Appendix for "FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space"

Quande Liu[1], Cheng Chen[1], Jing Qin[2], Qi Dou[1], Pheng-Ann Heng[1]
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2] School of Nursing, The Hong Kong Polytechnic University
{qdliu, cchen, qdou, pheng}@cse.cuhk.edu.hk, harry.qin@polyu.edu.hk

1. Datasets Details:

The retinal fundus images adopted in the experiments are collected from four different clinical centers out of three public datasets. Among these data, samples of sites A are from Drishti-GS [10] dataset; samples of site B are from RIM-ONE-r3 [3] dataset; samples of site C, D are from REFUGE [9] dataset. Note that the REFUGE dataset includes two different data sources, so we decompose them in our federated learning setting. Among the six data sources in the prostate MRI segmentation task, samples of Site A, B are from NIC-ISBI13 [1] datasets; samples of Site C are from I2CVB [5] datasets; and samples of Site D, E, F are from PROMISE12 [7] dataset. Similarly, since the NIC-ISBI13 and PROMISE12 contain data from multiple data sources, we separate them and consider each data source as an individual client in the federated scenario. Details of the scanners and imaging protocols of these data are illustrated in Table 1 and Table 2 respectively.

Table 1. Details of the scanning protocols for different data sources in fundus image segmentation.

Task	Dataset	Manufactor
Fundus Image Segmentation	Site A [10]	(Aravind eye hospital)
	Site B [3]	Nidek AFC-210
	Site C [9]	Zeiss Visucam 500
	Site D [9]	Canon CR-2

Table 2. Details of the scanning protocols for different data sources in prostate MRI segmentation.

Task	Dataset	Manufactor	Field strength(T)	Endorectal Coil
Prostate MRI Segmentation	Site A [1]	Siemens	3	Surface
	Site B [1]	Philips	1.5	Endorectal
	Site C [5]	Siemens	3	No
	Site D [7]	Siemens	1.5 and 3	No
	Site E [7]	GE	3	Endorectal
	Site F [7]	Siemens	1.5	Endorectal

2. Standard Division and Statistical Analysis

We calculate the standard division (std) for the generalization results of different comparison methods. The results of the two tasks are shown in Table 3 and Table 4 respectively. We notice that in fundus image segmentation (cf. Table 3), the std with considering site A and site B as unseen sites are relatively higher than the others. The reason could be that generalizing to these two sites when training with remaining three sites are more difficult, causing that the generalization results present a larger cross-subject variance. For prostate MRI segmentation (cf. Table 4), the std are relatively stable across different generalization settings compared with the the fundus image segmentation task.

We also conduct paired t-test between our approach and different comparison methods to analyze whether the performance improvement of our approach is significant. We adopt Dice as the evaluation measurement and set the significance level as 0.05. For each method, the statistical tests are conducted by jointly considering the prediction results of each unseen site setting on overall generalization performance. The results are listed in Table 5. It is observed that all paired t-test results present p-value smaller than 0.05, demonstrating that our improvements over these state-of-the-art domain generalization methods are significant.

3. Visualization of Transformed Data

We visualize the appearances of transformed images under different interpolation ratio λ for the two tasks. As shown in Fig. 1, the appearance of local source image is indeed gradually transformed to the style (i.e. distribution) of target image of other clients as we increase the interpolation ratio from 0 to 1, while the semantic content of the image is unchanged. Such continuous interpolation mechanism helps to enrich the multi-source distributions to a dedicated dense distribution space, hence benefits the model to gain domain-invariance in a more continuous latent space to improve the generalizability.

Table 3. Comparison results on fundus images for Optic Disc/Cup segmentation (with standard division).

Task	Optic Disc Segmentation					Optic Cup Segmentation					Overall
Unseen Site	A	B	C	D	Avg.	A	B	C	D	Avg.	
	Dice Coefficient (mean±std) ↑										
JiGen [2]	93.92±3.33	85.91±9.41	92.63±3.26	94.03±3.72	91.62	82.26±10.14	70.68±19.70	83.32±9.98	**85.70±8.31**	80.47	86.06
BigAug [11]	93.49±3.93	86.18±6.80	92.09±3.72	93.67±6.18	91.36	81.62±10.20	69.46±20.11	82.64±9.95	84.51±9.28	79.56	85.46
Epi-FCR [6]	94.34±3.38	86.22±9.29	92.88±3.43	93.73±3.18	91.79	83.06±10.88	70.25±20.19	83.68±8.70	83.14±9.72	80.03	85.91
RSC [4]	94.50±2.84	86.21±9.67	92.23±3.69	94.15±2.94	91.77	81.77±11.02	69.37±20.77	83.40±9.15	84.82±9.47	79.84	85.80
FedAvg [8]	92.88±4.36	85.73±9.34	92.07±3.75	93.21±4.69	90.97	80.84±10.44	69.71±20.94	82.28±9.44	83.35±9.96	79.05	85.01
ELCFS (Ours)	**95.37±2.39**	**87.52±5.36**	**93.37±3.38**	**94.50±2.80**	**92.69**	**84.13±11.22**	**71.88±19.91**	**83.94±8.63**	85.51±8.45	**81.37**	**87.03**
	Hausdorff Distance (mean±std) ↓										
JiGen [2]	13.12±10.26	20.18±16.29	11.29±5.73	8.15±6.99	13.19	20.88±10.95	23.21±21.35	11.55±9.59	9.23±7.24	16.22	14.71
BigAug [11]	16.91±12.80	19.01±13.26	11.53±6.45	8.76±9.01	14.05	21.21±11.42	23.10±21.54	12.02±9.28	10.47±9.52	16.70	15.39
Epi-FCR [6]	13.02±9.58	18.97±18.95	**10.67±6.55**	8.47±5.82	12.78	19.12±10.74	21.94±18.37	11.50±6.87	10.86±8.69	15.86	14.32
RSC [4]	19.44±14.86	19.26±16.43	13.47±8.21	8.14±4.99	15.08	23.85±12.09	24.01±21.99	11.38±5.77	9.79±9.45	17.25	16.16
FedAvg [8]	17.01±11.95	20.68±19.01	11.70±6.64	9.33±8.26	14.68	20.77±11.83	26.01±22.91	11.85±6.48	10.03±9.01	17.17	15.93
ELCFS (Ours)	**11.36±8.83**	**17.10±10.05**	10.83±7.31	**7.24±4.34**	**11.63**	**18.65±11.28**	**19.36±13.10**	**11.17±6.42**	**8.91±6.01**	**14.52**	**13.07**

Table 4. Comparison results on prostate MRI segmentation (with standard division).

Unseen Site	A	B	C	D	E	F	Average
	Dice Coefficient (mean±std) ↑						
JiGen [2]	89.95±2.59	85.81±5.40	84.06±9.50	87.34±3.08	81.32±7.40	89.11±3.47	86.26
BigAug [11]	89.63±2.45	84.62±7.07	83.86±9.58	87.66±3.19	81.20±5.12	88.96±3.16	85.99
Epi-FCR [6]	89.72±2.52	85.39±6.31	84.97±8.78	86.55±3.13	80.63±6.46	89.76±3.17	86.17
RSC [4]	88.86±2.73	85.56±5.96	84.36±9.11	86.21±3.21	79.97±6.87	89.80±3.03	85.80
FedAvg [8]	89.02±2.87	84.48±8.60	84.11±9.48	86.30±3.79	80.38±6.32	89.15±3.71	85.57
ELCFS (Ours)	**90.19±2.65**	**87.17±5.36**	**85.26±9.75**	**88.23±3.35**	**83.02±5.46**	**90.47±2.14**	**87.39**
	Hausdorff Distance (mean±std) ↓						
JiGen [2]	10.51±4.69	11.53±8.83	11.70±5.73	11.49±4.57	14.80±5.88	9.02±2.22	11.51
BigAug [11]	10.68±5.11	11.78±9.12	12.07±6.92	**10.66±6.32**	13.98±4.88	9.73±3.05	11.48
Epi-FCR [6]	10.60±5.38	12.31±9.10	12.29±5.89	12.00±4.72	15.68±5.16	8.81±3.04	11.95
RSC [4]	10.57±4.46	11.84±8.96	14.76±8.05	13.07±5.62	14.79±5.67	8.83±2.26	12.31
FedAvg [8]	11.64±5.19	12.01±9.45	14.86±9.28	11.80±5.52	14.90±5.11	9.30±3.17	12.42
ELCFS (Ours)	**10.30±5.18**	**11.49±9.08**	**11.50±5.57**	11.57±5.74	**11.08±3.65**	**8.31±1.93**	**10.88**

Table 5. P-value for statistical analysis between our approach and different comparison methods on overall Dice score.

	JiGen [2]	BigAug [11]	Epi-FCR [6]	RSC [4]	FedAvg [8]
Optic disc	3.6e-20	7.6e-22	3.5e-12	2.1e-9	1.2e-8
Optic cup	1.1e-16	0.0026	1.6e-7	0.0003	2.0e-21
Prostate	0.0004	2.3e-7	9.2e-5	5.2e-8	2.9e-8

References

[1] N Bloch, A Madabhushi, H Huisman, J Freymann, J Kirby, M Grauer, A Enquobahrie, C Jaffe, L Clarke, and K Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370, 2015. 1

[2] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2

[3] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011. 1

[4] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *ECCV*, 2020. 2

[5] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 1

[6] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1446–1455, 2019. 2

[7] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vin-

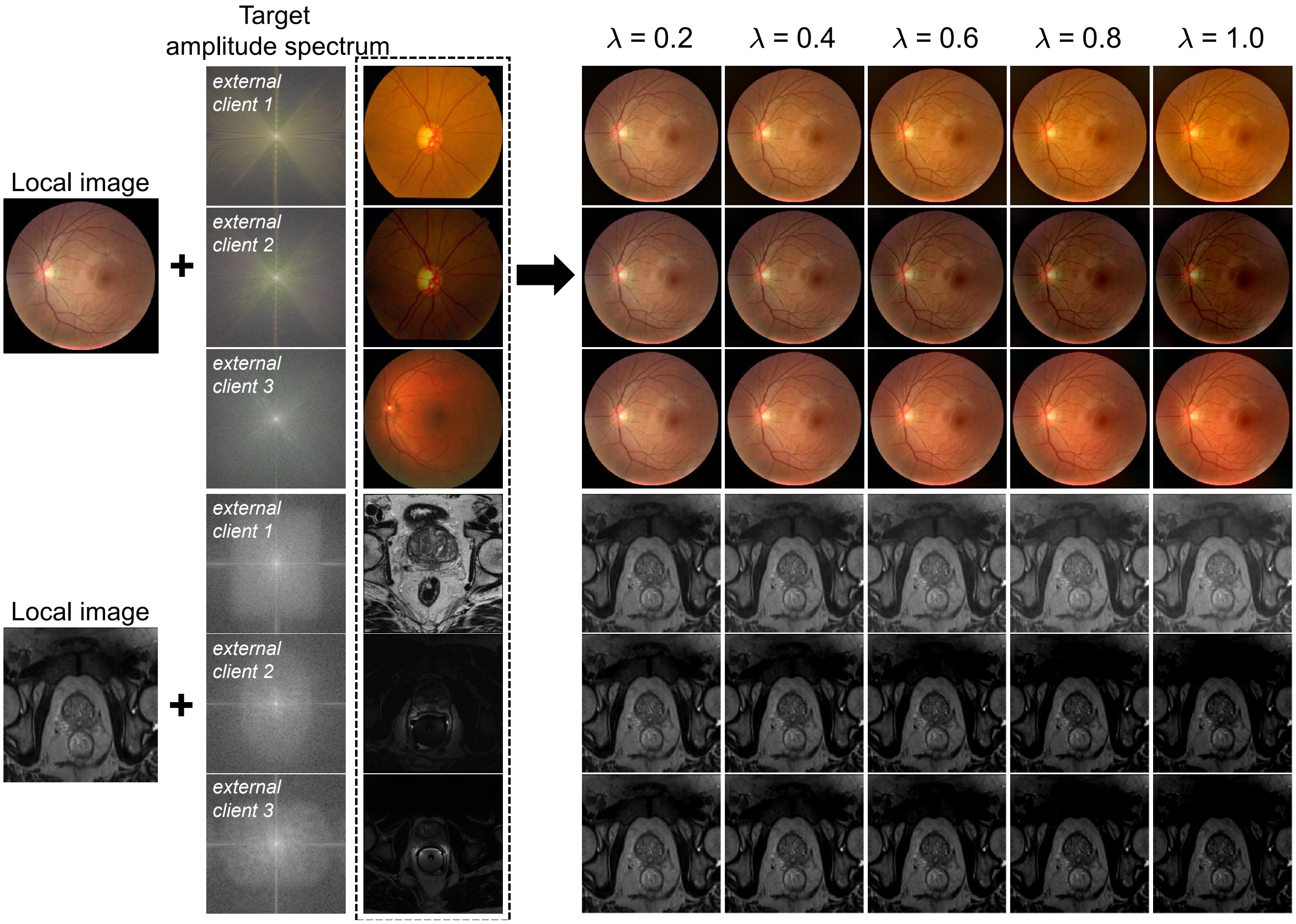

Figure 1. Visualization of transformed images under different interpolation ratio λ.

cent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 1

[8] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017. 2

[9] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 1

[10] Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1):1004, 2015. 1

[11] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 2020. 2