This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DELTA: Decoupling Long-Tailed Online Continual Learning

Siddeshwar Raghavan
raghav12@purdue.edu

Jiangpeng He
he416@purdue.edu

Fengqing Zhu
zhu0@purdue.edu

School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana USA

Abstract

A significant challenge in achieving ubiquitous Artificial Intelligence is the limited ability of models to rapidly learn new information in real-world scenarios where data follows long-tailed distributions, all while avoiding forgetting previously acquired knowledge. In this work, we study the under-explored problem of Long-Tailed Online Continual Learning (LTOCL), which aims to learn new tasks from sequentially arriving class-imbalanced data streams. Each data is observed only once for training without knowing the task data distribution. We present DELTA, a decoupled learning approach designed to enhance learning representations and address the substantial imbalance in LTOCL. We enhance the learning process by adapting supervised contrastive learning to attract similar samples and repel dissimilar (out-of-class) samples. Further, by balancing gradients during training using an equalization loss, DELTA significantly enhances learning outcomes and successfully mitigates catastrophic forgetting. Through extensive evaluation, we demonstrate that DELTA improves the capacity for incremental learning, surpassing existing OCL methods. Our results suggest considerable promise for applying OCL in real-world applications. Code is available online [1]

1. Introduction

The process through which humans and Artificial Intelligence (AI) systems acquire knowledge and experiences differs significantly. Over their lives, humans learn and accumulate knowledge from encountering sequential streams of temporally correlated information, mostly made up of unlabeled observations, and rarely experiencing the same scenario multiple times [10, 48]. Furthermore, humans are adept at learning, remembering knowledge, and solving multiple tasks concurrently by applying the learned knowledge. In contrast, AI systems have a smaller task focus [37] and a multi-stage learning approach, focusing on learning from static (non-changing) datasets through batches. Online continual learning [3, 10, 16, 34, 40, 43] strives to push the boundaries of AI by empowering agents to acquire knowledge continuously from a never-ending stream of data. However, a significant challenge continual learning systems face is how to mitigate the effects of catastrophic forgetting [1, 35, 39, 46] of previously learned information. Furthermore, recent work focuses on continual learning in the online scenario. In the online configuration [24], each data sample is used only once to train the model. Despite being more challenging, the online setting aligns with real-world limitations concerning data accessibility and compu-

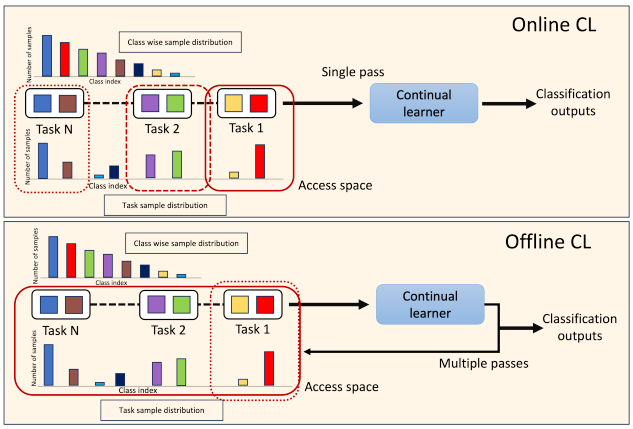

Figure 1. Illustration depicts online and offline setups for continual learning with a long-tailed distribution. In the continual learning process, tasks appear sequentially, one at a time. In the "online" scenario, the model only accesses the current task and its distribution, while the "offline" scenario grants access to the complete task set and their distributions. Additionally, the "online" approach involves training task data with a single pass, while the "offline" approach involves multiple passes across the entire dataset.

[1] Link to code - https://gitlab.com/viper-purdue/delta

tational capabilities, rendering it more appropriate for real-world applications [19, 56].

While existing online continual learning techniques have achieved impressive advancements in the context of image classification tasks, they operate under the assumption that the distribution of samples for each class is uniform [2, 3, 16, 40, 43, 53]. This assumption limits their applicability in real-world scenarios where data distribution tends to be long-tailed with extreme imbalances. A concrete example of this can be observed in real-world applications such as animal species recognition [47, 68], medical image diagnosis [31, 67], and food image classification [25, 45, 51]. In such cases, a minority of these images are encountered more frequently than others, leading to pronounced class imbalances. We focus on LTOCL as shown in figure 1, wherein data sampled from such distributions emerges sequentially in a stream over time.

Though previous studies have highlighted imbalanced data distributions within the context of continual learning [10, 17, 34], the majority of these investigations have primarily dealt with milder imbalance ratios or investigated the offline scenario [41]. Class imbalance signifies a scenario where the number of instances across different classes significantly varies, often leading to certain classes having a notably higher number of samples than others. However, a long-tailed distribution poses a more intricate challenge, as it involves a setup where a small subset of classes contains a substantial number of samples. In contrast, numerous classes are represented by only a limited number of samples. The distinction between online and offline CL settings is characterized by training the model once on each data sample in the online scenario, as opposed to the potential for multiple passes through the data in the offline scenario. In offline setups, the model can access the complete dataset, thereby obtaining the data distribution of classes across all tasks as seen in Figure 1. In sharp contrast, in online scenarios, data arrives in batches, restricting the model's access solely to specific batches or subsets of data, rather than the entire dataset or complete class distributions. As a result, the data distribution for each subsequent task becomes uncertain in the context of online continual learning, and online exemplar selection becomes notably more complex within a long-tailed distribution. The presence of class-imbalanced exemplars can worsen the overfitting problem and lower the model performance during online continual learning.

Our work fills the gap between online continual learning with severe long-tailed data distribution in the image classification task. We formulate scenarios with tasks containing significant data imbalances and do not provide task identifiers during training or testing. To address these challenging scenarios, we introduce the DELTA framework. The proposed method utilizes decoupled representations in a dual-stage learning strategy, integrating contrastive learning, and Equalization Loss to re-calibrate weights in the feature space, promoting an efficient learning process.

We assess the effectiveness of our approach by comparing it against current methods in the field of OCL under long-tailed distribution conditions. Our technique consistently outperforms various experimental configurations, maintaining robustness with changes in exemplar and incremental step sizes. The main contributions of our research are summarized as follows.

- We present DELTA, a dual-stage training approach that combines contrastive learning with equalization loss, tailored for Long-Tail Online Continual Learning (LTOCL) situations.
- We propose a multi-exemplar pairing strategy to demonstrate the potential for performance enhancement in LTOCL scenarios.
- We evaluate well-established OCL methods, compare them to DELTA in the proposed long-tailed setting under various experimental setups and report the findings.

2. Related Work

In this section, we review and summarize the existing methods that are most relevant to our work, including (1) online continual learning, (2) long-tailed classification and (3) Contrastive Learning, which are illustrated in Section 2.1, Section 2.2 and Section 2.3 respectively.

2.1. Online Continual Learning

Continual Learning (CL) is a machine learning strategy where a model is trained to progressively incorporate new classes/ categories over time without forgetting the learned knowledge. CL has been studied under different scenarios, including (i) *Online (OCL)* and *Offline* settings. The former involves the model accessing solely the specific batch of data within a task [2], enabling only a single pass over the data. In contrast, the latter scenario allows the model to access the complete dataset (task-aware), permitting multiple passes over its contents [41]. (ii) *task-incremental* and *class-incremental* approaches characterize another distinction, where the former necessitates a task index during both training and inference, while the latter does not. In this work, we focus on class-incremental learning in the online scenario. The objective is to learn new classes from the sequentially available data streams by using each data only once to update the model and classify all classes seen so far during the inference phase.

Online class-incremental learning methods can be grouped into two main categories including *Regularization* and *Memory* based approaches. The **Regularization** methods aim to limit parameter changes to preserve learned

[2]Sequential stream of data is encountered as a set of tasks

knowledge [6, 21, 24, 35, 42]. Conversely, **Memory** methods [3, 10, 16, 20, 34, 40, 43] combat catastrophic forgetting by storing task data as exemplars and replaying knowledge during learning. In the online setting, each new batch combines current task data with exemplars from memory for model updates. Memory-based methods are generally more effective than regularization-based ones [44]. Furthermore, the procedures for buffer retrieval and storage differ notably between the "online" and "offline" settings, owing to the distinct data access constraints inherent in each setup.

Recent studies have emphasized Class Incremental Learning (CIL) setups that simulate realistic scenarios characterized by ambiguous task boundaries and imbalanced datasets [17, 18, 23, 41]. Similarly, research has been into realistic Class Incremental Learning involving data imbalances [10, 34]. However, these investigations have predominantly centered on the offline setting [23, 41], or often featuring moderate imbalances in the online context. Recent works have begun to explore logit adjustment within the online setting [29, 60]. However, these investigations primarily concentrate on the conventional setting(equi-sample distribution across classes), addressing distribution shifts present in the online setting rather than tackling external, realistic data imbalances. Notably, exploring severe data imbalances within the online setting remains relatively underexplored.

2.2. Long-Tailed Classification

The long-tailed classification addresses the extreme class imbalance issue where many training classes contain a few training samples while the testing samples are class-balanced. The major challenge is the classification bias towards instance-rich (head) classes and poor generalization ability in classifying instance-rare (tail) classes. As the long-tailed classification has been widely studied over decades [69], we review the existing (*i.e.* end-to-end) approaches that are most related to our work. **Re-sampling** based methods aim to address class-imbalance issues by generating balanced training distribution. The typical work includes over-sampling [59] the instance-rare classes and under-sampling [4, 22] the instance-rich classes. The most recent work [25, 26, 49] further applies CutMix [65] as data augmentation to mitigate the over-fitting and under-fitting issues caused by naive sampling strategies. **Re-weighting** based method balances the loss gradients by assigning higher weights on instance-rare and lower weights on instance-rich classes or data samples. Specifically, the class level weights in [28, 61] are generated based on the inverse of class frequency. Furthermore, there are Class-balanced loss [12], label-distribution-aware-margin loss [5], balanced softmax [54] and LADE loss [27], which aim to balance the loss gradients. **Two-stage methods** [13, 32, 64, 70] focus on decoupling imbalanced feature learning from balanced classifier learning [58]. However, all the existing methods require knowing the data distribution in the entire training set, which is not feasible in the online continual learning scenario where the new data comes sequentially in a stream over time. In this study, we draw inspiration from re-sampling and re-weighting approaches used in Long-Tail image classification to design an Equalization loss to mitigate the severe imbalances present without knowing the entire distribution of the dataset.

2.3. Contrastive Learning

Contrastive learning effectively learns meaningful data representations by pulling similar samples closer and pushing dissimilar ones apart. This approach enables the model to uncover and leverage the underlying structures and semantics of the data, resulting in representations beneficial for various downstream tasks [7, 30, 33, 52, 63]. Researchers have demonstrated that contrastive learning is applicable in both supervised and unsupervised settings. In unsupervised contrastive learning, data augmentations act as similar samples, and randomly selected samples from the target batch act as dissimilar samples [8, 38]. Conversely, in supervised contrastive learning, samples from the same class are treated as similar, and those from different classes are considered dissimilar [14, 33].

3. Preliminaries for Long-Tailed Online Continual Learning

We examine a data distribution characterized by a significant long-tailed nature in the context of online continual learning for supervised image classification tasks. The long-tailed distribution follows an exponential decay in sample sizes across classes [5, 54]. This decay is parameterized by ρ, the ratio between the most and least appearing classes. In the OCL setting, the model encounters a continuous stream of data at each task t. In this setup, at any task t, k^t represents the number of classes, and n_j^t represents the number of samples in class j. Thus, the number of classes learned up to task t can be represented as $k^{1:t}$. The training samples at task t, denoted as $\mathcal{X}_t = \{x_i^t, y_i^t\}; i \in \{1, 2, .., k^t\}$, are non-independent and identically distributed (non i.i.d), drawn from the current distribution D_t (x_i^t represents an image and y_i^t represents its corresponding label). Within the Online Continual Learning (OCL) setting, the model's access is confined to the data of a specific batch within task t, given the sequential flow of data. Consequently, it remains aware solely of the data distribution up to that point. This distribution, referred to as D_t, is subject to change between tasks, transitioning from D_t to D_{t+1}.

ER methods involve the utilization of a fixed-size memory buffer denoted as B. This buffer stores a limited set of samples from the learned tasks. This stored subset is drawn

upon for knowledge rehearsal during the learning of subsequent tasks. After training on each task t in the continual learning process, the model is tested on a balanced held-out test set denoted as $\overline{\mathcal{X}_t} = \{\overline{x}_i^t, \overline{y}_i^t\}$; $i \in \{1, 2, .., k^t\}$. This test set includes only the classes encountered so far. The total number of training samples in task t is denoted as n^t, and it satisfies the condition $\sum_{j=1}^{k^t} n_j^t = n^t$.

4. Method

Decoupling representation learning from the classification task has been shown effective in two-stage network architectures tackling the problem of long-tailed recognition [11, 57, 71]. This concept is adapted for the OCL setting with long-tailed data by developing a two-stage approach that incorporates contrastive learning in the first stage and implements an Equalization Loss in the second stage to address the pronounced imbalances in the data.

The structure of DELTA is depicted in figure 2. We have included the pseudo code in the supplementary material.

- **Stage 1** comprises contrastive learning in the supervised setting due to the availability of labels in OCL. The motivation to utilize contrastive learning in the long-tailed setting is to cluster similar samples and push apart dissimilar samples to aid in effective feature learning.
- **Stage 2** consists of training just the classification layer of the network using an Equalization Loss to re-weight the samples, to fine-tune and obtain a better classifier.

4.1. Stage 1 - Representation Learning

In Online Continual Learning (OCL), where data undergoes single-pass processing and exhibits a pronounced imbalance, particularly in long-tail distributions, effectively learning the underlying feature representations is essential.

Various methodologies within contrastive learning, such as Barlow Twins [66], SimSiam [9] and BYOL [15], primarily focus on leveraging positive samples for learning. However, our methodology draws inspiration from SimCLR [7], which employs both positive and negative samples advantageously. In long-tailed distributions, where many classes have few samples, the diversity of negative samples from the more populated classes can help the model learn more discriminating features for the underrepresented classes. Additionally, SimCLR heavily relies on aggressive data augmentation strategies to generate positive pairs. This approach can help mitigate the effects of class imbalance by ensuring that the model learns robust features for each class, regardless of its frequency in the dataset. There are three main components in the contrastive learning pipeline [7, 43]:

- Data augmentation on the input sample is performed, denoted as $\widehat{x} = Aug(x)$.
- The encoder network, represented as $Encoder(.)$, transforms an image sample into a vector embedding, $e = Encoder(x) \in \mathbb{R}^{D_N}$, with normalization to the unit space in $\mathbb{R}^{D_N}$.
- The projection network, indicated by $Projection(.)$, takes the embedding and maps it to a projected vector, $v = Projection(e) \in \mathbb{R}^{D_P}$, which is then normalized using the L2 norm.

The contrastive loss is defined as,

$$L_{contrastive}(Z_T) = \sum_{j \in T} \frac{-1}{|P(j)|} \sum_{p \in P(j)} \frac{exp(v_j \cdot v_p/\tau)}{\sum_{k \in A(j)} exp(v_j \cdot v_p/\tau)} \tag{1}$$

$B_T = B_t \cup Aug(B_t)$ represents the samples obtained from the buffer for task t, where B_t is the set of original buffer samples and $Aug(B_t)$ is their augmented versions. The index set T identifies elements of B_T excluding the ith sample. $\mathcal{X}_t(i)$ refers to all elements in B_T excluding sample i. The function $P(j)$ identifies the set of positive samples in B_T that share the same label as sample i, but does not include sample i itself. The parameter $\tau \in \mathbb{R}$ serves as a temperature factor to modulate class separation, and the symbol $\cdot$ denotes the dot product operation.

4.2. Stage 2 - Balanced Classifier Learning

In the DELTA framework's second stage, the goal is to establish a balanced training environment for the classifier by decoupling the learning of feature representations from the classification task. In scenarios with long-tailed distributions, where class distribution is highly imbalanced, a model trained to minimize empirical risk often underperforms on a balanced test dataset due to distribution mismatch. Cross-entropy loss in such scenarios can result in learning-biased batches. While many Online Class Incremental Learning (OCL) methods [44], employ Cross-Entropy (CE) loss, this loss function inherently favors the classification of majority classes, potentially at the expense of minority class accuracy. To counteract the issue of biased batch learning, we propose the Equalization Loss (L_{EQ}), an innovative technique designed for online continual learning environments. This method is inspired by the balanced softmax loss [54], which requires the entire data distribution of the dataset. In Long-Tailed Online Continual Learning (LTOCL) settings, we cannot access the entire dataset but obtain the data via continuous streams. Thus, we incorporate a task-specific distribution vector, $P(k^t)$, which is updated after encountering a data stream within a task, as shown in Equation 2. In this context, D_t represents the distribution of samples for task t, encompassing both the training inputs and the exemplars retrieved from the buffer.

$$D_t = [n_1, n_2, ...n_h] \tag{2}$$

$$P(k^t) = [\frac{n_1}{\sum_{i=1}^{h} n_i}, \frac{n_2}{\sum_{i=1}^{h} n_i}, ..., \frac{n_h}{\sum_{i=1}^{h} n_i}] \tag{3}$$

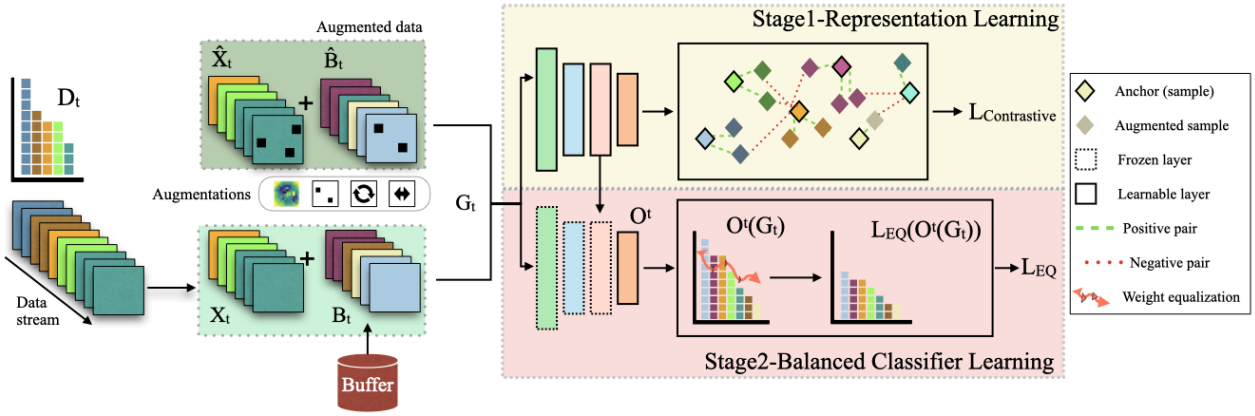

Figure 2. An overview of the DELTA framework: At task t, the current batch of samples(X_t) and samples retrieved from the memory buffer (B_t) undergo augmentation ($\hat{X}_t$, $\hat{B}_t$) and are then combined (G_t). This combined data is directed sequentially through a dual-stage training pipeline. In the first stage, the framework utilizes contrastive learning to generate effective data representations involving a contrastive loss ($L_{contrastive}$). During the second stage, the learning approach is decoupled by keeping all layers frozen except for the classification layer (O^t). This targeted training employs the weight equalization loss (L_{EQ}) to train a balanced classifier and reduce the shift in future data representations.

where where $\{n_h; h = 1, 2, .., k^{1:t}\}$ denotes the number of samples in class h. This vector dynamically characterizes the sample distribution within each incoming training batch for a given task t.

The logits at the output stage of the classifier is expressed as $O^t(I_x) = [O^1(I_x), O^2(I_x), ..., O^{1:t-1}(I_x), ..., O^t(I_x)]$, respectively, where I_x represents the input image sample to the classifier. For each incoming training batch, we calculate the temporary probability distribution. We then adjust the gradients to reflect this distribution by incorporating $P(k^t)$ as a prior vector in the output, a process detailed in Equation 4.

$$\mathcal{L}_{EQ}(O^t(I_x)) = \sum_{i=1}^{k^{1:t}} -I_{y_{(i)}} \sigma([\log[(P(k^t)] + O^t(I_x))]) \quad (4)$$

where I_y is the corresponding label for input Ix and σ represents the softmax function. Therefore, the more frequent class of the current training batch with a larger value in the prior vector $P(k^t)$ achieves smaller gradients when we compute the cross-entropy using the adjusted logits and vice versa. Therefore, effectively address the biased loss gradients issue in the long-tailed online continual learning without requiring knowing the data distribution beforehand.

The cornerstone of our Equalization Loss (L_{EQ}) is to correct gradient disparities by utilizing the temporary distribution vector, thereby bridging the gap between the training data's long-tailed nature and the testing data's balanced nature. In turn, it improves the model's capacity to effectively retain and generalize knowledge across different classes.

The overall loss in stage1 is represented in Equation 5.

$$\mathcal{L}_{stage1}(O^t(G_t)) = \mathcal{L}_{contrastive}(O^t(G_t)) \quad (5)$$

G represents the input samples, the buffer retrieved samples and their augmented versions. $G_t = B_t \bigcup \widehat{B_t} \bigcup \mathcal{X}_t \bigcup \widehat{\mathcal{X}_t}$

In **stage2**, as shown in figure 2, we freeze all the layers except the classification layer to train the network using the Equalization Loss to obtain the best classification accuracy, represented as,

$$\mathcal{L}_{stage2}(O^t(G_t)) = \mathcal{L}_{EQ}(O^t(G_t)) \quad (6)$$

4.3. Multi-Exemplar Learning

The discrepancies between long-tailed training distributions and balanced test distributions pose a significant challenge, leading to a bias in the learning algorithm towards the training data due to uneven sample sizes exacerbated in the online setting. To address this, we propose an exemplar selection strategy pairing more than one exemplar with each training sample to balance the batch composition and mitigate bias. This approach preserves the data's inherent randomness and ensures that exemplar representation aligns with the distribution vector. As shown in figure 4, the x-axis denotes the number of exemplars paired per input data sample.

Our work is among the earliest in exploring multi-exemplar pairing within the context of OCL. Traditional approaches [2, 3, 10, 34, 41, 55, 62] typically limit themselves to matching a single exemplar from the memory buffer with each sample from the current batch. However, this may not be advantageous when the number of tasks increases or data exhibits a high variability. This practice often results in suboptimal performance due to the one-off nature of the process. Moving beyond the one-exemplar-per-sample restriction for each new batch opens a new direction in OCL. Our DELTA method incorporates data augmentations before the training phase, which helps mitigate overfitting despite multiple encounters with previously learned samples from the

buffer. Moreover, this repeated exposure to past samples plays a crucial role in combating catastrophic forgetting and enhances the accuracy of learned representations.

The empirical risk for any given task t can be formulated as

$$R_{emp}f(\theta) = \mathbb{E}_{(G_{t\_x}, G_{t\_y}) \sim D_{G_t}} \mathcal{L}(f(G_{t\_x}; \theta), G_{t\_y}) \tag{7}$$

Where $R_{emp}f(\theta)$ signifies the empirical risk of the model f for a given task t, $\mathbb{E}$ represents the expected value, $\mathcal{L}$ denotes the loss function, and G_{t_x}, G_{t_y} correspond to the images and their respective labels, consisting of input, buffer data and their augmented pairs. However, the empirical risk evaluated on the training dataset does not equate to the true risk on the test dataset, given that $D_{\mathcal{X}} \neq D_{\overline{\mathcal{X}}}$, where $D_{\mathcal{X}}$ and $D_{\overline{\mathcal{X}}}$ denote the distributions of the training and test sets, respectively.

$$R_{true}f(\theta) = \mathbb{E}_{(\overline{x}^t, \overline{y}^t) \sim D_{\overline{\mathcal{X}}}} \mathcal{L}(f(\overline{x}^t; \theta), \overline{y}^t) \tag{8}$$

Increasing the number of paired exemplars and incorporating augmentations while balancing each batch are aimed at enhancing the model's robustness to input variations and improving its generalization to a held-out test set. By training with a dataset enriched to reflect the data's inherent variability better, the continual learner is less biased toward the long-tailed nature of the training data. This approach leads to more accurate and unbiased estimates of the gradient and reduces the variance in model updates, facilitating smoother convergence towards the empirical loss minimum. We aim to balance the data distribution within each batch more accurately; the likelihood of the model overfitting to specific training data features is decreased, resulting in improved generalization performance on unseen data.

5. Experiments

In this section, we first introduce the datasets and our experimental setup. Then, we comprehensively analyze the performance of the current OCL methods in the conventional and the LT setup. Finally, we conduct ablation studies to show the effectiveness of each component in our proposed framework.

5.1. Datasets

We use two publicly available datasets, CIFAR-100 [36] (100 classes), and the VFN-LT [25] (74 classes). We create the long-tailed version of Split CIFAR-100 with the imbalance factor $\rho = 0.01$, where ρ represents the ratio between most frequent and least frequent classes [41]. The smaller the value of ρ, the more pronounced the imbalance. Overall, the Split CIFAR-100 has over 10K training images with a maximum of 500 images and a minimum of 5 images per class. The VFN-LT dataset reflects the real-world food distribution compared to other datasets. It is long-tailed, containing over 15,000 training images across 74 classes, representing commonly consumed food categories in the United States based on the WWEIA database [3].

5.2. Implementation Detail

Our implementation is PyTorch [50] based. We use ResNet-32 for Split CIFAR-100 and ResNet-18 for VFN-LT dataset, which acts as the Encoder network and for the projection network we use a fully connected layer to map the representations from the encoder to 128-dimensional latent space [7]. We train the networks from scratch and split the datasets using fixed seed 1993. The input image size set for Split CIFAR-100 is 32×32, and 224×224 for VFN-LT following the settings suggested in [44]. For CIFAR-100-LT, we evaluate two configurations: one with 20 tasks, each comprising five unique classes, and another with 10 tasks, each containing ten unique classes. For VFN-LT, the division is into 15 tasks, where the initial task contains four classes, and each subsequent task consists of five classes, and another configuration with seven tasks, where the first task involves 14 classes, and each of the following tasks includes ten classes. We use a stochastic gradient descent optimizer with a fixed learning rate of 0.1 and a weight decay of 10^{-4}. The training batch size is 16, and the testing batch size is 128. The data (except exemplars) is seen only once by the model to train for all the experiments. For our experiments, we implement three memory buffer sizes (0.5K, 1K, and 2K) for various experience replay methods and configure single exemplar pairing. For the buffer, we use random sampling and reservoir update mechanism extended from [44]. We set the temperature parameter τ as 0.09, obtained via grid search. We run each experiment 5 times and report the average accuracy in Table 1. We run all our experiments on a single NVIDIA A40 GPU.

In our work, we employ the publicly accessible implementations of all existing OCL methods, as referenced in [10, 34, 41, 44, 62], for our comparative analysis. We have considered an offline OCL method [41] for comparison and we implement it in the online setting by running the method for one epoch and consider only the task-agnostic setting for fair evaluation. Additionally, we have integrated a long-tailed data loader into this framework and have implemented our proposed method, DELTA, for a comprehensive comparison.

5.3. Evaluation Metrics

In this work, we employ Average Accuracy to evaluate performance. Average Accuracy assesses the overall performance across the testing sets from previously encountered

[3] https://data.nal.usda.gov/dataset/what-we-eat-america-wweia-database

Methods	CIFAR100-LT						VFN-LT					
	20 tasks	20 tasks	20 tasks	10 tasks	10 tasks	10 tasks	15 tasks	15 tasks	15 tasks	7 tasks	7 tasks	7 tasks
	M=0.5K	M=1K	M=2K	M=0.5K	M=1K	M=2K	M=0.5K	M=1K	M=2K	M=0.5K	M=1K	M=2K
OnPRO[ICCV '23]	14.02 0.44	16.28 ± 0.81	18.01 ± 0.22	16.53 ± 0.55	16.92 ± 0.08	18.85 ± 0.32	11.93 ± 0.04	12.77 ± 0.07	13.50 ± 0.05	**8.02 ± 0.60**	9.38 ± 0.21	11.84 ± 0.49
SCR[CVPRW '21]	12.22 ± 0.72	13.48 ± 0.90	15.88 ± 0.79	16.65 ± 0.90	17.02 ± 0.77	17.58 ± 0.66	11.55 ± 0.17	11.82 ± 0.10	12.39 ± 0.73	7.71 ± 0.49	9.19 ± 0.46	9.48 ± 0.47
ASER[AAAI '21]	8.86 ± 0.30	7.86 ± 0.61	8.18 ± 0.31	12.68 ± 0.70	13.76 ± 0.01	15.90 ± 0.91	6.85 ± 0.34	7.61 ± 0.38	7.22 ± 0.36	7.46 ± 1.18	7.52 ± 1.09	6.35 ± 0.19
PRS[ECCV '20]	7.61 ± 0.09	7.54 ± 0.21	7.03 ± 0.13	7.34 ± 0.92	8.95 ± 0.33	9.01 ± 0.39	7.17 ± 0.83	8.72 ± 0.15	8.39 ± 0.19	7.85 ± 0.50	8.66 ± 0.22	9.21 ± 0.30
CBRS[ICML '20]	8.51 ± 0.19	8.66 ± 0.61	8.91 ± 0.33	9.50 ± 0.48	7.22 ± 0.43	7.31 ± 0.08	8.12 ± 0.94	8.35 ± 0.33	8.18 ± 0.44	7.52 ± 0.11	7.64 ± 0.08	7.92 ± 0.34
GSS[NeurIPS '19]	5.16 ± 0.10	5.22 ± 0.22	5.09 ± 0.21	8.97 ± 0.65	10.12 ± 0.02	9.96 ± 0.47	5.86 ± 0.30	6.01 ± 0.91	5.86 ± 0.06	5.92 ± 0.54	4.30 ± 0.22	4.66 ± 0.60
LT-CIL(offline)	3.01 ± 0.77	2.67 ± 0.04	2.43 ± 0.02	1.76 ± 0.11	2.36 ± 0.25	3.76 ± 0.22	1.82 ± 0.45	2.02 ± 0.44	2.38 ± 0.08	3.08 ± 0.71	2.92 ± 0.04	1.99 ± 0.31
DELTA (ours)	**16.53 ± 0.01**	**17.71 ± 0.11**	**19.93 ± 0.07**	**20.25 ± 0.71**	**21.06 ± 0.23**	**22.47 ± 0.51**	**12.5 ± 0.01**	**13.45 ± 0.02**	**13.84 ± 0.01**	8.00 ± 0.39	**10.41 ± 0.52**	**12.84 ± 0.54**

Table 1. **Average Accuracy** (%) ± **standard deviation** (↑) in the **long tailed** scenario on Split CIFAR-100-LT, VFN-LT with single exemplar pairing. The best accuracy results are highlighted in **boldface**

tasks. We have included the forgetting metrics in the supplementary section. Let $a_{i,j}$ be the model's performance on the held-out testing set of task j after the model is trained from task 1 to i [44]. For a total of T tasks :

$$\text{Average Accuracy -}A_T = \frac{1}{T}\sum_{j=1}^{T} a_{T,j} \tag{9}$$

5.4. Discussion of Results

In this section, we evaluate the discussed OCL methods under the long-tailed condition (with ρ = 0.01), using varying memory buffer sizes across two datasets: Split CIFAR-100 [36], and the VFN-LT [25] dataset. Table 1 reveals that the average accuracy of current OCL methods is comparatively low in these long-tailed scenarios. In contrast, as shown in Table 2, these methods demonstrate improved performance in conventional settings ($\rho = 1$) where the class distribution is balanced. The diminished performance in long-tailed scenarios is because existing approaches are not designed to handle the significant imbalances commonly present in real-world data. Learning tail classes is challenging in the online scenario with a single pass over the data, often resulting in reduced accuracy, as evidenced in Table 1. Among the existing OCL methods, SCR and CBRS outperform others. SCR achieves this through tightly clustering related class embeddings and using an NCM classifier, while CBRS benefits from class-balanced sampling. Variations in exemplar size, ranging from 0.5K to 2K, reveal that the performance of existing OCL methods is inconsistent with increased buffer size. In contrast, our approach demonstrates consistency and resilience against variations in exemplar size and task sizes. We attribute the performance gain to our DELTA method, which incorporates a dual-stage decoupled learning pipeline with contrastive learning and equalization loss structure. Utilizing contrastive learning, we effectively cluster long-tailed samples and integrate EQ loss results in more accurately learned representations. In the subsequent stage, the decoupling of representation learning allows for a more effective classification task facilitated by the EQ loss (L_{EQ}) that addresses data imbalances. We showcase our method is significantly less biased towards the long-tailed data after training on the last task in Figure 3, whereas other methods are biased towards the classes appearing in the task. Our approach demonstrates consistent performance in terms of accuracy (Table 1)across varying buffer sizes, task sizes, and imbalance ratios(Table 2).

5.5. Ablation Study

Effectiveness of dual-stage approach DELTA and Equalization Loss To verify the effectiveness of the dual-stage approach, we compare the performance in various imbalance ratios ranging from mild to severe imbalance representative of real-world scenarios. We showcase the consolidated results in Table 2. Additionally, we replace the Equalization Loss (detailed in section 4.2) with Cross Entropy (CE) to showcase the enhancements, whereas CE is biased toward classes with more number of samples as it averages the loss over all samples and do not effectively promote learning in the long-tailed scenario as shown in Table 3. In comparison Equalization loss effectively adjusts the prediction scores based on class frequency making the model less biased towards the majority classes leading to improved performance.

Imbalance ratio (ρ)	SCR	CBRS	DELTA *(ours)*
0.005	3.59 ± 0.64	7.60 ± 0.06	**18.02 ± 0.79**
0.03	4.11 ± 0.42	8.88 ± 0.47	**20.21 ± 0.22**
0.07	8.34 ± 0.04	9.47 ± 0.61	**23.60 ± 0.09**
0.1	6.12 ± 0.35	10.26± 0.38	**24.28 ± 0.60**
1.0 (Conventional)	20.74 ± 0.40	15.79 ± 0.33	**33.23 ± 0.97**

Table 2. Ablation study for average accuracy (%) with different imbalance ratios on Split CIFAR-100 for *long-tailed* distributions with a fixed exemplar size 2K and with single-exemplar pairing. Compared against the best performing methods. Smaller the value of ρ, greater the imbalance between most frequent and least frequent classes.

$L_{contrastive}$	L_{CE}	L_{EQ}	**Split CIFAR-100**	**VFN-LT**
✓	✓		16.62 ± 0.91	9.91 ± 0.28
✓		✓	**19.93 ± 0.07**	**13.84 ± 0.01**

Table 3. Ablation study for average accuracy (%) on the loss function used in stage 1 and stage 2 of DELTA in addition to the Contrastive Loss ($L_{contrastive}$).

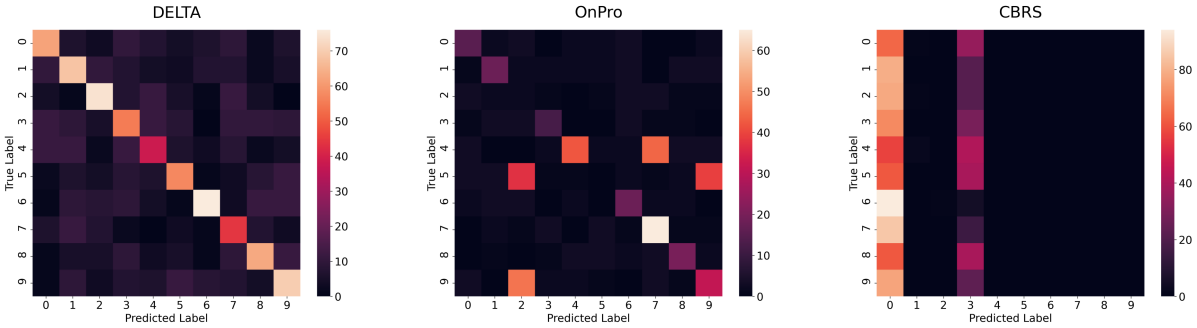

Figure 3. Confusion matrices for DELTA, OnPro [62], and CBRS [10] on CIFAR100-LT with a memory buffer of 2,000 show distinct patterns. Single-stage methods(OnPro, CBRS) are prone to a bias towards recent tasks, particularly with long-tailed samples, often misclassifying numerous samples as belonging to the latest task classes. DELTA exhibits a reduced bias thanks to its unique decoupled learning architecture that incorporates a contrastive learner and employs an equalization loss.

Analysis of multi-exemplar pairing As explained in Section 4.3, we analyze the effect of multi-exemplar pairing within our DELTA method and demonstrate the superiority of our method in terms of learning accuracy, memory efficiency, and time trade-off (during training) through multi-exemplar pairing. For every input image sample, we pair it with one or more exemplars from the buffer, effectively enlarging the training batch size. This strategy aims to mitigate the variance of gradient estimates compared to smaller data batches. This reduced variance means the direction of the gradient descent steps is more consistent and stable, leading to more reliable training progress, even if the data within those batches is not perfectly balanced. The depicted figure 4 presents the average accuracy, average forgetting, and the duration of training for 20 tasks. We observe a decline in performance when each input sample is paired with more than ten exemplars, indicating the onset of overfitting beyond this threshold. It is to be noted that, the performance is improved even in the conventional setting with multi-exemplar pairing as shown in Figure 4 (bottom), with an average accuracy of 55% with ten times exemplar paired. With multi-exemplar pairing in the conventional scenario, we observe from Figure 4 (bottom) that as the system becomes increasingly plastic, this comes at the expense of stability, leading to heightened forgetting. We believe this is linked to the rapid parameter update and increased data flow during rehearsal. Although this facilitates quick learning, it simultaneously risks overwriting the weights responsible for encoding previously acquired knowledge. Additionally, in the conventional case, using equalization loss may not be entirely advantageous when the inherent randomness present in the data is adjusted.

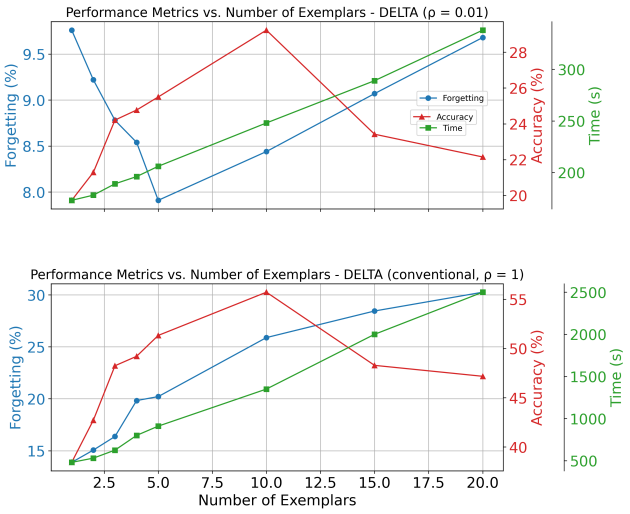

Figure 4. Performance of DELTA at $\rho = 0.01$ (top), and DELTA at $\rho = 1$ (conventional) with an increasing number of paired exemplars. The graph displays CIFAR100-LT utilizing a 2K buffer across 20 tasks.

6. Conclusion

In this work, we focus on long-tailed online class incremental learning for image classification. We introduce a new two-stage method, DELTA, that decouples the learning of features from the classification task in the online setting using contrastive learning and an equalization loss. Additionally, we present early-stage work on multi-exemplar pairing in the LTOCL scenario. Our method shows significantly improved accuracy compared to existing OCL methods, showing a great potential to deploy online continual learning in real-life applications. For future work, we plan to explore representative exemplar selection within multi-exemplar pairing in the LTOCL setting to strike a balance between the stability and plasticity of the continual learner.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. *Proceedings of the European Conference on Computer Vision*, pages 144–161, 2018. 1

[2] Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019. 2, 5

[3] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Proceedings of the conference in Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3, 5

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 3

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 3

[6] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. *Proccedings of the International Conference on Learning Representations*, 2019. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3, 4, 6

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 3

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 4

[10] Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. *Proceedings of the 37th International Conference on Machine Learning*, 119: 1952–1961, 2020. 1, 2, 3, 5, 6, 8

[11] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. *CoRR*, abs/2008.03673, 2020. 4

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. pages 9260–9269, 2019. 3

[13] Keqi Deng, Gaofeng Cheng, Runyan Yang, and Yonghong Yan. Alleviating asr long-tailed problem by decoupling the learning of representation and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:340–354, 2022. 3

[14] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021. 3

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. 2020. 4

[16] Yanan Gu, Xu Yang, Kun Wei, and Cheng Deng. Not just selection, but exploration: Online class-incremental continual learning via dual view consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7442–7451, 2022. 1, 2, 3

[17] Tyler L. Hayes and Christopher Kanan. Online continual learning for embedded devices, 2022. 2, 3

[18] Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. REMIND your neural network to prevent catastrophic forgetting. *CoRR*, abs/1910.02509, 2019. 3

[19] Jiangpeng He and Fengqing Zhu. Online continual learning for visual food classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2337–2346, 2021. 2

[20] Jiangpeng He and Fengqing Zhu. Online continual learning via candidates voting. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3154–3163, 2022. 3

[21] Jiangpeng He and Fengqing Zhu. Exemplar-free online continual learning. *2022 IEEE International Conference on Image Processing*, pages 541–545, 2022. 3

[22] Jiangpeng He and Fengqing Zhu. Single-stage heavy-tailed food classification. *2023 IEEE International Conference on Image Processing*, pages 1115–1119, 2023. 3

[23] Jiangpeng He and Fengqing Zhu. Gradient reweighting: Towards imbalanced class-incremental learning, 2024. 3

[24] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13926–13935, 2020. 1, 3

[25] Jiangpeng He, Luotao Lin, Heather Eicher-Miller, and Fengqing Zhu. Long-tailed food classification. *arXiv preprint arXiv:2210.14748*, 2022. 2, 3, 6, 7

[26] Jiangpeng He, Luotao Lin, Jack Ma, Heather A Eicher-Miller, and Fengqing Zhu. Long-tailed continual learning for visual food recognition. *arXiv preprint arXiv:2307.00183*, 2023. 3

[27] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. pages 6626–6636, 2021. 3

[28] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 3

[29] Zhehao Huang, Tao Li, Chenhe Yuan, Yingwen Wu, and Xiaolin Huang. Online continual learning via logit adjusted softmax, 2023. 3

[30] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2021. 3

[31] Lie Ju, Xin Wang, Lin Wang, Tongliang Liu, Xin Zhao, Tom Drummond, Dwarikanath Mahapatra, and Zongyuan Ge. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 3–12, Cham, 2021. Springer International Publishing. 2

[32] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *CoRR*, abs/1910.09217, 2019. 3

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3

[34] Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. *Proceedings of European Conference on Computer Vision*, 12345:411–428, 2020. 1, 2, 3, 5, 6

[35] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 1, 3

[36] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 6, 7

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *ACM*, 60(6):84–90, 2017. 1

[38] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. *CoRR*, abs/2005.04966, 2020. 3

[39] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. 1

[40] H. Lin, B. Zhang, S. Feng, X. Li, and Y. Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. pages 24246–24255, 2023. 1, 2, 3

[41] Xialei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. Long-tailed class incremental learning. In *European Conference on Computer Vision*, pages 495–512. Springer, 2022. 2, 3, 5, 6

[42] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6470–6479, 2017. 3

[43] Zheda Mai, Ruiwen Li, Hyunwoo J. Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3584–3594, 2021. 1, 2, 3, 4

[44] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomput.*, 469(C):28–51, 2022. 3, 4, 6, 7

[45] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. Visual aware hierarchy based food recognition. *Proceedings of the International Conference on Pattern Recognition Workshop*, pages 571–598, 2021. 2

[46] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989. 1

[47] Zhongqi Miao, Ziwei Liu, Kaitlyn M Gaynor, Meredith S Palmer, Stella X Yu, and Wayne M Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10):885–895, 2021. 2

[48] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 1

[49] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896, 2022. 3

[50] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[51] Siddeshwar Raghavan, Jiangpeng He, and Fengqing Zhu. Online class-incremental learning for real-world food image classification. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8195–8204, 2024. 2

[52] Nishant Rai, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Cocon: Cooperative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3384–3393, 2021. 3

[53] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017. 2

[54] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 3, 4

[55] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9630–9638, 2021. 5

[56] Ghalib Ahmed Tahir and Chu Kiong Loo. An open-ended continual learning for food recognition using class incremental extreme learning machines. *IEEE Access*, 8:82328–82346, 2020. 2

[57] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[58] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 3

[59] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. *Proceedings of the 24th international conference on Machine learning*, pages 935–942, 2007. 3

[60] Quanziang Wang, Renzhen Wang, Yichen Wu, Xixi Jia, and Deyu Meng. Cba: Improving online continual learning via continual bias adaptor. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19036–19046, 2023. 3

[61] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017. 3

[62] Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. Online prototype learning for online continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18764–18774, 2023. 5, 6, 8

[63] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *CoRR*, abs/2008.05659, 2020. 3

[64] Haiyang Yu, Ningyu Zhang, Shumin Deng, Zonggang Yuan, Yantao Jia, and Huajun Chen. The devil is the classifier: Investigating long tail relation classification with decoupling analysis. *CoRR*, abs/2009.07022, 2020. 3

[65] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[66] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 4

[67] Ruru Zhang, Haihong E, Lifei Yuan, Jiawen He, Hongxing Zhang, Shengjuan Zhang, Yanhui Wang, Meina Song, and Lifei Wang. Mbnm: Multi-branch network based on memory features for long-tailed medical image recognition. *Computer Methods and Programs in Biomedicine*, 212:106448, 2021. 2

[68] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10795–10816, 2021. 2

[69] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021. 3

[70] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. *CoRR*, abs/1912.02413, 2019. 3

[71] Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *AAAI Conference on Artificial Intelligence*, 2022. 4