

# SELF-ORGANIZING PATHWAY EXPANSION FOR NON-EXEMPLAR INCREMENTAL LEARNING ( SUPPLEMENTARY MATERIALS)

**Anonymous authors**

Paper under double-blind review

## A. DETAILED EXPLANATION

### A.1. STANDARD DEVIATION OF THE INCREMENTAL PERFORMANCE (ERROR BARS)

All results of the average incremental accuracy and average forgetting are evaluated on three different runs. To show the stability of our method, we report its standard deviation on three runs. As shown in Table 1 and 2, random factors have little impact on our scheme.

Methods		Average Accuracy ( $\uparrow$ )			Average Forgetting ( $\downarrow$ )		
		$P=5$	$P=10$	$P=20$	$P=5$	$P=10$	$P=20$
(1) $E=20$	iCaRL-CNN*	51.07	48.66	44.43	42.13	45.69	43.54
	iCaRL-NCM*	58.56	54.19	50.51	24.90	28.32	35.53
	EEIL*	60.37	56.05	52.34	23.36	26.65	32.40
	UCIR* Hou et al. (2019)	63.78	62.39	59.07	21.00	25.12	28.65
	PODNet <sup>†</sup>	64.88	63.05	61.62	19.12	22.55	25.64
(2) $E=0$	LwF_MC	45.93	27.43	20.07	44.23	50.47	55.46
	MUC Yu et al. (2020a)	49.42	30.19	21.27	40.28	47.56	52.65
	SDC <sup>‡</sup> Yu et al. (2020b)	56.77	57.00	58.90	6.96	7.50	10.77
	PASS Zhu et al. (2021b)	63.47	61.84	58.09	25.20	30.25	30.61
	IL2A <sup>‡</sup> Zhu et al. (2021a)	65.72	62.69	59.90	27.25	37.35	39.27
	ABD <sup>‡</sup> Yin et al. (2020)	63.85	62.46	57.40	23.12	27.34	33.42
	Ours	<b>66.64<math>\pm</math>0.01</b>	<b>65.84<math>\pm</math>0.07</b>	<b>61.83<math>\pm</math>0.12</b>	<b>6.50<math>\pm</math>0.13</b>	<b>3.30<math>\pm</math>0.39</b>	<b>9.14<math>\pm</math>1.42</b>

Table 1: Comparisons with other methods on CIFAR-100 dataset. P represents the number of phases and E represents the number of exemplars. Models with an asterisk \* represent the reproduced results in Zhu et al. (2021b). Models with a marker <sup>†</sup> represent the reproduced results of ours. The blue footnotes in the last row represent the values of error bars.

Methods		TinyImageNet			ImageNet-Subset
		$P=5$	$P=10$	$P=20$	$P=10$
(1) $E=20$	iCaRL-CNN*	34.64	31.15	27.90	50.53
	iCaRL-NCM*	45.86	43.29	38.04	60.79
	EEIL* Castro et al. (2018)	47.12	45.01	40.50	63.34
	UCIR* Hou et al. (2019)	49.15	48.52	42.83	66.16
(2) $E=0$	LwF_MC Rebuffi et al. (2017)	29.12	23.10	17.43	31.18
	MUC Yu et al. (2020a)	32.58	26.61	21.95	35.07
	MAS Aljundi et al. (2018)	18.97	11.82	7.17	19.11
	EWC Kirkpatrick et al. (2017)	19.64	16.18	17.09	27.32
	PASS Zhu et al. (2021b)	49.55	47.29	42.07	61.80
	Ours	<b>53.69<math>\pm</math>0.14</b>	<b>52.88<math>\pm</math>0.05</b>	<b>51.94<math>\pm</math>0.28</b>	<b>69.22<math>\pm</math>0.05</b>

Table 2: Comparisons of the average incremental accuracy (%) with other methods on 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 Zhu et al. (2021b). The blue footnotes in the last row represent the values of error bars.

## A.2. DETAILED SETTING

We use an Adam optimizer, in which the initial learning rate is set to 0.001 and the attenuation rate is set to 0.0002. The batch size is set to 128. The model stops training after 160 epochs and 60 epochs during the initial phase and incremental phases, respectively. We adopt ResNet18 and 3 standard convolution blocks as the backbone of feature extractor  $f_{\theta_t}$  and pathway planner  $f_{\alpha_t}$ , respectively.

In the main text, the maximum value of sparse rate in Equation 10 is set to 0.4. The values of  $e_1$  and  $e_2$  in Equation 10 are set to 0.08 and 0.75, respectively. The values of  $L$  and  $K$  in Section 3.3 are set to 4 and 16, respectively. One NVIDIA GTX2080Ti gpu is utilized for CIFAR-100 and TinyImageNet datasets. Two NVIDIA GTX3090 and eight NVIDIA Tesla A100 gpu are utilized for ImageNet-Sub and ImageNet-Full datasets, respectively. All datasets adopted in this paper are open to the public.

## A.3. OPTIMIZATION EXPLANATION IN OUR SCHEME

The optimization of feature representation  $\theta_t$  is mainly guided by the classification loss function  $\mathcal{L}_{cls}$  and feature distillation loss function  $\mathcal{L}_{kd}$ . Assuming that the optimal solution at the incremental phase  $t-1$  is taken when  $\theta_{t-1} = \theta_{t-1}^*$ . As  $\theta_t$  is initialized by the value of  $\theta_{t-1}^*$ , it can be assumed that  $\theta_t$  is close to  $\theta_{t-1}^*$ . Then the Taylor expansion on  $\theta_t$  can be written as follows,

$$\begin{aligned} f(\theta_t) &= f(\theta_{t-1}^*) + \left( \frac{\partial f(\theta)}{\partial \theta} \Big|_{\theta=\theta_{t-1}^*} \right) (\theta_t - \theta_{t-1}^*) + \\ &\quad \frac{1}{2} (\theta_t - \theta_{t-1}^*)^T \left( \frac{\partial^2 f(\theta)}{\partial^2 \theta} \Big|_{\theta=\theta_{t-1}^*} \right) (\theta_t - \theta_{t-1}^*) + o(\theta_{t-1}^*). \end{aligned} \quad (1)$$

The first order component is constrained to zero by the gradient descent, and the ones higher than second order can be ignored. The subscript  $t$  can be omitted for brevity, and Equation 1 can be approximated as follows:

$$\begin{aligned} f(\theta) &= f(\theta^*) + \frac{1}{2} (\theta - \theta^*)^2 f''(\theta^*) = f(\theta^*) + \frac{1}{2} \Omega (\theta - \theta^*)^2 = \\ &\quad f(\theta^*) + \frac{1}{2} (\Omega_{cls} + \Omega_{kd}) (\theta - \theta^*)^2, \end{aligned} \quad (2)$$

where  $\Omega_{cls}$  and  $\Omega_{kd}$  represents the importance of parameter space on the classification and distillation tasks, which is commonly estimated in different incremental methods Kirkpatrick et al. (2017); Aljundi et al. (2018). To mitigate the interference between the two objectives, we can improve their respective weight sparsity (*i.e.*, the sparsity of  $\Omega_{cls}$  and  $\Omega_{kd}$ ), and reduce the shared space of important parameters.

## A.4. DETAILED VALUES OF THE CURVES

To facilitate the fair comparison of subsequent work, we report the detailed values of incremental accuracy for each phase in Table 3, 4 and 5. The average accuracy is consistent with the one in Table 3 and 4 of the main text.

## A.5. MORE RESULTS ON VISUALIZATION.

To better demonstrate the role of CPO and PFU during optimization, we show more corresponding visualization results. In Fig. 1 (a), the center of the circle represents the novel class, and the surrounding represents the five different base classes. The middle values represent the intersection of union (IoU) of pathways between the new and old classes. It can be seen the pathways are class-specific, and the similarity is also positively related to the class relationship. As shown in Fig. 1 (b), the features of shared and unshared pathways are visualized by Grad-CAM Selvaraju et al. (2017). To further distinguish between the old and novel class, the novel one expands new pathways to learn representative features.

## A.6. CONFUSION MATRIX.

To evaluate performance of both old and new classes during training, we compare their accuracy on two setting (*i.e.* 5 and 10 incremental phases). As shown in Fig. 2, our method achieves similar

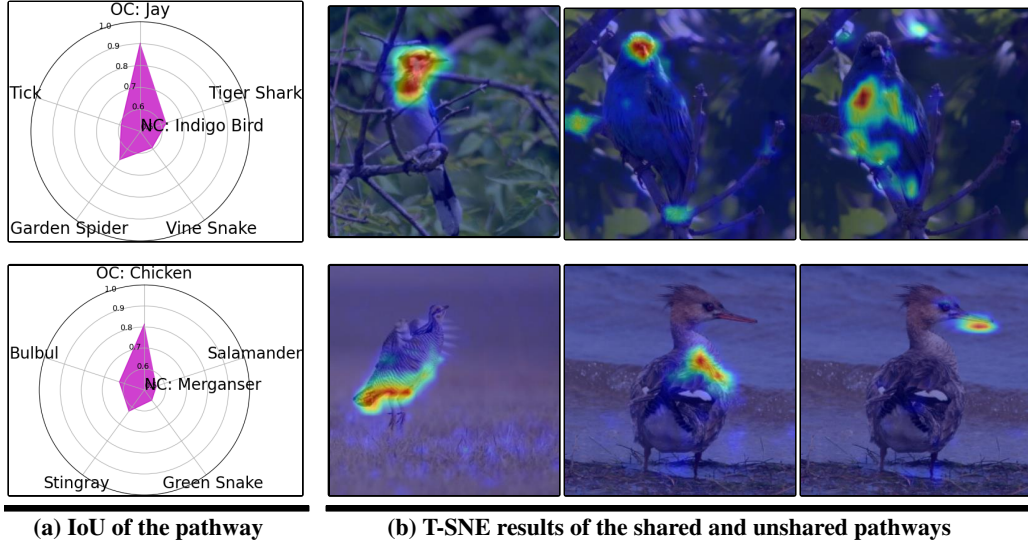


Figure 1: Effect of our scheme on the pathway learning. (a) CPO realizes the organization of distinguishable pathways, thus mitigating the overlap between the incremental classes and the old ones. NC represents the novel classes. (b) PFU promotes the pathway expansion of similar classes. The first two columns represent the shared pathways, and the last represents the unshared ones.

Dataset	Phase									
	0	1	2	3	4	5	6	7	8	9
A	82.40	78.23	74.87	72.34	68.62	67.96	65.52	64.84	62.57	60.83
B	62.70	59.92	58.55	57.01	55.25	54.42	53.18	52.74	52.27	51.70

Dataset	Phase									
	10	11	12	13	14	15	16	17	18	19
A	59.76	58.85	57.39	55.72	54.66	53.54	53.21	52.55	52.24	51.54
B	51.25	50.76	50.19	49.25	48.71	47.95	47.66	47.09	46.66	45.57

Dataset	Phase
	20
A	50.82
B	45.03

Table 3: Detailed values of classification accuracy under the setting of 20 incremental phases. A and B represent the CIFAR-100 and TinyImageNet datasets, respectively.

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

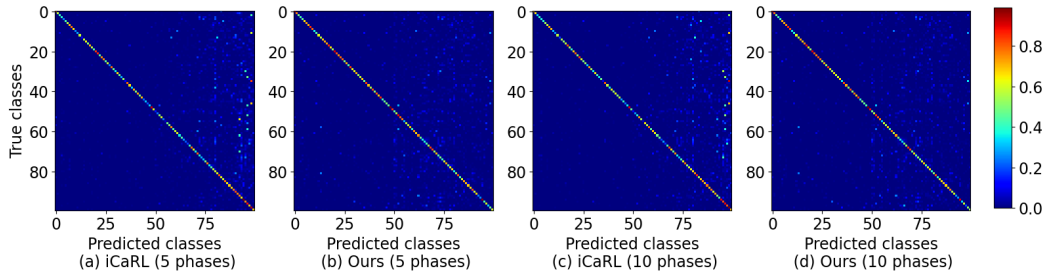


Figure 2: Confusion matrices of different methods on CIFAR-100. 5 phases and 10 phases settings are considered to evaluate the stability of our method on the old and novel classes.

Dataset	Phase										
	0	1	2	3	4	5	6	7	8	9	10
CIFAR-100	80.90	76.15	73.13	69.58	66.73	64.59	63.01	60.81	58.68	56.16	54.50
TinyImageNet	62.70	59.05	56.00	54.25	53.03	52.37	51.21	50.13	48.73	47.81	46.41
ImageNet-Subset	83.40	77.29	73.93	71.75	69.64	68.56	67.61	65.07	63.01	61.22	59.91
ImageNet-Full	76.46	67.59	64.92	62.89	60.61	58.72	57.12	55.75	54.17	52.30	51.65

Table 4: Detailed values of classification accuracy under the setting of 10 incremental phases.

Dataset	Phase					
	0	1	2	3	4	5
CIFAR-100	80.90	72.63	65.87	62.94	60.76	56.76
TinyImageNet	62.70	57.35	54.30	52.04	48.96	46.79

Table 5: Detailed values of classification accuracy under the setting of 5 incremental phases.

#### A.7. RELATED WORK ON FILTER PRUNING METHODS.

Network pruning Liebenwein et al. (2019); Sui et al. (2021); Gao et al. (2018) is an important technology to reduce memory size and bandwidth. Recently, various network pruning techniques have been proposed, which can be classified from the structural aspect, *i.e.*, the structured and unstructured pruning. Specifically, structured methods remove parameters in groups by pruning neurons, filters, or channels. Classical filter pruning methods Sui et al. (2021); Gao et al. (2018) make up the prominent family of structured methods for CNNs. Different from the pruning methods designed for the network efficiency, our scheme aims at the mitigation of update interference. The concept of group in this paper is slightly different as only the output channels of features are divided for the organization of pathway.

#### A.8. LIMITATION AND SOCIETAL IMPACT

The division way of the standard classification model in our method is too simple, which constrains the adjustment of some factors. As shown in Fig. 4, the maximum sparsity can only be kept below 0.5, which deserves the further improvement. Our non-exemplar method avoids the issue of privacy but an old model needs to be maintained during the training, which poses a risk of information leak. This calls for future research that addresses this aspect.

### B. ADDITIONAL RESULTS

#### B.1. PATHWAY VISUALIZATION

To better demonstrate the role of self-organizing pathway expansion scheme during optimization, we show more visualization results on the pathways of different classes. For the simplicity of viewing, we plot the most important group (*i.e.*, vertical coordinate) in each module (*i.e.*, horizontal coordinate). As shown in Fig. 3 (a), at the initial phase, two different classes (*i.e.*, the classes in the first and second columns) tend to utilize different pathways to extract the corresponding features. At the incremental phase, the novel class (*i.e.*, the class in the third column) tends to utilize the novel pathway to optimize the incremental features, and the whole pathway is similar to the semantically close class (*i.e.*, the class in the second column). The observation is consistent with the one from Fig. 5 in the main text.

#### B.2. THE IMPACT OF THE SET EPOCHS.

To explore the effect of the set epochs on the incremental performance, we conduct multiple experiments with different start epochs (*i.e.*  $e_1$ ) and end epochs (*i.e.*  $e_2$ ) in Equation 10 of the main text on CIFAR-100. As shown in Fig. 4 (b), the performance with larger start epoch is obviously worse than those with other values. In this case, due to the increase of initial classification accuracy with full pathways, it is more difficult to separate the class-specific pathway, which influences the overall performance. At the same time, the values of the end epoch have almost no effect on the incremental performance, which is more robust to the optimization process.

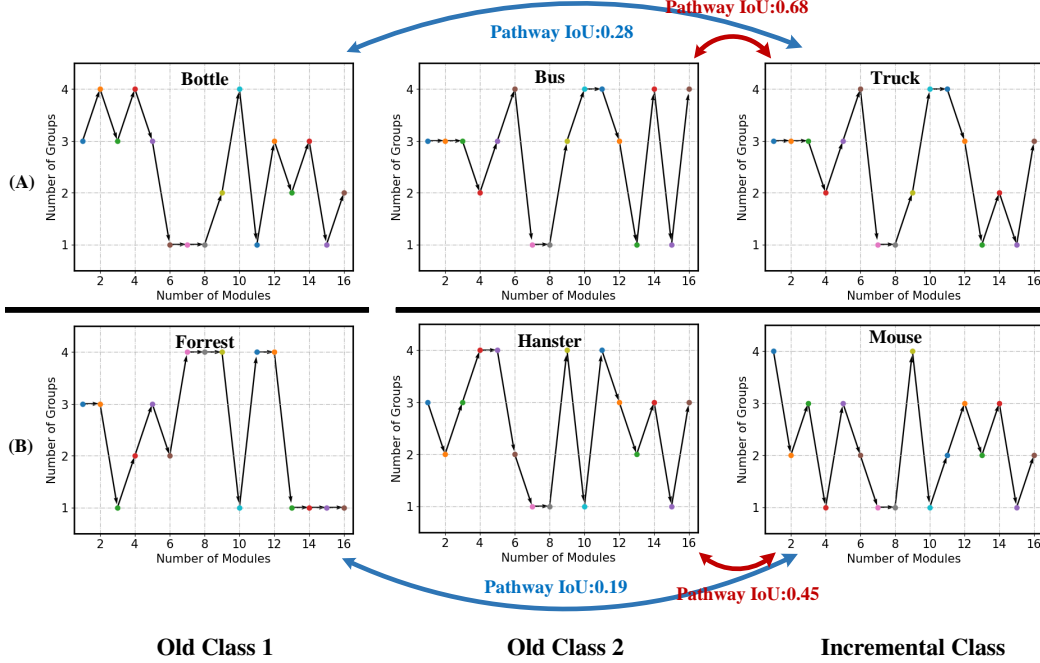


Figure 3: More visualization results on the pathway. The first and second columns represent the pathways of two old classes, which differ significantly in semantics. The third column represents the pathway of the incremental class, which is semantically closer to the one in the second column. For the simplicity of viewing, we plot the most important group (*i.e.*, L) in each module (*i.e.*, K). Pathway IoU represents the overlap rate of corresponding class-specific pathways.

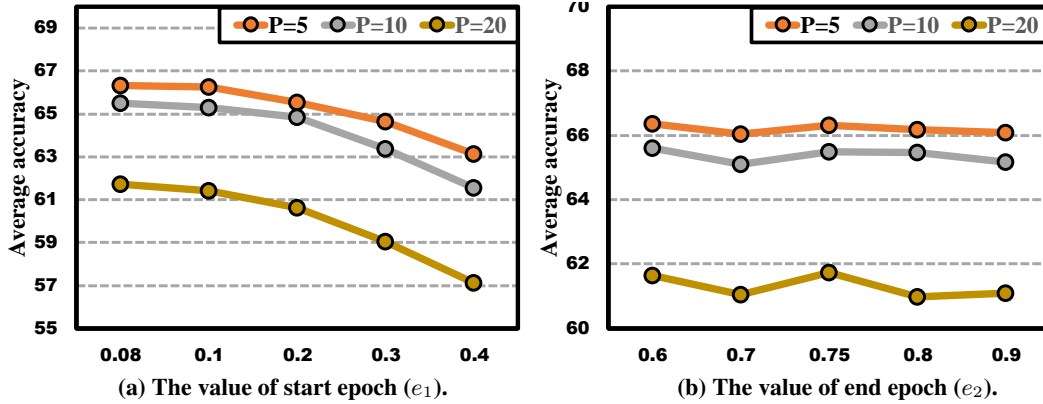


Figure 4: The impact of the values of the set epochs (*i.e.*,  $e_1$  and  $e_2$  in Equation 10 of the main text) on the incremental performance in the three-step strategy.

### B.3. FURTHER ANALYSIS OF MORE CIL METHODS.

In the comparative experiments of the main text, we compare with some classical CIL methods at two different settings, demonstrating that our method reduces the gap between the two settings. At the same time, most of the classical methods are not applicable to the NECIL settings, let alone the latest CIL methods. For example, we adapt the latest CIL method dynamic expandable network Yan et al. (2021) to the NECIL setting (*i.e.* NDER), and its performance is poor as shown in Table 6. Due to the lack of old samples, it is difficult to perform effective optimization with such large expanding parameters.

Method	CIFAR-100		
	5 phases	10 phases	20 phases
NDER	29.08	21.13	13.10
Ours	66.64	65.84	61.83

Table 6: Further analysis on the CIL method.

#### B.4. GENERALIZATION TO THE CIL SETTING

To further prove the effectiveness and generalization of our method, we introduce it into the CIL setting. As Douillard et al. (2020) is one of the SOTA methods in CIL setting, we modify its implementation with our self-organized pathway expansion scheme directly. As shown in Table 7, our method achieves average improvement of 2 points. Even if the effect of incremental samples on the overall performance is weakened by exemplars in CIL setting, our scheme still brings a boost to the existing method Douillard et al. (2020). It can be seen that our method has great potential for the CIL setting, which will serve as our future work.

#### B.5. COMPARISON WITH SOTA ON IMAGENET-FULL DATASET.

To better assess the overall performance of our scheme on larger dataset, we compare it to the SOTA of NECIL (PASS) and some classical methods of exemplar-based CIL (iCaRL, UCIR and PODNet) on ImageNet-Full.

As shown in Table 8, compared to the SOTA of non-exemplar methods (*i.e.*,  $E=0$ ), our method achieves average improvement of 2 points on the average accuracy. The performance of our method is comparable to the classical exemplar-based methods (*i.e.*,  $E=20$ ), which shows that our method further mitigate the gap between the two settings on larger dataset.

Method	CIFAR-100 (B50)	
	5 phases	10 phases
Podnet	64.88	63.05
Ours	66.64	65.84

Table 7: Comparisons of the average incremental accuracy (%) under the CIL setting.

Methods		ImageNet-Full
		$P=10$
$E=20$	iCaRL Rebuffi et al. (2017)	46.72
	UCIR Hou et al. (2019)	63.27
	PODNet Douillard et al. (2020)	64.17
$E=0$	iCaRL <sup>†</sup> Rebuffi et al. (2017)	32.43
	UCIR <sup>†</sup> Hou et al. (2019)	53.27
	PODNet <sup>†</sup> Douillard et al. (2020)	50.67
	PASS <sup>†</sup> Douillard et al. (2020)	55.90
	SSRE <sup>†</sup> Zhu et al. (2022)	58.12
	Ours	<b>60.20</b>

Table 8: Comparisons of the average incremental accuracy (%) with other methods on ImageNet-Full. P represents the number of phases and E represents the number of exemplars. Models with an asterisk <sup>†</sup> represent the reproduced results by this paper.

## REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

- 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)*, pp. 233–248, 2018.
- 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*, pp. 86–102. Springer, 2020.
- Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, and Cheng-zhong Xu. Dynamic channel pruning: Feature boosting and suppression. In *International Conference on Learning Representations*, 2018.
- 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*, pp. 831–839, 2019.
- 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.
- Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. In *International Conference on Learning Representations*, 2019.
- 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*, pp. 2001–2010, 2017.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.
- Yang Sui, Miao Yin, Yi Xie, Huy Phan, Saman Aliari Zonouz, and Bo Yuan. Chip: Channel independence-based pruning for compact neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.
- 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*, pp. 3014–3023, 2021.
- Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niranjan K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8715–8724, 2020.
- 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*, 2020a.
- 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*, pp. 6982–6991, 2020b.
- Fei Zhu, Zhen Cheng, Xu-yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34, 2021a.
- 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*, pp. 5871–5880, 2021b.

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