# Appendix for Flatten Long-Range Loss Landscapes for Cross-Domain Few-Shot Learning

Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, Ruixuan Li*
School of Computer Science and Technology, Huazhong University of Science and Technology
{yixiongz, smnight, m202273659, idcliyuhua, rxli}@hust.edu.cn

Figure 1. Samples of the *mini*ImageNet datasets.

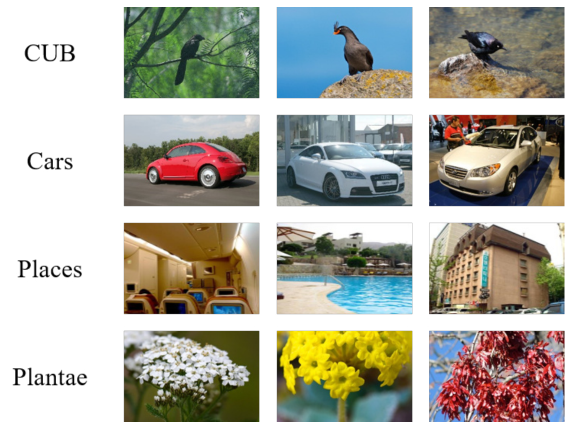


Figure 2. Samples of the CUB, Cars, Places, and Plantae datasets.

## A. Detailed Dataset Setups

*mini*ImageNet [13] is a subset of the ImageNet dataset [5], containing 100 classes randomly sampled from ImageNet, and each class contains 600 images. Following current works [2, 6], we utilize its base classes as the source-domain dataset, where 64 classes and 38,400 images are involved. Different with the ordinary few-shot learning works [13], cross-domain few-shot learning (FSL) utilize the raw image from ImageNet following the same data list, instead of resizing each image to the size of 84 $\times$ 84. Image samples can be found in Fig. 1.

CUB [14] is a fine-grained dataset of bird classification. Following current works [2, 6], we utilize its novel-class split to be one of our target datasets, which contains 50 classes and 2,953 samples in all. Sampled images can be found in Fig. 1.

Cars [9] is a fine-grained dataset of car classification. It contains images of cars with 49 classes and 2,027 images in all.

Places [17] collects images of different places such as the airplane, coffee bar and so on. It contains 19 classes and 3,800 images in all.

Plantae [8] is a dataset of plant classification. It contains 50 classes and 3,800 images in all.

CropDiseases [10] is a dataset for recognizing agricultural diseases. It contains 19 classes and 43,456 images in all. Sampled images can be found in Fig. 1. The above 5 datasets are all in natural images, which is close to the *mini*ImageNet dataset. Below we will also introduce three datasets that are in the distant domains.

EuroSAT [7] contains satellite imagery of the earth. It contains 10 classes and 27,000 images in all.

ISIC2018 [4] contains skin lesion images for lesion classification. It contains 7 classes and 10,015 images in all.

ChestX [15]) is the most challenging dataset with the X-ray images for chest classification. Since its images are very different from that of the *mini*ImageNet dataset, it is very hard to transfer knowledge to it. It contains 7 classes and 25,847 images in all.

The $k$-way $n$-shot classification refers to sampling episodes for few-shot training and evaluation. Each episode can be understood as a small dataset, which a training set (a.k.a. support set) contains $k$ classes and $n$ training samples in each class, and a test set (a.k.a. query set) containing un-overlapping samples from the given $k$ classes. Typically, we have the 5-way 1-shot and 5-way 5-shot settings. Since

*Corresponding author. Code is at https://github.com/Zoilsen/FLoR.

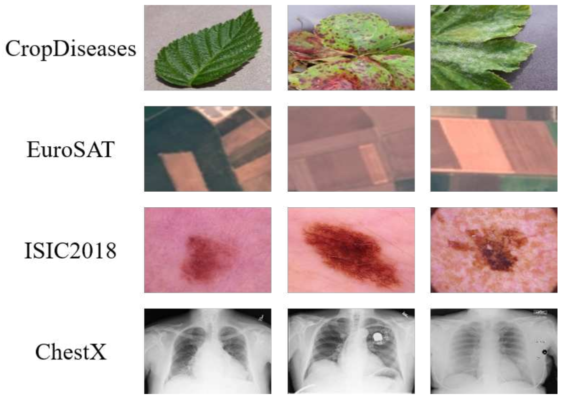


Figure 3. Samples of the CropDiseases, EuroSAT, ISIC2018, and ChestX datasets.

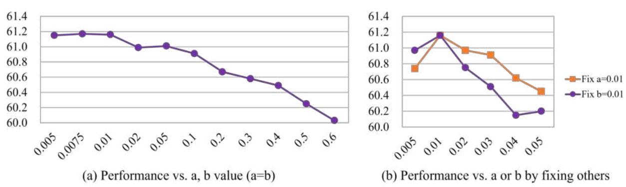


Figure 4. Sensitivity study of hyper-parameter choices by (a) keeping a = b and (b) by fixing a or b and tuning the other one. The best hyper-parameter choice is a = b = 0.1, and the performance is stable when hyper-parameter changes.

the transductive setting utilize the query set as an unlabeled training set, the size of the query set is also an important issue for fair comparisons. Typically, the query set contains 15 samples for each class, leading to 75 query samples in each query set in total.

# B. Sensitivity Study

We report the hyper-parameter choice of our model in Fig. 4. Since there are only two hyper-parameters in our method, we first test to keep their values the same (Fig. 4a), and then fix one value to search for the other value (Fig. 4b). We can see the optimal value is a=b=0.01, and the performance stably changes when altering a or b, which means our model is not sensitive to the hyper-parameter choice.

# C. More Validations

## C.1. Comparison with more BN + IN methods

We implement more methods based on BN + IN in Tab. 1, and compare with them both quantitatively and technically. Technically, (1) our analysis and instantiations **are not limited to normalization layers** (see Tab. 2) **or specific network structures** (CNN, ViT); (2) we **randomly** sample intermediate points between outputs to **cover more high-loss regions**, instead of setting fixed or learnable ratios like existing works, which is verified to be more effective in paper Tab.7.

Table 1. Comparison with more BN + IN methods.

| Method | CropDisease | EuroSAT | ISIC2018 | ChestX | Ave. |
|---|---|---|---|---|---|
| Meta BIN [3] | 86.66 ±0.19 | 78.52 ±0.28 | 46.06 ±0.31 | 25.28 ±0.16 | 59.13 ±0.12 |
| TaskNorm [1] | 87.95 ±0.23 | 79.32 ±0.32 | 43.15 ±0.43 | 26.48 ±0.19 | 59.23 ±0.20 |
| BIN [11] | 86.72 ±0.22 | 77.87 ±0.28 | 48.46 ±0.32 | 25.90 ±0.14 | 59.28 ±0.12 |
| Ours | **89.35** ±0.17 | **79.40** ±0.27 | **50.75** ±0.30 | **26.57** ±0.16 | **61.52** ±0.12 |

## C.2. Why selecting normalization layers

**Our analysis is not limited to normalization layers.** However, as normalization layers are easy to produce effective but distinct representations (i.e., different minima in landscapes), it is easier to be applied to flatten the long-range loss landscapes. We try to produce distinct representations through applying different convolutions in Tab. 2, which also improves the performance and verifies our analysis. However, the improvements are marginal compared with normalization layers.

Table 2. Comparison with more different instantiations.

| Method | CropDisease | EuroSAT | ISIC2018 | ChestX | Ave. |
|---|---|---|---|---|---|
| Baseline (Conv3x3) | 85.80 ±0.27 | 78.01 ±0.22 | 39.10 ±0.33 | 26.13 ±0.17 | 57.26 ±0.13 |
| Conv3x3 + Conv5x5 | 86.29 ±0.33 | 78.79 ±0.29 | 41.52 ±0.31 | 25.85 ±0.19 | 58.11 ±0.19 |
| Conv1x1 + Conv7x7 | 86.27 ±0.22 | 76.36 ±0.33 | 44.03 ±0.18 | 25.80 ±0.19 | 58.12 ±0.19 |
| Ours | **89.35** ±0.17 | **79.40** ±0.27 | **50.75** ±0.30 | **26.57** ±0.16 | **61.52** ±0.12 |

## C.3. Randomness of model parameters

Although randomness is verified to be beneficial (paper Tab.7), our improvements also originate from **where to import randomness** (i.e., intermediate points between normalized representations). We follow FWT [25] to compare with randomness-based works by the 5-way 1-shot accuracy in Tab. 3, showing our design is vital.

Table 3. Comparison with randomness-based methods.

| Method | Randomness Location | CUB | Cars | Places | Plantae | Ave. |
|---|---|---|---|---|---|---|
| Dropout | Single Output Feature | 35.86 ±0.51 | 30.72 ±0.43 | 37.47 ±0.62 | 29.22 ±0.47 | 33.32 |
| FWT | BN weight and bias | 45.69 ±0.68 | 31.79 ±0.51 | 53.10 ±0.80 | 35.60 ±0.56 | 41.55 |
| Ours | Intermediates of multiple features | **49.99** ±0.18 | **37.41** ±0.31 | **53.18** ±0.28 | **40.10** ±0.42 | **45.17** |

## C.4. Generalization to other normalization layers

Our method **can also generalize to the combination of other normalizations**. We report the performance of different combinations in Tab. 4. Since IN are more similar to GroupNorm (GN) [16], the minima produced by them are closer, making the flattened range smaller than BN + GN or BN + IN. Therefore, the improvements of GN + IN are smaller than others, although GN or IN shows better performance than BN individually.

## C.5. Analysis experiments on real-world data

We use remote sensing images in EuroSAT [13] and medical images in ISIC [5] as the real-word data in Fig. 5, and

Table 4. Comparison with more normalizations.

| Method | CropDisease | EuroSAT | ISIC2018 | ChestX | Ave. |
|---|---|---|---|---|---|
| Baseline (BN) | 85.80 ±0.27 | 78.01 ±0.22 | 39.10 ±0.33 | 26.13 ±0.17 | 57.26 ±0.13 |
| GN | 85.06 ±0.22 | 76.28 ±0.29 | 46.74 ±0.40 | 24.45 ±0.19 | 58.13 ±0.16 |
| IN | 86.67 ±0.20 | 76.17 ±0.24 | 47.25 ±0.21 | 24.79 ±0.15 | 58.72 ±0.10 |
| GN + IN | 87.22 ±0.28 | 76.96 ±0.31 | 48.77 ±0.29 | 24.94 ±0.18 | 59.47 ±0.22 |
| BN + GN | 89.28 ±0.17 | **80.79** ±0.22 | 46.26 ±0.17 | 25.66 ±0.19 | 60.50 ±0.18 |
| Ours (BN + IN) | **89.35** ±0.17 | 79.40 ±0.27 | **50.75** ±0.30 | **26.57** ±0.16 | **61.52** ±0.12 |

use the level of image styles being shifted as the perturbation level. **Results are consistent with the experiments in Fig.2 and Tab.1**.

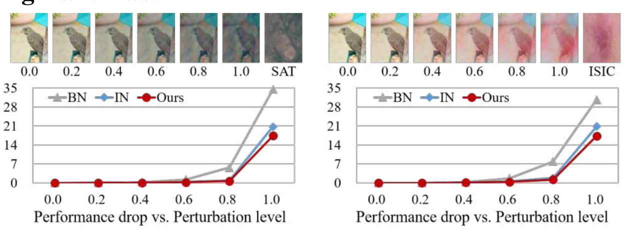


Figure 5. Analysis experiments on real-world data.

## C.6. Ablate parameter-space perturbation from FLoR

We first perturb only the parameters in the BN layer in Tab. 5, we can see the improvements are limited. We then remove the learnable parameters in the BN and IN layers. We can see the results are close to ours, verifying that **the improvement is predominantly due to the representation-space flatness**.

Table 5. Ablation study of parameter-space perturbations.

| Method | CropDisease | EuroSAT | ISIC2018 | ChestX | Ave. |
|---|---|---|---|---|---|
| Baseline | 85.80 ±0.27 | 78.01 ±0.22 | 39.10 ±0.33 | 26.13 ±0.17 | 57.26 ±0.13 |
| Perturb only BN Params | 88.23 ±0.33 | 77.65 ±0.40 | 42.02 ±0.34 | 26.52 ±0.28 | 58.61 ±0.20 |
| Ours (w/o learnable param) | 87.50 ±0.19 | **79.98** ±0.28 | 48.71 ±0.22 | 25.85 ±0.14 | 60.51 ±0.13 |
| Ours (w/ learnable param) | **89.35** ±0.17 | 79.40 ±0.27 | **50.75** ±0.30 | **26.57** ±0.16 | **61.52** ±0.12 |

## C.7. Comparison with the sharpness-based work (F2M [12])

We differ with F2M in (1) we flatten loss landscapes in the **representation space**, but F2M is in the **parameter space**; (2) our flattening is achieved by randomly sampling intermediate points between **multiple** local minima, but F2M is by adding perturbations to model parameters (a **single** minimum); (3) our performance is significantly higher. We implement F2M and compare with it in Tab. 6.

Table 6. Comparison with sharpness-based work.

| Method | CropDisease | EuroSAT | ISIC2018 | ChestX | Ave. |
|---|---|---|---|---|---|
| F2M | 86.37 ±0.15 | 75.05 ±0.17 | 43.52 ±0.14 | 26.06 ±0.11 | 57.75 ±0.10 |
| Ours | **89.35** ±0.17 | **79.40** ±0.27 | **50.75** ±0.30 | **26.57** ±0.16 | **61.52** ±0.12 |

## References


[1] John Bronskill, Jonathan Gordon, James Requeima, Sebastian Nowozin, and Richard Turner. Tasknorm: Rethinking batch normalization for meta-learning. In *Proceedings of the International Conference on Machine Learning*, pages 1153–1164. PMLR, 2020. 2

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representations*, 2018. 1

[3] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification, 2021. 2

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. 1

[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*, pages 248–255. Ieee, 2009. 1

[6] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 124–141. Springer, 2020. 1

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019. 1

[8] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018. 1

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[10] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7(September), 2016. Publisher Copyright: © 2016 Mohanty, Hughes and Salathé. 1

[11] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 2

[12] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima, 2021. 3

[13] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 3637–3645, 2016. 1

[14] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[15] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 1

[16] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 1