Appendix for *DTL: Disentangled Transfer Learning for Visual Recognition*

Implementation Details

Data Augmentation

Following Lian et al. (2022); Zhang, Zhou, and Liu (2022), we directly resize the input image to 224 $\times$ 224 for VTAB-1K (Zhai et al. 2019) benchmark. In few-shot learning and domain generalization experiments, we follow Zhang, Zhou, and Liu (2022) to apply color-jitter and RandAugmentation for fine-tuning, and then resize the input images into 256 $\times$ 256 with center crop 224 $\times$ 224 to conduct evaluation.

Fine-tuning Details

Following (Lian et al. 2022; Zhang, Zhou, and Liu 2022), we take AdamW (Loshchilov and Hutter 2019) as the optimizer, with weight decay 0.05. The cosine scheduler is adopted to decay the learning rate. By default, all pre-trained models are fine-tuned by 100 plus 10 warm-up epochs with batch size 32. Following Lian et al. (2022), the learning rate is selected according to the accuracy on the validation set for different datasets. Note that unlike previous methods (Zhang, Zhou, and Liu 2022; Lian et al. 2022), except for the standard data augmentation, we do not use any additional tricks such as mixup (Zhang et al. 2018), cutmix (Yun et al. 2019) or label smoothing (Szegedy et al. 2016) to boost the recognition accuracy.

In CSN, the d' in linear mapping, is 2 for ViT and 4 for Swin-B. β in Swish is fixed to 100. We set M of DTL and DTL+ as 7 for the ViT backbone to adapt half of the later blocks' output. For Swin-B, there are in total 24 blocks, in this case similarly M is set to 15 for adapting roughly half of the later blocks. Finally, for all models, we take the average of all patch tokens as the input of task-specific classification head to generate predictions.

References

Lian, D.; Zhou, D.; Feng, J.; and Wang, X. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, 109–123.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 1–18.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Yun, S.; Han, D.; Chun, S.; Oh, S.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision*, 6022–6031.

Zhai, X.; Puigcerver, J.; Kolesnikov, A.; Ruyssen, P.; Riquelme, C.; Lucic, M.; Djolonga, J.; Pinto, A. S.; Neumann, M.; Dosovitskiy, A.; et al. 2019. A large-scale study of representation learning with the Visual Task Adaptation Benchmark. *arXiv preprint arXiv:1910.04867*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. *mixup*: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 1–13.

Zhang, Y.; Zhou, K.; and Liu, Z. 2022. Neural prompt search. *arXiv preprint arXiv:2206.04673*.