

Self-Sustaining Representation Expansion for Non-Exemplar Class-Incremental Learning

Kai Zhu¹ Wei Zhai¹ Yang Cao^{1,3,†} Jiebo Luo² Zheng-Jun Zha¹

¹ University of Science and Technology of China ² University of Rochester

³ Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{zkzy@mail., wzhai056@mail., forrest@}ustc.edu.cn jluo@cs.rochester.edu zhazj@ustc.edu.cn

Abstract

Non-exemplar class-incremental learning is to recognize both the old and new classes when old class samples cannot be saved. It is a challenging task since representation optimization and feature retention can only be achieved under supervision from new classes. To address this problem, we propose a novel self-sustaining representation expansion scheme. Our scheme consists of a structure reorganization strategy that fuses main-branch expansion and side-branch updating to maintain the old features, and a main-branch distillation scheme to transfer the invariant knowledge. Furthermore, a prototype selection mechanism is proposed to enhance the discrimination between the old and new classes by selectively incorporating new samples into the distillation process. Extensive experiments on three benchmarks demonstrate significant incremental performance, outperforming the state-of-the-art methods by a margin of 3%, 3% and 6%, respectively.

1. Introduction

Since deep neural networks have made great advances in fully supervised conditions, research attention is increasingly turning to other aspects of learning. An important aspect is the ability to continuously learn new tasks as the input stream is updated, which is often the case in real applications. In recent years, class-incremental learning (CIL) [11, 27], a difficult type in continual learning, has attracted much attention, which aims to recognize new classes without forgetting the old ones that have been learned.

In this case, re-training the old and new class samples jointly in each phase is time-consuming and laborious, not to mention that the old class samples may not be fully available. A simple alternative is to fine-tune the network using the new class, however, it will cause the catastrophic forgetting problem [8]. That is, during the optimization process,

[†]Corresponding Author

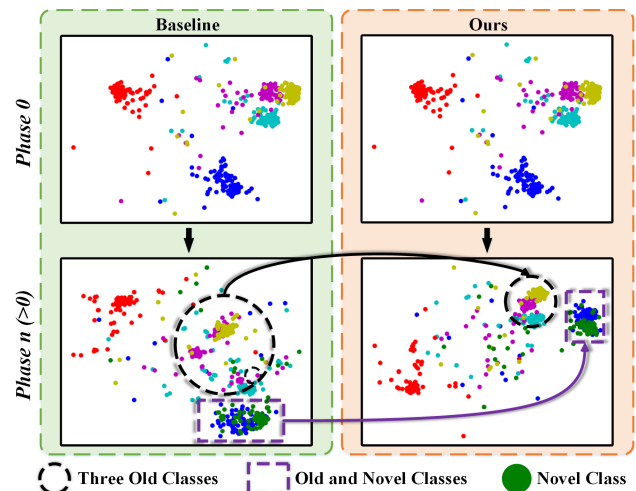


Figure 1. The t-SNE visualization. Compared to the baseline in Section 4.1, (1) the representations of the old classes in our method are better maintained (circular area), (2) and the novel class is more discriminating from the old classes (rectangular area).

the entire representation and the classifier become biased toward the new class, resulting in a sharp drop in the performance for the old class. To deal with it, recent CIL methods maintain the past knowledge by preserving some representative samples (*i.e.*, exemplars [27]) and introducing various distillation losses [7], and correct the bias caused by number imbalance by calibrating the classifier [11].

However, most of the existing methods [21, 30] assume that a certain number (*e.g.*, 2000) of exemplars can be stored in memory, which is usually difficult to satisfy in practice due to user privacy or device limitations. This fact poses great difficulties to incremental learning, because the optimization of the representation and the correction of the classifier will degenerate directly from the imbalance between the old and new classes. To this end, this paper focuses on this ability of incrementally learning new classes where old class samples cannot be preserved, which is called non-

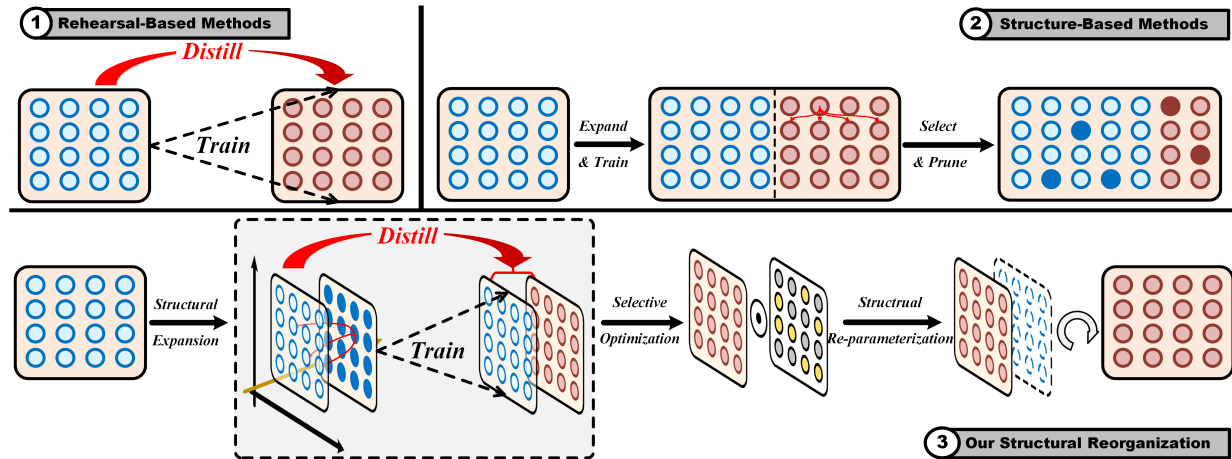


Figure 2. Motivation of our method. In NECIL, the rehearsal-based and structure-based methods suffer from the unreliability of distillation in the absence of exemplars and continuously expanding structure, respectively. DSR is proposed to drive network to expand from a structurally recoverable direction, thus maintaining the discrimination during the new optimization process. On this foundation, we utilize MBD to exploit the ability of distillation-based methods to balance old and new class knowledge.

exemplar class-incremental learning (NECIL [44]).

A natural idea for this problem is to directly transfer the existing CIL framework (*i.e.*, rehearsal-based and structure-based methods in Section 2.1) to NECIL, but the experimental results show that this way leads to performance degradation and parameter explosion. On one hand, in rehearsal-based methods, due to the lack of old class samples, the distillation that the new class samples participate in is the only one that can help maintain the representation of old classes. However, for new samples, it is impractical to provide the same complete old class distribution as the exemplars, so it is difficult to effectively promote the knowledge transfer in the distillation process. Consequently, representative features learned in the old phase are lost phase by phase with the decrease of relevance to the new class.

On the other hand, the idea of structure-based methods is to leave the old model for inference and expand a new model for training at each new phase [28, 35]. Although this strategy maintains the performance of the old class completely, demonstrating strong performance [35], the network parameters that increase linearly with phase (*i.e.*, 5, 10 and 20 in this paper) during training are discouraging. Besides, although a large amount of data can be used to learn the discriminative features among new classes, it is easy to confuse with similar ones from the old distribution. The augmentation of prototypes [44] can only improve the selection of the optimal boundary for the classifier, but cannot essentially improve the discrimination of the old and new classes in the feature representation. As shown in Fig. 1, the representations of old classes obtained by the standard CIL method are more confused compared to the initial phase, because they may gradually overlap with similar classes due to the lack of effective supervision. At the same time, the new

class may directly overlap with the old cluster, resulting in serious confusion to the subsequent optimization process.

To address this problem, we propose a self-sustaining representation expansion scheme to learn a structure-cyclic representation, promoting the optimization from the expanded direction while integrating the overall structure at the end of each phase. As shown in Fig. 2, the preservation of the old classes is reflected in both the structure and feature aspects. First, we adopt a dynamic structure reorganization (DSR) strategy, which leaves structured space for the learning of new class while stably preserving the old class space through maintaining heritage at the main-branch and fusing update at the side-branch. Second, on the basis of the expandable structure, we employ a main-branch distillation (MBD) to maintain the discrimination of the new network with respect to the old features by aligning the invariant distribution knowledge on the old classes.

Specifically, we insert a residual adapter in each block of the old feature extractor to map the old representation to a high-dimensional embedding space, forcing the optimization flow to only pass through the expanding branches unrelated to the old class. After the optimization, we adopt the structural re-parameterization technique to fuse the old and new features and map them back to the initial space losslessly. Furthermore, to reduce the confusion between the newly incremental classes and the original classes, we add a prototype selection mechanism (PSM) during the distillation process. The normalized cosine is first used to measure the similarity between the new representation and the old prototype. Then samples similar to the old classes are used for distillation, maintaining the old knowledge with a soft label that retains the old class statistical information, while those samples dissimilar to the old classes are used for new

class training. This mechanism improves the performance of forward transfer and mitigates the lack of joint optimization to some extent. Our main contributions are as follows:

1) A self-sustaining representation expansion scheme is proposed for non-exemplar incremental learning, in which a cyclically expanding optimization is accomplished by a dynamic structure reorganization strategy, resulting in a structure-invariant representation.

2) A prototype selection mechanism is proposed, which combinatorially co-uses the preserved invariant knowledge and the incoming new supervision to reduce the feature confusion among the similar classes.

3) Extensive experiments are performed on benchmarks including CIFAR-100, TinyImageNet and ImageNet-Subset, and the results demonstrate the superiority of our method over the state of the art.

2. Related Work

2.1. Incremental Learning

As deep learning research advances [34, 41], there is a growing demand for continual learning of neural networks, which requires the network to learn new tasks without forgetting the old knowledge to achieve the stability-plasticity trade-off. CIL [32, 45] is the most difficult scenery in continual learning and has received more attention recently. Current methods can broadly be divided into the following three classes. Regularization-based methods [15, 39] estimate the importance of the network parameters learned in the past tasks and constrain their optimization accordingly. Rehearsal-based methods [1, 7, 12, 22, 33] preserve exemplars of fixed memory size to maintain the distribution of old classes in the incremental phases, and adopt the distillation skills to retain the discriminative features of the old task. [11] incorporates three components, cosine normalization, less-forget constraint, and inter-class separation, to address the imbalance between previous and new data. The techniques on exemplar and distillation in rehearsal-based methods are widely used in class-incremental learning. Structure-based methods [13, 29] select and expand different sub-network structures involved in the optimization process of the incremental tasks. [24] progressively chooses optimal paths for the new tasks while encouraging parameter sharing, which promotes the forward knowledge transfer. [35] freezes the previously learned representation and augments it with additional feature dimensions from a new mask-based feature extractor. The structure-based methods are often mixed with other techniques such as exemplar and distillation, and have achieved good results.

Recently, some works [36, 38, 44] focus on a challenging but practical non-exemplar class-incremental learning problem, where no past data can be stored due to equipment limits or privacy security. [38] estimates the seman-

tic drift of incremental features and compensates the prototypes in each test phase. [44] adopts prototype augmentation to maintain the decision boundary of previous tasks, and employ self-supervised learning to learn more transferable features for future tasks. We follow their NECIL settings. However, different from their work considering generalizable features and augmented prototypes, we mainly consider the adjustment for joint representation learning and distillation process in the absence of exemplars.

2.2. Residual Block

Residual block has been widely used in convolutional neural network as the basic structure of ResNet [10], which improves the network depth and prevents vanishing gradient. Further improvements have been investigated for superior dynamic performance [19] and inference efficiency [5, 6, 9] recently. In domain adaptation [20, 43], residual adapter [18, 25, 26] is proposed to learn style information related to new domains, thus improving the overall generalization performance of the network. In these efforts, residual block is used to improve joint optimization performance or statistical domain information. Instead, we consider dynamically incremental residual blocks to learn new knowledge efficiently while maintaining old features.

3. Problem Description

The NECIL problem is defined as follows. Here we denote X , Y and Z as the training set, the label set and the test set, respectively. Our task is to train the model from a continuous data stream, *i.e.*, training sets X^1, X^2, \dots, X^n , where samples of a set X^i ($1 \leq i \leq n$) are from the label set Y^i , and n represents the incremental phase. It should be mentioned that all the incremental classes are disjoint, that is, $Y^i \cap Y^j = \emptyset$ ($i \neq j$). Except that there are sufficient samples in the current phase X^i , no old samples are available in memory for old classes. To measure the performance of models in NECIL task, we calculate the classification accuracy on the test set Z^i at each phase i . Different from the training set, the classes of the test set Z^i are from all the seen label sets $Y^1 \cup Y^2 \dots \cup Y^i$.

4. Methodology

First of all, we demonstrate the paradigms of standard CIL and how we adapt it to the NECIL setting as the baseline. Then we analyze the optimization flow of the overall pipeline and explain why it doesn't work well. Finally, two proposed core components dynamic structure reorganization and prototype selection are introduced.

4.1. Standard NECIL Paradigm

[27] first proposed a practical strategy for decoupling representation and classifiers learning in the CIL setting,

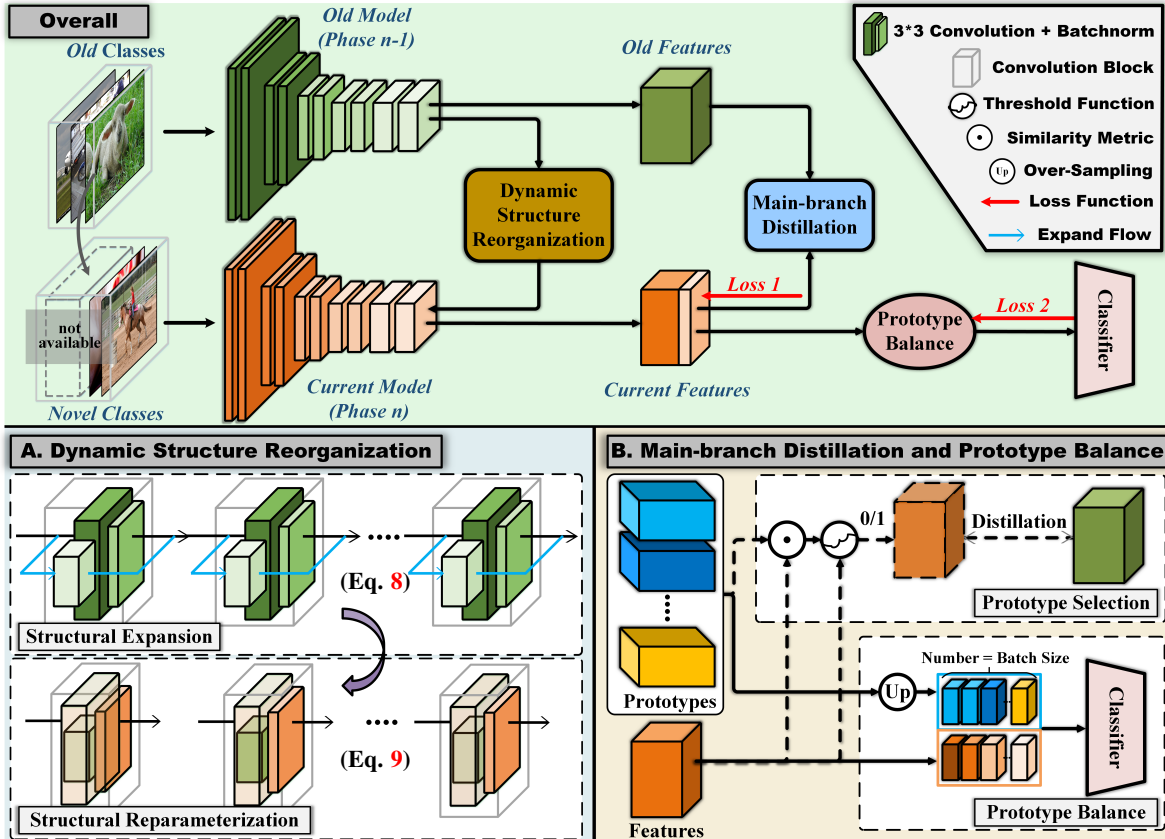


Figure 3. Our proposed self-sustaining representation expansion scheme for NECIL: (a) overview of our scheme, (b) dynamic structure reorganization, and (c) main-branch distillation and prototype balance. The source code will be made available to the public.

which is followed by most subsequent work. The three main components are representation learning using knowledge distillation and prototype rehearsal, prioritized exemplar selection, and classification by a balance calibration.

Incremental Representation Learning. As the exemplar cannot be saved in the NECIL setting, the representation learning will be slightly different, mainly in terms of cross-entropy loss and distillation loss. At the first phase, a standard classification model f_{θ}^1 consisting of the feature extractor f_e^1 and classifier g_c^1 is optimized under the full supervision, *i.e.*, X^1 and Y^1 . At the incremental phase ($n > 1$), the input of the current model is only the predicted images Q from X^n without old samples. A base feature extractor f_e^n such as VGG [31] or ResNet [10] parameterized by θ_e^n is utilized to learn the corresponding representation:

$$r_q^n = f_e^n(Q; \theta_e^n). \quad (1)$$

Then, to learn the discriminative features among the novel classes, the obtained representation is optimized under the supervision of the class label y_q^n from Y^n . We adopt the fully connected layer as the classifier g_c^n to map the repre-

sentation to the label space,

$$s_q^n = g_c^n(r_q^n; \theta_c^n), L_{ce} = F_{ce}(s_q^n, y_q^n), \quad (2)$$

F_{ce} represents the standard cross-entropy loss. Finally, to maintain the useful information learned with the old classes, the knowledge distillation is used to measure the similarity between the obtained representation and that of the previous model f_e^{n-1} ,

$$r_q^{n-1} = f_e^{n-1}(Q; \theta_e^{n-1}), L_{kd} = F_{kd}(r_q^n, r_q^{n-1}), \quad (3)$$

F_{kd} represents Euclidean distance the same as [44].

Incremental Classifier Calibration. A common way to overcome the imbalance between the exemplars and new samples in CIL is to under-sample a balanced subset for fine-tuning. As there is no exemplars in NECIL, we memorize one prototype in the deep feature space for each class, which is consistent with PASS [44]. Different from PASS augmenting the prototypes via Gaussian noise, we choose to over-sample (*i.e.*, Up_B) prototypes to the batch size (*i.e.*, B), achieving the calibration of the classifier, which is the simplest way adopted in the long-tail recognition [2],

$$p_B = Up_B(Prototype), L_{proto} = F_{ce}(p_B, y_B), \quad (4)$$

y_B is the over-sampled label set of the initial prototypes. The final loss for current model is their addition:

$$L = L_{ce} + \lambda L_{kd} + \gamma L_{proto}, \quad (5)$$

λ and γ are loss weights, and we set them to 10.

4.2. Optimization

Different from the previous work focusing on the effect on the classifier, this paper tries to analyze the representation. In the CIL, Equation 2 can be turned into two parts:

$$L_{ce} = F_{ce}(s_q^n, y_q^n) + F_{ce}(s_e^n, y_e^n), \quad (6)$$

s_e^n represents the saved exemplars, whose number is much lower than that of s_q^n . While this imbalance can bias the optimization process towards features that are more discriminative for the new class, the added distillation in Equation 3 can alleviate this problem,

$$L_{kd} = F_{kd}(r_q^n, r_q^{n-1}) + F_{kd}(r_e^n, r_e^{n-1}). \quad (7)$$

r_e^n represents the representation of exemplars. In this case, the features that are significant for the old and new classes will be maintained. However, note that there is no exemplars involved in the above NECIL setting. It means that the joint optimization on the old and new class representations completely collapses into feature optimization that is relevant only to incremental classes. What is reflected in the first part is that the cross-entropy loss will only focus on the features that facilitate the recognition of the new class, while in the second part, it will focus on the maintenance of the features related to the new class, which both accelerate the forgetting of the representative features of the old class. Suppose the forgetting rate (Fr) of distillation part in the initial phase is α . Note that the distillation loss is based on the overall representation of the previous phase, and the error will accumulate exponentially with the phase, that is, $Fr_n \geq \alpha^{n-1}$. Therefore, it is necessary to correct this error from the representational level.

4.3. Self-Sustaining Representation Expansion

Dynamic Structure Reorganization. To retain the representation [42] of the old class and guarantee the unbiased training of the new class, we propose a dynamic structure reorganization strategy. In general, as shown in Fig. 3, we firstly adopt the structural expansion to add the side branch to the current model by block for the optimization of new classes. Specifically, we insert a residual adapter to each convolution block of the fixed feature extractor from previous phase. The optimized flow propagates only through the adapter, updating the most discriminating position while maintaining the old features,

$$\begin{aligned} f_e^n(Q; \theta_e^n) &= F_{transform}(f_e^{n-1}(Q; \theta_e^{n-1})) \\ &= f_e^{n-1}(Q; \hat{\theta}_e^{n-1} \oplus \Delta\theta_e^n), \end{aligned} \quad (8)$$

where $\hat{\cdot}$ represents the fixed parameters, and \oplus represents the structural expansion operation. After training, we use the structural re-parameterization [6] to integrate the side-branch information into the main branch losslessly, ensuring that the number of network parameters does not increase at the end of each phase. Specifically, the parameters in the residual structures are fused with the parameters of the original convolution kernel and BatchNorm [14] through the zero-padding operation and linear transformation, and finally the adapters are removed to keep the network structure unchanged for the next update,

$$\begin{aligned} f_e^{n-1}(Q; \hat{\theta}_e^{n-1} \oplus \Delta\theta_e^n) \\ = f_e^n(Q; \theta_e^{n'} \oplus 0) = f_e^n(Q; \theta_e^{n'}). \end{aligned} \quad (9)$$

Prototype Selection. While new features are learned based on the old structure, the old class features are maintained in coordination with the main-branch distillation. To reduce the feature confusion in the distillation part, we adopt a prototype selection mechanism based on the expandable embedding space. In general, based on the similarity between the representation of new samples and the old prototypes, dissimilar samples are involved in the update of residual adapter to learn new features, and similar samples are involved in the distillation to retain the old discriminative features maintained in the main branch at previous phase. Specifically, after mapping all new samples to the learned embedding space [3, 40], we compute the normalized cosine scores Si between them and all prototypes,

$$Si = \text{Cosine}(N(r_q^n), \text{Nor}(\text{Prototype})), \quad (10)$$

Nor represents the normalization operation. We then set a threshold value, and attach a mask to the corresponding position of its distillation loss ($Mask_{kd}$) if greater than the threshold σ , and add a mask to the corresponding part of its cross-entropy loss ($Mask_{ce}$) if less than the threshold. Finally, the two losses are summed with the prototype balance loss as the final optimization function for the new phase.

$$L = Mask_{ce}(L_{ce}) + \lambda Mask_{kd}(L_{kd}) + \gamma L_{proto}. \quad (11)$$

5. Experiments

5.1. Dataset and Settings

Dataset. To evaluate the performance of our proposed method, we conduct comprehensive experiments on three datasets CIFAR-100 [16], TinyImageNet [17] and ImageNet-Subset. CIFAR-100 contains 60000 images of 32×32 size from 100 classes, and each class includes 500 training images and 100 test images. TinyImageNet contains 200 classes, and each class contains 500 training images, 50 validation images and 50 test images. It provides more phases and incremental classes to compare the sensitivity of different methods. ImageNet-Subset is a 100-class

DSR	MBD	PSM	CIFAR-100		
			5 phases	10 phases	20 phases
			61.11	57.08	51.04
✓			64.86	63.25	54.09
	✓		62.70	62.60	58.57
✓	✓		65.10	63.87	60.60
✓	✓	✓	65.88	64.69	61.61

Table 1. Ablation study of our method on CIFAR-100.

Method	CIFAR-100		
	5 phases	10 phases	20 phases
3×3 conv	64.28	63.47	60.81
1×1 conv + bn	65.88	64.84	60.72
1×1 conv	65.87	65.12	61.60

Table 2. Performance under different expanding structures.

subset of ImageNet-1k [4], which is much larger. For the order and division of all dataset classes in our experiments, we followed exactly the settings in [44].

Setting. As adopted in [44], we use ResNet-18 as the backbone network. The difference is that we use standard supervised training for the whole optimization process instead of involving self-supervised learning. For a fair comparison, we achieve the same accuracy as [44] at the first phase for all datasets. We use an Adam optimizer, in which the initial learning rate is set to 0.001 and the attenuation rate is set to 0.0005. The model stops training after 100 epochs, and batch size is set to 128.

Evaluation Metrics. Following [44], we report average incremental accuracy and average forgetting, and our performance is evaluated on three different runs. Average incremental accuracy is computed as the average accuracy of all the incremental phases (including the first phase), which compares the overall incremental performance of different methods fairly. Average forgetting is computed as the average forgetting of different tasks throughout the incremental process, which directly measures the ability of different methods to resist catastrophic forgetting.

5.2. Ablation Study

To prove the effectiveness of our proposed method, we conduct several ablation experiments on CIFAR-100. The performance of our scheme is mainly attributed to two prominent components: the dynamic structure reorganization strategy (DSR) and the main-branch distillation (MBD). To clarify the function of DSR, we replace the dynamic representation with the structurally invariant representation, which is adopted in most CIL methods [7, 44]. To clarify the function of MBD, we replace the distillation process with the one that interacts with the continuously optimized representation. As can be seen in Table 1, the

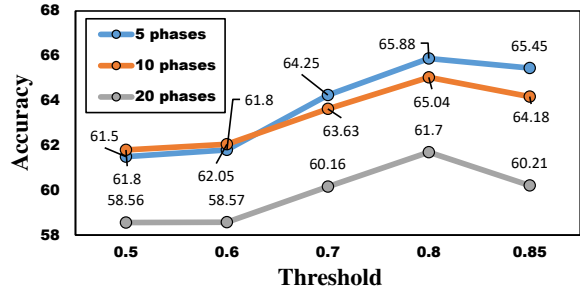


Figure 4. Illustration of the role of the selection mechanism.

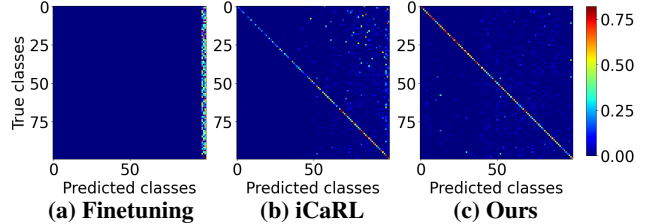


Figure 5. Confusion matrices of different methods on CIFAR-100.

dynamic representation and main-branch distillation separately bring a 4.3% and 4.8% improvement in overall performance. It demonstrates that the two parts are far more useful than standard representation and common distillation respectively in NECIL. It is worth noting that the former plays a greater role when there are fewer incremental phases (*i.e.*, 5 and 10 phases), while the latter shines more brightly when there are more incremental phases (*i.e.*, 20 phases). It demonstrates that keeping the old class features helps to improve the overall performance of incremental learning in the short term. However, as analyzed in the introduction, if the distilled network keep decaying or fixed, the errors will accumulate as the incremental phase increases. At this point, how to reasonably correct the distillation loss is the key to ensure the long-term effect.

5.3. Analysis

The impact of the adapter structure. To explore the impact of the structure of residual adapter on expandable representation during training, we design the following experiments. We adopt three different convolution blocks to the residual part: 1×1 convolution only, 3×3 convolution only and the combination of 1×1 convolution and Batch-Norm. As shown in Table 2, the results of 1×1 convolution and the combination are similar, and that of 3×3 convolution is one point lower. It suggests that the 1×1 convolution structure is good enough to learn the representation of the new class without needing more parameters.

The impact of the threshold in prototype selection. To verify the role of the PSM, we conduct data statistics on the incremental samples. As shown in Fig. 8, the new classes have a large difference in similarity. And the intra-class

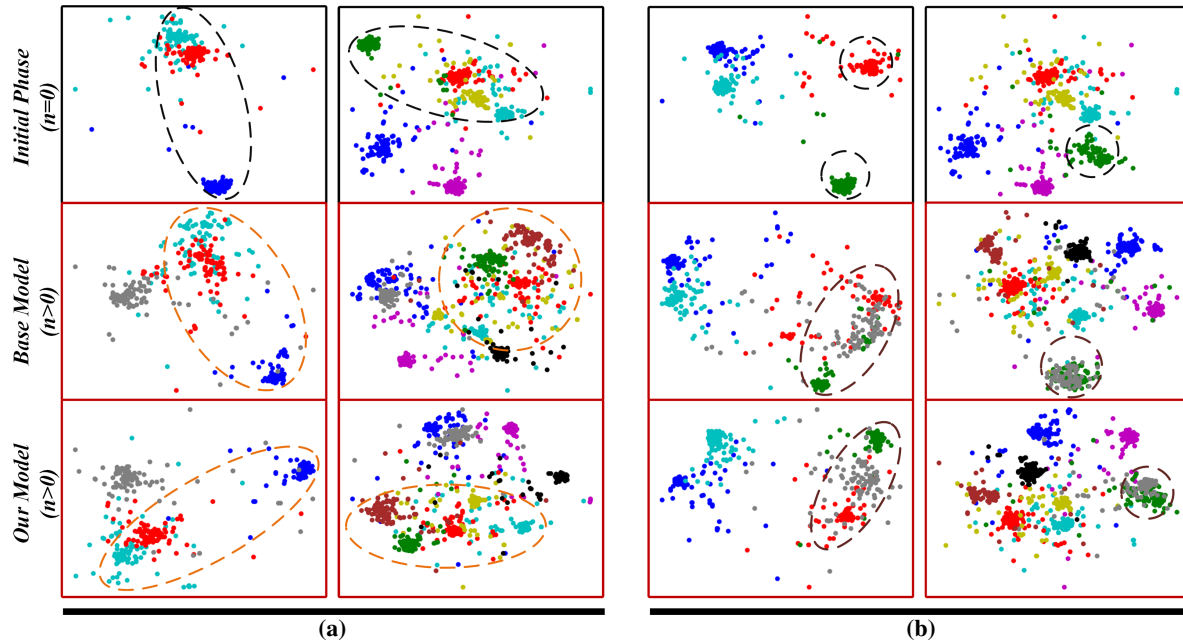


Figure 6. Effect of our scheme on the representation. (a) DSR maintains the discriminative features and inter-relations of old classes, thus enhancing the clustering and separation of the distribution of old classes. (b) MBD results in a better distinction between similar classes.

Methods		CIFAR-100			TinyImageNet			ImageNet-Subset
		$P=5$	$P=10$	$P=20$	$P=5$	$P=10$	$P=20$	$P=10$
(1) $E=20$	iCaRL-CNN*	51.07	48.66	44.43	34.64	31.15	27.90	50.53
	iCaRL-NCM* [27]	58.56	54.19	50.51	45.86	43.29	38.04	60.79
	EEIL* [1]	60.37	56.05	52.34	47.12	45.01	40.50	63.34
	UCIR* [11]	63.78	62.39	59.07	49.15	48.52	42.83	66.16
(2) $E=0$	EWC* [15]	24.48	21.20	15.89	18.80	15.77	12.39	20.40
	LwF_MC* [27]	45.93	27.43	20.07	29.12	23.10	17.43	31.18
	MUC* [37]	49.42	30.19	21.27	32.58	26.61	21.95	35.07
	SDC [38]	56.77	57.00	58.90	-	-	-	61.12
	PASS [44]	63.47	61.84	58.09	49.55	47.29	42.07	61.80
Ours		65.88+2.41	65.04+3.20	61.70+2.80	50.39+0.84	48.93+1.64	48.17+6.10	67.69+5.89

Table 3. Comparisons of the average incremental accuracy (%) with other methods on CIFAR-100, TinyImageNet, and ImageNet-Subset. P represents the number of phases and E represents the number of exemplars. Models with an asterisk * represent the reproduced results in [44]. The red footnotes in the last row represent the relative improvement compared with the results of SOTA.

fluctuations are also large, so different classes and samples involved in the optimization process will bring different changes. Therefore it is important to reasonably place them in the two potentially conflicting processes of old feature distillation and new feature learning. To demonstrate the sensitivity of the threshold on the distillation effect, we plot its fluctuation curve. As shown in Fig. 4, all curves rise to a peak at a threshold of 0.8, then gradually fall and lose distillation effect. It suggests that in the absence of the exemplars, fine-grained optimization of the new samples can better maintain the old features and learn the new features.

Classification accuracy of old and novel classes. To evaluate performance of both old and new classes during training, we compare their accuracy at each phase. As shown in Fig. 5, our method achieves similar performance

between the old and new classes without favoring one side due to overfitting, which is a prerequisite for a good incremental learning system.

5.4. Visualization

To better demonstrate the role of DSR and MBD during optimization, we show the visualization results with t-SNE [23]. As shown in Fig. 6 (a), although the old classes have slightly changed in the representation after multi-phase optimization, their discrimination and relative relationship almost do not decline with our DSR. As shown in Fig. 6 (b), newly incremental classes are easily confused with some of the old classes. Owing to our MBD, the optimized features are promoted to differentiate from the old class, thus improving the separation of novel clusters.

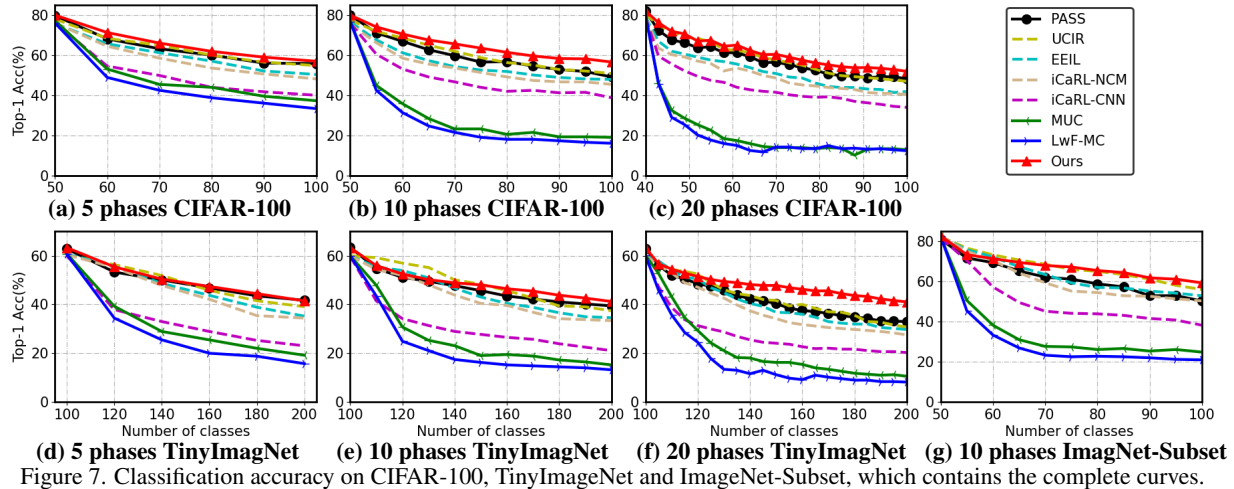


Figure 7. Classification accuracy on CIFAR-100, TinyImageNet and ImageNet-Subset, which contains the complete curves.

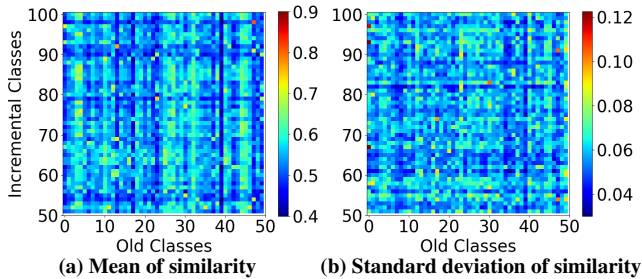


Figure 8. Statistics of similarity on the incremental samples.

Method	CIFAR-100			TinyImageNet		
	5	10	20	5	10	20
iCaRL-CNN	42.13	45.69	43.54	36.89	36.70	45.12
iCaRL-NCM	24.90	28.32	35.53	27.15	28.89	37.40
EEIL	23.36	26.65	32.40	25.56	25.91	35.04
UCIR	21.00	25.12	28.65	20.61	22.25	33.74
LwF_MC	44.23	50.47	55.46	54.26	54.37	63.54
MUC	40.28	47.56	52.65	51.46	50.21	58.00
PASS	25.20	30.25	30.61	18.04	23.11	30.55
Ours	18.37	19.48	19.00	9.17	14.06	14.20

Table 4. Results of average forgetting on 5, 10 and 20 phases.

5.5. Comparison with SOTA

To better assess the overall performance of our scheme, we compare it to the SOTA of NECIL (EWC*, LwF_MC*, MUC*, SDC and PASS) and some classical methods of exemplar-based CIL (iCaRL*, EEIL* and UCIR*).

Average accuracy and average forgetting. As shown in Table 3, compared to the SOTA of non-exemplar methods (E=0), our method achieves average improvement of 3, 3 and 6 points on CIFAR-100, TinyImageNet and ImageNet-Subset, respectively. The performance of our method is comparable to the classical exemplar-based methods (E=20), which shows that our scheme further reduces the impact of exemplars on CIL models. To provide further insight into the behaviors of different methods, we compare their average forgetting of all phases. As shown in Table 4, our method achieves much lower average forgetting, resisting catastrophic forgetting well in the absence of exemplars.

Trend of accuracy. To analyze the trend of different methods, we show the detailed accuracy curves on three datasets. As shown in Fig. 7, our method is superior at almost all phases, striking a better stability-plasticity balance. It can be seen that the difficulty increases as the number of incremental phases (P) increases. In this process, the advantage of our method are even expanding, such as in TinyImageNet. Whether in the smaller CIFAR-100 or the

larger ImageNet-Subset dataset, our method has a notable advantage, demonstrating its robustness.

6. Conclusion and Discussion

In this paper, a novel self-sustaining representation expansion scheme is presented for the NECIL task. A dynamic structure reorganization strategy is first proposed to optimize the newly incremental features in a side branch while maintaining the old feature distribution from the structurally expanded direction, and then the distillation process is arranged in the main branch. In particular, a prototype selection mechanism is integrated into the joint training to enhance the distinction between the old and new classes. Experimental results show that our method is superior in both performance and adaptability to the state-of-the-art methods, especially in the multi-phase process.

Acknowledgments. Supported by National Key R&D Program of China under Grant 2020AAA0105700, National Natural Science Foundation of China (NSFC) under Grants 61872327 and U19B2038, Major Special Science and Technology Project of Anhui (No. 012223665049), the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025.

References

- [1] 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] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [3] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10010–10019, 2021.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [5] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and J. Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1911–1920, 2019.
- [6] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021.
- [7] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020.
- [8] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.
- [9] Shuxuan Guo, José Manuel Álvarez, and Mathieu Salzmann. Expandnets: Linear over-parameterization to train compact convolutional networks. *arXiv: Computer Vision and Pattern Recognition*, 2020.
- [10] 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.
- [11] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.
- [12] Xinting Hu, Kaihua Tang, Chunyan Miao, Xiansheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3956–3965, 2021.
- [13] Steven C. Y. Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compact-ing, picking and growing for unforgetting continual learning. *ArXiv*, abs/1910.06562, 2019.
- [14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [15] 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.
- [16] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.
- [17] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.
- [18] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Improving task adaptation for cross-domain few-shot learning. *ArXiv*, abs/2107.00358, 2021.
- [19] Yunsheng Li, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Ye Yu, Lu Yuan, Zicheng Liu, Mei Chen, and Nuno Vasconcelos. Revisiting dynamic convolution via matrix decomposition. *arXiv preprint arXiv:2103.08756*, 2021.
- [20] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7202–7211, 2019.
- [21] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2544–2553, 2021.
- [22] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12245–12254, 2020.
- [23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [24] Jathushan Rajasegaran, Munawar Hayat, Salman Khan, Fahad Shabbaz Khan, and Ling Shao. Random path selection for incremental learning. *Advances in Neural Information Processing Systems*, 2019.
- [25] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *arXiv preprint arXiv:1705.08045*, 2017.
- [26] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018.
- [27] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

- [28] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- [29] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *ArXiv*, abs/1606.04671, 2016.
- [30] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1591–1600, 2021.
- [31] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.
- [32] Gido M. van de Ven and A. Tolias. Three scenarios for continual learning. *ArXiv*, abs/1904.07734, 2019.
- [33] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019.
- [34] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitaev: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021.
- [35] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021.
- [36] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deep-inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020.
- [37] L. Yu, S. Parisot, G. Slabaugh, J. Xu, and T. Tuytelaars. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. *European Conference on Computer Vision*, 2020.
- [38] Lu Yu, Bartłomiej 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.
- [39] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.
- [40] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, and Tat-Seng Chua. Robust (semi) nonnegative graph embedding. *IEEE transactions on image processing*, 23(7):2996–3012, 2014.
- [41] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022.
- [42] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [43] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5310–5319, 2021.
- [44] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.
- [45] Kai Zhu, 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.