A. Appendix

SEAL is not sensitive to forgetting frequencies: In this experiment, we evaluate the sensitivity of SEAL to different forgetting frequencies. For this, we test multiple values for the number of epochs per generation E. The fewer the number of epochs in a generation, the more forgetting stages. We do not modify any other hyper-parameter. As summarized in Table 11, we observe that our method is not sensitive to the forgetting frequency and significantly improves over Normal training with any forgetting frequency. Please note that in this experiment, the maximum number of training epochs is different as each model is trained for E epochs for 10 generations.

Evaluating on Smaller Models: In this experiment, we evaluate both the in-domain and transfer learning performance of SEAL, LLF, and normal training using ResNet-18 on Tiny-ImageNet. We do not modify any of the hyper-parameters that were used for ResNet-50. We observe that SEAL outperforms both LLF and Normal training on this model as well (Table 8). Furthermore, we see the same transfer learning improvements as we saw with ResNet-50 (Table 12).

Generation	Normal	LLF	Ours
Gen=1	50.47	-	-
Gen=3	48.32	52.36	**53.69**
Gen=10	46.66	53.64	**54.75**

Table 8. Comparison of our method with normal training and LLF on Tiny-ImageNet with ResNet-18. Please note that the behavior of the first generation for all methods is the same. We outperform standard long training and LLF on ResNet-18 as well.

Comparison to Self-Distillation: We now compare our method to self-distillation approaches on CIFAR-100 [18] dataset. We include direct comparisons to published results in the state-of-the-art [28] and the classical Born Again neural networks (BAN) [10,43]. For each work, our experiments used the exact same model and hyperparameters (epochs, optimizers, learning rates, etc). Our method outperforms BAN consistently in each generation, and our best model over 10 generations outperforms both [10,43] (table 9). Furthermore, Pham et al. [28] is a state-of-the-art self-distillation result. Applying SEAL directly in their setting, our method outperforms them when trained for the same number of epochs (table 10). Our experiments show that our method surpasses the state-of-the-art self-distillation methods in both an inferior hyper-parameter setting, and a well-tuned hyper-parameter setting.

Generation	Furlanello et al. [10]	Yang et al. [43]	Ours
Gen=0	71.55	–	–
Gen=1	71.41	–	**72.83**
Gen=2	72.30	–	**73.53**
Gen=3	72.26	–	**73.88**
Gen=4	72.52	–	**74.18**
Gen=10 (Best)	72.61	73.72	**75.43**

Table 9. Comparison with [10,43] on CIFAR100 using ResNet-110. The last row lists the best accuracy of each method throughout 10 generations. The hyperparameters and baseline accuracies are adopted (and not changed) from [43] to ensure fairness.

Generation	Pham et al. [28]	SEAL (Ours)
Gen=0 (Teacher)	76.30	76.15
Gen=Last (Student)	77.32	**78.50**

Table 10. Comparison with Pham et al. [28] on CIFAR100 using ResNet-18. We train our method for the same number of epochs and under the same hyper-parameter setting as [28] to ensure fairness.

Gen	E=160 (default)	E=60	E=70	E=80	E=90	E=100	E=120	E=200
Gen1	54.37	53.33	53.62	53.59	53.66	53.84	53.92	54.07
Gen3	58.25	54.00	55.36	57.04	57.8	57.21	57.59	57.78
Gen10	**59.22**	**56.56**	**57.49**	**58.35**	**58.37**	**59.36**	**59.50**	**59.47**

Table 11. Dependency of SEAL on forgetting frequency in ResNet-50. Numbers in the columns indicate the number of epochs per generation E. Every E epochs, we perform gradient ascent for $k = \frac{E}{4}$ epochs. Each model is trained for $G = 10$ generations. We can see that our method has significant positive impact in every forgetting frequency.

Method	Tiny-ImageNet	Flower	CUB	Aircraft	MIT	Stanford Dogs
Normal	50.47	31.47	7.47	7.14	28.20	11.85
Normal (long)	46.66	19.11	5.36	4.80	21.71	8.17
LLF	53.64	31.66	7.19	6.09	25.67	11.64
SEAL (Ours)	**54.75**	**40.68**	**9.87**	**8.85**	**33.65**	**14.61**

Table 12. Transferring features learned from Tiny-ImageNet with ResNet-18 to other datasets using linear probe. Normal, and Normal (long) refer to $G = 1$ and $G = 10$ generations of training, respectively. LLF and SEAL were trained for $G = 10$ generations. Our method, after $1,600$ epochs, surpasses both LLF and normal training. This demonstrates that our method learns much more generalizable features as compared to Normal training or LLF.