Single-source Domain Generalization in Deep Learning Segmentation via Lipschitz Regularization

Mazlum Ferhat Arslan*(✉), Weihong Guo†, Shuo Li†

* Middle East Technical University, † Case Western Reserve University
mferhata@gmail.com, {wxg49,shuo.li11}@case.edu

Abstract. Deep learning methods have proven useful in medical image segmentation when deployed on independent and identically distributed (iid) data. However, their effectiveness in generalizing to previously unseen domains, where data may deviate from the iid assumption, remains an open problem. In this paper, we consider the single-source domain generalization scenario where models are trained on data from a single domain and are expected to be robust under domain shifts. Our approach focuses on leveraging the spectral properties of images to enhance generalization performance. Specifically, we argue that the high frequency regime contains domain-specific information in the form of device-specific noise and exemplify this case via data from multiple domains. Overcoming this challenge is non-trivial since crucial segmentation information such as edges is also encoded in this regime. We propose a simple regularization method, Lipschitz regularization via frequency spectrum (LRFS), that limits the sensitivity of a model's latent representations to the high frequency components in the source domain while encouraging the sensitivity to middle frequency components. This regularization approach frames the problem as approximating and controlling the Lipschitz constant for high frequency components. LRFS can be seamlessly integrated into existing approaches. Our experimental results indicate that LRFS can significantly improve the generalization performance of a variety of models.

Keywords: Single source domain generalization · Lipschitz regularization · medical image segmentation · frequency spectrum

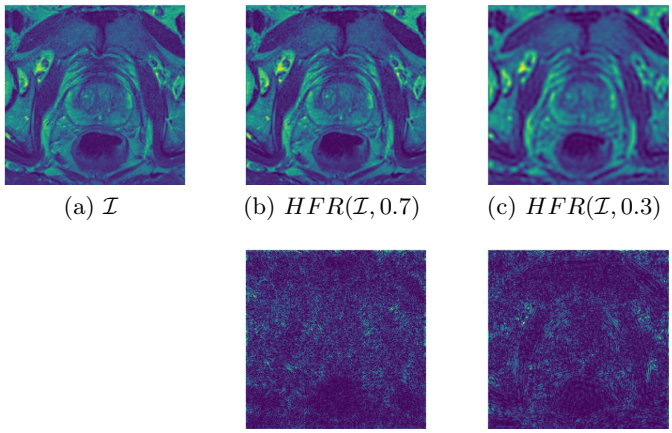
1 Introduction

Biomedical artificial intelligence (AI) technologies have achieved significant improvements over the past two decades thanks to the recent development of deep learning (DL) models. These technologies have the potential to bring a large impact on the health care system by automating the time consuming and labor intensive tasks to improve the efficiency and minimize the intra- or inter-reader variability. The success of DL relies on availability of large size training data (source domain) that need to be drawn from the same distribution with the

testing data (target domain). However, data heterogeneity has become a major challenge in biomedical AI when large training data is required. The heterogeneity is widespread amongst different medical image datasets due to scanners, scanning parameters and subject cohorts, etc. A deep learning model trained on one or multiple source domains might not work well on an unseen new domain due to distribution discrepancy. Therefore, there is an unmet need to develop reliable and effective models and algorithms to address domain inhomogeneity in biomedical deep learning. We aim to address training data uncertainty and tackle the data inhomogeneity problem from domain generalization perspective. Domain generalization refers to the task of training a model on multiple source domains and then applying it to an unseen target domain, without any or very minimal domain-specific adaptation. The goal is to make deep learning models more robust and applicable in real-world scenarios where the data may come from diverse sources. The data scarcity and distribution discrepancy problems are most prominent in the case of single-source domain generalization where data from only a single source is available. This causes DL models that rely on the existence of large training data to perform poorly when deployed on datasets that deviates from the source domain distribution.

In the literature, frequency space methods have been utilized to increase the generalizability of the models. Since phase spectrum of frequency space data contains high-level semantics while the amplitude spectrum relate to domain-specific features such as the style, in [14, 9, 6], researchers consider linear interpolations of amplitude spectrum for data augmentation. In these works low frequency (LF) components are especially targeted as these components relate more to the overall domain-specific features such as the contrast of the images. Another approach is taken by [13, 2] where LF and high frequency (HF) components of the original image are separated, and fused using a neural layer. In [7], the spectral properties of feature maps at different layers of a neural network are utilized to learn attention masks which helps in enhancing generalizable components.

In this paper, we also focus on domain-specific properties of frequency space representations. Different from the existing work, our approach aims at overcoming over-reliance on the high HF components while encouraging the networks to utilize information encoded in middle frequency (MF) components. The reason that we aim at avoiding the overfit to information encoded in HF regime is because measurement device-specific noise is encoded in this regime. However, crucial information for medical image segmentation, *i.e.* the edges, are also encoded in HF components. This necessitates a careful treatment of HF information. Therefore, in the proposed approach we intend only to reduce the sensitivity of the model to HF information and not discard it completely. Also, we argue that MF components pertain to the structural information that shows more domain-invariant characteristics, remarking that both LF and HF components are likely to encode domain-specific information. Hence, we promote the sensitivity to MF information during training. For these purposes, we consider the Lipschitz constant as an indicator of a model's sensitivity, and propose to regularize it by means of approximations of this quantity. The proposed approach,



(a) $\mathcal{I}$ (b) $HFR(\mathcal{I}, 0.7)$ (c) $HFR(\mathcal{I}, 0.3)$

(d) $|\mathcal{I} - HFR(\mathcal{I}, 0.7)|$ (e) $|\mathcal{I} - HFR(\mathcal{I}, 0.3)|$

Fig. 1. Sample (a) original ($\mathcal{I}$) and high-frequency removed (HFR) images for removed frequencies thresholds (b) $\nu = 0.7$ and (c) $\nu = 0.3$ for a prostate slice from the RUNMC dataset. Note the blurring edges with the increase in the removed HF components. Corresponding differences with the original image are shown in (d) and (e). Images are normalized for better visualization (best seen in digital).

Lipschitz **R**egularization via **F**requency **S**pectrum (LRFS), is model-agnostic and our experiments advocate for its effectiveness.

In Fig. 1 we visualize a slice from an MRI image (a), together with images for which HF (b) and MF (c) information of the original are removed. The proposed LRFS aims to reduce the sensitivity of a model to changes similar to that shown in (d) while increasing the sensitivity for changes of type (e).

In what follows, we present the proposed method in Section 2, experimental setup and results in Section 3, and our conclusions and future work in Section 4.

2 Proposed Method

The discrete Fourier transform of an image $\mathcal{I}$ of height H and width W is defined as

$$\mathcal{F}(\mathcal{I})(u,v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathcal{I}(h,w) \exp\left\{-2\pi i \left(\frac{h}{H}u + \frac{w}{W}v\right)\right\}. \tag{1}$$

We refer to the transformed image $\mathcal{F}(\mathcal{I})$ as the k-space data. We define a high frequency removed (HFR) image as

$$HFR(\mathcal{I}(h,w), \nu) = |\mathcal{F}^{-1}(LP(\mathcal{F}(\mathcal{I}); \nu))(h,w)|, \tag{2}$$

where $\mathcal{F}^{-1}$ denotes the inverse discrete Fourier transform, LP is a low-pass function

$$LP(\mathcal{F}(\mathcal{I}); \nu)(u,v) = \begin{cases} 0 & \text{if } k(u,v) \geq \nu \\ \mathcal{F}(\mathcal{I})(u,v) & \text{otherwise,} \end{cases} \tag{3}$$

and $k(u,v) = \frac{1}{2}\left(\frac{u^2}{U^2} + \frac{v^2}{V^2}\right)^{1/2}$ where $U \coloneqq H$, $V \coloneqq W$ and the constant multiplier serves to restrict the range of k to $[0,1]$ rather than $[0,2]$. In the proposed method, as explained below, we make use of HFR images that correspond to middle and high frequency removed images.

High-frequency removal may appear as a data augmentation method. Yet, rather than introducing new synthetic information to a sample, existing information in the k-space data is removed, which limits the number of possible images to be generated. Further, HFR images do not necessarily represent realistic variations of the original sample; see for example Fig. 1 (c). Therefore, HFR may not provide useful augmentations for end-to-end training. For these reasons, we use HFR images to regulate only the feature extracting branch, and prevent training the segmentation branch with possibly unrealistic images.

Definition. A function $f : X \to Y$ is called Lipschitz continuous with a Lipschitz constant $\kappa > 0$ if for all $x_1, x_2 \in X$, $\|f(x_1) - f(x_2)\|_Y \leq \kappa \|x_1 - x_2\|_X$ where $\|.\|_X$ denotes the metric on X, and similarly for $\|.\|_Y$.

Computing the Lipschitz constant of a model with ReLU activations is an NP-hard problem [12]. Thus, denoting the feature extracting (encoder) branch of a segmentation network as Φ and the segmentation (decoder) branch as Ψ, we *approximate* the Lipschitz constant of Φ for a fixed ν as

$$\kappa_\nu = \frac{\|\Phi(\mathcal{I}) - \Phi(HFR(\mathcal{I},\nu))\|_F}{\|\mathcal{I} - HFR(\mathcal{I},\nu))\|_F} \tag{4}$$

where Φ is assumed to be Lipschitz. We consider the approximated quantity κ_ν as an indicator of the encoder's sensitivity to high frequency removals for frequency ν. The higher the κ_ν, the more sensitive the encoder, and vice versa.

In order to reduce the sensitivity of Φ to high frequency (HF) components we propose the following regularization term

$$\mathcal{L}_{HF} = \text{ReLU}\left(\frac{1}{\hat{\kappa}_{\nu_{HF}}}\frac{\|\Phi(\mathcal{I}) - \Phi(HFR(\mathcal{I},\nu_{HF}))\|_F}{\|\mathcal{I} - HFR(\mathcal{I},\nu_{HF}))\|_F} - 1\right) = \text{ReLU}\left(\frac{\kappa_{\nu_{HF}}}{\hat{\kappa}_{\nu_{HF}}} - 1\right) \tag{5}$$

where $\hat{\kappa}_{\nu_{HF}}$ is the reference Lipschitz constant chosen for HF components. However, it is possible for a network to trivially minimize this loss term without any functional change by simply scaling Φ by a constant, which can be inverted by Ψ. For example, for $s \in \mathbb{R}\backslash\{0\}$, the encoder-decoder pair $(\tilde{\Phi}, \tilde{\Psi})$ defined as $\tilde{\Phi}(x) \coloneqq \Phi(x)/s$ and $\tilde{\Psi}(z) \coloneqq \Psi(sz)$ can reduce the Lipschitz constant of Φ without changing the behavior of $\Psi \circ \Phi$.

Thus, we introduce a second loss term that aims at *increasing* the sensitivity to middle frequency (MF) components with the premise that MF components carry structural information:

$$\mathcal{L}_{MF} = \text{ReLU}\left(1 - \frac{1}{\hat{\kappa}_{\nu_{MF}}}\frac{\|\Phi(\mathcal{I}) - \Phi(HFR(\mathcal{I},\nu_{MF}))\|_F}{\|\mathcal{I} - HFR(\mathcal{I},\nu_{MF}))\|_F}\right) = \text{ReLU}\left(1 - \frac{\kappa_{\nu_{MF}}}{\hat{\kappa}_{\nu_{MF}}}\right) \tag{6}$$

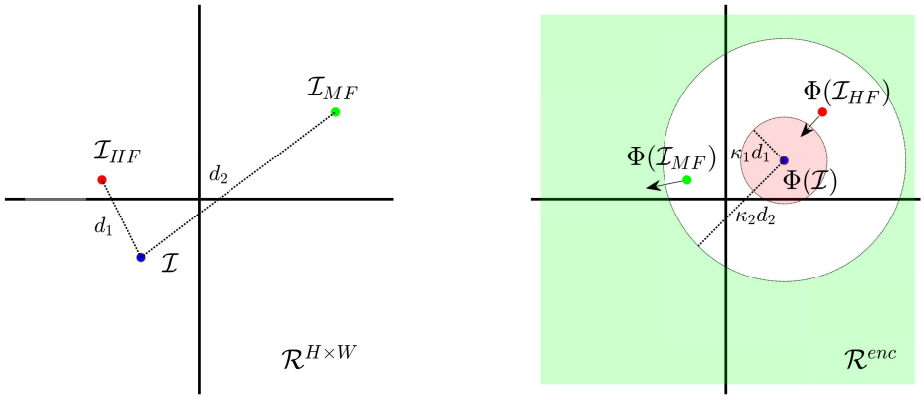


Fig. 2. The encoder Φ maps the image space of dimensions $H \times W$ to the latent space of dimensions *enc*. The circular regions in the latent space are determined by image space distances d_1 and d_2, and the predetermined Lipschitz constants κ_1 and κ_2 for HF and MF components, respectively. For this example, our regularization scheme would push $\Phi(\mathcal{I}_{HF})$ towards the pink region and $\Phi(\mathcal{I}_{MF})$ to the green region (indicated by arrows).

where $\hat{\kappa}_{\nu_{MF}}$ is the reference Lipschitz constant for MF components.

The loss function $\mathcal{L}_{MF}$ is non-zero only when the estimated Lipschitz constant $\kappa_{\nu_{MF}}$ for MF components is lesser than $\hat{\kappa}_{\nu_{MF}}$, while $\mathcal{L}_{HF}$ activates only when $\kappa_{\nu_{HF}}$ is larger than $\hat{\kappa}_{\nu_{HF}}$. For example, in Fig. 2, the image space distances d_1 and d_2 between the original image and HFR images determine circular regions of radii $\kappa_1 d_1$ and $\kappa_2 d_2$ in the latent space where κ_1 and κ_2 are some predetermined Lipschitz constants. In the given example, the latent representations $\Phi(I_{HF})$ and $\Phi(I_{MF})$ would be pushed towards the red and green regions, respectively, by the proposed regularization losses. Once the latent representations fit into their corresponding target regions, no more regularization is applied on them. When $\kappa_1 = \kappa_2$, the white region in-between is determined by the information encoded in the MF regime of the original image $\mathcal{I}$.

Note that the proposed framework is similar to contrastive learning which aims to keep positive pairs, *i.e.* a pair of samples sharing similar properties, close to each other and keep negative pairs distant from one another. Despite the similarity, our approach is different due to being based on Lipschitz constants rather than absolute distances between latent representations.

The proposed objective function for an arbitrary network with loss function $\mathcal{L}$ is

$$\mathcal{L}_{proposed} = \mathcal{L} + \lambda_{MF}\mathcal{L}_{MF} + \lambda_{HF}\mathcal{L}_{HF}. \tag{7}$$

In practice, we used $\lambda_{MF} = \lambda_{HF}$. The introduced losses $\mathcal{L}_{MF}$ and $\mathcal{L}_{HF}$ are optimized through $\Phi(I)$ alone, that is, $\Phi(HFR(I,.))$ are treated as constants.

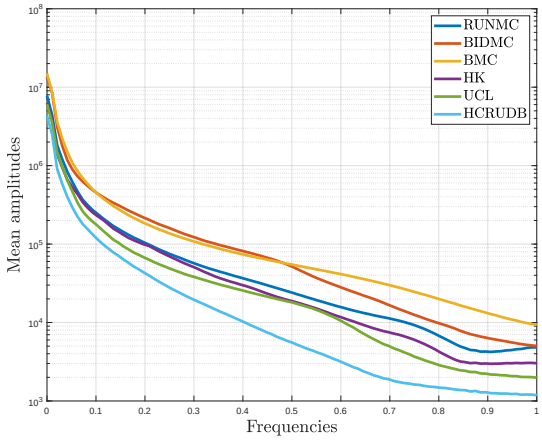


Fig. 3. Histograms of mean amplitudes of frequencies are domain-specific.

3 Experimental Setup and Results

3.1 Datasets

In our experiments we use T2-weighted prostate MRI data collected from six different sites. The data from sites RUNMC and BMC are from NCI-ISBI13 dataset [1], HCRUDB is from I2CVB dataset [5], and UCL, BIDMC and HK are from PROMISE12 dataset [8]. We utilize the preprocessed data used in [10][1]. During training, we only use slices with non-zero labels. We use RUNMC, which is comprised of scans from 30 patients, as our source domain and follow a 70%/10%/20% train-validation-test split strategy to train the models.

We present amplitude versus frequency histograms of the considered datasets in Fig. 3 to demonstrate domain-specific characteristics. In the presented plot, the frequency range is indicated in normalized k-space coordinate norms. We sampled $k(u, v)$ at 100 discrete points $\{k_i\}_{i=1}^{100}$ (bins of the histogram) from slices that include a part of the target organ and calculate mean amplitudes $|\mathcal{F}(\mathcal{I})(u, v)|$ where $k_i \leq k(u, v) < k_{i+1}$.

3.2 Implementation details

The experiments are implemented using PyTorch and run on Colab's T4 and V100 GPUs. For the implementations of UNet [11] and UNet with residual connections (ResUNet) [15] we utilize the MONAI framework. For both, we use instance normalization (IN) rather than batch normalization, as IN yielded better generalization performance in our experiments. UNet is initialized with default parameters of the framework, and ResUNet is initialized with number of residual units set to 2, and channel sizes set to $(16, 32, 64, 128, 256)$ for each level where

[1] `https://liuquande.github.io/SAML/`

a stride of 2 is used for downsampling between the levels. For ResUNet++ [4] we use the official implementation[2]. For UNet, ResUNet, and ResUNet++ the number of prediction channels are set to 2. While evaluating a network's performance, we use argmax predictions to calculate Dice scores. All three of the models are trained with $\mathcal{L}$ in Eq. (7) set to Dice loss, a batch size of 16 for 1000 epochs with Adam optimizer (with a momentum of 10^{-4}), polynomial learning rate decay (initial learning rate is 10^{-3} and power of the polynomial is 0.9), and a linear warm-up scheduler used during the first 5 epochs. For BayeSeg [3], we use the official implementation[3]. Due to memory limitations, we use a batch size of 16. The rest of the training configuration are as in the original paper.

We use rotation, scaling, translation, elastic transformation, intensity normalization, and Gaussian noise addition for data augmentation. Since the features learned during the early stages are of generalizable nature, we apply Lipschitz regularization after the 100th epoch for each model in order not to interfere with or limit the learning process at early stages. For UNet, ResUNet and ResUNet++, we used $\lambda_{MF} = \lambda_{HF} = 3 \times 10^{-4}$ and for BayeSeg 1×10^{-2} is used, *cf.* Table 2. We considered the $k(u, v)$ values in the interval $[0, 0.3)$ as LF, $[0.3, 0.7)$ as MF, and $[0.7, 1.0]$ as HF regimes, hence used $\nu_{MF} = 0.3$ and $\nu_{HF} = 0.7$. Favoring simplicity, we set $\hat{\kappa}_{\nu_{MF}} = \hat{\kappa}_{\nu_{HF}} = 1$ and $\lambda_{HF} = \lambda_{MF}$ in our experiments.

3.3 Results

We tested the proposed method in a controlled experiment fashion and compared model performances[4] with and without the Lipschitz regularization. We tabulate our results in Table 1 in terms of Dice scores. For all four of the models, Lipschitz regularization provided significant improvements on average target domain performances. Notably, for all models except ResUNet++, the Dice scores on the source domain improved as well.

Although $\lambda_{HF} = \lambda_{MF} = 3 \times 10^{-4}$ may work well with different model architectures, one might need to search for optimal parameters when the loss function or the weight decay parameter is changed. In Table 2 we present how the performance of BayeSeg, which uses cross-entropy loss together with a custom variational loss, changes with respect to these parameters. Observe that the average performance on targets suggest a consistent trend. The parameters yield no distinguishable performance change for $\lambda = 3 \times 10^{-4}$ with all changes being well within the respective standard deviations, while $\lambda = 1 \times 10^{-1}$ yield a clearly worse model. The optimal parameters for this model are probably included in the range $[3 \times 10^{-3}, 1 \times 10^{-2}]$, though we opted for 1×10^{-2} while searching for parameters for this model. We also note that the standard deviations on target domains are lowered when Lipschitz regularization is applied, especially for the dataset HCRUDB ($\sigma = 20.3$ for baseline vs. $\sigma = 7.6$ for the best model), indicating a more robust prediction across different patient and slice data.

[2] `https://github.com/DebeshJha/ResUNetPlusPlus`
[3] `https://github.com/obiyoag/BayeSeg`
[4] `https://github.com/kaptres/LRFS`

Table 1. Single-source domain generalization performances in terms of Dice scores for a variety of models with and without Lipschitz regularization.

	(source)	(targets)					Avg on
	RUNMC	BIDMC	BMC	HK	UCL	HCRUDB	targets
UNet	88.09	47.19	78.34	77.98	82.40	80.20	73.22
+Lipschitz	88.51	60.51	80.44	85.85	81.56	79.75	**77.62**
ResUNet	86.97	56.12	79.03	82.75	78.61	77.57	74.82
+Lipschitz	88.61	62.94	80.32	83.62	79.86	80.74	**77.51**
ResUNet++	89.17	60.71	69.95	86.85	78.20	66.87	72.51
+Lipschitz	88.22	64.52	73.31	85.93	77.33	78.74	**75.97**
BayeSeg	85.0	71.2	82.6	82.0	79.7	73.5	77.8
+Lipschitz	87.4	76.9	83.3	84.8	82.2	82.5	**81.94**

Table 2. Model performances in terms of Dice scores for different loss weight parameters used with BayeSeg. The standard deviations are calculated on the Dice scores of all the slices in a domain.

	(source)	(targets)					Avg. on
$\lambda_{HF} = \lambda_{MF}$	RUNMC	BIDMC	BMC	HK	UCL	HCRUDB	targets
0	84.9±3.6	71.2±7.6	82.6±5.5	82.0±3.5	79.7±5.6	73.5±20.3	77.8
3×10^{-4}	85.9±4.0	70.7±8.1	82.4±5.5	79.4±5.9	80.7±4.5	74.7±15.6	77.58
3×10^{-3}	86.1±5.3	73.2±9.5	**84.1**±4.7	**84.8**±4.6	**82.6**±5.6	82.4±7.6	81.42
1×10^{-2}	87.4±4.0	**76.9**±5.8	83.3±5.0	**84.8**±3.8	82.2±5.1	**82.5**±7.6	**81.94**
3×10^{-2}	**87.5**±4.7	59.6±12.8	78.5±8.1	82.7±5.2	78.2±5.2	80.8±6.9	75.96
1×10^{-1}	83.1±6.7	62.8±10.4	72.3±11.4	78.9±8.1	74.4±8.0	77.4±9.9	73.16

4 Conclusions and Future Work

In this work, we proposed Lipschitz regularization via frequency spectrum (LRFS) which is a simple yet effective regularization strategy for improving the generalizability of DL models. LRFS is generic in its nature and can be utilized to regularize various model architectures to acquire significant performance improvements without compromising the performance on the source domain, as demonstrated by our experiments.

Despite being initially tailored for medical image segmentation LRFS can provide generalizability for other DL tasks. Thus, for future work, we plan to explore LRFS's utility in diverse DL applications.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K.: Adversarial consistency for single domain generalization in medical image segmentation. In: The Cancer Imaging Archive (2015), http://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv
2. Fu, G., Jimenez, G., Loizillon, S., Chougar, L., Dormont, D., Valabregue, R., Burgos, N., Lehéricy, S., Racoceanu, D., Colliot, O., et al.: Frequency disentangled learning for segmentation of midbrain structures from quantitative susceptibility mapping data. arXiv preprint arXiv:2302.12980 (2023)
3. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. arXiv preprint arXiv:2303.01710 (2023)
4. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE international symposium on multimedia (ISM). pp. 225–2255. IEEE (2019)
5. Lemaître, G., Martí, R., Freixenet, J., Vilanova, J.C., Walker, P.M., Meriaudeau, F.: 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)
6. Li, H., Li, H., Zhao, W., Fu, H., Su, X., Hu, Y., Liu, J.: Frequency-mixed single-source domain generalization for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 127–136. Springer (2023)
7. Lin, S., Zhang, Z., Huang, Z., Lu, Y., Lan, C., Chu, P., You, Q., Wang, J., Liu, Z., Parulkar, A., et al.: Deep frequency filtering for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11797–11807 (2023)
8. Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., Van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis **18**(2), 359–373 (2014)
9. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1013–1023 (2021)
10. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE Transactions on Medical Imaging (2020)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
12. Virmaux, A., Scaman, K.: Lipschitz regularity of deep neural networks: analysis and efficient estimation. Advances in Neural Information Processing Systems **31** (2018)
13. Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y.S., Loy, C.C.: Masked frequency modeling for self-supervised visual pre-training. arXiv preprint arXiv:2206.07706 (2022)

14. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14383–14392 (2021)
15. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018)