Appendix for
The Devil is in Low-Level Features for Cross-Domain Few-Shot Segmentation

Yuhan Liu, Yixiong Zou*, Yuhua Li, Ruixuan Li
School of Computer Science and Technology, Huazhong University of Science and Technology
{yuhan_liu, yixiongz, idcliyuhua, rxli}@hust.edu.cn

Figure 1. Samples of the Pascal VOC 2012 dataset.

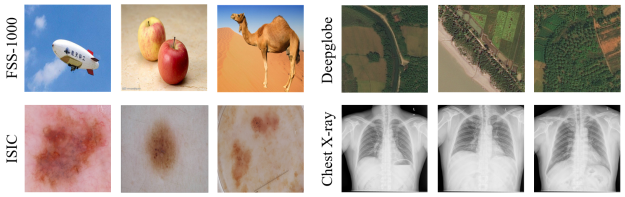

Figure 2. Samples of the FSS-1000, Deepglobe, ISIC, and Chest X-ray datasets.

A. Detailed Dataset Setups

Following the settings in PATNet [11], we use PASCAL VOC 2012 [5] with SBD augmentation [7] as source domain for training. Sampled images can be found in Fig. 1. Then we evaluate the trained models on four target domains: FSS-1000 [12], Deepglobe [4], ISIC [3, 14], and Chest X-ray [1, 9]. Sampled images can be found in Fig. 2.

FSS-1000 [12] is a natural image dataset for few-shot segmentation, comprising 1,000 natural image categories, each containing 10 samples. Following the official split for semantic segmentation, our experiments evaluate performance on the designated testing set, which includes 240 classes and 2,400 images.

Deepglobe [4] contains densely annotated satellite images spanning seven categories: urban, agriculture, rangeland, forest, water, barren, and unknown. The dataset contains 5,666 images, each resized to 408×408 pixels.

*Corresponding author.

ISIC [3, 14] focuses on lesion images for skin cancer screening, with each image featuring a single primary lesion. The dataset comprises 2,596 images, all resized to 512×512 pixels.

Chest X-ray [1, 9] is a dataset of X-ray images for Tuberculosis diagnosis, comprising 566 images with a resolution of 1024×1024 pixels. These images are sourced from 58 Tuberculosis cases and 80 normal cases.

B. Sharpness-Aware Minimization(SAM)

In our work, we use the sharpness of loss landscape as an entry point to explore the connection between shallow layers and early stops, and propose a novel sharpness-aware minimization method to flatten the loss landscapes for low-level features. Following the previous work [6, 17], we present a detailed introduction to sharpness-aware minimization(SAM).

Consider a family of models parameterized by $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$, the core idea of SAM is to seek parameters $\boldsymbol{w}$ that lie in neighborhoods with uniformly low loss values, rather than focusing solely on parameters with low individual loss values.

Define the training set loss $L_S(\boldsymbol{w})$ and the population loss $L_D(\boldsymbol{w})$. The goal of model training is to select model parameters $\boldsymbol{w}$ having low population loss $L_D(\boldsymbol{w})$. According to PAC-Bayesian Generalization Bound Theorem:

For any $\rho > 0$ and any distribution D, with probability $1-\delta$ over the choice of the training set $\mathcal{S} \sim D$,

$$L_D(\boldsymbol{w}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + \sqrt{\frac{k \log\left(1 + \frac{\|\boldsymbol{w}\|_2^2}{\rho^2}\left(1+\sqrt{\frac{\log(n)}{k}}\right)^2\right) + 4\log\frac{n}{\delta} + \tilde{O}(1)}{n-1}} \tag{1}$$

where $n = |\mathcal{S}|$, k is the number of parameters and we assumed $L_D(\boldsymbol{w}) \leq \mathbb{E}_{\epsilon_i \sim (0,\rho)}[L_D(\boldsymbol{w}+\boldsymbol{\epsilon})]$.

Eq. 1 can be simplified as:

$$L_D(\boldsymbol{w}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + h(\|\boldsymbol{w}\|_2^2/\rho^2), \tag{2}$$

where $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function (under some technical conditions on $L_D(\boldsymbol{w})$).

The right-hand side of the inequality above can be rewritten as:

$$[\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) - L_{\mathcal{S}}(\boldsymbol{w})] + L_{\mathcal{S}}(\boldsymbol{w}) + h(\|\boldsymbol{w}\|_2^2/\rho^2), \quad (3)$$

The term in square brackets, i.e. $\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_{\mathcal{S}}(\boldsymbol{w} + \boldsymbol{\epsilon}) - L_{\mathcal{S}}(\boldsymbol{w})$ quantifies how rapidly the training loss increases when moving from $\boldsymbol{w}$ to a nearby parameter value, which is termed sharpness. Therefore, the right-hand side of Eq. 1 can be viewed as sharpness, the training loss value itself, and a regularizer on the magnitude of $\boldsymbol{w}$. In other words, the optimization goal of SAM is to seek parameter values by simultaneously minimizing loss values and loss sharpness.

C. Centered Kernel Alignment(CKA)

In our ablation study, we use Centered Kernel Alignment (CKA) [10, 16, 18–20] to measure domain similarity to demonstrate the effectiveness of LEM. The core idea of CKA is to compare the inner product matrices of the feature representations and to center them, removing the bias introduced by the data's mean.

Given two feature matrices $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$, where n is the number of samples, and d_1 and d_2 are the dimensions of the two feature spaces, CKA measures their similarity by comparing the kernel alignment of their inner product spaces.

First, calculate the inner product matrices for both feature matrices:

$$K_X = X^T X, \quad K_Y = Y^T Y \quad (4)$$

Then, center the kernel matrices to remove the effect of the mean. For the matrix K_X, the centered matrix C_X is obtained by:

$$C_X = K_X - \frac{1}{n}\mathbf{1}_n K_X - \frac{1}{n} K_X \mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n K_X \mathbf{1}_n \quad (5)$$

where $\mathbf{1}_n$ is an n-dimensional column vector of ones, representing the number of samples. Similarly, C_Y can also be obtained.

Finally, CKA is computed as the following formula:

$$\text{CKA}(X, Y) = \frac{\langle C_X, C_Y \rangle_F}{\|C_X\|_F \|C_Y\|_F} \quad (6)$$

where $\langle C_X, C_Y \rangle_F$ is the Frobenius inner product (sum of element-wise products of the matrices), and $\|C_X\|_F$ and $\|C_Y\|_F$ are the Frobenius norms (square root of the sum of squared elements) of C_X and C_Y, respectively.

D. Compared with Other Domain Generalization Methods

Domain Generalization (DG) aims to train models that can generalize to diverse, unseen target domains, particularly when target domain data is unavailable during training, which aligns with the objective of CDFSS. Instance Normalization (IN) [15] normalizes features within each individual sample by adjusting its mean and variance. This helps in reducing style-specific variations, making it particularly effective in tasks like domain generalization, style transfer and domain-invariant representation learning. We compare our LEM with IN in Tab.1 and demonstrate that our LEM outperforms IN in preserving domain-invariant information by incorporating randomly synthesized domains. [2] introduced a domain generalization approach based on augmentation, leveraging amplitude-phase recombination to direct the model's attention toward the phase spectrum. The results shown in Tab.1 validate that the random convolution in our LEM module is a more effective way to simulate different domains. Besides, our method can be seen as a novel sharpness-aware minimization method. Many existing SAM methods focus on analyzing loss landscapes in the parameter space, rather than in the representation space. We compare our method with directly perturbing low-level features in Tab.1 and demonstrate the superiority of our approach.

Table 1. Compared with other domain generalization methods.

Method	FSS-1000	deepglobe	ISIC	Chest X-ray	Average
baseline	78.91	40.00	35.49	74.44	57.21
instance normalization	78.39	39.99	36.85	76.27	57.88
amplitude-phase recombination	78.62	39.63	35.53	75.56	57.34
perturbtion	78.72	39.50	35.82	76.77	57.70
ours	78.83	40.05	36.31	78.21	58.35

E. Sensitivity Study

As shown in Tab.2 and Tab.3, we investigated the effects of using our LCM and LEM at various positions in the shallow layers. We inserted our modules at before stage 1, stage 1 block 1, stage 2 block 1, stage 2 block 2, and stage 2 block 3. The improvement in performance demonstrates the effectiveness of our modules.

Table 2. The effect of LCM in different shallow layers.

Method	FSS-1000	deepglobe	ISIC	Chest X-ray	Average
baseline	78.91	40.00	35.49	74.44	57.21
baseline+LCM(before stage1)	78.84	44.22	38.12	77.93	59.78
baseline+LCM(stage1 block1)	78.91	42.09	36.20	78.14	58.84
baseline+LCM(stage2 block1)	78.17	44.74	37.85	78.13	59.72
baseline+LCM(stage2 block2)	78.51	42.96	37.18	78.74	59.35
baseline+LCM(stage2 block3)	77.67	45.77	37.98	76.95	59.59

Table 3. The effect of LEM in different shallow layers.

Method	FSS-1000	deepglobe	ISIC	Chest X-ray	Average
baseline	78.91	40.00	35.49	74.44	57.21
baseline+LEM(before stage1)	78.83	40.05	36.31	78.21	58.35
baseline+LEM(stage1 block1)	78.76	39.26	36.04	77.22	57.82
baseline+LEM(stage2 block1)	78.74	39.35	36.96	76.51	57.89
baseline+LEM(stage2 block2)	78.66	39.95	36.21	75.19	57.50
baseline+LEM(stage2 block3)	78.82	39.81	35.32	75.89	57.46

Table 4. Applying our method to the CDFSS baseline can further enhance performance.

Method	FSS-1000		Deepglobe		ISIC		Chest X-ray		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PATNet	78.59	81.23	37.89	42.97	41.16	53.58	66.61	70.20	56.06	61.99
ABCDFSS	74.60	76.20	42.60	49.00	45.70	53.30	79.80	81.40	60.67	64.97
DRA	79.05	80.40	41.29	50.12	40.77	48.87	82.35	82.31	60.86	65.42
PATNet+Ours	78.70	81.20	42.08	47.42	42.80	53.96	74.25	76.60	59.46	64.80
ABCDFSS+Ours	78.42	79.93	**44.87**	49.56	**46.24**	**54.08**	82.10	82.65	62.91	66.56
DRA+Ours	**80.92**	**81.35**	44.66	**50.89**	42.29	50.11	**84.28**	**84.76**	**63.04**	**66.78**

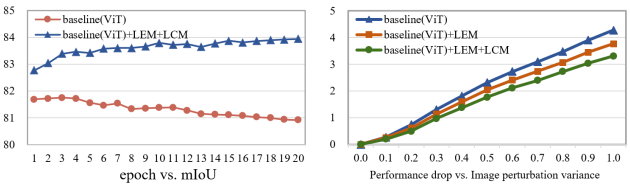

Figure 3. Our exploration of low-level features also works on ViT.

F. More Validations

F.1. CDFSS methods as the baseline

In Tab.4, we apply our method on CDFSS baselines[8, 11, 13] and observe improved performance, consistent with the improvement of using FSS as the baseline.

F.2. ViT as the backbone

In Fig. 3, we present the mIoU trend analysis and sharpness evaluation using ViT as the backbone, demonstrating that both CNN and ViT encounter similar challenges in cross-domain tasks. Our analysis and method effectively address these shared issues.

References

[1] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2013. 1

[2] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021. 2

[3] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 1

[4] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 1

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 1

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 1

[7] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 1

[8] Jonas Herzog. Adapt before comparison: A new perspective on cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23605–23615, 2024. 3

[9] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2013. 1

[10] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 2

[11] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 73–90. Springer, 2022. 1, 3

[12] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2869–2878, 2020. 1

[13] Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24036–24045, 2024. 3

[14] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018. 1

[15] D Ulyanov. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[16] Yixiong Zou, Shanghang Zhang, Yuhua Li, and Ruixuan Li. Margin-based few-shot class-incremental learning with class-level overfitting mitigation. *Advances in neural information processing systems*, 35:27267–27279, 2022. 2

[17] Yixiong Zou, Yicong Liu, Yiman Hu, Yuhua Li, and Ruixuan Li. Flatten long-range loss landscapes for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23575–23584, 2024. 1

[18] Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:116332–116354, 2024. 2

[19] Yixiong Zou, Shuai Yi, Yuhua Li, and Ruixuan Li. A closer look at the cls token for cross-domain few-shot learning. *Advances in Neural Information Processing Systems*, 37:85523–85545, 2024.

[20] Yixiong Zou, Shanghang Zhang, Haichen Zhou, Yuhua Li, and Ruixuan Li. Compositional few-shot class-incremental learning. *arXiv preprint arXiv:2405.17022*, 2024. 2