent tasks. A simple way to address partial supervision in MLCIL is to use pseudo labels from the previous model. Inspired by the old and new knowledge tokens from KRT [10], we divide the general CM into two parts shown in Fig. 2: the old task CM $\mathbf{A}_g^{1:t-1}$ is inherited from the previous tasks, preserving old task label relationships, and the new task CM $\mathbf{A}_g^t$ establishes relationships in the new task label space. We employ the new supervision from the training set and the old supervision, *i.e.* pseudo-labels, from previous model, to jointly update $\mathbf{A}_g$ through gradient descent, with lr as the learning rate and L representing the loss function, formulated as:

$$\mathbf{A}_g = \mathbf{A}_g - lr * \frac{\partial L}{\partial \mathbf{A}_g}, \quad \text{where} \quad \frac{\partial L}{\partial \mathbf{A}_g} = \frac{\partial L}{\partial \mathbf{V}_2}\frac{\partial \mathbf{V}_2}{\partial \mathbf{A}_g} = \frac{\partial L}{\partial \mathbf{V}_2}\frac{\partial \mathbf{V}_2}{\partial \mathbf{V}_1}\frac{\partial \mathbf{V}_1}{\partial \mathbf{A}_g}, \quad (1)$$

enabling $\mathbf{A}_g$ to learn cross-task label relationships, where L takes both ground truth and pseudo-labels as input. $\mathbf{V}_1$ and $\mathbf{V}_2$ denote the outputs of the general and specific GCN shown in Fig. 2, respectively, which will be discussed next.

In previous GCN-based multi-label classifications [7,8,15,50,51], the CMs are shared across all images, and directly counting pseudo-labels like AGCN [15] will lead to significant error accumulation. Motivated by this, we design an adaptive CM that does not rely on pseudo-label statistics, providing unique relationships for each image without being significantly affected by error accumulation, thus to mitigate catastrophic forgetting. As shown in Fig. 2, after augmenting the general CM, the specific CM $\mathbf{A}_s$ is adaptively estimated from input graph nodes $\mathbf{V}_1$, automatically expanding with the expansion of $\mathbf{V}_1$. As shown in Fig. 2, the specific CM is also divided into new and old parts, offering greater flexibility to adapt to continually changing CIL compared to fixed statistical CMs without compromising the performance of rare label combinations. In detail, we begin by performing global pooling and applying a convolutional layer to $\mathbf{V}_1$ to derive its global feature $\mathbf{v} \in \mathbb{R}^{D_1}$. This feature is then replicated for $|\mathcal{C}^{1:t}|$ times and concatenated with $\mathbf{V}_1$ shown in Fig. 2, yielding $\mathbf{V}_1' \in \mathbb{R}^{|\mathcal{C}^{1:t}|\times 2D_1}$. Subsequently, we pass $\mathbf{V}_1'$ through another convolutional layer with parameters $\mathbf{W} \in \mathbb{R}^{2D_1 \times |\mathcal{C}^{1:t}|}$ to compute the sample-specific CM $\mathbf{A}_s \in \mathbb{R}^{|\mathcal{C}^{1:t}|\times|\mathcal{C}^{1:t}|}$:

$$\mathbf{A}_s = \sigma(\mathbf{V}_1'\mathbf{W}), \quad (2)$$

where σ is the $Sigmoid(\cdot)$ function.

General and Specific GCN. As shown in Fig. 2, the CI-GCN takes class-related label representations $\mathbf{V}_0$ as graph nodes, calculated by:

$$\mathbf{V}_0 = \mathbf{M}^\top \otimes \mathbf{F}, \quad (3)$$

where $\mathbf{F} \in \mathbb{R}^{H\times W\times D_0}$ is class-agnostic feature maps, and $\mathbf{V}_0 \in \mathbb{R}^{|\mathcal{C}^{1:t}|\times D_0}$. The class activation maps $\mathbf{M} \in \mathbb{R}^{H\times W\times|\mathcal{C}^{1:t}|}$ are generated by the convolution classifier with $Softmax(\cdot)$ normalization shown in Fig. 2. Given the two CMs computed by Eqs. (1) and (2), the general and specific GCN can be defined as:

$$\mathbf{V}_1 = LReLU(\mathbf{A}_g\mathbf{V}_0\mathbf{W}_g), \quad \mathbf{V}_2 = LReLU(\mathbf{A}_s\mathbf{V}_1\mathbf{W}_s), \quad (4)$$

(a) The quantitative results
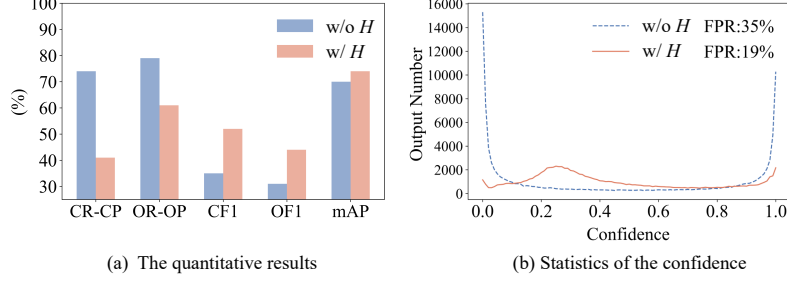
(b) Statistics of the confidence

Fig. 3: (a) The quantitative results without or with max-entropy based on CSC framework after training on the final task. CR-CP represents per-class recall minus precision and OR-OP represents overall recall minus precision. (b) The corresponding multi-label confidence statistics on the entire test set, as the information entropy H increases from 1.61 (w/o H) to 5.36 (w/ H) in {B4-C2} of VOC 2007.

where $LReLU(\cdot)$ represents the activation function LeakyReLU, which makes the whole operation nonlinear. $\mathbf{W}_g \in \mathbb{R}^{D_0 \times D_1}$ and $\mathbf{W}_s \in \mathbb{R}^{D_1 \times D_2}$ denote the state-update parameters for general and specific GCN. $\mathbf{V}_1 \in \mathbb{R}^{|\mathcal{C}^{1:t}| \times D_1}$ is the output of the general GCN, serving as the input notes for the specific GCN. To enhance model robustness, the specific GCN is employed to generate sample-specific label relationships for each image under the partial label condition. $\mathbf{V}_2 \in \mathbb{R}^{|\mathcal{C}^{1:t}| \times D_2}$ represents the output of the specific GCN. With the two GCNs, we complete the missing labels via two designed CMs. However, as shown in Fig. 3, we observe that the recall is generally much higher than precision, indicating that the confidence for each label is generally high, leading to many F-P errors. This phenomenon is also very evident in other CIL methods [1,12,15,21,39]. We propose to calibrate this multi-label confidence via an entropy regularization.

3.4 Max-Entropy Regularization

Building upon the calibrated label relationships, we proceed to calibrate multi-label confidence. Given that the current task classes are partially labeled, the model struggles to effectively discern between labeled and unlabeled class features. Consequently, the output distribution becomes over-confident, leading to numerous F-P errors.

From the quantitative perspective: when incrementally training multiple labels, we observe a significant imbalance between the precision and recall values of the model (often with precision markedly lower than recall) shown in Fig. 3 (a) CR-CP and OR-OP w/o H. This discrepancy indicates a high frequency of F-P errors from an alternate perspective. From the qualitative perspective: we define $P(l_{\mathrm{person}}|l_{\mathrm{tie}})$ as the conditional probability of the "person" label appearing when the "tie" label is present. Given the proximity of the relationship between these two labels, $P(l_{\mathrm{person}}|l_{\mathrm{tie}})$ tends to be high [8,50,51]. Suppose the "person" label originates from another task and remains unlabeled, while the "tie" is labeled. During training on the "tie", the model tends to conflate the features of

"tie" and "person". Consequently, during the testing phase, even if only "person" is present in the image, the model tends to predict high values for both labels, resulting in an F-P error occurring on "tie". As more such label combinations are trained during class-incremental learning, the output distribution develops multiple peaks and becomes over-confident, leading to low uncertainty.

High confidence predictions correspond to output distributions with low uncertainty and vice versa [34,35]. Under the partial label setting, Shannon entropy serves as a well-defined metric to quantify uncertainty. Thus, we measure uncertainty for each old class in MLCIL using the following formula in task t:

$$H = -\mathbb{E}_{c \in \mathcal{C}^{1:t-1}} \left[\hat{y}_c^t \log(\hat{y}_c^t) \right], \tag{5}$$

where c denotes the class index, and $\hat{y}_c^t$ represents each confident score output by the model for task t. During the model optimization process, we negate the entropy H to achieve maximum entropy regularization on the output distribution. The MLCIL model is calibrated by penalizing over-confident output distribution using the max-entropy regularization. Fig. 3 (b) shows the statistics of confidence without or with max-entropy. Our method significantly reduces the F-P rate (FPR) from 35% to 19%, thus to mitigate catastrophic forgetting. As a result, the max-entropy boosts the CF1, OF1, and mAP metrics shown in Fig. 3 (a).

3.5 Loss Function

For the output, depicted in Fig. 2, we derive cross-task graph representations $\hat{y}_{\text{gcn}}^t \in \mathbb{R}^{|\mathcal{C}^{1:t}|}$ by passing $\mathbf{V}_2$ through a fully connected layer, aiding in the final classification $\hat{y}^t$ by calibrating the label relationships. Then, the multi-label confidence in task t is calculated by:

$$\hat{y}^t = \hat{y}_{\text{cls}}^t + \hat{y}_{\text{gcn}}^t, \tag{6}$$

where class-related outputs $\hat{y}_{\text{cls}}^t \in \mathbb{R}^{|\mathcal{C}^{1:t}|}$ are obtained by passing class activation maps $\mathbf{M}$ through global pooling. Similar to [2,10,15], we combine cross-entropy with knowledge distillation (KD) to formulate our classification loss function:

$$L_{\text{cls}} = \underbrace{-\alpha\mathbb{E}_{c \in \mathcal{C}^t} \left[y_c^t \log(\hat{y}_c^t) \right]}_{\text{cross-entropy}} - \underbrace{(1-\alpha)\mathbb{E}_{c \in \mathcal{C}^{1:t-1}} \left[\hat{y}_c^{t-1} \log(\hat{y}_c^t) \right]}_{\text{knowledge distillation}}, \tag{7}$$

where α represents the loss weight, y_c^t denotes the ground truth for $c \in \mathcal{C}^t$ and $\hat{y}_c^t$ signifies the prediction in task t. $\hat{y}_c^{t-1}$ represents the predictions from the previous task, utilized to mitigate catastrophic forgetting. Notably, $\hat{y}_c^{t-1}$ incorporates unlabeled supervisory information regarding old classes due to the partial label issue, rendering it more effective in multi-label tasks than in single-label ones, as previously verified in [2, 10, 15]. Because $\mathbf{A}_g$ can learn old task relationships from KD, it can more effectively suppress catastrophic forgetting.

Combining the H in Eq. (5), our final loss function can be defined as:

$$L = L_{\text{cls}} - \beta H, \tag{8}$$

where β is used to control the strength of max-entropy regularization.

4   Experiments

4.1   Experimental Setting

In this section, we provide a quantitative comparison between our approach and current state-of-the-art methods, along with comprehensive ablation studies.
Datasets. We follow the experimental setting of [10], using MS-COCO 2014 [23] and PASCAL VOC 2007 [16] datasets to demonstrate the effectiveness of our method in MLCIL. MS-COCO serves as a widely used benchmark for multi-label classification. It consists of 82,081 images for training and 40,504 for validation with 80 classes. The label average number for every image is 2.9. PASCAL VOC 2007 dataset comprises 20 classes. The training set contains 5011 images, the test set contains 4952 images. On average, each image contains 2.4 labels.

Similar to these CIL works [4, 10, 12], we define different MLCIL scenarios using {Bx-Cy}, where "x" represents the number of classes trained in the base task, and "y" denotes the number of classes trained in each subsequent incremental task. We evaluate our method on MS-COCO with two different scenarios, i.e., {B0-C10 and B40-C10} as well as two more challenging scenarios, {B0-C5 and B20-C4}. Similarly, for VOC 2007, we evaluate our method with two scenarios, i.e., {B0-C4 and B10-C2} as well as two more challenging scenarios, {B5-C3 and B4-C2}. The incremental learning is the lexicographical order of class names, which is the same as [10].
Evaluation Metrics. Similar to [10, 12], we adopt the widely used average accuracy (Avg.Acc) and last accuracy (Last Acc) for CIL task. Following the MLCIL works [10, 15], we use the per-class F1 (CF1), overall F1 (OF1) as well as the mean average precision (mAP) to evaluate the MLCIL model. The mAP is divided into the average mAP and the last mAP, which means the average of the mAP of all tasks and the final task mAP, respectively. During the testing phase, the label space for the last task is $\mathcal{Y}_{\text{tst}}^T = \mathcal{C}^{1:T}$, thus evaluation is conducted on the entire test set for the last task.
Implementation Details. For a fair comparison, we follow the experimental setting in [10] for all experiments. We adopt TResNetM [37] pre-trained on ImageNet-21k as the backbone. We train the model with a batch size of 64 for MS-COCO 2014 and PASCAL VOC 2007. We optimize the network using Adam [18] and OneCycleLR scheduler with a weight decay of 1e-4. We use an initial learning rate of 3e-5 for the base model and 5e-5 for the following tasks on MS-COCO, and use the learning rate of 4e-5 for all tasks on PASCAL VOC. We use the same data augmentation of [10].

4.2   Comparison Methods

We select several advanced methods for comprehensive comparison.
Baseline Methods. Following KRT [10] and AGCN [15], we select Fine-Tuning and Joint as baseline methods. Fine-Tuning performs training without any anti-forgetting technique. It can be regarded as the performance lower bound. Joint is the upper bound, which trains the model jointly using all task data.

Table 1: Class-incremental results on MS-COCO datasets. Buffer size of 0 means no replay is applied and the * symbol denotes the results reported in [10], same as below.

Method	Method Type	Buffer Size	MS-COCO B0-C10				MS-COCO B40-C10			
			Last Acc			Avg.Acc	Last Acc			Avg.Acc
			mAP	CF1	OF1	mAP	mAP	CF1	OF1	mAP
Joint	Baseline	-	81.8	76.4	79.4	-	81.8	76.4	79.4	-
Fine-Tuning	Baseline		16.9	6.1	13.4	38.3	17.0	6.0	13.6	35.1
oEWC [40]	SL		24.3	6.7	13.4	46.9	27.3	11.1	16.5	44.8
LwF [21]	SL		42.4	45.3	43.7	61.2	51.7	47.0	45.7	64.8
AGCN [15]	ML	0	61.4	53.9	56.6	72.4	69.1	58.7	59.9	73.9
KRT* [10]	ML		65.9	55.6	56.5	74.6	74.0	64.4	63.4	77.8
CSC	ML		**72.8**	**64.9**	**66.8**	**78.0**	**75.0**	**65.7**	**67.0**	**78.2**
PODNet [12]	SL		53.4	13.6	17.3	65.7	57.8	24.2	23.4	65.4
DER++ [1]	SL		54.6	33.3	36.7	68.1	59.0	41.9	43.7	69.6
AGCN-R [15]	ML	5/class	63.6	56.7	58.5	72.9	69.7	59.8	61.3	74.5
KRT-R* [10]	ML		68.3	60.0	61.0	75.8	74.3	66.0	65.9	78.0
CSC-R	ML		**73.7**	**67.3**	**68.1**	**79.2**	**75.6**	**68.1**	**69.0**	**78.4**
ER [39]	SL		47.2	40.6	43.6	60.3	61.6	58.6	61.1	68.9
PODNet [12]	SL		58.8	45.2	48.7	70.0	64.2	46.6	42.1	71.0
DER++ [1]	SL	20/class	63.1	45.2	48.7	72.7	66.3	51.5	53.5	73.6
AGCN-R [15]	ML		66.0	59.5	60.3	73.2	71.7	64.1	65.2	75.2
KRT-R* [10]	ML		70.2	63.9	64.7	76.5	75.2	67.9	68.9	78.3
CSC-R	ML		**74.8**	**67.8**	**68.6**	**79.6**	**76.7**	**68.2**	**69.4**	**78.7**
PRS [17]	ML		27.9	8.5	14.7	48.8	33.2	9.3	15.1	50.8
OCDM [22]	ML		28.5	8.6	14.9	49.5	34.0	9.5	15.5	51.3
AGCN-R [15]	ML	1000	59.0	59.4	65.9	73.0	71.1	63.1	64.8	75.0
KRT-R* [10]	ML		69.3	61.6	63.6	75.7	75.1	67.5	68.5	78.3
CSC-R	ML		**73.9**	**67.5**	**68.5**	**79.3**	**76.0**	**67.8**	**69.7**	**78.5**

Single-Label Class-Incremental Methods. We run six single-label (SL) class-incremental learning methods with our MLCIL setting, including the most representative regularization-based methods oEWC [40] and LwF [21], and the widely recognized replay-based methods ER [39], PODNet [12] and DER++ [1]. We also compare with the recent SLCIL prompt-based method L2P [47] based on ViT-B/16 backbone, to show the gap from the respective upper bound.
Multi-Label Class-Incremental Methods. We select four best-performing multi-label (ML) class-incremental methods to compare, including replay-based methods PRS [17] and OCDM [22], the regularization-based methods AGCN [15] and KRT [10]. Note that KRT [10] is the current SOTA. We compare the results with those reported in KRT [10] under strictly the same experimental setting.
Our Method. CSC is our method without a replay buffer, and CSC-R is CSC with the replay buffer for a fair comparison with replay-based methods.

4.3   Comparison Results

MS-COCO. The results in {B0-C10 and B40-C10} of MS-COCO are shown in Tab. 1. CSC outperforms other advanced methods in all scenarios on MS-COCO. We have the following observations: 1) CSC achieves the best final mAP

Table 2: Class-incremental results on PASCAL VOC datasets.

Method	Buffer Size	VOC B0-C4		VOC B10-C2		VOC B5-C3		VOC B4-C2	
		Last mAP	Avg.mAP	Last mAP	Avg.mAP	Last mAP	Avg.mAP	Last mAP	Avg.mAP
Joint	-	93.6	-	93.6	-	93.6	-	93.6	-
Fine-Tuning		62.9	82.1	43.0	70.1	49.4	74.4	37.0	60.4
AGCN [15]	0	73.4	84.3	65.1	79.4	69.2	80.9	53.4	72.1
CSC		**85.1**	**90.4**	**83.8**	**89.0**	**82.1**	**88.0**	**74.1**	**83.3**
ER [39]	2/class	67.7	82.6	64.3	78.8	65.3	79.0	57.9	73.7
PODNet [12]		72.3	84.1	66.2	80.2	70.3	81.4	62.5	75.7
DER++ [1]		74.4	85.0	66.2	80.7	68.1	78.0	61.6	77.0
AGCN-R [15]		76.0	86.5	69.3	82.8	71.8	82.1	59.3	74.3
KRT-R* [10]		83.4	90.7	80.5	87.7	-	-	-	-
CSC-R		**87.9**	**92.4**	**87.8**	**91.6**	**87.5**	**91.9**	**86.6**	**90.4**

Table 3: Comparative experimental results under different backbones.

Method	Backbone	Param.	MS-COCO B0-C10				MS-COCO B40-C10			
			Last Acc			Avg.Acc	Last Acc			Avg.Acc
			mAP	CF1	OF1	mAP	mAP	CF1	OF1	mAP
Joint	ViT-B/16	86.0M	83.2	78.3	80.5	-	83.2	78.3	80.5	-
L2P [47]			68.0	58.8	56.0	73.4	71.1	62.8	62.3	73.7
Joint	TResNetM	29.4M	81.8	76.4	79.4	-	81.8	76.4	79.4	-
KRT* [10]			65.9	55.6	56.5	74.6	74.0	64.4	63.4	77.8
CSC			**72.8**	**64.9**	**66.8**	**78.0**	**75.0**	**65.7**	**67.0**	**78.2**

of **74.8%** and **76.7%** when the buffer size is 20/class. 2) When the buffer size gets smaller (5/class), our method outperforms the SOTA method KRT more: the improvements on final mAP, CF1, and OF1 are **5.4%** (68.3%→73.7%), **7.3%** (60.0%→67.3%), and **7.1%** (61.0%→68.1%) in {B0-C10} scenario; **1.3%** (74.3%→75.6%), **2.1%** (66.0%→68.1%) and **3.1%** (65.9%→69.0%) in {B40-C10} scenario. 3) It is evident that {B0-C10} scenario is more challenging than {B40-C10}. In more challenging scenarios, our method is superior to others. When the buffer size is set to 0, our method achieves a more significant leap in performance, surpassing the current SOTA by a wide margin, improved by **6.9%** (65.9%→72.8%), **9.3%** (55.6%→64.9%), and **10.2%** (56.6%→66.8%) on the mAP, CF1, and OF1 in {B0-C10}, respectively. And 4) compared to KRT-R (20/class, **70.2%** in mAP), CSC without replay buffer (0, **72.8%** in mAP) also achieves superior results in the more challenging scenario {B0-C10}.

The results on the two more challenging scenarios {B0-C5 and B20-C4} of MS-COCO are shown in Fig. 4. These mAP curves demonstrate the significant superiority of our method during the CIL process. Moreover, our method exhibits a closer proximity to the upper bound. This demonstrates that our method performs effectively even in long-term class-incremental processes.

PASCAL VOC. The results in {B0-C4, B10-C2, B5-C3 and B4-C2} of PAS-CAL VOC are shown in Tab. 2. Just as on MS-COCO, our method significantly outperforms the state-of-the-art in all scenarios. Specifically, the final mAP improved by **4.5%** (83.4%→87.9%) and **7.3%** (80.5%→87.8%) in {B0-C4} and {B10-C2}, respectively. Furthermore, our method exhibits better robustness across different scenarios. For instance, from {B0-C4} to {B4-C2}, the final mAP

Table 4: The ablation study for effectiveness of each component.

Method	Max-Entropy	CI-GCN	MS-COCO B0-C10				MS-COCO B0-C5			
			mAP	CF1	OF1	Avg. mAP	mAP	CF1	OF1	Avg. mAP
Baseline			42.4	45.3	43.7	61.2	35.9	30.0	28.0	54.9
(a) CSC	✓		47.6	50.3	49.5	63.1	39.2	35.4	33.1	57.3
(b) CSC		✓	69.3	59.0	59.5	75.6	64.8	42.6	39.5	73.4
(c) CSC	✓	✓	**72.8**	**64.9**	**66.8**	**78.0**	**67.3**	**50.5**	**47.2**	**75.0**

Table 5: Correlation matrix study. "Z": fixed statistical CM, "G": general CM and "S": specific CM.

Structure	Last Acc			Avg.Acc
	mAP	CF1	OF1	mAP
Z → Z (sigmoid)	64.1	57.1	59.1	73.7
Z → S (sigmoid)	69.8	62.6	65.3	75.6
G → S (sigmoid)	71.2	64.0	65.5	76.5
G → S (softmax)	**72.8**	**64.9**	**66.8**	**78.0**

Table 6: Sensitive study of max-entropy weight β and buffer size from 5/class to 50/class in {B0-C10} of MS-COCO.

β	mAP	CF1	OF1	Buffer Size	mAP	CF1	OF1
1e-3	72.97	62.54	62.92	5/class	73.68	67.33	68.09
2e-3	**73.02**	63.84	64.24	10/class	73.93	67.54	68.53
4e-3	72.84	**64.89**	**66.83**	20/class	74.78	67.82	68.56
6e-3	71.71	63.57	66.66	30/class	75.01	67.89	68.90
8e-3	71.13	60.56	65.53	50/class	**75.26**	**68.08**	**68.97**

of CSC-R changes from 87.9% to 86.6% (↓ **1.3%**). However, the final mAP of AGCN-R changes from 76.0% to 59.3% (↓ **16.7%**). This demonstrates that, owing to CI-GCN and max-entropy regularization, our method possesses better environmental adaptability and robustness. The every task result in {B5-C3 and B4-C2} scenarios of PASCAL VOC is shown in Fig. 4. Consistent with the conclusions drawn from the MS-COCO, our method outperforms other comparative methods across all class-incremental tasks.

Comparison of different backbones. Tab. 3 illustrates the comparison between our method and the prompt-based method L2P [47]. The backbone of L2P is the pre-trained ViT-B/16. Our method and KRT [10] are based on the pre-trained TResNetM. It can be observed that L2P performs better overall compared to KRT in the more challenging {B0-C10} scenario. However, our method achieves better results with fewer parameters than L2P in both scenarios.

4.4   Ablation Study

Effectiveness of Each Component. Tab. 4 shows the detailed analysis of our contributions in two challenging scenarios. Under the task-level partial label setting, the knowledge distillation provides supervised information for missing labels, thus we naturally utilize it as a baseline, similar to [2,10,15]. Firstly, when we calibrate the baseline model confidence with max-entropy (row a), the mAP, CF1 and OF1 of the model are significantly improved. Then, when the disjoint label spaces are connected by constructing cross-task relationships (row b), the improvement is remarkable. For example, the mAP, CF1 and OF1 of the model in {B0-C10} are improved by **26.9%** (42.4%→69.3%), **13.7%** (45.3%→59.0%) and **15.8%** (43.7%→59.5%). Finally, we add the max-entropy regularization to the model (row c), this component can further improve the model's performance
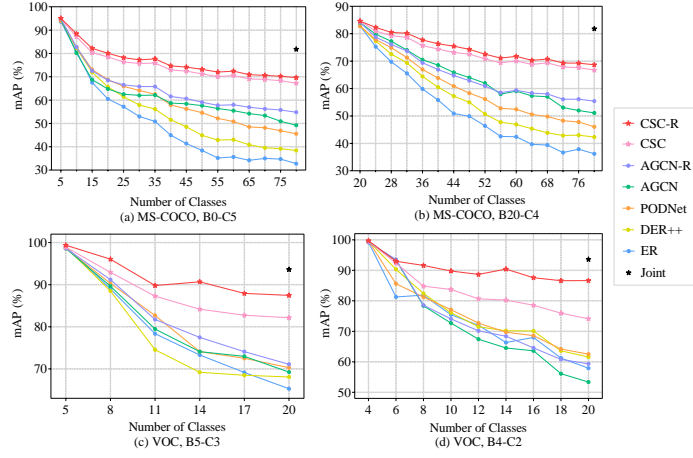
Fig. 4: Comparison results on MS-COCO and VOC datasets in challenging scenarios. There are more incremental tasks in these scenarios.
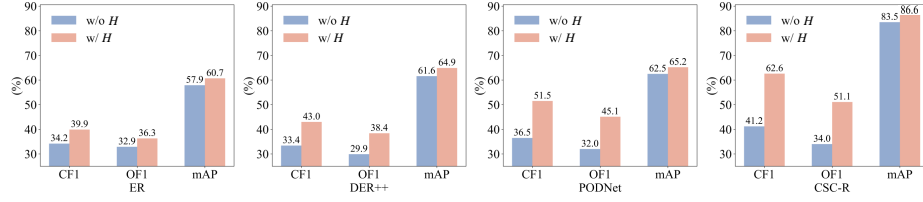

Fig. 5: Max-entropy analysis in {B4-C2} of VOC.

by calibrating the multi-label confidence: **64.8%** vs **67.3%** in mAP, **42.6%** vs **50.5%** in CF1 and **39.5%** vs **47.2%** in OF1 of {B0-C5}. Furthermore, in horizontal comparisons, we found that even in more challenging scenarios, our approach still demonstrates stable superiority **72.8%** vs **67.3%** in mAP. These results demonstrate that CSC framework can significantly suppress the negative impact of partial label and make multi-label confidence well-calibrated.

Correlation Matrix Study. To verify the effectiveness of the general and specific CM, we show the performances of different combinations of CM in Tab. 5. "Z" denotes the CM constructed using a statistical way like [7,15], which is fixed for each sample. "G" means our general CM and "S" means specific CM. Tab. 5 also shows ablation studies on regularization $Sigmoid(\cdot)$ and $Softmax(\cdot)$ functions, which are used for regularization of class activation maps $\mathbf{M}$ in Eq. (3). When the combination way is a general CM followed by a specific CM (G → S) with $Softmax(\cdot)$, the results are superior. The results show that constructing class-incremental label relationships in CI-GCN is more effective than fixed statistical label relationships.

Sensitive Study. As shown in Tab. 6, to further demonstrate the robustness of our approach, we conduct the sensitive study. We first set α in Eq. (7) to 0.15.
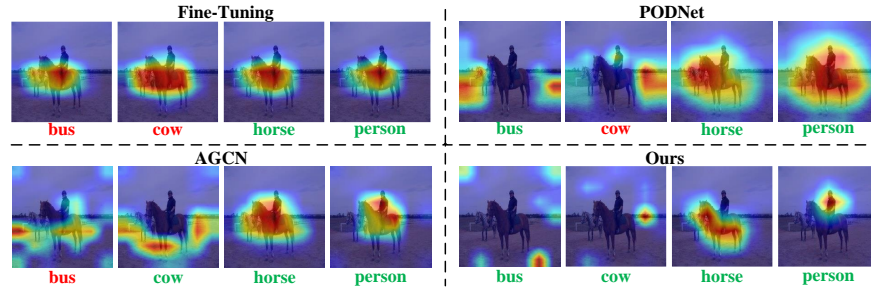
Fig. 6: Visualization of CIL methods. Labels in red: F-P error, green: correctness.

Then, we analyze the influence of β in Eq. (8), when the β is set to 4e-3, the overall performance of the model is superior. We increase the buffer size from 5/class to 50/class, the final mAP increases from 73.68% to 75.26%. Furthermore, when the buffer size is set to 0, the final mAP is 72.84%. These results indicate the robustness of our method, demonstrating its capability to effectively preserve knowledge and mitigate forgetting even without a buffer.

Max-Entropy Analysis. Fig. 5 demonstrates the orthogonality of max-entropy calibration across various class-incremental methods under different evaluation metrics. These results provide evidence that max-entropy regularization can significantly enhance the performance of different methods.

4.5   Visualization

Fig. 6 shows several attention maps for four methods. The labels "bus", "cow", "horse" and "person" are sequentially trained in different tasks. After training on "person", given a test image with "horse" and "person", the model outputs attention maps for the four classes. It can be seen that our calibrated model can output appropriate confidence to achieve correct predictions and accurately capture the features of each class.

5   Conclusion

In this paper, we studied the less-explored task of multi-label class-incremental learning (MLCIL). In MLCIL, only the new classes are labeled at training time, the absence of past and future labels results in a *task-level partial label* issue for the current task. In this paper, we refined the multi-label confidence calibration in MLCIL and proposed a Confidence Self-Calibration (CSC) framework for addressing this issue. Firstly, we introduced a class-incremental graph convolutional network (CI-GCN) to calibrate the label relationships by bridging the isolated label spaces. Then, we presented a max-entropy regularization for over-confident output distribution to facilitate confidence self-calibration. Our approach achieved new SOTA results in MLCIL tasks on both MS-COCO and PASCAL VOC datasets.

Acknowledgments

References

1. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 15920–15930 (2020)
2. Cermelli, F., Mancini, M., Bulo, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9233–9242 (2020)
3. Cha, S., Cho, S., Hwang, D., Hong, S., Lee, M., Moon, T.: Rebalancing batch normalization for exemplar-based class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20127–20136 (2023)
4. Cha, S., Yoo, Y., Moon, T., et al.: Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 10919–10930 (2021)
5. Chen, T., Lin, L., Hui, X., Chen, R., Wu, H.: Knowledge-guided multi-label few-shot learning for general image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3), 1371–1384 (2020)
6. Chen, T., Xu, M., Hui, X., Wu, H., Lin, L.: Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 522–531 (2019)
7. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019)
8. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Learning graph convolutional networks for multi-label recognition and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 6969–6983 (2023)
9. De Lange, M., Tuytelaars, T.: Continual prototype evolution: Learning online from non-stationary data streams. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8250–8259 (2021)
10. Dong, S., Luo, H., He, Y., Wei, X., Cheng, J., Gong, Y.: Knowledge restore and transfer for multi-label class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18711–18720 (2023)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (2021)
12. Douillard, A., Cord, M., Ollion, C., Robert, T., Valle, E.: Podnet: Pooled outputs distillation for small-tasks incremental learning. In: Proceedings of the European Conference on Computer Vision. pp. 86–102 (2020)

13. Douillard, A., Ramé, A., Couairon, G., Cord, M.: Dytox: Transformers for continual learning with dynamic token expansion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9285–9295 (2022)
14. Du, K., Lyu, F., Hu, F., Li, L., Feng, W., Xu, F., Fu, Q.: Agcn: Augmented graph convolutional network for lifelong multi-label image recognition. In: IEEE International Conference on Multimedia and Expo. pp. 01–06 (2022)
15. Du, K., Lyu, F., Li, L., Hu, F., Feng, W., Xu, F., Xi, X., Cheng, H.: Multi-label continual learning using augmented graph convolutional network. IEEE Transactions on Multimedia **26**, 2978–2992 (2024)
16. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision (2010)
17. Kim, C.D., Jeong, J., Kim, G.: Imbalanced continual learning with partitioning reservoir sampling. In: Proceedings of the European Conference on Computer Vision. pp. 411–428 (2020)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2015)
19. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. National Academy of Sciences **114**(13), 3521–3526 (2017)
20. Li, W., Lyu, F., Shang, F., Wan, L., Feng, W.: Long-tailed learning as multi-objective optimization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
21. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017)
22. Liang, Y.S., Li, W.J.: Optimizing class distribution in memory for multi-label online continual learning. arXiv preprint arXiv:2209.11469 (2022)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
24. Liu, D., Lyu, F., Li, L., Xia, Z., Hu, F.: Centroid distance distillation for effective rehearsal in continual learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2023)
25. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
26. Luo, Z., Liu, Y., Schiele, B., Sun, Q.: Class-incremental exemplar compression for class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11371–11380 (2023)
27. Lyu, F., Du, K., Li, Y., Zhao, H., Zhang, Z., Liu, G., Wang, L.: Variational continual test-time adaptation. arXiv preprint arXiv:2402.08182 (2024)
28. Lyu, F., Feng, W., Li, Y., Sun, Q., Shang, F., Wan, L., Wang, L.: Elastic multi-gradient descent for parallel continual learning. arXiv preprint arXiv:2401.01054 (2024)
29. Lyu, F., Sun, Q., Shang, F., Wan, L., Feng, W.: Measuring asymmetric gradient discrepancy in parallel continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11411–11420 (2023)
30. Lyu, F., Wang, S., Feng, W., Ye, Z., Hu, F., Wang, S.: Multi-domain multi-task rehearsal for lifelong learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8819–8827 (2021)

31. Mai, Z., Li, R., Kim, H., Sanner, S.: Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3589–3599 (2021)
32. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of Learning and Motivation, vol. 24, pp. 109–165 (1989)
33. Mohamed, A., Grandhe, R., Joseph, K., Khan, S., Khan, F.: D3former: Debiased dual distilled transformer for incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2420–2429 (2023)
34. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 33, pp. 15288–15299 (2020)
35. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: Proceedings of the International Conference on Learning Representations (2017)
36. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
37. Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., Friedman, I.: Tresnet: High performance gpu-dedicated architecture. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1400–1409 (2021)
38. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: Proceedings of the International Conference on Learning Representations (2021)
39. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 350–360 (2019)
40. Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y.W., Pascanu, R., Hadsell, R.: Progress & compress: A scalable framework for continual learning. In: Proceedings of the International Conference on Machine Learning. pp. 4528–4537 (2018)
41. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Proceedings of the Advances in Neural Information Processing Systems (2017)
42. Song, L., Wu, J., Yang, M., Zhang, Q., Li, Y., Yuan, J.: Handling difficult labels for multi-label image classification via uncertainty distillation. In: Proceedings of the International Conference on Multimedia. pp. 2410–2419 (2021)
43. Sun, Q., Lyu, F., Shang, F., Feng, W., Wan, L.: Exploring example influence in continual learning. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 35, pp. 27075–27086 (2022)
44. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2285–2294 (2016)
45. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–20 (2024)
46. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 464–472 (2017)

47. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)
48. Wu, Z., Baek, C., You, C., Ma, Y.: Incremental learning via rate reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1125–1133 (2021)
49. Ye, F., Bors, A.G.: Lifelong teacher-student network learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6280–6296 (2022)
50. Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: Proceedings of the European Conference on Computer Vision. pp. 649–665. Springer (2020)
51. Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., Li, J.: Transformer-based dual relation graph for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 163–172 (2021)
52. Zhao, L., Lu, J., Xu, Y., Cheng, Z., Guo, D., Niu, Y., Fang, X.: Few-shot class-incremental learning via class-aware bilateral distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11838–11847 (2023)
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)
54. Zhou, D.W., Ye, H.J., Zhan, D.C.: Few-shot class-incremental learning by sampling multi-phase tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)