Enhancing Pseudo Label Quality for Semi-Supervised Domain-Generalized Medical Image Segmentation

Huifeng Yao[1], **Xiaowei Hu**[2], **Xiaomeng Li**[1,3*]

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[3] The Hong Kong University of Science and Technology Shenzhen Research Institute
eehfyao@ust.hk, xwhu@cse.cuhk.edu.hk, eexmli@ust.hk

Abstract

Generalizing the medical image segmentation algorithms to unseen domains is an important research topic for computer-aided diagnosis and surgery. Most existing methods require a fully labeled dataset in each source domain. Although (Liu et al. 2021b) developed a semi-supervised domain generalized method, it still requires the domain labels. This paper presents a novel confidence-aware cross pseudo supervision algorithm for semi-supervised domain generalized medical image segmentation. The main goal is to enhance the pseudo label quality for unlabeled images from unknown distributions. To achieve it, we perform the Fourier transformation to learn low-level statistic information across domains and augment the images to incorporate cross-domain information. With these augmentations as perturbations, we feed the input to a confidence-aware cross pseudo supervision network to measure the variance of pseudo labels and regularize the network to learn with more confident pseudo labels. Our method sets new records on public datasets, *i.e.*, M&Ms and SCGM. Notably, without using domain labels, our method surpasses the prior art that even uses domain labels by 11.67% on Dice on M&Ms dataset with 2% labeled data. Code is available at https://github.com/XMed-Lab/EPL_SemiDG.

Introduction

Medical image segmentation is one of the fundamental tasks in computer-aided diagnosis and computer-aided surgery. In recently years, researchers have developed many convolutional neural networks for medical image segmentation, such as U-Net (Ronneberger, Fischer, and Brox 2015), DenseUNet (Li et al. 2018a), nnUNet (Isensee et al. 2021), and HyperDenseNet (Dolz et al. 2018). Medical images are usually collected from different clinical centers with different scanners (Li et al. 2020c; Puyol-Anton et al. 2021; Wang et al. 2020). As a result, they may have apparent domain shifts due to variation in patient populations, scanners, and scan acquisition settings; see examples in Figure 1. However, the above methods generate inferior results when testing on images from unseen domains; see results in Table 1 and Table 2 in the experiments. Hence, it is crucial to strengthen the model's generalization ability over different domain shifts.

*Corresponding Author: eexmli@ust.hk

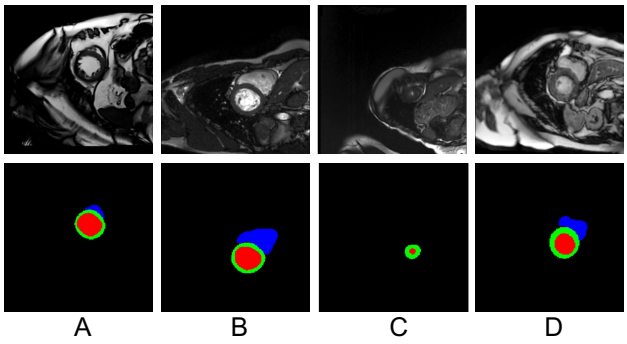

Figure 1: Some 2D slices of MRI from M&Ms dataset. The first row is images from four different domains (A,B,C,D) and the second row is the corresponding segmentation masks. The red, blue, and green colors refer to the left ventricle blood pool,the right ventricle blood pool, and the left ventricular myocardium, respectively.

One naive solution to address the domain shifts is to obtain and annotate as much data as possible. However, the annotation cost is quite expensive to the community. Another solution is to train the model on the source domains and generalize it to the target domain with some information from the target domain, namely domain adaptation (DA) (Bian et al. 2020; Pomponio et al. 2020). For example, Pomponio et al. (2020) developed cross-site MRI harmonization to enforce the source and target domains to share similar image-specific characteristics. Domain generalization (DG) is a more strict setting, where the difference with DA is that the model *does not use any information from the target domain*. In this paper, we focus on this challenging but more practical problem, *i.e.*, train the model on the source domains (A, B, and C) and test it on an unseen domain (D).

Existing domain generalization methods have been developed for various medical image applications, such as cardiac segmentation (Liu et al. 2020), skin lesion classification (Li et al. 2020a), spinal cord gray matter segmentation (Li et al. 2020a) and prostate MRI segmentation (Liu, Dou, and Heng 2020). For example, Liu et al. (2020) proposed two data augmentation methods: "Resolution Augmentaion" and "Factor-based Augmentation" to generate more diverse data.

Li et al. (2020a) proposed to learn a representative feature space through variational encoding to capture the semantic-consistent information among multiple domains. However, the above methods require that each domain should have fully-labeled datasets, and are not applicable if only partial data in each domain is labeled. Recently, Liu et al. (2021b) considered addressing this problem and presented a meta-learning approach with disentanglement for semi-supervised domain generalization. However, their method requires the source domain labels, which may not easily be obtained in clinical practice. Here, we consider a more practical setting for semi-supervised domain-generalized medical image segmentation: *training data consists of labeled and unlabeled images from three source domains without knowing domain labels, and test data is from an unknown distribution*.

One straightforward solution is to directly use semi-supervised semantic segmentation methods (Chen et al. 2021; Lai et al. 2021; Lee, Kim, and Yoon 2021). For example, Chen et al. (2021) introduced a cross pseudo supervision method, where two segmentation networks are randomly initialized and supervised separately by the corresponding pseudo labels. The unlabeled images in semi-supervised semantic segmentation usually are from the same distribution as the labeled one; thus, the pseudo labels can be directly used to refine another segmentation module. However, in our problem, the unlabeled images are from an unknown distribution, leading to a biased pseudo label.

To this end, we introduce a novel confidence-aware cross pseudo supervision algorithm. The key idea is that two segmentation networks that shared the same structure are initialized differently. Hence, we can encourage consistent outputs from two networks. For the unlabeled images, each segmentation network can generate a pseudo label, which can be used as an additional signal to supervise the other segmentation network. To improve the quality of pseudo labels for unlabeled images from unknown domains, we propose to use Fourier transformation for unlabeled images, which can not only help obtain the low-level statistic information across different domains, but also augment the image samples by incorporating the cross-domain information. Then, we develop the confidence-aware regularization to measure the pseudo variances generated by the original image and the image augmented by Fourier transformation. It helps improve the quality of pseudo labels, thus facilitating the learning process of the whole framework.

This paper has the following contributions:

- We present a simple yet effective method for semi-supervised domain-generalized medical image segmentation. Our key idea is to enhance the quality of pseudo labels for unlabeled images from unknown distributions.
- We introduce two innovative modules: Fourier-transform-based data augmentation to augment the samples with cross-domain information, and confidence-aware cross pseudo supervision to measure the variances of pseudo labels.
- Our method achieves the new state-of-the-art performance on M&Ms and SCGM dataset. Notably, without using domain labels, our method surpasses the prior art that even uses domain labels by 11.67% on Dice on M&Ms dataset with only 2% labeled data.

Related Work

Domain Generalization

Domain generalization (DG) trains the model with multiple source domains and generalizes it to unseen target domains. Existing DG methods can be categorized into three classes: representation learning (Zhou et al. 2020), learning strategy design (Yeo, Kar, and Zamir 2021), and data manipulation (Tobin et al. 2017). Representation learning methods mainly follow the idea of domain adaptation by learning domain-invariant features or explicitly feature alignment between domains. For example, Zhou et al. (2020) aligned the marginal distribution of different source domains via optimal transport by minimizing the Wasserstein distance to achieve domain-invariant feature space. Learning strategy methods focus on exploiting the general learning strategy to promote the generalization capability. For example, Yeo, Kar, and Zamir (2021) used the ensemble learning method to ensembling the middle domains into one strong prediction with uncertainty as to their weight. Data manipulation aims to manipulate the inputs to assist in learning general representations. For example, Tobin et al. (2017) first used this method to generate more training data from the simulated environment for generalization in the real environment.

Recently, the Fourier transformation method has also proven to be very effective for DG. Xu et al. (2021) proposed a Fourier-transform-based framework for domain generalization by replacing the amplitude spectrum of a source image with that of another image from a different source domain. Their method achieves a remarkable performance in classification tasks. Liu et al. (2021a) also used the Fourier transformation method in federated learning and proves that it is a useful augmentation for medical image segmentation under federated learning.

Unlike the above methods, we use the Fourier transformation as a data augmentation to get low-level statistic information among different domains and incorporate cross-domain information for unlabeled images. With these perturbations, we can measure the confidences of pseudo labels, and then enhance the model performance with reliable pseudo labels.

Semi-supervised Semantic Segmentation

Unlike the image classification task, manually labeling pixel-wise annotations for the segmentation task is expensive and time-consuming. Existing methods for semi-supervised segmentation can be broadly classified into two categories: self-training (Lee et al. 2013) and consistency learning (Tarvainen and Valpola 2017; Li et al. 2018b; Yu et al. 2019; Li et al. 2020b). The self-training method mainly uses the pseudo-label (Lee et al. 2013) method to improve performance. This is initially developed for using unlabeled data in classification tasks. Then, it is also applied for semi-supervised segmentation (Feng et al. 2020; Ibrahim et al. 2020) recently. This method uses the pseudo segmentation maps of unlabeled data acquired by the previously trained

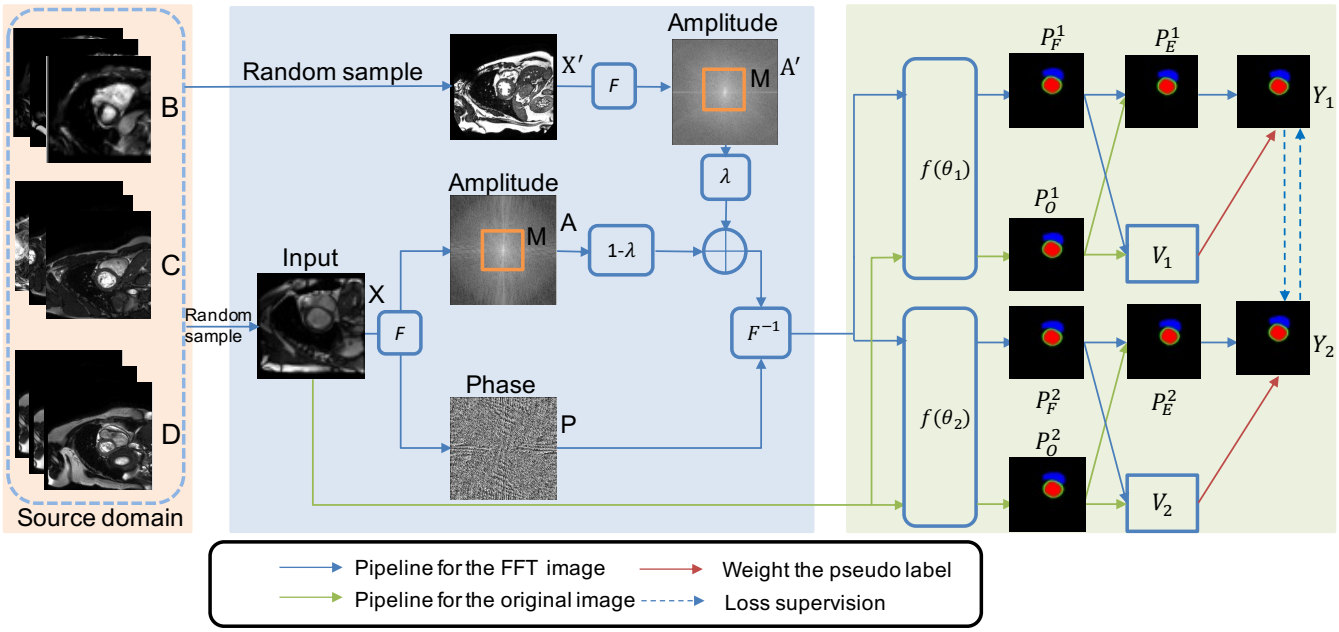

Figure 2: The overall architecture of the proposed confidence-aware cross-pseudo supervision network. B, C, and D are the source domains and A is the target domain. Note that the network does not use domain labels (*i.e.*, B, C, and D) during training.

segmentation model on labeled data to retrain the segmentation model. The process can be iterated multiple times. Consistency learning encourages the model to have a similar output for the same input after different transformations. It imposes the consistency constraint between the predictions of different augmented images so that the decision function lies in the low-density region. Mean teacher (Tarvainen and Valpola 2017) is a famous work of consistency learning in the classification task. Many works (Kim, Jang, and Park 2020; French et al. 2019) also use this in semi-supervised segmentation tasks. Recently, many methods combined these two ideas to get better performance in segmentation tasks. PseudoSeg (Zou et al. 2020) performed both strong and weak augmentation on the same input images and used the weakly augmented image as the pseudo label. Cross pseudo supervision(CPS) (Chen et al. 2021) also employed the pseudo-label method to expand the dataset and used consistency learning to learn a more compact feature encoding. Unlike these methods where unlabeled images are from the same domain, our method involves unlabeled images from an out-of-domain distribution, so the pseudo label generated by the network may not be reliable. Hence, we develop two innovative modules to enhance the quality of pseudo labels: one is Fourier-transform to augment cross-domain information for a single input from the source domain; the other is to measure the confidence of pseudo labels to get reliable supervision signals.

Methodology

Data Augmentation by Fourier Transformation

Figure 2 illustrates the overall architecture of our method. During the training process, we randomly take an image X from source domain as the input image without knowing the domain label, and then perform the Fourier transformation F to transfer the image to the frequency domain and obtain an amplitude spectrum A and a phase image P, where the amplitude spectrum contains low-level statistics while the phase image includes high-level semantics of the original signal. To improve the capability of domain generalization, we randomly select another sample X' from the source domain, perform the Fourier transformation, and obtain another amplitude A'. Then, we augment the first image X by incorporating the amplitude information of the second image X' by:

$$\mathcal{A}_{new} = (1-\lambda)A * (1-\mathcal{M}) + \lambda\mathcal{A}' * \mathcal{M}\,, \qquad (1)$$

where $\mathcal{A}_{new}$ is the newly generated phase image; λ is a parameter that used to adjust the ratio between the phase information of X and X'; and M is a binary mask to control the spatial range of amplitude spectrum to be exchanged and we set M as the central region of the amplitude spectrum that contains low-frequency information. After that, we transform the merge sample from the frequency domain to the image domain through F^{-1} to obtain the image sample Z that augmented through Fourier transformation and incorporated the low-level information from another sample:

$$Z = \mathcal{F}^{-1}\left(\mathcal{A}_{new}, \mathcal{P}\right). \qquad (2)$$

Confidence-Aware Cross Pseudo Supervision

Next, we take both the original image X and the transformed image Z into two parallel segmentation networks $f(\theta_1)$ and $f(\theta_2)$, where these two networks have the same structure but their weights are initialized differently. For each segmentation network, we obtain the predictions of the the original

image X and the transformed image Z by:

$$\begin{aligned} P_F^1 &= f(\theta_1)(Z), \\ P_O^1 &= f(\theta_1)(X), \\ P_F^2 &= f(\theta_2)(Z), \\ P_O^2 &= f(\theta_2)(X). \end{aligned} \tag{3}$$

For the unsupervised part, since we do not have the supervision signals for the unlabeled data, after obtaining the predictions from different networks, we aim to adopt the prediction from one network to supervise the prediction from another network. This technique is proposed by (Chen et al. 2021) and named as the cross pseudo supervision. However, in (Chen et al. 2021), the label and unlabeled data are from the same domain with less variation, hence, the generated pseudo labels usually have a high quality. In contrast, the data in our task is from multiple domains and the large variance in the training samples from different domains may lead to low-quality pseudo labels. To improve the label quality and reduce the influence of low-quality labels, in this work, we present a confident-aware cross pseudo supervision mechanism.

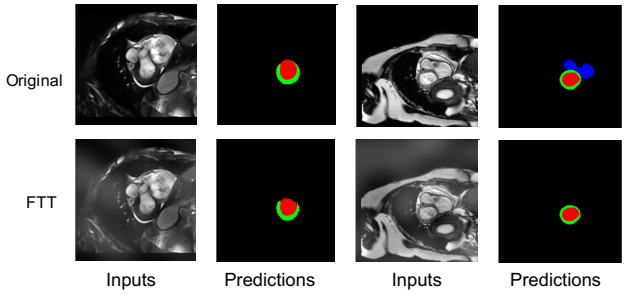

Figure 3: These images show that the predictions of the original image and transformed image may have a large variance. Hence, we present the confidence-aware cross pseudo supervision mechanism to reduce the weight of training loss in such cases.

As shown in Figure 2, after obtaining the predictions P_O^1 and P_F^1 from the original image and transformed image, we compute the average value of P_O^1 and P_F^1 as the ensemble prediction result P_E^1:

$$P_E^1 = (P_O^1 + P_F^1)/2, \tag{4}$$

and we obtain the ensemble prediction result P_E^2 of the second network by:

$$P_E^2 = (P_O^2 + P_F^2)/2. \tag{5}$$

To selectively use the pseudo labels as the supervisions and reduce the low-quality labels, we first compute the variance of the predictions of the original image and transformed image to measure the quality of the pseudo labels. The variance is computed as the KL-divergence of these two predictions:

$$\begin{aligned} V_1 &= E[\, P_F^1 \, log(\frac{P_F^1}{P_O^1})\,], \\ V_2 &= E[\, P_F^2 \, log(\frac{P_F^2}{P_O^2})\,], \end{aligned} \tag{6}$$

where E compute the expectation value. If the difference between these two predictions is large, the computed variance value is also large, which reflects these two prediction have a relative low quality, and vice versa. Figure 3 shows some visual examples, where the predictions between the original image and the transformed image have a large variance.

After obtaining the variances of the prediction of these two network, we formulate the confidence-aware loss function $L_{cacps} = L_a + L_b$ to optimize the networks by using the cross supervision signals:

$$\begin{aligned} L_a &= E[\, e^{-V_1} L_{ce}(P_E^2, Y_1) + V_1\,], \\ L_b &= E[\, e^{-V_2} L_{ce}(P_E^1, Y_2) + V_2\,], \end{aligned} \tag{7}$$

where Y_1 and Y_2 are the one-hot vectors generated from the probability maps P_E^1 and P_E^2, and L_{ce} denotes the cross-entropy loss.

For the supervised part, we use dice loss as loss function. We define the supervision loss as L_s.

$$L_s = E[L_{Dice}(P_O^1, G_1) + L_{Dice}(P_O^2, G_2)], \tag{8}$$

where L_{Dice} is the dice loss function and $G_1(G_2)$ is the ground truth.

So, the whole training objective is written as:

$$L = L_s + \beta * L_{cacps}, \tag{9}$$

where β is the CACPS weight. Its goal is to put the two losses into balance.

During the test procedure, we use the ensemble of two models' predictions as the final results.

Experimental Results

Implementation Details

We implemented the model on Pytorch1.8 and trained it by using two NVidia 3090 GPUs with 377GB RAM on the Ubuntu20.04 system. We implemented two segmentation networks by using DeepLabv3+ (Chen et al. 2017) with ResNet50 (He et al. 2016) backbone, and adopted the weights trained on ImageNet (Deng et al. 2009) for classification to initialize the parameters of the backbone network and other layers were initialized by random noise. For M&Ms dataset, we leveraged AdamW to optimize the network with the weight decay of 0.1, the learning rate of 0.0001, and the batch size of 32. We trained the whole architecture for 20 epochs and the images were cropped to 288×288. We set β in equation 9 as three to balance the supervision loss and our proposed CACPS loss. We set λ in equation 1 as 1. For SCGM dataset, we leveraged AdamW to optimize the network with the weight decay of 0.1, the learning rate of 0.0001, and the batch size of eight. We trained the whole architecture for 50 epochs and the images were cropped to 288×288. We set β in equation 9 as 1.5 to balance the supervision loss and our proposed CACPS loss. We set λ in equation 1 as 0.8. We also adopt the random rotation, random scaling, random crop, and random flip as the data augmentation strategies.

Table 1: Dice (%) results and the standard deviations on M&Ms dataset using 5% labeled data. For "SDNet+Aug", "Meta", and our method, the training data contains all unlabeled data and 5% of labeled data from the source domains. The other models are trained by using 5% labeled data only. Bold numbers denote the best performance.

Method \ Target	A	B	C	D	Average
nnUNet (Isensee et al. 2021)	65.30±17	79.73±10	78.06±11	81.25±8.3	76.09±6.3
SDNet+Aug (Liu et al. 2020)	71.21±13	77.31±10	81.40±8.0	79.95±7.8	77.47±3.9
LDDG (Li et al. 2020a)	66.22±9.1	69.49±8.3	73.40±9.8	75.66±8.5	71.29±3.6
SAML (Liu, Dou, and Heng 2020)	67.11±10	76.35±7.9	77.43±8.3	78.64±5.8	74.88±4.6
Meta (Liu et al. 2021b)	72.40±12	80.30±9.1	82.51±6.6	83.77±5.1	79.75±4.4
Ours	**83.3±5.83**	**85.04±6.49**	**87.14±4.74**	**87.38±4.49**	**85.72±1.67**

Table 2: Dice (%) results and the standard deviations on M&Ms dataset using 2% labeled data. For "SDNet+Aug", "Meta", and our method, the training data contains all the unlabeled data and 2% of labeled data from source domains. The other models are trained by using 2% labeled data only. Bold numbers denote the best performance.

Method \ Target	A	B	C	D	Average
nnUNet (Isensee et al. 2021)	52.87±19	64.63±17	72.97±14	73.27±11	65.94±8.3
SDNet+Aug (Liu et al. 2020)	54.48±18	67.81±14	76.46±12	74.35±11	68.28±8.6
LDDG (Li et al. 2020a)	59.47±12	56.16±14	68.21±11	68.56±10	63.16±5.4
SAML (Liu, Dou, and Heng 2020)	56.31±13	56.32±15	75.70±8.7	69.94±9.8	64.57±8.5
Meta (Liu et al. 2021b)	66.01±12	72.72±10	77.54±10	75.14±8.4	72.85±4.3
Ours	**82.35±6.24**	**82.84±7.59**	**86.31±5.47**	**86.58±4.78**	**84.52±1.94**

Table 3: Dice (%) results and the standard deviations on SCGM dataset using 20% labeled data. For "SDNet+Aug", "Meta", and our method, the training data contains all the unlabeled data and 20% of labeled data from source domains. The other models are trained by using 20% labeled data only. Bold numbers denote the best performance.

Method \ Target	1	2	3	4	Average
nnUNet (Isensee et al. 2021)	59.07±21	69.94±12	60.25±7.2	70.13±4.3	64.85±5.2
SDNet+Aug (Liu et al. 2020)	83.07±16	80.01±5.2	58.57±10	85.27±2.2	76.73±11
LDDG (Li et al. 2020a)	77.71±9.1	44.08±12	48.04±5.5	83.42±2.7	63.31±17
SAML (Liu, Dou, and Heng 2020)	78.71±25	75.58±12	54.36±7.6	85.36±2.8	73.50±12
Meta (Liu et al. 2021b)	**87.45±6.3**	81.05±5.2	61.85±7.3	87.96±2.1	79.58±11
Ours	87.13±1.4	**87.31±2.02**	**78.75±9.44**	**91.73±1.28**	**86.23±4.69**

Datasets and Evaluation Metrics

We adopt the multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset (Campello et al. 2021) to evaluate of our method. It contains 320 subjects, which are scanned at six clinical centres in three different countries by using four different magnetic resonance scanner vendors, *i.e.*, Siemens, Philips, GE, and Canon, and we consider the subjects scanned from different vendors are from different domains (domains A,B,C,D). For each subject, only the end-systole and end-diastole phases are annotated. The resolutions of the voxels range from 0.85×0.85×10 mm to 1.45×1.45×9.9 mm. In total, there are 95 subjects in domain A, 125 subjects in domain B, 50 subjects in domain C, and another 50 subjects in domain D.

We also adopt the spinal cord gray matter segmentation (SCGM) dataset(Prados et al. 2017) to evaluate of our method. This dataset contains single channel Spinal Cord MRI data with gray matter labels from four different centers. Data is collected from four centers (UCL, Montreal, Zurich, Vanderbilt) using three different MRI systems (Philips Acheiva, Siemens Trio, Siemens Skyra) with institution specific acquisition parameters. So it has four domains (domains A,B,C,D) in total. The voxel resolutions range from 0.25 × 0.25 × 2.5 mm to 0.5 × 0.5 × 5 mm. Each domain has 10 labeled subjects and 10 unlabelled subjects.

We evaluate the segmentation performance by using Dice (%) score, which is defined as:

$$Dice\,(P, G) = \frac{2 \times |P \cap G|}{|P| + |G|}, \tag{10}$$

where P and G are the predicted segmentation result and ground truth image, respectively. $|P \cap G|$ denotes the overlapped region between P and G, while $|P| + |G|$ represents the union region. In general, a better segmentation result has a larger $Dice$.

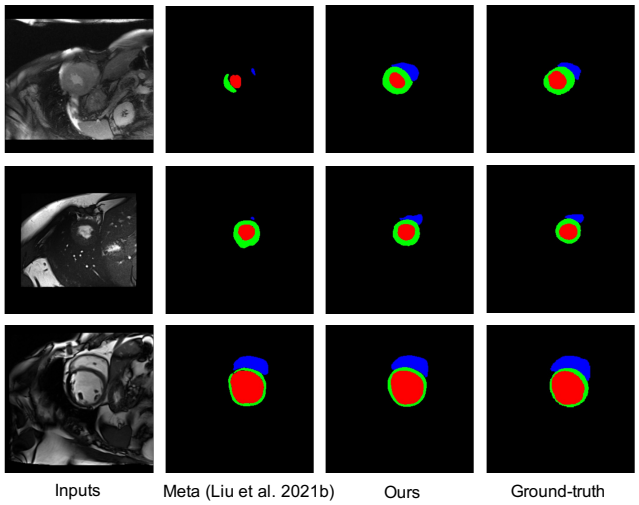

Figure 4: Visual comparison results on the M&Ms dataset.

Experiments on the M&Ms Dataset

Comparison with the State-of-the-art Methods. We compare our method with the following state-of-the-art methods:

(i) **nnUNet (Isensee et al. 2021)**: is a self-adapting framework based on 2D and 3D U-Nets(Ronneberger, Fischer, and Brox 2015) for fast and effective segmentation. But it is not designed for domain generalization.

(ii) **SDNet+Aug (Liu et al. 2020)**: first generates diverse data by rescaling the images to different resolutions within a range spanning different scanner protocols, and then generates more diverse data by projecting the original samples onto disentangled latent spaces and combining the learned anatomy and modality factors from different domains.

(iii) **LDDG (Li et al. 2020a)**: presents a method to learn a representative feature space through variational encoding with a novel linear-dependency regularization term to capture the shareable information among medical data, which is collected from different domains. It is the latest state-of-the-art model for domain-generalized medical image analysis in a fully supervised setting.

(iv) **SAML (Liu, Dou, and Heng 2020)**: is a gradient-based meta-learning approach, which constrains the compactness and smoothness properties of segmentation masks across meta-train and meta-test sets in a fully supervised manner.

(v) **Meta (Liu et al. 2021b)**: is a semi-supervised meta-learning framework, which models the domain shifts by using the disentanglement and extracts robust anatomical features for predicting segmentation masks in a semi-supervised manner by applying multiple constraints with the gradient-based meta-learning approach. It is the latest state-of-the-art model for domain-generalized medical image analysis in a semi-supervised setting.

Results on the M&Ms dataset Table 1 and Table 2 report the comparison results on the M&Ms dataset, where our method achieves the best performance on all different settings. Specifically, our method shows the improvements $\approx 6\%$ and $\approx 12\%$ in terms of Dice scores on the settings of 5% labeled data and 2% labeled data, respectively, by comparing with the previous best method. Note that by comparing with other methods, our method gives a significant improvement on the setting that we adopt the image samples from domains B, C, D as the training data, and evaluate the trained model on the domain A. This is because the image samples in domain A are less similar to other domain, as also shown in Figure 1, and our method has a great generalization capability.

Figure 4 shows more visual results by comparing with the previous best method, *i.e.*, Meta (Liu et al. 2021b). From the results, we can see that our method generates the results that are more consistent with the ground truth images, while Meta (Liu et al. 2021b) tend to mis-segment some unobvious regions of the input images.

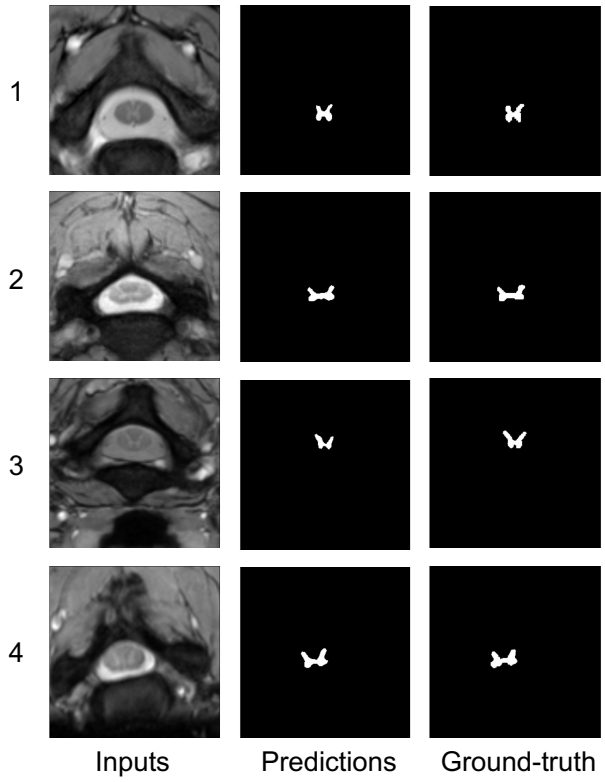

Figure 5: Visual comparison results on the SCGM dataset. We take the images from four different domains as the example visualization.

Results on the SCGM dataset Table 3 reports the comparison results on the SCGM dataset, where our method shows the improvement $\approx 7\%$ in terms of Dice score on the setting of 20% labeled data, by comparing with the previous best method. Figure 5 shows the visual results produced by our method in different domains, where we can see that our method successfully generates the prediction results that are consistent with the ground truth images on all the four domains.

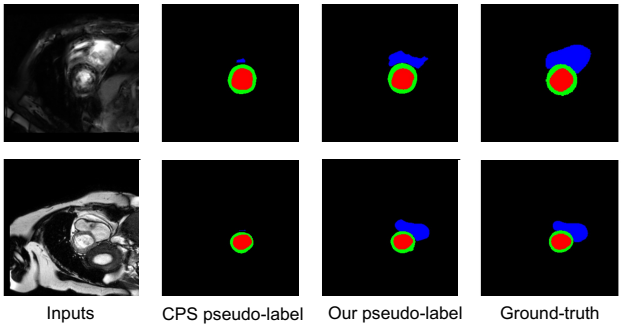

Figure 6: We compare the pseudo labels generated by CPS (Chen et al. 2021) and by our method, where our method can generate the pseudo labels that are more consistent with the ground-truth images.

Table 4: Ablation Study on the M&Ms dataset by using 2% labeled data. We consider domain A as the target domain and others as the source domain.

CPS	Fourier	CACPS	Dice(%)
✓			80.46±4.67
✓	✓		81.47±4.66
✓	✓	✓	82.61±4.15

Ablation Study We perform an ablation study to evaluate the effectiveness of each component of our proposed method on the M&Ms dataset. First, we build our first baseline by using CPS (Chen et al. 2021) designed for semi-supervised learning to perform segmentation task on different domains. Then, we construct the second model by further using the Fourier transformation to perform data augmentation and further use CPS to train the network. Third, we consider the confidence-aware cross pseudo supervision (CACPS) instead of simple CPS to optimize the network, which is the full pipeline of our proposed method. Moreover, we visually compare the pseudo labels generated by CPS and our method in Figure 6, where we can see that our pseudo labels are more consistent with the ground truth images. Table 4 reports the results, where we can see that each component gives a clear improvement on the results.

Conclusion

This paper presents a semi-supervised domain generalization method for medical images by formulating two novel techniques. One is the Fourier-transform-based data augmentation and another is the confidence-aware cross pseudo supervision. The Fourier-transform-based data augmentation helps to obtain the low-level statistic information from different domains, which augments the image samples by incorporating the cross-domain information. The confidence-aware cross pseudo supervision algorithm measures the variance between the original image sample and the image sample augmented by Fourier transformation, which helps to improves the quality of pseudo labels, thus facilitating the learning process of the whole framework. Finally, we evaluate our method on two public benchmark datasets, compare our method with various methods, and show its superiority over the other state-of-the-art methods. In the future, we will jointly leverage the semi-supervised and weakly-supervised algorithms to improve the performance of medical image segmentation by exploring the knowledge of unlabeled data and the data with weaker labels, such as bounding boxes. Moreover, we will explore our framework for more medical image segmentation tasks and integrate the network into the artificial intelligent systems on medical diagnostics.

Acknowledgments

This work was supported by a research grant from Shenzhen Municipal Central Government Guides Local Science and Technology Development Special Funded Projects (2021Szvup139) and a research grant from HKUST Bridge Gap Fund (BGF.027.2021).

References

Bian, C.; Yuan, C.; Wang, J.; Li, M.; Yang, X.; Yu, S.; Ma, K.; Yuan, J.; and Zheng, Y. 2020. Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64: 101732.

Campello, V. M.; Gkontra, P.; Izquierdo, C.; Martín-Isla, C.; Sojoudi, A.; Full, P. M.; Maier-Hein, K.; Zhang, Y.; He, Z.; Ma, J.; et al. 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, 9458279.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dolz, J.; Gopinath, K.; Yuan, J.; Lombaert, H.; Desrosiers, C.; and Ayed, I. B. 2018. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. *IEEE transactions on medical imaging*, 38(5): 1116–1126.

Feng, Z.; Zhou, Q.; Cheng, G.; Tan, X.; Shi, J.; and Ma, L. 2020. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 1(2): 5.

French, G.; Aila, T.; Laine, S.; Mackiewicz, M.; and Finlayson, G. 2019. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations.

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

Ibrahim, M. S.; Vahdat, A.; Ranjbar, M.; and Macready, W. G. 2020. Semi-supervised semantic image segmentation with self-correcting networks. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12715–12725.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Kim, J.; Jang, J.; and Park, H. 2020. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint arXiv:2001.04647*.

Lai, X.; Tian, Z.; Jiang, L.; Liu, S.; Zhao, H.; Wang, L.; and Jia, J. 2021. Semi-Supervised Semantic Segmentation With Directional Context-Aware Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1205–1214.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Lee, J.; Kim, E.; and Yoon, S. 2021. Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4071–4080.

Li, H.; Wang, Y.; Wan, R.; Wang, S.; Li, T.-Q.; and Kot, A. C. 2020a. Domain generalization for medical imaging classification with linear-dependency regularization. *arXiv preprint arXiv:2009.12829*.

Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; and Heng, P.-A. 2018a. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE transactions on medical imaging*, 37(12): 2663–2674.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; and Heng, P.-A. 2018b. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020b. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 523–534.

Li, Y.; Chen, J.; Xie, X.; Ma, K.; and Zheng, Y. 2020c. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 614–623. Springer.

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021a. 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*, 1013–1023.

Liu, Q.; Dou, Q.; and Heng, P.-A. 2020. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 475–485. Springer.

Liu, X.; Thermos, S.; Chartsias, A.; O'Neil, A.; and Tsaftaris, S. A. 2020. Disentangled Representations for Domain-generalized Cardiac Segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 187–195. Springer.

Liu, X.; Thermos, S.; O'Neil, A.; and Tsaftaris, S. A. 2021b. Semi-supervised Meta-learning with Disentanglement for Domain-generalised Medical Image Segmentation. *arXiv preprint arXiv:2106.13292*.

Pomponio, R.; Erus, G.; Habes, M.; Doshi, J.; Srinivasan, D.; Mamourian, E.; Bashyam, V.; Nasrallah, I. M.; Satterthwaite, T. D.; Fan, Y.; et al. 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208: 116450.

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

Puyol-Anton, E.; Ruijsink, B.; Piechnik, S. K.; Neubauer, S.; Petersen, S. E.; Razavi, R.; and King, A. P. 2021. Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. *arXiv preprint arXiv:2106.12387*.

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

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 23–30. IEEE.

Wang, J.; Zhou, S.; Fang, C.; Wang, L.; and Wang, J. 2020. Meta Corrupted Pixels Mining for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 335–345. Springer.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A Fourier-based Framework for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Yeo, T.; Kar, O. F.; and Zamir, A. 2021. Robustness via Cross-Domain Ensembles. *arXiv preprint arXiv:2103.10919*.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 605–613. Springer.

Zhou, F.; Jiang, Z.; Shui, C.; Wang, B.; and Chaib-draa, B. 2020. Domain generalization with optimal transport and metric learning. *arXiv preprint arXiv:2007.10573*.

Zou, Y.; Zhang, Z.; Zhang, H.; Li, C.-L.; Bian, X.; Huang, J.-B.; and Pfister, T. 2020. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*.