D^{3}Former: Supplementary Materials

Training Details

Zeroth phase: The zeroth phase training starts with 10 warm up epochs for small scale datasets, and 20 epochs for large scale datasets.

Incremental phases: In each incremental phase, the new classes classifier weights are initialized following weight imprinting introduced in [2], old classes classifier weights are frozen. The learning rate starts from 2.5e-4 for the feature extractor and 2.5e-3 for the classifier. Both learning rates follow a cosine annealing scheduler that decays the weight till it reaches zero at the final epoch. The number of epochs for each phase is 250 in case of 10 classes per task and 5 classes per task, while for 2 classes per task the number of epochs is kept at 150.

Knowledge distillation factor λ is increased every phase as a factor of number of classes as follows:

$$\lambda_t = \lambda_{t-1} \times \sqrt{\frac{B+C}{C}}, \tag{1}$$

Where B is the number of base classes, and C is the number of new added classes every phase. The classes exemplars are chosen following the same herding method of [3].

Effect of Augmentations

NesT uses augmentations such as Mixup, RandomErasing and RandAugment. These augmentations have been shown to be useful in stabilizing training and improve performance of hybrid ViTs [1, 4]. The importance of these augmentations has also been discussed in the NesT paper. We show the effect of these augmentations when used in the incremental phases.

Table 1. Effect of augmentations when used in incremental phases of 5 tasks setting for CIFAR100

Augmentations	**Average accuracy**
With Mixup	71.85
With Randaug, RandomErasing	71.82
With all augmentations	72.23

Qualitative analysis

Figure 1 shows some qualitative results in the form of Grad-CAMs with increasing number of incremental tasks. It is observed that the model does not forget much and makes use of the discriminatory regions in an image to make the correct prediction.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *In ICCV*, 2021. 1

[2] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *In CVPR*, pages 5822–5830, 2018. 1

[3] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *In CVPR*, 2017. 1

[4] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan O. Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. *In AAAI*, 2022. 1

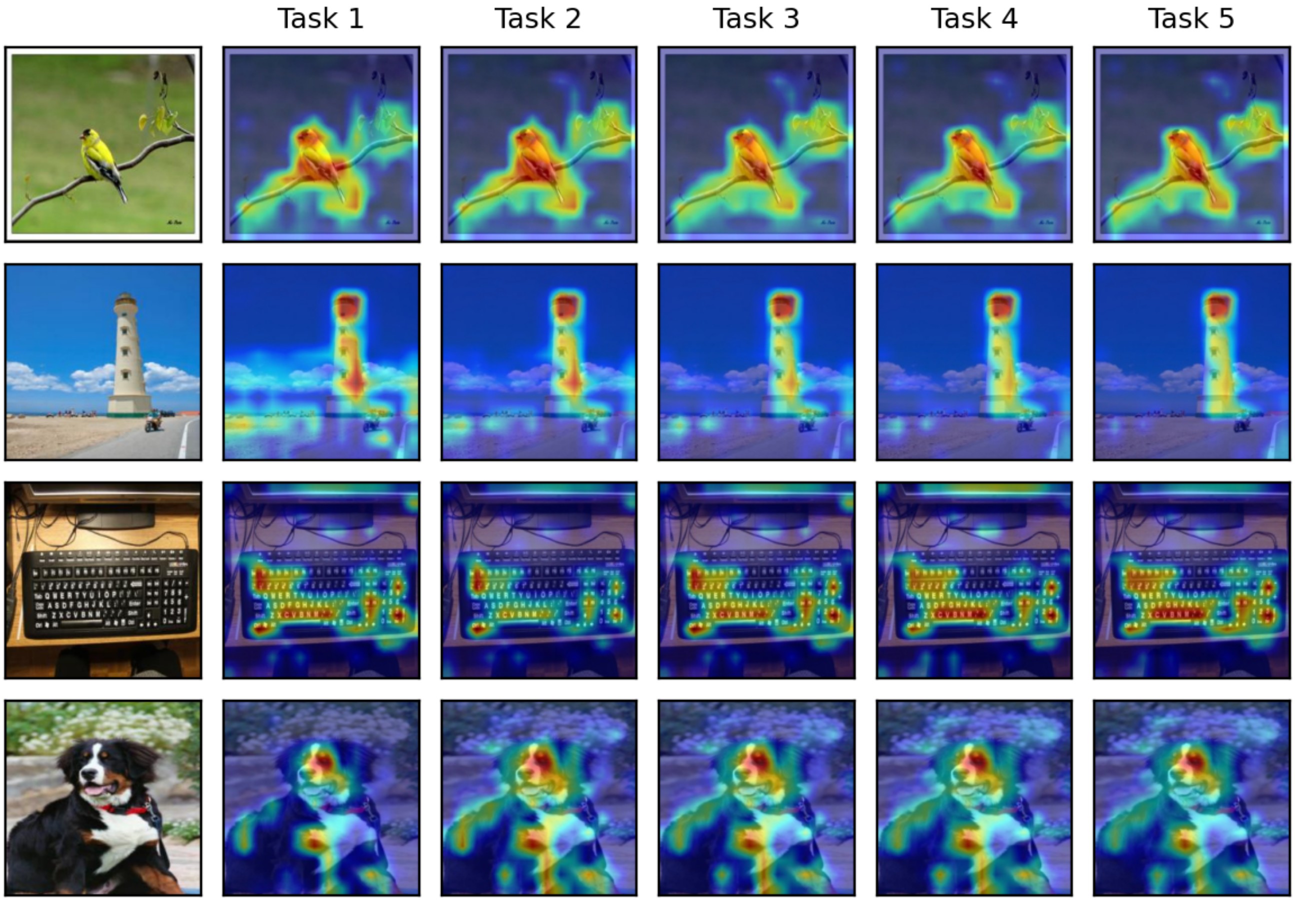

Figure 1. Grad-CAMs for images from ImageNet subset-100 as incremental learning progresses. This shows that Grad-CAM distillation helps D^{3}Former maintain attention on discriminative patches. (*figure best viewed with zoom-in*)