Appendices

A. Batch normalization calibration

As described in [10], if we want to use the extracted sub-net from a weight-sharing supernet directly, batch normalization (BN) requires calibration because of inconsistency between training and testing. At training time, input features of BN layers are normalized with mean and variance of the current mini-batch. While at test time, the global statics of mean and variance which are the moving average of corresponding counterpart in training steps are adopted for batch normalization. We vary the architecture of student branch during training, hence the global statics can not be utilized for any specific sub-net extracted from supernet. Concretely, after extracting a sub-net, we fix its weight and only update the global statics in each BN layer with 1000 training steps.

B. ImageNet semi-supervised classification

Following [3]. At training time, we apply random crops with resize to 224 $\times$ 224 pixels and random flips. At test time, the images are resized to 256 pixels, after which a 224 $\times$ 224 center crop is applied. We optimize the loss using SGD with Nesterov momentum. We use a batch size of 256, a momentum of 0.9. The learning rate and number of epochs are selected from $\{$0.01,0.005,0.02,0.05,0.1$\}$ and $\{$30,50,80$\}$ according to the performance on local validation set. The weight decay is not used in this setting.

C. Architectures for classification tasks

Model	Params	Depth	Width
R18	11.7M	[2, 2, 2, 2]	[64, 64, 128, 256, 512]
R34	21.6M	[3, 4, 6, 3]	[64, 64, 128, 256, 512]
R50	25.5M	[3, 4, 6, 3]	[64, 64, 128, 256, 512]
R101	44.7M	[3, 4, 23, 3]	[64, 64, 128, 256, 512]
Group	Params	Depth	Width
1G$\sim$2G	14.7M	[2, 2, 5, 4]	[48, 48, 96, 192, 384]
2G$\sim$3G	19.5M	[3, 2, 7, 3]	[48, 64, 96, 192, 512]
3G$\sim$4G	33.5M	[4, 4, 5, 4]	[32, 48, 128, 192, 640]
4G$\sim$5G	37.0M	[4, 2, 11, 3]	[32, 48, 128, 256, 640]
5G$\sim$6G	43.4M	[4, 6, 21, 4]	[32, 64, 96, 192, 640]
6G$\sim$7G	45.4M	[4, 6, 23, 4]	[32, 64, 128, 192, 640]
7G$\sim$8G	55.8M	[2, 2, 23, 4]	[48, 48, 96, 256, 640]

Table 1. Network architectures searched for image classification task in each budget group. Information about standard ResNets$\{$18,34,50,101$\}$ is also reported. Note that ResNet18 and ResNet34 are composed of *basic block*, while ResNet50 and ResNet101 are composed of *bottleneck*.

D. Transfer to other classification datasets

D.1. Details of datasets

In section 4.3, we transfer DATA to more diverse classification tasks in VTAB benchmark [11]. These tasks include CIFAR-10/100 [5], Oxford-IIIT Pet [8], Oxford Flowers-102 [7], DMLab [1], EuraSAT [4], CAMELYON [9], DTD [2], and smallNORB [6]. For the last five datasets, following the VTAB-1k [11] setting, we only use 1k examples per task to evaluate adaption with limited data. We summarize them in Table 2

Dataset	Train size	Classes
CIFAR-10	50,000	100
CIFAR-100	50,000	10
Oxford-IIIT Pet	3,680	37
Oxford Flowers-102	1,020	102
DMLab	1,000	6
EuraSAT	1,000	10
CAMELYON	1,000	2
DTD	1,000	47
smallNORB	1,000	5

Table 2. Description of datasets used in transferring to more diverse classification tasks.

D.2. Training details

For CIFAR-10, due to its low-resolution, we follow its common setting. Concretely, at training time, we apply padding to images to get them into 36 $\times$ 36 and random crop to 32 $\times$ 32. At test time, we don't apply any transformation. For the others, we follow similar procedure in semi-supervised classification. Specifically, at training time, we apply random crops with resize to 224 $\times$ 224 pixels and random flips. At test time, images are resized to 256 pixels, after which a 224 $\times$ 224 center crop is applied. After applying the augmentations, we optimized the loss using SGD with Nesterov momentum for 2000 steps with a batch size of 512 and a momentum of 0.9. The learning rate and weight decay are selected from $\{$0.001, 0.01, 0.03, 0.05, 0.1.$\}$ and $\{10^{-5}, 10^{-4}, 10^{-3}\}$ as well as no weight decay on local validation dataset.

References

[1] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab. abs/1612.03801, 2016. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 1

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207. IEEE, 2018. 1

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[6] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR*, 2004. 1

[7] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008. 1

[8] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 1

[9] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*. Springer, 2018. 1

[10] Jiahui Yu and Thomas Huang. Universally slimmable networks and improved training techniques (supplementary). 2019. 1

[11] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1