A Detailed Descriptions for the Evaluation Datasets

A.1 Image Classification

We show the detailed descriptions of image classification as follows. The train/val/test splits and the classes are shown in Table 1.

Dataset	Description	#Classes	Train size	Val size	Test size
Fine-Grained Visual Classification (FGVC)					
CUB-200-2011 [31]	Fine-grained bird species recognition	200	5,394*	600*	5,794
NABirds [29]	Fine-grained bird species recognition	55	21,536*	2,393*	24,633
Oxford Flowers [27]	Fine-grained flower species recognition	102	1,020	1,020	6,149
Stanford Dogs [20]	Fine-grained dog species recognition	120	10,800*	1,200*	8,580
Stanford Cars [8]	Fine-grained car classification	196	7,329*	815*	8,041
Visual Task Adaptation Benchmark (VTAB-1k) [34]					
CIFAR-100 [21]	Natural	100	800/1,000	200	10,000
Caltech101 [7]		102			6,084
DTD [4]		47			1,880
Flowers102 [27]		102			6,149
Pets [28]		37			3,669
SVHN [26]		10			26,032
Sun397 [32]		397			21,750
Patch Camelyon [30]	Specialized	2	800/1,000	200	32,768
EuroSAT [14]		10			5,400
Resisc45 [3]		45			6,300
Retinopathy [12]		5			42,670
Clevr/count [19]	Structured	8	800/1,000	200	15,000
Clevr/distance [19]		6			15,000
DMLab [1]		6			22,735
KITTI/distance [9]		4			711
dSprites/location [25]		16			73,728
dSprites/orientation [25]		16			73,728
SmallNORB/azimuth [22]		18			12,150
SmallNORB/elevation [22]		9			12,150
General Image Classification Datasets					
CIFAR-100 [21]	General image classification	100	50,000	-	10,000
ImageNet-1K [6]		1,000	1,281,167	50,000	150,000
Robustness and Out-of-Distribution Dataset					
ImageNet-A [17]	Robustness & OOD	200	7,500		
ImageNet-R [15]		200	30,000		
ImageNet-C [16]		1,000	$75 \times 50{,}000$		

Table 1: The statistics of the various datasets. *: Since there are no public train/val splits in these datasets, we follow VPT [18] for random train/val split. This table is partially borrowed from VPT [18].

FGVC. Following VPT [18], we employ five Fine-Grained Visual Classification (FGVC) datasets to evaluate the effectiveness of our proposed SSF, which consists of CUB-200-2011 [31], NABirds [29], Oxford Flowers [27], Stanford Dogs [20] and Stanford Cars [8].

VTAB-1k. VTAB-1k benchmark is introduced in [34], which contains 19 tasks from diverse domains: i) Natural images that are captured by standard cameras; ii) Specialized images that are captured by non-standard cameras, *e.g.*, remote sensing and medical cameras; iii) Structured images that are synthesized from simulated environments. This benchmark contains a variety of tasks (*e.g.*, object counting, depth estimation) from different image domains and each task only contains 1,000 training samples, thus is extremely challenging.

General Image Classification Datasets. We also validate the effectiveness of SSF on general image classification tasks. We choose the CIFAR-100 [21] and ImageNet-1K [6] datasets as evaluation datasets, where CIFAR-100 contains 60,000 images with 100 categories. ImageNet-1K contains 1.28M training images and 50K validation images with 1,000 categories, which are very large datasets for object recognition.

A.2 Robustness and OOD

ImageNet-A is introduced in [17], where 200 classes from 1,000 classes of ImageNet-1K are chosen and the real-world adversarial samples that make the ResNet model mis-classified are collected.

ImageNet-R [15] contains rendition of 200 ImageNet-1K classes and 30,000 images in total.

ImageNet-C [16] consists of the corrupted images, including noise, blur, weather, *etc.* The performance of model on ImageNet-C show the robustness of model.

A.3 Detection and Segmentation

We also conduct experiments on downstream tasks beyond image classification, such as object detection, instance segmentation and semantic segmentation. We employ the COCO dataset [23] for evaluation based on mmdetection [2] framework for the object detection and instance segmentation. COCO contains 118K training images for training and 5K images for validation, which is one of the most challenging object detection datasets. We use Mask R-CNN [13] with Swin Transformer backbone to perform our experiments, following the same training strategies as Swin Transformers [24]. For semantic segmentation, we employ the ADE20K dataset [35] for evaluation based on mmsegmentation [5] framework. ADE20K contains 20,210 training images and 2,000 validation images. Following Swin Transformer [24], we use UperNet [33] with Swin Transformer backbone . All models are initialized with weights pre-trained on ImageNet-1K for detection and segmentation.

Method \ Dataset	COCO with Mask R-CNN						ADE20K with UPerNet	
	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	mIoU	MS mIoU
Full fine-tuning	43.7	66.6	47.7	39.8	63.3	42.7	44.5	45.8
Linear probing	31.7	55.7	32.5	31.2	53.0	32.2	35.7	36.8
VPT-Deep [18]	33.8	57.6	35.3	32.5	54.5	33.9	37.0	37.9
SSF (ours)	34.9	58.9	36.1	33.5	55.8	34.7	38.9	39.8

Table 2: Performance of different fine-tuning methods on the COCO val2017 dataset and ADE20K dataset, where AP^b and AP^m are the average precision of object detection and instance segmentation, respectively. mIoU and MS mIoU show single-scale and multi-scale inference of semantic segmentation.

B Experiments on Detection and Segmentation

We also conduct experiments on broader downstream tasks, *e.g.*, object detection, instance segmentation, and semantic segmentation. For object detection and instance segmentation, we perform experiments on the COCO dataset with Mask R-CNN [13], where Swin-T pre-trained on ImageNet-1K is adopted as the backbone. The specific hyper-parameter setup and data augmentation refer to Swin Transformer [24] and mmdetection [2]. We perform i) full fine-tuning; ii) linear probing, where the weights at the backbone layers are frozen and only weights at the neck and head layers are updated; iii) VPT-Deep; iv) SSF. All models are trained with 1x schedule (12 epochs). The results are shown in Table 2. We can see that SSF outperforms linear probing and VPT-Deep [18] on the COCO dataset in terms of object detection and instance segmentation. For semantic segmentation, we perform experiments on the ADE20K dataset with UperNet [33] and Swin-T pre-trained on ImageNet-1K. The results in Table 2 show that SSF outperforms linear probing and VPT-Deep [18]. However, for both datasets, SSF still has a large gap compared to the full fine-tuning, which might be due to the fact that detection and segmentation tasks are fundamentally different from classification tasks. Only fine-tuning a few parameters in the backbone will result in inferior performance. How to introduce trainable parameters for parameter-efficient fine-tuning in object detection and segmentation will be the future work.

C Visualizations

C.1 Feature Distribution

We also visualize the feature distribution of different fine-tuning methods via t-SNE on the CIFAR-100 dataset. All fine-tuning methods are based on a ViT-B/16 pre-trained on the ImageNet-21K datasets. The results are shown in Figure 1. Our SSF achieves better feature clustering results compared to linear probing and VPT-Deep. Besides, since our method and full fine-tuning have similar accuracy

(93.99% *vs.* 93.82%), it is difficult to distinguish them in terms of feature distribution, which also shows the effectiveness of our method.

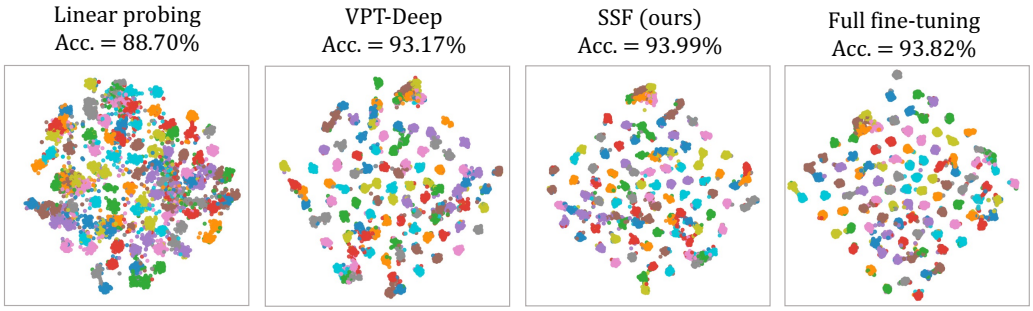

Figure 1: t-SNE visualization of different fine-tuning methods, including linear probing, VPT-Deep, our SSF, and full fine-tuning (best viewed in color).

C.2 Attention Map

We also visualize the attention maps of different fine-tuning methods, as shown in Figure 2. All models are fine-tuned on ImageNet-1K with ViT-B/16 pre-trained on ImageNet-21K. The specific experimental results refer to the main text. We find that VPT-Deep has more concentrated attention on the object in some images (*e.g.*, the first two lines), but lacks suitable attention on some other images (*e.g.*, the last two lines). In contrast, SSF tends to obtain attention similar to the full fine-tuning but also generates the failure prediction such as the second row.

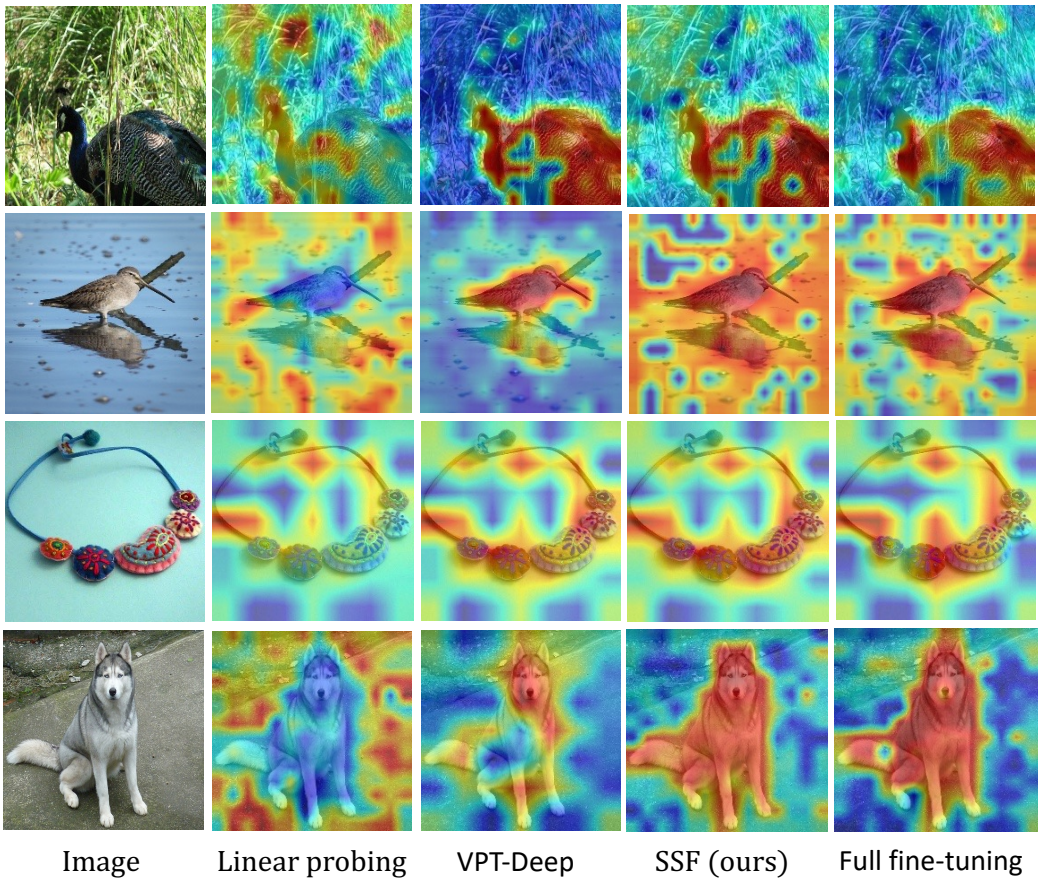

Figure 2: Visualization of attention maps. From left to right, each column shows the RGB image, linear probing, VPT-Deep, our SSF and full fine-tuning.

D Limitations and Societal Impacts

Regarding the limitations of this work, we currently focus on sharing backbone parameters among different tasks while treating each task independently of the rest of the tasks involved. However, some recent papers (*e.g.*, [11, 10]) show that by correlating multiple tasks together during the fine-tuning, the performance for every single task can be further improved. However, recent works treat this relationship among tasks as a black box that inevitably suffers a huge computational cost. Thus, we believe an efficient method to find positive task relationships could be a meaningful direction for further exploration.

This work has the following societal impact. SSF can effectively save parameters compared to the full fine-tuning so that the approach can quickly transfer large models pre-trained on large datasets to downstream tasks, which saves computational resources and carbon emissions. Thanks to the linear transformation and re-parameterization, we do not need to change the deployed backbone architecture when the model is transferred to the downstream task. Only a set of weights need to be replaced, which is also more convenient compared to the methods that introduce additional parameters such as VPT [18]. However, like other fine-tuning methods, SSF is also based on a pre-trained model, which will probably also cause a violation of the use of fine-tuning methods if this upstream pre-trained model is trained on some illegal data.

References

[1] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[5] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[7] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[8] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. Fine-grained car detection for visual census estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[10] Andrea Gesmundo and Jeff Dean. An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems. *arXiv preprint arXiv:2205.12755*, 2022.

[11] Andrea Gesmundo and Jeff Dean. munet: Evolving pretrained deep neural networks into scalable auto-tuning multitask systems. *arXiv preprint arXiv:2205.10937*, 2022.

[12] Ben Graham. Kaggle diabetic retinopathy detection competition report. *University of Warwick*, pages 24–26, 2015.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.
[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.
[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
[20] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.
[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
[22] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.
[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.
[24] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.
[25] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset, 2017. *URL https://github. com/deepmind/dsprites-dataset*, 2020.
[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.
[27] 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*, pages 722–729. IEEE, 2008.
[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.
[29] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.
[30] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.
[31] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
[32] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.
[33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.
[34] 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.
[35] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.