Few-Shot Incremental Multi-modal Learning via Touch Guidance and Imaginary Vision Synthesis

Lina Wei[1,2], **Yuhang Ma**[1], **Zhongsheng Lin**[1], **Fangfang Wang**[3], **Canghong Jin**[1,2], **Hanbin Zhao**[4,*], **Dapeng Chen**[5]

[1]School of Computer Science and Computing, Hangzhou City University
[2]Zhejiang Provincial Engineering Research Center for Real-Time SmartTech in Urban Security Governance
[3]School of Information Science and Technology, Hangzhou Normal University
[4]College of Computer Science and Technology, Zhejiang University
[5]School of Automation, Nanjing University of Information Science and Technology
weiln@hzcu.edu.cn, monkeyc4281@163.com, 32201298@stu.hzcu.edu.cn, wangff@hznu.edu.cn, jinch@hzcu.edu.cn, zhaohanbin@zju.edu.cn, dpchen@nuist.edu.cn

Abstract

Multimodal perception, which integrates vision and touch, is increasingly demonstrating its significance in domains such as embodied intelligence and human-computer interaction. However, in open-world scenarios, multimodal data streams face significant challenges, including catastrophic forgetting and overfitting, during few-shot class incremental learning (FSCIL), leading to a severe degradation in model performance. In this work, we propose a novel approach named Few-Shot Incremental Multi-modal Learning via Touch Guidance and Imaginary Vision Synthesis (TIFS). Our method leverages vision imagination synthesis to enhance the semantic understanding and integrates touch and vision fusion to improve the problem of modal imbalance. Specifically, we introduce a framework that employs touch-guided vision information for cross-modal contrastive learning to address the challenges of few-shot learning. Additionally, we incorporate multiple learning mechanisms, including regularization, memory mechanisms, and attention mechanisms, to mitigate catastrophic forgetting during multi-incremental step learning. Experimental results on the Touch and Go and VisGel datasets demonstrate that the TIFS framework exhibits robust continuous learning capabilities and strong generalization performance in touch-vision few-shot incremental learning tasks. Our code is available at https://github.com/Vision-Multimodal-Lab-HZCU/TIFS.

1 Introduction

In recent years, multi-modal learning has achieved significant advancements in various practical applications. However, in open-world scenarios where multi-modal data arrives as a continuous stream, there is an urgent need for machine learning paradigms that can incrementally acquire knowledge of new categories without forgetting previously learned information, specifically multi-modal class-incremental learning (CIL) [Yu *et al.*, 2024a]. CIL faces two primary challenges. First, "catastrophic forgetting" occurs when models lose previously acquired knowledge while learning new information, leading to performance degradation on old tasks. Second, ensuring high-quality data and efficient fusion of multi-modal data is challenging. For example, due to dataset limitations or the requirement for multiple incremental training steps, models may encounter insufficient samples for new categories, resulting in fewer available samples at each step, known as few-shot class-incremental learning (FSCIL) [Tian *et al.*, 2024]. FSCIL faces the dual challenges of learning from limited labeled samples and mitigating catastrophic forgetting, making it particularly difficult. In addition to these challenges, the significant disparity in the number of samples between old and new categories can cause the model to bias towards the larger set of old-class training samples during both training and prediction. This imbalance between base and novel class samples further complicates the model's ability to effectively learn new categories.

To address these challenges, mainstream FSCIL frameworks leverage a variety of techniques, including transfer learning, meta-learning, data augmentation [Mumuni and Mumuni, 2022], metric learning, prototypical networks, relation networks, generative adversarial networks (GANs), self-supervised learning [Krishnan *et al.*, 2022], and memory mechanisms. Among these, data augmentation enhances the generalization ability of few-shot training models by generating additional training samples to expand the original dataset. Therefore, designing an effective CIL method has become crucial. In particular, methods that utilize data augmentation and can mitigate catastrophic forgetting have garnered significant attention from researchers. Some effective methods for addressing the few-shot problem involve using predefined transformations to generate additional samples, such as adding noise, blurring, cropping, scaling, and rotating [Song

* is the corresponding author.

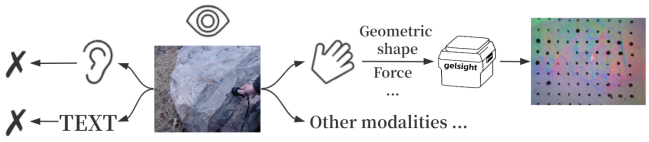


Figure 1: Touch provides supplemental information for the vision.

et al., 2023]. However, these methods face challenges in forming one-to-one corresponding sample pairs after data augmentation, particularly in touch and vision FSCIL scenarios.

Furthermore, research indicates that integrating multi-modal data enables models to learn richer semantic information, thereby alleviating the limitations imposed by few-shot scenarios and enhancing both learning effectiveness and generalization capabilities. Multi-modal data combines information from various sensory channels, such as vision, touch, and auditory inputs, allowing models to capture complex features and relationships that are difficult to achieve with a single modality. Specifically, in real-world object recognition and classification tasks, vision information provides color, shape, and other appearance features, When other modalities are unavailable, touch input offers detailed physical properties such as hardness and texture, aiding models in distinguishing objects that are challenging to differentiate based on vision cues alone, as shown in Figure 1. Consequently, touch and vision perception are essential components of physical reasoning and play significant roles in embodied intelligence, multi-modal learning, and other fields involving interaction with the environment. With the continuous increase of touch and vision data streams in open-world settings, this cross-modal information fusion enables models to comprehensively understand the essential features of objects, maintaining high recognition accuracy and generalization capabilities even with limited samples.

However, as the model learns new concepts, catastrophic forgetting remains a significant issue. This phenomenon is particularly pronounced in scenarios with limited high-quality data and during vision-touch fusion. This motivates us to explore whether it is possible to jointly rectify the model by enhancing the efficient fusion of vision and touch data to mitigate catastrophic forgetting. Based on this, we propose a novel FSCIL framework, Few-Shot Incremental Multi-modal Learning via Touch Guidance and Imaginary Vision Synthesis (TIFS), which integrates touch and vision information along with imaginative vision synthesis to provide richer semantic information. Additionally, we incorporate regularization, memory mechanisms, attention mechanisms, and other incremental learning techniques to address the challenges of FSCIL. Experiments on the Touch and Go dataset [Yang *et al.*, 2022] demonstrate that our TIFS framework achieves significant performance improvements in touch-vision few-shot class incremental learning tasks, outperforming existing methods in terms of average accuracy. Experiments on the VisGel dataset demonstrate that our TIFS framework exhibits excellent generalization capability for new datasets in touch-vision few-shot class incremental learning tasks, exceeding the average accuracy of current methods. Furthermore, we designed a new evaluation metric to assess the generalization ability of class incremental learning models for new categories. Our contributions can be summarized as follows:

- We propose an innovative continuous contrastive learning framework that integrates touch guidance and imagination-based vision synthesis (IVS), specifically designed to address the catastrophic forgetting problem and modality imbalance issue in few-shot category incremental learning (FSCIL).
- The proposed method combines attention retention mechanisms and memory augmentation with touch-vision cross-modal contrastive learning. This ensures semantic alignment between touch and vision features during incremental steps, thereby helping the model retain its previously acquired attention capabilities and reducing the risk of forgetting learned touch-vision relationships in future tasks or categories.
- Experimental results on the Touch and Go and VisGel datasets demonstrate that our approach significantly outperforms state-of-the-art multi-model FSCIL methods.

2 Related Work

2.1 Few-shot Class Incremental Learning

Class-Incremental Learning (CIL) is a technique aimed at solving the problem of "catastrophic forgetting" that models may encounter when learning new categories. In CIL, regularization parameters, knowledge distillation, and dynamic architectures play crucial roles. Regularization parameter methods [Liu *et al.*, 2022] effectively update model parameters by assigning appropriate weights based on their importance during the incremental learning process. Knowledge distillation techniques [Pan *et al.*, 2024; Zheng *et al.*, 2024] ensure consistency by extracting information from previous learning phases and minimizing discrepancies between representations produced by the model or output probability distributions. Exemplar/memory replay methods [Kong *et al.*, 2024; Zhang *et al.*, 2024b] preserve past learning by storing samples of old tasks or classes in memory. Furthermore, dynamic architecture approaches enhance the model's capability to manage new categories by incorporating incremental modules, thereby reducing the computational burden associated with continually adding new components. Currently, incremental learning has been widely applied in various domains, including image classification [Qazi *et al.*, 2024], action recognition [Wei *et al.*, 2024; Nawal *et al.*, 2023], semantic segmentation [Cermelli *et al.*, 2023; Zhu *et al.*, 2023], object detection [Zhang *et al.*, 2024a; Deng *et al.*, 2024], and language or joint vision-language tasks [Yu *et al.*, 2024b; Shi *et al.*, 2024]. Furthermore, it has found applications in self-supervised representation learning and pre-training [Magistri *et al.*, 2024].

Due to limitations in the dataset or training requirements, incremental models may encounter challenges such as having a limited number of samples for new categories or significant data heterogeneity among samples from old categories. This scenario is referred to as few-shot class incremental learning (FSCIL) [Tao *et al.*, 2020; Ganea *et al.*, 2021;

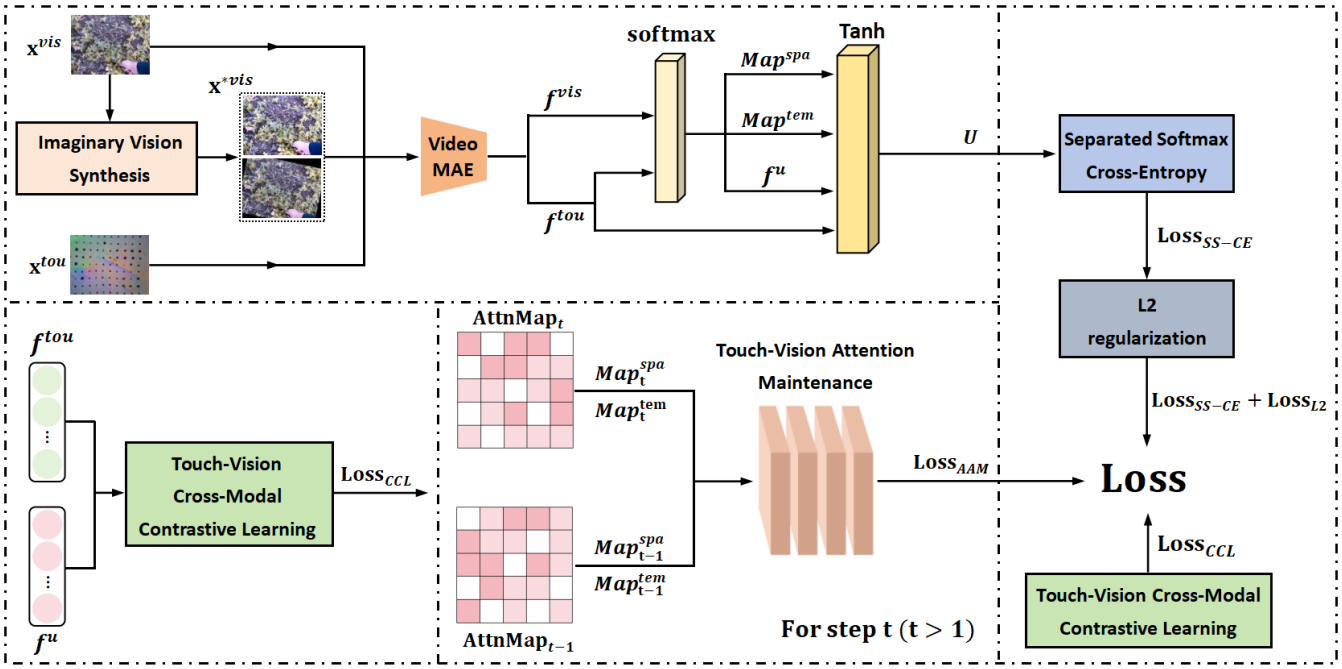


Figure 2: Overview of our proposed TIFS, which consists of four main components: Imaginary Vision Synthesis (IVS), Touch-Vision Cross-Modal Contrastive Learning (TVCC), Touch-Vision Attention Maintenance (TVAM).

Perez-Rua *et al.*, 2020; Seo *et al.*, 2023]. To solve these problems, Weight imprinting [Qi *et al.*, 2018] allows the model to have good classification performance on new categories by directly setting the final layer weights of the ConvNet classifier for new low-sample categories. Dynamic few-shot learning [Gidaris and Komodakis, 2018] relies on the use of an attention-based generator for generating weights for few-shot classification, allowing the model to recognize new categories and basic categories in a unified manner. Similarly, attention attractor network [Ren *et al.*, 2019] uses a regulator based on the attention attractor network to ensure the model's learning ability for few samples. Based on the pre-trained model, task-adaptive representation [Yoon *et al.*, 2020] uses three basic modules of meta-training to extract new features that the pre-trained backbone network cannot capture, and then combines these new features with the basic features captured by the backbone network. By expanding it into an end-to-end counterpart, synthesizing few-shot classifiers [Ye *et al.*, 2021] can simultaneously learn and compare few-shot and multi-shot classifiers, and observe the confidence calibration that benefits both types of classifiers. Furthermore, meta module generation [Xie *et al.*, 2019] uses meta-learning to learn a set of meta-modules in order to quickly adapt to new tasks. Although FSCIL holds significant potential in practical applications and has garnered considerable attention from researchers, the challenge of scarce high-quality data remains a formidable obstacle.

2.2 Touch-Vision based Multi-model CIL

In the field of multi-modal continuous learning (MMCL), the main methods include regularization, architecture, replay, and prompt methods. Regularization methods such as ZSCL [Zheng *et al.*, 2023] and Mod-X [Ni *et al.*, 2023] reduce catastrophic forgetting through explicit and implicit constraints, which have the advantage of being simple and easy to implement, but may face performance bottlenecks and computational overhead. Architecture methods such as MoE-Adapters [Yu *et al.*, 2024b] and CLAP [Wu *et al.*, 2023] dynamically adjust the model structure to adapt to new tasks, which is flexible and task-specific, but complex and resource-consuming. Replay methods such as IncCLIP [Yan *et al.*, 2022] use historical data replay to mitigate forgetting, which is effective and flexible, but has high storage requirements and may have privacy concerns. Prompt methods such as CPE-CLIP [D'Alessandro *et al.*, 2023] improve model adaptability by designing prompts, which have the advantages of high efficiency and maintaining pre-training knowledge, but prompt design is complex and the effect depends on specific tasks.

Touch-Vision Fusion is an important research direction in multi-modal learning, aiming to effectively combine touch and vision information to improve the recognition and classification of objects in models [Varadarajan *et al.*, 2022; Smarandache *et al.*, 2023]. Research has shown that touch information can significantly complement vision information, especially when processing objects that are difficult to distinguish by sight alone. As two different modes of perception, touch and vision each have unique advantages: vision information provides features such as the shape, color, and position of an object, while touch information provides information such as the texture, temperature, and hardness of an object. When these two kinds of information are merged, the model is able to understand the object features more

fully. Through cross-modal information fusion, the model can not only maintain a high recognition accuracy in the case of sparse samples, but also improve its generalization ability, which means that the model can still effectively recognize and classify new objects even when the training samples are insufficient. This feature is particularly important in real-world applications where data acquisition is expensive or where data is scarce. In recent years, the application of touch-vision fusion in incremental class learning has gradually attracted attention, which aims to make models keep good memory of old classes while constantly receiving new ones. Relevant studies have shown that the method of integrating touch and vision can significantly improve the learning effect and stability of the model, reduce the forgetting phenomenon caused by the introduction of new categories, and accelerate the model's adaptability to new categories.

3 Method

3.1 Problem Formulation

In the CIL process, we need to divide all categories equally into several preset incremental steps. In the incremental step t, the data that the model needs to learn includes n_t pairs of touch sample x^{tou} and vision sample x^{vis}, which can be denoted as $\mathcal{D}_t^{tra} = \{(x_{t,i}^{\text{tou}}, x_{t,i}^{vis}, c_{t,i})\}_{i=1}^{n_t}$, where $c_{t,i}$ is the category of the i^{th} sample pair. Our objective is to train a model $\mathcal{F}_{Object_t}$ on the partitioned incremental training set, which is parameterized by Obj_t. In the incremental step t, the training process can be described as follows:

$$Obj_t = \underset{Obj_{t-1}}{\operatorname{argmin}}[\mathbf{Loss}(\mathcal{F}_{Obj_{t-1}}(x^{tou}, x^{vis}), c)], \quad (1)$$

where $\mathbf{Loss}$ represents the loss function that compares the output of the model with the actual categories.

To alleviate the catastrophic forgetting problem in CIL, we introduce a memory mechanism. We set up a fixed-size memory buffer $\mathcal{D}_t^{mem}$ that can store data from previous incremental steps. In this way, the model can also review old categories while learning new ones, in order to consolidate the knowledge that has been learned.

3.2 Imaginary Vision Synthesis

Figure 2 illustrates our proposed method. To address the issue of insufficient learnable training samples in few-shot incremental learning, we conducted Imaginary Vision Synthesis (IVS) on the fewest several categories in the training set, enabling the augmented vision samples to continue to be combined with touch ones to form sample pairs, aiming to enhance the model's ability to learn rich semantic information from limited data. Figure 3 shows an example of our Imaginary Vision Synthesis for an instance object, which is generated through random color dithering, cropping, and rotation of x^{*vis}. Since the touch information collected by the touch sensor is independent of vision color or viewing angle, we do not process the touch data, and the vision information of Imaginary Vision Synthesis still corresponds to the original touch information. In the incremental step t, the vision samples generated by IVS are combined with the touch samples to form $\mathcal{D}_t^{ima}$, and the data used for training can finally be updated as $\mathcal{D}_t = \mathcal{D}_t^{tra} \cup \mathcal{D}_t^{ima} \cup \mathcal{D}_t^{mem}$.

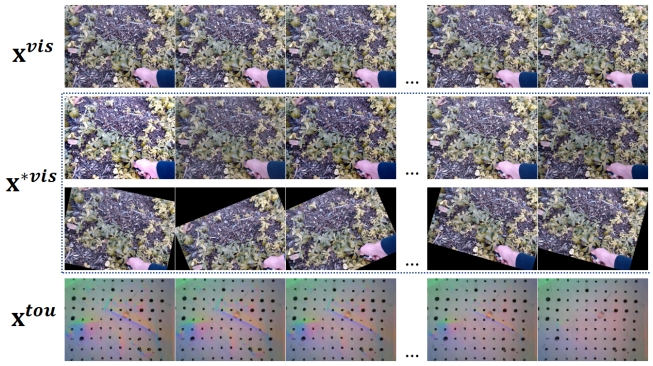


Figure 3: An example of our Imaginary Vision Synthesis for an instance object.

3.3 Touch Guided Features Fusion

In the Touch and Go and VisGel datasets, vision information are videos of real objects captured by a camera, while touch information are videos of thin film images collected and converted by the GelSight sensor. The GelSight sensor can convert subtle deformations of the contact surface into high-resolution vision images, allowing touch data to be processed and analyzed in the form of images. Based on this, we use the self-supervised pre-training model VideoMae [Tong *et al.*, 2022] to extract the touch features f^{tou} of touch samples x^{tou} and vision features f^{vis} of vision samples x^{vis}. Note that, the extraction strategy of VideoMAE involves using 16 frames uniformly distributed throughout a video as the representation for that video.

Then we use an attention interaction mechanism to make the model adaptively learn the similarities between touch and vision features. For each pair of touch and vision video frames, calculate their spatial attention maps:

$$\boldsymbol{Map}_i^{spa} = Softmax(W_i^{tou} \odot W_i^{vis}), \quad (2)$$

where W_i^{tou} and W_i^{vis} are the mappings of the touch feature f_i^{tou} and vision feature f_i^{vis} of the i^{th} frame processed by the nonlinear activation unit, respectively. And $\odot$ is the Hadamard product. Next weighted sum the time attention maps:

$$\boldsymbol{WS}_i = \sum(\boldsymbol{Map}_i^{spa} \odot \boldsymbol{weight}_i), \quad (3)$$

where $\boldsymbol{weight}_i$ is is the feature weight of the i^{th} frame. Based on this, calculate the spatial attention maps:

$$\boldsymbol{Map}_i^{tem} = Softmax(\boldsymbol{WS}_1, \ldots, \boldsymbol{WS}_i). \quad (4)$$

Next, combine the calculated time and spatial attention maps:

$$\boldsymbol{f}^u = \sum_{i=1}^{16} \boldsymbol{Map}_i^{tem} \odot \sum(\boldsymbol{f}_i^{vis} \odot \boldsymbol{Map}_i^{spa}). \quad (5)$$

Finally, we calculate the fusion features:

$$\boldsymbol{U} = \mathbf{Tanh}(\boldsymbol{f}^{tou}\boldsymbol{Map}^{tou}) + \mathbf{Tanh}(\boldsymbol{f}^{u}\boldsymbol{Map}^{vis}). \quad (6)$$

methods	Acc_1	Acc_2	Acc_3	Acc_4	Acc_5	Acc_6	Acc_7	Acc_8	**ave.Acc**	**ADR**
Fine_tuning-touch	72.59	58.47	47.61	43.14	40.13	36.49	32.92	29.85	45.15	11.80
Fine_tuning-vision	87.63	69.76	54.55	50.15	43.98	39.45	36.59	31.02	51.64	13.62
LwF-touch	79.65	63.47	58.75	46.72	42.31	39.13	38.12	36.49	50.58	10.29
LwF-vision	89.03	76.83	62.31	57.25	50.96	46.34	41.12	34.49	57.29	12.59
iCaRL-touch	81.26	66.15	52.21	48.63	41.46	39.13	38.31	36.18	50.42	10.65
iCaRL-vision	91.24	78.42	65.01	56.74	51.98	45.48	40.11	36.55	58.19	12.21
SSIL-touch	83.84	64.15	52.21	49.21	45.46	42.68	40.74	38.97	52.16	**10.07**
SSIL-vision	94.73	82.75	69.02	57.96	49.14	46.47	43.42	40.86	60.54	11.20
AFC-touch	80.95	68.81	61.13	52.79	46.96	43.49	40.78	37.71	54.08	10.28
AFC-vision	94.06	81.39	66.09	58.73	52.32	47.48	44.22	41.12	**60.76**	11.06
ours	93.45	81.39	68.83	64.67	57.17	53.37	50.93	46.92	**64.59**	**9.30**

Table 1: Accuracy (%) and Average Drop Rate (%) for different methods across incremental steps. Compared to the baseline, our method achieved the highest average accuracy of 64.59 and the lowest average Drop Rate of 9.31.

3.4 Touch-Vision Cross-Modal Contrastive Learning

In the incremental step t, the model needs to learn several new data of different categories. In order to make the model better generalize to new categories, we introduce the Touch-Vision Cross-Modal Contrastive Learning (TVCC). Our strategy is to maximize the similarity between features of different categories to facilitate category separation:

$$\mathbb{I}(i,j) = \begin{cases} 1, & \text{if } c_i = c_j \\ 0, & \text{if } c_i \neq c_j \end{cases},$$

$$\mathcal{D}_{t,i}^{Class} = \mathcal{D}_t \left[\log \frac{\sum_{j=1}^{N} e^{\boldsymbol{f}_{t,i}^{tou} \boldsymbol{f}_{t,j}^{vis\,\mathrm{T}}/\tau} \cdot \mathbb{I}(i,j)}{\left(\sum_{j=1}^{N} e^{\boldsymbol{f}_{t,i}^{tou} \boldsymbol{f}_{t,j}^{vis\,\mathrm{T}}/\tau}\right)\left(\sum_{j=1}^{N} \mathbb{I}(i,j)\right)} \right],$$

$$\mathbf{Loss}_{Class} = -\mathbb{E}_{(\boldsymbol{x}_i^{\text{tou}}, \boldsymbol{x}_i^{vis}, y_i) \sim \mathcal{D}_{t,i}^{Class}}, \tag{7}$$

where τ is a temperature hyperparameter. In FSCIL, due to the limited number of training samples in the same category, the model may have difficulty fully capturing the differences between different samples in this category. Therefore, we also make the model conduct contrastive learning on different samples within the same category, in order to enhance the model's understanding of the diversity of features in that category:

$$\mathcal{D}_{t,i}^{Sample} = \mathcal{D}_t \left[\log \frac{e^{\boldsymbol{f}_{t,i}^{tou} \boldsymbol{f}_{t,i}^{u\,\mathrm{T}}/\tau}}{\sum_{j=1}^{N} e^{\boldsymbol{f}_{t,i}^{tac} \boldsymbol{f}_{t,i}^{u\,\mathrm{T}}/\tau}} \right],$$

$$\mathbf{Loss}_{Sample} = -\mathbb{E}_{(\boldsymbol{x}_i^{tou}, \boldsymbol{x}_i^{vis}) \sim \mathcal{D}_{t,i}^{Sample}}. \tag{8}$$

Finally, the total loss of this part can be expressed as:

$$\mathbf{Loss}_{CCL} = \lambda_{Class} \mathbf{Loss}_{Class} + \lambda_{Sample} \mathbf{Loss}_{Sample}, \tag{9}$$

where λ_{Class} and λ_{Sample} are the hyperparameters.

3.5 Touch-Vision Attention Maintenance

In the process of incremental learning, the model not only forgets the knowledge of old categories it has already learned, but also may forget the cross-modal attention capabilities it has already established, leading to a reduction in the learning effect of fusion features. To alleviate this problem, we introduce the Touch-Vision Attention Maintenance (TVAM), enabling the model to continuously and efficiently learn the fusion features. We use the Kullback-Leibler (KL) divergence function to establish the temporal and spatial correlation between two adjacent incremental steps:

$$\mathbf{Loss}_{\text{spa}} = \mathbb{E}_{(\boldsymbol{x}^{\text{tou}}, \boldsymbol{x}^{\text{vis}}) \sim \mathcal{D}_t^{\text{mem}}} \left[KL(\boldsymbol{Map}_t^{\text{spa}} || \boldsymbol{Map}_{t-1}^{\text{spa}}) \right],$$

$$\mathbf{Loss}_{\text{tem}} = \mathbb{E}_{(\boldsymbol{x}^{\text{tou}}, \boldsymbol{x}^{\text{vis}}) \sim \mathcal{D}_t^{\text{mem}}} \left[KL(\boldsymbol{Map}_t^{\text{tem}} || \boldsymbol{Map}_{t-1}^{\text{tem}}) \right], \tag{10}$$

when $t > 1$. Finally, the total loss of this part can be expressed as:

$$\mathbf{Loss}_{AAM} = \lambda_{AAM} \mathbf{Loss}_{spa} + (1 - \lambda_{AAM}) \mathbf{Loss}_{tem}, \tag{11}$$

where λ_{AAM} is a hyperparameter.

3.6 Regularized Final Loss Function

When the model learns new categories, we use the Softmax Cross-Entropy (SS-CE) loss function to optimize it, to improve its generalization ability for new categories while reducing the penalty for prediction errors for old categories. Following the method in [Ahn *et al.*, 2020], we use $\boldsymbol{U}_t$ and $\boldsymbol{U}_{t-1}$ to calculate $\mathbf{Loss}_{SS-CE}$. At the same time, to avoid the model from overfitting when learning old categories, which may lead to a weakening of its generalization ability for new categories, we introduce the L2 regularization. Following the method in [Cortes *et al.*, 2009], we calculate $\mathbf{Loss}_{L2}$.

Finally, combining all parts together, our final loss function can be expressed as:

$$\mathbf{Loss} = \mathbf{Loss}_{CCL} + \mathbf{Loss}_{AAM} + \mathbf{Loss}_{SS-CE} + \mathbf{Loss}_{L2}. \tag{12}$$

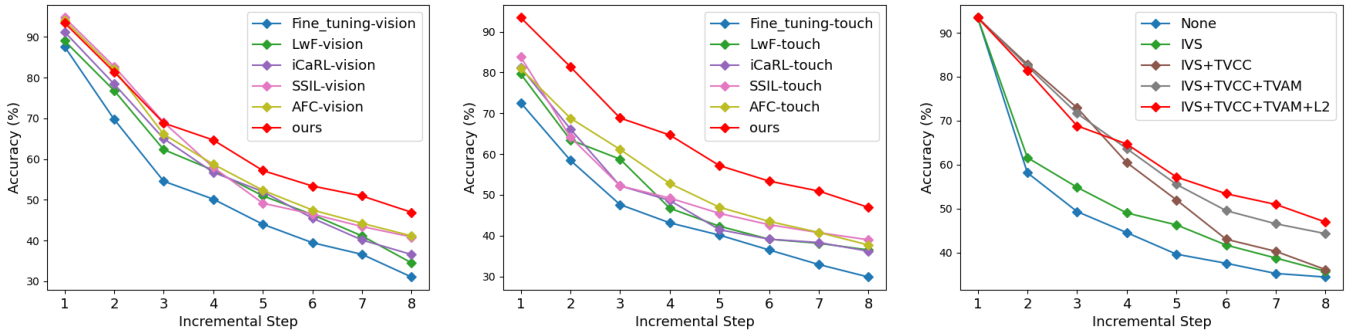
4 Experiments

4.1 Datasets

We compared our approach with state-of-the-art incremental learning frameworks using the Touch and Go and VisGel

IVS	TVCC	TVAM	L2	Acc_1	Acc_2	Acc_3	Acc_4	Acc_5	Acc_6	Acc_7	Acc_8	**ave.Acc**	**ADR**
×	×	×	×	93.45	58.14	49.31	44.52	39.61	37.54	35.22	34.41	49.03	12.49
✓	×	×	×	93.45	61.59	54.84	48.99	46.28	41.69	38.73	35.72	52.66	12.29
✓	✓	×	×	93.45	**82.83**	**72.93**	60.45	51.96	43.01	40.26	36.11	60.13	12.63
✓	✓	✓	×	93.45	82.41	71.65	63.63	55.49	49.52	46.53	44.29	63.37	10.07
✓	✓	✓	✓	**93.45**	81.39	68.83	**64.67**	**57.17**	**53.37**	**50.93**	**46.92**	**64.59**	**9.31**

Table 2: Comparison of accuracy (%) and Average Drop Rate (%) in the ablation experiment.



(a) Comparison of our method with the baseline of the vision modality.

(b) Comparison of our method with the baseline of the touch modality.

(c) Ablation Study.

Figure 4: Visualization chart of experimental results. (a) shows a visual comparison of our method and the baseline of the vision modality. (b) shows a visual comparison of our method and the baseline of the touch modality. (c) shows a visual comparison of the performance of activating different modules in our method. These results demonstrate that our approach significantly outperforms state-of-the-art methods.

datasets. The Touch and Go dataset comprises 20 distinct categories of touch and vision data, from which we selected 18 categories with significant instances, totaling 3,378 instances. These instances exhibit a long-tail distribution: the first six categories have an average of 310 instances per category, the middle six categories have an average of 168 instances per category, and the final six categories have an average of only 85 instances per category. In particular, the category with the fewest instances contains just five samples, which aligns well with the characteristics of FSCIL. In contrast, the VisGel dataset consists solely of unlabeled data. To simulate FSCIL conditions, we manually annotated six new categories not present in the Touch and Go dataset, each containing 10 instances. This manual annotation ensured that these categories represented novel classes for incremental learning purposes. Overall, our experimental dataset encompasses 24 distinct categories of touch and vision data pairs, totaling 3,438 instances, and we randomly divide the instance objects of each category into training sets, validation sets, and test sets in a ratio of 7:1.5:1.5.

4.2 Implementation Details

We compared our proposed method with the most representative and advanced incremental learning frameworks iCaRL [Rebuffi *et al.*, 2016], LwF [Li and Hoiem, 2016],SSIL [Ahn *et al.*, 2020], and AFC [Kang *et al.*, 2022]. Given the current lack of methods that simultaneously utilize touch and vision information for CIL tasks, we conducted experiments on the benchmark methods of touch and vison information separately. We also included the results of the fine-tuning. We divided 24 different categories into 8 incremental steps in ascending order, with each incremental step containing 3 categories. We trained our model on one RTX A6000 GPU and the hyperparameters τ, λ_{Sample}, λ_{Class}, λ_{AAM} are set to 0.05, 1.0, 0.5, 0.5, respectively. Simultaneously, the size of the memory buffer was set to 200.

We conducted accuracy tests on the knowledge that we have already learned at each incremental step. We also used the average accuracy metric to evaluate the overall performance of the model:

$$\mathbf{ave.Acc} = (Acc_1 + ... + Acc_8)/8, \tag{13}$$

where Acc_1 denotes testing accuracy of all seen classes after completing the training on the first incremental step. To evaluate the ability of the average model to generalize to new models, we propose a new evaluation metric, the Average Drop Rate (ADR):

$$\mathbf{ADR} = \frac{1}{7}\sum_{i=1}^{7}\frac{Acc_i - Acc_{i+1}}{Acc_i}. \tag{14}$$

The smaller the value of ADR, the slower the decline in the test accuracy of the model as the incremental steps proceed, and the better its generalization ability to new categories.

4.3 Experimental Results

In this section, we compare TIFS with other state-of-the-art (SOTA) methods across two datasets and backbone weights. The main experimental results are presented in Table 1 and Figures 4(a)(b). From the quantitative results, as shown

in Figure 4(a), which compares our method with the vision modality baseline, the accuracy of our method decreases more slowly as incremental steps increase and remains consistently higher than all other baseline methods. This demonstrates the effectiveness of the TIFS method. Similarly, as shown in Figure 4(b), when compared with the touch modality baseline, TIFS also exhibits superior performance, significantly improving accuracy in incremental learning.

Table 1 provides the Accuracy and Average Drop Rate for each algorithm at each incremental step. In the initial few incremental steps, our method did not achieve the best performance. For example, SSIL-vision outperformed our method in the first three steps, and AFC-vision had higher accuracy in the first two steps. However, as the incremental steps progressed, starting from the fourth step, our method's accuracy surpassed that of all baseline methods in every subsequent step. Particularly in the last incremental step, while the highest accuracy among the baseline methods was 41.12 for AFC, our method maintained an accuracy of 46.92. This indicates that our method has excellent generalization ability for new category data and effectively mitigates the catastrophic forgetting problem that occurs in incremental learning as the steps increase.

λ_{Sample}	1.0	0.5	1.0	0.5	1.0
λ_{Class}	1.0	1.0	0.5	0.5	0.5
λ_{AAM}	1.0	1.0	1.0	1.0	0.5
ave.Acc	62.36	61.65	63.15	62.23	**64.59**

Table 3: Comparison of average accuracy (%) for different hyperparameter combinations.

4.4 Ablation Study

In this section, we conduct an ablation study by incrementally adding each component to evaluate their effectiveness within TIFS. Table 2 and Figure 4 (c) show the experimental results, from which the following conclusions can be drawn: Imaginary Vision Synthesis (IVS) demonstrated a significant improvement in accuracy when dealing with incremental steps with a small number of samples, for example, in the fifth incremental step, the model's accuracy improved by about 6.5 percentage points after activating IVS. Meanwhile, Touch-Vision Cross-Modal Contrastive Learning (TVCC) significantly improved accuracy in each incremental step, increasing average accuracy by about 11 percentage points compared to when only IVS is activated. On this basis, when Touch-Vision Attention Maintenance (TVAM) is activated, the decline in model accuracy is significantly mitigated as the incremental learning process progresses. Finally, the introduction of L2 regularization results in a slight decrease in the accuracy of the model in the initial few incremental steps. This is because it corrects the overfitting phenomenon, but this adjustment enhances the model's generalization ability for new categories, ultimately achieving the maximum average accuracy of 64.59 and the minimum ADR of 9.31. We conducted comparative experiments with different combinations of hyperparameters and determined a set of optimal hyperparameter combinations. Due to the synergy between λ_{Sample} and λ_{Class}, we first fixed λ_{AAM} to determine the optimal combination of λ_{Sample} and λ_{Class}, and then further determine the optimal value of λ_{AAM}. However, due to space constraints, we present only the most representative combinations of the five most influential hyperparameters in Table 3, all initialized with a value of 1.0.

5 Conclusion

This paper introduces an innovative multi-modal few-shot class incremental learning framework, which named Few-Shot Incremental Multi-modal Learning via Touch Guidance and Imaginary Vision Synthesis (TIFS). By leveraging Imaginary Vision Synthesis (IVS) and touch guidance to integrate cross-modal features, the framework enhances the semantic understanding and addresses the challenge of insufficient new learnable samples in few-shot class incremental learning. Through Touch-Vision Cross-Modal Contrastive Learning (TVCC) and L2 regularization, it improves the generalization capability for new categories of the model. Additionally, by combining Touch-Vision Attention Maintenance (TVAM) with a memory mechanism, we introduce a separation cross-entropy loss to mitigate the catastrophic forgetting problem in few-shot incremental learning. Experimental results on the Touch and Go and VisGel datasets demonstrate that the TIFS framework significantly outperforms existing multi-model FSCIL methods. The effectiveness of each module has been validated, providing a robust new approach to addressing the challenges of few-shot incremental learning.

Acknowledgments

This work is supported by National Key R&D Program of China (2022ZD0119100), National Natural Science Foundation of China (Grant No: 62476238 , 62202436, 62402430, 62473200), the Natural Science Foundation of Zhejiang Province of China under Grant(No. LY24F020012, LQN25F020008, LHZSD24F020001), and Aeronautical Science Foundation of China 20240048076001.

References

[Ahn *et al.*, 2020] Hongjoon Ahn, Jihwan Kwak, Su Fang Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 824–833, 2020.

[Cermelli *et al.*, 2023] Fabio Cermelli, Matthieu Cord, and Arthur Douillard. Comformer: Continual learning in semantic and panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3010–3020, 2023.

[Cortes *et al.*, 2009] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 regularization for learning kernels. *AUAI Press*, 2009.

[D'Alessandro *et al.*, 2023] Marco D'Alessandro, Alberto Alonso, Enrique Calabrés, and Mikel Galar. Multimodal parameter-efficient few-shot class incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3393–3403, 2023.

[Deng *et al.*, 2024] Jieren Deng, Haojian Zhang, Kun Ding, Jianhua Hu, Xingxuan Zhang, and Yunkuan Wang. Zero-shot generalizable incremental learning for vision-language object detection. *arXiv preprint arXiv:2403.01680*, 2024.

[Ganea *et al.*, 2021] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1185–1194, 2021.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[Kang *et al.*, 2022] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16050–16059, 2022.

[Kong *et al.*, 2024] Jiangtao Kong, Zhenyu Zong, Tianyi Zhou, and Huajie Shao. Yooop: You only optimize one prototype per class for non-exemplar incremental learning, 2024.

[Krishnan *et al.*, 2022] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

[Li and Hoiem, 2016] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2935–2947, 2016.

[Liu *et al.*, 2022] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[Magistri *et al.*, 2024] Simone Magistri, Joost van de Weijer, Andew D Bagdanov, et al. An empirical analysis of forgetting in pre-trained models with incremental low-rank updates. *arXiv preprint arXiv:2405.18069*, 2024.

[Mumuni and Mumuni, 2022] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[Nawal *et al.*, 2023] Yala Nawal, Mourad Oussalah, Belkacem Fergani, and Anthony Fleury. New incremental svm algorithms for human activity recognition in smart homes. *Journal of Ambient Intelligence and Humanized Computing*, 14(10):13433–13450, 2023.

[Ni *et al.*, 2023] Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *International Conference on Machine Learning*, pages 26129–26149. PMLR, 2023.

[Pan *et al.*, 2024] Wensheng Pan, Timin Gao, Yan Zhang, Xiawu Zheng, Yunhang Shen, Ke Li, Runze Hu, Yutao Liu, and Pingyang Dai. Semi-supervised blind image quality assessment through knowledge distillation and incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4388–4396, 2024.

[Perez-Rua *et al.*, 2020] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13846–13855, 2020.

[Qazi *et al.*, 2024] Mohammad Areeb Qazi, Anees Ur Rehman Hashmi, Santosh Sanjeev, Ibrahim Almakky, Numan Saeed, and Mohammad Yaqub. Continual learning in medical imaging from theory to practice: A survey and practical analysis. *arXiv preprint arXiv:2405.13482*, 2024.

[Qi *et al.*, 2018] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018.

[Rebuffi *et al.*, 2016] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, G. Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2016.

[Ren *et al.*, 2019] Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard Zemel. Incremental few-shot learning with attention attractor networks. *Advances in neural information processing systems*, 32, 2019.

[Seo *et al.*, 2023] Juwon Seo, Ji-Su Kang, and Gyeong-Moon Park. Lfs-gan: Lifelong few-shot image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11356–11366, 2023.

[Shi *et al.*, 2024] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

[Smarandache *et al.*, 2023] Florentin Smarandache, Jean Dezert, and Albena Tchamova. Advances and applications of dsmt for information fusion (collected works. volume 5). 2023.

[Song *et al.*, 2023] Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Liuliang Yuan, and Yonghong Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24183–24192, 2023.

[Tao *et al.*, 2020] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12183–12192, 2020.

[Tian *et al.*, 2024] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2024.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022.

[Varadarajan *et al.*, 2022] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.

[Wei *et al.*, 2024] Wei Wei, Tom De Schepper, and Kevin Mets. Benchmarking sensitivity of continual graph learning for skeleton-based action recognition. *arXiv preprint arXiv:2401.18054*, 2024.

[Wu *et al.*, 2023] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Xie *et al.*, 2019] Shudong Xie, Yiqun Li, Dongyun Lin, Tin Lay Nwe, and Sheng Dong. Meta module generation for fast few-shot incremental learning. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[Yan *et al.*, 2022] Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. Generative negative text replay for continual vision-language pretraining. In *European Conference on Computer Vision*, pages 22–38. Springer, 2022.

[Yang *et al.*, 2022] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[Ye *et al.*, 2021] Han-Jia Ye, Hexiang Hu, and De-Chuan Zhan. Learning adaptive classifiers synthesis for generalized few-shot learning. *International Journal of Computer Vision*, 129(6):1930–1953, 2021.

[Yoon *et al.*, 2020] Sung Whan Yoon, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. Xtarnet: Learning to extract task-adaptive representation for incremental few-shot learning. In *International Conference on Machine Learning*, pages 10852–10860. PMLR, 2020.

[Yu *et al.*, 2024a] Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S Yu, and Irwin King. Recent advances of multimodal continual learning: A comprehensive survey. *arXiv preprint arXiv:2410.05352*, 2024.

[Yu *et al.*, 2024b] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024.

[Zhang *et al.*, 2024a] Hongquan Zhang, Bin-Bin Gao, Yi Zeng, Xudong Tian, Xin Tan, Zhizhong Zhang, Yanyun Qu, Jun Liu, and Yuan Xie. Learning task-aware language-image representation for class-incremental object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7096–7104, 2024.

[Zhang *et al.*, 2024b] Xinni Zhang, Yankai Chen, Chenhao Ma, Yixiang Fang, and Irwin King. Influential exemplar replay for incremental learning in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9368–9376, 2024.

[Zheng *et al.*, 2023] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19125–19136, 2023.

[Zheng *et al.*, 2024] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *arXiv preprint arXiv:2406.06391*, 2024.

[Zhu *et al.*, 2023] Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3082–3092, 2023.