

# $M^2KD$ : Incremental Learning via Multi-model and Multi-level Knowledge Distillation

Peng Zhou<sup>1</sup>  
pengzhou@umd.edu

Long Mai<sup>2</sup>  
malong@adobe.com

Jianming Zhang<sup>2</sup>  
jianmzha@adobe.com

Ning Xu<sup>2</sup>  
nxu@adobe.com

Zuxuan Wu<sup>1</sup>  
zxwu@cs.umd.edu

Larry S. Davis<sup>1</sup>  
lsd@umiacs.umd.edu

<sup>1</sup> University of Maryland  
College Park, MD, USA.

<sup>2</sup> Adobe Research  
345 Park Avenue  
San Jose, CA, USA.

---

## Abstract

Incremental learning targets at achieving good performance on new categories without forgetting old ones. Knowledge distillation has been shown critical in preserving the performance on old classes. Conventional methods, however, sequentially distill knowledge only from the penultimate model, leading to performance degradation on the old classes in later incremental learning steps. In this paper, we propose a multi-model and multi-level knowledge distillation strategy. Instead of sequentially distilling knowledge only from the penultimate model, we directly leverage all previous model snapshots. In addition, we incorporate an auxiliary distillation to further preserve knowledge encoded at the intermediate feature levels. To make the model more memory efficient, we adapt mask based pruning to reconstruct all previous models with a small memory footprint. Experiments on standard incremental learning benchmarks show that our method improves the overall performance over standard distillation techniques.

## 1 Introduction

Deep neural networks perform well on many visual recognition tasks [1, 19, 23] given specific training data. However, problem arises when adapting networks to unseen categories while remembering seen ones, which is known as catastrophic forgetting [9, 17, 22]. To tackle this issue, there is a growing research attention on incremental learning where the new training data is not provided upfront but added incrementally. The target of incremental learning is to achieve good performance on new data without sacrificing the performance on old and it has been widely explored across different tasks such as classification [21, 51] and detection [54].

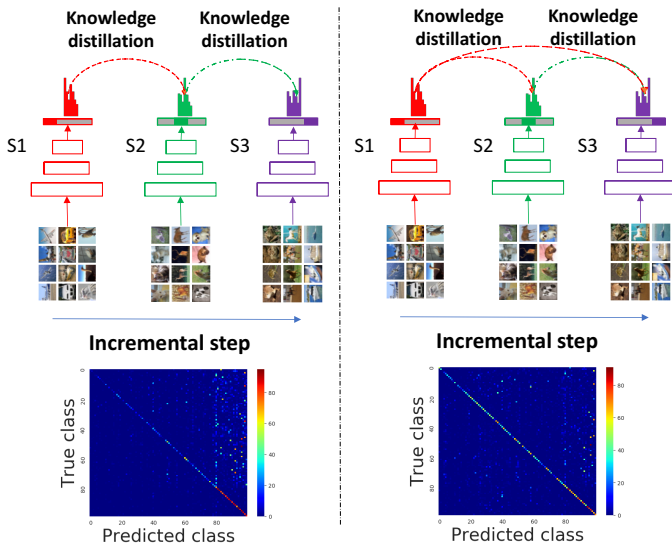


Figure 1: Concept overview. We propose to distill knowledge from all previous models efficiently to preserve old data information rather than sequentially applying distillation only to the last model. (e.g. using both S1 and S2 in S3 for distillation instead of sequentially using S1 for S2 and then S2 for S3). The confusion matrix is LWF-MC [51] on the left and our method on the right for the exemplar-free incremental setting.

To alleviate catastrophic forgetting in incremental learning, one possibility is to maintain a subset of old data to avoid over fitting on new data [9, 20, 51, 55]. However, an issue in practice is that when models embedded in a product are delivered to customers, they no longer have access to trained data for privacy purposes. To tackle the situation, a stricter exemplar-free setting was introduced in [21], which requires no exemplar set for previous categories and only distills previous knowledge from the current categories.

Prior methods typically apply knowledge distillation [13] sequentially during the incremental procedure to preserve previous knowledge. Since they apply distillation only to the penultimate model, it is difficult to maintain all past knowledge completely (the left side of Figure 1). From that observation, we propose using all the model snapshots. Prior knowledge is preserved better through our approach (the right side of Figure 1). However, saving all previous models may incur a great penalty in memory storage and without somehow compressing this historical information would not be practical. To address this, we reconstruct previous outputs using only “necessary” parameters during training.

To this end, we propose an end-to-end Multi-model and Multi-level Knowledge Distillation ( $M^2KD$ ) framework as depicted in Figure 2. We introduce a multi-model distillation loss which leverages the snapshots of all previous models to serve as teacher models during distillation, and then directly matches the outputs of a network with those from the corresponding teacher models. To make the pipeline more efficient, we adapt mask based pruning methods to reconstruct the previous models. We prune the network after each incremental training step and identify significant weights to reconstruct the model. This allows us to reconstruct previous models on-the-fly and utilize them as teacher models in our multi-model distillation. To further enhance the distillation process, we also include an auxiliary distillation loss to preserve more intermediate features of previous models. Additionally, our approach

addresses catastrophic forgetting in sequential distillation, and thus generalizes well for both exemplar based and exemplar-free settings.

To show the effectiveness of our approach, we evaluate our model on Cifar-100 [18] and a subset of ImageNet [19]. We achieve state-of-the-art performance for all the datasets in exemplar-free setting. We also show improvement when adapting to exemplar-based incremental learning and our exemplar-free setting outperforms iCaRL [50] with a 200 exemplar budget.

In summary, our contributions are three fold. First, we propose a multi-model distillation loss, which directly matches logits of the current model with those from the corresponding teacher models. Secondly, for efficiency, we reconstruct historical models via mask based pruning such that model snapshots can be reconstructed with low memory footprint. Experiments on standard incremental learning benchmarks show that our method achieves state-of-the-art performance in exemplar-free incremental setting.

## 2 Related Work

The ultimate goal of incremental learning is to achieve good performance on new data while preserving the knowledge about old data. Generally, two types of evaluation settings [4] have been considered. One is multi-head incremental learning which utilizes multiple classifiers at inference, and the other is single-head incremental learning which only utilizes one classifier at inference.

**Multi-head incremental learning.** The evaluation setting in this stream is that a specific classifier is selected during testing according to the tasks or categories. With this prior information, no confusion exists across different classifiers, and thus the target becomes how to adapt the old model for new tasks or categories. Research has been focused on utilizing an episodic memory to trace back previous tasks [0, 5, 24], or constraining the important weights on old tasks [17]. In addition, Mallya *et al.* [25, 26] learn a mask for pruning to further constrain the weights on old tasks. Hou *et al.* [14] distill the knowledge from the old model when adapting to new tasks. Different from this setting, we do not assume the task or category information is known during inference and follow the setting of single-head incremental learning. Even though we apply pruning in our approach, our goal is different from [25, 26] as the masks are utilized to reconstruct previous models and our approach requires no mask selection at inference.

**Single-head incremental learning.** Single-head evaluation uses only one classifier to predict both the old and the new classes. This setting is more challenging [4] compared to the multi-head counterpart because of the confusion between old and new categories. Knowledge distillation [3] is frequently utilized to preserve information. Li *et al.* [21] distill the knowledge from the penultimate model. Dhar *et al.* [6] introduce Grad-CAM [53] in the loss function. A relaxed setting is to introduce exemplar set [30] for the old data and match previous logits through distillation. Castro *et al.* [8] explore the balance between old and new data during training. Li *et al.* [20] focus on constructing exemplar set and Caselles *et al.* [0, 30] replay the seen categories with GANs [8]. Instead of saving exemplars, we save the parameters of previous models for reconstruction. With that, this paper can be considered a complement research direction. In fact, as knowledge distillation is an important component in these methods, they can potentially benefit from our approach as well. Additionally, Javed *et al.* [16] alleviate the bias in knowledge distillation by introducing a scaling vector to trained classifier, however, our approach is agnostic to classifier and achieves better

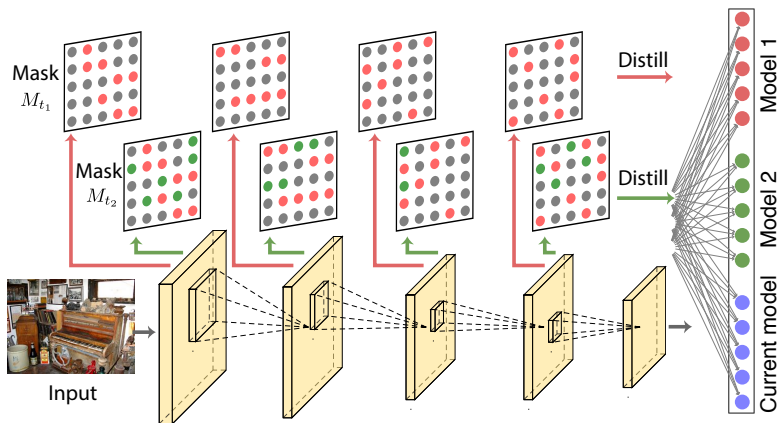


Figure 2: Framework overview. Given images from the current training data, we preserve previous knowledge directly from the reconstructed output through matching the logits with the corresponding model (e.g., model 1 and 2) and classifying the current data with its ground truth. The gray dots represent the weights to be trained on the current data. The red and green dots are fixed during training, denoting the weights retained from the first and second incremental step respectively.

performance.

**Network pruning.** Considerable research has explored this area to reduce network redundancy. Han *et al.* [10, 11] propose to compress network through quantization and Huffman coding. Yu *et al.* [66] compress the weights according to their scores. Other methods [15, 22, 29] explore compression for fast inference. In contrast to these methods, we leverage network redundancy and use pruning to reconstruct all previous models in incremental learning with low memory footprint.

## 3 Approach

We propose novel distillation losses to preserve previous information without introducing too much memory overhead (See Figure 2). The model is agnostic to the backbone architecture and generalizes well to both exemplar based and exemplar-free methods.

### 3.1 Multi-model Distillation

Single-head incremental learning consists of a sequence of incremental class inclusion process, referred to as incremental steps. Samples from a batch of new classes  $C_k$  are added at the  $k$ -th incremental step. For instance, 20 classes will be added per incremental step in a 20-class batch setting. Accordingly, the network assigns new logits (output nodes) for the incremental classes. At inference, the maximum logit score in the output is treated as the final decision.

The knowledge distillation used in incremental learning [21, 51] mainly aims to match the output of the current model to a concatenation of the penultimate model logits and ground truth labels. Formally, it optimizes the cross entropy for both the old and new logits,

$$L_D = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{C_o} s'_{ij} \log(s_{ij}) - \frac{1}{N} \sum_{i=1}^N \sum_{j=C_o+1}^C y_{ij} \log(s_{ij}), \quad (1)$$

where  $N$  and  $C$  denotes the number of samples and the total class number so far respectively, and  $C_o$  denotes the old classes.  $s_{ij}$  is the output score of the network obtained by applying Sigmoid function to the output logits for sample  $i$  at logit  $j$ .  $s'_{ij}$  denotes the old score obtained by the penultimate model.  $y_{ij}$  denotes the ground truth.

Treating the penultimate model as the teacher and applying this distillation sequentially helps preserve historical information, especially when no previous exemplar set is stored, which is the protocol for prior methods [3, 6, 21, 30]. However, the historical information will be gradually lost in this sequential pipeline as the current model must reconstruct all the prior information from the penultimate model alone. To address this limitation, we propose multi-model distillation, which directly leverages all previous models as our teacher model set. Since we mainly have current training data and labels for both settings, the network is more confident on current classes than old ones. Therefore, matching the previous logits of the current model directly with their corresponding old models preserves information better than always using the penultimate model. Formally, we minimize the cross entropy for the logits between the current model and corresponding teacher models from previous incremental steps,

$$L_{MMD} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{P-1} \sum_{j=C_{k-1}+1}^{C_k} s'_{ijk} \log(s_{ijk}) - \frac{1}{N} \sum_{i=1}^N \sum_{j=C_{P-1}+1}^C y_{ij} \log(s_{ij}), \quad (2)$$

where classes from  $C_{k-1} + 1$  to  $C_k$  belong to the  $k$ -th incremental step and  $P$  denotes the number of incremental steps. Classes from  $C_{P-1} + 1$  to  $C$  belong to the current categories.  $s_{ijk}$  is the output score of the current model for sample  $i$  at logit  $j$  in the  $k$ -th incremental step.  $s'_{ijk}$  denotes the output score of the  $k$ -th previous model.

At inference, we directly choose the maximum among the output logits, which acts as an ensemble of all the previous teacher models and the current model.

## 3.2 Auxiliary Distillation

Previous incremental learning methods preserve old class information through matching the final output. However, the features from intermediate layers also contain useful information. Inspired by the auxiliary loss in segmentation task [30], we propose an auxiliary distillation loss to preserve the intermediate statistics of previous models. Similar to using the final output to represent network statistics, the prediction made by lower level features also represents intermediate feature statistics. Following the main branch classification, we extract lower level features and use an auxiliary classifier to conduct classification based on intermediate features (See Figure 3).

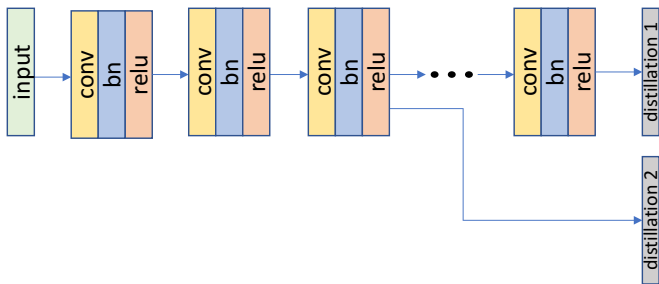


Figure 3: Illustration of auxiliary distillation. We extract the intermediate features and connect directly with an auxiliary classifier to preserve middle level knowledge.

Also, a multi-model distillation loss is added to this auxiliary classifier for the purpose of preserving prior lower level features, and a standard cross entropy loss is included for classifying the current data. Formally,

$$\begin{aligned}
 L_{AD} = & -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{P-1} \sum_{j=C_{k-1}+1}^{C_k} a'_{ijk} \log(a_{ijk}) \\
 & - \frac{\alpha}{N} \sum_{i=1}^N \sum_{j=1}^C y_{ij} \log(a_{ij}),
 \end{aligned} \tag{3}$$

where  $a'_{ijk}$  denotes the output score from previous auxiliary classifiers,  $a_{ijk}$  or  $a_{ij}$  is the output score of the auxiliary branch,  $\alpha$  is the ratio between the distillation and cross entropy loss. Notice that all the logits in ground truth labels are utilized in the classification cross entropy to enforce the correct prediction of current data.

The total loss function of the network becomes,

$$L_{total} = L_{MMD} + \lambda L_{AD}, \tag{4}$$

where  $\lambda$  is the ratio between the main classification multi-model distillation and the auxiliary classification distillation. This auxiliary classification branch is only used during training. At inference time, we only use the main branch classifier for prediction.

### 3.3 Model Reconstruction

One drawback of multi-model distillation in its original form is that it utilizes all previous models, requiring additional memory storage for the models. However, we observe that distillation aims to match logits. Therefore it is only necessary to preserve the outputs of previous networks, not the entire networks themselves. Our idea is to save only a small set of the necessary parameters from which we can approximate the output. By that way, all the models can be recovered on-the-fly without large memory penalty.

To determine the necessary parameters, we adapt mask based pruning [25] for model reconstruction. Specifically, after training each incremental step we sort the magnitude of weights in each layer, freeze the important ones to reach a specified pruning ratio, and use

the residual weights to train the next incremental class set. We repeat this procedure for all future incremental steps until all the incremental classes are included.

Formally, the output of a network with  $n$  convolutional layers is obtained from its classifier (the last layer) and features,

$$s = \Psi(f^{(n)}), \quad (5)$$

where  $\Psi$  denotes the classifier and  $f^{(n)}$  denotes the features in the  $n$ -th layer and can be generally written as

$$f^{(n)} = \sigma(w^{(n)}f^{(n-1)} + b^{(n)}), \quad (6)$$

where  $w$  and  $b$  are weights and biases respectively,  $\sigma$  denotes the activation function and  $f^{(0)}$  is the input.

We use a mask  $M$  to identify the important weights of each layer for all previous incremental steps. After each pruning procedure, we update the mask for the current incremental step, *e.g.*, using  $k$  to mark the weights for the  $k$ -th step. With the mask  $M_k$  for the  $k$ -th incremental step, we reconstruct the corresponding features by:

$$f_k^{(n)} = \sigma(w_k^{(n)}\delta(M_k^{(n)} \leq k)f_k^{(n-1)} + b_k^{(n)}), \quad (7)$$

where  $M_k^{(n)}$  denotes the mask in the  $n$ -th layer at incremental step  $k$ ,  $f_k^{(n)}$  denotes the feature in the  $n$ -th layer in  $k$ -th incremental step, and  $\delta$  denotes delta function.

With the saved biases, batch normalization and classifier parameters, we can reconstruct all previous models from the pre-updated model on-the-fly. The output of the  $k$ -th model is reconstructed by

$$s_k = \Psi_k(f_k^{(n)}), \quad (8)$$

where  $s_k$  and  $\Psi_k$  denote the output of the network and the classifier for the  $k$ -th incremental step respectively.

## 4 Experiments

We first evaluate our method in the exemplar-free setting. Then we extend our method to the exemplar-based setting. For more analysis, we also compare our memory cost with other methods.

### 4.1 Implementation Details

We use PyTorch for implementation. The network architecture follows prior works [9, 21, 31]: we use ResNet-32 [22] with input size  $32 \times 32$  for Cifar-100 and ResNet-18 with input size  $224 \times 224$  for iILSVRC-small. We extract the output of the second residual block for auxiliary distillation and empirically set  $\alpha$  to 0.5. We train 80 epochs for Cifar-100 and 60 epochs for iILSVRC-small. Following the setting in [31], we use the training batch size of 128 and the initial learning rate of 2.0 to train the model. The learning rate decays by a factor of 5 every 40 epochs for Cifar-100 and 20 epochs for iILSVRC-small. Weight decay with a factor of  $1e-5$  is applied for the first incremental step and 0 for the rest in our full model to ensure weights from previous models remain the same. We optimize the network using standard Stochastic Gradient Descent (SGD) with a momentum 0.9. The pruning ratio

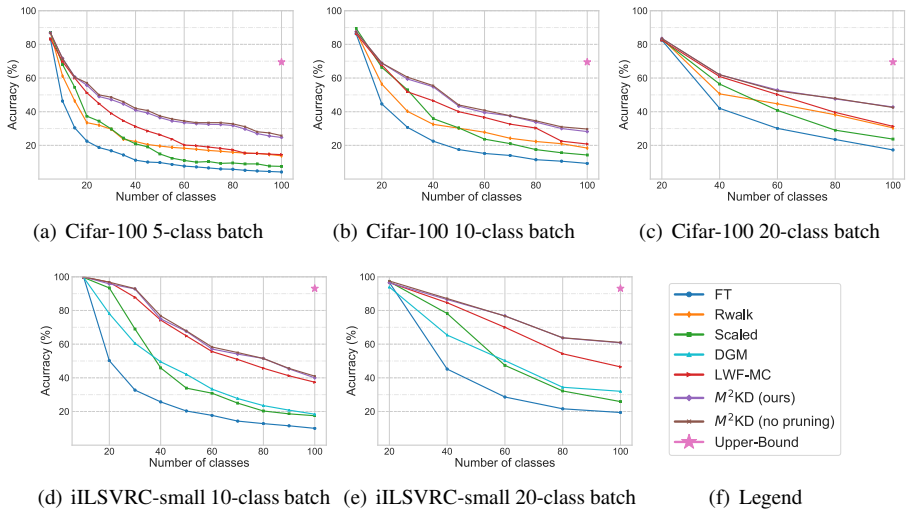


Figure 4: Performance on iLSVRC-small and Cifar-100 dataset in exemplar-free setting. (a) Top-1 accuracy on Cifar-100 (5-class batch). (b) Top-1 accuracy on Cifar-100 (10-class batch). (c) Top-1 accuracy on Cifar-100 (20-class batch). (d) Top-5 accuracy on iLSVRC-small (10-class batch). (e) Top-5 accuracy on iLSVRC-small (20-class batch).

is 0.75 for class batch less than 20 and 0.8 for 20 groups. After cutting off insignificant weights in the pruning stage, we fine tune the network for another 15 epochs.  $\lambda$  is set to be 1.0 to balance the losses. Only random horizontal flipping is applied as data augmentation for all experiments.

## 4.2 Datasets and Evaluation Metrics

The evaluation is conducted on iLSVRC-small [62] and Cifar-100 [48]. Dataset details are provided in **supplementary material**.

**Evaluation Metrics.** Following the same metrics in prior methods [21, 50], the top-1 classification accuracy is reported for Cifar-100 and top-5 classification accuracy is reported for iLSVRC-small.

## 4.3 Exemplar-free setting

We evaluate our methods in exemplar-free single-head setting. For evaluation, we also compare with state-of-the-art single-head approaches— **Scaled** [46], **DGM** [30], **Rwalk** [9], **LWF-MC** [34] — and the following baselines. (DGM hasn’t released code for Cifar-100 and thus is not compared in the Cifar-100 experiment.)

**FT:** A baseline approach that only applies cross entropy loss to fine-tune the penultimate model on new coming incremental classes. Knowledge distillation is not applied.

**$M^2KD$  (ours):** Our full model applying multi-model, auxiliary distillation along with pruning to save memory storage.

**$M^2KD$  (no pruning):** Our full model which directly loads all the previous snapshots for multi-model distillation.

**Upper-Bound:** The upper bound which directly trains all classes together.



Step	1	2	3	4	5		iLSVRC-small	Cifar-100
No pruning	<b>83.5</b>	<b>61.8</b>	<b>52.5</b>	<b>51.5</b>	42.1	LWF-MC	0	0
Ratio 0.6	82.9	59.6	52.2	46.5	40.1	iCaRL	68.0	9.4
Ratio 0.7	83.5	61.7	52.5	50.0	<b>42.8</b>	Ours	9.80	0.84
Ratio 0.8	83.5	58.5	52.0	49.3	42.0			
Ratio 0.9	83.0	58.0	49.7	47.3	39.9			

Table 1: Top-1 accuracy comparison among different pruning ratios on Cifar-100 (20 classes per incremental step).

Table 2: Memory compensation comparison (MB). Each entry is the additional memory requirement for methods across different datasets based on the memory footprint of LWF.

Figure 4 highlights our performance compared to state-of-the-art methods. For Cifar-100, our method consistently outperforms other methods from 5-class to 20-class batch per incremental step. The margin becomes larger as more incremental steps are added. This demonstrates the advantage of multi-model distillation as it avoids accumulating loss of historical information. Similar observation can be made when evaluating on iLSVRC-small. It is interesting to note that our model with pruning achieves comparable performance with the no-pruning version. This indicates the effectiveness of the pruning procedure in terms of saving memory while maintaining performance. Even though the residual active weights decrease gradually due to pruning, we still preserve the performance up to 20 incremental steps. To further analyze the performance, we provide ablation analysis in **supplementary material**.

#### 4.4 Analysis on pruning ratio

We compare the results corresponding to different pruning ratios to investigate the robustness of our approach. Table 1 summarizes the results. Marginal performance variation (around 3%) is observed for different pruning ratios. Even though a higher (0.9) pruning ratio affects the performance as the active weights decrease in the current incremental step and a lower (0.6) ratio affects the performance as available weights decrease in the future steps, the relatively trivial influence indicates that a large redundancy exists in the network architecture. Benefitting from it, our approach shows robustness to different pruning ratios.

#### 4.5 Exemplar Based Setting

Our approach can also be applied to exemplar based incremental learning methods which use distillation sequentially on the output of networks [3, 20, 31]. To evaluate our model in this setting, we add exemplar selection to our approach and compare with exemplar based methods.

**iCaRL [31]**: A prominent exemplar based incremental learning approach which constructs exemplar set for the old data according to the feature means and do distillation on the penultimate model. A nearest class mean classifier [28] is applied at inference.

**iCaRL aux**: Adding auxiliary distillation to iCaRL.

**iCaRL  $M^2KD$** : Change the original distillation function which only matches logits from the penultimate model to our multi-model distillation. Auxiliary distillation is also appended for a better performance.

The results are shown in Figure 5. With the introduction of multi-model and auxiliary distillation, the performance of iCaRL improves. It indicates that with direct access to all the previous models for distillation, the knowledge preserves better even with exemplar set.

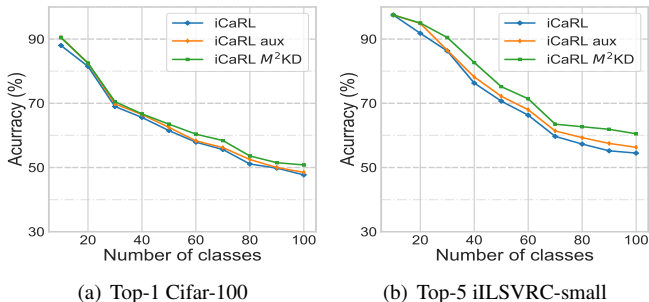


Figure 5: Performance comparison in exemplar based setting (10-class batch). (a) Top-1 accuracy performance on Cifar-100. (b) Top-5 accuracy performance on iILSVRC-small.

## 4.6 Memory Comparison

Starting from the memory footprint of LWF as our baseline, we compare the extra memory storage between exemplar based method such as iCaRL [51] and our approach. The memory is calculated in the 10-class incremental step setting for both iILSVRC-small and Cifar-100. For our approach, we directly calculate the storage difference between the last and the initial step. For iCaRL, the memory is approximately calculated by the average size of image for 2000 samples (*i.e.* the default exemplar size), and the compensation for saving the record of exemplar set. To optimize the memory consumption of iCaRL, we resize the images in iILSVRC-small to  $256 \times 256$  and compress to JPG with quality 95 to match their network input size during training.

Table 2 shows the memory compensation for different methods. It indicates that our approach has approximately  $7\times$  smaller memory compensation on iILSVRC-small and  $10\times$  smaller on Cifar-100 than iCaRL. On average, for each incremental step, our approach only takes 0.98 MB and 0.08 MB for iILSVRC-small and Cifar-100 respectively. The memory advantage to exemplar based methods might become larger as higher resolution images take more storage. See **supplementary material** for further overhead analysis.

## 5 Conclusion and Discussion

This paper presents a novel distillation strategy that mitigates catastrophic forgetting in single-head incremental learning setting. We introduce multi-model distillation which directly guides the model to distill knowledge from the corresponding teacher models. To further improve our performance, we incorporate auxiliary distillation to preserve intermediate features. More efficiently, we avoid saving all the model snapshots through reconstructing all previous models using mask based pruning algorithm. Extensive experiments on standard incremental learning benchmarks demonstrate the effectiveness of our approach. Incremental learning is still far from solved. A significant gap between one-step training versus incremental training still exists. It remains to be an open question how to reduce the confusion between different incremental steps especially without access to previous data, which might be a future exploration for our research.

**Acknowledgement** This work was partly funded by Adobe. The authors acknowledge the Maryland Advanced Research Computing Center (MARCC) for providing computing resources.

## References

- [1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018.
- [2] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Continual state representation learning for reinforcement learning using generative replay. *NeurIPS*, 2018.
- [3] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018.
- [4] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018.
- [5] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *ICLR*, 2019.
- [6] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyang Wu, and Rama Chelappa. Learning without memorizing. *arXiv preprint arXiv:1811.08051*, 2018.
- [7] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [9] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *ICLR*, 2014.
- [10] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *NeurIPS*, 2015.
- [11] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, et al. Dsd: Dense-sparse-dense training for deep neural networks. *ICLR*, 2016.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [14] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *ECCV*, 2018.
- [15] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *BMVC*, 2014.
- [16] Khurram Javed and Faisal Shafait. Revisiting distillation and incremental classifier learning. In *ACCV*, 2018.

- [17] 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. *PNAS*, 2017.
- [18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.
- [20] Yu Li, Zhongxiao Li, Lizhong Ding, Peng Yang, Yuhui Hu, Wei Chen, and Xin Gao. Supportnet: solving catastrophic forgetting in class incremental learning with support data. *arXiv preprint arXiv:1806.02942*, 2018.
- [21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 2018.
- [22] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *ICCV*, 2017.
- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [24] David Lopez-Paz et al. Gradient episodic memory for continual learning. In *NeurIPS*, 2017.
- [25] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 2018.
- [26] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *ECCV*, 2018.
- [27] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. 1989.
- [28] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*. 2012.
- [29] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *ICLR*, 2016.
- [30] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *CVPR*, 2019.
- [31] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.
- [32] 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. *IJCV*, 2015.

- [33] 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 *ICCV*, 2017.
- [34] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *CVPR*, 2017.
- [35] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, 2019.
- [36] Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S Davis. Nisp: Pruning networks using neuron importance score propagation. In *CVPR*, 2018.
- [37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.