CauSSL: Causality-inspired Semi-supervised Learning for Medical Image Segmentation

Juzheng Miao[1], Cheng Chen[2], Furui Liu[3]*, Hao Wei[4], Pheng-Ann Heng[1,5]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Center for Advanced Medical Computing and Analysis,
Harvard Medical School and Massachusetts General Hospital
[3]Zhejiang Lab, [4]Department of Biomedical Engineering, The Chinese University of Hong Kong
[5]Institute of Medical Intelligence and XR, The Chinese University of Hong Kong

Abstract

Semi-supervised learning (SSL) has recently demonstrated great success in medical image segmentation, significantly enhancing data efficiency with limited annotations. However, despite its empirical benefits, there are still concerns in the literature about the theoretical foundation and explanation of semi-supervised segmentation. To explore this problem, this study first proposes a novel causal diagram to provide a theoretical foundation for the mainstream semi-supervised segmentation methods. Our causal diagram takes two additional intermediate variables into account, which are neglected in previous work. Drawing from this proposed causal diagram, we then introduce a causality-inspired SSL approach on top of co-training frameworks called CauSSL, to improve SSL for medical image segmentation. Specifically, we first point out the importance of algorithmic independence between two networks or branches in SSL, which is often overlooked in the literature. We then propose a novel statistical quantification of the uncomputable algorithmic independence and further enhance the independence via a min-max optimization process. Our method can be flexibly incorporated into different existing SSL methods to improve their performance. Our method has been evaluated on three challenging medical image segmentation tasks using both 2D and 3D network architectures and has shown consistent improvements over state-of-the-art methods. Our code is publicly available at: https://github.com/JuzhengMiao/CauSSL.

1. Introduction

Data-driven deep learning methods have shown remarkable performance in medical image segmentation [18, 38]. However, they typically require a large number of high-quality labeled data, which is extremely costly and difficult to obtain for the pixel-wise annotations of medical images requiring domain expertise. To solve this problem, semi-supervised learning (SSL) has become more and more popular and achieved remarkable success in medical image segmentation with limited annotations and large amounts of unlabeled data [1, 14, 25, 55]. Current semi-supervised segmentation methods can be mainly divided into two categories. The first one is self-training or pseudo-labeling methods [1, 14], which utilize pseudo-labels as supervision for unlabeled images. The other mainstream SSL methods are based on consistency regularization. These methods apply consistency regularization on the predictions between different models or branches based on the popular Mean Teacher (MT) [25, 55] or co-training [36, 52] frameworks.

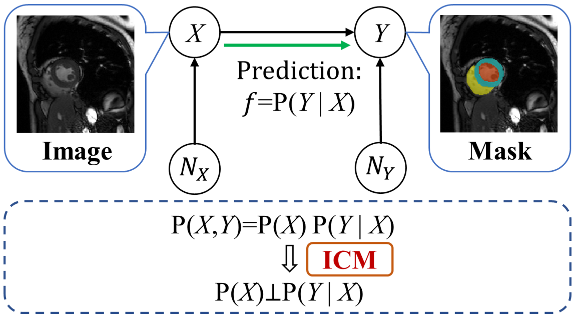

Figure 1. Previous causal diagram for medical image segmentation. The cause is the input image (X), based on which experts give the corresponding effect, i.e. segmentation mask (Y). The green arrow represents the prediction direction for the segmentation network. $\perp$ indicates algorithmic independence defined on two functions or distributions. According to ICM, unlabeled data can help improve the estimation of $P(X)$ but fail to improve the network $P(Y|X)$ since they are algorithmically independent.

Although many works have demonstrated the success

*Corresponding author.

of SSL methods on medical image segmentation tasks, Kügelgen *et al.* [23] argue that semi-supervised segmentation should be expected ineffective based on the principle of independent causal mechanisms (ICM) [13, 35, 42]. ICM claims that the causal generative process consists of independent mechanisms that do not share information with each other. Fig. 1 depicts the causal diagram of the annotation generation process for medical image segmentation following [9]. In this process, experts manually delineate the corresponding segmentation mask for a given medical image based on visual inspection of the image content and intensity contrast. As a result, the image X is the cause, while the annotation Y is the effect [9]. According to ICM, for a causal prediction task, i.e., $X \rightarrow Y$, the image generation process $P(X)$ should be algorithmically independent with respect to the annotation generation process $P(Y|X)$. In this regard, adding unlabeled data can provide more knowledge about the data generation mechanism $P(X)$, but brings no helpful information about $P(Y|X)$, parameterized by the segmentation network since there is no link between them. This challenges the theoretical explanation of the success in SSL segmentation.

To address this issue, we propose a new causal diagram (see Fig. 2) by introducing two intermediate nodes which can provide a better explanation for the mainstream semi-supervised segmentation methods. These intermediate nodes denote pseudo-labels or predictions of another network/branch to assist the network learning on unlabeled data which are common in current SSL methods but are neglected in Fig. 1. The detailed analysis can be found in Section 3. Based on the new diagram, we further demonstrate that the algorithmic independence in a co-training framework can be even beneficial to the segmentation performance. However, the formalized measurement of the algorithmic independence, i.e., Kolmogorov complexity, is not computable, and proxies are often used in specific applications, such as the minimum description length for NLP tasks [19]. Nonetheless, a proxy for segmentation networks has not yet been explored. In this work, we propose a novel statistical quantification of the algorithmic independence specialized for deep convolutional networks, based on which, we design a min-max optimization process to further enhance the independence in co-training frameworks.

In summary, our main contributions are four-folds:

- This study proposes a novel causal diagram which is in compatibility with mainstream SSL methods in segmentation. The diagram sheds light on the effectiveness of semi-supervised segmentation from a causal perspective, and thus provides a theoretical foundation for understanding and further improving the performance of SSL in medical image segmentation.

- Based on our proposed causal diagram and ICM, we give comprehensive explanations of the limitations of the vanilla self-training and MT-based methods compared to co-training frameworks. This deepens the researchers' understanding of the SSL framework from the viewpoint of algorithmic information and provides justifications for the significance of considering algorithmic independence for model learning.

- We propose a novel statistical quantification of the uncomputable algorithmic independence, specialized for deep convolutional networks, named as network independence. This defines the algorithmic independence on the preservation of matrix ranks, treating the convolution kernels as matrices. A min-max optimization framework is then utilized to enable end-to-end metric learning and validated on both co-training and MT-based learning scenarios.

- We evaluate our method with extensive experiments on three public medical image segmentation tasks by using both 2D and 3D network architectures. The superior performance of our method provides empirical evidence for the claim that semi-supervised medical image segmentation can be improved by causal-diagram-induced algorithmic independence.

2. Related Work

SSL for Medical Image Segmentation. In recent years, SSL has made significant progress in leveraging unlabeled data to improve the segmentation performance under limited annotations. Previous methods can be broadly categorized into self-training methods [1, 14], and consistency-regularization methods [25, 55]. Bai *et al.* [1] developed a representative self-training framework for cardiac MR image segmentation. It includes the network predictions for unlabeled data as pseudo-labels and updates the training network iteratively. Under the framework of consistency regularization, Li *et al.* [25] proposed to enhance the consistency between predictions of inputs under different data augmentations on top of the MT framework [46]. On the other hand, Xia *et al.* [52] demonstrated the effectiveness of the co-training strategy by training a segmentation network on each view of volume data and encouraging the multi-view consistency among these networks.

Causality in Medical Image Analysis. Improving the models' performance on medical image analysis from the view of causality has received significant attention recently. Causality-inspired learning models have been applied to discover causal links of various neural processes [41], provide explanations for network performances [6, 20], and improve fairness [43]. Another interesting direction is to generate images of the potential appearance if a patient was healthy using counterfactual techniques [15]. Moreover, a lot of works focus on improving the robustness and generalization abilities of their networks using causal reasoning, such as domain adaptation and Out-of-Distribution detection [48]. For example, Kouw *et al.* [22] introduced a causal Bayesian prior to enhance the cross-center segmentation performance on MRI data. Ouyang *et al.* [32] proposed a causality-inspired data augmentation approach and leveraged causal intervention to improve the model robustness on the single-source domain generalization problem.

Causality in Semi-supervised Learning. Most causality-related works on SSL focus on how the causal direction can affect the learning performance. Schölkopf *et al.* [42] first pointed out that SSL works better when predicting the cause variables from its effects (anticausal learning) or from confounded inputs (confounded learning) and should be impossible when predicting the target labels from the causes (causal learning). Based on this conclusion, Kügelgen *et al.* [23] further proposed a new framework for semi-supervised classification by conducting the prediction using both cause and effect features simultaneously, creating an anti-causal learning setting. However, the pessimistic conclusion cannot explain the promising achievements of SSL in segmentation tasks which is a classic causal learning setting [9]. In our work, we aim to investigate the causal diagram of SSL segmentation and demonstrate that the principle of independent causal mechanisms is not always detrimental to causal learning settings like segmentation tasks. Instead, we show proper statistical quantification and further enhancement of the algorithmic independence property is helpful for improving the segmentation performance.

3. Causal Modeling of SSL Segmentation

In this section, we propose a novel causal diagram that is compatible with the current SSL frameworks. Based on the theoretical foundation, plausible explanations for the effectiveness of SSL segmentation methods are presented, and some key factors of SSL segmentation performance are naturally highlighted with the lens of causality.

3.1. Causal Diagram for SSL Segmentation

In Fig. 1, only the label annotation process in the SSL segmentation tasks is considered. However, actual implementations of SSL methods tend to introduce some intermediate variables such as pseudo-labels or predictions of another network/branch to assist the network learning on unlabeled data. The incomplete causal diagram in previous works [23, 35, 42] results in the pessimistic conclusion on SSL segmentation. Therefore, we introduce intermediate variables into our proposed causal diagram to better describe the general learning process in most SSL methods. As shown in Fig. 2, the input image X, the mask of the target organ Y, and the original predictions of segmentation networks $\hat{P}$ and $\hat{P}^{'}$ are observable variables. Following Castro *et al.* [9], we consider X as the cause of Y. $\hat{P}$ and $\hat{P}^{'}$ are the approximations of the target mask generated by the segmentation networks f and $f^{'}$, respectively, and thus determined by both the image and the mask.

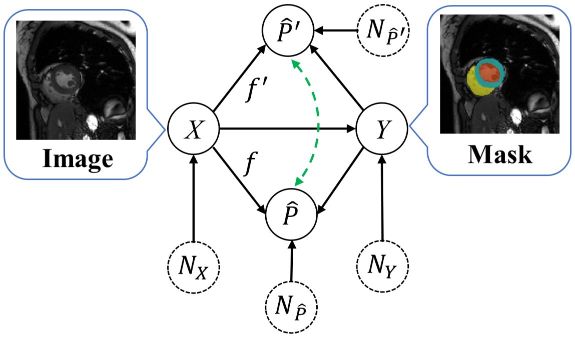

Figure 2. The causal diagram for semi-supervised medical image segmentation, where the variables with solid line boundaries are observable variables while those with dotted line boundaries are unobservable variables. The green arrow presents the loss dependence on unlabeled data. $\hat{P}$ and $\hat{P}^{'}$ mean the network predictions.

3.2. Consistency with Mainstream Methods

In the SSL setting, the training dataset $\boldsymbol{D}$ consists of $M_{\mathcal{L}}$ labeled data and $M_{\mathcal{U}}$ unlabeled data, denoted as $\mathcal{L} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{M_{\mathcal{L}}}$ and $\mathcal{U} = \{(\boldsymbol{x}_i)\}_{i=M_{\mathcal{L}}+1}^{M_{\mathcal{L}}+M_{\mathcal{U}}}$, where $\boldsymbol{x}_i \in \mathcal{R}^{H_{in} \times W_{in}}$ denotes an image and $\boldsymbol{y}_i \in \{0,1\}^{H_{in} \times W_{in} \times C}$ represents the corresponding ground-truth label for labeled data, with C meaning the number of semantic classes. For the labeled data, Y can be directly observed in Fig. 2 and a supervised loss L_s is utilized to help $\hat{P}$ and $\hat{P}^{'}$ approximate the mask directly. On the other hand, since the ground-truth labels are not available for unlabeled data, $\hat{P}^{'}$ is usually adopted as a proxy of Y and used to guide the learning of $\hat{P}$, as indicated by the green arrow in Fig. 2.

In self-training methods, $\hat{P}$ is the prediction generated by the segmentation network, while $\hat{P}^{'}$ is the output for the same unlabeled image by the same network predicted in previous iterations and functions as a pseudo label to supervise $\hat{P}$ with supervised loss L_s. By contrast, $\hat{P}^{'}$ can be predicted by another network or branch and guide the learning of $\hat{P}$ via a consistency regulation loss in the MT/co-training framework. For instance, methods based on the MT framework usually adopt the teacher model as $f^{'}$ to generate $\hat{P}^{'}$ and enforce the consistency between the teacher and student using a mean squared error (MSE) loss. In the co-training framework, the CPS method [10] uses another independent segmentation network with the same architecture but different weight initializations to generate $\hat{P}^{'}$. $\hat{P}$ and $\hat{P}^{'}$ then function as the pseudo-labels for each other via a cross-entropy loss.

3.3. Effectiveness Explanation and Key Components for SSL Segmentation

We conjecture the remarkable progress of the SSL segmentation methods can be largely attributed to the green arrow in Fig. 2. More unlabeled data help us obtain more information about $P(X)$. This helps provide more information about $\hat{P}^{'}$ through $f^{'}$, which is a noisy estimation of $P(Y)$. Naturally, if the approximation is precise enough, the SSL problem becomes a supervised one, where a good performance can be ensured. Therefore, with more pairs of input images and segmentation masks, the network predictions $\hat{P}$ can be improved if the quality of the approximation is good enough. Also, from the causal perspective, the loss between $\hat{P}$ and $\hat{P}^{'}$ introduces a learning direction from $\hat{P}$ to $\hat{P}^{'}$, which is a confounded learning setting and should help information sharing among different mechanisms and improve the segmentation performance [23, 42].

With the help of the proposed causal diagram, it's also easier for us to identify the key components that have significant influences on the medical image segmentation performance in an SSL framework. As presented in Fig. 2, the learning of $\hat{P}$ is directly affected by ground-truth labels and $\hat{P}^{'}$ for labeled and unlabeled data, respectively. Therefore, the quality of $\hat{P}^{'}$, as well as the learning constraint between $\hat{P}$ and $\hat{P}^{'}$ are of great importance. The former has been noticed in [44, 47, 54]. In addition, how to design an appropriate loss to ensure a thorough consistency between $\hat{P}$ and $\hat{P}^{'}$ such as shape-aware constraints rather than using the pixel-wise MSE can be a promising direction. Moreover, uncertainty estimation can be integrated into the constraint loss to reweight the contribution from different regions of unlabeled data and avoid the harmful guidance from $\hat{P}^{'}$ when the quality of $\hat{P}^{'}$ is not good enough [31, 45, 49, 52, 55].

In the following section, we will further demonstrate that the algorithmic independence between f and $f^{'}$ also significantly affects the segmentation performance.

4. Method

4.1. Structural Causal Model for SSL Segmentation

The structural causal model (SCM) framework [33] is adopted to describe our proposed causal diagram for causal analysis. The observed variables $X_i \in O = \{X, \hat{P}, \hat{P}^{'}, Y\}$ are determined by their parents PA_i and noise variables N_i, using a deterministic function f_i: $X_i := f_i(PA_i, N_i)$. Also, an independent assumption is often made over the unobserved noise variables following [23], indicating there are no hidden confounders. Then, the joint distribution over the observed variables can be factorized as:

$$\begin{aligned} P(X, \hat{P}, \hat{P}^{'}, Y) &= \prod_{X_i \in O} P\left(X_i \mid PA_i\right) \\ &= P(X)P(\hat{P}|X,Y)P(\hat{P}^{'}|X,Y)P(Y|X) \end{aligned} \tag{1}$$

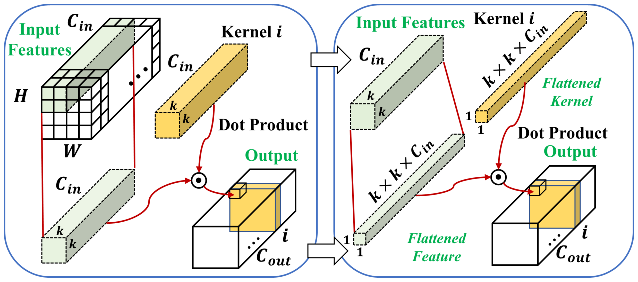

Figure 3. An illustration of the 2D convolution process.

According to ICM [13, 35, 42], mechanisms $P(\hat{P}|X,Y)$ and $P(\hat{P}^{'}|X,Y)$ should be algorithmically independent and do not inform or influence each other. Since they are highly related to the segmentation models, the two models should be algorithmically independent to some extent as well. This also aligns with our intuition that two different networks should provide complementary help for each other.

The importance and the effectiveness of the algorithmic independence between the two networks align with the practical observations in SSL segmentation tasks. For self-training and MT-based methods, the high dependence between the assistant network $f^{'}$ and the training network f leads to the performance bottleneck [10, 17, 21]. In an extreme case when $f^{'}$ is the same as f, the $\hat{P}^{'}$ in our proposed causal diagram (Fig. 2) disappears. This degenerates into the causal learning setting as shown in Fig. 1, making SSL segmentation models not so useful since unlabeled data $P(X)$ don't contain helpful information for network f. This explains the performance bottleneck of vanilla self-training and MT-based frameworks from a causal view. By contrast, the co-training framework breaks such limits and obtains a better performance by using two independent networks with different initial parameters [10], introducing different decoders with different upsampling strategies on top of the same encoder [50, 51], constructing different network architectures [27], and even adopting adversarial samples [34]. Despite the efficacy, these methods are intuitive improvements on the algorithmic independence and fail to propose metrics to directly measure the independence.

4.2. Network Independence

Our work focuses on finding a reasonable and computable proxy for the algorithmic independence metric (Kolmogorov complexity) in the scenery of convolutional networks. For simplicity, we only consider the case of different networks/branches with the same convolutional architecture. As mentioned in Section 1, Kolmogorov complexity $K(x)$ describes the compression length of x. Therefore, we design our proxy metric based on the Minimum Description Length (MDL) principle [16, 37] from the aspect of compression similar to [8, 19].

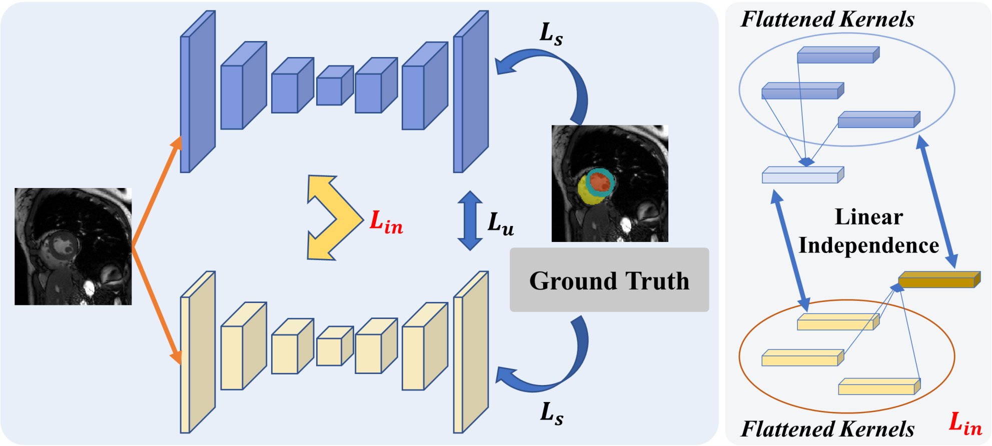

Figure 4. The pipeline of our CauSSL framework. Two networks are optimized by minimizing the combination of the supervised loss L_s on labeled data, the unsupervised loss L_u on unlabeled data, and the network independence loss L_{in} between network parameters.

Fig. 3 illustrates how a 2D convolutional kernel is applied to the input features and generates a filtered result. For a $k \times k$ convolutional kernel, the actual number of the kernel weights is $k \times k \times C_{in}$, where C_{in} means the channel number of input features. Then, such an extended 3D kernel calculates the dot product with part of features, and slides over the whole feature map, generating the result of an output channel. With C_{out} different kernels similar to this, we can generate a total of C_{out} output channels. From the view of signal processing, each kernel can be seen as a template pattern to find specific patterns. Features with patterns similar to the kernel will generate a high activation value by the dot product. Thus, the flattened kernel can be considered as a base vector in linear algebra to detect the similarity of input features on this pattern. Then, a convolutional layer with $C_{out} \times (k \times k \times C_{in})$ parameters can be viewed as a $C_{out} \times d$ matrix ($d = k \times k \times C_{in}$), and each row in this matrix is a d-dimensional base vector for a certain pattern.

With the view of vectors and matrices, the compressibility of a neural network is thus transferred to the compressibility of matrices and the latter can be naturally related to the matrix rank. According to the MDL principle, the algorithmic independence between two matrices indicates that the length of describing matrix A and B together equals the sum of the separate description, i.e. $rank([A, B]) = rank(A) + rank(B)$. $[A, B]$ is the extended $(C_{out,A} + C_{out,B}) \times d$ matrix by concatenating a $C_{out,A} \times d$ matrix A and a $C_{out,B} \times d$ matrix B along the row dimension. This holds if any row vector in A and B cannot be represented by the linear combination of the row vectors in the other matrix. Based on this linear independence condition, we propose a novel statistical quantification proxy for Kolmogorov complexity to measure the network dependence (the opposite of independence).

We first define the dependence between the same convolutional layers from two networks, denoting the matrix version of kernel weights as A and B with a size of $C_{out} \times d$:

$$L_{in}(A, B; G_B) = \frac{1}{C_{out}} \sum_{i=1}^{C_{out}} \left(\frac{\boldsymbol{v}_{A,i} \cdot \boldsymbol{q}_{B,i}}{|\boldsymbol{v}_{A,i}| \times |\boldsymbol{q}_{B,i}|} \right)^2 \qquad (2)$$

$$\boldsymbol{q}_{B,i} = (G_B \times B)_i$$

where $\boldsymbol{v}_{A,i}$ is the i-th row vector in matrix A, and $\boldsymbol{q}_{B,i}$ is the optimal linear combination vector using the vector group of B that can approximate $\boldsymbol{v}_{A,i}$ as close as possible. G_B is the optimal coefficient matrix whose elements are the optimal linear combination coefficients, with a size of $C_{out} \times C_{out}$. In this way, the lower the L_{in} is, the higher the network independence it indicates.

We then define the network dependence between two neural networks or branches with the same architecture by taking the average over all the convolutional layers:

$$L_{in}(\theta_1, \theta_2; G_2) = \frac{1}{\#\text{ layers}} \sum_{i=1}^{\#\text{layers}} L_{in}(\theta_{1,i}, \theta_{2,i}; G_{2,i}) \qquad (3)$$

where $\theta_{1,i}$, $\theta_{2,i}$, and $G_{2,i}$ are the weight parameters in the format of matrices and the optimal coefficient matrix of the i-th convolutional layer, respectively. Only convolutional layers are considered in this work.

4.3. Causality-inspired SSL

Based on the network dependence, we propose a causality-inspired SSL framework to further enhance the algorithmic independence on top of the co-training framework called CauSSL (Fig. 4), via a min-max optimization framework to improve the SSL segmentation performance:

$$\min_{\theta_1, \theta_2} \max_{G_1, G_2} L_{in}(\theta_1, \bar{\theta_2}; G_2) + L_{in}(\theta_2, \bar{\theta_1}; G_1) \qquad (4)$$

where G_1 and G_2 are the set of linear coefficient matrices which are only used in training and will be discarded during

Algorithm 1 Pseudocode of CauSSL

```
Input: labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$ and hyperparameters $\lambda_1$ and $\lambda_2$.
Output: Two independent segmentation networks or branches parameterized by $\theta_1$ and $\theta_2$, respectively.
Randomly initialize the network weights $\theta_1$, $\theta_2$ and linear coefficients $G_1$, $G_2$. // initialization
$i = 0$ // iteration number
while $i \leq$maximum iterations do // training
    for $j$=1:$s_{max}$ do // maximize
        Fix $\theta_1$, $\theta_2$. Update $G_1$, $G_2$ by maximizing
        $L_{in}(\theta_2, \bar{\theta_1}; G_1)$ and $L_{in}(\theta_1, \bar{\theta_2}; G_2)$.
    for $j$=1:$s_{min}$ do // minimize
        Fix $G_1$, $G_2$. Update $\theta_1$, $\theta_2$ by minimizing
        $L_{total,1}$ and $L_{total,2}$ using Equation 5.
        $i = i + 1$.
Return $\theta_1$ and $\theta_2$.
```

inference. $\bar{\theta_1}$ and $\bar{\theta_2}$ represent weights copy without gradient flows. In this standard bilevel optimization problem, we want to find the best linear combination coefficients that can maximize L_{in} to provide an accurate estimation of the network dependence on one hand, and enhance the algorithmic independence between two networks by minimizing L_{in} on the other hand. To this end, we update the linear combination coefficients and network weights in an alternative way as shown in Algorithm 1. In each round, we first fix the parameters of networks and update the linear combination coefficients to maximize L_{in} for s_{max} steps. Then, we fix the linear coefficients and update the network by minimizing L_{total} for s_{min} steps, which is a combination of the supervised loss L_s on the labeled data, the unsupervised loss L_u on the unlabeled data, and the network independence loss L_{in} between different networks or branches:

$$\begin{aligned} L_{total,1} &= L_{s,1} + \lambda_1 L_{u,1} + \lambda_2 L_{in}(\theta_1, \bar{\theta_2}; G_2) \\ L_{total,2} &= L_{s,2} + \lambda_1 L_{u,2} + \lambda_2 L_{in}(\theta_2, \bar{\theta_1}; G_1) \end{aligned} \tag{5}$$

$$L_{s,i}\left(\hat{\boldsymbol{y}}_{\mathcal{L},i}, \boldsymbol{y}\right) = \frac{1}{2}\left[L_{dice}\left(\hat{\boldsymbol{y}}_{\mathcal{L},i}, \boldsymbol{y}\right) + L_{ce}\left(\hat{\boldsymbol{y}}_{\mathcal{L},i}, \boldsymbol{y}\right)\right] \tag{6}$$

where $i \in [1, 2]$. λ_1 and λ_2 are balancing coefficients. $\lambda_1(t) = 0.1 * e^{-5\left(1-\frac{t}{t_{max}}\right)^2}$ is adopted following [55] considering the quality of predictions from the assistant network might not be good enough in the initial training stage. L_{dice} and L_{ce} indicate the Dice loss and the cross-entropy loss, respectively. $\hat{\boldsymbol{y}}_{\mathcal{L},i}$ represents the predicted probability maps of the i-th network for labeled data. The unsupervised loss L_u can be either cross-entropy loss between the network prediction and pseudo-labels generated by another network/branch or MSE loss between two probability maps:

$$L_{u,1} = L_{ce}(\hat{\boldsymbol{y}}_{\mathcal{U},1}, \tilde{\boldsymbol{y}}_{\mathcal{U},2}) \text{ or } L_{u,1} = \text{MSE}\left(\hat{\boldsymbol{y}}_{\mathcal{U},1}, \hat{\boldsymbol{y}}_{\mathcal{U},2}\right) \tag{7}$$

where $\hat{\boldsymbol{y}}_{\mathcal{U},i}$ is the predicted probability maps on the unlabeled data generated by the two networks, and $\tilde{\boldsymbol{y}}_{\mathcal{U},i}$ means the corresponding one-hot pseudo-labels. Details of how to extend our method to a framework with 3 branches can be found in Appendix Section 1.

5. Experiments

5.1. Datasets and Evaluation Metrics

Our proposed method is validated on three public datasets with different imaging modalities and segmentation tasks, i.e., the Automatic Cardiac Diagnosis Challenge dataset (ACDC) [7], Pancreas-CT dataset [12, 39, 40] and Multimodal Brain Tumor Segmentation Challenge 2019 (BraTS'19) dataset [3, 4, 5, 29]. The details of these datasets and the preprocessing steps are described in Appendix Section 2. Four metrics were used for evaluation, including the Dice similarity coefficient (DSC), Jaccard (JC), 95% Hausdor Distance (95HD), and the average surface distance (ASD). We have highlighted the results in bold when our proposed CauSSL outperforms the original counterparts and underlined the best results. Also, standard deviations are reported in parentheses.

5.2. Implementation Details and Baselines

We applied our proposed CauSSL on top of two popular co-training methods, CPS [10] and MC-Net+ [50], and denote the modified methods as CPSCauSSL and MC-CauSSL, respectively. Both of them are trained using Algorithm 1 but the CPS method utilizes cross-entropy loss as the unsupervised loss whereas MC-Net+ uses the MSE loss (Equation 7). We applied these two independence-enhanced methods for various network architectures to demonstrate the efficacy of our method, including 2D U-Net [38], 3D V-Net [30], and 3D U-Net [11]. Specific settings on each dataset are described in Appendix Section 2.

In all experiments of our proposed method, we empirically updated the network weights and linear coefficient matrices alternatively, with 60 steps for each. In addition, the linear coefficient matrices were optimized using an Adam optimizer with a fixed learning rate of 0.02. Moreover, our method was compared with fully supervised learning only (SL), MT [46], uncertainty-aware Mean Teacher (UA-MT) [55], SASSNet [24], DTC[26], URPC [28], CPS [10], and MC-Net+ [50], which were re-implemented in the identical environment and used the same training configurations for a fair comparison. Also, we compared our method with BCP (CVPR'23) [2] on the ACDC dataset and FUSS-Net (MICCAI'22) [53] on the Pancreas-CT dataset, which are the state-of-the-art (SOTA) methods in their respective datasets. If the same training setting (dataset and data split) is used, we directly reported the results from their original paper. Otherwise, we re-ran their publicly available code

Table 1. Comparisons with other methods on the ACDC dataset with 10% and 20% labeled data.

Labeled%	Method	DSC (%) ↑	JC (%) ↑	95HD (voxel) ↓	ASD (voxel) ↓
100%	SL (upper bound)	$91.53_{(2.89)}$	$84.76_{(4.62)}$	$2.41_{(5.28)}$	$0.59_{(1.08)}$
10%	SL	$77.66_{(13.10)}$	$66.40_{(14.53)}$	$11.68_{(12.30)}$	$3.31_{(3.65)}$
	MT	$81.11_{(9.65)}$	$69.99_{(11.72)}$	$8.99_{(10.25)}$	$2.70_{(3.14)}$
	UA-MT	$80.71_{(9.69)}$	$69.58_{(12.02)}$	$13.69_{(16.54)}$	$4.50_{(6.00)}$
	SASSNet	$82.56_{(8.94)}$	$71.90_{(11.42)}$	$9.13_{(9.99)}$	$2.64_{(2.71)}$
	DTC	$84.32_{(6.92)}$	$74.04_{(9.29)}$	$9.47_{(11.61)}$	$2.63_{(3.00)}$
	URPC	$82.41_{(10.15)}$	$71.69_{(12.91)}$	$5.83_{(9.09)}$	$1.65_{(2.87)}$
	CPS	$84.24_{(6.85)}$	$73.91_{(9.37)}$	$8.26_{(9.68)}$	$2.45_{(2.90)}$
	CPSCauSSL	$\mathbf{85.25}_{(6.43)}$	$\mathbf{75.31}_{(8.98)}$	$\mathbf{6.05}_{(8.87)}$	$\mathbf{1.97}_{(2.64)}$
	MC-Net+	$86.14_{(6.13)}$	$76.61_{(8.32)}$	$6.04_{(9.02)}$	$1.85_{(2.50)}$
	MCCauSSL	$\mathbf{86.80}_{(5.34)}$	$\mathbf{77.48}_{(7.62)}$	$\mathbf{5.73}_{(9.26)}$	$\mathbf{1.83}_{(2.56)}$
	BCP	$88.84_{(/)}$	$80.62_{(/)}$	$3.98_{(/)}$	$1.17_{(/)}$
	BCPCauSSL	$\underline{\mathbf{89.66}}_{(3.82)}$	$\underline{\mathbf{81.79}}_{(5.93)}$	$\underline{\mathbf{3.67}}_{(8.16)}$	$\underline{\mathbf{0.93}}_{(1.27)}$
20%	SL	$84.62_{(8.74)}$	$74.85_{(11.13)}$	$6.32_{(9.20)}$	$1.79_{(2.58)}$
	MT	$85.46_{(7.28)}$	$75.89_{(9.97)}$	$8.02_{(10.55)}$	$2.39_{(3.20)}$
	UA-MT	$85.16_{(7.41)}$	$75.49_{(9.93)}$	$5.91_{(8.95)}$	$1.79_{(2.71)}$
	SASSnet	$86.45_{(6.77)}$	$77.20_{(9.53)}$	$6.63_{(8.52)}$	$1.98_{(2.40)}$
	DTC	$87.10_{(6.18)}$	$78.15_{(8.76)}$	$6.76_{(10.83)}$	$1.99_{(3.10)}$
	URPC	$85.44_{(9.29)}$	$76.36_{(11.27)}$	$5.93_{(9.04)}$	$1.70_{(2.87)}$
	CPS	$86.85_{(7.05)}$	$77.96_{(9.41)}$	$5.48_{(9.13)}$	$1.64_{(2.70)}$
	CPSCauSSL	$\mathbf{87.24}_{(6.18)}$	$\mathbf{78.44}_{(8.55)}$	$5.57_{(9.13)}$	$1.73_{(2.65)}$
	MC-Net+	$87.10_{(6.45)}$	$78.21_{(9.03)}$	$5.04_{(8.49)}$	$1.56_{(2.43)}$
	MCCauSSL	$\mathbf{87.84}_{(6.31)}$	$\mathbf{79.32}_{(8.84)}$	$\mathbf{4.37}_{(8.04)}$	$\mathbf{1.28}_{(2.30)}$
	BCP	$89.52_{(4.20)}$	$81.62_{(6.44)}$	$3.69_{(7.02)}$	$1.03_{(1.89)}$
	BCPCauSSL	$\underline{\mathbf{89.99}}_{(3.65)}$	$\underline{\mathbf{82.34}}_{(5.77)}$	$\underline{\mathbf{3.60}}_{(8.62)}$	$\underline{\mathbf{0.88}}_{(1.73)}$

Table 2. Comparisons with other methods on the Pancreas-CT dataset with 6 and 12 volumes having annotations.

#Labeled	Method	DSC (%) ↑	JC (%) ↑	95HD (voxel) ↓	ASD (voxel) ↓
62/62	SL (upper bound)	$82.80_{(6.32)}$	$71.10_{(8.87)}$	$5.58_{(4.18)}$	$1.26_{(0.98)}$
6/62	SL	$56.59_{(21.31)}$	$42.24_{(19.99)}$	$23.79_{(16.35)}$	$7.18_{(5.45)}$
	MT	$68.61_{(13.86)}$	$53.71_{(15.07)}$	$18.64_{(16.45)}$	$5.28_{(4.12)}$
	UA-MT	$66.96_{(14.43)}$	$51.89_{(15.32)}$	$21.65_{(14.12)}$	$6.25_{(3.25)}$
	SASSNet	$66.69_{(14.86)}$	$51.66_{(15.51)}$	$18.88_{(11.55)}$	$5.76_{(2.70)}$
	DTC	$67.28_{(17.37)}$	$52.86_{(17.63)}$	$17.74_{(18.58)}$	$\underline{1.97}_{(0.89)}$
	URPC	$64.73_{(15.36)}$	$49.62_{(16.57)}$	$21.90_{(9.83)}$	$7.73_{(3.02)}$
	FUSSNet	$72.55_{(10.66)}$	$57.95_{(13.03)}$	$18.45_{(19.22)}$	$5.23_{(5.96)}$
	CPS	$66.97_{(13.94)}$	$51.93_{(15.17)}$	$14.73_{(8.90)}$	$4.49_{(2.25)}$
	CPSCauSSL	$\mathbf{67.33}_{(13.59)}$	$\mathbf{52.28}_{(14.65)}$	$16.16_{(8.38)}$	$5.21_{2.27)}$
	MC-Net+	$68.18_{(12.49)}$	$52.94_{(13.67)}$	$16.35_{(11.05)}$	$4.13_{(2.80)}$
	MCCauSSL	$\underline{\mathbf{72.89}}_{(8.90)}$	$\underline{\mathbf{58.06}}_{(10.84)}$	$\underline{\mathbf{14.19}}_{(11.20)}$	$4.37_{(2.88)}$
12/62	SL	$72.72_{(11.32)}$	$58.25_{(13.33)}$	$19.23_{(14.83)}$	$5.77_{(3.99)}$
	MT	$76.39_{(9.80)}$	$62.73_{(12.44)}$	$9.91_{(9.54)}$	$2.56_{(2.57)}$
	UA-MT	$77.42_{(8.68)}$	$63.91_{(11.16)}$	$\underline{7.96}_{(5.44)}$	$1.87_{(1.00)}$
	SASSNet	$78.06_{(7.40)}$	$64.59_{(9.89)}$	$12.76_{(15.78)}$	$3.15_{(3.51)}$
	DTC	$76.82_{(12.53)}$	$63.70_{(13.95)}$	$8.69_{(10.38)}$	$\underline{1.28}_{(0.42)}$
	URPC	$79.09_{(7.39)}$	$65.99_{(9.86)}$	$11.68_{(13.80)}$	$3.31_{(2.62)}$
	FUSSNet	$80.37_{(5.93)}$	$67.57_{(8.16)}$	$13.75_{(20.92)}$	$3.46_{(3.87)}$
	CPS	$78.16_{(7.33)}$	$64.74_{(9.83)}$	$9.54_{(9.11)}$	$2.63_{(2.14)}$
	CPSCauSSL	$\mathbf{78.58}_{(7.52)}$	$\mathbf{65.32}_{(9.97)}$	$\mathbf{8.30}_{(6.22)}$	$\mathbf{2.34}_{(1.57)}$
	MC-Net+	$79.36_{(6.54)}$	$66.23_{(8.87)}$	$10.22_{(9.59)}$	$2.66_{(2.21)}$
	MCCauSSL	$\underline{\mathbf{80.92}}_{(5.20)}$	$\underline{\mathbf{68.26}}_{(7.30)}$	$\mathbf{8.11}_{(9.24)}$	$\mathbf{1.53}_{(1.30)}$

using the default hyperparameters.

5.3. Comparison on Organ Segmentation Tasks

Table 1 and Table 2 tabulate the quantitative comparison results on the ACDC and Pancreas-CT datasets, respectively, from which several observations can be found. First, methods with a co-training framework (CPS and MC-Net+) generally outperform the vanilla MT framework (MT and UA-MT). For example, CPS surpasses MT and UA-MT with a margin of 3.13% and 3.53% DSC, respectively on the ACDC dataset. This is consistent with findings in [10, 17, 21] and demonstrates the importance of algorithmic independence in the SSL framework. Nonetheless, it is worth noting that incorporating other intricate modules into the MT-based method can also yield outstanding results, as demonstrated by approaches like BCP [2].

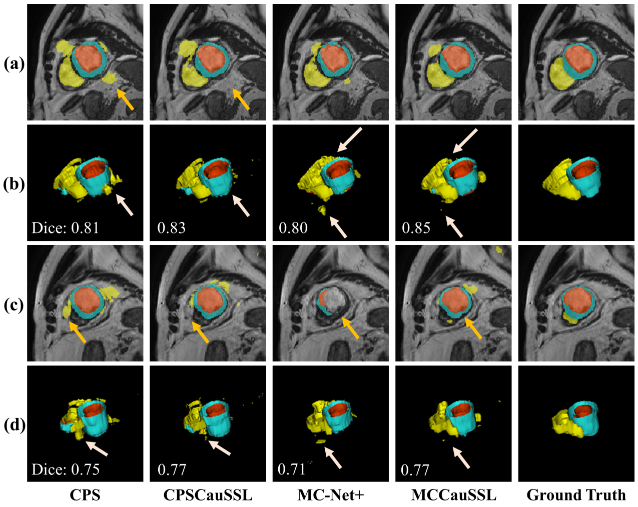

Figure 5. Visualization of segmentation results on the ACDC testing dataset trained with 10% labeled images.

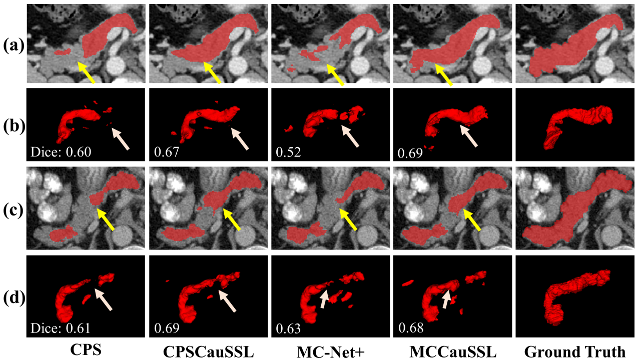

Figure 6. Visualization of segmentation results on the Pancreas-CT testing dataset trained with 6 labeled volumes.

Second, by introducing a causality-inspired independence constraint into the co-training framework, we can further achieve a performance improvement (highlighted in bold) and outperform other SOTA methods. Almost all the metrics are improved under different ratios of labeled data on both datasets and our proposed CauSSL obtains the best results across various settings. On the ACDC dataset, the performance gain using our proposed CauSSL is 1.01% DSC and 1.40% JC for the CPS method with 10% labeled data. MCCauSSL also outperforms the original version by about 0.7% DSC and 0.9% JC. When 20% annotations are used, the performance improvement for the CPS method is narrowed (0.39% DSC), whereas the gain is still stable (0.74% DSC and 1.11% JC) for MC-Net+. On the Pancreas-CT dataset, the gap between MC-Net+ and MCCauSSL is even larger, with a margin of 4.71% and 1.56% DSC using 6 and 12 annotated volumes, respectively. Trained by only about 20% labeled data, MCCauSSL even closely approaches the upper bound (80.92 vs 82.80 DSC).

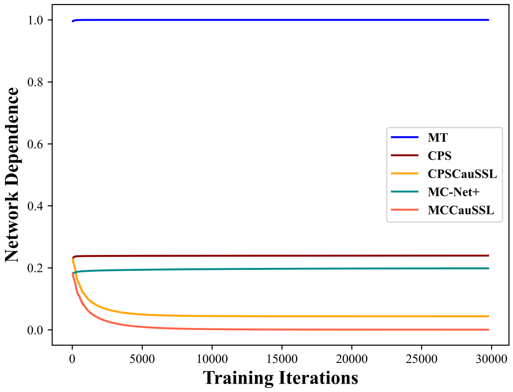

Figure 7. Network dependence of different methods during training on the ACDC dataset using 10% labeled training data.

The effectiveness of our proposed method can also be shown in some hard examples (See Fig. 5 and Fig. 6). Two cases are presented here on each dataset. The first and third rows are the predictions for a certain slice while the second and fourth rows show 3D visualizations for each case. In these examples, CPS and MC-Net+ tend to generate false predictions (Fig. 5, row (a),(b)) or incomplete structures (Fig. 6), whereas the introduction of independence constraint can mitigate these problems and obtain a more plausible segmentation result.

More comparisons with other SOTA methods further demonstrate the efficacy of our proposed method. For example, on the Pancreas-CT dataset, our proposed MCCauSSL is superior to the FUSSNet [53] under various ratios of labeled data, with an improvement of 0.55% DSC trained with 12 labeled volumes (See Table 2). By contrast, although the MCCauSSL obtains a lower DSC compared to BCP [2] on the ACDC dataset as shown in Table 1, applying our proposed CauSSL on top of the BCP method can further improve the performance by 0.82% DSC when 10% labeled data are used. It also approaches the upper bound with all the labels (91.53%) only with a margin of 1.87%. In spite of the performance improvement for BCP by introducing another 10% annotations, the DSC result (89.52%) is even lower than our BCPCauSSL with half labels (89.66%).

5.4. Comparison on a Tumor Segmentation Task

To validate the generalizability of our method, we also tried our method with the challenging brain tumor segmentation on the BraTS'19 dataset using a 3D U-Net backbone. Table 3 shows results with 10% labeled data for different methods. Similar to the results on organ segmentation tasks, semi-supervised learning methods are superior to the baseline using labeled data alone, with a DSC improvement of over 10% for both CPS and MC-Net+. In the meanwhile, our proposed CauSSL can further improve the performance of the SSL methods, as shown in bold in Table 3. For example, our CauSSL scheme obtains 1% and 1.7% DSC improvements over CPS and MC-Net+, respectively, demonstrating the efficacy of our proposed method on another more challenging setting and network architecture.

Table 3. Comparisons with other methods on the BraTS'19 dataset with 10% annotations.

Method	DSC (%) ↑	JC (%) ↑	95HD (voxel) ↓	ASD (voxel) ↓
SL	$72.84_{(17.53)}$	$60.05_{(20.52)}$	$41.64_{(36.39)}$	$2.56_{(1.45)}$
CPS	$82.56_{(12.90)}$	$72.18_{(17.11)}$	$14.41_{(16.36)}$	$2.31_{(2.01)}$
CPSCauSSL	$\mathbf{83.56}_{(12.84)}$	$\mathbf{73.60}_{(16.82)}$	$\mathbf{11.91}_{(14.17)}$	$\mathbf{2.06}_{(1.70)}$
MC-Net+	$81.84_{(15.21)}$	$71.65_{(18.99)}$	$13.82_{(17.04)}$	$2.43_{(2.17)}$
MCCauSSL	$\mathbf{83.54}_{(12.41)}$	$\mathbf{73.46}_{(16.61)}$	$\mathbf{12.53}_{(15.94)}$	$\mathbf{1.98}_{(1.53)}$

Table 4. DSC results of applying our CauSSL to the vanilla MT.

Dataset	#Labeled	MT	MTCauSSL
ACDC	7/70	81.11%	82.89%
	14/70	85.46%	86.35%
Pancreas-CT	6/62	68.61%	71.36%
	12/62	76.39%	77.63%

5.5. Application to MT-based Methods

Although our proposed CauSSL is originally designed on top of co-training methods to further improve the network independence, it can also be applied to MT-based methods (named as MTCauSSL) thanks to its plug-and-play nature by adding the network independence loss to the student training. In other words, we only keep one item in Equation 4 for MTCauSSL. As shown in Table 4, our MTCauSSL can obtain a stable and significant improvement over the vanilla MT method on both the ACDC and the Pancreas-CT datasets. Especially, the DSC improvement is 2.75% when only 6 labeled data are used on the Pancreas-CT dataset. Moreover, by applying our proposed method to the SOTA method BCP, we can further improve its performance from DSC 88.84% to 89.66% on the ACDC dataset with 10% labeled data as shown in Table 1.

5.6. Analysis of Network Dependence

To demonstrate the efficacy of our proposed min-max framework in enhancing the algorithmic independence, we further measure the dependence using the metric defined in Equation 3. We take the average of all the networks or branches in a method as the final measurement. Fig. 7 illustrates the network dependence of different methods on the ACDC dataset during the training process.

First of all, the network dependence of the MT method is extremely high, approaching 1 at last. This is due to the use of the exponential moving average strategy. Using the limit theory, Ke *et al.* [21] have proved that the weights of the teacher network and the student model will converge to the same target given infinite iteration steps. Such a dependence explains the performance bottleneck of MT-based methods. By contrast, CPS and MC-Net+ have a much

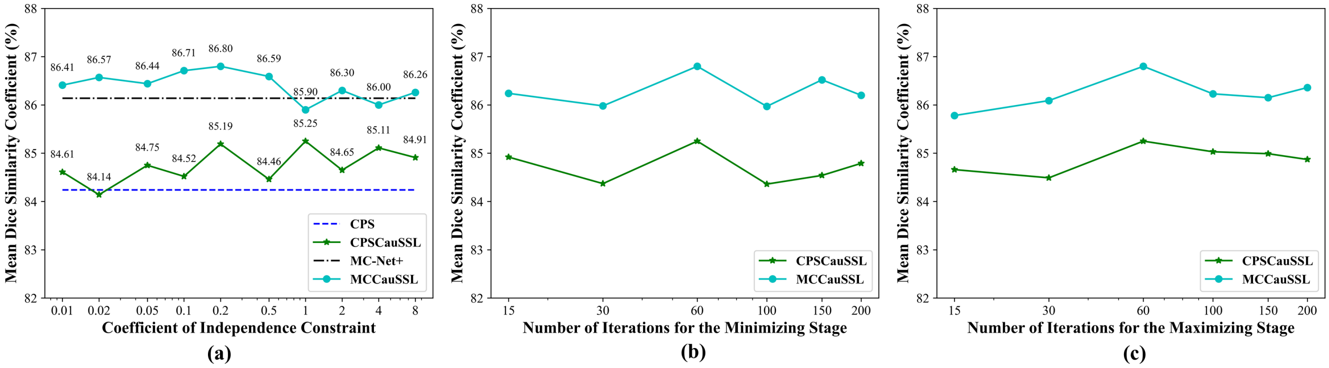

Figure 8. DSC performances on the ACDC dataset with different hyperparameters under the setting of 10% labeled data.

lower dependence and also achieve a superior result on this dataset, proving the necessity of independence constraint.

Moreover, the proposed min-max framework can further reduce the algorithmic dependence and improve the segmentation performance. The original dependence score for CPS and MC-Net+ are 0.24 and 0.20, respectively. After integrating the independence constraint into the training process, the dependence of CPS is significantly reduced to 0.04, whereas the metric for MC-Net+ becomes 4e-4. The improvement of the algorithmic dependence is consistent with the performance gain of the two co-training methods.

5.7. Impacts of Hyperparameters

We first compare the mean DSC of different independence constraint coefficients using 10% labeled data on the ACDC dataset. As shown in Fig. 8 (a), no matter which coefficient is taken, both CPSCauSSL and MCCauSSL surpass their counterparts without the independence constraint in most cases, indicated by the horizontal dotted lines, demonstrating the efficacy of our proposed CauSSL framework. Second, the number of minimizing or maximizing iterations is observed to have a similar effect on both methods. 60 steps for minimizing obtains the highest DSC, whereas too many iterations might introduce over enhancement of the network independence and weak enforcement of the independence might not fulfill the potential of the co-training framework if the number of minimizing iterations is not big enough (See Fig. 8 (b)). Moreover, according to Fig. 8 (c), when the number of maximizing steps is too small (such as 15 and 30), the linear coefficients might fail to match well, leading to an underestimated network dependence and possibly wrong independence optimization. On the other hand, although more maximizing steps don't bring higher segmentation performance, it is better than insufficient dependence measurement.

5.8. Analysis of Training Efficiency

As shown in Equation 2 and 3, compared to the original SSL methods, the additional computation introduced by our method mainly includes matrix multiplication, normalization, MSE, and average calculation over all the convolutional layers. All of these can be implemented in Pytorch with high efficiency and just slightly increase the overall training time. For example, on the ACDC dataset with 10% labeled data using an NVIDIA RTX 3090 GPU, the training duration of our CPSCauSSL was 4.71 hours, which is comparable to CPS taking 4.40 hours. Likewise, MCCauSSL requires just an additional 0.34 hours compared to MC-Net+ (4.61 vs. 4.27 hours).

6. Conclusion

This paper proposes a novel causal diagram to provide plausible explanations for the effectiveness of SSL medical image segmentation. Based on the diagram, the importance of the algorithmic independence is noticed and a novel statistical quantification is designed for convolutional networks to approximate the uncomputable algorithmic independence. Then, we propose a causality-inspired SSL framework to further enhance the algorithmic independence and thus improve the SSL segmentation performance. Comparisons on three datasets and three network architectures demonstrate the effectiveness of our proposed method.

7. Acknowledgement

This work described in this paper was supported in part by the Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089. The work was also partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Number: T45-401/22-N) and by the following grants from the Hong Kong Innovation and Technology Fund (Project number: MHP/085/21 and GHP/080/20SZ).

References

[1] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew

King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017.

[2] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11514–11524, 2023.

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[4] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[5] Spyridon (Spyros) Bakas. Brats miccai brain tumor dataset, 2020.

[6] Mehri Baniasadi, Mikkel V Petersen, Jorge Gonçalves, Andreas Horn, Vanja Vlasov, Frank Hertel, and Andreas Husch. Dbsegment: Fast and robust segmentation of deep brain structures considering domain generalization. *Human Brain Mapping*, 44(2):762–778, 2023.

[7] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[8] Kailash Budhathoki and Jilles Vreeken. Causal inference by compression. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 41–50. IEEE, 2016.

[9] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[10] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[12] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[13] Povilas Daniušis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 143–150, 2010.

[14] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020.

[15] Pedro M Gordaliza, Juan José Vaquero, and Arrate Muñoz-Barrutia. Translational lung imaging analysis through disentangled representations. *arXiv preprint arXiv:2203.01668*, 2022.

[16] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[17] Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li, and Qi Tian. Atso: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1235–1244, 2021.

[18] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[19] Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schölkopf. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513. Association for Computational Linguistics, 2021.

[20] Maxime Kayser, Roger D Soberanis-Mukul, Anna-Maria Zvereva, Peter Klare, Nassir Navab, and Shadi Albarqouni. Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting. *arXiv preprint arXiv:2002.02883*, 2020.

[21] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6728–6736, 2019.

[22] Wouter M Kouw, Silas N Ørting, Jens Petersen, Kim S Pedersen, and Marleen de Bruijne. A cross-center smoothness prior for variational bayesian brain tissue segmentation. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 360–371. Springer, 2019.

[23] Julius Kügelgen, Alexander Mey, Marco Loog, and Bernhard Schölkopf. Semi-supervised learning, causality, and the conditional cluster assumption. In *Conference on Uncertainty in Artificial Intelligence*, pages 1–10. PMLR, 2020.

[24] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020.

[25] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

[26] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021.

[27] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 820–833. PMLR, 2022.

[28] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021.

[29] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[31] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. Asdnet: attention based semi-supervised deep networks for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 370–378. Springer, 2018.

[32] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022.

[33] Judea Pearl. *Causality*. Cambridge university press, 2009.

[34] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107:107269, 2020.

[35] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[36] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018.

[37] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[39] Holger R Roth, Amal Farag, E Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. Data from pancreas-ct. the cancer imaging archive. *IEEE Transactions on Image Processing*, 2016.

[40] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015.

[41] Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, and Clark Glymour. Identification of effective connectivity subregions. *arXiv preprint arXiv:1908.03264*, 2019.

[42] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

[43] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

[44] Zhiqiang Shen, Peng Cao, Hua Yang, Xiaoli Liu, Jinzhu Yang, and Osmar R Zaiane. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. *arXiv preprint arXiv:2301.04465*, 2023.

[45] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2021.

[46] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[47] Bethany H Thompson, Gaetano Di Caterina, and Jeremy P Voisey. Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[48] Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. A review of causality for learning algorithms in medical image analysis. *arXiv preprint arXiv:2206.05498*, 2022.

[49] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 450–460. Springer, 2021.

[50] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Enforcing mutual con-

sistency of hard regions for semi-supervised medical image segmentation. *arXiv preprint arXiv:2109.09960*, 2021.

[51] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–306. Springer, 2021.

[52] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis*, 65:101766, 2020.

[53] Jinyi Xiang, Peng Qiu, and Yang Yang. Fussnet: Fusing two sources of uncertainty for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–491. Springer, 2022.

[54] Huifeng Yao, Xiaowei Hu, and Xiaomeng Li. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3099–3107, 2022.

[55] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.