Supplementary Material

A. Proof of the Proposition

Proposition 1. *For downstream datasets, we can hardly learn new information about the input when most features are frozen, then if we learn representation z_{new} in a way that information is lost, i.e., $I(z_{new}, x) < I(z, x)$, the representation will also lose information about the label y_2 as $I(z_{new}, y_2) < I(z, y_2)$.*

Proof. According to the assumption. 1 and 2, we have:

$$\begin{aligned} I(z,x) = I(z,y_1) = I(z,y_2) \\ I(z_{new},x) = I(z_{new},y_1) = I(z_{new},y_2) \end{aligned} \tag{S-1}$$

As $I(z_{new}, x) \leq I(z, x)$ holds, then we can achieve $I(z_{new}, y_2) < I(z, y_2)$ naturally. □

Proposition 2. *For $z_1, z_2 \in \mathbb{R}^N$, the mutual information $I(z_1, z_2)$ equals to $I(z_1, z_1)$ when the mapping $z_2 = f_\psi(z_1), f_\psi : \mathbb{R}^N \to \mathbb{R}^N$ is invertible and smooth.*

Proof. When the mapping $z_2 = f_\psi(z_1), f_\psi : \mathbb{R}^N \to \mathbb{R}^N$ is invertible and smooth, this means $z_1 \to z_2$ is a one-to-one mapping. Then $I(z_1, z_2) = H(z_1)$. For variable z_1, $I(z_1, z_1) = H(z_1)$, we have $I(z_1, z_2) = I(z_1, z_1)$. □

Proposition 3. *When $\ln \left|\det \frac{\partial f}{\partial z}\right| \leq 0$ holds, the Lipschitz constant for f will meet the constrain as $K(f(\cdot)) \leq 1$.*

Proof. Here, we consider our used planar flow. The log Jacobian for planar transformation is

$$\ln \left|\det \frac{\partial f}{\partial z}\right| = \ln \left| I + \lambda^T \cdot h'(\gamma^T \cdot z + \beta) \cdot \gamma \right| \tag{S-2}$$

When $\ln \left|\det \frac{\partial f}{\partial z}\right| \leq 0$ holds, we have $-2 \leq \lambda^T \cdot h'(\gamma^T \cdot z + \beta) \cdot \gamma \leq 0$. Considering two input $z_1 > z_2$, we have $f(z_1) - f(z_2) = z_1 - z_2 + \lambda \cdot \left[h(\gamma^T \cdot z_1 + \beta) - h(\gamma^T \cdot z_2 + \beta)\right]$. As the gradient of $\lambda \cdot h(\gamma^T \cdot z + \beta)$ has the constraint, then $\lambda \cdot \left[h(\gamma^T \cdot z_1 + \beta) - h(\gamma^T \cdot z_2 + \beta)\right]$ is also constrained following 1st order approximation. Thus we have

$$z_2 - z_1 \leq f(z_1) - f(z_2) \leq z_1 - z_2 \tag{S-3}$$

This means for any z_1, z_2, the constrain is:

$$K(f(\cdot)) = \frac{||f(z_1) - f(z_2)||}{||x_1 - x_2||} \leq 1 \tag{S-4}$$

In this way the constrain $K(f(\cdot)) \leq 1$ holds. □

B. Datasets and Implementation Details

Datasets. We introduce the details of our used datasets in Tab. S-1.

Augmentation. For VTAB-1k and domain generalization, we follow its default augmentation settings, implementing the resizing and normalization for input images. For few-shot learning and other FGVC datasets, different from NOAH, which uses strong augmentation, such as color-jitters and RandAugmentation, we employ very simple random center crop and random horizontal flip.

Hyper-parameters. The batch-size is set as 128 for all the experiments and the learning rate for Few-shot and FGVC datasets is 2×10^{-3}, while the learning rate for ImageNet is 5×10^{-4}. For VTAB-1k, we follow VPT [12] and search for a superior hyper-parameter (learning rate and weight decay) from a learning rate list:$\{1 \times 10^{-3}$,2×10^{-3},5×10^{-5},5×10^{-3},1×10^{-2},$5 \times 10^{-2}\}$ and weight decay list: $\{1 \times 10^{-2}$, 5×10^{-3},1×10^{-3}, 1×10^{-4},$5 \times 10^{-5}\}$.

C. Extension Experiments

C.1 Attention Map

We further visualize the attention map in Fig S-1. The model is fine-tuned with SNF-shallow or linear layer on ImageNet-1k with 16 examples per-class training. The top-1 accuracy for linear probing is 70.7 and the top-1 accuracy for SNF-shallow is 78.5. Obviously, SNF can adjust the feature attention with information-keeping adaption on the shortcut.

C.2 With or without regularization on log Jacobian.

We perform ablation studies to verify the effectiveness of the regularization on log Jacobian and the results on Caltech101 (VTAB-1k). As illustrated in Tab. S-3. We notice that our approach can already achieve great performance without regularizing on log Jacobian, and the model will actively reduce the log Jacobian during the adapting process to reduce error propagation, as shown in Fig. S-2. The above phenomenon confirms our analysis of error propagation. In Tab. S-3, we also show that explicitly imposing regularization on log Jacobian can further improve performance.

Table S-1. Specifications of datasets evaluated. Different from VPT [12], the val-train split is only employed for VTAB-1k benchmark.

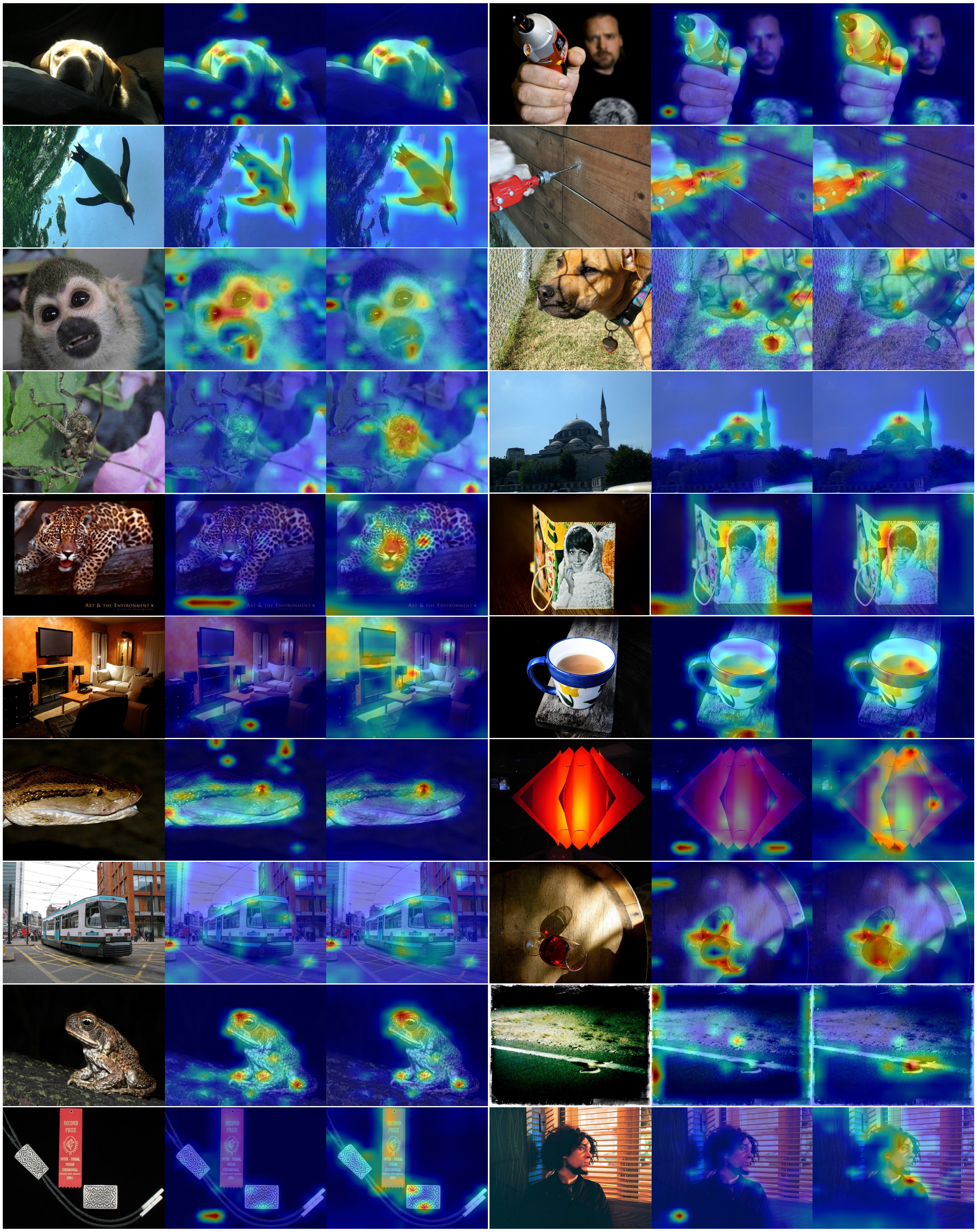
Dataset	Description	#Classes	Train	Val	Test
Fine-grained visual recognition tasks (FGVC)					
CUB-200-2011 [29]	Fine-grained bird species recognition	200	5,994	-	5,794
NABirds [27]	Fine-grained bird species recognition	555	23,929	-	24,633
Oxford Flowers [23]	Fine-grained flower species recognition	102	2,040	-	6,149
Stanford Dogs [15]	Fine-grained dog species recognition	120	12,000	-	8,580
Stanford Cars [7]	Fine-grained car recognition	196	8,144	-	8,041
Visual Task Adaptation Benchmark (VTAB-1k)					
Cifar100 [17]	Natural	100	800/1000	200	10,000
Caltech101 [6]		102			6,084
DTD [4]		47			1,880
Oxford-Flowers102 [22]		102			6,149
Oxford-Pets [24]		37			3,669
SVHN [21]		10			26,032
Sun397 [31]		397			21,750
Patch Camelyon [28]	Specialized	2	800/1000	200	32,768
EuroSAT [9]		10			5,400
Resisc45 [3]		45			6,300
Retinopathy [14]		5			42,670
Clevr/count [13]	Structured	8	800/1000	200	15,000
Clevr/distance [13]		6			15,000
DMLab [1]		6			22,735
KITTI-Dist [8]		4			711
dSprites/location [20]		16			73,728
dSprites/orientation [20]		16			73,728
SmallNORB/azimuth [18]		18			12,150
SmallNORB/elevation [18]		9			12,150
Few-shot Learning					
Food-101 [2]	Daily fine-grained food recognition	101	(1/2/4/8/16)*(#Classes)	-	25,250
Stanford Cars [16]	Daily fine-grained car recognition	196		-	8,041
Oxford-Flowers102 [22]	Daily fine-grained flower species recognition	102		-	6,149
FGVC-Aircraft [19]	Daily fine-grained Aircraft species recognition	100		-	3,333
Oxford-Pets [24]	Daily fine-grained pet species recognition	37		-	3,669
Domain Generalization					
ImageNet-V2 [25]	Variants of ImageNet with domain shifts	1000	-	-	10,000
ImageNet-Sketch [30]		1000	-	-	50,889
ImageNet-A [11]		1000	-	-	7,500
ImageNet-R [10],		1000	-	-	30,000
Other Visual Recognition Tasks					
ImageNet [5]	Other general visual recognition	1,000	16*(#Classes)	50,000	150,000
Cifar-100 [17]		100	50,000	-	10,000

C.3 Whether using affine as the first layer.

The shallow SNF is an affine transformation of the shortcut connection following [26]. However, it is not necessary and we can replace the affine transformation with the planar transformation used in SNF-deep. As shown in Tab. S-2, using affine or not has almost no difference in accuracy.

C.4 Feature Visualization

As shown in Fig S-3, We visualize the feature distribution learned from the pretrained model and SNF-shallow via t-SNE on the Cifar-100 dataset. It reveals SNF-shallow can achieve impressive feature clustering results compared with the pretrained backbone by adapting the shortcut with only



(a) Origin **(b) Linear** **(c) SNF-shallow** **(a) Origin** **(b) Linear** **(c) SNF-shallow**

Figure S-1. The attention for images sampled from val split of ImageNet. (a) The origin image (b) The attention map of the linear probing (c) The attention map of our SNF-shallow.

Table S-2. Whether using affine as the first layer.

Methods	Caltech101		Cifar100	
	SNF-s	SNF-d	SNF-s	SNF-d
w/ affine	93.5	**94.0**	**84.3**	84.0
w/o affine	**93.6**	**94.0**	83.9	**84.2**

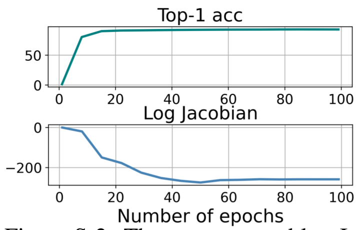
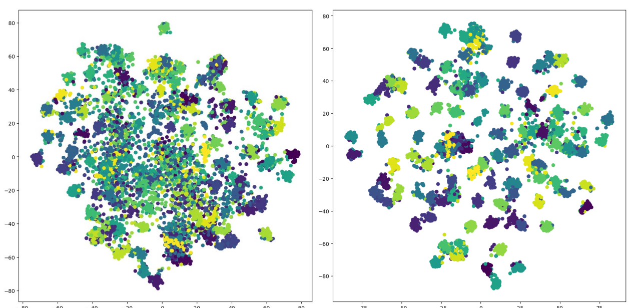


Figure S-2. The accuracy and log Jacobian curve when no constraint on log Jacobian is added in the loss function.

Method	w/	w/o
SNF-s	93.5	93.1
SNF-d	94.0	93.5

Table S-3. The performance on Caltech101 (VTAB-1k) when with or without log Jacobian constrain.



(a) ImageNet-21k pretrained **(b) SNF-shallow**

Figure S-3. t-SNE visualization of the feature learned by pretrained backbone and SNF-shallow respectively. Different colors represent different ground truth labels.

Table S-4. More ablation studies on different backbones and different layer lengths.

Methods	**ViT-L**		**ResNet-50**	
	Cifar100	Caltech101	Cifar100	Caltech101
Linear	64.7	91.8	33.3	86.2
Full	75.7	92.1	43.9	89.1
SNF-s	83.0	92.5	44.7	88.8
3-layer	84.3	**93.2**	48.0	89.4
SNF-d	**85.4**	**93.2**	49.3	89.5
7-layer	**85.5**	**93.3**	**50.0**	**89.7**

0.036 M parameters,

C.5 More ablation studies on different backbones and layer lengths

We further provide more ablation studies on different backbones and layer lengths using Caltech101 and Cifar100. Results are shown in Tab. S-4.

D. Limitation and Future Work

The assumptions in this work is not rigorously proven, which is also prohibitive in deep learning. We present these assumptions and derive SNF to adapt the feature distribution without losing information. The success of SNF can verify our assumptions to some extent. However, it still leaves an challenging problem for future research that how can we adapt the model with few parameters when the information captured by the pre-trained model is not enough. Maybe we can use small networks to adapt the pre-trained model as well as learn new information from target 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

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 2

[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. 2

[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. 2

[5] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 2

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 2

[7] 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. 2

[8] 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. 2

[9] 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*, 2019. 2

[10] 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. 2

[11] 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. 2

[12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 1, 2

[13] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017. 2

[14] Kaggle and EyePacs. Kaggle diabetic retinopathy detection. 2015. 2

[15] 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. 2

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[18] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–104. IEEE, 2004. 2

[19] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[20] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset, 2017. 2

[21] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 2

[22] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006. 2

[23] 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. 2

[24] 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. 2

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 2

[26] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[27] 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. 2

[28] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018. 2

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[30] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[31] 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. 2