Semantic Drift Compensation for Lifelong Learning (Supplementary Material)

A. Visualization of E-LwF and E-MAS

In Fig. 8 we show examples of the drift vectors which are estimated by SDC in the case of E-LwF and E-MAS to supplement Fig. 4 in the main paper.

B. Pre-trained Model without Birds

The results presented in Table. 1 in the main paper are based on ResNet18 pre-trained from the ImageNet dataset. As some of the categories of birds are present in the ImageNet, we conducted additional experiments. We removed 59 classes in total form the original dataset, including birds (e.g. macaw, flamingo, black swan) and similar species (e.g. cock, hen, king penguin), and then trained ResNet18 network on the constrained dataset. Table. 2 shows the average incremental accuracy results pre-trained from ImageNet without birds. It can be seen that there is no significant difference with fine-tuning, while slightly worse with the other three methods compared to Table. 1, where bird categories were used in the pre-trained network.

C. Results with Multi-similarity Loss and Angular Loss

A triplet loss is used in the main paper as a default metric loss function. Additionally, we investigated two newer versions of metric losses: Multi-similarity [8] and Angular loss [7] on CUB-200-2011 with a class-IL setting. The results are shown in Table. 3. We can see that the accuracy for the first task with the angular loss is 5.0% lower than for the triplet loss, while the multi-similarity starts with 4.0% higher accuracy. For E-FT method, a multi-similarity loss can achieve much better average incremental accuracy after training six tasks with a 13.3% improvement compared to the triplet loss. It is interesting to note that after adding our SDC, it achieves 56.1% after the final task, which is even better than other methods with regularization and SDC except for E-EWC+SDC. For the angular loss E-FT and E-FT+SDC present slightly lower results in comparison to the others regularized and regularized with SDC methods. Despite addressing some of the triplet loss function shortcomings, both of new losses obtain similar results for class-IL to the triplet loss used for all experiments in the main paper.

Table 2. Average incremental accuracy for CUB-200-2011 datasets with constrained pre-trained ImageNet.

Pre-trained ImageNet (w/o birds)	T1	T2	T3	T4	T5	T6
FT	79.1	33.5	23.2	17.3	14.3	10.0
E-FT	86.3	74.6	63.2	54.8	43.8	37.5
LwF	79.1	51.7	37.0	28.7	24.8	19.5
E-LwF	86.3	76.4	67.7	60.1	55.7	50.8
EWC	79.1	37.8	27.3	18.0	14.6	10.2
E-EWC	86.3	73.9	63.2	59.0	53.4	50.7
MAS	79.1	44.5	32.1	27.2	23.2	19.4
E-MAS	86.3	73.2	61.1	55.9	51.1	48.6

D. Confusion Matrix

We show confusion matrix of CUB-200-2011 and Flowers-102 dataset with Fine-tuning respectively in Fig. 9, for further insight of our SDC method. The left figures are the confusion matrices before applying SDC, the right ones are after applying SDC. We can see that our SDC method is able to compensate the forgetting of the previous tasks to some extent.

E. Experiments on VGG

To be able to compare our method to R-EWC and validate its generalization ability, we follow the protocol of Liu et al. [5] and implement our method on a VGG16 [6]. The CUB-200 dataset is divided into four equal tasks; the same setting as in Table.1. The comparison of different methods is shown in Fig. 10. We can see that our E-EWC surpasses EWC [3] and R-EWC [5] with clear superiority, improving with 30.1% and 22.1% respectively. SDC contributes an additional 1.6% gain.

F. Classification with Embedding Networks on Cars-196 Dataset

Cars-196 dataset [4] contains $16,185$ images of 196 cars classes. ResNet-18 [1] is adopted as the backbone network pretrained from ImageNet for Cars-196 dataset as well. We train our model with learning rate $1e{-}5$ for 100 epochs on cars, the other settings are the same as birds and flowers. Results are shown in Table. 4 after training the last task (T6). The conclusion is consistent with the CUB-200 and Flowers-102 datasets.

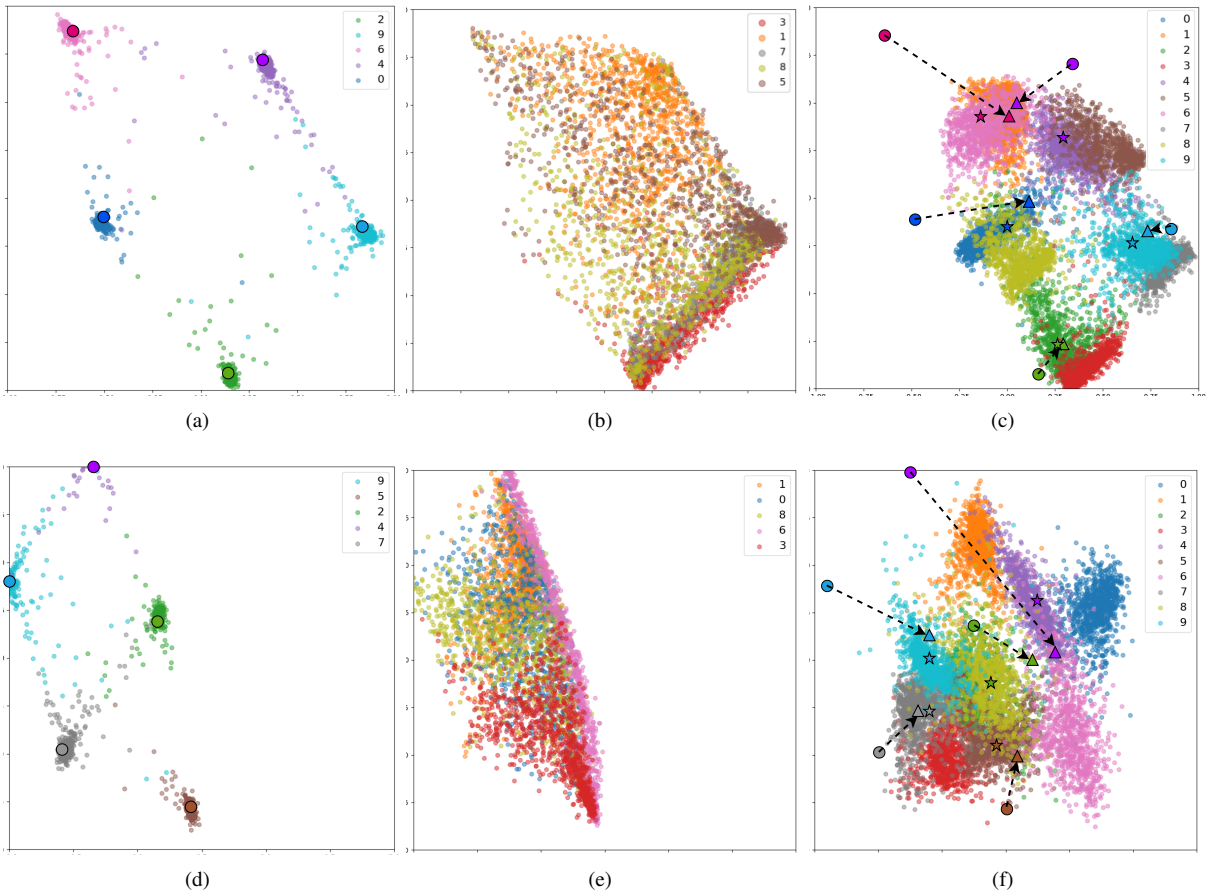

Figure 8. Examples of the drift vectors in the cases of E-LWF (top) and E-MAS (bottom). (a) and (d) represent the embedding of 5 classes of task 1 after training task 1; (b) and (e) represent the embedding of another 5 classes of task 2 after training task 1; (c) and (f) show the embeddings of 10 classes of two tasks together. The saved prototypes of the previous task(indicated by round) are estimated to new positions (indicated by triangle) by our proposed SDC in the new model which is observed to be closer to the real mean (indicated by star). The dotted arrows are the SDC vectors.

Table 3. Average incremental accuracy for CUB-200-2011 datasets with Multi-similarity and Angular loss loss.

	Multi-similarity						Angular					
	T1	T2	T3	T4	T5	T6	T1	T2	T3	T4	T5	T6
E-FT	88.1	74.4	65.5	59.8	52.2	50.7	79.1	61.7	50.9	48.1	40.9	40.5
E-FT+SDC	88.1	76.4	69.9	63.0	59.5	56.1	79.1	65.8	57.6	53.4	49.6	45.5
E-LwF	88.1	74.3	66.5	61.6	56.6	50.9	79.1	70.8	61.9	56.0	50.9	45.4
E-LwF+SDC	88.1	74.7	66.9	61.3	57.4	51.5	79.1	69.6	60.6	55.5	51.2	46.6
E-EWC	88.1	75.2	66.3	62.0	55.2	52.9	79.1	66.3	57.5	53.4	48.3	44.6
E-EWC+SDC	88.1	76.5	67.9	64.0	60.4	57.7	79.1	67.7	59.9	55.5	51.2	48.6
E-MAS	88.1	74.9	64.9	59.9	54.1	51.2	79.1	68.0	59.1	54.1	46.4	46.4
E-MAS+SDC	88.1	76.1	66.8	63.0	58.6	55.7	79.1	67.9	60.8	56.5	52.0	49.0

G. Experiments on CIFAR100 and ImageNet-Subset

We show the details of the average accuracy of our methods on CIFAR100 and ImageNet-Subset followed by the eleven-task evaluation protocol [2] in Table 5 (E-EWC+SDC is shown in Fig. 7 in the main paper). Batch nomalization is fixed after training the first task. It can be seen that E-LwF, E-EWC and E-MAS outperform E-FT on both datasets. Also we can observe that SDC improves the results of all methods even further except for E-LwF, especially for E-FT with 7.4% on CIFAR100, and 3.5% on ImageNet-Subset. Essentially, E-EWC and E-MAS indi-

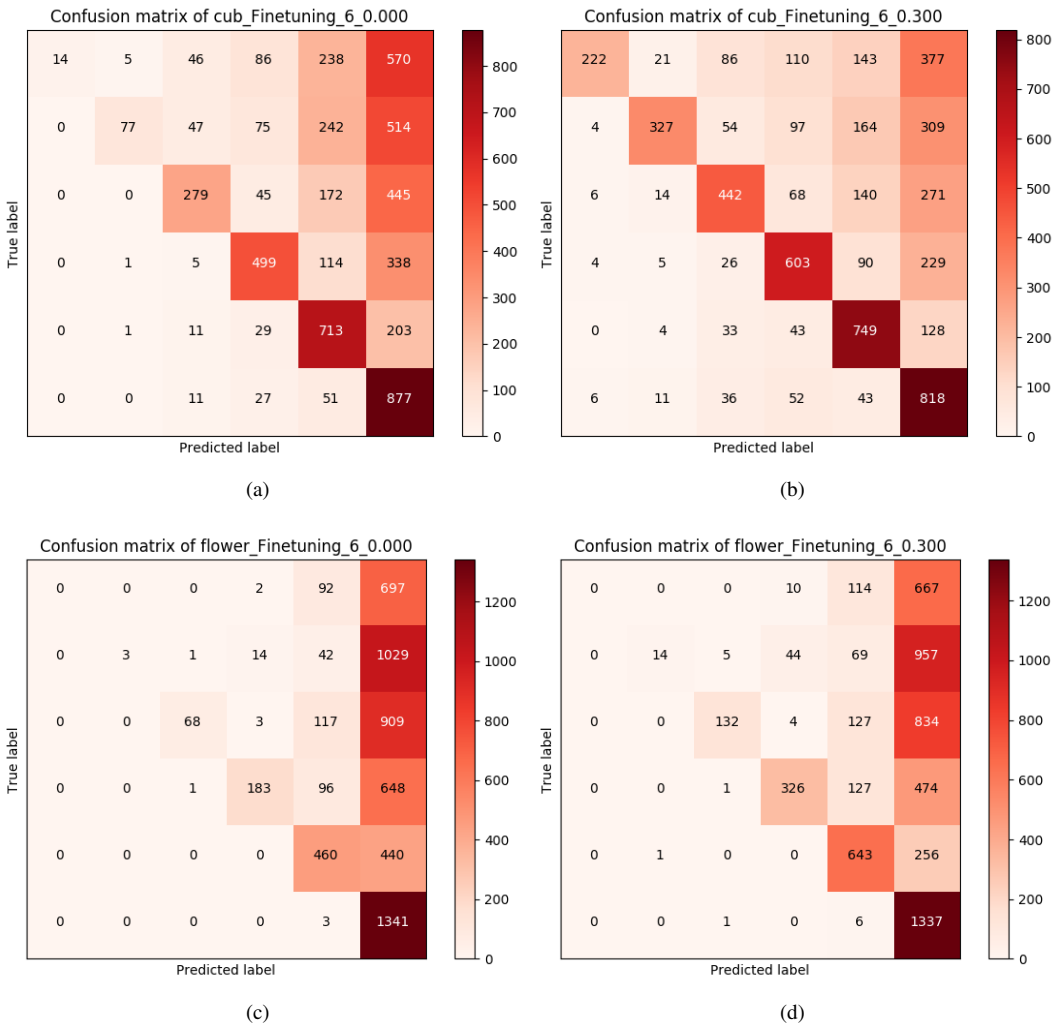

Figure 9. Confusion matrix of CUB-200-2011 and Flowers-102 with Fine-tuning method before applying SDC (a, c) and after applying SDC (b, d).

rectly limit the drift of the embedding by constraining the important weights, whereas E-LwF is directly constraining the embedding, which in the end results in less drift.

As discussed in the main paper, the good results of E-Fix for these more difficult datasets shows that continual learning methods without exemplars have difficulty outperforming this baseline (and even some methods which use exemplars like iCaRL). In Fig. 11 we also show the accuracy of each task after training the eleven tasks for E-Fix (in cyan) and E-EWC (in red). We can see that E-EWC always outperforms E-Fix except for the first task. It means even though the average accuracy of the eleven tasks with E-Fix and E-EWC is similar, freezing the first model does not have any positive forward transfer.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[2] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, pages 831–839, 2019. 2

[3] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proc. Nat. Acad. Sci. USA*, page 201611835, 2017. 1

[4] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[5] X. Liu, M. Masana, L. Herranz, J. Van de Weijer, A. M. Lopez, and A. D. Bagdanov. Rotate your networks: Better weight

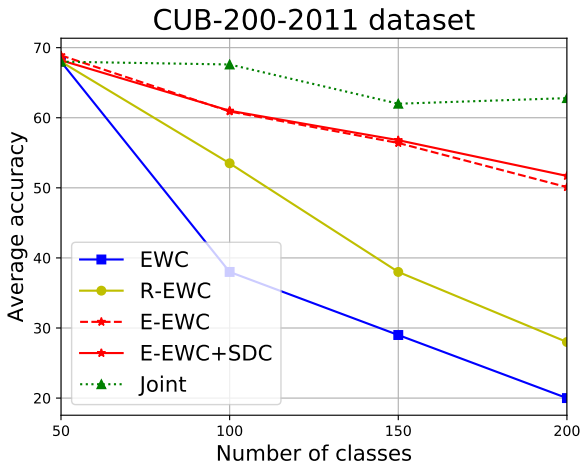

Figure 10. Comparison of four-task with VGG16 network.

Table 4. Average incremental accuracy for Cars-196 dataset.

	Cars-196					
	T1	T2	T3	T4	T5	T6
E-Pre	44.0	34.5	27.4	24.8	23.5	22.3
E-Fix	58.2	45.9	38.6	33.8	32.1	30.5
FT	67.5	33.0	24.2	19.6	15.0	13.6
E-FT	58.2	44.8	34.7	30.2	23.6	17.3
E-FT+SDC	58.2	**50.3**	**41.8**	**34.0**	**25.4**	**18.2**
LwF	67.5	40.3	33.3	30.1	26.7	21.9
E-LwF	58.2	48.2	40.9	36.2	34.2	32.0
E-LwF+SDC	58.2	**47.2**	**41.8**	**36.8**	**35.4**	**33.9**
EWC	67.5	30.8	25.8	19.9	16.5	15.6
E-EWC	58.2	47.0	39.6	35.1	32.9	30.7
E-EWC+SDC	58.2	**48.1**	**40.9**	**36.4**	**34.0**	**32.2**
MAS	67.5	37.1	27.7	22.9	20.2	17.0
E-MAS	58.2	**46.3**	38.3	33.6	31.4	28.8
E-MAS+SDC	58.2	**46.3**	**39.0**	**34.0**	**31.8**	**30.7**

Table 5. Average incremental accuracy for CIFAR100 and ImageNet-Subset.

	CIFAR100	ImageNet-Subset
	T11	T11
E-Fix	46.3	50.5
E-FT	37.4	47.4
E-FT+SDC	**44.8**	**50.9**
E-LwF	46.1	**51.5**
E-LwF+SDC	46.1	50.5
E-EWC	40.8	49.5
E-EWC+SDC	**46.1**	**51.5**
E-MAS	43.1	50.8
E-MAS+SDC	**46.3**	**51.2**

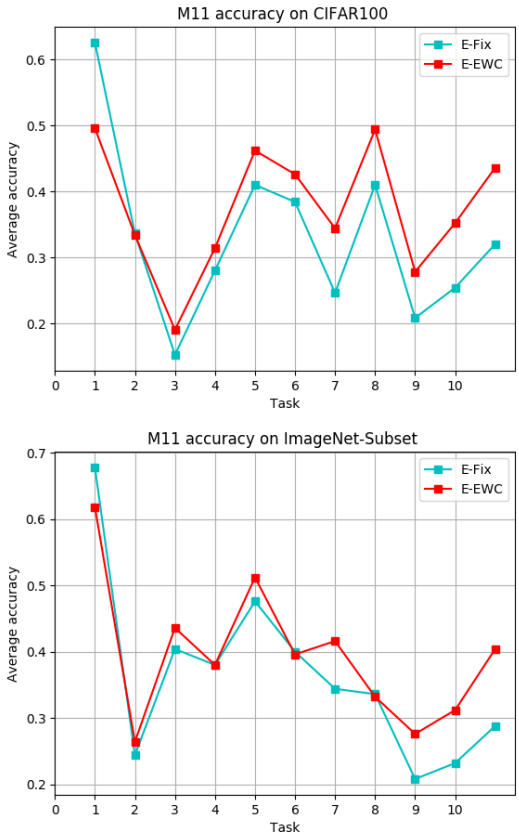

Figure 11. Accuracy of each of the eleven tasks with E-Fix and after training all tasks with E-EWC on CIFAR100 and ImageNet-Subset dataset.

consolidation and less catastrophic forgetting. In *Proc. ICPR*, 2018. 1

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[7] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *ICCV*, pages 2612–2620. IEEE, 2017. 1

[8] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019. 1