Cross-model Mutual Learning for Exemplar-based Medical Image Segmentation

Qing En[*] **Yuhong Guo**[*†]
[*]School of Computer Science, Carleton University, Ottawa, Canada
[†]Canada CIFAR AI Chair, Amii, Canada
qingen@cunet.carleton.ca, yuhong.guo@carleton.ca

Abstract

Medical image segmentation typically demands extensive dense annotations for model training, which is both time-consuming and skill-intensive. To mitigate this burden, exemplar-based medical image segmentation methods have been introduced to achieve effective training with only one annotated image. In this paper, we introduce a novel Cross-model Mutual learning framework for Exemplar-based Medical image Segmentation (CMEMS), which leverages two models to mutually excavate implicit information from unlabeled data at multiple granularities. CMEMS can eliminate confirmation bias and enable collaborative training to learn complementary information by enforcing consistency at different granularities across models. Concretely, cross-model image perturbation based mutual learning is devised by using weakly perturbed images to generate high-confidence pseudo-labels, supervising predictions of strongly perturbed images across models. This approach enables joint pursuit of prediction consistency at the image granularity. Moreover, cross-model multi-level feature perturbation based mutual learning is designed by letting pseudo-labels supervise predictions from perturbed multi-level features with different resolutions, which can broaden the perturbation space and enhance the robustness of our framework. CMEMS is jointly trained using exemplar data, synthetic data, and unlabeled data in an end-to-end manner. Experimental results on two medical image datasets indicate that the proposed CMEMS outperforms the state-of-the-art segmentation methods with extremely limited supervision.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Medical image analysis has gained significant attention in clinical diagnostics and complementary medicine due to the rapid advancements in medical image technology (Duncan and Ayache, 2000; Anwar et al., 2018). In particular, medical image segmentation is a fundamental and challenging task that involves determining the category of each pixel in medical images (Sharma et al., 2010; Ronneberger et al., 2015). While existing fully supervised methods have produced satisfactory results, obtaining numerous fine-grained annotations is both time-consuming and labour-intensive, which is a major barrier to the advancement of medical image analysis. Several approaches have been proposed to alleviate this burden by using limited annotations to achieve complex tasks and overcome the limitations of fully supervised methods (Qian et al., 2019; Tang et al., 2018; Lin et al., 2016).

Although existing methods have made considerable progress, they still struggle to overcome the barriers presented by labeling. They typically require either class and location information of each image or a portion of fine annotations during the training and testing phases (Roy et al., 2020; Tang et al., 2021; Luo et al., 2021a; Seibold et al., 2022; Wu et al., 2021). Unlike the above methods, exemplar learning-based medical image segmentation can enjoy the ability to complete the segmentation task with only one expert-annotated image, owing to the unique properties of medical images (En and Guo, 2022). Additionally, this technique can utilize new unlabeled data to continuously improve the model's effectiveness. Consequently, we concentrate on segmenting medical images in the exemplar-learning scenario, in which the only annotated image

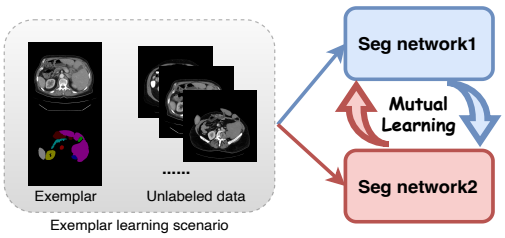

Figure 1: Illustration of the proposed idea. The proposed CMEMS leverages two mutual learning models to excavate implicit information from unlabeled data for exemplar-based medical image segmentation.

serves as an exemplar containing all organ categories present in the dataset.

Recently, an exemplar-learning based method called ELSNet (En and Guo, 2022) has been proposed by synthesizing annotated images and applying a contrastive approach to learn segmentation models from exemplars. However, the use of static pseudo-labels restricts it from taking full advantage of the knowledge of unlabelled data. Furthermore, it is inappropriate to directly apply existing semi-supervised methods to the ELSNet as they ignore the diverse granularity of consistency information and do not leverage the complementary knowledge offered by multiple models simultaneously, which can lead to a dilemma in the exemplar learning scenario.

The fundamental challenges of exemplar-learning based medical semantic segmentation can be attributed to the following two aspects: (1) With only one annotated exemplar image, it is difficult to train the model effectively when inaccurate pseudo-labels produced by unlabeled images hinder the training process. (2) The lack of guidance from complementary information, combined with insufficient label diversity and unsatisfactory quality of pseudo labels, also results in confirmation bias and aggravates the difficulty of the task. Humans have the ability to exhibit different patterns of behaviour, whether in the face of integral or separate features (Zhu et al., 2011). Inspired by the idea, we propose integrating cross-model learning at different granularities to obtain consistency and complementary information across various levels of detail.

In this paper, we propose a novel Cross-Model Mutual Learning framework for Exemplar-based Medical image Segmentation (CMEMS), as shown in Figure 1. The core of this framework is to leverage two mutual segmentation models that exploit implicit information from unlabeled data at various granularity levels to produce more precise pseudo-labels, thereby enhancing the performance of both models through mutual supervision. CMEMS can alleviate confirmation bias and enable the acquisition of complementary information by promoting consistency across various granularities of unlabeled images and facilitating collaborative training of multiple models (Arazo et al., 2020). Specifically, we devise a cross-model image perturbation based mutual learning mechanism that leverages high-confidence pseudo-labels obtained from weakly perturbed unlabeled images by one model to supervise the predictions of strongly perturbed unlabeled images generated by the other model. In this case, the two models can jointly pursue prediction consistency at the image granularity by computing the cross-model image perturbation loss. Moreover, we present a cross-model multi-level feature perturbation based mutual learning mechanism by letting pseudo-labels supervise predictions of perturbed multi-level features to broaden the perturbation space and strengthen the robustness of our framework. This is achieved by computing the cross-model multi-layer feature perturbation loss, which can enable the framework to maintain feature-level consistency across models. Finally, the supervised segmentation losses calculated from the exemplar and the generated synthetic dataset are combined with the two losses mentioned above to optimize the proposed CMEMS collaboratively. Extensive experimental results demonstrate the effectiveness of the proposed CMEMS framework. In summary, the key contributions of our paper are as follows:

- We propose a novel CMEMS framework for exemplar-based medical image segmentation by leveraging two mutual learning models to excavate implicit information from unlabeled data at different granularities.
- We devise a cross-model image perturbation based mutual learning mechanism and a cross-model multi-level feature perturbation based mutual learning mechanism by enforcing consistency across unlabeled images and features to alleviate confirmation bias and enable the acquisition of complementary information.
- Experimental results demonstrate that the proposed CMEMS framework achieves state-of-the-art performance on the Synapse and ACDC medical image datasets in exemplar learning scenarios.

2 RELATED WORKS

2.1 Medical Image Segmentation

The medical image segmentation task aims to identify the organ or lesion area of the input medical im-

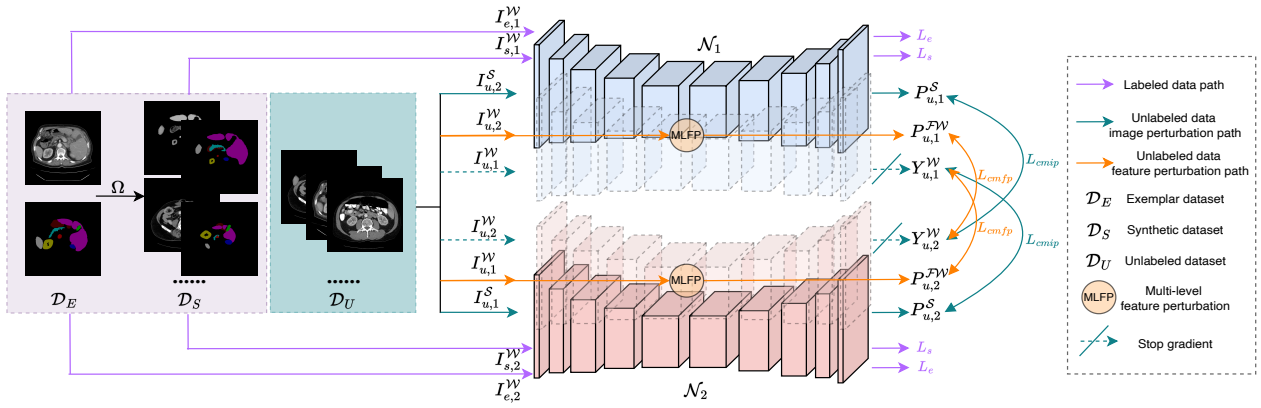

Figure 2: An overview of the proposed CMEMS. Firstly, a synthetic dataset $\mathcal{D}_S$ is generated from an exemplar dataset $\mathcal{D}_E$ using the Ω method. Both datasets are then fed into segmentation networks $\mathcal{N}_m(m \in \{1, 2\})$ to calculate L_e and L_s. Next, the unlabeled image I_u is fed into two segmentation networks using weak and strong perturbations, respectively, to calculate L_{cmip} by cross-model image perturbation based mutual learning and to calculate L_{cmfp} via cross-model multi-level feature perturbation based mutual learning. Finally, the two segmentation networks are optimized collaboratively by computing all loss functions.

age. CNN-based methods (Milletari et al., 2016; Ronneberger et al., 2015) and transformer-based methods (Chen et al., 2021a; Wang et al., 2022a; Li et al., 2022) have been widely used to solve this problem with promising results. UNet (Ronneberger et al., 2015) is the most widely used encoder-decoder structured CNN model that can efficiently solve medical image segmentation tasks. Its variants have produced better segmentation results (Zhang et al., 2019; Huang et al., 2020; Rahman and Marculescu, 2023). Besides, transformer-based methods have achieved notable success in medical image segmentation by capturing long-range dependencies (Cao et al., 2022; Wang et al., 2022a). However, the abovementioned methods depend on many labeled images, whereas we intend to address the challenging scenario where only one labeled image is available.

2.2 Semi-Supervised Semantic Segmentation

Semi-supervised semantic segmentation has been extensively investigated and gained success (Zou et al., 2021; Ouali et al., 2020; Chen et al., 2021b; Wang et al., 2022b; Yuan et al., 2023; Yang et al., 2023). Recent attempts have focused on consistency regularization (French et al., 2020; Chen et al., 2021b; Hu et al., 2021; Lai et al., 2021) and entropy minimization (Yuan et al., 2021; Yang et al., 2022; Guan et al., 2022). Some methods (Luo et al., 2022; Chen et al., 2021b; Zheng et al., 2022) use the same images for co-training or uncertainty consistency across models but overlook varying levels of perturbations. Instead, we aim to solve the problem of segmenting medical images based on exemplar learning in a unique, more challenging experimental setting than natural images.

2.3 Medical Image Segmentation with Limited Supervision

Many methods have been proposed to achieve medical image analysis using limited supervision. Some works utilized the mean-teacher model (Yu et al., 2019) and incorporated the self-ensembling framework (Cui et al., 2019; Li et al., 2020; Seibold et al., 2022; You et al., 2022) to explore the prediction information of the unlabeled images. Several uncertainty constraints and co-training methods (Luo et al., 2021b; Wu et al., 2022; Luo et al., 2022; Xia et al., 2020) have been attempted to enable the model to generate invariant results by minimizing the discrepancy of outputs. One step further, the first and so far only exemplar learning method for medical image segmentation, ELSNet (En and Guo, 2022), has been proposed, which uses one annotated image for medical image semantic segmentation. Unlike ELSNet's single-model approach, our framework uses two mutual learning models to guide the network away from incorrect directions. Besides, while ELSNet relies on static pseudo-labels, our framework dynamically generates them through image and multi-level feature perturbation, exploiting implicit information from unlabeled data at various granularity levels for more precise labeling.

3 METHOD

In this section, we present the proposed CMEMS framework for exemplar-based medical image segmentation. We first introduce the experimental setup and

architecture of the proposed CMEMS. Next, we briefly introduce exemplar-based data synthesis. We then present the cross-model mutual learning framework for exemplar-based medical image segmentation, including cross-model image perturbation based mutual learning and cross-model multi-level feature perturbation based mutual learning. Finally, we present the collaborative optimization procedure.

3.1 Overview

In the exemplar learning scenario, a single exemplar image is defined as an image containing one segmentation instance for each category present in the dataset. We aim to train two good segmentation models using a single labeled training image (i.e. exemplar) and T unlabeled training images, represented as $\mathcal{D}_E = (I_e, Y_e)$ and $\mathcal{D}_U = \{(I_u^t)\}_{t=1}^T$, respectively. Let $I \in \mathbb{R}^{1\times H\times W}$ denote a 2D input image and $Y_e \in \{0,1\}^{K\times H\times W}$ denote the corresponding segmentation labels, where K is the number of classes (i.e., categories of organs), and H and W represent the height and width of the input image, respectively.

The architecture of the proposed CMEMS is presented in Figure 2, comprising two segmentation networks with identical structures but distinct parameters. We initially generate a synthetic dataset $\mathcal{D}_S$ from the exemplar $\mathcal{D}_E$ to enhance the diversity of the training set. Next, given the unlabeled dataset $\mathcal{D}_U$, cross-model image perturbation learning is accomplished by using pseudo-labels produced from weakly perturbed images by one model to supervise predictions of strong perturbed unlabeled images by the other model, which can make two models jointly pursue prediction consistency at the image granularity. Furthermore, pseudo-labels are used to supervise the predictions obtained from perturbed multi-level features in cross-model multi-level feature perturbation learning. This broadens the perturbation space and enhances the robustness of the framework at the feature granularity. Finally, both segmentation networks are jointly optimized using the exemplar, synthetic, and unlabeled datasets. Each of the segmentation networks, denoted as $\mathcal{N}_m(m \in 1,2)$, comprises of an encoder $f_m : \mathbb{R}^{1\times H\times W} \to \mathbb{R}^{c\times h\times w}$ and a decoder $g_m : \mathbb{R}^{c\times h\times w} \to \mathbb{R}^{K\times H\times W}$, where c, h, and w denote the number of channels, height, and width of the encoded features, respectively. The proposed CMEMS framework is trained end-to-end.

3.2 Exemplar-based Data Synthesis

We first employ an efficient method to synthesize segmentation instances of each label category from an exemplar into various backgrounds to augment the diversity of annotated data (En and Guo, 2022). For an exemplar image I_e and its annotation Y_e, we use a series of geometric transformations $\mathcal{G}$ and intensity transformations $\mathcal{I}$ to generate a synthetic image I_s and its annotation Y_s through a copy-and-paste strategy as follows:

$$(I_s, Y_s) = \Omega(\mathcal{G}(\mathcal{I}(I_e)), \mathcal{G}(Y_e), \mathcal{G}(\mathcal{I}(I_b))). \quad (1)$$

where Ω denotes the operation of copying, transforming and pasting the exemplar organ onto the background image I_b and generating the corresponding annotation. Using this operation, we construct a synthetic dataset $\mathcal{D}_S = \{(I_s^b, Y_s^b)\}_{b=1}^B$, where $I_s^b \in \mathbb{R}^{1\times H\times W}$, $Y_s^b \in \{0,1\}^{K\times H\times W}$, and B is the number of synthetic images.

3.3 Cross-model Mutual Learning for Exemplar-based Medical Image Segmentation

Although the synthetic dataset can increase the diversity of annotated data, it still provides limited supervision information. It is desirable to utilize the abundant unlabeled images with predicted pseudo-labels to assist the learning of accurate segmentation models. Despite ELSNet (En and Guo, 2022) attempts to increase the discriminative ability of the segmentation network to obtain better pseudo-labels, the model is still negatively impacted by noise from static pseudo-labels and its own confirmation bias. Therefore, we propose to leverage two mutual learning segmentation networks to dynamically generate pseudo-labels through image and feature perturbation learning across models, aiming to learn complementary information and eliminate confirmation bias. This is realized through two mutual learning mechanisms elaborated below.

3.3.1 Cross-model image perturbation based mutual learning

Given a set of unlabeled images $\mathcal{D}_U = (I_u^t)_{t=1}^T$, we conduct cross-model mutual learning at the image granularity by deploying week and strong perturbations, defined as $\mathcal{W}$ and $\mathcal{S}$, respectively. Firstly, we apply the weak perturbation $\mathcal{W}$ to an unlabeled image I_u twice to generate two weakly augmented images $\{I_{u,m}^{\mathcal{W}} : m \in \{1,2\}\}$, which are then separately fed into the two segmentation networks $\mathcal{N}_m(m \in \{1,2\})$ to obtain probabilistic predictions $P_{u,m}^{\mathcal{W}} \in [0,1]^{K\times H\times W}$ $(m \in \{1,2\})$, such that:

$$P_{u,m}^{\mathcal{W}} = \text{softmax}(\mathcal{N}_m(I_{u,m}^{\mathcal{W}})), \quad (2)$$

where softmax represents a class-wise softmax function that is used to compute the probability of each pixel belonging to one of the K categories. Next, we

filter out the low-probability predictions and keep only the high-probability ones, ensuring that the obtained pseudo-labels $Y^{\mathcal{W}}_{u,m} \in \mathbb{R}^{1\times H\times W}$ consists only of confident predictions:

$$Y^{\mathcal{W}}_{u,m} = \text{argmax}(P^{\mathcal{W}}_{u,m}) \cdot \mathbb{I}\left[\sum\nolimits_{dim=0} \mathbb{I}\left[P^{\mathcal{W}}_{u,m} \geq \tau\right]\right], \tag{3}$$

where τ represents a predefined threshold used to filter out noisy pseudo-labels; $\sum_{dim=0}$ indicates summation over the first dimension of the 3D matrix, and $\mathbb{I}[\cdot]$ denotes the indicator function. Then, we apply the strong perturbation $\mathcal{S}$ to each weak perturbed unlabeled image $I^{\mathcal{W}}_{u,m}$ to generate $I^{\mathcal{S}}_{u,m}$. Each $I^{\mathcal{S}}_{u,\bar{m}}$ is cross-input into the segmentation network $\mathcal{N}_m$ to produce predictions $P^{\mathcal{S}}_{u,m} \in [0,1]^{K\times H\times W}$:

$$P^{\mathcal{S}}_{u,m} = \text{softmax}(\mathcal{N}_m(I^{\mathcal{S}}_{u,\bar{m}})), \tag{4}$$

where $\bar{m} \in \{1,2\}$ denotes the complement of $m \in \{1,2\}$ in the set $\{1,2\}$. Subsequently, the cross-model image perturbation consistency loss function is formulated over the predictions produced by the two segmentation networks for the perturbed unlabeled images as follows:

$$L_{cmip} = \sum_{m=1}^{M} L_{seg}(P^{\mathcal{S}}_{u,m}, Y^{\mathcal{W}}_{u,\bar{m}}) \tag{5}$$

where M denotes the number of segmentation networks such as $M = 2$, and $\mathcal{L}_{seg}$ denotes the segmentation loss function commonly used in medical image segmentation that includes both a cross-entropy loss function L_{ce} and a Dice loss function L_{dice}:

$$\mathcal{L}_{seg}(P,Y) = \frac{1}{2}L_{ce}(P,Y) + \frac{1}{2}L_{dice}(P,Y). \tag{6}$$

Here P represents the predictions and Y indicates the target labels. With the proposed consistency loss L_{cmip}, by alternately acting as teachers and students for each other, the two segmentation networks can robustly promote cross-model prediction consistency and mitigate confirmation bias, producing better pseudo-labels and enhancing segmentation performance (Arazo et al., 2020).

3.3.2 Cross-model multi-level feature perturbation based mutual learning

Although cross-model image perturbation learning can enhance a network's prediction robustness, it can only exploit perturbations that originate in the input space and may obstruct the exploration of consistency in various level features across models. Therefore, we propose a straightforward yet effective cross-model mutual learning through multi-level feature perturbations to further enhance consistency in the feature granularity of our segmentation models. Specifically, we

Algorithm 1 Training process of the proposed method

Input: $\mathcal{D}_E = (I_e, Y_e)$, $\mathcal{D}_U = \{(I^t_u)\}^T_{t=1}$
Output: Trained segmentation networks $\mathcal{N}_m (m \in \{1,2\})$
Generate the synthetic dataset $\mathcal{D}_S = \{(I^b_s, Y^b_s)\}^B_{b=1}$ by Eq. (1)
for $iteration = 1, MaxIter$ **do**
  Sample a batch: $(I_e, Y_e), (I_s, Y_s), I_u$
  $(I^{\mathcal{W}}_{e,1}, Y^{\mathcal{W}}_{e,1}), (I^{\mathcal{W}}_{e,2}, Y^{\mathcal{W}}_{e,2}) = \mathcal{W}(I_e, Y_e), \mathcal{W}(I_e, Y_e)$;
  $(I^{\mathcal{W}}_{s,1}, Y^{\mathcal{W}}_{s,1}), (I^{\mathcal{W}}_{s,2}, Y^{\mathcal{W}}_{s,2}) = \mathcal{W}(I_s, Y_s), \mathcal{W}(I_s, Y_s)$;
  $I^{\mathcal{W}}_{u,1}, I^{\mathcal{W}}_{u,2} = \mathcal{W}(I_u), \mathcal{W}(I_u)$;
  $I^{\mathcal{S}}_{u,1}, I^{\mathcal{S}}_{u,2} = \mathcal{S}(I^{\mathcal{W}}_{u,1}), \mathcal{S}(I^{\mathcal{W}}_{u,2})$;
  $Y^{\mathcal{W}}_{u,1} \xleftarrow{\mathcal{N}_1} I^{\mathcal{W}}_{u,1}, Y^{\mathcal{W}}_{u,2} \xleftarrow{\mathcal{N}_2} I^{\mathcal{W}}_{u,2}$ by Eq.(2), Eq.(3);
  $P^{\mathcal{S}}_{u,1} \xleftarrow{\mathcal{N}_1} I^{\mathcal{S}}_{u,2}, P^{\mathcal{S}}_{u,2} \xleftarrow{\mathcal{N}_2} I^{\mathcal{S}}_{u,1}$ by Eq.(4);
  Compute L_{cmip} by Eq.(5);
  $P^{\mathcal{FW}}_{u,1} \xleftarrow{\mathcal{N}_1} I^{\mathcal{W}}_{u,2}, P^{\mathcal{FW}}_{u,2} \xleftarrow{\mathcal{N}_2} I^{\mathcal{W}}_{u,1}$ by Eq.(7);
  Compute L_{cmfp} by Eq. (8);
  $P_{e,1} \xleftarrow{\mathcal{N}_1} I^{\mathcal{W}}_{e,1}, P_{e,2} \xleftarrow{\mathcal{N}_2} I^{\mathcal{W}}_{e,2}$;
  $P_{s,1} \xleftarrow{\mathcal{N}_1} I^{\mathcal{W}}_{s,1}, P_{s,2} \xleftarrow{\mathcal{N}_2} I^{\mathcal{W}}_{s,2}$;
  Compute L_e and L_s by Eq.(10);
  Compute L_{total} by Eq.(9);
  Update prameters of $\mathcal{N}_m (m \in \{1,2\})$;
end for

cross-input the weak perturbed unlabeled image $I^{\mathcal{W}}_{u,\bar{m}}$ into the encoder f_m to generate its multi-level features $X^{\mathcal{W}}_{u,\bar{m}} = \{x^{\mathcal{W},l}_{u,\bar{m}}\}^{L=5}_{l=1}$, where $L = 5$ denotes the total number of layers. We then apply a feature perturbation operation $\mathcal{F}$ to the multi-level features, and fed the perturbed features into the corresponding layers of the decoder g_m to obtain the predictions $P^{\mathcal{FW}}_{u,m}$ as follows:

$$P^{\mathcal{FW}}_{u,m} = \text{softmax}(g_m(\mathcal{F}(X^{\mathcal{W}}_{u,\bar{m}})). \tag{7}$$

The perturbation operation $\mathcal{F}$ is defined as randomly dropping out channels with a probability of 0.5. Subsequently, we formulate the cross-model multi-level feature perturbation consistency loss over unlabeled images as follows:

$$L_{cmfp} = \sum_{m=1}^{M} L_{seg}(P^{\mathcal{FW}}_{u,m}, Y^{\mathcal{W}}_{u,\bar{m}}). \tag{8}$$

The proposed L_{cmfp} maintains cross-model segmentation consistency with broadened perturbation space, strengthening the robustness of our framework. Moreover, the multi-level feature perturbation based learning is performed on features with different resolutions. It allows the framework to capture useful semantic information at various depths and complement our cross-model image perturbation based learning with various levels of detail, enhancing the segmentation models.

Table 1: Quantitative comparison results on the Synapse dataset. We report the class average DSC and HD95 results and the DSC results for all individual classes.

Method	HD95.Avg↓	DSC.Avg↑	Aor	Gal	Kid(L)	Kid(R)	Liv	Pan	Spl	Sto
UNet (Ronneberger et al., 2015)	132.42	0.160	0.026	0.167	0.177	0.154	0.649	0.015	0.059	0.033
MTUNet (Wang et al., 2022a)	154.60	0.112	0.066	0.108	0.155	0.053	0.352	0.008	0.046	0.102
MLDS (Reiß et al., 2021)	159.26	0.221	0.057	0.147	0.306	0.183	0.638	0.038	0.306	0.090
FixMatch (Sohn et al., 2020)	154.49	0.235	0.009	0.174	0.349	0.126	0.697	0.037	0.403	0.084
CPS (Chen et al., 2021b)	140.27	0.212	0.014	0.267	0.202	0.153	0.693	0.093	0.225	0.048
ELSNet (En and Guo, 2022)	109.70	0.315	0.319	**0.372**	0.381	0.219	0.784	0.067	0.276	0.104
CMEMS	**55.02**	**0.597**	**0.840**	0.230	**0.802**	**0.757**	**0.833**	**0.153**	**0.873**	**0.291**
FullySup	39.70	0.769	0.891	0.697	0.778	0.686	0.934	0.540	0.867	0.756

3.4 Collaborative Optimization Procedure

The proposed CMEMS is trained end-to-end. In each iteration, we sample a batch of images, which includes the annotated exemplar image (I_e, Y_e), an annotated synthetic image (I_s^b, Y_s^b), and an unlabeled image I_u^t, to update the segmentation networks by minimizing a joint loss function. Specifically, the overall loss function for training the segmentation networks $\mathcal{N}_m (m \in \{1, 2\})$ contains four terms:

$$L_{total} = L_e + L_s + \lambda_{cmip} L_{cmip} + \lambda_{cmfp} L_{cmfp}, \quad (9)$$

where λ_{cmip} and λ_{cmfp} are trade-off hyperparameters. L_e and L_s represent an exemplar segmentation loss and a synthetic segmentation loss calculated from the exemplar and synthetic images, respectively:

$$L_e = \sum_{m=1}^{M} \mathcal{L}_{seg}(P_{e,m}, Y_e), \quad L_s = \sum_{m=1}^{M} \mathcal{L}_{seg}(P_{s,m}, Y_s), \quad (10)$$

where $P_{e,m}$ and $P_{s,m}$ denote the predictions of $\mathcal{N}_m$ on the exemplar and synthetic images, respectively. L_{cmip} indicates the cross-model image perturbation consistency loss defined in Eq.(5), and L_{cmfp} is the cross-model multi-level feature perturbation consistency loss defined in Eq.(8). The training process of the proposed framework is described in Algorithm 1.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets and evaluation metrics The proposed CMEMS framework is evaluated on the Synapse multi-organ CT dataset and the Automated Cardiac Diagnosis Challenge (ACDC) dataset. The Synapse dataset comprises 30 abdominal CT cases containing 2,211 images with eight abdominal organs. We followed the experimental setup of Wang et al. (2022a) and Chen et al. (2021a) and used 18 cases for training and 12 cases for testing. The ACDC dataset contains 100 cardiac MRI cases with 1,300 images, and each image is labeled with three organ categories. We used 70 cases for training, 20 for validation, and 10 for testing. We used the DSC and the HD95 as evaluation metrics (Wang et al., 2022a; Chen et al., 2021a) to assess the performance of the proposed CMEMS framework.

Implementation details We adopt the UNet (Ronneberger et al., 2015) as our base network due to its effectiveness and efficiency in medical image segmentation. We randomly initialize the weights of the two segmentation networks, $\mathcal{N}_1$ and $\mathcal{N}_2$, and resize the input images to 224×224. We define the weak perturbation operation $\mathcal{W}$ as random rotation and flipping, and define the strong perturbation operation $\mathcal{S}$ as color jittering with an intensity factor α. We set the batch size to 12, set τ to 0.8, and set α to 1.0 and 0.2 on ACDC and Synapse, respectively. For the ACDC dataset, we set λ_{cmip} and λ_{cmfp} to 1 and 0.09, respectively, while for the Synapse dataset, we set them to 0.1 and 0.09, respectively. To optimize the proposed framework, we use the Adam optimizer with a weight decay of 0.0001 and a learning rate of 3e-4. We use the same settings as ELSNet (En and Guo, 2022) for the synthetic dataset, i.e., ten synthetic images for the same background on the Synapse dataset and fifteen on the ACDC dataset. For evaluation, we followed the experimental setup described in the literature (Chen et al., 2021a), where all 3D volumes are split into individual images for inference. Our experiments are conducted using NVIDIA GTX 2080Ti GPU.

4.2 Quantitative Evaluation Results

The proposed CMEMS is compared with six existing state-of-the-art methods on the Synapse and ACDC datasets. To ensure a fair comparison, we utilize the same exemplar, synthetic datasets, unlabeled datasets, backbone architecture and perturbation operations for all the comparison methods.

Table 2: Quantitative comparison results on the ACDC dataset. We report the class average results and the results for individual classes in terms of DSC and HD95.

Method	DSC.Avg↑	RV	Myo	LV	HD95.Avg↓	RV	Myo	LV
UNet (Ronneberger et al., 2015)	0.142	0.140	0.112	0.174	43.30	63.76	35.60	30.80
MT-UNet (Wang et al., 2022a)	0.142	0.119	0.126	0.182	74.20	83.91	61.48	77.22
MLDS (Reiß et al., 2021)	0.189	0.144	0.165	0.258	50.03	72.13	30.20	47.77
FixMatch (Sohn et al., 2020)	0.529	0.291	0.606	0.691	43.18	85.80	11.02	32.73
CPS (Chen et al., 2021b)	0.194	0.130	0.180	0.271	66.08	85.12	62.89	50.23
ELSNet (En and Guo, 2022)	0.410	0.293	0.374	0.563	26.64	47.63	16.58	15.73
CMEMS	**0.817**	**0.759**	**0.793**	**0.900**	**7.35**	**12.91**	**3.48**	**5.67**
FullySup	0.898	0.882	0.883	0.930	7.00	6.90	5.90	8.10

Table 3: Ablation study of the proposed components on the ACDC and Synapse datasets. We report the class average DSC and HD95 results. SD: using synthetic dataset. CM: cross-model mutual learning. IP: using image perturbations. FP: using multi-level feature perturbations. Data: datasets used in the training process.

Data					Synapse		ACDC	
	SD	CM	IP	FP	DSC.Avg↑	HD95.Avg↓	DSC.Avg↑	HD95.Avg↓
$\mathcal{D}_E$	-	-	-	-	0.160	132.42	0.142	43.30
$\mathcal{D}_E$+$\mathcal{D}_S$	√	-	-	-	0.256	122.49	0.359	15.16
$\mathcal{D}_E$+$\mathcal{D}_U$	-	√	-	-	0.212	140.27	0.194	66.08
$\mathcal{D}_E$+$\mathcal{D}_U$	-	-	√	-	0.235	154.49	0.529	43.18
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	-	-	0.378	113.07	0.516	10.43
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	-	√	-	0.429	106.76	0.670	20.67
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	√	-	0.523	101.93	0.807	8.07
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	√	√	**0.597**	**55.02**	0.817	7.35

Table 4: Ablation study on using different or same weak perturbations for the two segmentation networks on the Synapse and ACDC datasets.

Method	Synapse		ACDC	
	DSC↑	HD95↓	DSC↑	HD95↓
Same $I^{\mathcal{W}}_{e,m}$	0.550	74.25	0.805	6.84
Same $I^{\mathcal{W}}_{s,m}$	0.542	80.73	0.800	6.93
Same $I^{\mathcal{W}}_{u,m}$	0.583	71.36	0.812	6.94
Different	0.597	55.02	0.817	6.35

Comparison results on Synapse The test results of the proposed CMEMS and the other comparison methods on the Synapse dataset are reported in Table 1. The results show that CMEMS achieves superior performance, outperforming all the other compared methods in terms of both DSC and HD95 metrics. MLDS (Reiß et al., 2021), FixMatch (Sohn et al., 2020), and CPS (Chen et al., 2021b) are semi-supervised methods without effective integration of complementary information among models and perturbation information from multiple levels. The proposed CMEMS surpasses these methods in terms of class average DSC by 0.376, 0.362, and 0.385, respectively. Besides, the proposed CMEMS outperforms the second-best method, ELSNet (En and Guo, 2022), by 0.282 in terms of class average DSC, while reducing the class average HD95 value from 109.7 to 55.02. Moreover, Gal, Pan and Sto exhibit small organ areas and distinctive shape variations, resulting in general lower performance. Nevertheless, our method still works better than most existing methods in these categories. These experimental results illustrate the effectiveness of the proposed CMEMS.

Comparison results on ACDC The test results of the proposed CMEMS and the other comparison methods on the ACDC dataset are reported in Table 2. The results indicate that the proposed CMEMS produces substantial improvements over other methods in terms of both DSC and HD95 metrics, achieving the best class average DSC and HD95 scores of 0.817 and 7.35, respectively. FixMatch (Sohn et al., 2020) is the second-best-performing method, yet it performs poorly in terms of HD95. Although ELSNet (En and Guo, 2022) performs the second best in terms of HD95, it does not perform well in terms of DSC. Surprisingly, the performance of our proposed CMEMS is close to that of the fully supervised method. These findings again demonstrate the effectiveness of CMEMS for exemplar-based medical image segmentation.

4.3 Ablation Study

Impact of different components To investigate the contributions of each component to the performance of CMEMS, we conducted a set of experiments on Synapse and ACDC. The results are reported in Table 3. On Synapse, the experimental results (from the

Table 5: Ablation study of cross-model versus single-model multi-level feature perturbations on Synapse and ACDC in terms of DSC and HD95.

Method	Synapse		ACDC	
	DSC↑	HD95↓	DSC↑	HD95↓
Individual-model	0.587	66.51	0.812	6.53
Cross-model	0.597	55.02	0.817	6.35

second to the fourth row) show that using the synthetic dataset (SD), conducting cross-model mutual learning without perturbations (CM), and using image perturbations (IP) can improve the DSC result from the base 0.16 to 0.256, 0.212, and 0.235, respectively. Employing the synthetic dataset (SD) together with CM or IP can further improve the experimental results, while substantial improvements can be obtained by considering both synthetic and unlabeled data and combining both CM and IP, as shown in the second last row. By further taking the multi-level feature perturbations (FP) into consideration, the full model with all of the components yields the best results in terms of both DSC and HD95. Similar observations can be made on the ACDC dataset. In summary, the experiment results illustrated the effectiveness of each component.

Impact of using different weak perturbations We summarize the impact of using different weak image perturbations as inputs for the two segmentation networks in Table 4. The first three rows indicate that we applied a single weak perturbation to the exemplar image (I_e), synthetic image (I_s), and unlabeled image (I_u), respectively, before feeding them into the two segmentation networks, i.e., $I^{\mathcal{W}}_{d,1} = I^{\mathcal{W}}_{d,2}$ with $d \in \{e, s, u\}$. The results show that the best segmentation performances on both datasets are obtained when using two different weak perturbations on each input image to produce inputs for the two networks, as shown in the last row. In contrast, segmentation performance is degraded when the same weak perturbations are deployed on any type of images. We attribute this to the fact that applying different weak perturbations to each of the input images, together with our proposed cross-model perturbation based mutual learning, can significantly increase the diversity of training data, leading to more robust segmentation networks.

Impact of cross-model versus single-model multi-level feature perturbation We summarize the effect of cross-model versus single-model multi-level feature perturbations on the Synapse and ACDC datasets, as shown in Table 5. The "Individual-model" variant indicates that we input $I^{W}_{u,1}$ into $\mathcal{N}_1$ to obtain $P^{\mathcal{FW}}_{u,2}$, and input $I^{W}_{u,2}$ into $\mathcal{N}_2$ to obtain $P^{\mathcal{FW}}_{u,1}$. In this case, the two models can independently use their own pseudo-labels for the feature perturbed predictions through the loss term L_{cmfp}. The results indicate that using the cross-model feature perturbation based predictions is preferable than using the individual single-model based predictions, improving the DSC results on the Synapse and the ACDC datasets by 1% and 0.5%, respectively. These experimental results validate that the cross-model feature perturbation based learning is effective.

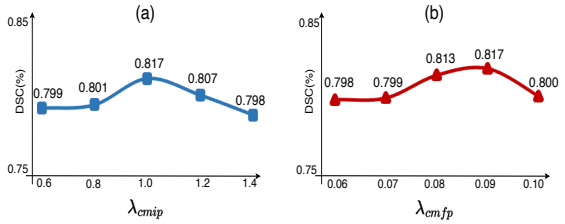

Figure 3: Impact of the weight (a) λ_{cmip} and (b) λ_{cmfp} on the performance on the ACDC dataset.

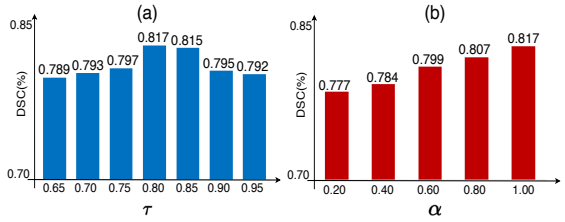

Figure 4: Impact of (a) the confidence threshold τ, and (b) the intensity factor α on the performance on the ACDC dataset.

4.4 Parameter Analysis

Impact of the weights of mutual learning loss functions We summarize the impact of the values of λ_{cmip} and λ_{cmfp} on the model performance on the ACDC dataset in Figure 3. The results in Figure 3 (a) show that the best performance is obtained when λ_{cmip} is set to 1, achieving a DSC value of 0.817, while decreasing or increasing λ_{cmip} leads to performance degradation. This suggests that it is important to exploit the unlabeled images through cross-model image perturbation based mutual learning, but overly emphasizing the unlabeled loss L_{cmip} is not desirable. In Figure 3 (b), it is evident that the performance improves gradually when λ_{cmfp} increases from 0.06 to 0.09, and then degrades when the λ_{cmfp} value further increases to 0.1. This indicates that a relatively small λ_{cmfp} value leads to desirable results, suggesting that L_{cmfp} should only be used as a slightly weighted auxiliary loss.

Impact of confidence threshold τ We conducted experiments on ACDC to investigate the impact of the

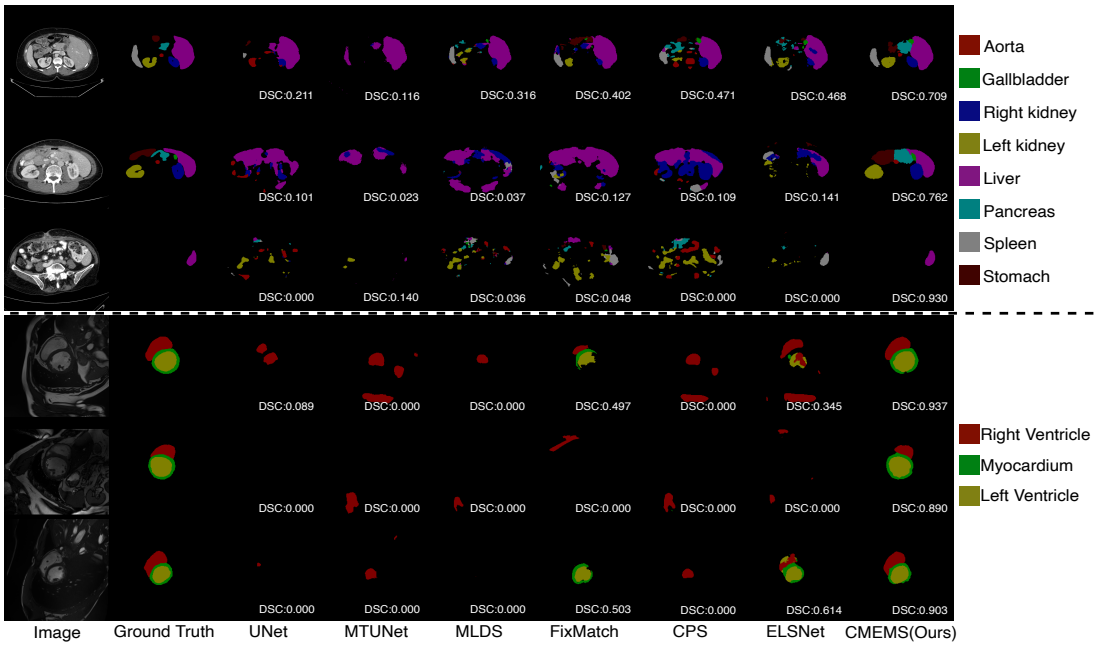

Figure 5: Visual examples of the segmentation results obtained by the proposed CMEMS framework and other state-of-the-art methods on the Synapse and ACDC datasets. The first two columns display the input images and the corresponding ground truth labels. The last column shows visualization of the segmentation results generated by CMEMS. The remaining columns show the results obtained by other methods.

confidence threshold τ on the model performance and report the results in Figure 4 (a). The parameter τ plays a crucial role in balancing the quality and quantity of the generated pseudo-labels. An optimal value of 0.8 for τ yields the best segmentation result of 0.817 in terms of DSC. Decreasing or increasing τ to 0.65 or 0.95 results in reduced experimental performance. This underscores the significance of both the quality and quantity of the pseudo-labels for learning effective segmentation models.

Impact of intensity factor α We summarize the impact of the intensity factor α for strong perturbations (i.e., colour jittering) over model performance on the ACDC dataset in Figure 4 (b). The α value controls the range of varying intensity of brightness, contrast and saturation. The experimental results indicate that a larger intensity factor leads to better results, with the best results obtained when α=1.0, reaching a DSC value of 0.817. This suggests that a wider range of colour jittering operations could help with cross-model mutual learning on unlabeled data.

4.5 Qualitative Evaluation Results

To demonstrate the effectiveness of the proposed method, we present the visualization comparisons with existing state-of-the-art methods in Figure 5. The results demonstrate that the proposed CMEMS outperforms all the other methods in terms of visual segmentation on both datasets. The background clutter and the low brightness phenomenon in medical images can greatly affect the segmentation results of the existing methods, making them prone to misclassifying background regions as foreground organs. Surprisingly, the proposed CMEMS can obtain segmentation results nearly as good as the ground-truth, benefiting from its ability to mitigate confirmation bias and learn complementary information.

5 CONCLUSION

In this paper, we proposed a novel framework, CMEMS, to achieve exemplar-based medical image segmentation by utilizing two mutual learning models to excavate implicit information from unlabeled data at multiple granularities. The CMEMS enables the cross-model image perturbation based mutual learning by using pseudo-labels generated by one model from weakly perturbed images to supervise predictions of the other model over strongly perturbed images. Moreover, the cross-model multi-level feature perturbation based mutual learning is designed to broaden the perturbation space and further enhance the robustness of the proposed framework. The experimental results demonstrate that the proposed CMEMS substantially outperforms the state-of-the-art methods.

References

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., and Khan, M. K. (2018). Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42:1–13.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks (IJCNN)*.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision (ECCV)*, pages 205–218. Springer.

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

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021a). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. (2021b). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., and Ye, C. (2019). Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *Information Processing in Medical Imaging (IPMI)*.

Duncan, J. S. and Ayache, N. (2000). Medical image analysis: Progress over two decades and the challenges ahead. *IEEE transactions on pattern analysis and machine intelligence*, 22(1):85–106.

En, Q. and Guo, Y. (2022). Exemplar learning for medical image segmentation. In *The British Machine Vision Conference (BMVC)*.

French, G., Laine, S., Aila, T., Mackiewicz, M., and Finlayson, G. (2020). Semi-supervised semantic segmentation needs strong, varied perturbations. In *The British Machine Vision Conference (BMVC)*.

Guan, D., Huang, J., Xiao, A., and Lu, S. (2022). Unbiased subclass regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., and Wang, L. (2021). Semi-supervised semantic segmentation via adaptive equalization learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., and Jia, J. (2021). Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

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

Li, Z., Li, D., Xu, C., Wang, W., Hong, Q., Li, Q., and Tian, J. (2022). Tfcns: A cnn-transformer hybrid network for medical image segmentation. In *International Conference on Artificial Neural Networks (ICANN)*.

Lin, D., Dai, J., Jia, J., He, K., and Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

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

Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., and Zhang, S. (2021b). Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D. N., and Zhang, S. (2022). Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*.

Ouali, Y., Hudelot, C., and Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Qian, R., Wei, Y., Shi, H., Li, J., Liu, J., and Huang, T. (2019). Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Rahman, M. M. and Marculescu, R. (2023). Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Reiß, S., Seibold, C., Freytag, A., Rodner, E., and Stiefelhagen, R. (2021). Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N., and Wachinger, C. (2020). 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587.

Seibold, C. M., Reiß, S., Kleesiek, J., and Stiefelhagen, R. (2022). Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Sharma, N., Aggarwal, L. M., et al. (2010). Automated medical image segmentation techniques. *Journal of medical physics*, 35(1):3.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tang, H., Liu, X., Sun, S., Yan, X., and Xie, X. (2021). Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., and Boykov, Y. (2018). On regularized losses for weakly-supervised cnn segmentation. In *European conference on computer vision (ECCV)*.

Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.-H., Chen, Y.-W., and Tong, R. (2022a). Mixed transformer u-net for medical image segmentation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., and Le, X. (2022b). Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Wu, H., Chen, G., Wen, Z., and Qin, J. (2021). Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., and Cai, J. (2022). Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530.

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

Yang, L., Qi, L., Feng, L., Zhang, W., and Shi, Y. (2023). Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Yang, L., Zhuo, W., Qi, L., Shi, Y., and Gao, Y. (2022). St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

You, C., Zhao, R., Staib, L. H., and Duncan, J. S. (2022). Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

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

Yuan, J., Ge, J., Wang, Z., and Liu, Y. (2023). Semi-supervised semantic segmentation with mutual knowledge distillation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Yuan, J., Liu, Y., Shen, C., Wang, Z., and Li, H. (2021). A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., and Shao, L. (2019). Et-net: A generic edge-attention guidance network for medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Zheng, X., Fu, C., Xie, H., Chen, J., Wang, X., and Sham, C.-W. (2022). Uncertainty-aware deep co-training for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 149:106051.

Zhu, X., Gibson, B., and Rogers, T. (2011). Co-training as a human collaboration policy. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zou, Y., Zhang, Z., Zhang, H., Li, C.-L., Bian, X., Huang, J.-B., and Pfister, T. (2021). Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations (ICLR)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Additional Ablation Study Results

We present the detailed ablation study results regarding the model components on the Synapse and ACDC datasets in Table 6 and Table 7, respectively. Table 6 demonstrates substantial enhancements in the DSC values for each category on the Synapse dataset with the incorporation of the various components. Notably, even for Aor, the organ with the smallest area in this dataset, the DSC value exhibits remarkable improvements, increasing from 0.026 to a final value of 0.840. A similar noteworthy enhancement is observed for the Spl category as well. These results highlight the effectiveness of CMEMS in significantly improving the segmentation outcomes, particularly for challenging categories. Furthermore, Table 7 reveals significant improvements in terms of both DSC and HD95 metrics for each category of the ACDC dataset.

Table 6: Ablation study of the proposed components on the Synapse dataset. We report the class average DSC and HD95 results and the DSC results for all individual classes. SD: using synthetic dataset. CM: cross-model mutual learning. IP: using image perturbations. FP: using multi-level feature perturbations. Data: datasets used in the training process.

Data	SD	CM	IP	FP	HD95.Avg↓	DSC.Avg↑	Aor	Gal	Kid(L)	Kid(R)	Liv	Pan	Spl	Sto
$\mathcal{D}_E$	-	-	-	-	132.42	0.160	0.026	0.167	0.177	0.154	0.649	0.015	0.059	0.033
$\mathcal{D}_E$+$\mathcal{D}_S$	√	-	-	-	122.49	0.256	0.112	0.273	0.321	0.129	0.792	0.008	0.360	0.051
$\mathcal{D}_E$+$\mathcal{D}_U$	-	√	-	-	140.27	0.212	0.014	0.267	0.202	0.153	0.693	0.093	0.225	0.048
$\mathcal{D}_E$+$\mathcal{D}_U$	-	-	√	-	154.49	0.235	0.009	0.174	0.349	0.126	0.697	0.037	0.403	0.084
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	-	-	113.07	0.378	0.658	0.328	0.510	0.079	0.689	0.141	0.526	0.094
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	-	√	-	106.76	0.429	0.700	0.256	0.499	0.185	0.826	0.143	0.533	0.289
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	√	-	101.93	0.523	0.783	**0.414**	0.585	0.468	0.828	0.027	0.695	**0.380**
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	√	√	**55.02**	**0.597**	**0.840**	0.230	**0.802**	**0.757**	**0.833**	**0.153**	**0.873**	0.291

Table 7: Ablation study of the proposed components on the ACDC dataset. We report the class average DSC and HD95 results and the DSC and HD95 results for all individual classes. CM: cross-model mutual learning. IP: using image perturbations. FP: using multi-level feature perturbations. Data: datasets used in the training process.

Data	SD	CM	IP	FP	DSC.Avg↑	RV	Myo	LV	HD95.Avg↓	RV	Myo	LV
$\mathcal{D}_E$	-	-	-	-	0.142	0.140	0.112	0.174	43.30	63.76	35.60	30.80
$\mathcal{D}_E$+$\mathcal{D}_S$	√	-	-	-	0.359	0.193	0.347	0.535	15.16	21.19	10.98	13.32
$\mathcal{D}_E$+$\mathcal{D}_U$	-	√	-	-	0.194	0.130	0.181	0.272	66.08	85.13	62.90	50.23
$\mathcal{D}_E$+$\mathcal{D}_U$	-	-	√	-	0.529	0.291	0.606	0.691	43.18	85.80	11.02	32.73
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	-	-	0.516	0.284	0.572	0.693	10.43	26.98	2.50	1.83
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	-	√	-	0.670	0.618	0.611	0.781	20.67	27.71	16.66	17.64
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	√	-	0.807	0.737	0.785	0.900	8.07	13.05	5.10	6.05
$\mathcal{D}_E$+$\mathcal{D}_S$+$\mathcal{D}_U$	√	√	√	√	**0.817**	**0.759**	**0.793**	**0.900**	**7.35**	**12.91**	**3.48**	**5.67**

B Additional Parameter Analysis Results

B.1 Impact of different exemplars

We summarize the results with different exemplars on the ACDC dataset in terms of DSC values in Table 8. We randomly selected three exemplar samples and compared our proposed CMEMS with UNet and ELSNet. The results indicate that the performance of our proposed CMEMS is more stable than that of the other methods when using different exemplars, and CMEMS substantially outperforms the other two compared methods regardless of the exemplar used. In particular, ELSNet uses the same data as our CMEMS, while CMEMS outperforms ELSNet by more than 0.4 in terms of class average DSC.

Table 8: Results with different exemplars on ACDC in terms of DSC.Avg.

	Exemplar1	Exemplar2	Exemplar3
UNet (Ronneberger et al., 2015)	0.160	0.114	0.145
ELSNet (En and Guo, 2022)	0.410	0.399	0.409
CMEMS(Ours)	0.817	0.809	0.812

Table 9: Results with different types of strong perturbations on ACDC in terms of DSC.Avg. CJ: color jittering. GB: Gaussian blur. AS: sharpness adjustment.

	CJ	CJ+GB	CJ+GB+AS
CMEMS(Ours)	0.817	0.818	0.818

B.2 Impact of different types of strong perturbations

We summarize the results with different types of strong perturbations on the ACDC dataset in terms of DSC values in Table 9. In addition to colour jittering, we tested Gaussian blur and sharpness adjustment. The experimental results show that the performance of our proposed method does not change much by adding more strong perturbation methods. It also validates that using colour jittering is sufficient.

C Additional Quantitative Evaluation Results

To further illustrate the efficacy of our proposed CMEMS, we conducted a comparison with a self-supervised learning approach for semi-supervised medical image segmentation (Chaitanya et al., 2020). The results are presented in Table 10, which show that our CMEMS method outperforms the compared approach when evaluated on the ACDC dataset. Our proposed CMEMS achieves superior results even when trained with only a single labeled image (i.e., the exemplar), surpassing the performance of the compared method trained with approximately 100 labeled images. Moreover, our method surpasses the performance of the compared method augmented with the Mixup technique, further underscoring the robustness and effectiveness of the proposed CMEMS in the context of medical image segmentation.

D Details of the Model Architecture

Our base network is the UNet, which contains an encoder and a decoder. The encoder contains 1*ConvBlock and 4*DownBlock. ConvBlock: Conv3*3 → BatchNorm → LeakyReLU → Dropout → Conv3*3 → BatchNorm → LeakyReLU. DownBlock: MaxPool2*2 → ConvBlock. The decoder contains 4*UpBlock and 1*Final Layer. UpBlock: Conv1*1 → Upsample → Concat → ConvBlock. Final Layer: Conv3*3. The encoder feature channels

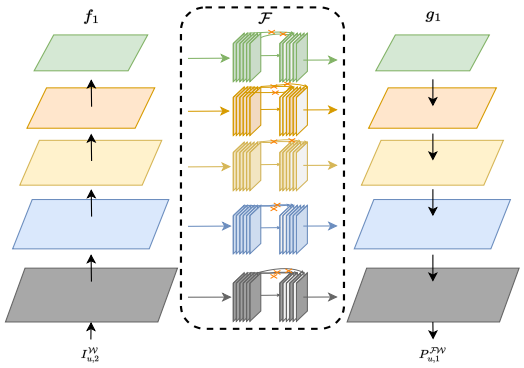

Figure 6: The architecture of multi-level feature perturbation.

Table 10: Additional quantitative comparison results on the ACDC dataset.

Method	(Chaitanya et al., 2020)	(Chaitanya et al., 2020)+Mixup	CMEMS(Ours)
Num of labels	~ 100	~ 100	1
DSC.Avg	0.725	0.757	**0.817**

and dropout rates are {16, 32, 64, 128, 256} and {0.05, 0.1, 0.2, 0.3, 0.5}, while the decoder has {256, 128, 64, 32, 16} channels and a dropout rate of 0.

In addition, we present the architecture of the multi-level feature perturbation operation on UNet in Figure 6, where the decoder takes the multi-level feature outputs from the encoder as inputs at the corresponding levels.