Table 1. Comparison on more architectures. # Re-Init refers to the number of re-initializations. We use RIFLE-A for most architectures except for DenseNet-169.

	Inception-v3		
Datasets	Fine-tuning	RIFLE	# Re-Init
Stanford-Dogs	77.07±0.37	77.75±0.26	3
FGVC-Aircraft	82.80±0.49	84.31±0.35	3
Indoor-67	68.63±0.37	70.47±0.62	3
DTD	61.31±0.54	63.32±0.52	3
	MobileNet-v2		
Datasets	Fine-tuning	RIFLE	# Re-Init
Stanford-Dogs	75.42±0.27	76.03±0.26	3
FGVC-Aircraft	83.34±0.18	84.66±0.25	3
Indoor-67	72.68±0.26	73.34±0.47	3
DTD	62.98±0.36	64.42±0.24	3
	DenseNet-169		
Datasets	Fine-tuning	RIFLE	# Re-Init
Stanford-Dogs	80.71±0.37	80.98±0.28	2
FGVC-Aircraft	84.86±0.19	85.44±0.35	2
Indoor-67	76.59±0.55	77.36±0.37	2
DTD	64.52±0.25	67.15±0.19	2
	EfficientNet-v3		
Datasets	Fine-tuning	RIFLE	# Re-Init
Stanford-Dogs	86.22±0.29	86.41±0.01	2
FGVC-Aircraft	84.49±0.36	86.14±0.22	3
Indoor-67	77.31±0.28	77.86±0.20	2
DTD	69.54±0.19	71.36±0.68	3

Table 2. Comparison over more tasks. # Re-Init refers to the number of re-initializations.

	Object Detection		
Datasets	Fine-tuning	RIFLE	# Re-Init
Sheep (mAP)	0.5279	0.5309	3
Chair (mAP)	0.3658	0.3699	3
TV-Monitor (mAP)	0.6102	0.6135	3
	Semantic Segmentation		
Datasets	Fine-tuning	RIFLE	# Re-Init
CrackForest (AUROC)	0.8270	0.8360	2
Penn-Fudan (F1-Score)	0.7699	0.7764	2

1. Analysis of the Period of Random Re-initialization

We do additional ablation analysis with various settings on the period of random re-initialization. The analysis is carried out using CUB-200-2011 dataset with a training procedure of 48 epochs. We find that, performing RIFLE in the early or middle stages of the training procedure would be better than doing random re-initialization in the latter stage. In detail, for the first case we re-initialize the FC layer with an exponential increasing number of re-initialization periods. That is, the FC layer is re-initialized at epoch 3, 6, 12 and 24 if the total number of re-initializations is 4. Analogously, we re-initialize the FC layer at epoch 24, 36, 42 and 45 for evaluating the second case. Results show that, we obtain the top-1 accuracy of 80.99%(-0.14%) for the case of increased period, slightly worse than original RIFLE (81.13%). However, if we re-initialize the FC layer only twice, using an increased period gets the accuracy of 81.25%(+0.12%), slightly better than the original version. This indicates that there may exist an optimal selection of the re-initialization periods. However, using the decreased period usually hurts the performance, especially when an increased number of re-initializations are performed. For example, the top-1 accuracy of using decreased period is 80.65%(-0.48%) and 79.22%(-1.91%) if we re-initialize 2 and 3 times, respectively. If the number of re-initializations increases to 4, the accuracy dramatically drops to 49.36%, because the FC layer is not able to converge (only 3 epochs left) after the last re-initialization.

We also evaluate RIFLE with adaptive re-initialization, with respect to the number of epochs and the loss. We find that, regardless of the RIFLE procedures (adaptively triggered or predefined), the more often re-initialization takes place in the late stage of training, the lower top-1 accuracy will be. In detail, we design Ada-RIFLE, which adaptively determines the opportune moment of re-initialization according to the training loss. Specifically, we empirically set a threshold value L_t of the training loss computed using the standard cross entropy loss function. The FC layer is re-initialized when the current training loss is lower than L_t. In order to prevent the FC layer from under-fitting, we set a closing time E_c, meaning that no re-initialization will be performed after E_c epochs. E_c is set to 0.75 times the total number of epochs, that is 36 in our experiment. Results on the CUB-200-2011 dataset show that further improvements are obtained by a proper selection of L_t. We test choices of L_t from [0.02, 0.05, 0.1, 0.2]. The top-1 accuracy of these settings are 81.36%(+0.26%), 81.44%(+0.31%), 81.01%(-0.12%) and 80.95%(-0.18%). We investigate the internal heuristic process of these experiments. When using the best choice of 0.05, the FC layer is re-initialized 3 times, at epoch 18, 24 and 30 respectively. This result indicates that a better choice to re-initialize FC may be at the middle stage of the training process. We also show the detail of $L_t = 0.2$. Since the loss threshold is larger, more re-initializations are performed in this case, at epoch 15, 19, 22, 25, 28, 31 and 34. Despite the fact that the FC layer is easy to converge, this may bring negative effects to the learning of the feature extractor.

2. Experiments on More Architectures

We further evaluate RIFLE on more modern architectures including Inception-v3, MobileNet-v2, DenseNet-169 and EfficientNet-b3, over four datasets covering animals, objects, textures and scenes. We run all experiments three times and report the average accuracy. As shown in Table 1, most of the tasks significantly benefit from RIFLE.

We also implement additional experiments for object de-

tection with Fast-RCNN using ResNet-50+FPN as backbone and semantic segmentation with the DeepLab-V3 model. We used L^2+RIFLE for transfer learning and COCO datasets as the source dataset to obtain the pre-trained weights. As shown in Table 2, for both object detection and semantic segmentation tasks, L^2+RIFLE performs better.

3. Advanced Simulation Experiments

For the asymptotic studies, we evaluate RIFLE and fine-tuning approach with noise on W_1 (as the oracle to generate target datasets). The noise scale is set to 5% of the average scale of W_1. We run the same simulation experiment as in the main paper and get similar result that, OT(W_1^*, W_1)=0.1257 $\leq$ OT($W_1^{'}$,W_1)=0.1489.