A Detail of Architectures and Experimental Settings

A.1 Experimental Setting for Skin Lesion Classification Task

We use a ResNet18 model [1] pretrained on ImageNet without the FC layer as the feature extractor Q_θ with the input size 224×224 for our proposed method as well as other baselines. For our method, the network before average pooling is used as the feature extractor. We insert a variational encoding network between the feature extractor and the final fully connected layer which acts as the classifier T_ϕ. The variational encoding network F_ω is implemented using two separated networks with the same architecture, including a fully connected layer with the output dimension as 512, a Relu activation layer and a fully connected layer with the output dimension as 80. We then conduct reparameterization trick to obtain the output of variational encoding network. The classification network T_ϕ is a fully connected layer with the output size as 7.

For the hyperparameters, we choose $\lambda_1 = 0.001$ and $\lambda_2 = 0.4$ for all settings. Due to the class imbalance within and across datasets, we adopt the focal loss [2] as the classification objective for our proposed method as well as other baseline techniques. For implementation, the alternate form proposed in [2] is adopted as it is an extension of cross-entropy loss and is also bounded and convex, which satisfies our assumption in the manuscript. During training, the Adam optimizer is used with learning rate as 0.0001, weight decay as 0.001 and the size of minibatch as 32. We train the models for 200 epochs and the learning rate is decreased by a factor 10 after every 80 epochs. For evaluation on testing set, we use the best performing model on the validation set.

A.2 Experimental Setting for Spinal Cord Gray Matter Segmentation Task

We adopt 2D-UNet [4] (without the last 1×1 convolutional layer) as the backbone network by considering the MRI axial slice as input. The variational encoding network F_ω is implemented using two separated network with an identical architecture which includes a latent layer using 1×1 convolutional layer with output channel as 64, a Relu activation layer and a 1×1 convolution layer to predict mean and standard deviation layers of the distribution with output channel as 8 . We then conduct reparameterization trick to obtain the output of variational encoding network. We further adopt a $1{\times}1$ convolution layer as the classification network T_ϕ with the output channel size as 2 for segmentation purpose.

For the hyperparameters, we use $\lambda_1 = 0.001$ and $\lambda_2 = 0.01$. We adopt the weighted binary cross-entropy loss for classification, where the weight of a positive sample is set to the reciprocal of the positive sample ratio in the region of interest. We use Adam algorithm with learning rate as 1e-4, weight decay as 1e-8 and the batch size as 8 for each domain for training. We train the model for 200 epochs, where the learning rate is decreased every 80 epoch with a factor of 10. For data processing, the 3D MRI data is first sliced into 2D in axial slice view and then center cropped to 160×160. We further conduct random cropping which leads to the size as 144×144 for training.

B BigAug [5] for Segmentation

We present here by considering the data-augmentation based domain generalizing method BigAug [5], which stacked different types of transformations, including sharpness, blurriness, noise, brightness, contrast, rotation, scaling, etc., by considering 2D-UNet for spinal cord gray matter segmentation task [3]. The results are shown in Table 1 (a). As we can observed, the performances are not desired by directly adopting the default parameters for augmentation in [5], which are even worse than the “DeepAll" baseline in terms of DSC and JI.

To understand the reason why BigAug [5] with default parameter setting leads to negative transfer, we visualize in Figure 1 some examples of the transformed input and groundtruth pairs by considering both the groundtruth of spinal cord and gray matter. As we can see, by conducting the augmentation with default parameters in [5], the quality of input deteriorates and the boundary can be oversmooth, which may further lead to more discrepancy between source and target domain.

We further consider to adopt the same augmentation in [5] by tuning the parameters in a wide range to report the best segmentation performance for this task. The results are shown in Table 1 (b). We observe that there exists some improvement compared with “DeepAll" baseline, but our proposed

Table 1: Domain generalization results on gray matter segmentation task using BigAug [5].

(a) Default Parameters [5]

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.7675	38.47	0.6250	0.7798	0.1286
1,3,4	2	0.7542	34.50	0.6061	0.9187	0.1013
1,2,4	3	0.5468	-76.2	0.3809	0.6381	1.9013
1,2,3	4	0.8706	70.18	0.7712	0.9232	0.0437
Average		0.7348	16.74	0.5958	0.8150	0.5437

(b) Tuned Parameters

source	target	DSC	CC	JI	TPR	ASD
2,3,4	1	0.8438	62.02	0.7334	0.8600	0.1613
1,3,4	2	0.7703	40.17	0.6269	0.8866	0.1802
1,2,4	3	0.5556	-73.7	0.3905	0.6282	1.5560
1,2,3	4	0.8891	74.94	0.8009	0.8827	0.0362
Average		0.7647	25.86	0.6379	0.8144	0.4834

method can still outperform [5] with parameter tuning, which is reasonable as it is difficult to choose a suitable augmentation type and magnitude for different medical imaging classification tasks.

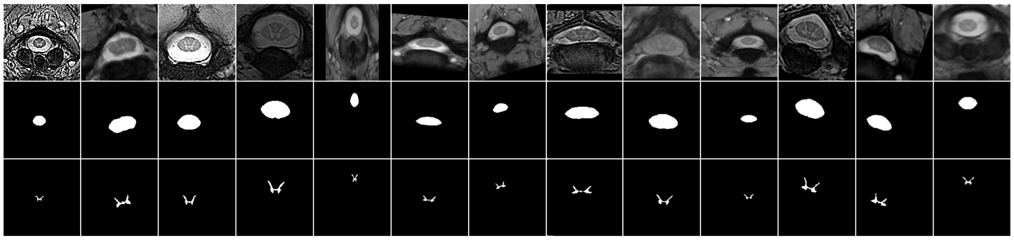

Figure 1: The samples of input and ground truth pairs generated from BigAug using the default hyper parameters. The first row shows the input, the second row shows the groundtruth of spinal cord, and the last row shows the groundtruth gray matters.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[3] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener, et al. Spinal cord grey matter segmentation challenge. *Neuroimage*, 152:312–329, 2017.

[4] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[5] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, H. Roth, A. Myronenko, D. Xu, and Z. Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 2020.