Generative Multi-modal Models are Good Class-Incremental Learners

Xusheng Cao[1], Haori Lu[1], Linlan Huang[1], Xialei Liu[2,1]*, Ming-Ming Cheng[2,1]
[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian
{caoxusheng, luhaori, huanglinlan}@mail.nankai.edu.cn, {xialei, cmm}@nankai.edu.cn

Abstract

In class-incremental learning (CIL) scenarios, the phenomenon of catastrophic forgetting caused by the classifier's bias towards the current task has long posed a significant challenge. It is mainly caused by the characteristic of discriminative models. With the growing popularity of the generative multi-modal models, we would explore replacing discriminative models with generative ones for CIL. However, transitioning from discriminative to generative models requires addressing two key challenges. The primary challenge lies in transferring the generated textual information into the classification of distinct categories. Additionally, it requires formulating the task of CIL within a generative framework. To this end, we propose a novel generative multi-modal model (GMM) framework for class-incremental learning. Our approach directly generates labels for images using an adapted generative model. After obtaining the detailed text, we use a text encoder to extract text features and employ feature matching to determine the most similar label as the classification prediction. In the conventional CIL settings, we achieve significantly better results in long-sequence task scenarios. Under the Few-shot CIL setting, we have improved by at least 14% accuracy over all the current state-of-the-art methods with significantly less forgetting. Our code is available at https://github.com/DoubleClass/GMM.

1. Introduction

Deep neural networks [19, 33, 56] have made remarkable strides in numerous applications, primarily owing to the vast amounts of data and computational resources at their disposal. Nonetheless, these accomplishments are predominantly contingent on having access to all the required data simultaneously for training on various tasks. In cases where data is acquired incrementally, these networks often encounter the challenge of catastrophic forgetting [43]. Hence, the capacity to seamlessly incorporate new knowledge while retaining previously acquired knowledge is a highly desirable attribute for future artificial intelligence systems. Continual learning [45, 67, 77, 84] is a subject of study aimed at advancing the evolution of neural networks toward this goal.

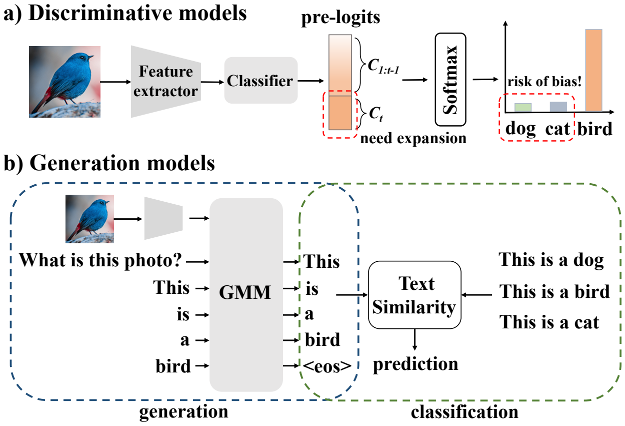


Figure 1. Illustration of conventional discriminative models for class-incremental learning (CIL) and our generative multi-modal models (GMM) for CIL. Discriminative models pose potential risk of classifier bias toward current task with network expasion. Our GMM framework consists generation and classification phases. And It is adapted to CIL based on the similarity of generated text and the true category names.

Numerous studies have delved into continual learning, categorizing their approaches into three main groups [14]: rehearsal-based, architecture-based, and regularization-based methods. Additionally, hybrid methods are gaining popularity as they combine insights from different perspectives. Within this research landscape, three primary scenarios [66] have received extensive attention, with class-incremental learning (CIL) [41] being one of the most demanding settings. In our work, we concentrate on CIL, wherein each task comprises a distinct set of classes, and the primary challenge is to enable the network to recognize new classes without forgetting knowledge of previously encountered ones.

The majority of existing research on class-incremental

*Corresposing author.

learning (CIL) focuses on training models from scratch, relying solely on data from the current tasks [7, 15, 16, 24, 26, 29, 48, 74, 76, 89]. In contrast, humans accumulate knowledge over extended periods, drawing upon a wealth of prior world knowledge. Consequently, there has been a growing interest in pre-trained models for CIL [55, 71, 72, 81, 86], harnessing the knowledge acquired from extensive, pre-existing datasets to address the immediate tasks at hand. For instance, prompt-based methods [55, 71, 72] utilize prompt-tuning to summrize task-specific knowledge from prior pre-trained knowledge. SLCA [81] and ADAM [85], on the other hand, solely fine-tuning the pre-trained model to adapt the pre-existing knowledge into the objectives of immediate tasks.

To tackle image classification downstream tasks, pre-trained models are traditionally derived from discriminative tasks, like supervised learning on datasets such as ImageNet-21K [49], or they may stem from self-supervised learning efforts [6, 9, 10, 22]. However, in our research, we venture into the paradigm of generative multi-modal models to address image classification tasks. Generative models like GPT4 [44] and LLaVa [32] have garnered significant attention in recent years due to their capacity to produce highly informative descriptions of input images. On the one hand, it can harness the wealth of semantic correspondences between texts and images, while on the other hand, there's no requirement to expand the classifier with each new task, unlike in the case of discriminative models for CIL.

Nonetheless, harnessing the knowledge from pre-trained generative models for downstream class-incremental learning (CIL) tasks presents a nontrivial endeavor. The primary challenge lies in transferring the generated textual information into the classification of distinct categories. Additionally, there's the task of formulating CIL within a generative framework, which poses a second significant challenge. Shao et al. [52] presents the VAG system, which formulates CIL as a continual label generation problem, preserving the language model's ability to learn new classes. However, it only works in the field of Natural Language Processing (NLP), which is inherently suited to Large Language Models (LLM). To the best of our knowledge, we are the first to apply this generative approach to incremental learning in the field of image classification.

In this work, we propose Generative Multi-modal Models (GMM) for class-incremental learning. As illustrated in Fig. 1 (a), conventional discriminative methods extract image features with a network backbone, then forward them to a classifier to obtain the probability of the image belonging to each label, with the label having the highest probability being the output of the discriminative models. While in Fig. 1 (b), we adopt a generative approach, which directly produces a descriptive sentence for the given image, which is then compared with the actual label texts with a text encoder. The most similar label becomes the predicted result of our generative model. This approach allows us to leverage the rich pre-training knowledge in generative multi-modal models while avoiding the use of the expanded classification head, which mitigates the risk of the model bias towards the current task and reduces catastrophic forgetting.

The main contributions of this paper are:

- We propose a novel generative approach (GMM) to address class-incremental learning by leveraging multi-modal models.
- We reformulate GMM for image classification and adapt it for the downstream benchmarks. Without an expanded classification head like in discriminative models, our model significantly mitigates the issue of bias towards current tasks, resulting in significantly reduced forgetting in CIL.
- Our model achieves state-of-the-art performance across multiple datasets in both conventional and few-shot CIL settings.

2. Related Work

2.1. Class-Incremental Learning

In Class-Incremental Learning, tasks arrive sequentially, and each class is exclusive to a specific task without any overlap. The goal is to acquire knowledge from new classes while preserving information from previously encountered classes. There are three primary branches in CIL [14], including rehearsal-based, architecture-based, and regularization-based methods. Rehearsal-based methods [1, 7, 48, 75] store a small set of data derived from old classes to represent knowledge from previous tasks. These exemplar data can be either original data [48], generative data [18, 54] or hidden features [20]. Architecture-based methods focus on modifying network architecture to alleviate forgetting. Approaches include learning redundant network architecture [17, 47], learning different expert networks [3, 50] or parameters [38, 40, 51] for each task, dynamically expanding network parameters to accumulate incremental knowledge [76]. Regularization-based methods introduce an additional regularization term to restrict network updates when adapting to new tasks. In such cases, EWC [26], SDC [78] and Rotated-EWC [35] expect that parameters essential for the old tasks should not be updated excessively. Moreover, from the perspective of network output consistency, numerous studies [25, 30, 36, 61, 80] incorporate distillation to prevent forgetting.

Few-shot CIL Few-shot Class-Incremental Learning (FSCIL) [42, 62] explores few-shot learning in an incremental context, with all data samples available for base session and very limited data in each incremental session. Some FSCIL methods [11, 62, 83] train the model in both base and incremental sessions, aiming to mitigate overfitting challenges caused by the limited data in incremental learning. Other strategies [53, 79, 90] primarily train the model in the base session and make minimal adjustments in the incremental sessions, thereby reducing forgetting but may come at the cost of decreased precision in the incremental sessions.

2.2. Pre-trained models for CIL

There are many methods [63, 72, 73] having shown that pre-trained models are effective for continual learning. One main branch trains a set of prompts to retain previous knowledge [55, 68, 71, 72]. A selected subset of prompts are fed into the model during forward to prompt model the past knowledge. Additionally, methods like SLCA [81] and ADAM [85] fine-tune pre-trained models, achieving impressive results with less forgetting. Continual-CLIP [63] demonstrates that the CLIP [46] model is capable of performing continual learning without any extra training. This highlights the significant potential of multi-modal pre-trained models in the realm of continual learning. Inspired by this, many methods [37, 86] employ CLIP as the backbone to utilize the multi-modal information. However, if not using a classifier directly, these approaches need to utilize expanded text features to calculate distances with image features for classification. This will exacerbate the model's bias towards current data, consequently leading to the forgetting of previously acquired knowledge. To avoid this bias, we use a generative model to directly generate prediction text. The fixed text decoder will function as the classifier, significantly alleviating the bias.

2.3. Vision Language Models

In recent years, vision-language multi-modal models have made significant progress and achieved impressive results in various downstream tasks [5, 28, 34, 70]. Traditional vision-language models employ different types of encoders to extract information from vision and language models, including single-stream [57], dual-stream [39] and fusion [60] encoders. A key aspect of vision-language models is the alignment of multi-modal features. CLIP [46], for example, extracts image and text features separately using respective encoders and enforces alignment through a contrastive loss, ensuring alignment between positive image-text pairs in the feature space. VisualGPT [8] and Frozen [65] leverage pre-trained models as encoders for visual-language tasks. From then on, the utilization of pre-trained models in vision-language tasks became more and more popular. For instance, Flamingo [2] and BLIP-2 [27] align the pre-trained image and text encoders employing gated cross-attention and Q-Former, respectively. Furthermore, LLaVA [32] and MiniGPT-4 [88] leverage more robust Large Language Models (LLM) [13, 64] as text encoders, while only training a projection layer for alignment. With the increasing popularity of LLM, an increasing number of studies [4, 69, 87] explore the potential of multi-modal LLMs for vision-language tasks.

3. Method

In this section, we introduce the preliminaries of class-incremental learning and generative multi-modal models. Then, we present our approach to leverage generative models for CIL and the corresponding learning process.

3.1. Preliminaries

Class-Incremental Learning. Given N tasks $T = \{T_1, T_2, ..., T_N\}$, the goal of class-incremental learning is to learn each task T_t with its associated data $\{\mathbf{X}_t, \mathbf{Y}_t\}$ in a sequential order. For each task, it contains samples $\{\mathbf{x}_i, \mathbf{y}_i\}, i = 1, ..., n_t$, where $\mathbf{x}_i$ is the images and $\mathbf{y}_i$ is the corresponding one-hot labels. Typically, $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset, \forall i \neq j$. At inference, the model is tested on all seen tasks without task IDs. In some scenarios, fixed memory storage is set to keep a few samples of previous tasks to prevent forgetting.

Normally, a CIL model consists of a feature extractor and a classifier head $F = \{f_\theta, \mathcal{H}_\phi\}$ which are parameterized by $\{\theta, \phi\}$. In traditional class-incremental learning, θ is usually a modified ResNet [21] with all parameters tunable. In pre-trained or prompt-based methods, θ represents fewer trainable parameters like a linear adaptor or a couple of prompts. ϕ is a linear classifier head projecting image features to probability predictions, which has to be expanded for each new task in order to make predictions for the new classes. The conventional Cross Entropy loss is often used for updating θ and ϕ, which for task t is:

$$\mathcal{L}_{\text{CE}}(\mathbf{X}_t, \mathbf{Y}_t; \theta, \phi) = -\frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{y}_i \cdot \log \mathcal{H}\left(f\left(\mathbf{x}_i; \theta\right); \phi\right). \tag{1}$$

In the continual learning process, the parameter ϕ can easily deviate to the data of the current task due to the absence or scarcity of old samples in the previous tasks, resulting in forgetting the previously acquired knowledge and deteriorating the overall performance.

Generative Multi-Modal Models (GMM). Multi-modal models have demonstrated exceptional performance in generating detailed image descriptions by incorporating both visual and textual information. Notably, GPT-4 [44] stands out as an advanced model proficient in generating comprehensive image descriptions and providing explanations for the depicted content. Furthermore, MiniGPT-4 [88] proposes a two-stage fine-tuning process that aligns image features and large language models, enabling LLaMa [64] to recognize images and conduct further dialogue based on the image content.

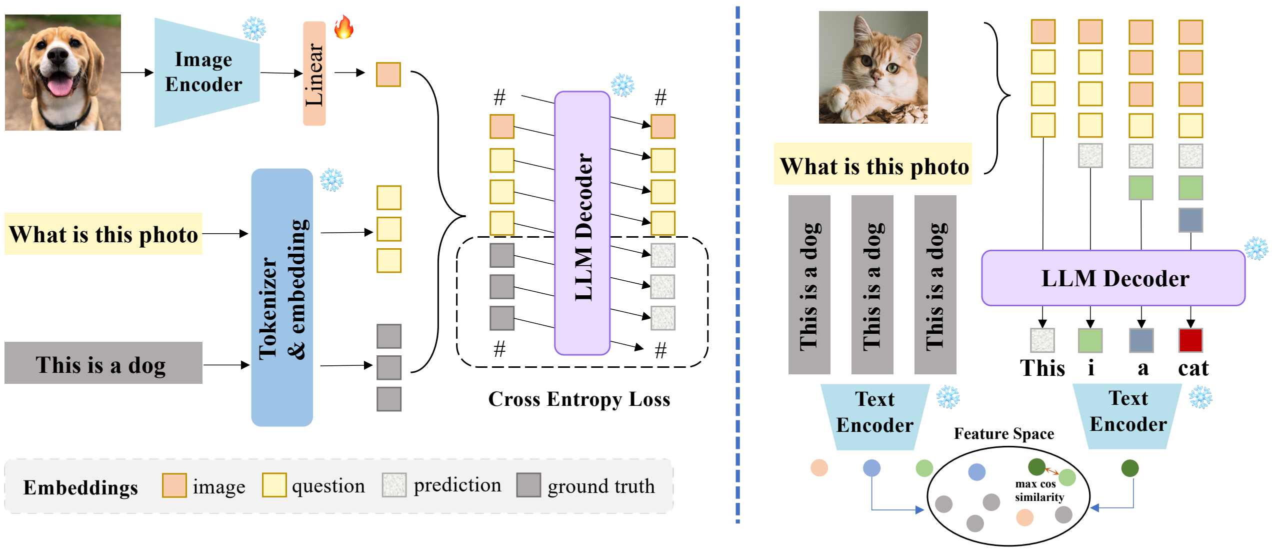


Figure 2. The overview structure of our proposed method. The conceptual illustration of the generative multi-modal model (GMM) is shown on the left. In order to adapt this model for CIL, We have to turn the GMM model for classification and further adapt to our objective benchmark for learning (see Sec. 3.2). On the right side, we demonstrate how the final evaluation is taken for CIL with all seen classes. Text encoder is used for obtaining the embeddings for similarity prediction.

As shown in Fig. 2, these models consist of an encoder f_{enc} to generate content embeddings including image embedding and text embedding, which is further used as the input of an auto-aggressive decoder f_{dec} to generate image descriptions. Input image $\mathbf{x}_i$ is encoded as an image embedding e_i, and question embeddings $\mathbf{q}$ with $q_1, ..., q_l$ can be concatenated with the image embedding together to generate answer embeddings $\mathbf{s}$ with $s_1, ..., s_m$. The output tokens are generated one by one with the condition of previously generated tokens. For instance, s_m is generated with all previous $m-1$ tokens (see the decoder in Fig. 2).

3.2. Generative Multi-Modal Models for CIL

We adhere to the foundational settings of MiniGPT-4, incorporating a frozen image encoder f_{enc} followed by a trainable projection layer for adaptation to downstream tasks, as shown in Fig. 2. Our primary innovation involves the direct utilization of generative models to produce text, which can then serve as a basis for discriminative classification. However, two major challenges need addressing. First, there is the issue of adapting generative multi-modal models for classification, given that the generated text may differ significantly from class names. Second, we must devise a mechanism for our classification benchmarks to learn in a manner consistent with generative multi-modal models. We introduce these two aspects as follows.

Turning GMM for classification. We employ a distance metric to bridge the gap between generative and discriminative models. During training, we use the real label's text to encourage the model to predict the label of an image with a concise and accurate sentence in the format of "This is a photo of [CLS]." avoiding detailed descriptions of all contents in the image. During testing, the model follows the format to output the category text for a given image. We extract the content in "[CLS]", then obtain its text features using the CLIP [46] text encoder f_{text}, and compute the distance with text features of all categories seen by now. The closest class was then considered the final prediction of the generative model.

Converting CIL Benchmarks for adaptation. CIL is usually evaluated on ImageNet, CIFAR-100, and ImageNet-R datasets. These datasets usually consist of images and corresponding one-hot labels $\{\mathbf{X}_t, \mathbf{Y}_t\}$. Using the CIFAR100 dataset as an example, we pair each image with a sentence to form an image-text pair format $\{\mathbf{X}_t, \mathbf{S}_t\}$ with the template: "This is a photo of [CLS]", where "[CLS]" is the label name of that category, such as apple, dog, etc. Next, we partitioned the 100 classes into various tasks based on different settings and fed them into the model sequentially. After completing the training on task T, the model should be capable of classifying all the classes encompassed from task 0 to task T. Note that only the linear projection layer is updated for further adaptation.

3.3. Optimization and Inference

Optimization. For each task t, we obtain the current task's image-text pair $\{\mathbf{X}_t, \mathbf{S}_t\}$, where $\mathbf{S}_t$ contains the corresponding sentence of each image. During training, we

Type	Method	Exemplar	Tiny-ImageNet						ImageNet-R
			5 tasks		10 tasks		20 tasks		10 tasks
			Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**	Last
Conventional	EWC [26]	✗	19.01	6.00	15.82	3.79	12.35	4.73	35.00
	LwF [29]	✗	22.31	7.34	17.34	4.73	12.48	4.26	38.50
	iCaRL [48]	✓	45.95	34.60	43.22	33.22	37.85	27.54	-
	EEIL [7]	✓	47.17	35.12	45.03	34.64	40.41	29.72	-
	UCIR [24]	✓	50.30	39.42	48.58	37.29	42.84	30.85	-
	PASS [89]	✗	49.54	41.64	47.19	39.27	42.01	32.93	-
	DyTox [16]	✓	55.58	47.23	52.26	42.79	46.18	36.21	-
Discriminative PT models	Continual-CLIP[63]	✗	70.49	66.43	70.55	66.43	70.51	66.43	72.00
	L2P [72]	✗	83.53	78.32	76.37	65.78	68.04	52.40	72.92
	L2P [72]	✓	80.24	72.89	80.08	72.61	79.44	70.41	59.78
	DualPrompt [71]	✗	85.15	81.01	81.38	73.73	73.45	60.16	68.82
	DualPrompt [71]	✓	79.92	72.83	79.15	73.21	80.17	71.74	57.02
	CODA-Prompt [55]	✗	**85.91**	**81.36**	82.80	75.28	77.43	66.32	73.88
	Linear Probe	✗	74.38	65.40	69.73	58.31	60.14	49.72	45.17
	Linear Probe	✓	70.10	61.11	69.35	64.19	71.64	70.50	55.72
Generative PT models	Zero-shot	✗	58.16	53.72	58.10	53.72	58.13	53.72	67.38
	GMM (Ours)	✗	83.42	76.98	82.49	76.51	81.70	76.03	80.72
	GMM (Ours)	✓	84.16	78.46	**83.95**	**78.64**	**84.23**	**79.17**	**89.41**

Table 1. Comparison results of our method with other conventional baselines and methods learned with discriminative pre-trained (PT) models on Tiny-ImageNet and ImageNet-R under the conventional CIL setting. "Avg" represents the averaged performance after training each task, and "Last" represents the performance on all test samples after training the last task.

first utilize a tokenizer to tokenize and acquire the embedding of the questions and answers. We leverage pre-trained encoder f_{enc} and the projection layer to obtain the corresponding features for the input images:

$$\mathbf{e}_i = f_{enc}(\mathbf{x}_i; \theta_{enc}). \tag{2}$$

Then, the question embedding and the ground-truth embedding of this question, e.g., "This is a photo of [CLS]", is concatenated with image embedding. The final input of LLM Decoder f_{dec} is:

$$\hat{\mathbf{e}}_i = \text{CONCATE}(bos, \mathbf{e}_i, \mathbf{q}, \mathbf{s}, eos). \tag{3}$$

bos is the symbol of the sentence beginning, and eos is the symbol for the end of the sentence. This encourages tokens at positions $m-1$ to predict token m:

$$P(\hat{s}_1, \hat{s}_2, ..., \hat{s}_m | \mathbf{x}_i, \mathbf{q}, \mathbf{s}) = \prod_{j=1}^{m-1} P(s_j | \mathbf{e}_i, \mathbf{q}, s_1, s_2, \ldots, s_{j-1}), \tag{4}$$

where s_j indicates the ground-truth answer token and $\hat{s}_m$ is the generated prediction. Then, we can compute the Cross Entropy loss as follows:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{m} \sum_{j=1}^{m} s_j \cdot \log \hat{s}_j. \tag{5}$$

Inference. During inference, we use the updated projection layer in conjunction with the pre-trained encoder to obtain image features. These image features, combined with the question embeddings are then passed to the LLM Decoder to obtain the text output.

$$pred = \text{argmax} < f_{\text{text}}(\mathbf{s}), f_{\text{text}}(\hat{\mathbf{s}}) >, \tag{6}$$

where f_{text} is the text encoder, $<, >$ is the cosine similarity used to calculate the final predictions $pred$.

4. Experiments

4.1. Experimental setups

Datasets and Baselines. We conduct experiments in both conventional CIL and Few-shot CIL scenarios. In conventional CIL, we evaluate on three datasets. CIFAR100, Tiny-ImageNet and ImageNet-R. CIFAR100 contains 60,000 images of 32x32 pixels in 100 categories. Each category has 600 images, of which 500 are for the training set, and 100 are for the test set. We experiment with two settings, B0-n and B50-n. The former splits 100 classes into n tasks,

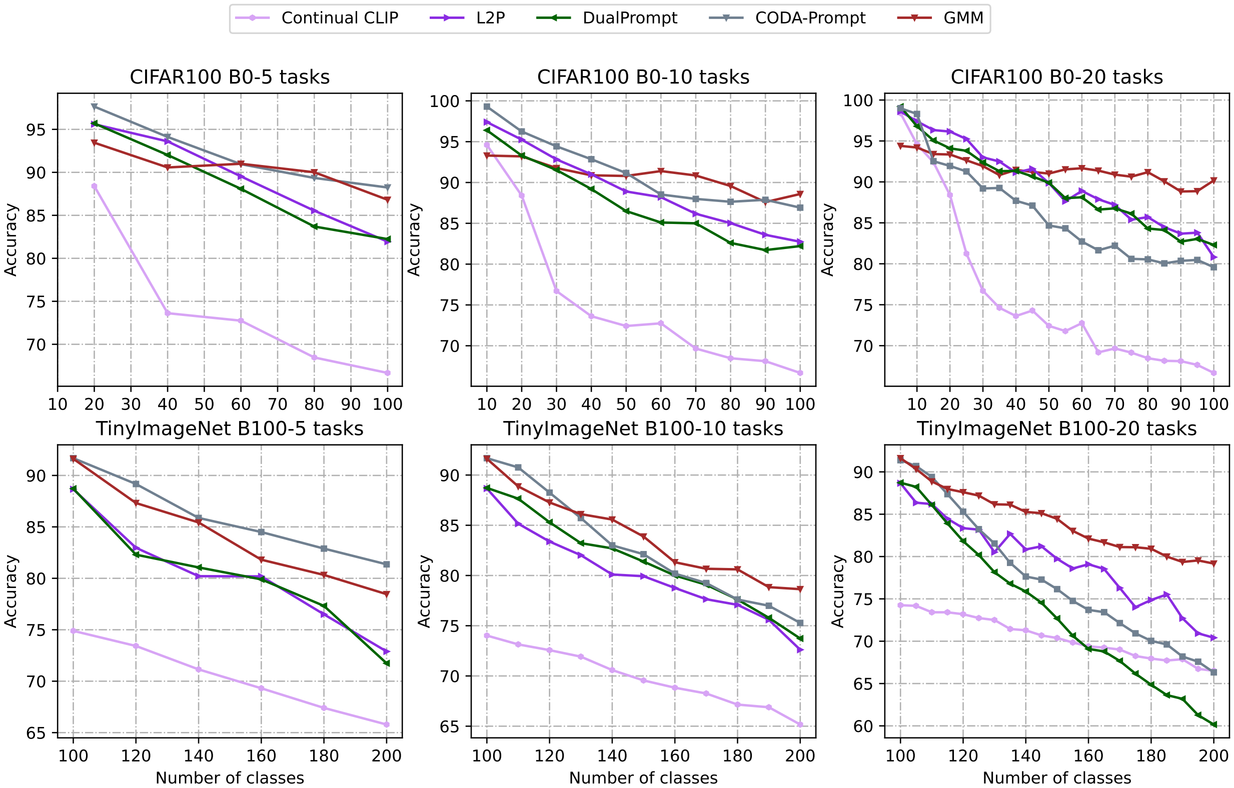


Figure 3. Comparison of our method with other SOTA baselines on CIFAR100 and Tiny-ImageNet under the conventional CIL setting.

while the latter trains on 50 classes first and then distributes the other 50 classes across 5/10 tasks.

Tiny-ImageNet contains 200 classes out of the original 1000 classes in ImageNet, with 550 images per class, of which 500 are in the training set and 50 are in the test set. The images are down-sampled to 64 × 64 pixels, which makes them easier to process and analyze. We train the first task in half 100 classes and split the other 100 classes into 5/10/20 tasks following [89].

ImageNet-R [23] contains 200 classes of images, which are included in the original ImageNet 1000 classes. However, many images are newly added and have various styles, such as sketch, painting, misc, etc. The dataset poses a great challenge for continual learning, as it has a wide diversity of image categories and styles and an uneven distribution of samples ranging from 45 to 500 per category. We follow [71] to split the dataset into 10 tasks, each containing 20 classes.

In Few-shot CIL, we use CIFAR100 and *mini*-ImageNet [49] following the split proposed by [62]. For both datasets, we partition the data into two parts: base session and incremental sessions. The base session comprises 60 classes with all data available, while the incremental session follows a 5-way 5-shot setting, meaning that each session consists of only 5 classes with 5 samples each.

In both conventional and few-shot scenarios, we compare our method with some of the current state-of-the-art approaches, including conventional methods [7, 15, 16, 24, 29, 48, 53, 74, 76, 82, 89], pre-trained and prompt-based methods [55, 71, 72], and some specifically designed methods [12, 53, 62, 79] for few-shot scenarios. Additionally, we compare with a linear probe baseline, wherein features obtained from the image encoder are connected to a classifier for classification. We also consider the Zero-shot approach, where the generated text is directly used for classification without further fine-tuning.

Implementation Details. We follow BLIP2 [27] to use the EVA-CLIP [59] pre-trained ViT-g/14 and BLIP2 pre-trained Qfomer. We also used the MiniGPT-4 pre-trained projection layer checkpoint as our initial parameters. Under the many-shot "B0" setting, we employ a learning rate of 3e-7 and use a scheduler with cosine decay. The total training process consists of 2 epochs only. In B50 or B100 settings, we first train the linear layer with a learning rate of 3e-6 on the base classes, and then on the subsequent tasks, we adopt a lower learning rate of 3e-7, both employing a cosine decay scheduler. For the few-shot setting, we use a learning rate of 3e-6 for both base task and incremental tasks. We train one epoch for the base task and two epochs for incremental tasks.

	0	1	2	3	4	5	6	7	8	PD↓
iCaRL [48]	61.31	46.32	42.94	37.63	30.49	24.00	20.89	18.80	17.21	44.10
EEIL [7]	61.31	46.58	44.00	37.29	33.14	27.12	24.10	21.57	19.58	41.73
LUCIR [24]	61.31	47.80	39.31	31.91	25.68	21.35	18.67	17.24	14.17	47.14
TOPIC [62]	61.31	50.09	45.17	41.16	37.48	35.52	32.19	29.46	24.42	36.89
CEC [79]	72.00	66.83	62.97	59.43	56.70	53.73	51.19	49.24	47.63	24.37
F2M [53]	72.05	67.47	63.16	59.70	56.71	53.77	51.11	49.21	47.84	24.21
MetaFSCIL [12]	72.04	67.94	63.77	60.29	57.58	55.16	52.90	50.79	49.19	22.85
Entropy-reg [31]	71.84	67.12	63.21	59.77	57.01	53.95	51.55	49.52	48.21	23.63
L2P* [72]	94.12	87.20	80.99	75.67	70.94	66.76	63.11	59.81	56.83	37.29
DualPrompt* [71]	93.97	86.85	80.67	75.31	70.61	66.44	62.77	59.58	56.80	37.17
CODA-Prompt* [55]	**95.37**	**88.86**	82.69	77.87	74.47	70.16	66.46	63.73	61.14	34.23
Zero-shot	58.08	58.95	57.76	57.89	58.19	57.42	56.26	54.82	54.95	**3.13**
GMM (Ours)	89.35	88.40	**86.11**	**85.07**	**83.61**	**81.35**	**78.97**	**77.34**	**75.18**	14.17

Table 2. Comparison results of our method with other SOTA baselines on *mini*-ImageNet under the few-shot CIL setting. Phase 0 is the base task with all samples available, and the following 1-8 phases are the incremental 5-way 5-shot tasks. The reported accuracy is the test result on all seen classes after each session of training. "PD" represents Performance Prop between session 0 and session 8, with lower values indicating lower forgetting. * indicates our re-implementation based on PILOT [58].

4.2. Experiments on Conventional CIL

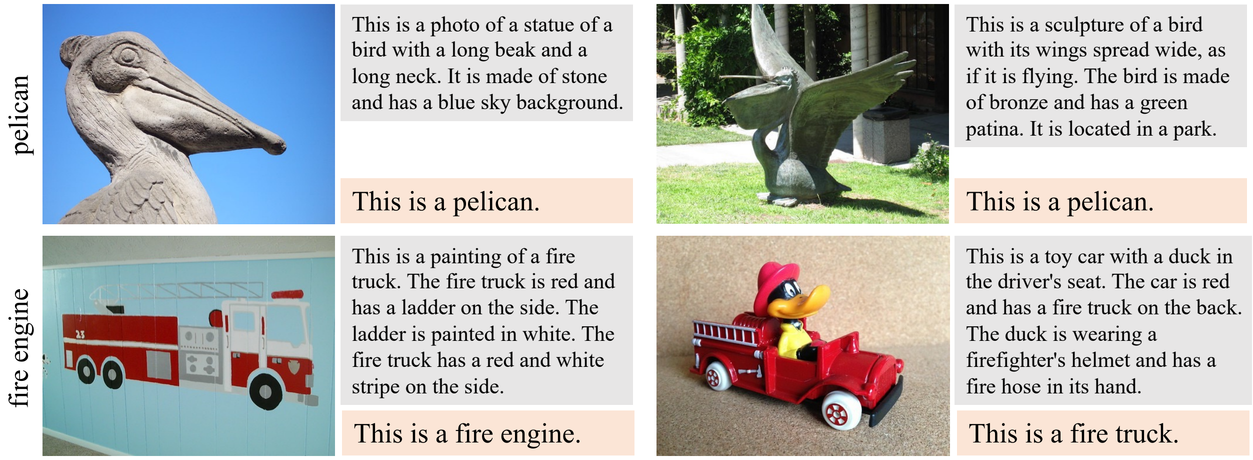
In Table 1, we can see that our method outperforms all conventional methods by a large margin, including ResNet-based method DER and ViT-based DyTox. Note that without exemplar, our performance is a bit lower than DualPrompt and CODA-Prompt at B100-5 setting. We argue that their performance is mainly due to the backbone pre-trained on ImageNet-21K, which largely overlaps with CIFAR100 and Tiny-ImageNet. Another interesting observation is that our method has better performance than all baselines under longer sequence settings (B100-10, B100-20). We believe this is because generative models do not rely on classification heads, making them less prone to bias toward the current task, resulting in less forgetting of past tasks. The Linear probe setting performs less than our method, indicating that our main contribution is not from the Large pre-trained ViT but the generation pipeline. In addition, the Zero-shot performance is superior to many traditional baselines, meaning that the Generative Multi-modal Models are indeed efficient Class-Incremental learners, but its output is less concise without fine-tuning (see Fig. 4).

In Fig. 3, we compared our approach with some pre-trained models on CIFAR100 and Tiny-ImageNet in terms of last task accuracy (all baselines are based on PILOT [58] and use 2000 exemplars). It can be observed that our method does not outperform other approaches in the initial tasks (0-2) and short sequence settings (B0-5, B100-5). This is because we do not rely on a supervised ImageNet-21K pre-trained backbone. Besides, we trained each task for only 1-2 epochs to ensure efficiency without sacrificing generalization. However, our method exhibits significant advantages in long sequences and later tasks. For instance, under the CIFAR100 B0-20 setting, we outperform CODA-Prompt by 10 points and DualPrompt by 7 points.

4.3. Experiments on Few-shot CIL

In Table 2, we compare our method with several baselines in the few-shot setting on *mini*-ImageNet. The evaluation metric is the model's accuracy across all the classes it has encountered so far. Our method outperforms conventional methods by a substantial margin in the final task, achieving an increase of more than 26%. Furthermore, we surpass the best discriminative pre-trained approach CODA-Prompt by more then 14% points. It's important to note that our accuracy in the first task (89.35) might not be as high as CODA-Prompt (95.37). However, in subsequent sessions, we consistently perform better than CODA-Prompt due to our ability to learn new tasks and retain knowledge of old tasks simultaneously.

In Table 3, our method outperforms all other baselines on the CIFAR100 dataset of the few-shot setting, achieving a remarkably lower Performance Drop (PD) of 10.06. Furthermore, the Zero-shot baseline can achieve a very low PD due to the absence of forgetting. However, its overall performance is not very satisfying, as the length and con-



This is a photo of a statue of a bird with a long beak and a long neck. It is made of stone and has a blue sky background.

This is a pelican.

This is a sculpture of a bird with its wings spread wide, as if it is flying. The bird is made of bronze and has a green patina. It is located in a park.

This is a pelican.

This is a painting of a fire truck. The fire truck is red and has a ladder on the side. The ladder is painted in white. The fire truck has a red and white stripe on the side.

This is a fire engine.

This is a toy car with a duck in the driver's seat. The car is red and has a fire truck on the back. The duck is wearing a firefighter's helmet and has a fire hose in its hand.

This is a fire truck.

Figure 4. Visual comparison examples of our method against frozen MiniGPT-4 [88] (Zero-shot). Text with a gray background is generated by MiniGPT-4 based on the image, while text with a orange background represents our method's output. The ground-truth labels are displayed on the left side of each row of images. All images displayed here are random sampled from ImageNet-R.

Method	0	4	8	PD↓
iCaRL [48]	64.10	27.93	13.73	50.37
EEIL [7]	64.10	28.96	15.85	48.25
LUCIR [24]	64.10	31.61	13.54	50.56
TOPIC [62]	64.10	40.11	29.37	34.73
CEC [79]	73.07	58.09	49.14	23.93
F2M [53]	71.45	57.76	49.35	22.06
MetaFSCIL [12]	74.50	59.48	49.97	24.53
Entropy-reg [31]	74.40	59.71	50.14	24.26
L2P* [72]	91.22	68.66	54.89	36.33
DualPrompt* [71]	91.08	68.45	54.67	36.41
CODA-Prompt* [55]	**93.55**	71.91	59.32	34.23
Zero-shot	74.13	72.59	67.93	**6.20**
GMM (Ours)	91.53	**85.65**	**81.47**	10.06

Table 3. Comparison results of our method with other SOTA baselines on CIFAR100 under the few-shot CIL setting.

tent of its output are inconsistent and unpredictable without fine-tuning.

4.4. Visualizations

In Fig. 4, we present some comparison examples of our methods against GMM without fine-tuning [88]. We can see that GMM without fine-tuning provides an intuitive description of the overall image content with varying lengths of output text. However, it tends to recognize only broad categories (e.g., biar, car) and struggles with fine-grained categorization (e.g., pelican, truck). The descriptions are sometimes somewhat repetitive (e.g., first fire engine). In contrast, our fine-tuned method accurately identifies the image's real category, even if there are occasional discrepancies with the true labels (e.g., ”fire engine” vs. ”fire truck”). Besides, with the assistance of the text encoder during the testing phase, our model can achieve correct classification results even when predicting similar but not identical text.

5. Conclusion

In this paper, we propose GMM to use generative models for class-incremental learning. By fine-tuning the Generative Multi-modal Model (GMM), we directly generate the label text of the images to be classified. Then we select the label most similar to the generated text by its features. Our experiments demonstrate that this method, which does not require a classification head, is highly effective in addressing classification biases in continual learning.

Limitations. Since we are the first to introduce generative models to class-incremental learning, the overall design of our method is embarrassingly simple. We believe that with more focused efforts in this direction, there will be significant advancements in the field of continual learning.

Broader impact. We believe that introducing GMM into continual learning (CL) is both necessary and urgent. With the rapid development of GMM, we can leverage their capabilities to improve the performance of continual learning. Besides, integrating CL methods into the training process of GMM could significantly reduce training costs.

Acknowledgments. This work is funded by NSFC (NO. 62206135, 62225604), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), and the Fundamental Research Funds for the Central Universities (Nankai Universitiy, 070-63233085). Computation is supported by the Supercomputing Center of Nankai University.

References

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *ICCV*, 2021. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2(3):4, 2023. 3

[5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[7] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018. 2, 5, 6, 7, 8

[8] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 3

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[10] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 2

[11] Ali Cheraghian, Shafin Rahman, Pengfei Fang, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2534–2543, 2021. 2

[12] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14166–14175, 2022. 6, 7, 8

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 3

[14] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *TPAMI*, 2021. 1, 2

[15] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, 2020. 2, 6

[16] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *CVPR*, 2022. 2, 5, 6

[17] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017. 2

[18] Rui Gao and Weiwei Liu. Ddgr: Continual learning with deep diffusion-based generative replay. In *ICML*, 2023. 2

[19] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 1

[20] Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *ECCV*, 2020. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 6

[24] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 2, 5, 6, 7, 8

[25] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *CVPR*, 2021. 2

[26] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 5

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 6

[28] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019. 3

[29] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2, 5, 6

[30] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2018. 2

[31] 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. 7, 8

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3

[33] Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. *Science China Information Sciences*, 66(5):151101, 2023. 1

[34] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 3

[35] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *ICPR*, 2018. 2

[36] 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

[37] Xialei Liu, Xusheng Cao, Haori Lu, Jia-wen Xiao, Andrew D Bagdanov, and Ming-Ming Cheng. Class incremental learning with pre-trained vision-language models. *arXiv preprint arXiv:2310.20348*, 2023. 3

[38] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *CVPR*, 2021. 2

[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[40] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *ECCV*, 2018. 2

[41] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022. 1

[42] Pratik Mazumder, Pravendra Singh, and Piyush Rai. Few-shot lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2337–2345, 2021. 2

[43] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 1

[44] OpenAI. Gpt-4 technical report, 2023. 2, 3

[45] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. Recent advances of continual learning in computer vision: An overview. *arXiv preprint arXiv:2109.11369*, 2021. 1

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 4

[47] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Random path selection for incremental learning. In *NIPS*, 2019. 2

[48] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. 2, 5, 6, 7, 8

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2, 6

[50] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *ICML*, 2018. 2

[51] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, 2018. 2

[52] Yijia Shao, Yiduo Guo, Dongyan Zhao, and Bing Liu. Class-incremental learning based on label generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1263–1276, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[53] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in neural information processing systems*, 34: 6747–6761, 2021. 3, 6, 7, 8

[54] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 2

[55] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023. 2, 3, 5, 6, 7, 8

[56] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[57] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 3

[58] Hai-Long Sun, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Pilot: A pre-trained model-based continual learning toolbox. *arXiv preprint arXiv:2309.07117*, 2023. 7

[59] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 6

[60] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[61] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *ECCV*, 2020. 2

[62] 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. 2, 6, 7, 8

[63] Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022. 3, 5

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[65] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[66] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *NIPS Workshops*, 2019. 1

[67] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023. 1

[68] Liyuan Wang, Jingyi Xie, Xingxing Zhang, Mingyi Huang, Hang Su, and Jun Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[69] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 3

[70] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20 (4):447–482, 2023. 3

[71] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, 2022. 2, 3, 5, 6, 7, 8

[72] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. *CVPR*, 2022. 2, 3, 5, 6, 7, 8

[73] Tz-Ying Wu, Gurumurthy Swaminathan, Zhizhong Li, Avinash Ravichandran, Nuno Vasconcelos, Rahul Bhotika, and Stefano Soatto. Class-incremental learning with strong pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2022. 3

[74] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019. 2, 6

[75] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *ICCV*, 2019. 2

[76] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *CVPR*, 2021. 2, 6

[77] Yang Yang, Zhiying Cui, Junjie Xu, Changhong Zhong, Wei-Shi Zheng, and Ruixuan Wang. Continual learning with bayesian model based on a fixed pre-trained feature extractor. *Visual Intelligence*, 1(1):5, 2023. 1

[78] 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. 2

[79] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2021. 3, 6, 7, 8

[80] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022. 2

[81] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118*, 2023. 2, 3

[82] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *CVPR*, 2020. 6

[83] Hanbin Zhao, Yongjian Fu, Mintong Kang, Qi Tian, Fei Wu, and Xi Li. Mgsvf: Multi-grained slow vs. fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[84] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Deep class-incremental learning: A survey. *arXiv preprint arXiv:2302.03648*, 2023. 1

[85] Da-Wei Zhou, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained mod-

els: Generalizability and adaptivity are all you need, 2023. 2, 3

[86] Da-Wei Zhou, Yuanhan Zhang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *arXiv preprint arXiv:2305.19270*, 2023. 2, 3

[87] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*, 2023. 3

[88] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 8

[89] 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. 2, 5, 6

[90] Kai Zhu, Yang Cao, Wei Zhai, Jie Cheng, and Zheng-Jun Zha. Self-promoted prototype refinement for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6801–6810, 2021. 3