Confusion Mixup Regularized Multimodal Fusion Network for Continual Egocentric Activity Recognition

Hanxin Wang* Shuchang Zhou* Qingbo Wu† Hongliang Li Fanman Meng Linfeng Xu
Heqian Qiu
University of Electronic Science and Technology of China
{hxwang09, sczhou}@std.uestc.edu.cn {qbwu, hlli, fmmeng, lfxu, hqqiu}@uestc.edu.cn

Abstract

Continual egocentric activity recognition aims to understand diverse first-person activities from the multimodal data of a wearable device captured in streaming environments, which is an emerging and challenging task. Existing continual learning methods ignore the dynamic change of multiple modalities' correlation and hardly learn discriminative representations for the sequentially isolated activity classes from different stages. In this paper, we propose a Confusion Mixup Regularized Multimodal Fusion Network (CMR-MFN) to address this issue. Firstly, CMR-MFN is composed of a ternary-modality-input dynamic expansion architecture, which progressively grows additional branches for in-stage class recognition. Each input owns a frozen modality-specific backbone to avoid forgetting caused by parameter shifts. Secondly, CMR-MFN captures the dynamics of multimodal inputs via learnable self-attention layers. We augment unknown classes by linearly mixing up the samples from two known classes and assigning a biased weight to one of them, which makes the unknown class samples confusing toward the known class with a higher weight. By learning from the current and augmented training data together, we regularize the multimodal fusion representation to distinguish the in-stage classes from their confusing samples of unknown classes, which implicitly pushes the out-stage classes' samples far from the in-stage classes' ones when they are similar to each other. Experiments show that the proposed method significantly outperforms state-of-the-art methods for multimodal continual egocentric activity recognition. Our code is available at `https://github.com/Hanna-W/CMR-MFN`.

1. Introduction

Multimodal egocentric activity recognition [12, 11, 23, 36] refers to the task of recognizing and understanding human activities from a first-person perspective using multiple modalities of data, such as visual, audio, and inertial sensor data [19, 18, 17, 21, 20]. This field of research focuses on developing algorithms and models that can analyze and interpret the actions and behaviors of individuals captured through wearable devices, like head-mounted cameras or smart glasses. However, in practical applications, the training data is typically acquired in stages rather than being obtained all at once as in the traditional training paradigm. Therefore, continual egocentric activity recognition is highly desirable in practical applications. In this paper, we explore a multimodal continuous learning method for egocentric activity recognition.

The main challenge of Continual Learning (CL) is how to strike a balance between acquiring new knowledge and preserving old knowledge, which is also known as the stability-plasticity dilemma [7]. The root of this problem lies in two aspects: data isolation and unified architecture [28], which force the model to overfit the data at the current stage and lead to catastrophic forgetting. More intuitively, catastrophic forgetting is the result of confusion between the representations of data from different stages in the feature space [44]. This problem becomes more serious for multimodal data, which may confuse with each other when even only one modality is overlapped in the feature space. The current continual learning methods primarily focus on single modality data and do not take into account the representation discriminability of fusing multiple modalities in streaming environments. As a result, multimodal continual learning becomes even more challenging.

Most of the existing methods attempt to tackle the aforementioned issues of data isolation and unified architecture through rehearsal and dynamic architecture. Rehearsal-based works [16, 32, 10, 37, 42, 44, 45, 43] involve the storage of prototypes or exemplar instances from previous

[*] These authors contributed equally to this work.

[†] Corresponding author.

This work was supported in part by the National Key R&D Program of China (2021ZD0112001), the National Natural Science Foundation of China (No.61971095), and the Independent Research Project of Civil Aviation Flight Technology and Flight Safety Key Laboratory (FZ2022ZZ06).

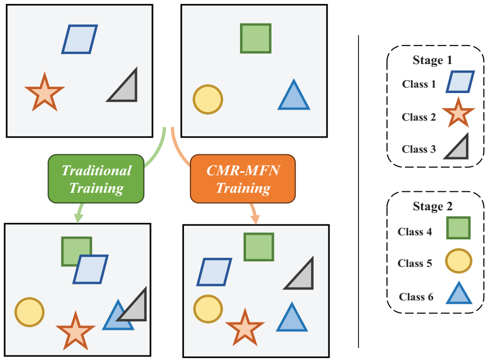


Figure 1. Top: In-stage classes have discriminative representations. Bottom: Traditional training hardly learns discriminative representations for isolated in-stage and out-stage classes, while CMR-MFN training will alleviate the confusion between in-stage and out-stage classes.

tasks, which are subsequently replayed during the training of new tasks. Unfortunately, this is particularly true in the context of first-person behavior recognition, where privacy considerations are paramount. Recent approaches based on dynamic architecture [34, 26, 35, 15] have demonstrated impressive performance by incorporating new modules for learning new tasks while preserving the knowledge of old ones. Thereby, designing based on dynamic architecture enables better handling of a growing training distribution while maintaining the learned parameters associated with previous classes fixed.

To address these issues, in this paper, we aim to present a multimodal continuous learning method that incorporates three properties: 1. rehearsal-free: The method does not rely on the use of any replay samples or prototypes during the training process; 2. dynamic expansion: The method utilizes a network architecture that dynamically expands; 3. generalized multimodal fusion: The method incorporates a learnable fusion network that is capable of acquiring more discriminative multimodal representations.

To this end, we initially employ a dedicated pre-trained transformer backbone for each modality and keep them frozen as a strong prior. Then we adopt the strategy of stage isolation which allows each stage to independently learn the fusion network and classifier. This approach enables the fusion network and classifier to achieve the optimal performance in their stage as shown at the top of Fig. 1. Each fusion network consists of a self-attention layer, enabling it to learn better feature representations by dynamically capturing the intrinsic connections among different modalities. However, traditional training would lead to stage-level overfitting. As shown in the bottom-left of Fig. 1: the data outside the stage may not learn a discriminative representation and it is easy to overlap with the data in the stage. Hence, inspired by Mixup [39], we synthesize confusing samples by linearly combining any two known categories and giving a higher weight to one of them. Through the joint training of the current and augmented data, regularization is applied to the representation learning of the multimodal fusion network so that the learned multimodal representation can effectively distinguish the in-stage and out-stage data.

In conclusion, we propose a rehearsal-free multimodal continuous learning method for egocentric activity recognition called CMR-MFN, our main contributions are summarized as follows:

- We design a dynamic expansion fusion architecture to ensure the data within each stage can learn the optimal multimodal representation.

- We introduce a called confusion mixup regularized multimodal fusion network that can capture the dynamic change of correlation from different modalities and help alleviate the confusion between the in-stage data and out-stage data in the feature space.

- Our method significantly outperforms SOTA unimodal methods on existing multimodal continual learning benchmarks for egocentric activity recognition.

2. Related Work

2.1. Multimodal Egocentric Activity Recognition

Given the wide range of modalities through which egocentric activities can be represented, an increasing amount of research is dedicated to exploring the multimodal domain for first-person activity recognition. **Audio data** provide complementary information to appearance and motion in visual data. TBN [12] draws inspiration from TSN [27] and utilizes a temporal binding window to fuse audiovisual features. This approach combines modalities before temporal aggregation, using shared modality and fusion weights over time. [11] presents MTCN, a transformer-based model that learns to focus on surrounding activities and model multimodal temporal context. **Inertial sensors data** from accelerometers and gyroscopes have been used for egocentric activity recognition, allowing recognition beyond the limited field of view of vision-based sensors. A hierarchical fusion framework is presented in [23, 36], utilizing LSTM and CNN based on motion sensor data and photo streams at different levels, respectively. MKL [2] is proposed to adaptively weigh the visual, audio, and sensor features, additionally, feature and kernel weighting and recognition tasks are performed simultaneously. [9] introduces a first-view multimodal framework based on knowledge-driven approaches, GCN and LSTM.

2.2. Continual Learning

Regularization-based methods can be categorized into two aspects based on parameter regularization and knowledge distillation. Parameter regularization-based methods [13, 38, 1] aim to retain previous class knowledge by penalizing changes to former classes during model updates. In [13], the importance of parameters is assessed using the Fisher information matrix, and significant parameters are restricted in their updates. However, conflicts arise due to the differing importance matrices for each task, which subsequently impacts the effectiveness of the algorithm. On the other hand, knowledge distillation-based methods [14, 5, 40, 30] employ implicit regularization by applying knowledge distillation techniques to continuous learning. BiC [30] and WA [40] propose effective solutions to address the issue of classifier bias that arises after distillation.

Rehearsal-based methods primarily rely on utilizing old-task data to mitigate catastrophic forgetting. [16, 30, 40, 8] allocate specific memory to store exemplars of past tasks, allowing access to a portion of the old data which is then replayed to reinforce previous knowledge during the learning of new tasks. Furthermore, some approaches [31, 32, 22] involve generating old-task data using separate generative models for replay, rather than directly storing the original data. However, it is important to consider that storing large amounts of old-task data can be memory-intensive and may raise privacy concerns. Additionally, generative models often face challenges such as instability and inefficiency during training. To address these issues, some researchers have focused on leveraging class-representative prototypes of old data. For example, [45] adopts a prototype selection strategy while [44] focuses on prototype augmentation. In addition, [43] proposes classAug on the basis of prototype augmentation to prevent the bias of prototype representation.

Dynamic-networks-based methods design dynamic modules to satisfy evolving training distributions without task identifiers. [34] continuously expands feature extractors, which are subsequently fed into a unified classifier. Furthermore, the network undergoes pruning after model learning. Similarly, [6] introduces a task dynamic strategy based on the transformer architecture, where task tokens are continuously expanded without requiring any hyperparameter adjustments to control network expansion. Another notable contribution is made by [29], which proposes a novel framework for continual learning through prompt tuning. Additionally, [26] presents a two-stage learning paradigm that utilizes dynamic expansion modules and compression models based on the gradient boosting algorithm. However, conflicts can arise between different modules during dynamic expansion. To address this, [25] introduces a unified energy-based theory and framework to mitigate conflicts in the expansion process.

3. Proposed Method

3.1. Problem Definition

Continuous learning aims to enable the model to learn a continuous stream of information, only accessing the current stage samples during training, and effectively categorizing the test samples from all previously learned stages during testing. Let $\left\{\mathcal{D}^1, \mathcal{D}^2, .., \mathcal{D}^T\right\}$ be continuous data stream, where$\{1, 2, .., T\}$ is the sequence of stages. The incoming data at stage t is denoted as $\mathcal{D}^t = \{x_b^t, y_b^t\}_{b=1}^{N^t}$, which have N^t labeled samples of this stage. Specifically, the model receives $x_b^t = \{v_b^t, a_b^t, g_b^t\}$ as the input of multimodal data, where v_b^t, a_b^t, g_b^t represent the visual signal (R), acceleration signal (A) and gyroscope signal (G), respectively. y_b^t is the corresponding label for x_b^t which is expressed in the form of one-hot encoding: $y_b^t = [l_1, l_2, \cdots, l_k]$, where k is the number of category, $l_i = 1$ when it belongs to class i otherwise $l_i = 0$. For class incremental learning, y_b^t is selected from $\mathcal{Y}^t$ where $\mathcal{Y}^t$ represents the label space of class groups without overlapping classes. In the t-th incremental step, the model learns knowledge from the available training samples $\mathcal{D}^t$, and then is expected to perform well on all seen classes $\tilde{\mathcal{Y}}^t = \cup_{i=1}^t \mathcal{Y}^i$.

3.2. Overview of Framework

In this paper, we present a rehearsal-free multimodal continual learning approach. The framework of our method is shown in Fig. 2. During the training phase, we use frozen pre-trained transformers to extract features for each modal input. Before multimodal fusion, we introduce confusion samples generated by mixup in current training data to regularize multimodal fusion network learning. Then we train the fusion network and classifier independently at each stage. During the testing phase, the test samples choose the prediction result by selecting the highest score from all classifiers.

The detail of the feature extractor is shown in Fig.3. At step t, for the visual data, we uniformly sample 8 RGB frames v_b^t from the video. Subsequently, we feed them into the feature extractor $\mathcal{F}_v$, yielding the RGB feature f_v^t. Regarding the inertial sensor data, considering that both the accelerometer and gyroscope provide three-dimensional data a_b^t and g_b^t, three spectrograms are generated after STFT, whereafter individually passed to the corresponding feature extractors $\mathcal{F}_a$ and $\mathcal{F}_g$ to obtain their respective features f_a^t and f_g^t. For simplicity, for image frames that contain time information, we employ pre-trained TimeSformer [3] as the backbone, while for the spectrograms derived from inertial sensor data, we utilize pre-trained ViT-B/16 [4].

In the following sections, we will first provide a comprehensive introduction to the confusion mixup strategy, which effectively addresses the confusion between in-stage

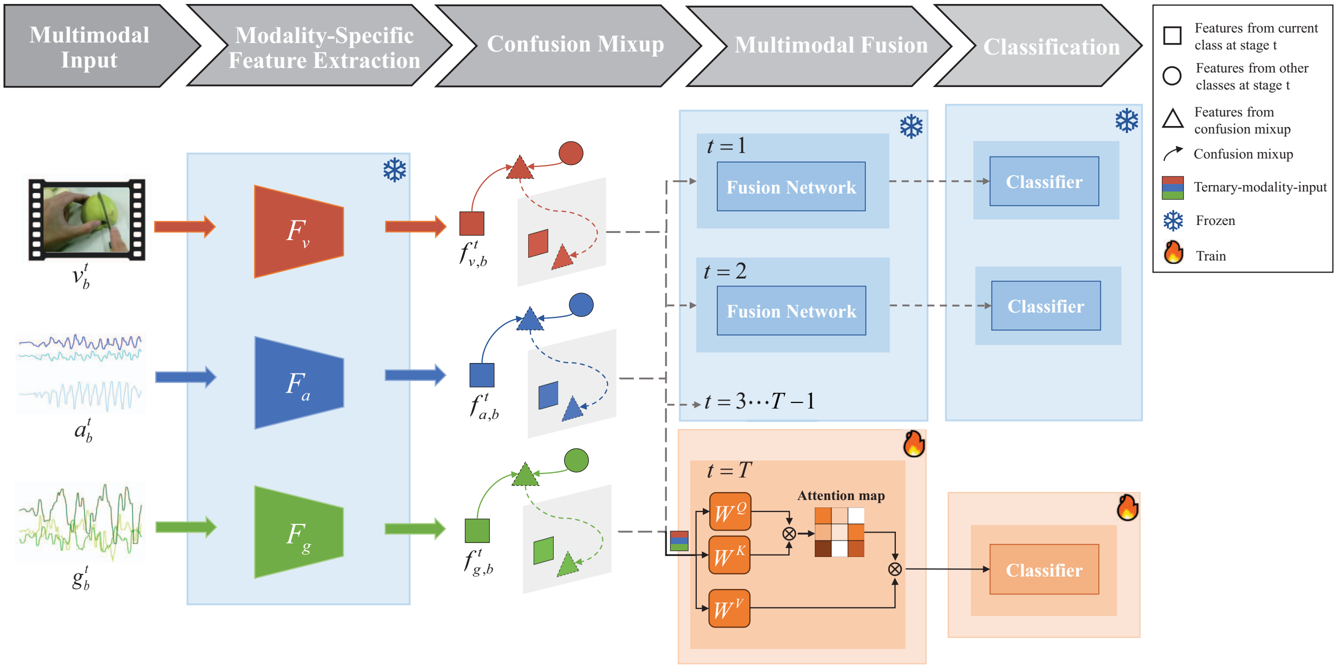


Figure 2. Framework of CMR-MFN. The method involves several steps during training: feature extraction for each modality, addition of generated confusing samples and independent training of the fusion network and classifier at each stage. In the inference phase, the test samples select the highest score from all classifiers as the prediction result.

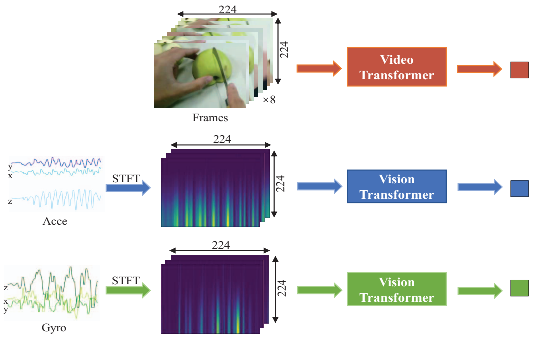


Figure 3. The specific architecture of the feature extractor.

and out-stage data caused by data isolation. Additionally, we will present the self-attention based multimodal fusion process. By independently training the fusion network and classifiers, they can achieve optimal performance at their respective stages without being influenced by other stages.

3.3. Confusion Mixup

In the training phase, only the data from the current stage is available, while all previously learned classes are seen during testing. Traditional training methods would lead to overfitting at the stage level, causing confusion in feature space between in-stage and out-stage data and resulting in catastrophic forgetting. Typically, these confusing out-stage data exhibit similar characteristics to a specific category within the stage. To address this issue, we explore the use of mixup to synthesize similar samples based on available data and regularize the fusion network to learn more generalized multimodal representations.

Specifically, at stage t, the incoming data $\mathcal{D}^t$ contains a total of k classes. For each class n, we assign a confusion class $n+k$. During training, we fuse the embedding of the pair (x_i^t, x_j^t) from two different classes a and b in the mini-batch as a confusion sample $x^{t'} = \left\{f_v^{t'}, f_a^{t'}, f_g^{t'}\right\}$:

$$f_m^{t'} = \lambda_m \mathcal{F}_m(m_i^t) + (1-\lambda_m)\mathcal{F}_m(m_j^t), \qquad (1)$$

where $m \in [v, a, g]$ and λ is sampled from Beta(α, α). In contrast to the setting described in [43], we impose a restriction on the sampling of λ, confining it to the interval of $[0.5, 1]$. As shown in Fig. 4 and 5, our confusion mixup ensures that the synthesized samples of the new class are closer to one class of the original data.

And the sample generated by Eq. 1 would be labeled as:

$$y^{t'} = [l_1, \cdots, l_{a+k}, \cdots, l_{2k}], \qquad (2)$$

where l_{a+k} is equal to 1 and the others are equal to 0. We represent the synthetic data set as $\mathcal{D}^{t'}$ while the corresponding label set as $\mathcal{Y}^{t'}$. Therefore, the original k-class problem in the current stage is transformed into a $2k$-class problem. Furthermore, in order to maintain sample balance, the number of each synthesis will align with the batch size. In our all experiments, α is set to 0.2.

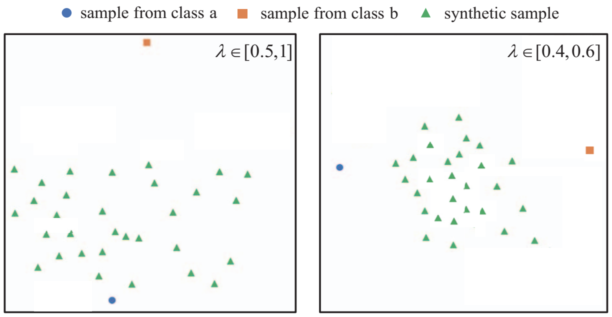


Figure 4. The illustration of mixup under different λ settings, taking RGB as an example. We randomly select two samples from class a and class b, generate 50 new samples through mixup, and plot them in a 2D space for visualization.

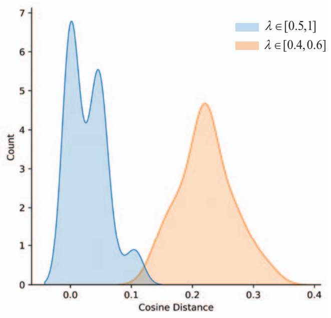


Figure 5. The distance distribution between the synthetic samples and a specific original sample. The y-axis is the count of mixup samples, and the x-axis is the cosine distance to the class a.

3.4. Self-attention based Multimodal Fusion

Following data augmentation through confusing mixup, we employ a self-attention layer [24] as the fusion network to fuse the features from different modalities, denoted as $\mathcal{SA}$. $\mathcal{SA}$ consists of three parts: query Q, key K, and value V. The input $Z = [f_v||f_a||f_g]$, where $||$ denotes the concatenation of the features for each modality, will be projected into the same space, represented as $[Q, K, V] = [W^Q Z, W^K Z, W^V Z]$. The final output of $\mathcal{SA}(Z)$ can be computed as follows:

$$\mathcal{SA}(Z) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k/h}}\right)V, \tag{3}$$

where d_k represents the embedding dimension, and h indicates the number of attention heads. Moreover, training a single fusion network by fine-tuning parameters allows it to accommodate new tasks but possibly results in forgetting the knowledge acquired from previous ones.

To this end, we introduce expandable multimodal fusion networks. Initially, we have only one fusion network denoted as $\mathcal{SA}_1$. As we progress to each new stage, we propose expanding the parameter space by creating a new fusion network while retaining the previous ones. Correspondingly, we extend the classifier at each new stage. This implies that at step t, we train a new fusion network $\mathcal{SA}_t$ independently, along with a corresponding classifier $\mathcal{H}_t$. Our expansion would add approximately 0.7% new parameters in each stage. Additionally, the incremental learning process only involves training these new parameters, leading to a significantly faster model training.

3.5. Optimization Objective of CMR-MFN

By combining the aforementioned techniques, we obtain a complete loss function of CMR-MFN which consists of two terms:(1) the cross entropy loss $\mathcal{L}_{\text{ce}}$ of the original data D^t, (2) the cross entropy loss $\mathcal{L}'_{\text{ce}}$ of the synthetic data $D^{t'}$. The total loss is presented as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}'_{\text{ce}}. \tag{4}$$

During the training process of stage t, $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}'_{\text{ce}}$ can be calculated as:

$$\mathcal{L}_{\text{ce}} = F_{ce}(\mathcal{H}_t(\mathcal{SA}_t(\mathcal{F}(\mathcal{D}^t)); \mathcal{Y}^t), \tag{5}$$

$$\mathcal{L}'_{\text{ce}} = F_{ce}(\mathcal{H}_t(\mathcal{SA}_t(\mathcal{F}(\mathcal{D}^{t'})); \mathcal{Y}^{t'}), \tag{6}$$

where F_{ce} represents the standard cross entropy function.

When the testing phase at stage t, the test sample $\{x, y\} \in \cup_{i=1}^{t} \mathcal{D}^i$ traverses through a series of fusion networks $\{\mathcal{SA}_1, \mathcal{SA}_2, .., \mathcal{SA}_t\}$ and the corresponding classifiers $\{\mathcal{H}_1, \mathcal{H}_2, .., \mathcal{H}_t\}$. It should be noted that the additional class nodes generated by confusion mixup in the classifier will be discarded. Thus, the output of each classifier can be formulated as:

$$P_{\mathcal{H}_i}(y \mid x) = \text{Softmax}(\mathcal{H}_i(\mathcal{SA}_i(\mathcal{F}(x)))[:k]). \tag{7}$$

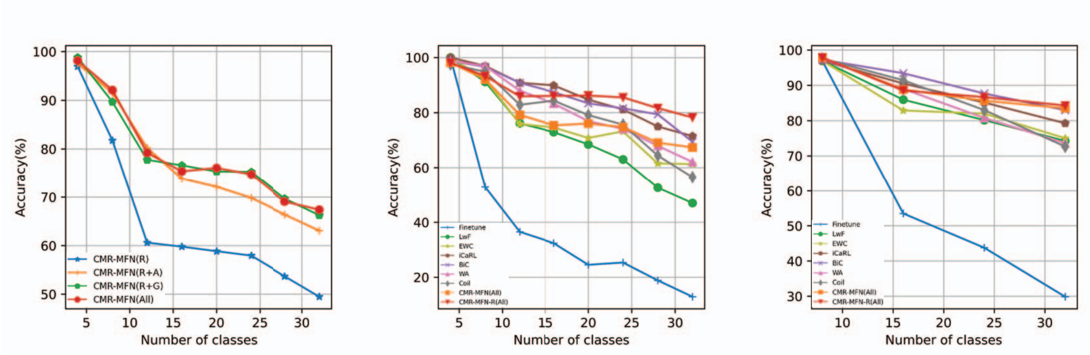
Ultimately, we integrate the outputs from all the classifiers and select the category with the highest confidence as the final prediction result:

$$\hat{y} = \arg\max P_{\mathcal{H}_1||\cdots||\mathcal{H}_t}(y \mid x), \quad \hat{y} \in \tilde{\mathcal{Y}}_t. \tag{8}$$

4. Experiments

4.1. Benchmarks & Implementation

Benchmarks. We evaluate our model on UESTC-MMEA-CL [33]. UESTC-MMEA-CL is the first multi-modal dataset for continual egocentric activity recognition. It contains 30.4 hours of video clips, accelerometer data and gyroscope data. UESTC-MMEA-CL comprises 32 daily activities including basic human movements, indoor work tasks, leisure activities, etc. The standard continual scenario in UESTC-MMEA-CL has 8 steps and 4 steps. Thus we compare performances on 4 classes per step and 8 classes per step on UESTC-MMEA-CL.



(a)

(b)

(c)

Figure 6. (a) Results of CMR-MFN with different modal inputs on UESTC-MMEA-CL with 8 steps. (b) Results on UESTC-MMEA-CL with 8 steps. (c) Results on UESTC-MMEA-CL with 4 steps.

Comparison Methods. We compare our proposed method against the state-of-the-art CL methods. Comparison methods include rehearsal-free methods LwF [14], EWC [13] and ESN [28], as well as rehearsal-based methods iCaRL [16], BiC [30], WA [40], Coil [42] and DyTox [6]. DyTox and ESN are recently published transformer-based methods. Besides, we use joint training performance as the upper bound for CMR-MFN. For fair comparison, we use the same ImageNet pre-trained transformer backbone for all comparison methods and CMR-MFN. Specifically, all comparison methods rely on visual single-modal input with the RGB backbone being Timesformer, which is identical to CMR-MFN's. Moreover, we adopt the same settings as in the original work for DyTox and ESN, which are specially designed based on ViT. To ensure uniformity, we replace our RGB backbone with ViT-B/16, referred to as CMR-MFN*. In addition, we equip our approach with a rehearsal buffer called CMR-MFN-R for a fair comparison with rehearsal-based methods.

Implementation Details. For inertial sensor data processing, to address the outliers in the acceleration and gyroscope signals, we employ a median filter with a kernel size of 5. Additionally, we mitigate the bias drift in the gyroscope signals by subtracting the mean value. We extract 10.32s of inertial sensor data from a video and convert them into spectrograms using the STFT with a sampling frequency of 25Hz, window length of 4, overlap rate of 2, and nfft of 256. Finally, we generate two-dimensional spectrograms of size 224 × 224. For more training details, we implement our methods in PyTorch and PyCIL [41] with a single NVIDIA RTX 3090 GPU. All data streams (R, A, G) are trained by the Adam optimizer with a weight decay of 0.0005 and a learning rate of 0.001. The batch size is set to 16 and the dropout is 0.5. All networks are trained for 50 epochs, and the learning rate is decayed by a factor of 10 at epoch 10 and 20.

4.2. Metrics

Following [33], we use average accuracy and average forgetting as evaluation metrics, which are defined as follows:

Average Accuracy (AA) Define $a_k^j (j \leq k)$ as the accuracy evaluated on task j after training task k. Thus average accuracy on task k can be calculated as $A_k = \frac{1}{k}\sum_{j=1}^{k} a_k^j$.

Average Forgetting (F) Forgetting denotes the knowledge forgetting degree about the task throughout the learning process. It is defined as the difference between the maximum accuracy during the learning process and the current accuracy, which can be formulated as

$$f_k^j = \max_{l \in 1,\cdots,k-1} (a_l^j - a_k^j), \tag{9}$$

thus average forgetting on task k can be calculated as $F_k = \frac{1}{k-1}\sum_{j=1}^{k-1} f_k^j$.

4.3. Quantitative Results

We conduct experiments based on UESTC-MMEA-CL. The performance curves are illustrated in Fig. 6, while Table 1 and Table 2 report the average accuracy and average forgetting for 4 and 8 incremental tasks, respectively. Additionally, Table 3 displays the results of transformer-based methods.

Average accuracy. In Fig. 6 (a), we can clearly see that by incorporating the inertial sensor modality accelerometer and gyroscope, the multimodal prediction accuracy significantly surpasses the accuracy of the single modality, starting from task 2. However, as the number of modalities increases, the rate of improvement starts to diminish. Finally, CMR-MFN(All) achieves an accuracy of 67.4% for 4 class-incremental learning, which is over 18% higher than the accuracy obtained with the single RGB modality. Compared to other methods, CMR-MFN(All) also demonstrates state-of-the-art performance in average accuracy, both with

Methods	Memory Size	AA(↑)	F(↓)
iCaRL	150	71.5	30.97
BiC		69.45	30.17
WA		62.08	41.12
Coil		56.53	38.67
CMR-MFN-R(All)		**78.27**	**5.47**
FT	0	12.92	98.08
LwF		47.04	43.53
EWC		61.32	36.32
CMR-MFN(R)		49.47	**16.52**
CMR-MFN(R+A)		63.07	18.3
CMR-MFN(R+G)		66.26	18.91
CMR-MFN(All)		**67.4**	19.96
Upper-bound	-	94.6	1.9

Table 1. Results on UESTC-MMEA-CL for 4 class-incremental learning.

Methods	Memory Size	AA(↑)	F(↓)
iCaRL	150	79.26	24.52
BiC		82.98	16.8
WA		73.63	31.97
Coil		72.64	28.89
CMR-MFN-R(All)		**84.27**	**6.4**
FT	0	29.86	91.13
LwF		74.24	17.94
EWC		75	26.76
CMR-MFN(R)		67.4	**5.49**
CMR-MFN(R+A)		81.69	7.74
CMR-MFN(R+G)		80.09	8.34
CMR-MFN(All)		**83.51**	8.85
Upper-bound	-	95.14	1.22

Table 2. Results on UESTC-MMEA-CL for 8 class-incremental learning.

Methods	Memory Size	AA(↑)	F(↓)
DyTox	150	63.6	16.11
ESN	0	63.83	**15.22**
CMR-MFN*(All)		**73.02**	19.91
Upper-bound	-	95.29	1.22

Table 3. Results of transformer-based methods on UESTC-MMEA-CL for 4 class-incremental learning.

and without exemplars. Remarkably, our rehearsal-free method outperforms individual exemplar-based methods significantly, particularly with 8 incremental tasks. Specifically, CMR-MFN(All) exhibits a 0.53% superiority over the best performing exemplar-based method BiC.

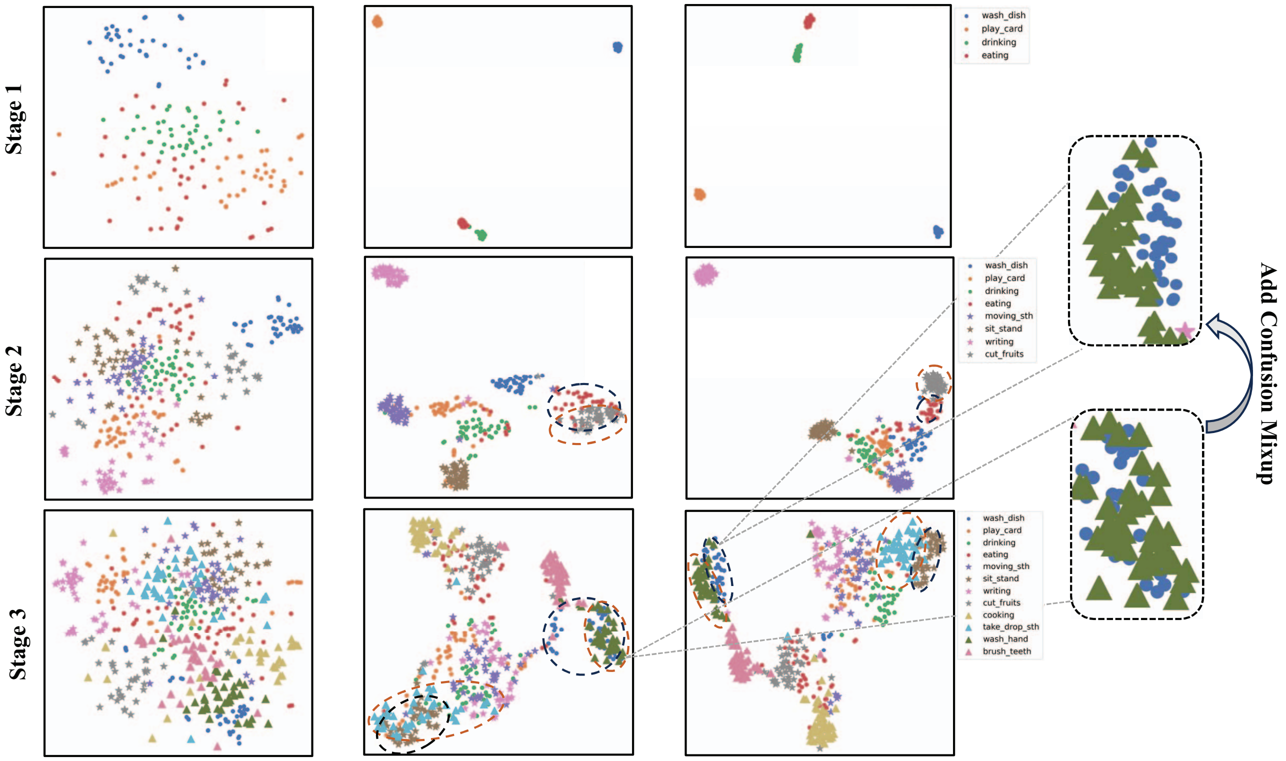
Average Forgetting. From Table 1 and Table 2, it can be observed that, similar to the average accuracy, the average forgetting also exhibits an increasing trend as modal input rises. This is because CMR-MFN(R) has a limited ability to acquire new knowledge, resulting in less forgetting. Nevertheless, CMR-MFN-R(All) also reaches 5.47% in average forgetting, which is in close proximity to the upper limit of 1.9%. CMR-MFN-R(All) demonstrates significantly lower average forgetting compared to methods utilizing exemplars. Even in the absence of exemplars, our method exhibits superior performance compared to most methods, showcasing its ability to resist catastrophic forgetting.

4.4. Ablation Study

To the efficacy of our proposed model, we conduct a series of ablation experiments on the UESTC-MMEA-CL dataset, and all experimental results are based on the ternary-modality-input. The performance improvement of our proposed model can be primarily attributed to two essential components: the learnable multimodal fusion and the confusion mixup strategy.

The Effect of learnable multimodal fusion. To demonstrate the functionality of the learnable multimodal fusion, we compare it with the baseline, which solely employs a fusion network to concatenate the multimodal data. As evident from Table 4, the multimodal fusion network with a self-attention layer brings a 19.3 % improvement on the average accuracy, while the average forgetting decreases by 2.66 %. Additionally, in Fig.7 (a) and (b), it is evident that the learnable multimodal fusion networks enhance the aggregation and discriminative capacity of representations.

The Effect of confusion mixup. To demonstrate the superiority of the confusion mixup strategy, we compared CMR-MFN's performance against the model that solely utilizes the learnable multimodal fusion. As shown in Table 4, the confusion mixup strategy led to a 3.65 % improvement on the average accuracy. Fig. 7 visually presents the impact of the confusion mixup strategy on the representation from the feature space. Specifically, in subfigure (b), the orange ellipse area and the black ellipse region overlap to a large extent, indicating that it is challenging to differentiate between in-stage and out-stage data. For instance, the classes ”wash_hand” (in-stage) and ”wash_dish” (out-stage) display substantial similarity at stage 3, leading to significant confusion when employing solely multimodal fusion networks. As the number of learning tasks increases, the likelihood of encountering similarities between classes from different stages also rises, exacerbating the confusion between classes. Conversely, in subfigure (c), the mixup strategy effectively enhances the discriminative capability



(a) (b) (c) (d)

Figure 7. The influence of our model on representations. Each stage encompasses four classes, visually represented by the orange dotted elliptical area indicating the in-stage classes and the black dotted ellipse area representing the out-stage classes. (a) Baseline. (b) Learnable multimodal fusion. (c) CMR-MFN. (d) Partial enlargement. We can clearly see that CMR-MFN results in a better distinction between in-stage and out-stage classes.

of the in-stage classes from the out-stage classes. In more detail, subfigure (d) clearly shows that CMR-MFN efficiently reduces the overlap between the ”wash_hand” and wash_dish classes. Furthermore, when incorporating the confusing samples into the data within stages, such as the classes ”cooking” and ”brush teeth”, where the representations in subfigure (b) already exhibit the ability to distinguish classes from different stages, the discriminatory capacity of the representations remains unaffected. This is evident in subfigure (c), where there is virtually no negative impact. The same observation holds true for stage 2.

Componets			
Learnble Fuison	**Confusion Mixup**	**AA(↑)**	**F(↓)**
✗	✗	44.45	22.60
✓	✗	63.75	**19.94**
✓	✓	**67.4**	19.96

Table 4. Ablation study of our method on UESTC-MMEA-CL.

5. Conclusion

In this paper, we propose a novel and effective method of CMR-MFN for continual egocentric activity recognition. CMR-MFN incorporates a ternary-modality-input dynamic expansion architecture with learnable self-attention layers. Furthermore, we employ a confusion mixup strategy to regularize the multimodal fusion representations. Exhaustive experiments conducted on the latest UESTC-MMEA-CL database demonstrate that our proposed method is significantly better than state-of-the-art approaches for continual egocentric activity recognition.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.

[2] Mehmet Ali Arabacı, Fatih Özkan, Elif Surer, Peter Jančovič, and Alptekin Temizel. Multi-modal egocentric activity recognition using multi-kernel learning. *Multimedia Tools and Applications*, 80(11):16299–16328, 2021.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020.

[6] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9275–9285, June 2022.

[7] Stephen T Grossberg. *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*, volume 70. Springer Science & Business Media, 2012.

[8] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019.

[9] Yi Huang, Xiaoshan Yang, Junyu Gao, Jitao Sang, and Changsheng Xu. Knowledge-driven egocentric multimodal activity recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(4):1–133, 2020.

[10] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 699–715. Springer, 2020.

[11] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

[12] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[13] 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, Mar. 2017.

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

[15] Zhuoyun Li, Changhong Zhong, Sijia Liu, Ruixuan Wang, and Wei-Shi Zheng. Preserving earlier knowledge in continual learning with the help of all previous feature extractors. *arXiv preprint arXiv:2104.13614*, 2021.

[16] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[17] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018.

[18] Hengcan Shi, Hongliang Li, Fanman Meng, Qingbo Wu, Linfeng Xu, and King Ngi Ngan. Hierarchical parsing net: Semantic scene parsing from global scene to objects. *IEEE Transactions on Multimedia*, 20(10):2670–2682, 2018.

[19] Hengcan Shi, Hongliang Li, Qingbo Wu, Fanman Meng, and King N Ngan. Boosting scene parsing performance via reliable scale prediction. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 492–500, 2018.

[20] Hengcan Shi, Hongliang Li, Qingbo Wu, and King Ngi Ngan. Query reconstruction network for referring expression image segmentation. *IEEE Transactions on Multimedia*, 23:995–1007, 2020.

[21] Hengcan Shi, Hongliang Li, Qingbo Wu, and Zichen Song. Scene parsing via integrated classification model and variance-based regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5307–5316, 2019.

[22] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

[23] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San, and Ngai-Man Cheung. Multimodal multi-stream deep learning for egocentric activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 24–31, 2016.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Han-Jia Ye, Yatao Bian, De-Chuan Zhan, and Peilin Zhao. Beef: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *The Eleventh International Conference on Learning Representations*, 2022.

[26] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning, July 2022.

[27] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[28] Yabin Wang, Zhiheng Ma, Zhiwu Huang, Yaowei Wang, Zhou Su, and Xiaopeng Hong. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10209–10217, 2023.

[29] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[30] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.

[31] Ziyang Wu, Christina Baek, Chong You, and Yi Ma. Incremental learning via rate reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1133, 2021.

[32] Ye Xiang, Ying Fu, Pan Ji, and Hua Huang. Incremental learning using conditional adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6619–6628, 2019.

[33] Linfeng Xu, Qingbo Wu, Lili Pan, Fanman Meng, Hongliang Li, Chiyuan He, Hanxin Wang, Shaoxu Cheng, and Yu Dai. Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning. *IEEE Transactions on Multimedia*, pages 1–15, 2023.

[34] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021.

[35] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

[36] Haibin Yu, Guoxiong Pan, Mian Pan, Chong Li, Wenyan Jia, Li Zhang, and Mingui Sun. A hierarchical deep fusion framework for egocentric activity recognition using a wearable hybrid sensor system. *Sensors*, 19(3):546, 2019.

[37] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6982–6991, 2020.

[38] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[40] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020.

[41] Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. Pycil: A python toolbox for class-incremental learning, 2023.

[42] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Co-transport for class-incremental learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 1645–1654, New York, NY, USA, Oct. 2021. Association for Computing Machinery.

[43] Fei Zhu, Zhen Cheng, Xu-yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021.

[44] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.

[45] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2022.