Medical Image Analysis 83 (2023) 102656

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation

Shuo Zhang [a,b], Jiaojiao Zhang [a,b], Biao Tian [a,b], Thomas Lukasiewicz [c], Zhenghua Xu [a,b,*]

[a] State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China

[b] Tianjin Key Laboratory of Bioelectromagnetic Technology and Intelligent Health, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China

[c] Department of Computer Science, University of Oxford, United Kingdom

ARTICLE INFO

Dataset link: https://zhenghuaxu.info/publications/

Keywords:
Semi-supervised learning
Multi-modal contrastive learning
Pseudo-label re-learning
Medical image segmentation

ABSTRACT

Semi-supervised learning has a great potential in medical image segmentation tasks with a few labeled data, but most of them only consider single-modal data. The excellent characteristics of multi-modal data can improve the performance of semi-supervised segmentation for each image modality. However, a shortcoming for most existing multi-modal solutions is that as the corresponding processing models of the multi-modal data are highly coupled, multi-modal data are required not only in the training but also in the inference stages, which thus limits its usage in clinical practice. Consequently, we propose a semi-supervised contrastive mutual learning (Semi-CML) segmentation framework, where a novel area-similarity contrastive (ASC) loss leverages the cross-modal information and prediction consistency between different modalities to conduct contrastive mutual learning. Although Semi-CML can improve the segmentation performance of both modalities simultaneously, there is a performance gap between two modalities, i.e., there exists a modality whose segmentation performance is usually better than that of the other. Therefore, we further develop a soft pseudo-label re-learning (PReL) scheme to remedy this gap. We conducted experiments on two public multi-modal datasets. The results show that Semi-CML with PReL greatly outperforms the state-of-the-art semi-supervised segmentation methods and achieves a similar (and sometimes even better) performance as fully supervised segmentation methods with 100% labeled data, while reducing the cost of data annotation by 90%. We also conducted ablation studies to evaluate the effectiveness of the ASC loss and the PReL module.

1. Introduction

In recent years, deep learning techniques have achieved some great successes in medical image analysis, where supervised learning is the most commonly adopted solution, and a large amount of manually annotated data are usually needed (Esteva et al., 2017; Krizhevsky et al., 2012; Ronneberger et al., 2015). However, annotating medical images is a highly professional task that can only be done by radiologists with extensive clinical experience. Due to the limited number, time, and annotating efficiency of professional radiologists, obtaining a large medical image dataset with accurate annotations is usually very difficult, which thus limits the use of supervised learning in real-world clinical practice. A classical solution to this problem is semi-supervised learning, where a few labeled data and numerous unlabeled data are used together for the learning of deep models to eliminate the need of massive expensive labeled data while maintaining a satisfactory performance (Arazo et al., 2020; Lee et al., 2013; Mittal et al., 2019; Sohn et al., 2020; Xie et al., 2019). However, the existing semi-supervised methods are mostly based on single-modal data, and lack of capabilities to utilize fruitful information in multi-modal medical images.

Specifically, it is known that some medical imaging methods can generate medical images with multiple modalities, e.g., magnetic resonance imaging (MRI) can generate images with four modalities (i.e., T1, T1CE, T2, and FLAIR), and positron emission tomography (PET) can be done together with computerized tomography (CT) to obtain both PET and CT scans. At present, many different fully-supervised multi-modal fusion classification and segmentation networks have been proposed to make full use of multi-modal data (Liu et al., 2021; Andrearczyk et al., 2020; Kumar et al., 2019; Dolz et al., 2018; Hu et al., 2020; Mo et al., 2020; Tseng et al., 2017; Zhou et al., 2019a; Zhao et al., 2021; Tang

[*] Correspondence to: State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, 300130, China.
E-mail address: zhenghua.xu@hebut.edu.cn (Z. Xu).

https://doi.org/10.1016/j.media.2022.102656
Received 11 January 2022; Received in revised form 4 October 2022; Accepted 12 October 2022
Available online 17 October 2022
1361-8415/© 2022 Published by Elsevier B.V.

et al., 2022). However, these methods cannot use a large amount of unlabeled multi-modal data, which limits their performances. Therefore, some studies tend to use semi-supervised learning to utilize unlabeled multi-modal data, but these methods mainly focus on multi-modal classification and clustering tasks (Yang et al., 2018; Sun et al., 2020; Du et al., 2021). To our knowledge, there are only few works (Mondal et al., 2018; Chartsias et al., 2020) aiming at multi-modal semi-supervised medical image segmentation. However, these multi-modal segmentation methods have a common drawback regardless of whether they use unlabeled multi-modal data. The drawback is that the fusion process of two modalities usually requires weight sharing or complex feature concatenation (i.e., high coupling) in the training stage, which leads to the need for both modalities to be fed simultaneously in the inference stage. This greatly limits the application of these models in clinical practice, because taking medical images of multiple modalities is very time-consuming and money-costly and usually only a single modality of medical images is obtained for the imaging diagnosis in practice. So the need for a multi-modal semi-supervised model that uses only one modality in the inference stage for accurate medical image segmentation is compelling.

In this paper, we propose to take advantage of large amounts of unlabeled multi-modal data to improve the segmentation performance of each modality by allowing networks to in-depth learn from different modalities, where a novel contrastive loss is proposed to minimize the prediction differences of two image modalities, and thus get multiple single-modal inference networks. We call this framework Semi-supervised Contrastive Mutual Learning (*Semi-CML*). Specifically, we first perform the mean square error (MSE) consistency loss on the prediction maps of different modalities to achieve a simple mutual learning between different modalities. However, we find that the MSE loss is not good enough, because (i) it does not consider the area context information within the images, which is usually important for medical image segmentation tasks, and (ii) it only aims to minimize the differences between positive sample pairs but is not able to maximize the differences between the negative samples. Therefore, we propose to utilize a contrastive loss to resolve this problem. Although the state-of-the-art contrastive loss, Noise Contrastive Estimation (NCE) loss (Chaitanya et al., 2020; Chen et al., 2020; He et al., 2020), can additionally maximize the differences between the negative samples, the area context information is still not considered in NCE, which thus limits its performance. Therefore, we propose an improved contrastive loss, named Area-Similarity Contrastive (*ASC*) loss. Differently from conventional contrastive learning methods, ASC has the following two advantages: (i) ASC creatively incorporates Dice similarity into the contrastive learning, making it possible to take into account the area context information in learning, and (ii) ASC can bypass the projection head and directly perform area-based contrastive learning on the predicted segmentation map without ground-truth, thus directly and effectively improving the performance of semi-supervised semantic segmentation. In the process of mutual learning of multi-modal images, the ASC loss can maximize the lower bound of the mutual information between the predicted segmentation maps of different modalities, which leads to a higher prediction consistency. Consequently, the loss makes networks deeply learn the potential complementary information between different image modalities, which plays a vital role in the mutual learning between the two modalities.

We note that there exist performance gaps between different modalities. Although Semi-CML can alleviate this problem, there is always a modality whose performance is relatively low. Therefore, to further improve the segmentation ability of the low-performance image modality model, we develop a soft pseudo-label re-learning (*PReL*) scheme by using the prediction of the unlabeled high-performance image modality. Specifically, the model of high-performance modality is used as a teacher model with higher confidence through an improved exponential moving average self-ensembling technology, called Best-model Moving Average (*BMA*). We use the BMA update strategy to get a more reliable teacher model, which is called Best-model Moving Average self-ensembling teacher model (*BMA Teacher*). Then, under the re-learning epoch, we use the BMA teacher model with Monte-Carlo dropout sampling to generate a soft pseudo-label with higher reliability and precision for the training of the low-performance modality. Consequently, the semi-supervised contrastive mutual learning framework with a soft pseudo-label re-learning using BMA Teacher is called *Semi-CML with PReL*. The contributions of this work are briefly as follows:

- We identify a common shortcoming of multi-modal segmentation methods: they need images of all modalities in the inference stage, which thus limit their applications in clinical practices. To alleviate this problem, in this work, we propose a new low-coupling multi-modal semi-supervised segmentation framework, Semi-CML with PReL, which can adequately consider potential correlations and differences between unlabeled multi-modal data in the training stage, and uses only one modality in the inference stage for accurate medical image segmentation.
- We first propose a novel ASC loss to better utilize the area context information in contrastive learning, which is proved to achieve much better segmentation performances in semi-supervised medical image segmentation tasks. In order to remedy the performance gaps between different modalities, we further propose a soft pseudo-label re-learning (PReL) scheme which, based on best-model moving average (BMA) method, utilizes the model of high-performance modality as a teacher model to generate high-precision soft pseudo-labels for the re-learning of the model of low-performance modality.
- We conducted extensive experiments on two public multi-modal medical image segmentation datasets. The results show that (i) Semi-CML with PReL achieves a similar (and sometimes even better) performance as fully supervised segmentation solutions with 100% labeled data and greatly outperforms the state-of-the-art semi-supervised segmentation methods, and (ii) the proposed ASC loss and PReL scheme are both effective and essential for our model to achieve superior performances.

2. Related work

2.1. Semi-supervised segmentation

In medical image segmentation tasks, there usually exists only limited number of annotation data. To solve this problem, many researchers have focused on the semi-supervised medical image segmentation methods, which can be divided into self-training methods (Zhu et al., 2020; Zou et al., 2018; Bai et al., 2017), adversarial training methods (Hung et al., 2018; Souly et al., 2017; Li et al., 2020a), co-training methods (Zhou et al., 2019b; Peng et al., 2020a; Xia et al., 2020; Wang et al., 2021a), and consistency regularization methods (Bortsova et al., 2019; Hang et al., 2020; Li et al., 2020c; Luo et al., 2021; Wang et al., 2020; Yu et al., 2019). Zhu et al. (2020) propose to train a teacher model using labeled data and infer pseudo-labels for unlabeled data, then mix the pseudo-labeled and real-labeled data into a new student model to perform semi-supervised semantic segmentation. Li et al. (2020a) propose a shape-aware semi-supervised segmentation method, which uses generative adversarial learning to perform geometric shape constraints on the output maps.

Currently, co-training based and consistency regularization based methods are of great significance and have been widely used in semi-supervised medical image segmentation. Co-training based methods focus on co-training with different views of the 3D volume. For example, Zhou et al. (2019b) propose deep multi-planar co-training, where three planes of the 3D volume are trained separately, and more reliable pseudo-labels are obtained by fusing predictions from different planes. Xia et al. (2020) propose a multi-view co-training method, which utilizes multi-view information from unlabeled data to jointly train different views of the 3D volume to improve semi-supervised segmentation performances. Then, Wang et al. (2021a) propose to use self-paced and self-consistent learning strategies for the co-training between different networks in semi-supervised image segmentation tasks. Consistency regularization methods focus on the consistency of model predictions under different perturbations. For example, Luo et al. (2021) propose to add a dual-task consistency regularization at the output of the network for model training. Another method, the mean teacher model, has proved to be very effective in semi-supervised learning (Tarvainen and Valpola, 2017). Cui et al. (2019) propose to encourage consistent segmentation predictions on the student model and the teacher model for the same input under different disturbances. Yu et al. (2019) propose to add uncertainty estimation in the teacher model to enable the teacher model to generate certain predictions and then carry out consistency training with the student model. Co-training methods usually need to train multiple view models and utilize multi-view prediction consistency to obtain progressive performance of semi-supervised segmentation, however, joint training of multi-view models also increases the training and inference costs of the algorithm. Consistency regularization methods usually only need to design a reasonable consistency regularization scheme to obtain satisfactory semi-supervised segmentation performance. In our work, we mainly use the multi-modal prediction consistency to improve the model performance, so directly using the consistency regularization method will be more suitable, while the co-training method will introduce more models and cause the model complexity to be too high. Furthermore, different from the previous methods, our proposed method is based on the consistency prediction of multi-modal data by using the difference and agreement of different modalities to overcome the limitation that the above methods cannot directly use multi-modal information. Therefore, we choose the state-of-the-art co-training and consistency regularization methods as our baselines to show the superior performances of our multi-modal based semi-CML.

2.2. Multi-modal semi-supervised medical image analysis

At present, many works have proved that special complementary information and synergistic information among multi-modal data can obtain an additional performance boost for image segmentation. Consequently, co-segmentation methods are proposed by researchers to identify similar foreground regions from multi-modal images. A co-segmentation model based on Markov random fields (MRFs) is first proposed by Rother et al. (2006); then, many subsequent co-segmentation methods have been proposed (Daryanto et al., 2017), which can be roughly divided into unsupervised co-segmentation methods (Li et al., 2014; Dong et al., 2015; Wang et al., 2016; Meng et al., 2015) and interactive co-segmentation methods (Batra et al., 2010; Vicente et al., 2011; Batra et al., 2011; Tao et al., 2015). Furthermore, there also exists many deep learning based multi-modal fully-supervised segmentation methods (Zhou et al., 2019a); according to different multi-modal fusion strategies, multi-modal image segmentation networks can be divided into input-level fusion networks (Andrearczyk et al., 2020; Hu et al., 2020), layer-level networks (Kumar et al., 2019; Dolz et al., 2018; Tseng et al., 2017) and late-fusion networks (Andrearczyk et al., 2020; Mo et al., 2020; Zhao et al., 2021). However, these studies mainly focus on designing new information fusion strategies for the multi-modal data and their models are usually trained by full supervision.

Currently, for multi-modal semi-supervised works, related researches are mainly focusing on classification or clustering tasks (Yang et al., 2018; Sun et al., 2020; Du et al., 2021). For example, Sun et al. (2020) propose to use the total correlation gain maximization to predict Alzheimer's disease with multi-modal data. Differently from these works, we focus on semi-supervised segmentation tasks in medical images. To our knowledge, there are only few researches targeting at the semi-supervised segmentation tasks of multi-modal data. Mondal et al. (2018) have proposed a few-shot multi-modal segmentation method based on generative adversarial learning, which simply uses multi-modal data with a channel fusion method. Chartsias et al. (2020) have proposed a DAFNet, which uses incompletely labeled multi-modal data for image segmentation; this work uses disentanglement, alignment, and fusion to build a complex network to fuse multi-modal data to improve the performance of the target modal. However, these two existing semi-supervised multi-modal methods have the same disadvantage as other fully supervised multi-modal fusion models: the networks are highly coupled, which leads to the necessity of using multi-modal data as inputs simultaneously in the inference stage to obtain satisfactory final results. Differently, although our work also focuses on the task of semi-supervised multi-modal segmentation, we propose a method that uses multi-modal data in the training stage and generate multiple independent models that can be used to obtain satisfactory segmentation results in the inference stage using only single-modal data. To show the superior performances of our method, the few-shot multi-modal segmentation method and DAFNet are both used as the baselines in our experiments.

2.3. Contrastive learning

Recently, great progresses have been made in self-supervised learning based on contrastive losses (Chaitanya et al., 2020; Chen et al., 2020; He et al., 2020; Khosla et al., 2020; Wang et al., 2021b), and its performance has gone beyond supervised learning (Chen et al., 2020; He et al., 2020). A contrastive loss can increase the mutual information of similar samples by maximizing the similarity of positive samples and minimizing the similarity of negative samples. For example, Chen et al. (2020) propose to use data augmentation to get more meaningful positive and negative samples to achieve the competitive performance of downstream tasks. Recently, the contrastive loss have been successfully applied in medical imaging (Chaitanya et al., 2020; Iwasawa et al., 2020). Chaitanya et al. (2020) use the structural similarity of 3D medical data to design a global and local contrastive loss. In addition, contrastive learning is also used in the fully supervised semantic segmentation task, and a pixel-level contrastive algorithm is proposed in Wang et al. (2021b). In this paper, we take the contrastive loss as the instructor of mutual learning of different modalities in the semi-supervised setting and take the prediction results of two modalities as a positive sample pair to maximize the mutual learning ability between different image modalities.

2.4. Narrowing performance gap and EMA-weighted model

We propose the PReL algorithm to narrow the performance differences between different modalities based on the proposed Semi-CML. Similarly, domain adaptation (DA) is an existing method that can also be used to narrow the performance gaps. DA usually uses a generative adversarial network (GAN) (Radford et al., 2015; Arjovsky et al., 2017) to make the distributions of the target domain as close as possible to those of the source domain to narrow the performance gaps. In medical imaging, there are often multiple modalities of data due to variations in imaging protocols. To reduce the annotation cost of multi-modal data, an unsupervised domain adaptation method can be used to utilize cross-modal data for image segmentation (Kamnitsas et al., 2017; Dou et al., 2018; Zeng et al., 2021). For example, Dou et al. (2018) propose a cross-modality domain adaptation method, which designs a domain adaptation module and a domain critic module to transfer the MRI segmenter to CT data for narrowing their performance gaps. Comparing to DA solution, our PReL method has the following three advantages for the performance gap narrowing: First, PReL is more efficient. It can use reliable prediction maps for direct re-learning, while DA indirectly narrows the performance gap by narrowing the distribution differences. Second, PReL is more flexible. It can perform re-learning simultaneously on low- and high-performance modalities, whereas the

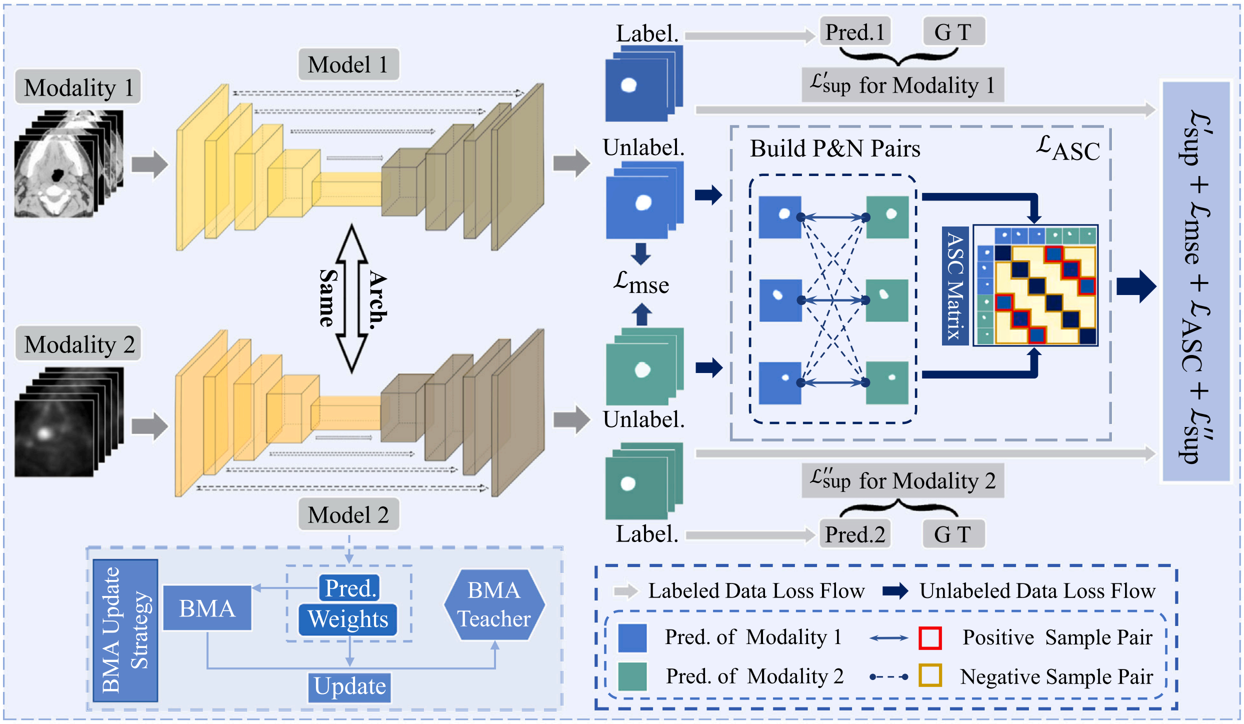


Fig. 1. Our semi-supervised contrastive mutual learning (Semi-CML) segmentation framework using multi-modal data. The framework consists of a dual-modal supervised loss, and cross-modal MSE and ASC unsupervised losses. For the ASC loss, we construct positive sample pairs from the paired cross-modalities data (solid arrow), and construct negative sample pairs from the unpaired cross-modalities data (dashed arrow) and the unpaired same-modalities data (not shown). According to the built positive and negative sample pairs, we use Dice similarity to calculate the ASC matrix. Finally, our optimization goal is, by minimizing the ASC loss, to make the positive sample pair (red area) in the matrix closer (darker color), and push the negative sample pair (brown area) farther (lighter color) in the matrix. The light blue dashed box in the lower-left corner illustrates the process of using a high-performance model to generate the BMA teacher model in the Semi-CML iteration. The details of the BMA update strategy are shown in Fig. 2.

target domain in DA usually does not get additional learning. Finally, PReL is more stable and simpler. It only needs to use the training-free teacher model to perform re-learning, while DA usually needs to train additional discriminator, which may lead to training instability.

The exponential moving average (EMA) can often be used as an update method for model weight ensemble, which makes the model weight smoother and more stable (Laine and Aila, 2017; Tarvainen and Valpola, 2017; Cui et al., 2019; Yu et al., 2019; Verma et al., 2019). For example, Tarvainen and Valpola (2017) propose to use EMA to aggregate the model weights obtained by multiple prior networks into a single ensemble model. Although EMA method can obtain a more stable teacher model, the weights of the teacher model are updated in each mini-batch, making the EMA-weighted model updated when the weight obtained in some training stage is still poor, and the network cannot dynamically adjust the update ratio according to the model training quality. Differently from the previous work, we propose to use the best model pool (BMP) to decide whether to update the teacher model, and we use a Best-model Moving Average (BMA) strategy to update the weights of the teacher model using the best model parameters for PReL.

3. Methodology

3.1. Semi-supervised contrastive mutual learning

Inspired by the potential value of intrinsic correlation in multi-modal data and contrastive self-supervised learning, and to overcome the high coupling problem in the multi-modal fusion model, we propose a novel low-coupling semi-supervised contrastive mutual learning framework, named Semi-CML, for multi-modal image segmentation, as shown in Fig. 1. The Semi-CML framework can perform cross-modal knowledge mutual learning on multi-modal data via low-coupling consistency losses using a large amount of unlabeled data. Concretely, first, we build two U-Nets (Ronneberger et al., 2015) with identical structures as segmentation network backbones for two different modalities. Then, two mini-batches of prediction maps in different modalities are obtained by forward propagation of the two U-Nets, where the labeled batches of two modalities perform for supervised learning. Second, for mutual learning between two modalities, the unlabeled batches of two modalities are fed into the MSE loss for the simple mutual learning and proposed area-similarity contrastive loss for the in-depth mutual learning.

3.1.1. Dual-modal supervised learning

Given a dual-modal dataset $\mathcal{D}'$ and $\mathcal{D}''$, the number of labeled data is N, and the number of unlabeled data is M. We define their labeled and unlabeled datasets $\mathcal{D}'_L$, $\mathcal{D}'_U$ and $\mathcal{D}''_L$, $\mathcal{D}''_U$ as follows:

$$\mathcal{D}'_L = \left\{\left(x'_i, y_i\right)\right\}_{i=1}^{N}, \mathcal{D}'_U = \left\{x'_i\right\}_{i=N+1}^{N+M}, \tag{1}$$

$$\mathcal{D}''_L = \left\{x''_i, y_i\right\}_{i=1}^{N}, \quad \mathcal{D}''_U = \left\{x''_i\right\}_{i=N+1}^{N+M}, \tag{2}$$

where $x'_i, x''_i \in \mathbb{R}^{H\times W}$ are different image modalities with the size of $H \times W$, and $y_i \in \{0,1\}^{C\times H\times W}$ is the ground-truth with C classes. In particular, the two image modalities x'_i, x''_i correspond to the same ground-truth mask y_i. Our semi-supervised architecture has as input two different image modalities with the same annotation in a training stage and gets their prediction results at the output end of the network at the same time. So, we define segmentation networks $F(\cdot)$ and $G(\cdot)$ with the same structure but different parameters for the two modalities as Model 1 and Model 2, respectively. For the supervised training of two different models, the predicted masks with annotations of each mini-batch are taken out to calculate the supervision loss. The supervision loss consists of the weighted sum of dice loss and binary cross-entropy loss, which is defined as follows:

$$\mathcal{L}_{\text{sup}}\left(\hat{y}, y\right) = \beta\mathcal{L}_{\text{bce}}\left(\hat{y}, y\right) + \gamma\mathcal{L}_{\text{dice}}\left(\hat{y}, y\right), \tag{3}$$

where $\mathcal{L}_{\mathrm{bce}}$ is the binary cross-entropy loss, $\mathcal{L}_{\mathrm{dice}}$ is the dice loss (Milletari et al., 2016), and β and γ are the weights of $\mathcal{L}_{\mathrm{bce}}$ and $\mathcal{L}_{\mathrm{dice}}$, respectively. Finally, we obtain the two supervised loss functions generated by different modalities, and their corresponding segmentation networks are optimized at the same time:

$$\min_{F,G} \mathcal{L}_{\mathrm{sup}}^{\mathrm{Total}}(F,G) = \mathbb{E}_{x',x'',y}\left[\mathcal{L}'_{\mathrm{sup}}\left(F\left(x'\right),y\right) + \mathcal{L}''_{\mathrm{sup}}\left(G\left(x''\right),y\right)\right], \tag{4}$$

where $\mathcal{L}'_{\mathrm{sup}}$ and $\mathcal{L}''_{\mathrm{sup}}$ optimize the parameters of segmentation networks $F(\cdot)$ and $G(\cdot)$, respectively, to learn the modality-specific information of each modality.

3.1.2. Cross-modal knowledge contrastive mutual learning

Consistency regularization has a great potential in semi-supervised learning. Previous works have proposed a variety of consistency regularization methods (Li et al., 2020c; Peng et al., 2020b; Tarvainen and Valpola, 2017; Xie et al., 2019). Differently from previous methods, in our proposed method, the consistency regularization term is designed based on the correlation and difference information between different modalities. The consistency of their predictions is used as the driving force of semi-supervised cross-modal knowledge mutual learning. This allows two different models to learn from each other by building a bridge between different modalities for information complementarity. Based on this, we design two cross-modal consistency loss functions. One is the MSE consistency loss for the simple mutual learning, and the other is a contrastive loss based on area similarity for the in-depth mutual learning. The cross-modal MSE consistency loss is defined as follows:

$$\min_{F,G} \mathcal{L}_{\mathrm{mse}}(F,G) = \mathbb{E}_{x',x''}\left[\left\|F\left(x'\right) - G\left(x''\right)\right\|^{2}\right], \tag{5}$$

which can achieve the simple mutual learning by minimizing the difference of the prediction results between two modalities.

However, the MSE loss is not good enough on the increase of the consistency lower bound of the two modalities mutual information (Tschannen et al., 2020), thereby leading to performance limitations in cross-modal complementary knowledge learning. Specifically, first, the MSE usually only constructs a distance error between paired predictions. Especially in the cross-modal prediction consistency, only the prediction similarity of paired modalities can be paid attention to, but the prediction dissimilarity for unpaired cross-modal data and unpaired same-modal data cannot be paid attention to. This causes the network to easily fall into over-fitting, because it only focuses on the easy-to-learn paired modal similarity, and at the same time causes the network to fail to learn more in-depth unpaired cross-modal and same-modal complementary information. Second, the MSE only considers the Euclidean distance metric between each pixel in two predictions and cannot pay attention to the edge and area context information of the prediction target. This is a disadvantage for image segmentation that needs to focus on the edge and region of the prediction target.

Therefore, we also design a novel area-similarity contrastive (ASC) loss to overcome these problems. Specifically, inspired by the promising self-supervised learning based on a contrastive loss (Chaitanya et al., 2020; Chen et al., 2020), we propose to use a contrastive learning algorithm based on positive and negative sample pairs to learn the cooperative information in the paired cross-modal data and the complementary information in the unpaired cross-modal data and the unpaired same-modal data. On the one hand, this cross-modal mutual learning algorithm based on contrastive learning can learn the consistency information between paired modalities by maximizing the prediction similarity between paired different modalities. On the other hand, we can learn the difference information between unpaired modalities by minimizing the similarity between negative sample pairs (unpaired cross-modal data and unpaired same-modal data). Therefore, we can not only force the network to focus on the prediction similarity of paired modalities, but indirectly focus on the prediction dissimilarity in unpaired cross-modal data and unpaired same-modal data by constructing negative sample learning, thereby alleviating the network falling into overfitting.

However, current contrastive learning usually uses image embedding vectors and the cosine similarity for representation learning (Chaitanya et al., 2020; Chen et al., 2020; You et al., 2021), which is not suitable for the segmentation task that requires dense pixel-wise classification. Specifically, first, if the segmentation task uses the embedding vector obtained by the additional projection head for the contrastive representation learning, it cannot directly learn the semantic information of the segmentation target. Second, although the cosine similarity focuses not only on the Euclidean distance between corresponding pixels like MSE, the cosine similarity cannot directly focus on the edge and area context information of the segmentation target. Therefore, our proposed ASC loss does not use a projection head but directly performs the contrastive learning on predicted segmentation maps to learn pixel-level information. At the same time, we use Dice similarity instead of cosine similarity as the similarity measurement function to pay attention to the area context information of the segmentation target.

Below, we give a detailed description of the proposed ASC loss. In the process of constructing positive and negative sample pairs, we treat the different classification regions of each sample as the same object (usually there are only two to three classes in medical image segmentation tasks), and construct positive and negative sample pairs between different modalities and different samples to focus on cross-modal and same-modal collaboration information and difference information. We have also noticed that the lesion area and location of the same disease in the medical image segmentation task may be a little similar between adjacent slices in the same patient scan. Therefore, we only randomly sample a small batch size from a large amount of unlabeled data to perform the contrastive learning, which can almost avoid the possibility of adjacent slices appearing in negative sample pairs. Specifically, we randomly selected K unlabeled samples as a mini-batch from two different image modalities, thus generating $2K$ data points as the input of contrastive loss. Among them, two image modalities corresponding to the same ground truth are taken as positive sample pairs (paired modalities). Now, K positive sample pairs are generated, defined as set Γ^{+}. For each positive pair, the remaining $2(K-1)$ data in the mini-batch are taken as the negative sample set, defined as Γ^{-}, which includes prediction results of different modalities and different slices (unpaired cross-modal data and unpaired same-modal data). Therefore, in a mini-batch of size K, the positive and negative sample sets are described as:

$$\Gamma^{+} = \left\{\left(F\left(x'_i\right), G\left(x''_i\right)\right)\right\}_{i=1}^{K}, \tag{6}$$

$$\Gamma^{-} = \left\{F\left(x'_j\right)\right\}_{j\neq i}^{K-1} \cup \left\{G\left(x''_j\right)\right\}_{j\neq i}^{K-1}. \tag{7}$$

Next, we give the specific formula of the proposed ASC loss. First, in order to measure the area context similarity between positive and negative sample pairs, we use Dice coefficient similarity as a similarity measurement function, which is defined as follows:

$$S_{\mathrm{dice}}\left(y_1, y_2\right) = \frac{2 \times \sum_{\substack{\dot{y_1}\in y_1,\\ \dot{y_2}\in y_2}} \dot{y_1}\,\dot{y_2}}{\sum_{\substack{q_1\in y_1,\\ q_2\in y_2}} \left(\dot{y_1} + \dot{y_2}\right)}, \tag{8}$$

where $\dot{y_1}$ and $\dot{y_2}$ denote the values of each pixel of two predictions y_1 and y_2, respectively. Second, we define our ASC loss for a positive pair as follows:

$$l_{\mathrm{asc}}(\hat{y}, \tilde{y}) = -\log \frac{\exp\left(S_{\mathrm{dice}}(\hat{y}, \tilde{y})\right)}{\exp\left(S_{\mathrm{dice}}(\hat{y}, \tilde{y})\right) + \sum_{h\in\Gamma^{-}} \exp\left(S_{\mathrm{dice}}(\hat{y}, h)\right)}, \tag{9}$$

where $(\hat{y}, \tilde{y})$ is a positive sample pair in Γ^{+}, and h is a negative sample corresponding to $\hat{y}$ in Γ^{-}. Then, the total ASC loss for a minibatch of K unlabeled images is as follows:

$$\min_{F,G} \mathcal{L}_{\mathrm{ASC}}(F,G) = \mathbb{E}_{x',x''}$$

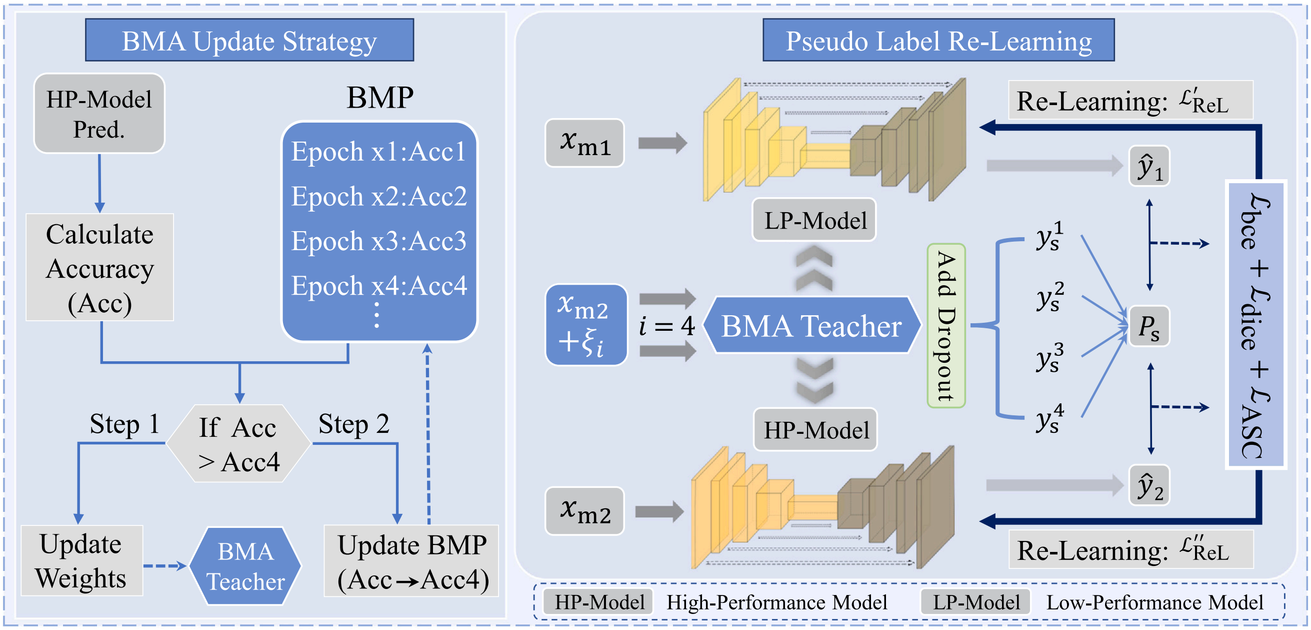


Fig. 2. Our proposed soft pseudo-label re-learning algorithm using the BMA teacher model based on the Semi-CML framework (Semi-CML with PReL). The left box shows the detailed workflow of using the BMA update strategy to generate a teacher model, where the generation of the BMA teacher model is only run in Stage 1 (before the L_1th epoch). The box on the right shows the pipeline that uses the previously generated BMA teacher model to perform soft pseudo-label re-learning for the low and high performance modality, where the re-learning supervision is run in Stage 2 (after the L_1th epoch).

$$\left[\frac{1}{2K}\sum_{(F',G'')\in\Gamma^+}\left(l_{\mathrm{asc}}\left(F',G''\right)+l_{\mathrm{asc}}\left(G'',F'\right)\right)\right], \tag{10}$$

where $F' = F(x')$ and $G'' = G(x'')$. Each sample in a positive sample pair has to calculate the similarity with the samples in the negative sample set, so a positive sample pair needs to calculate l_{asc} twice.

In addition, to balance the MSE loss and the ASC loss, the weight coefficients w_1 and w_2 are added, where w_1 is a ramp-up function to adjust the weight value according to the epoch number, just like Tarvainen and Valpola (2017), and w_2 is a scalar. Finally, the overall loss of the Semi-CML framework for the training of Stage 1 is defined as follows:

$$\min_{F,G}\mathcal{L}_{\mathrm{CML}}(F,G)=\mathcal{L}_{\mathrm{sup}}^{\mathrm{Total}}+w_1\mathcal{L}_{\mathrm{mse}}+w_2\mathcal{L}_{\mathrm{ASC}}. \tag{11}$$

3.2. Soft pseudo-label re-learning using the BMA teacher model

Although the above methods greatly improve the segmentation performance of both modalities, our experiments show that there exists a gap between the performance of two modalities, i.e., there is a modality whose segmentation performance is usually better than that of the other. To further improve the segmentation precision of the low-performance modality, we design a soft pseudo-label re-learning strategy by using the model with high-performance modality, as shown in Fig. 2 and Algorithm 1. To be specific, motivated by EMA-weighted models (Tarvainen and Valpola, 2017) and dropout as a Bayesian approximation (Gal and Ghahramani, 2016), we first design a novel best-model moving average (BMA) self-ensembling technology to generate an optimal and reliable teacher model during Stage 1 (i.e., before a certain epoch when Semi-CML reaches convergence, L_1). The teacher model is called the best-model moving average self-ensembling teacher model (BMA Teacher). Second, in Stage 2 (after the L_1th epoch), we use the BMA teacher model and Monte-Carlo dropout sampling to generate soft pseudo-labels with higher reliability. Then, re-learning is performed for the low-performance modality using the soft pseudo-labels. In addition, the teacher model is also helpful for the high-performance modality. Therefore, the re-learning process is also executed for the high-performance model but starts after the warm-up epoch (L_2, lags the L_1th epoch). We assume that Model 1 and Model 2 are the low and high performance models.

3.2.1. Best-model moving average teacher model

The student–teacher model is widely used in semi-supervised learning algorithms, where the teacher model usually uses the exponential moving average (EMA) to update network parameters (Tarvainen and Valpola, 2017; Yu et al., 2019). This method usually updates the model weights in every epoch or minibatch (though the model performance is poor), which results in a decrease in the performance of the teacher model. This is because a deep model usually has performance shocks due to unstable training (as can be clearly seen from the 5th to 40th epoch in Fig. 7(b)), where the model weights that are obtained when the model falls into a low-performance state may be unreliable. At this time, if the weights of this model are updated to the teacher model, the reliability of the teacher model will be reduced. Therefore, to integrate only high-quality model weights, we propose a novel Best-model Moving Average (BMA) self-ensembling method to selectively update the weights of the teacher model, defined as $T(\theta)$. Specifically, we decide whether to update the teacher model based on the training or validation accuracy of each epoch, so that only the optimal (compared to the model performance of all previous epochs) or sub-optimal high-performance model weights are selected to update the teacher model weight. For ensuring that the optimal or suboptimal model weights are updated to the teacher model, we design a Best Model Pool (BMP), defined as set $\mathbb{R}_{\mathrm{pool}}$, to dynamically store p (BMP Number, e.g., 6) optimal or suboptimal accuracy values of different epochs.

Specifically, for Stage 1, the BMA teacher model is updated after m epochs (e.g., 10), because better model weights are usually not available in the first m epochs, and all updates are only completed in Stage 1 (i.e., before epoch L_1). First, in p epochs after the mth epoch, we initialize the BMP with a capacity of p using the accuracy of these p epochs. At the same time, in the $(m + p + 1)$th epoch, the weights of the high-performance model are directly used as the initial weights of the teacher model. Then, for each epoch between $(m + p + 1)$ and L_1

Algorithm 1 Soft Pseudo-Label Re-Learning algorithm.

```
Input: Batch of unlabeled examples of different modalities $\mathcal{D}'_U = (x'_i; i \in (1,\dots,N+M))$, $\mathcal{D}''_U = (x''_i; i \in (1,\dots,N+M))$, max epoch $t_{\max}$, BMP number $p$, start epoch of BMA update $m$, start epoch of PReL for Model 1 $L_1$, start epoch of PReL for Model 2 $L_2$. We assume that Model 2 is a high-performance model.
for $t = 0$ to $t_{\max}$ do
  if $t < L_1$ then
    Training for Semi-CML using Eq. (11)
    Calculate $Acc_t$ using the high-performance model and get weights $\theta''_t$ from the high-performance model {Make two preparations for updating of BMA Teacher.}
    if $m < t <= (m + p)$ then
      $Acc_t \Rightarrow \mathbb{R}_{\mathrm{pool}}$ {Initialize the BMA Pool using $Acc_t$.}
    else if $t = (m + p + 1)$ then
      $\theta_t = \alpha\theta_{t-1} + (1-\alpha)\theta''_t, \alpha = 0.99$ {Initialize the BMA Teacher using weights $\theta_t$.}
    else if $t > (m + p + 1)$ and $Acc_t > \min(\mathbb{R}_{\mathrm{pool}})$ then
      $Acc_t \Rightarrow \mathbb{R}_{\mathrm{pool}}$ {Continually update the BMA Pool.}
      $\theta_t = \alpha\theta_{t-1} + (1-\alpha)\theta''_t$, $\alpha$ is equal to Eq. (12) {Continually update the BMA Teacher using weights $\theta_t$.}
    end if
    Get BMA Teacher ($T$).
  else
    $P_s = \frac{1}{D}\sum_{i=1}^{D} T\left(x'' + \xi_i\right)$ {Generating pseudo-labels using the BMA Teacher.}
    Calculate loss $\mathcal{L}'_{\mathrm{ReL}}$ using Eq. (16) {Training Model 1 using the pseudo-labels $P_s$}
    if $L_2 < t < t_{\max}$ then
      Calculate loss $\mathcal{L}''_{\mathrm{ReL}}$ using Eq. (17) {Training Model 2 using the pseudo-labels $P_s$ after warm-up epoch.}
    end if
  end if
end for
```

epochs, we need to perform two steps after meeting certain conditions. For Step 1, we need to continuously update the BMP to ensure that it contains the optimal and sub-optimal accuracy values. For Step 2, we use the function of the best-model moving average to continuously update the weights of the teacher model. Concretely, we compare the accuracy of the current epoch with the minimum accuracy in the BMP to decide whether to update the BMP and the teacher model. If it is greater than the minimum value, we update the minimum value in the pool with the current accuracy to ensure that the accuracy value in the pool is the top p best accuracy in the previous L_1 epochs. At the same time, we use the BMA update function to update the weights of the teacher model. The weight update ratio of the teacher model changes dynamically according to the increase of accuracy, which can be defined as follows:

$$\alpha = \min\left(1 - \frac{Acc_t - Acc_{min}}{Acc_t}, \alpha_0\right), \tag{12}$$

where Acc_t is the accuracy at the training epoch t, $Acc_{\min} = \min\left(\mathbb{R}_{\mathrm{pool}}\right)$, and $\alpha_0 = 0.99$. Finally, all the update procedures of the BMP and the BMA teacher are described as follows:

$$\begin{cases} Acc_t \Rightarrow \mathbb{R}_{\mathrm{pool}} & \text{if } Acc_t > Acc_{min} \\ \theta_t = \alpha\theta_{t-1} + (1-\alpha)\theta''_t & \text{if } Acc_t > Acc_{min}, \end{cases} \tag{13}$$

where θ_t and θ''_t are the weights of the teacher model and the high-performance model (e.g., $G(\cdot)$) at the training epoch t, respectively. Updating the teacher model weights in this method can filter out model weights with poor performance and make the weight update ratio larger when the model is better, thereby maximizing the ensemble of the highest-quality model weights.

3.2.2. *Generating pseudo-labels and re-learning*

In Stage 2, we perform Monte-Carlo dropout sampling (Yu et al., 2019) on the BMA teacher model $T(\theta)$ to get more reliable prediction results as pseudo-labels for unlabeled data. In detail, we apply dropout layers on the teacher model as an approximation of the Bayesian neural network and apply Gaussian noise to the input. Then, the teacher model performs D random forward propagation and averages the obtained prediction results to obtain the final reliable soft pseudo-label. Formally, this process can be defined as:

$$P_s = \frac{1}{D}\sum_{i=1}^{D} T\left(x'' + \xi_i\right), \tag{14}$$

where P_s is the reliable soft pseudo-label, and ξ_i is the Gaussian noise. Then, we use the generated soft pseudo-labels to obtain a new loss. They are composed of two parts, including the supervision loss ($\mathcal{L}_{\mathrm{sup}}$) and the ASC loss ($\mathcal{L}_{\mathrm{ASC}}$) following Semi-CML. In the re-learning stage, the positive and negative sample pairs in the ASC loss consist of the prediction from one of two modalities and the soft pseudo-label generated by the BMA teacher model, which is defined as follows:

$$\mathcal{L}_{\mathrm{ASC}}^{\mathrm{ReL}}\left(\hat{y}, P_s\right) = \frac{1}{2K}\sum_{(\hat{y},P_s)\in\Gamma^+}\left(l_{asc}\left(\hat{y}, P_s\right) + l_{asc}\left(P_s, \hat{y}\right)\right), \tag{15}$$

where $\hat{y}$ is the output of the supervised model and K is the number of unlabeled images in a mini-batch. The pseudo-label re-learning process is mainly for low-performance modality, so the supervising formula is defined as follows:

$$\min_F \mathcal{L}'_{\mathrm{ReL}}(F) = \mathbb{E}_{x',P_s} \left[\alpha_1 \mathcal{L}_{\mathrm{ASC}}^{\mathrm{ReL}}\left(F\left(x'\right), P_s\right) + \left(1-\alpha_1\right)\mathcal{L}'_{\mathrm{sup}}\left(F\left(x'\right), P_s\right)\right], \tag{16}$$

where α_1 is a balance factor for the training of a low-performance modality. The pseudo-labels generated by the BMA teacher model are also helpful for the training of a high-performance modality. Therefore, we also perform a similar supervision for a high-performance modality, which is defined as follows:

$$\min_G \mathcal{L}''_{\mathrm{ReL}}(G) = \mathbb{E}_{x'',P_s} \left[\alpha_2 \mathcal{L}_{\mathrm{ASC}}^{\mathrm{ReL}}\left(G\left(x''\right), P_s\right) + \left(1-\alpha_2\right)\mathcal{L}''_{\mathrm{sup}}\left(G\left(x''\right), P_s\right)\right], \tag{17}$$

where α_2 is a balance factor for the training of high-performance modality. To ensure the low-performance modality has a more stable optimization during the re-learning process, the re-learning process of high-performance modality only starts after the warm-up epoch (L_2) based on epoch L_1. After the pseudo-label re-learning strategy, the segmentation performance of the two models will be further improved. In general, the accuracy of the low-performance modality will increase even more, because this learning process provides more cross-modal knowledge.

4. Experiments

4.1. *Datasets*

To verify the effectiveness of our approach, we have conducted extensive evaluations on two publicly available multi-modal datasets, MICCAI 2020 Hecktor Challenge (Andrearczyk et al., 2020; Yuan, 2020) and BraTS 2019 Challenge (Bakas et al., 2017, 2018; Menze et al., 2014). Each dataset is randomly divided into two parts based on patient-level 3D image volumes: 80% of the 3D image volumes are used as the training set and the other 20% are used as the test set. Then, we get sliced data from the z-axis in each 3D volume for 2D image segmentation, and remove the background only slices (no lesion included) to prevent data imbalance. Finally, we obtain 5788 slices for the Hecktor dataset and 13 850 slices for the BraTS dataset. The details of datasets and preprocessing are as follows.

Table 1
The results of Semi-CML with PReL and the state-of-the-art fully-supervised and single-modal semi-supervised segmentation methods on the Hecktor and BraTS datasets with 1%, 5%, and 10% labeled data in terms of DSC and Sens.

Methods		Hecktor				BraTS							
Modality		CT		PET		T2		T1CE		T1		FLAIR	
Metrics		DSC	Sens	DSC	Sens	DSC	Sens	DSC	Sens	DSC	Sens	DSC	Sens
1%	Sup	0.1106	0.1058	0.2807	0.3766	0.2161	0.2196	0.1358	0.1459	0.1249	0.1183	0.3296	0.3257
	MT	0.1423	0.1545	0.3031	0.4749	0.2429	0.2526	0.2628	0.3145	0.1586	0.2241	0.3461	0.3410
	ICT	0.1394	0.1577	0.3401	0.3365	0.2483	0.2767	0.2233	0.2514	0.1672	0.1523	0.3616	0.3798
	DTC	0.1621	0.2133	0.3215	0.3854	0.2442	0.2766	0.2563	0.3584	0.1614	0.2008	0.3374	0.3182
	DTML	0.1337	0.1541	0.3130	0.4449	0.2435	0.2495	0.2656	0.2980	0.1559	0.2037	0.3726	0.3753
	SASS	0.1797	0.1920	0.3116	0.4091	0.2391	0.3063	0.2253	0.2175	0.1745	0.2336	0.3331	0.3264
	UAMT	0.1222	0.1326	0.3195	0.3461	0.2823	0.3501	0.2350	0.2147	0.1841	0.2291	0.3565	0.3619
	UMCT	0.1546	0.3311	0.3648	0.4590	0.2445	0.2727	0.2544	0.2551	0.1804	0.2407	0.3304	0.3232
	SPCT	0.1410	0.3605	0.3321	0.4826	0.3092	**0.3556**	0.2647	0.2537	0.1855	0.2932	0.3582	0.3441
	Ours	**0.2837**	**0.3627**	**0.4260**	**0.4886**	**0.3391**	0.3470	**0.4316**	**0.4718**	**0.3184**	**0.3644**	**0.4337**	**0.4833**
5%	Sup	0.2375	0.2709	0.4776	0.5221	0.3229	0.4876	0.3852	0.3613	0.2255	0.2085	0.3987	0.4198
	MT	0.2669	0.3639	0.5103	0.6320	0.3857	0.3747	0.4504	0.4491	0.2713	0.3071	0.4090	0.4290
	ICT	0.2430	0.3420	0.5403	0.6785	0.3648	0.3307	0.4660	0.4340	0.2742	0.3026	0.4035	0.4150
	DTC	0.2513	0.3584	0.5163	0.6086	0.3770	0.3643	0.4562	0.4335	0.2730	0.2702	0.4011	0.4163
	DTML	0.2729	0.2939	0.5114	0.5885	0.3761	0.3769	0.4520	0.4308	0.2743	0.2764	0.3586	0.3506
	SASS	0.2663	0.3924	0.5224	0.5573	0.3624	0.3611	0.4508	0.4305	0.2728	0.2612	0.4154	0.4327
	UAMT	0.3029	0.3735	0.5244	0.5870	0.3993	0.3983	0.4697	0.4777	0.3104	0.3556	0.4209	0.4683
	UMCT	0.2712	**0.4549**	0.5419	0.6196	0.3942	0.3866	0.4681	0.4377	0.3093	0.3432	0.4410	0.4922
	SPCT	0.2899	0.3729	0.5410	0.6486	0.4135	0.4059	0.4581	0.4038	0.2959	0.3176	0.4312	0.4346
	Ours	**0.3835**	0.4411	**0.5868**	**0.6823**	**0.4806**	**0.5455**	**0.6121**	**0.6403**	**0.4049**	**0.4962**	**0.4854**	**0.5858**
10%	Sup	0.2866	0.3688	0.5080	0.5931	0.4368	0.4020	0.4920	0.4416	0.2724	0.2597	0.4264	0.4410
	MT	0.3034	0.3687	0.5320	0.6238	0.4468	0.4243	0.5627	0.5410	0.3328	0.3381	0.4507	0.5024
	ICT	0.2960	0.4496	0.5525	0.6642	0.4588	0.4572	0.6308	0.5938	0.3733	0.3805	0.4571	0.4976
	DTC	0.3075	0.4111	0.5419	0.6397	0.4600	0.4619	0.5765	0.5192	0.3265	0.3169	0.4585	0.4905
	DTML	0.2944	0.4225	0.5483	0.6870	0.4760	0.4483	0.5464	0.5023	0.3283	0.3003	0.4450	0.4701
	SASS	0.2969	0.4435	0.5507	0.6715	0.4553	0.4559	0.5530	0.5071	0.3229	0.3087	0.4609	0.4842
	UAMT	0.3098	0.3913	0.5501	**0.7266**	0.4521	0.4339	0.6157	0.5926	0.3424	0.3610	0.4736	0.5590
	UMCT	0.3257	**0.5269**	0.5545	0.6324	0.4670	0.4556	0.5937	0.5662	0.3549	0.3448	0.4897	0.5133
	SPCT	0.3281	0.4050	0.5617	0.6612	0.4783	0.4592	0.5908	0.5589	0.3474	0.3237	0.4979	0.5360
	Ours	**0.3942**	0.4874	**0.6072**	0.6897	**0.5612**	**0.6189**	**0.6656**	**0.6917**	**0.4508**	**0.4990**	**0.5302**	**0.6079**
Sup (50%)		0.3334	0.3713	0.5476	0.6351	0.5209	0.5033	0.6790	0.6397	0.4464	0.4322	0.5252	0.5755
Sup (100%)		0.4057	0.4390	0.5806	0.6410	0.5645	0.5654	0.7083	0.6749	0.4898	0.4760	0.5569	0.5778

Hecktor. The dataset is provided by a MICCAI 2020 challenge called HEad and neCK TumOR segmentation challenge (Hecktor), which is a multi-modal CT-PET dataset consisting of 201 3D image volumes. All CT volumes are clipped within the range $[-150, 150]$ Hounsfield Units (HU). Max–min normalization is performed for PET volumes in the range of 0 to 1. All slices for both modalities are cropped to the size of 144 × 144.

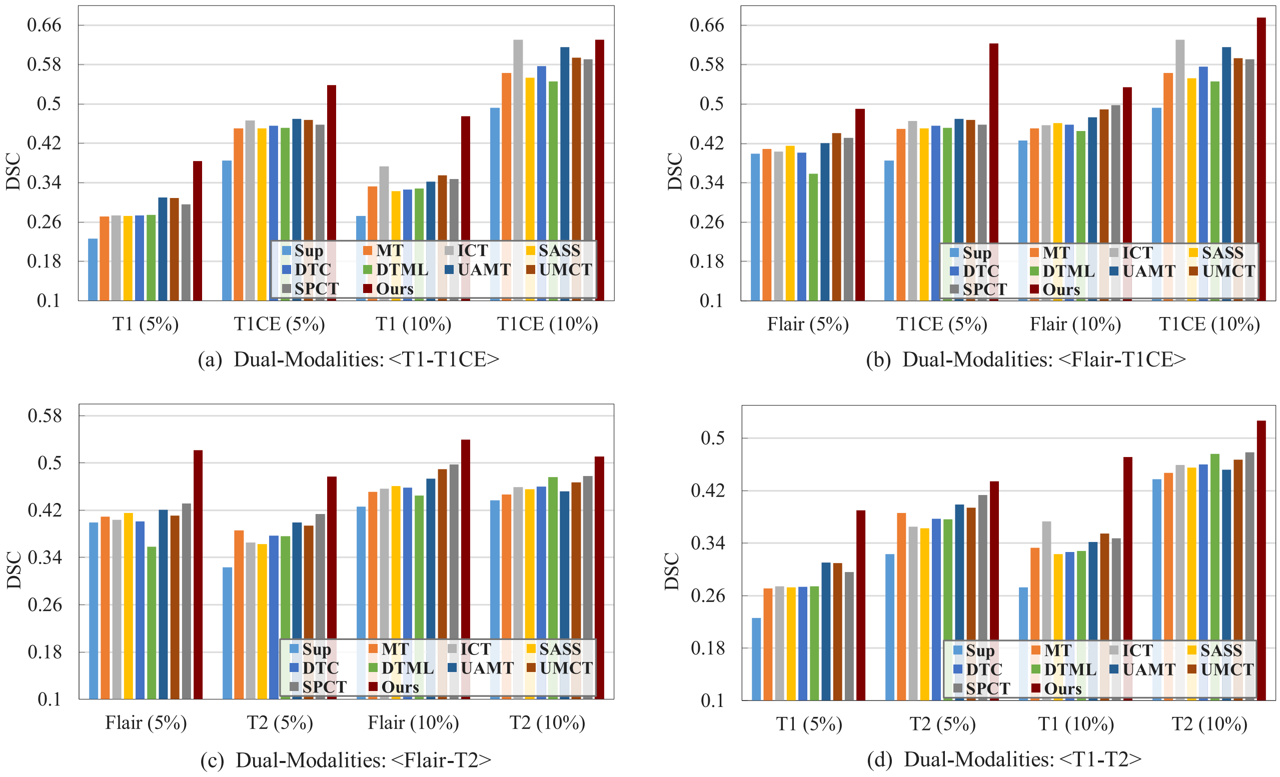
BraTS 2019**.** The dataset is designed for brain tumor segmentation using multi-modal magnetic resonance imaging (MRI) scans, containing 259 high-grade gliomas (HGG) data and 76 low-grade gliomas (LGG) data, and we only use the HGG data in our experiments. The dataset contains four modalities: T1, T1CE, T2, and FLAIR. The task is to segment three areas, namely, whole tumor (WT), enhanced tumor (ET), and tumor core (TC). Each volume is normalized to zero mean and unit variance; slices are center-cropped to the size of 160 × 160.

4.2. Implementation details

All models are implemented using PyTorch 1.6, CUDA 10.1, and are run on a GeForce RTX 2080 TI GPU. The training time of our method only needs about 40 min for the Hecktor dataset and 90 min for the BraTS dataset. Our method and all baselines use the same code base, including the same segmentation network backbone, training process, and evaluation methods. We fix a random seed to ensure that the same training result and evaluation result can be obtained under the same hyperparameters. For a fair comparison, we perform full hyperparameter tuning for each model under each case and report the optimal results. Specific experimental details are as follows.

The U-Net (Ronneberger et al., 2015) model is used as the segmentation backbone network for all methods. The U-Net encoder includes four maximum poolings to reduce the original image resolution by 16 times, and it includes five layers of double convolutional blocks. The feature numbers of these blocks are 32, 64, 128, 256, and 512, respectively. For the U-Net decoder, four transposed convolutions are performed to restore the original image resolution. In the semi-supervised setting, we randomly select labeled and unlabeled data at three different scales (1%, 5%, and 10%). We have two types of mini-batches in our work, the size of the mini-batch of labeled slices is 30 (resp., 20) and that of the mini-batch of unlabeled slices (i.e., K in Eqs. (10) and (15)) is 26 (resp., 4) for the Hecktor (resp., BraTS) dataset. The max epoch t_{max} (in Algorithm 1) is set to 81 for 5% and 10% labeled data, and to 121 for 1% labeled data.

Furthermore, the values of hyperparameters of our proposed method are determined by searching within some empirically defined value sets/ranges. Since the hyperparameter search is conducted for each segmentation task to get the optimal results, and our experiments evaluate several tasks, to keep it concise, we give here the search strategies of the hyperparameters instead of showing the specific values of all hyperparameters under all tasks (but the specific values can still be found in the online released codes). Specifically, the initial learning rate is searched within the value range $[0.0002, 0.008]$ with a step of 0.0004, and the learning strategy is warmup MultiStep (Goyal et al., 2017), which increases the learning rate slowly and then decreases in multiple steps. We use the Adam (Kingma and Ba, 2014) optimizer, and the weight decay coefficient is searched within the value set $\{1e-4, 1e-3, 1e-2, 1e-1\}$. β and γ in the supervised loss (in Eq. (3)) are 10.0 and 7.0, respectively. In Semi-CML, the weight of MSE w_1 (in Eq. (11)) follows a ramp-up function to adjust the weight value according to the epoch number (similar setting as in Tarvainen and Valpola (2017)), where the weight coefficient in ramp-up function is searched within



(a) Dual-Modalities: <T1-T1CE>

(b) Dual-Modalities: <Flair-T1CE>

(c) Dual-Modalities: <Flair-T2>

(d) Dual-Modalities: <T1-T2>

Fig. 3. The DSC results of Semi-CML with PReL and the state-of-the-art fully-supervised and single-modal semi-supervised segmentation methods using other combinations of two modalities from the four modalities in the BraTS dataset with 5% and 10% labeled data.

the value set $\{0.001, 0.05, 0.1, 1.0, 10, 50, 100\}$; the weight of ASC w_2 (in Eq. (11)) is searched within the value range $[2.5, 4]$ with a step of 0.1. In PReL, the start epoch of BMA update m (in Algorithm 1) is searched within the value set $\{5, 10\}$; the start epoch of PReL for Model 1, L_1 (in Algorithm 1), is searched within the value set $\{51, 55, 61, 71\}$; the start epoch of PReL for Model 2, L_2 (in Algorithm 1), is calculated by the function $L_2 = (t_{\max} - L_1) \times r + L_1$, where the value of r is searched within the value set $\{0.4, 0.6, 0.8\}$; the BMP number p (in Algorithm 1) is searched within the value set $\{1, 2, 4, 6\}$; the times of Monte Carlo samples D (in Eq. (14)) is set by searching the value set $\{4, 8\}$; and the balancing factors α_1 (in Eq. (16)) and α_2 (in Eq. (17)) are both searched within the value set $\{0.05, 0.1, 0.5, 0.9\}$ for the Hecktor dataset and within the value set $\{0.1, 0.5, 0.9, 0.95, 0.99\}$ for the BraTS dataset.

4.3. Evaluation metrics

We mainly use two evaluation metrics, dice similarity coefficient (DSC) and sensitivity (Sens), to verify the segmentation performances of all methods. The formal definition of DSC and Sens functions are as follows: $\text{DSC} = 2TP/(FP + 2TP + FN)$ and $\text{Sens} = TP/(TP + FN)$, where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively. Furthermore, to show the superior segmentation performances of our proposed method more comprehensively, a boundary-based evaluation metric Boundary IoU (BIoU) (Cheng et al., 2021) is additionally used in our experiments, which is more sensitive to the boundary errors of the target areas and does not overpunish the errors of smaller objects. The reported results of DSC and Sens are the average values among patients' 3D image volumes, while those of BIoU are the average results among 2D slices.

4.4. Main results

4.4.1. Comparison with state-of-the-art fully-supervised methods

We first compare our method with the fully-supervised method (denoted Sup) using the 1%, 5%, and 10% labeled data on both two datasets. The results in Table 1 show that our semi-supervised model has a significant improvement in any modalities of two datasets. For example, when only 1% of labeled Hecktor data are used, the DSC results of our method are 0.2837 and 0.4260 on the CT and PET modalities, respectively; but they are only 0.1106 and 0.2807 if using the fully-supervised learning method (Sup). Furthermore, the DSC results of our method are 0.1577 and 0.2269 higher than those of Sup for the T2 and T1CE modalities on BraTS using 5% labeled data; while the increases are 0.1784 and 0.1038 for the T1 and FLAIR modalities on BraTS using 10% labeled data. We then compare our method with the fully-supervised solution using more labels. The segmentation results of our model (with 10% labeled data) surpass most of the results of Sup using 50% labeled data and are close to the results of Sup using 100% labeled data. Especially, on the Hecktor dataset, our method (with 10% labeled data) have generally outperformed the fully-supervised method with 100% labeled data. These thus prove the effectiveness of our method in making use of a large amount of unlabeled data to help deep model achieve great performance improvements.

4.4.2. Comparison with state-of-the-art single-modal semi-supervised methods

Several state-of-the-art semi-supervised single-modal segmentation methods are also used as the baselines, including mean teacher (MT) (Tarvainen and Valpola, 2017), interpolation consistency training (ICT) (Verma et al., 2019), dual-task consistency (DTC) (Luo et al., 2021), dual-task mutual learning (DTML) (Zhang and Zhang, 2021), shape-aware semi-supervision (SASS) (Li et al., 2020b), uncertainty-aware mean teacher (UAMT) (Yu et al., 2019), uncertainty-aware multi-view co-training (UMCT) (Xia et al., 2020), and self-paced and self-consistent co-training (SPCT) (Wang et al., 2021a). Since these methods are designed for single-modal images, and our method only requires a single modality in the inference phase (multi-modal data are used only in the training phase), we choose the above methods and train their models with data of both modalities to compare with ours. The results are also shown in Table 1.

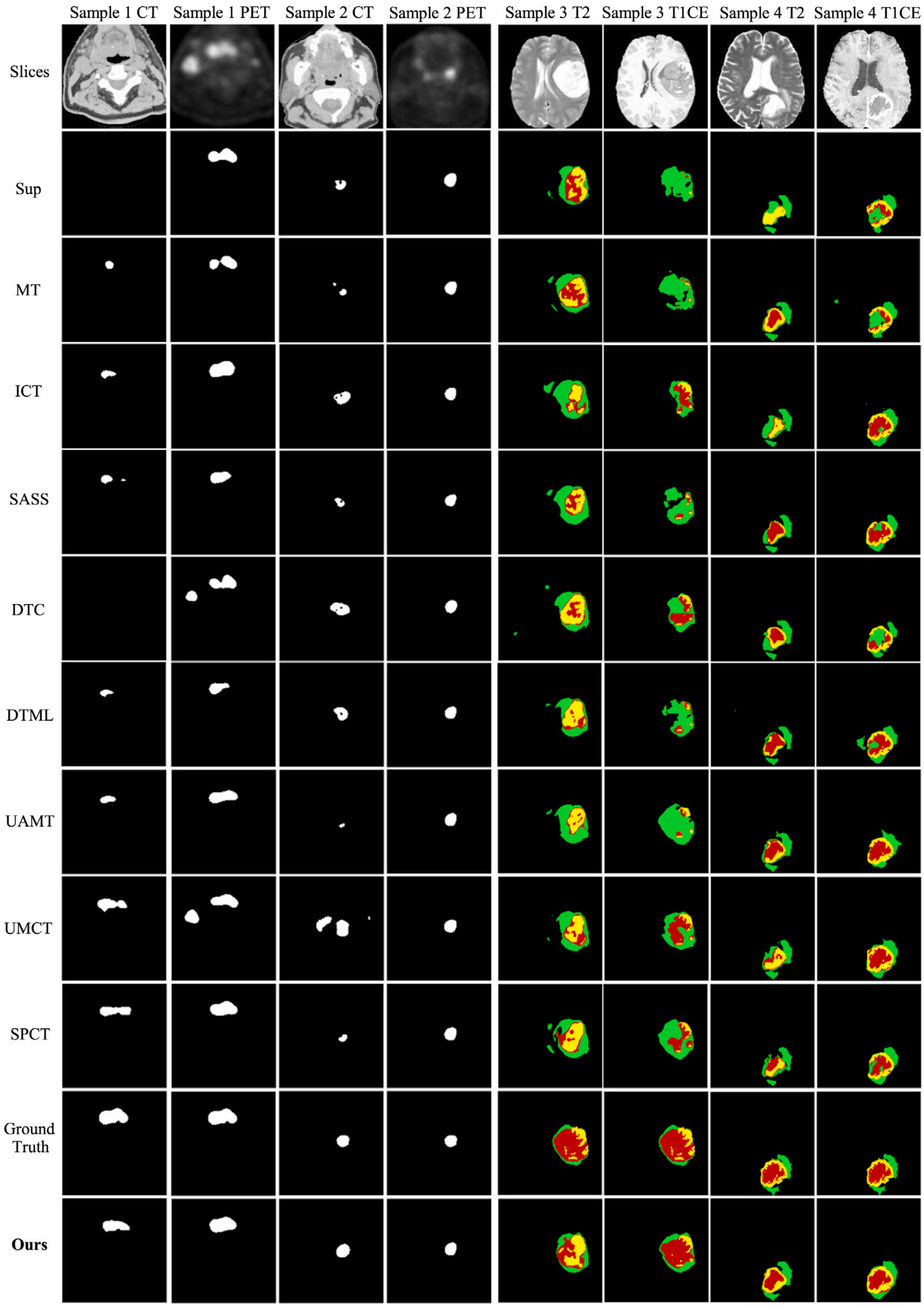


Fig. 4. Visualization of segmentation results using dual-modal images for the same slice in the Hecktor and BraTS datasets with 10% labeled data.

Generally, it can be witnessed that our method significantly outperforms all the state-of-the-art single-modal semi-supervised methods in almost all cases on both datasets in terms of both DSC and Sens, which proves the superior medical image segmentation performances of our work. Furthermore, some other observations are as follows. First, all the semi-supervised methods have obtained certain performance improvements comparing to the fully-supervised methods (Sup), indicating that semi-supervised segmentation methods can effectively utilize unlabeled data to improve the models' performances. Second, the co-training methods obtain better segmentation performances than other

Table 2
The segmentation results of our method and multi-modal semi-supervised methods on both datasets with 1%, 5%, and 10% labeled data in terms of DSC and Sens, where bold (resp., underlined) values are the best results of our method and the baselines using single (resp., dual) modality inference.

Methods		Inference modality	1% labeled data		5% labeled data		10% labeled data	
			DSC	Sens	DSC	Sens	DSC	Sens
Hecktor	MM-MT	CT-PET	0.3137	0.4360	0.5344	0.6162	0.5637	0.6972
		PET	0.2879	0.4040	0.3774	0.3455	0.5077	0.5928
	MM-ICT	CT-PET	0.3406	0.4520	0.5534	0.6543	0.5737	0.6883
		PET	0.0528	0.0311	0.2402	0.1994	0.1289	0.0963
	MM-DTC	CT-PET	0.3282	0.3887	0.5514	0.6211	0.5887	0.6920
		PET	0.2256	0.2437	0.3343	0.3076	0.5114	0.5260
	MM-DTML	CT-PET	0.3159	0.4055	0.5614	0.6802	0.5926	0.6785
		PET	0.2752	0.2560	0.2897	0.2327	0.3474	0.3055
	MM-SASS	CT-PET	0.3202	0.4451	0.5575	0.6067	0.5805	0.7170
		PET	0.2234	0.2243	0.4600	0.5287	0.2710	0.2456
	MM-UAMT	CT-PET	0.3307	0.4442	0.5543	0.6523	0.5961	0.7202
		PET	0.1472	0.1031	0.4196	0.5349	0.5429	0.6024
	MM-UMCT	CT-PET	0.3306	0.5315	0.5534	0.6897	0.5969	0.7361
		PET	0.2625	0.3667	0.2445	0.2706	0.3833	0.3518
	MM-SPCT	CT-PET	0.3454	0.4655	0.5885	0.6632	0.6083	0.6656
		PET	0.2977	0.3061	0.4662	0.4375	0.3947	0.3835
	DAFNet	CT-PET	0.4190	0.5580	0.5450	0.6600	0.5840	0.6930
		PET	0.3510	0.4320	0.3550	0.3070	0.3950	0.3730
	FewGAN	CT-PET	0.3388	0.4618	0.5216	0.7448	0.5773	0.6404
		PET	0.1307	0.2417	0.0424	0.0263	0.4432	0.4981
	Ours	PET	**0.4260**	**0.4886**	**0.5868**	**0.6823**	**0.6072**	**0.6897**
BraTS (T2-T1CE)	MM-MT	T2-T1CE	0.4579	0.4226	0.6269	0.5915	0.6946	0.6412
		T1CE	0.1265	0.3526	0.0895	0.4502	0.0909	0.3972
	MM-ICT	T2-T1CE	0.4522	0.4208	0.6324	0.5881	0.6992	0.6364
		T1CE	0.1238	0.3053	0.0968	0.3915	0.1417	0.3633
	MM-DTC	T2-T1CE	0.4378	0.4296	0.6071	0.5550	0.6607	0.6034
		T1CE	0.1049	0.4036	0.1036	0.3518	0.1159	0.4014
	MM-DTML	T2-T1CE	0.4455	0.4173	0.6033	0.5482	0.6777	0.6346
		T1CE	0.0628	0.3840	0.0860	0.3137	0.0943	0.4341
	MM-SASS	T2-T1CE	0.4119	0.3733	0.6187	0.5708	0.6814	0.6293
		T1CE	0.1175	0.3140	0.1059	0.2951	0.1135	0.1135
	MM-UAMT	T2-T1CE	0.4601	0.4368	0.6445	0.6181	0.7069	0.6873
		T1CE	0.1209	0.3881	0.0834	0.4129	0.1066	0.4793
	MM-UMCT	T2-T1CE	0.4629	0.4485	0.6440	0.6018	0.7187	0.6590
		T1CE	0.0643	0.4382	0.1122	0.4067	0.1636	0.4284
	MM-SPCT	T2-T1CE	0.4817	0.4257	0.6478	0.5894	0.7191	0.6614
		T1CE	0.1219	0.3612	0.1213	0.3587	0.1307	0.4154
	DAFNet	T2-T1CE	0.4960	0.4790	0.6120	0.5120	0.6830	0.6200
		T1CE	0.3940	0.3770	0.4980	0.3810	0.5390	0.4330
	FewGAN	T2-T1CE	0.4612	0.5043	0.6165	0.6773	0.6795	0.6296
		T1CE	0.0924	0.4118	0.0980	0.5515	0.1128	0.3739
	Ours	T1CE	**0.4316**	**0.4718**	**0.6121**	**0.6403**	**0.6656**	**0.6917**

(*continued on next page*)

semi-supervised methods, which proves the effectiveness of multi-view co-training. Third, the existing single-modal semi-supervised methods generally have poor performance improvements for the low-performance image modalities on both Hecktor and BraTS. For example, with 10% labeled data, the DSC improvements of the best baseline (SPCT) compared to the fully-supervised method (Sup) are only 0.0415, 0.0415, and 0.0750 in the low-performance modalities, CT, T2, and T1, while those are 0.0537, 0.0988, and 0.0715 in the corresponding high-performance modalities, PET, T1CE, and FLAIR, respectively. However, our method can achieve significant performance improvements in both low-performance and high-performance modalities. For example, with 10% labeled data, the DSC improvements of our method compared to the fully-supervised method are 0.1076, 0.1244, and 0.1784 in the low-performance modalities, CT, T2, T1, respectively; while those in the corresponding high-performance modalities, PET, T1CE, and FLAIR, are 0.0992, 0.1736, and 0.1038, respectively. These results fully demonstrate that, with the help of mutual learning in CML and PReL, our method can utilize the information in both modalities to complement each other and thus achieve great performance improvements in both modalities.

To further verify the effectiveness of our proposed method, other combinations of two modalities from the four modalities in the BraTS dataset are also used for evaluation; Fig. 3 exhibits the four groups of experimental results in DSC using 5% and 10% labeled data. Generally, our method still achieves better segmentation performances than the state-of-the-art fully-supervised and semi-supervised segmentation methods in any combination of two modalities. To qualitatively demonstrate the superior performances of our proposed method, we visualize the segmentation results of our method, the fully-supervised method, and the semi-supervised baselines with 10% labeled data in Fig. 4. We show the results of CT and PET images corresponding to the same slice number in the Hecktor dataset, and those of T2 and T1CE in the BraTS dataset. In Fig. 4, our method shows a higher rate of lesion area overlap and fewer false positives than the baselines. Furthermore, for a given

Table 2 (*continued*).

Methods		Inference modality	1% labeled data		5% labeled data		10% labeled data	
			DSC	Sens	DSC	Sens	DSC	Sens
BraTS (T1-FLAIR)	MM-MT	T1-FLAIR	0.3929	0.4132	0.4433	0.4582	0.4912	0.5073
		FLAIR	0.2407	0.2166	0.2178	0.2071	0.2946	0.3142
	MM-ICT	T1-FLAIR	0.3836	0.3865	0.4443	0.4630	0.5105	0.5435
		FLAIR	0.3229	0.3178	0.2475	0.2511	0.2917	0.4338
	MM-DTC	T1-FLAIR	0.3712	0.3825	0.4319	0.4515	0.5016	0.5174
		FLAIR	0.2042	0.1827	0.2248	0.2671	0.3023	0.3380
	MM-DTML	T1-FLAIR	0.3926	0.3962	0.4306	0.4602	0.5080	0.5149
		FLAIR	0.2459	0.2209	0.2466	0.2465	0.2809	0.3296
	MM-SASS	T1-FLAIR	0.3868	0.4081	0.4232	0.4416	0.4939	0.5216
		FLAIR	0.2744	0.2387	0.1932	0.2407	0.2667	0.4287
	MM-UAMT	T1-FLAIR	0.3956	0.4195	0.4615	0.5733	0.5129	0.5891
		FLAIR	0.2601	0.2275	0.3048	0.3087	0.4047	0.4895
	MM-UMCT	T1-FLAIR	0.4054	0.4243	0.4525	0.5602	0.5040	0.5230
		FLAIR	0.3372	0.3365	0.2610	0.2662	0.3089	0.3083
	MM-SPCT	T1-FLAIR	0.4002	0.4220	0.4638	0.4860	0.5131	0.5355
		FLAIR	0.2595	0.2357	0.2444	0.2393	0.2934	0.3684
	DAFNet	T1-FLAIR	0.4100	0.5900	0.5050	0.5300	0.5260	0.5880
		FLAIR	0.3460	**0.4960**	0.3720	0.4710	0.4540	0.5660
	FewGAN	T1-FLAIR	0.3892	0.5889	0.4390	0.5242	0.4805	0.5356
		FLAIR	0.3173	0.4565	0.2207	0.1968	0.2575	0.2544
	Ours	FLAIR	**0.4337**	0.4833	**0.4854**	**0.5858**	**0.5302**	**0.6079**

Table 3
Ablation studies of our method on the Hecktor and BraTS datasets with 10% labeled data using DSC and Sens as evaluation metrics.

Methods		Hecktor				BraTS			
		CT		PET		T2		T1CE	
		DSC	Sens	DSC	Sens	DSC	Sens	DSC	Sens
sup		0.2866	0.3688	0.5080	0.5931	0.4368	0.4020	0.4920	0.4416
CML	sup+mse	0.2916	0.3805	0.5078	0.5700	0.4667	0.4538	0.5039	0.4625
	sup+NCE	0.3346	0.4155	0.5097	0.6211	0.5006	0.4599	0.6225	0.5759
	sup+mse+NCE*	0.2825	0.3158	0.5044	0.5505	0.5044	0.4910	0.6151	0.5795
	sup+mse+NCE	0.3550	0.4264	0.5210	0.6433	0.5162	0.4886	0.6272	0.5858
	sup+ASC	0.3667	0.3978	0.5831	0.6551	0.5235	0.4925	0.6433	0.5981
	sup+mse+ASC	0.3758	0.4014	0.6011	0.6634	0.5314	0.5120	0.6574	0.6312
PReL	PReL (Base)	0.3884	0.4721	0.6023	0.6777	0.5429	0.5385	0.6614	0.6534
	PReL (EMA)	0.3897	0.4786	0.6033	0.6796	0.5436	0.5398	0.6626	0.6559
	PReL (BMA)	**0.3942**	**0.4874**	**0.6072**	**0.6897**	**0.5612**	**0.6189**	**0.6656**	**0.6917**

pair of dual-modal images, the segmentation results generated by our work are more consistent to each other than those of the baselines. This is because our method utilize mutual learning to help different modalities learn from each other and make their predictions better and closer. Consequently, we conclude that information in multi-modal data is very important and should be used for semi-supervised medical image segmentation tasks.

4.4.3. Comparison with state-of-the-art multi-modal semi-supervised methods

Furthermore, to prove that the superior performance of our work is not solely coming from multi-modal information, but also due to its technical superiority, we further compare our method with the existing multi-modal semi-supervised segmentation solutions, DAFNet (Chartsias et al., 2020) and few-shot GAN (denoted FewGAN) (Mondal et al., 2018), and also with the multi-modality extended version of MT, ICT, DTC, DTML, SASS, UAMT, UMCT, and SPCT ("MM-" is added as a prefix for distinction). Specifically, the MM extension is achieved by using a dual-modal fusion model (U-Net network with a dual encoder and a shared decoder) and feeding two-modal images into the network simultaneously. According to the results in Table 2, we first find that, compared with the original single-modal models in Table 1, the segmentation performances of multi-modal models have been greatly enhanced. For example, for 10% labeled data, the DSC of MM-UAMT in Table 2 using bimodal CT-PET is 0.046 higher than that of UAMT in Table 1 using PET only; similarly, the DSC performance of MM-UMCT (resp., MM-SPCT) using bimodal T1-FLAIR is 0.0143 (resp., 0.0152) higher than that of UMCT (resp., SPCT) using FLAIR only. Therefore, this proves that the use of multi-modal data will improve the models' performances, so it is important to utilize multi-modal information in medical image segmentation tasks. Second, our proposed method greatly outperforms the existing two multi-modal semi-supervised methods (i.e., DAFNet and FewGAN) and the extended multi-modal semi-supervised works in nearly all cases on both datasets when all models are trained using two modalities and inferred with only one modality (the best results are bold); furthermore, even when DAFNet, FewGAN, and the extended multi-modal semi-supervised methods are inferred using two modalities, our method (using only one modality for inference) can still achieve competitive and sometimes even better performances (especially, in the T1-FLAIR case) than their dual-modal inference results (the best results are underlined). This is because the multi-modal semi-supervised baselines are usually highly coupled fusion networks, which thus need bi-modal information in the inference stage to ensure satisfactory results; consequently, they usually have a great performance degradation when there is only one modality for inference, due to the lack of information of the other modality. For example, MM-UAMT has a DSC of 0.7069 when performing T2-T1CE dual-modal inference, but its DSC is only 0.1066 when using only T1CE for inference. However, for our method, the learned segmentation models for the two modalities are quite independent, so it only needs one modality for inference, and its single modal inference results are close to or sometimes even exceed the dual-modal inference results of the multi-modal baselines. Therefore, our method is much easier to be used in clinical practices, because it only needs one modality of medical images to obtain satisfactory segmentation results, which greatly reduces the patients' time and money costs.

4.5. Ablation studies

To verify the effectiveness of our semi-supervised segmentation method, we conducted ablation experiments on both two datasets and show the results in Table 3, where CML represents Semi-CML, and PReL represents Semi-CML with PReL.

4.5.1. Ablation experiments for CML

We first verified the effectiveness of the MSE loss and the ASC loss for the mutual learning in the Semi-CML framework. First, we find from Table 3 that: (i) the performances of using both the fully-supervised and MSE losses (denoted sup+mse) as the final loss are better than those of using solely the fully-supervised loss (denoted sup); (ii) the performances of using the fully-supervised, MSE, and NCE losses (denoted sup+mse+NCE) are better than those of using the fully-supervised and NCE losses (denoted sup+NCE); (iii) the performances of using the fully-supervised, MSE, and ASC losses (denoted sup+mse+ASC) are better than those of using the fully-supervised and ASC losses (denoted sup+ASC). Since the segmentation performances are always improved by additionally introducing the MSE loss into the final loss, we believe that the MSE loss is effective and essential for our method to achieve the superior performances. The effectiveness of the MSE loss may be for the following reason: although multi-modal medical images have different intensity characteristics, for any given patient, they will still share the same segmentation mask as the ground truth in the multi-modal segmentation tasks, i.e., when the multi-modal images are used as the inputs of the multi-modal segmentation models, they are expected to obtain the same (or at least similar) segmentation results; consequently, using the MSE loss for mutual learning will help the segmentation subnetwork of each modality to learn complementary knowledge from the other modality and generate as consistent segmentation results as possible for the multi-modal images. Therefore, we conclude that MSE is a different type of consistency loss from NCE and ASC; so introducing it into the final loss will help the multi-modal model to achieve a more comprehensive multi-modal consistency supervision in the mutual learning. Second, we implement the NCE loss in two ways, i.e., with and without using the projection head to encode the prediction map into the embedding vector (denoted NCE* and NCE, respectively); by comparing sup+mse+NCE* with sup+mse+NCE, a consistent performance degradation is witnessed when the projection head is used, proving that the projection head is harmful for the contrastive loss in segmentation tasks with dense predictions; so the projection head is not used in the proposed ASC loss. Third, we also find that sup+mse+NCE and sup+NCE are generally better than those of sup+mse in all cases, which thus proves the effectiveness of adding the contrastive loss NCE and the necessity of also maximizing the differences between the negative samples. Finally, sup+ASC and sup+mse+ASC respectively achieve better performances than sup+NCE and sup+mse+NCE, mainly because ASC pays extra attention to the area context information of the segmentation map. This proves that the ASC loss is a better choice than the traditional contrastive loss (NCE) in the multi-modal semi-supervised segmentation task.

Table 4
The BIoU results of our method and the single-modal semi-supervised baselines on Hecktor and BraTS with 10% labeled data.

Methods	Hecktor		BraTS	
	CT BIoU	PET BIoU	T1 BIoU	FLAIR BIoU
Sup	0.1497	0.3215	0.0923	0.2144
MT	0.1725	0.3260	0.1358	0.2374
ICT	0.1553	0.3487	0.1513	0.2348
DTC	0.1678	0.3437	0.1277	0.2331
DTML	0.1631	**0.3669**	0.1330	0.2181
SASS	0.1644	0.3518	0.1299	0.2278
UAMT	0.1722	0.3506	0.1365	0.2517
UMCT	0.1855	0.3513	0.1396	0.2475
SPCT	0.1783	0.3634	0.1366	0.2456
Ours	**0.2073**	0.3620	**0.2030**	**0.2876**

4.5.2. Ablation experiments for PReL

We further verify the effectiveness of the proposed PReL strategy in improving segmentation performances and reducing performance gaps between modalities. We have the following observations in Table 3. First, using the PReL strategy can further improve the segmentation performances of both modalities on both datasets, proving the effectiveness of the proposed soft pseudo-label re-learning strategy. Specifically, when generating a teacher model, using the EMA method has better performances than directly using the high-performance model (Base); however, EMA's performances are still worse than those of BMA. This is because EMA may introduce poor model weights, while the BMA strategy selects the best model weights when updating the teacher model. By using the BMA re-learning scheme, the segmentation performances of the low-performance modality on both datasets enhance about 0.02 for DSC and 0.09 for Sens, and the performances of the high-performance modality are also improved on both datasets. Second, the segmentation gaps between different modalities have been greatly narrowed with the help of PReL: When PReL is not used, the gaps between CT and PET in terms of DSC and Sens in the best model (sup+mse+ASC) are 0.2253 and 0.262, respectively; and the gaps between T2 and T1CE in DSC and Sens are 0.126 and 0.1192. After applying PReL, the gaps between CT and PET in terms of DSC and Sens in our final proposed model have been reduced to 0.213 (gap narrowed by 5.4%) and 0.2023 (gap narrowed by 22.8%); similarly, the gaps between T2 and T1CE in DSC and Sens are reduced to 0.1044 (gap narrowed by 17.1%) and 0.0728 (gap narrowed by 38.9%). Therefore, we can conclude that although PReL cannot fully close the gap, it indeed greatly narrows the performance gap between two modalities.

4.6. Additional experiments

4.6.1. Effectiveness in boundary-based evaluation metric

Boundary IoU (BIoU) (Cheng et al., 2021) is a boundary-based image segmentation evaluation metric, which is more sensitive to the boundary errors of the target areas and does not overpunish the errors of smaller objects. The formal definition of BIoU is:

$$\mathrm{BIoU} = \frac{\left|\left(G_d \cap G\right) \cap \left(P_d \cap P\right)\right|}{\left|\left(G_d \cap G\right) \cup \left(P_d \cap P\right)\right|}, \tag{18}$$

where G represents the ground truth binary mask, P represents the prediction binary mask, G_d and P_d indicate the pixel set of the boundary region of the binary mask, and d is the pixel width of the boundary region (Cheng et al., 2021). In our experiment, d is set to 4 for Hecktor and set to 5 for BraTS.

To show the superior segmentation performances of our proposed method more comprehensively, we additionally compare the BIoU-based segmentation results of our method with those of all the state-of-the-art single-modal semi-supervised methods in Table 4, and with those of the state-of-the-art multi-modal semi-supervised methods,

Table 5
The BIoU results of our method and the multi-modal semi-supervised baselines on Hecktor and BraTS with 10% labeled data, where bold (resp., underlined) values are the best results of our method and the baselines using single (resp., dual) modality inference.

Methods	Hecktor		BraTS	
	Inference modality	BIoU	Inference modality	BIoU
MM-MT	CT-PET	0.3975	T1-FLAIR	0.1388
	PET	0.3165	FLAIR	0.1494
MM-ICT	CT-PET	0.3612	T1-FLAIR	0.1766
	PET	0.0500	FLAIR	0.1588
MM-DTC	CT-PET	0.3997	T1-FLAIR	0.1546
	PET	0.2986	FLAIR	0.1504
MM-DTML	CT-PET	0.3761	T1-FLAIR	0.1692
	PET	0.1358	FLAIR	0.1545
MM-SASS	CT-PET	0.3775	T1-FLAIR	0.1635
	PET	0.1038	FLAIR	0.1484
MM-UAMT	CT-PET	0.3939	T1-FLAIR	0.1707
	PET	0.3165	FLAIR	0.1922
MM-UMCT	CT-PET	0.3999	T1-FLAIR	0.2701
	PET	0.1768	FLAIR	0.1714
MM-SPCT	CT-PET	0.3904	T1-FLAIR	0.2719
	PET	0.1992	FLAIR	0.1599
DAFNet	CT-PET	0.3780	T1-FLAIR	0.2220
	PET	0.2040	FLAIR	0.1880
FewGAN	CT-PET	0.3504	T1-FLAIR	0.2685
	PET	0.2337	FLAIR	0.1234
Ours	PET	**0.3620**	FLAIR	**0.2876**

Table 6
The DSC results of our method using different weight ratios for two modalities on Hecktor and BraTS with 10% labeled data.

R	Hecktor			BraTS		
	CT	PET	Avg.	T2	T1CE	Avg.
0.1	0.3735	0.5640	0.4688	0.5062	0.6588	0.5825
0.2	0.3896	0.5820	0.4858	0.5476	0.6715	0.6096
0.3	0.3978	0.5856	0.4917	0.5566	0.6618	0.6092
0.4	0.3843	0.5966	0.4905	0.5551	0.6654	0.6103
0.5	**0.3942**	**0.6072**	**0.5007**	**0.5612**	**0.6656**	**0.6135**
0.6	0.3895	0.5834	0.4864	0.5538	0.6698	0.6118
0.7	0.3507	0.5666	0.4586	0.5476	0.6675	0.6075
0.8	0.3534	0.5647	0.4591	0.5410	0.6705	0.6058
0.9	0.3717	0.5766	0.4742	0.5353	0.6411	0.5882

DAFNet and FewGAN, and the corresponding extended multi-modal semi-supervised methods in Table 5. It can be seen from the tables that the BIoU-based relative segmentation performances of our methods and the baselines are very similar to their DSC-based and Sens-based relative segmentation performances as shown in Tables 1 and 2. Consequently, we have the following findings: (i) our method is generally better than all single-modal semi-supervised baselines in BIoU; (ii) our method consistently outperforms the multi-modal semi-supervised baselines when these models are trained using two modalities but inferred with only one modality; and (iii) even when the multi-modal semi-supervised models use two modalities for inference, the performances of our method (using only one modality for inference) are still competitive and sometimes even better than those results. These findings thus prove our conclusion again: our method greatly outperforms the state-of-the-art semi-supervised segmentation methods and is much easier to be used in clinical practices.

4.6.2. Analysis for balancing the learning of two modalities in CML

In medical imaging, data in different modalities are usually very different due to the usage of different imaging equipments or methods, making images of different modalities contain different effective information. For multi-modal methods, it is also interesting to explore if the information contained in one modality is more informative than that of the other. Therefore, additional experiments are conducted by assigning different weight ratios to two supervision losses in Eq. (4). Specifically, we set the weight ratio of the supervised loss of the low-performance modality (i.e., CT in Hecktor, T2 in BraTS) to *R*, and the weight ratio of the other modality to $1 - R$. The experimental results are shown in Table 6, where we report not only the DSC results for the corresponding modalities but also the average DSC (denoted Avg.) among two modalities to show the overall performances of the multi-modal model.

As shown in Table 6, our method obtains relative better performances when the weights of two modalities are close (e.g., 0.4, 0.5, and 0.6); generally, the closer the weights the better the results, and when the weights are the same ($R = 0.5$), our model obtains the best performances on both datasets. This observation asserts that the information contained in both modalities are equally important for learning the multi-modal model, because their information is usually complementary. Therefore, in this work, we treat the losses of both modalities with the same importance in Eq. (4).

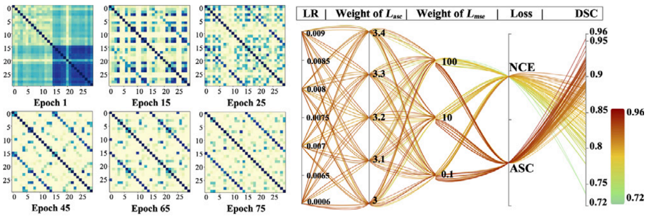
Table 7
The DSC results of using PReL and two adversarial learning based domain adaptation methods to narrow the performance gaps of two modalities on Hecktor and BraTS with 10% labeled data.

Methods	Hecktor		BraTS	
	CT	PET	T2	T1CE
CML	0.3754	0.5998	0.5243	0.6631
CML+DA (Adv)	0.3824	0.5998	0.5374	0.6631
CML+DA (AdvW)	0.3832	0.5998	0.5445	0.6631
CML+PReL	**0.3942**	**0.6072**	**0.5612**	**0.6656**

4.6.3. Effect of narrowing performance gaps using PReL and Domain Adaptation (DA)

Domain adaptation (DA) is an existing method that can be used to narrow the performance gaps between different modalities. In order to show that the proposed PReL is a better solution to narrow the performance gaps, additional experiments that use adversarial-learning-based domain adaptation solutions for performance gap narrowing are conducted, where the high-performance modality is treated as the source domain, the low-performance modality is treated as the target domain, their segmentation networks are used as generators, and generative adversarial learning is conducted to make the distributions of the target domain as close as possible to those of the source domain to narrow the performance gaps. Specifically, two adversarial-learning-based domain adaptation solutions are considered: (i) The first solution is DCGAN-based (Radford et al., 2015) (denoted adv), which uses the predicted output maps in the generators as the inputs of the discriminator to perform domain adaptation on the predicted segmentation maps. (ii) The second approach is similar to Dou et al. (2018), which uses the fused feature maps of different layers in the generators as inputs to the discriminator and use WGAN (Arjovsky et al., 2017) for adversarial learning. For a fair comparison, the starting epoch of domain adaptation keeps the same as that of PReL in all experiments.

The results in Table 7 show that two domain adaptation methods both effectively improve the DSC performances of low-performance modalities, i.e., CT and T2, which proves the effectiveness of domain adaptation methods in narrowing performance gaps. However, their resulting performances are worse than those of PReL in all modalities on both datasets, this may be because domain adaptation aims to narrow the distribution differences to indirectly narrow the performance gaps, making them not as effective in narrowing the performance gaps as PReL, which directly uses the segmentation results of the BMA teacher model for supervised re-learning. We also notice that the results of the high-performance modalities cannot be further improved using domain adaptation, because they are used as the target domains, and



(a) Dice Similarity matrix (b) The comparison for ASC and NCE loss

Fig. 5. Visualization of our proposed ASC loss on the Hecktor dataset. (a) Dice similarity matrix generated by the ASC loss; (b) Comparison between the ASC and NCE loss under different hyperparameters.

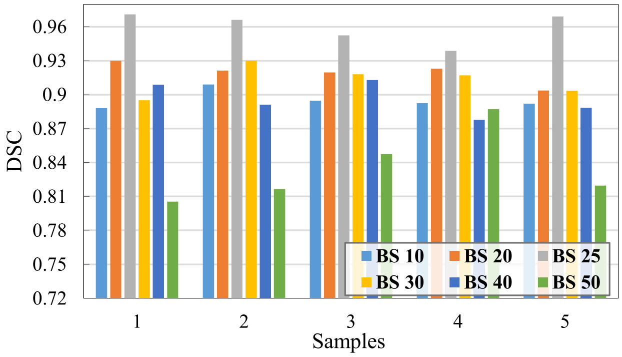
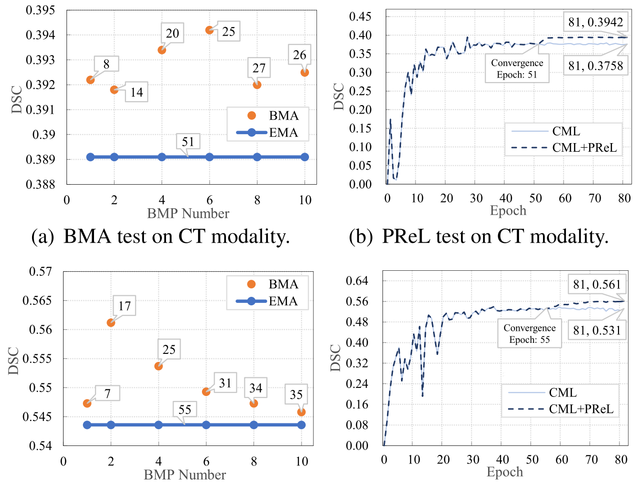


Fig. 6. Comparison of different batch sizes of unlabeled data (i.e., K in Eqs. (10) and (15)) in the ASC loss on Hecktor with 10% labeled data. The DSC is the sum of the results for two modalities.

their models will not be updated anymore; however, the results of the high-performance and low-performance modalities are both improved by PReL, because pseudo-label re-learning is conducted for both modalities. Finally, the adversarial-learning-based domain adaptation solutions needs to train additional discriminators, and the training is usually unstable. Consequently, all these findings conclude that PReL is a better choice for performance gap narrowing than domain adaptation solutions.

4.6.4. *Analysis for ASC loss*

We visualize the Dice similarity matrix in the ASC loss of the same mini-batch samples for unlabeled data in different epochs for 10% labeled data on Hecktor, as shown in Fig. 5(a). It shows that as the training epoch increases, the positive samples become more and more similar (darker in the 15th diagonal above and below the main diagonal position), while the negative samples can be gradually distinguished (lighter in other locations). This fully proves that our proposed area similarity contrastive loss plays a role in cross-modal knowledge mutual learning. Meanwhile, we conduct parallel experiments on the NCE loss and ASC loss under multiple combinations of hyperparameters, and show the sum of DSC of the two modalities. It shows that the DSC of the ASC loss is generally higher than the NCE loss. This is because NCE is based on the cosine similarity, which cannot utilize the valuable area context information of images; since the area context information is usually vital for medical image segmentation tasks, the segmentation performance improvements using NCE are generally limited. However, the ASC loss can resolve this problem, where Dice similarity instead of cosine similarity is used to ensure the model can take into account the area context information in contrastive learning. Therefore, we can conclude that Dice similarity, as a similarity measurement function in contrastive loss, is very beneficial to the semi-supervised segmentation task.



(a) BMA test on CT modality. (b) PReL test on CT modality.

(c) BMA test on T2 modality. (d) PReL test on T2 modality.

Fig. 7. Analysis of the PReL strategy under 10% labeled data. (a) (c): Comparison of BMA teacher with different BMP numbers and EMA teacher. (b) (d): Visualization of the DSC curve after using PReL.

4.6.5. *Analysis for batch size of unlabeled data in ASC loss*

In the contrastive self-supervised method, it has been proved that a large batch size of unlabeled data can bring more negative samples to the contrastive loss, which may help to generate better pre-training weights and improve the performance of downstream tasks (Chen et al., 2020; He et al., 2020; Misra and Maaten, 2020). In order to analyze the impact of batch sizes of unlabeled data (i.e., K in Eqs. (10) and (15)) on our proposed contrastive loss for multi-modal semi-supervised medical image segmentation, we compare the segmentation results of different batch sizes. We sample the experimental results under five different sets of hyperparameter combinations (including initial learning rate, the weight of the MSE consistency loss, and the weight of the ASC loss), and show the experimental results in Fig. 6. It can be seen from the figure that a large batch size of unlabeled data does not bring a better performance; actually, there exists a threshold for the optimal value of the batch size, i.e., too small or too large batch sizes cause performance degradation, and the best value is obtained in a medium range (at around 25). This finding is in line with the hypothesis in the methodology: the ASC loss requires a relative small mini-batch to construct positive and negative sample pairs to prevent the possibility of treating adjacent slices of the same patient as negative samples.

4.6.6. *Analysis for the BMA teacher re-learning strategy*

To give more details on the effect of BMA, we show the results of the BMA teacher model under different BMP numbers for Hecktor and BraTS in Figs. 7(a) and 7(c). We can see that the BMA update strategy can reduce the number of updates of the teacher model (the box in the figure indicates the number of updates for the teacher model) and achieve better results than the EMA update strategy under different numbers of BMP. This proves that BMA can automatically select the best model weights to update the teacher model based on the model performance in each epoch. Besides, we show the DSC metric curves of low-performance modalities CT and T2 in Figs. 7(b) and 7(d), respectively. The results show that, even after reaching the convergence epoch, the DSC of the low-performance model can be further significantly improved after using PReL with a BMA teacher model.

5. Discussion and future work

5.1. *Social impact for proposed algorithm*

The proposed model can be widely used in a lot of clinical scenarios, where the work of segmenting medical images is needed to effectively reduce the workload of doctors and improve the efficiency and accuracy of medical image segmentation. We take radiotherapy for cancer as an example, where doctors need to accurately delineate the outline of the tumor area on the patient's 3D CT or PET images as the radiotherapy target area. However, each 3D CT and PET is composed of hundreds of slices, and will take an experienced doctor several hours to annotate them one by one. Moreover, since the edge of the tumor is uneven and very difficult to delineate, to ensure the accuracy and comprehensiveness of labeling, it is usually necessary for multiple doctors to label the same image independently, and then gather them together as the final results. Consequently, the whole image segmentation process is very time-consuming and laborious; this not only greatly consumes the medical social resources (e.g., the time of experienced doctors), but may also bring long waiting times for the patient and delay the treatment. By applying our proposed automatic segmentation solution in such clinical practices, the model can generate the draft segmentation results automatically in seconds, which can then be sent to experienced doctors for fine-tuning. This thus greatly reduces the workload of doctors, and saves both time and money for patients.

More importantly, differently from the fully-supervised segmentation solutions that require a huge number of annotated data for training, our semi-supervised multi-modal segmentation solution can achieve an accurate segmentation using only a small amount of labeled training data. This thus greatly reduces the application requirements and enhances the deployment efficiency of automatic medical image segmentation systems in clinical practices. In addition, compared to the existing multi-modal semi-supervised segmentation solution, which requires the data of two modalities for inference, our proposed solution can use only one modality to achieve accurate inference. Consequently, this further enhances the segmentation model's usability and reduces the time and examination costs of patients in some clinical scenarios.

5.2. Limitations and future work

Due to the difference between the modalities, different image modalities may cause large accuracy differences under the same training settings. The fundamental reason may be that there is a large domain deviation between the two modalities, leading to a large difference in the distribution expressed by them in the potential space. Our proposed method has alleviated this problem to a large extent. That is, our method can greatly improve the accuracy of the low-performance modality, which makes the prediction results of the two modalities closer. However, we cannot make the results of the two modalities close to similar accuracy for the time being. Although this may be difficult, we think it is possible to achieve this by introducing new methods in the future. Therefore, a potential future work is to further explore how to make the prediction results of the modality that is more difficult to train closer to the one that is easier to train using mutual learning between the two modalities. In addition, the two networks, after learning from each other, can be fused in some new ways so that a higher precision can be obtained when the two modalities are input simultaneously.

In our task setting, images of different modalities need to be registered before they can be directly used in our method. Indeed, there exists some multi-modal medical images that cannot be easily registered in clinical practices. However, there are still many combinations of multi-modal medical images that can be registered using the existing regulation solutions, e.g., MR to CT registration (Mohammed and Hassan, 2016; Roy et al., 2014), registration of fluoroscopic X-ray to CT (Livyatan et al., 2003), PET to MRI registration (Shan et al., 2011), and preoperative magnetic resonance (MR) to intraoperative ultrasound registration (Machado et al., 2019). Therefore, even if not all, our proposed semi-supervised multi-modal segmentation solution can be applied for many clinical segmentation tasks by registering the corresponding multi-modal medical images before using them as the inputs. More importantly, after applying registration, we can use the same segmentation masks for both modalities, which thus reduces half of the annotation time cost.

Although our experiments are conducted on multi-modal medical data, we do not think this kind of learning schema is only applicable to medical imaging. We believe that it can also be used to segment general multi-modal images in daily life with proper registration, such as the multi-domain or multi-modal dataset mentioned in Cao et al. (2021), Martin et al. (2019), Sun et al. (2019) and Vu et al. (2019). Whether they are classification, detection, or segmentation, it is possible to improve the results of one of these data types using our proposed mutual learning method.

Our semi-supervised learning method mainly uses the mutual learning of two modalities to improve the accuracy of each modality, which is fundamentally different from other semi-supervised learning strategies. This makes it possible to add other semi-supervised strategies to our approach. Because almost all semi-supervised learning strategies are designed using a type of data, these designs can be integrated with our scheme.

6. Conclusion

In this paper, we proposed a multi-modal semi-supervised segmentation framework named Semi-CML with PReL. This architecture can achieve accurate medical image segmentation by using unlabeled multi-modal data for mutual supervised learning. Specifically, with the help of the area-similarity contrastive loss, one modality model can learn the complementary information from another modality, which simultaneously can improve the segmentation performances of all modalities. In addition, we designed a soft pseudo-label re-learning scheme based on the BMA teacher model to further improve the segmentation performances of the low-performance modality. We have conducted numerous experiments on multiple datasets and the results showed that the performances of our proposed method (using only a small portion of labeled data) are close to or sometimes even better than those of the fully-supervised method with 100% labeled data, and our proposed method also generally outperforms the state-of-the-art semi-supervised solutions. In addition, the inference of our work can be performed when only one modality data is available, and the corresponding segmentation results approach and even exceed those of the state-of-the-art semi-supervised multi-modal models that use multi-modal data for inference. Therefore, our method is much easier to be used in clinical practices and will greatly reduce the time and money costs of patients.

CRediT authorship contribution statement

Shuo Zhang: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Jiaojiao Zhang:** Validation, Investigation, Writing – review & editing, Visualization. **Biao Tian:** Validation, Investigation, Writing – review & editing, Visualization. **Thomas Lukasiewicz:** Supervision, Conceptualization, Writing – review & editing. **Zhenghua Xu:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code will be released on the corresponding author's homepage https://zhenghuaxu.info/publications/.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grant 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the Natural Science Foundation of Tianjin City, China, under the grant 19JCQNJC00400, by the "100 Talents Plan" of Hebei Province, China, under the grant E2019050017, by the Key Research and Development Project of Hainan Province, China, under the grant ZDYF2022SHFZ015, and by the Yuanguang Scholar Fund of Hebei University of Technology, China. This work was also partially supported by the AXA Research Fund, France.

References

Andrearczyk, V., Oreiller, V., Vallières, M., Castelli, J., Elhalawani, H., Jreige, M., Boughdad, S., Prior, J.O., Depeursinge, A., 2020. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: Proceedings of the Medical Imaging with Deep Learning. pp. 33–43.

Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K., 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–8.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 214–223.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A.P., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In: MICCAI.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data 4, 1–13.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. 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.

Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T., 2010. Icoseg: Interactive co-segmentation with intelligent scribble guidance. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3169–3176.

Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T., 2011. Interactively co-segmentating topically related images with intelligent scribble guidance. Int. J. Comput. Vis. 93 (3), 273–292.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 810–818.

Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y., 2021. ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. arXiv preprint arXiv:2108.10528.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. In: Proceedings of the Advances in Neural Information Processing Systems.

Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2020. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE Trans. Med. Imaging.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607.

Cheng, B., Girshick, R.B., Doll'ar, P., Berg, A.C., Kirillov, A., 2021. Boundary IoU: Improving object-centric image segmentation evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15329–15337.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 554–565.

Daryanto, T., Arif, S., Yang, S., 2017. Survey: recent trends and techniques in image co-segmentation challenges, issues and its applications. Int. J. Comput. Sci. Softw. Eng. 6 (5), 99.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B., 2018. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imaging 38, 1116–1126.

Dong, X., Shen, J., Shao, L., Yang, M.-H., 2015. Interactive cosegmentation using global and local energy optimization. IEEE Trans. Image Process. 24 (11), 3966–3977.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.-A., 2018. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. arXiv preprint arXiv:1804.10916.

Du, C., Du, C., He, H., 2021. Multimodal deep generative adversarial models for scalable doubly semi-supervised learning. Inf. Fusion 68, 118–130.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the International Conference on Machine Learning. pp. 1651–1660.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677.

Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K.-S., Qin, J., 2020. Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 562–571.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P., 2020. Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 772–781.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., Yang, M.-H., 2018. Adversarial learning for semi-supervised semantic segmentation. ArXiv, abs/1802.07934.

Iwasawa, J., Hirano, Y., Sugawara, Y., 2020. Label-efficient multi-task segmentation using contrastive learning. arXiv preprint arXiv:2009.11160.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging. pp. 597–609.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 18661–18673.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, Vol. 25. pp. 1097–1105.

Kumar, A., Fulham, M., Feng, D., Kim, J., 2019. Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Trans. Med. Imaging 39, 204–217.

Laine, S., Aila, T., 2017. Temporal ensembling for semi-supervised learning. In: Proceedings of the International Conference on Learning Representations.

Lee, D.-H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Proceedings of the International Conference on Machine Learning Workshops.

Li, H., Meng, F., Wu, Q., Luo, B., 2014. Unsupervised multiclass region cosegmentation via ensemble clustering and energy minimization. IEEE Trans. Circuits Syst. Video Technol. 24, 789–801.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A., 2020c. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst..

Li, S., Zhang, C., He, X., 2020a. Shape-aware semi-supervised 3D semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 552–561.

Li, S., Zhang, C., He, X., 2020b. Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 552–561.

Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., Shen, D., 2021. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med. Image Anal. 69, 101953.

Livyatan, H., Yaniv, Z., Joskowicz, L., 2003. Gradient-based 2-D/3-D rigid registration of fluoroscopic X-ray to CT. IEEE Trans. Med. Imaging 22 (11), 1395–1406.

Luo, X., Chen, J., Song, T., Wang, G., 2021. Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Machado, I., Toews, M., George, E., Unadkat, P., Essayed, W., Luo, J., Teodoro, P., Carvalho, H., Martins, J., Golland, P., et al., 2019. Deformable MRI-ultrasound registration using correlation-based attribute matching for brain shift correction: Accuracy and generality in multi-site data. Neuroimage 202, 116094.

Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., Stiefelhagen, R., 2019. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2801–2810.

Meng, F., Li, H., Zhu, S., Luo, B., Huang, C., Zeng, B., Gabbouj, M., 2015. Constrained directed graph clustering and segmentation propagation for multiple foregrounds cosegmentation. IEEE Trans. Circuits Syst. Video Technol. 25 (11), 1735–1748.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34, 1993–2024.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the International Conference on 3D Vision. pp. 565–571.

Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717.

Mittal, S., Tatarchenko, M., Brox, T., 2019. Semi-supervised semantic segmentation with high-and low-level consistency. IEEE Trans. Pattern Anal. Mach. Intell..

Mo, S., Cai, M., Lin, L., Tong, R., Chen, Q., Wang, F., Hu, H., Iwamoto, Y., Han, X.-H., Chen, Y.-W., 2020. Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 429–438.

Mohammed, H.A., Hassan, M.A., 2016. The image registration techniques for medical imaging (MRI-CT). Amer. J. Biomed. Eng. 6 (2), 53–58.

Mondal, A.K., Dolz, J., Desrosiers, C., 2018. Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. ArXiv, abs/1810.12241.

Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C., 2020a. Deep co-training for semi-supervised image segmentation. Pattern Recognit. 107, 107269.

Peng, J., Pedersoli, M., Desrosiers, C., 2020b. Mutual information deep regularization for semi-supervised segmentation. In: Proceedings of the International Conference on Medical Imaging with Deep Learning. pp. 601–613.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Rother, C., Minka, T., Blake, A., Kolmogorov, V., 2006. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1, IEEE, pp. 993–1000.

Roy, S., Carass, A., Jog, A., Prince, J.L., Lee, J., 2014. MR to CT registration of brains using image synthesis. In: Proceedings of the Medical Imaging 2014: Image Processing, Vol. 9034. 903419.

Shan, Z.Y., Mateja, S.J., Reddick, W.E., Glass, J.O., Shulkin, B.L., 2011. Retrospective evaluation of PET-MRI registration algorithms. J. Digit. Imaging 24 (3), 485–493.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 596–608.

Souly, N., Spampinato, C., Shah, M., 2017. Semi supervised semantic segmentation using generative adversarial network. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 5689–5697.

Sun, X., Xu, Y., Cao, P., Kong, Y., Hu, L., Zhang, S., Wang, Y., 2020. TCGM: An information-theoretic framework for semi-supervised multi-modality learning. In: Proceedings of European Conference on Computer Vision. pp. 171–188.

Sun, Y., Zuo, W., Liu, M., 2019. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. IEEE Robot. Autom. Lett. 4 (3), 2576–2583.

Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., Lasser, T., 2022. FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Med. Image Anal. 76, 102307.

Tao, W., Li, K., Sun, K., 2015. SaCoseg: Object cosegmentation by shape conformability. IEEE Trans. Image Process. 24 (3), 943–955.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1195–1204.

Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M., 2020. On mutual information maximization for representation learning. In: Proceedings of the International Conference on Learning Representations.

Tseng, K.-L., Lin, Y.-L., Hsu, W., Huang, C.-Y., 2017. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6393–6400.

Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D., 2019. Interpolation consistency training for semi-supervised learning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.

Vicente, S., Rother, C., Kolmogorov, V., 2011. Object cosegmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2217–2224.

Vu, T.-H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526.

Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C., Desrosiers, C., 2021a. Self-paced and self-consistent co-training for semi-supervised image segmentation. Med. Image Anal. 73, 102146.

Wang, Y., Yoon, B.-J., Qian, X., 2016. Co-segmentation of multiple images through random walk on graphs. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1811–1815.

Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020. Double-uncertainty weighted method for semi-supervised learning. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 542–551.

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L., 2021b. Exploring cross-image pixel contrast for semantic segmentation. arXiv preprint arXiv:2101.11939.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med. Image Anal. 65, 101766.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., Le, Q.V., 2019. Unsupervised data augmentation for consistency training. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 6256–6268.

Yang, Y., Zhan, D.-C., Sheng, X.-R., Jiang, Y., 2018. Semi-supervised multi-modal learning with incomplete modalities. In: Proceedings of IJCAI. pp. 2998–3004.

You, C., Zhao, R., Staib, L., Duncan, J.S., 2021. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. arXiv preprint arXiv:2105.07059.

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

Yuan, Y., 2020. Automatic head and neck tumor segmentation in PET/CT with scale attention network. In: Proceedings of the 3D Head and Neck Tumor Segmentation in PET/CT Challenge. pp. 44–52.

Zeng, G., Lerch, T.D., Schmaranzer, F., Zheng, G., Burger, J., Gerber, K., Tannast, M., Siebenrock, K., Gerber, N., 2021. Semantic consistent unsupervised domain adaptation for cross-modality medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention.

Zhang, Y., Zhang, J., 2021. Dual-task mutual learning for semi-supervised medical image segmentation. arXiv preprint arXiv:2103.04708.

Zhao, J., Li, D., Xiao, X., Accorsi, F., Marshall, H., Cossetto, T., Kim, D., McCarthy, D., Dawson, C., Knezevic, S., et al., 2021. United adversarial learning for liver tumor segmentation and detection of multi-modality non-contrast MRI. Med. Image Anal. 73, 102154.

Zhou, T., Ruan, S., Canu, S., 2019a. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3, 100004.

Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E.K., Yuille, A.L., 2019b. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 121–140.

Zhu, Y., Zhang, Z., Wu, C., Zhang, Z.-L., He, T., Zhang, H., Manmatha, R., Li, M., Smola, A., 2020. Improving semantic segmentation via self-training. ArXiv, abs/2004.14960.

Zou, Y., Yu, Z., Kumar, B.V.K.V., Wang, J., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV.