Computers in Biology and Medicine 176 (2024) 108570

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Semi-supervised multi-modal medical image segmentation with unified translation

Huajun Sun [a], Jia Wei [a,*], Wenguang Yuan [b], Rui Li [c]

[a] South China University of Technology, Guangzhou, 510006, China
[b] Huawei Cloud BU EI Innovation Laboratory, Dongguan, 523000, China
[c] Rochester Institute of Technology, Rochester, NY 14623, USA

ARTICLE INFO

Keywords:
Medical image segmentation
Semi-supervised
Multi-modal
Unified translation

ABSTRACT

The two major challenges to deep-learning-based medical image segmentation are multi-modality and a lack of expert annotations. Existing semi-supervised segmentation models can mitigate the problem of insufficient annotations by utilizing a small amount of labeled data. However, most of these models are limited to single-modal data and cannot exploit the complementary information from multi-modal medical images. A few semi-supervised multi-modal models have been proposed recently, but they have rigid structures and require additional training steps for each modality. In this work, we propose a novel flexible method, semi-supervised multi-modal medical image segmentation with unified translation (SMSUT), and a unique semi-supervised procedure that can leverage multi-modal information to improve the semi-supervised segmentation performance. Our architecture capitalizes on unified translation to extract complementary information from multi-modal data which compels the network to focus on the disparities and salient features among each modality. Furthermore, we impose constraints on the model at both pixel and feature levels, to cope with the lack of annotation information and the diverse representations within semi-supervised multi-modal data. We introduce a novel training procedure tailored for semi-supervised multi-modal medical image analysis, by integrating the concept of conditional translation. Our method has a remarkable ability for seamless adaptation to varying numbers of distinct modalities in the training data. Experiments show that our model exceeds the semi-supervised segmentation counterparts in the public datasets which proves our network's high-performance capabilities and the transferability of our proposed method. The code of our method will be openly available at https://github.com/Sue1347/SMSUT-MedicalImgSegmentation.

1. Introduction

The pixel-to-pixel level annotations of lesions and organs help surgeons make better diagnoses and give treatments precisely. However, in general, eliciting manual annotated medical images from radiology experts for every patient is laborious and time-consuming. An ideal solution is to enable machine learning to assist diagnosis with limited annotations [1]. Semi-supervised deep learning methods can leverage a large number of unlabeled data to improve the learning performance given limited labeled data, which has been a hot research topic in medical image analysis in recent years [2]. Common semi-supervised medical image segmentation based on deep learning can be roughly divided into two categories: iterative pseudo-label-based methods and consistency constraint training methods. The former methods generate pseudo-labels of partial unlabeled images from a supervised learning network trained with labeled data, and then the pseudo-labels can be used in the next round of training until eligible predictions are made for all the unlabeled data [3]. The latter strategies utilize some consistency constraints and regularize the model by minimizing the difference between these constrained elements on both labeled and unlabeled data. These constraints can be implemented through data perturbation, network dropout, self-ensemble models [4–6], parallel models [7,8], and GAN-based segmentation [9].

In radiology diagnosis, physicians examine various modalities of medical images for the diagnosis, such as computed tomography (CT), magnetic resonance imaging (MRI), magnetic resonance angiography (MRA), etc. [10]. Each modality of medical images contains anatomic structures of patients and, based on different imaging technologies, preserves different features that amplify different properties of organs and tissues [11]. CT images utilize X-ray absorption to analyze the structure of the body, providing accurate information about hard tissues, internal organs, and tumors. MR images can visualize normal and pathological

* Corresponding author.
E-mail addresses: 202120144084@mail.scut.edu.cn (H. Sun), csjwei@scut.edu.cn (J. Wei), yuanwenguang@huawei.com (W. Yuan), rxlics@rit.edu (R. Li).

https://doi.org/10.1016/j.compbiomed.2024.108570
Received 27 September 2023; Received in revised form 8 March 2024; Accepted 5 May 2024
Available online 8 May 2024

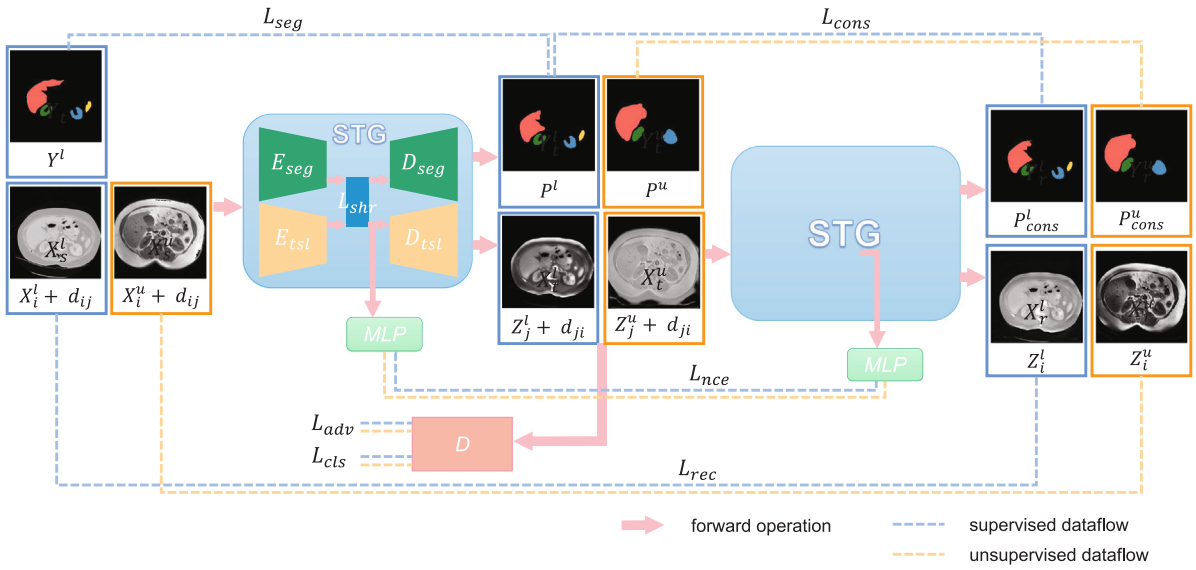

Fig. 1. The framework of SMSUT, where images and labels are selected from the abdominal dataset. The pink arrows show the training stream of our framework, the blue frames indicate the supervised training data, the yellow frames indicate the unsupervised training data, the blue dotted lines indicate the supervised losses and the yellow dotted lines indicate the unsupervised losses.

soft tissues with high resolution, aiding in the diagnosis of blood vessels, the brain, heart, and other internal organs. Integrating multiple data modalities improves diagnostic accuracy and enhances medical decision-making [12–14]. Generative adversarial network (GAN) [15] is a type of neural network model that has gained attention in medical image analysis due to its usefulness in counteracting domain shift, and effectiveness in generating new image samples. It is commonly used to perform medical images' modal translation, which can transform the source modal image into target modality without changing its anatomic content [16,17] in order to process the downstream tasks such as detection, and segmentation.

Multi-modal medical image segmentation methods typically necessitate a paired set of training images and achieve the fusion of multi-modal image information by either concatenating the input multi-modal images or fusing features in the latent space [18,19]. However, acquiring a paired set of medical images is challenging for medical image analysis due to various scanning protocols and unsuitable patient conditions. The transformer method which based on multi-head attention mechanism, has demonstrated state-of-the-art performance in various computer vision tasks [20,21]. Leveraging the attention mechanism, multi-modality information can be integrated without the need for multiple parallel structures and remains flexible to the number of modalities [19]. However, multi-head attention methods require a large dataset and significant computational resources, posing challenges in the context of medical image analysis.

Although semi-supervised learning methods effectively alleviate the insufficient label problem, most of them can only model one modality at one time. How to extend it to model multi-modal data simultaneously so as to leverage their complementary information remains an open question. The semi-supervised multi-modal medical image segmentation is a novel area of research, it is based on real scenes in which physicians tend to have limited time to annotate every patient's medical images for each modality [22,23]. Tackling the task can highly improve the efficiency of the doctors and possess a great prospect of application.

Most of the research uses the cross-modal strategy, in terms of solving the multi-modal medical images by tackling two modalities at one time. One of the strategies is utilizing translation [24–26] or disentangling [27,28] methods to enable cross-modal synthesis. These kinds of methods intend to simplify the semi-supervised or unsupervised cross-modal segmentation problems into semi-supervised single-modal segmentation problems. However, these methods do not combine the important complementary information extracted from multi-modal data to improve the semi-supervised segmentation task. Also, They are time-consuming and have complicated training details. Another common strategy is to combine translation and segmentation in order to force the model learning and sharing the feature level semantic information [29–31]. Zhu et al. [23] use the architecture that consists of two segmentation networks and two translation networks to generate pseudo-labels for unlabeled data to complete the cross-modal semi-supervised learning. These networks are commonly restricted to training each modality with one segmentation network and one translation network, which makes the model difficult to adapt to multiple modalities in reality. Maheshwari et al. [32] use cross-attention method to integrate multi-modal information bypass the translation procedure. However, it requires large databases for optimal performance. While proven effective in natural images, the approach faces challenges in multi-modal medical image segmentation due to the limited dataset.

To approach semi-supervised multi-modal medical image segmentation, we present a novel deep-learning framework, SMSUT, that combines a segmentation network with a unified multi-modal translation module and follows a unique semi-supervised procedure that can handle multiple modal information. The network is inspired by the inherent of the task, which requires the full utilization of multi-modal information to overcome the insufficient annotation problem. We integrate the segmentation network and translation network to exploit the complementary information within multi-modal data and the limited annotation information. Through the translation task, we force the network to capture the differences and significant features between each modality, and by composing consistency losses and contrastive learning loss, we achieve constraints on the model at the pixel level and feature level to cope with the lack of information and the multiple representations of semi-supervised multi-modal data. As a result, our model can integrate the complementary information from different modalities to propagate information from labeled data to unlabeled data. Also, we propose a unique procedure for training semi-supervised multi-modal medical images and using the idea of conditional translation. The method has a strong ability to be conveniently transferred to adapt to flexible numbers of different modalities of training data. Experiments show that our model exceeds the semi-supervised segmentation counterparts in both the public multi-modal healthy abdominal organ dataset and the multi-site prostate dataset. The abdominal dataset proves our network can efficiently solve the semi-supervised multi-modal segmentation task, and the prostate dataset proves our method's transferability, which can adapt to any number of modalities.

In summary, our contributions can be summarized as follows:

- Propose a semi-supervised multi-modal segmentation framework that can utilize the translation and segmentation tasks to exploit the complementary information of multi-modal images and assist in solving the insufficient annotation problem.
- Propose a semi-supervised multi-modal procedure with multiple consistency constraints, that can adapt to flexible numbers of modal information.
- Experiments show that our model exceeds the semi-supervised segmentation counterparts in the two public datasets, which proves our method's high performance and transferability.

2. Related works

2.1. Deep learning based medical image segmentation

Medical image segmentation is to separate a target area of interest from its background [33]. Ronneberger et al. [34] propose U-Net, a U-shaped architecture that consists of a contracting path to capture context and a symmetric expanding path that enables precise localization, for medical image segmentation. Considering the relatively simple structure of medical images and the importance of boundary information, U-Net corresponds the underlying texture information to high-level semantic information using skip connections. Through the encoder–decoder structure with skip connections, U-Net achieves promising segmentation results. Because of the simplicity and superior performance, various UNet-like methods are constantly emerging, such as swin-UNet [35], ST-UNet [36], PDAtt-UNet [37], and TDD-UNet [38]. Thus We adopt U-Net as the backbone structure of our framework SMSUT.

2.2. Deep learning based medical image translation

In medical image translation, Generative adversarial networks (GANs) [15] have been widely used in multi-modal image synthesis tasks, which can transform the source modal image into the target modality without changing its anatomic content [39,40]. Gulrajani et al. [41] propose gradient penalty loss to solve the consistently stable GAN training problem, which penalizes the norm of the gradient of the critic with respect to its input. To avoid image deformation generated from GAN, contrastive learning [42,43] is introduced to constrain the deformation at the feature level during training, which can provide another perspective of self-supervised learning. Mirza et al. [44] propose conditional GAN for multi-modal image generation with a single generator and a single discriminator, by generating descriptive tags which are not part of training labels. Jung et al. [45] integrating an additional module that ensures smooth and realistic transitions in 3D space to cGAN. Ziegler et al. [46] utilize cGAN to analyze the combination of image and tabular data for conditional 3D image synthesis.

We use the same conditional method for our translation sub-network, which enables our network to adapt to flexible numbers of modality data without extra training processes. The generator takes the source image and its corresponding modal information as inputs, and the discriminator has two outputs, one for distinguishing real images, and the other for resolving which modality the image comes from. Also, we apply contrastive learning to constrain the deformation of generative adversarial learning and provide guidance to the semi-supervised learning process.

2.3. Semi-supervised uni-modal medical image segmentation

Common semi-supervised medical image segmentation based on deep learning is a uni-modal segmentation task. The segmentation is based on the constraints such as data perturbation, network dropout, self-ensemble models [4,47,48], parallel models [7,8], and GAN-based segmentation [9]. Self-ensemble models and parallel models are extensively used in semi-supervised segmentation tasks [4,49–51]. Ouali et al. [52] propose cross-consistency training to achieve semi-supervised learning. The method enforces the consistency of predictions after different perturbations are applied to the outputs of the encoder and the consistency between the main decoder predictions and the auxiliary decoders. Basak et al. [53] propose a semi-supervised patch-based contrastive learning framework for medical image segmentation without using any explicit pretext task, they use the pseudo-labels generated from semi-supervised learning strategies to provide additional guidance to the contrastive learning method. We adopt the pixel-level consistency and the feature-level consistency to constrain the semi-supervised segmentation prediction. The supervised segmentation labels and the segmentation consistency provide the basic semi-supervised segmentation performance on the pixel level. Through the unified translation task, the extra modality information and the feature-level contrastive learning consistency give the feature-level constraint.

2.4. Semi-supervised multi-modal medical image segmentation

At first, the semi-supervised multi-modal medical image segmentation is mainly about semi-supervised cross-modal learning tasks, which focus on fully annotated source modal images (usually CT images) and zero-annotated target modal images (usually MRI). To approach cross-modal segmentation, it usually utilizes translation to transform the unsupervised cross-modal segmentation problem into a semi-supervised single-modal segmentation problem [24–26,54]. However, it is time-consuming and has complicated training details while translating the simulated images. It only focuses on the pixel level of constraining by using the segmentation label, it does not combine the important features extracted from translation into the segmentation network. The improved strategy is to combine translation and segmentation in order to force the model learning and sharing the feature level semantic information, [29,30]. Nonetheless, the cross-modal tasks are limited to only two modalities and require fully annotated source modal images.

Recent studies are focusing on a broader field and applying the improved strategies to adapt the multiple modalities. Chartsias et al. [27] use disentangled decomposition encoders that are dedicated to each modality to get anatomical and imaging factors from the original data. The Shared anatomical factors from the different inputs are jointly processed and fused into the segmentation prediction. MASS [28] uses cross-modal consistency to regularize deep segmentation models in aspects of both semantic and anatomical spaces. Zhang et al. [31] propose a semi-supervised contrastive mutual learning segmentation framework, and an area-similarity contrastive loss to conduct semi-supervised multi-modal segmentation.

However, these methods are still restricted by the cross-modal strategies, which require training encoders or decoders dedicated to each modality. This design choice makes the training procedures heavy and rigid. To address this limitation, we propose applying a conditional encoder that can adapt to multiple modalities with flexible numbers. We demonstrate this flexibility through experiments conducted on a four-modal abdominal dataset and a six-site prostate dataset. Additionally, we combine important complementary information extracted from multi-modal data to enhance the semi-supervised segmentation task. We provide the idea for utilizing the combination of the latent features of translation and segmentation and applying both pixel-level and feature-level restraints for multi-modal tasks to solve semi-supervised multi-modal segmentation problems. Due to the heavy training procedures involved in applying cross-modal methods to multi-modal data, we opt not to conduct comparison experiments with existing research that employs cross-modal strategies.

3. Method

Our framework SMSUT, as in Fig. 1, consists of Segmentation and unified Translation Generator (STG), Discriminator (D), and multi-layer perceptron (MLP) connecting to the shared middle layers of STG. STG consists of an encoder–decoder sub-structure $G_{seg} = \{E_{seg}, L_{shr}, D_{seg}\}$ for multi-modal segmentation, and an encoder–decoder sub-structure $G_{tsl} = \{E_{tsl}, L_{shr}, D_{tsl}\}$ for multi-modal translation, L_{shr} is a shared convolution block, sharing semantic information and content information between two tasks. Given multi-modal images $X = \{X_i^l, X_i^u\}, x \in R^{1\times H\times W}, i \in N_m$ and their annotations $Y = \{Y^l\}, y \in R^{1\times H\times W}$, where subscript i denotes the modality index, superscript l and u denote labeled and unlabeled, N_m denotes the number of modalities, H and W denote the height of width of images. Given a modal attribute vector $C = \{C_i, C_j \parallel c \in \{0,1\}_m^N\}$, where c_i denotes one-hot coding for a source modality, c_j denotes one-hot coding for a target modality. We use the Difference Attribute (DA) Vectors [55] to define the translation between a target modality and a source modality, DA vectors from the source modality to the target modality are D_{ij} and DA vectors from the target modality to the source modality are D_{ji}. The detailed procedure is in pseudocode 1.

Algorithm 1 The SMSUT procedure

Require: N_m = number of modalities in the training set
Require: $x_i = \{x_i^l, x_i^u\}$ = training multi-modal medical images $i \in N_m$
Require: y^l = partial annotations
Require: I_{max} = the maximum of training epochs
Require: $f_{\theta_1}(x, d) = STG$ network with trainable parameter θ_1, d is the DA vector
Require: $g_{\theta_2}(z, c) = D$ network with trainable parameter θ_2, z is the generated images, c is the modal attribute vector
Require: $k_{\theta_3}(f) = MLP$ network with trainable parameter θ_3, f is the latent feature after the shared middle layer
Ensure: Trained STG and D
for $t = 0 \rightarrow I_{max}$ **do**
  for each minibatch B **do**
    randomly set target modality j
    get DA vectors d_{ij} and d_{ji}
    $z_j, p, f_{mid1} \leftarrow f_{\theta_1}(x_i, d_{ij})$ ▷ get network outputs
    $z_i, p_{cons}, f_{mid2} \leftarrow f_{\theta_1}(z_j, d_{ji})$ ▷ get network cyclic outputs
    supervised loss $\leftarrow y^l$ and p^l ▷ use Equation (1)
    adversarial losses $\leftarrow g_{\theta_2}(x_i, c_i)$ and $g_{\theta_2}(z_j, c_j)$ ▷ use Equation (2), (3), (4), (5)
    reconstruction loss $\leftarrow x_i$ and z_i ▷ use Equation (6)
    contrastive learning loss $\leftarrow k_{\theta_3}(f_{mid1})$ and $k_{\theta_3}(f_{mid2})$ ▷ use Equation (7)
    if $t \geq I_{pre}$ **then** ▷ calculate the segmentation consistency loss after I_{pre} iterations
      segmentation consistency loss $\leftarrow p$ and p_{cons} ▷ use Equation (8)
    end if
    Update $\theta_1, \theta_2, \theta_3$ ▷ update network parameters
  end for
end for
Test Input: x_i = multi-modal images
Test Output: p = segmentation prediction
Set target modality j
get DA vectors d_{ij}
$z_j, p, f_m id1 \leftarrow f_{\theta_1}(x_i, d_{ij})$ ▷ get network outputs

3.1. Supervised segmentation loss

The limited labeled image set (x_i^l, y^l) is the only supervised learning data in our method. We make the supervised segmentation loss as the base guidance of our network, which is the combination of cross entropy loss and dice loss between the prediction p^l and the manual annotation label y^l:

$$\mathcal{L}_{seg} = \mathbb{E}_{p^l \rightarrow P^l, y^l \rightarrow Y^l}[-(y^l \log(p^l) + \frac{2p^l y^l}{\|p^l\|_1 + \|y^l\|_1})], \tag{1}$$

where $\| \ldots \|_1$ denoted the L_1 norm.

3.2. Unsupervised learning based on unified translation

Unified translation helps our model to use the different information among multiple modalities to complement the lack of information in semi-supervised learning. The module can be flexibly applied to tasks with any number of modalities.

In order to enable STG to change the modal features of the output image to misguide the discriminator's classification results. We use the loss functions described in WGAN-GP [41] that can differentiate both the authenticity and modality information of images, which can be denoted as L_{adv}, L_{cls}. L_{adv} computes whether the generated image z is a near-real medical image with the gradient penalty that enables the stable training of GAN, L_{cls} is to verify the modality of images and to guide the multi-modal translation. They are defined as:

$$\begin{aligned}\mathcal{L}_{adv}^D = &-\mathbb{E}_{x_i \rightarrow X_i}[D_{src}(x_i)] + \mathbb{E}_{z_j \rightarrow Z_j}[D_{src}(z_j)] \\ &+\lambda_{gp}\mathbb{E}_{\tilde{x} \rightarrow \tilde{X}}[(\|\nabla_{\tilde{x}} D_{src}(\tilde{x})\|_2 - 1)^2],\end{aligned} \tag{2}$$

$$\mathcal{L}_{adv}^G = -\mathbb{E}_{z_j \rightarrow Z_j}[D_{src}(z_j)], \tag{3}$$

where the superscripts D and G denotes the loss that used for the Discriminator D and for the STG, the generated image $Z = \{Z_j^l, Z_j^u\}, z \in R^{1\times H\times W}, j \in N_m$, λ_{gp} denotes the hyper-parameter of the gradient parallel constraints, $\tilde{x}$ denotes the linear interpolation result of sample x_i and z_j, and $\| \ldots \|_2$ denotes the L_2 loss.

The other sub-discriminator D_{cls} is for identifying the modality of images, so as to guide the generator's multi-modal translation. STG needs to change the modal features of the output image to misguide the discriminator's classification results. Specifically, the Discriminator classification losses are defined as:

$$\begin{aligned}\mathcal{L}_{cls}^D = &-\mathbb{E}_{x_i \rightarrow X_i, c_i \rightarrow C_i}[c_i \log(D_{cls}(c_i \mid x_i))] \\ &+\mathbb{E}_{z_j \rightarrow Z_j, c_j \rightarrow C_j}[c_j \log(D_{cls}(c_j \mid z_j))],\end{aligned} \tag{4}$$

$$\mathcal{L}_{cls}^G = -\mathbb{E}_{z_j \rightarrow Z_j, c_j \rightarrow C_j}[c_j \log(D_{cls}(c_j \mid z_j))]. \tag{5}$$

The cyclic consistency reconstruction loss $\mathcal{L}_{rec}$ is defined as the L_1 loss between X_i and X_j, which prevents STG from creating images with the same content information in the modal reconstruction step. We define the cyclic consistency reconstruction loss as:

$$\mathcal{L}_{rec} = \mathbb{E}_{x_i \rightarrow X_i, z_j \rightarrow Z_j}[\|x_i - z_j\|_1]. \tag{6}$$

where $\| \ldots \|_1$ denoted the L_1 norm.

The STG utilizes a generative adversarial strategy to fully exploit multi-modal semi-supervised data. Additionally, the cyclic consistency reconstruction loss ensures that the STG can extract modality information without discarding any important details. The unified translation module serves as an unsupervised method for the task, thereby enhancing the semi-supervised segmentation performance.

3.3. Unsupervised learning based on feature-level and pixel-level consistency

The feature-level consistency for semi-supervised training relies deeply on the shared middle layers of the STG, as depicted in Fig. 1. Other than that, we incorporate the contrastive learning loss [56] into our network. This loss function leverages the concept of similarity measurement, which has been proven to be highly effective in classification and translation tasks. It provides additional feature-level supervision and constrains the deformation caused by the generative network.

The idea of nce loss is to map the positive samples (patches with the same content at the same location of the different feature maps) to be close to each other and the negative sample to be far away in the latent feature space, and the equation is:

$$\mathcal{L}_{nce} = \mathbb{E}_{x_i \to X_i, z_j \to Z_j}[\mathcal{L}_{patchNCE}(G, H, X_i) + \mathcal{L}_{patchNCE}(G, H, Z_j)], \tag{7}$$

where G denotes the generator STG and H denotes the MLP multi-layer perceptron, that connects to the STG's middle layer.

For pixel-level consistency constraint of semi-supervised tasks, we propose segmentation consistency loss in our training procedure 1. This loss also ensures that, after modal translation, the organ position and the morphology of organs are consistent, and the corresponding segmentation labels remain unchanged. Thus it forces STG to learn the same semantic features imposed by different modalities. We define this segmentation consistency loss $\mathcal{L}_{cons}$ as the combination of cross entropy loss and dice loss with combination ratios of ($\lambda_{ce} = 0.5, \lambda_{dc} = 0.5$).

$$\mathcal{L}_{cons} = \mathbb{E}_{p_{cons} \to P_{cons}, y_j \to Y_j}[-\lambda_{ce}(y_j \log(p_{cons}) + \lambda_{dc} \frac{2 y_j p_{cons}}{\|y_j\|_1 + \|p_{cons}\|_1})]. \tag{8}$$

As suggested in [57], we apply the dynamic adjustment strategy of exponential growth in setting the hyperparameter of $\mathcal{L}_{cons}$, with

$$\lambda_{cons} = \lambda_{max} \exp[-5(1 - \frac{I}{I_{max}})^2], \tag{9}$$

where λ_{max} denotes the maximum of the training weights, and I denotes the current iteration epoch, I_{max} denotes the maximum of the iteration epochs.

3.4. Overall losses

The SMSUT is designed to fully exploit semi-supervised multi-modal data. It consists of a unified translation structure that can adapt to a flexible number of modality information and provide strong supervision in addition to the limited annotation data. Our cyclic training procedure allows for a comprehensive investigation of the semi-supervised multi-modal data and enables the network to achieve a better training model. This approach effectively solves the problem of insufficient labels and provides additional supervision through the training procedure.

The complete loss functions of SMSUT are:

$$\min \mathcal{L}^{STG} = \lambda_{adv}\mathcal{L}^{G}_{adv} + \lambda_{cls}\mathcal{L}^{G}_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{nce}\mathcal{L}_{nce}, \tag{10}$$

$$\min \mathcal{L}^{D} = \mathcal{L}^{D}_{adv} + \lambda_{cls}\mathcal{L}^{D}_{cls}, \tag{11}$$

where $\lambda_{adv}, \lambda_{cls}, \lambda_{rec}, \lambda_{seg}, \lambda_{cons}$ and λ_{nce} denote the hyperparameters.

4. Experiments

We evaluate our proposed method with two datasets, multi-modal abdominal medical image segmentation dataset [58,59] and multi-site MRI prostate medical image segmentation dataset [60], to verify the performance of our model on semi-supervised multi-modal tasks and the ability to transfer to any number of modalities tasks. The results achieve superior performance in both semi-supervised learning and supervised learning experiments over the other competing models. We further design ablation experiments to analyze the performance of our model's critical loss functions.

4.1. Datasets

The multi-modal abdominal dataset contains four modalities, {CT, MR T1-DUAL in-phase, MR T1-DUAL out-phase, and MRI T2w Spectral Pre-saturation Inversion Recovery}, which have {20, 20, 20, 30} cases of 3D sequences respectively (36 slices on average). The CT images of the multi-modal abdominal dataset are sourced from the Multi-Atlas Labeling Beyond the Cranial Vault–Workshop Challenge data of MICCAI 2015 [59], Each scanning image has a resolution of 512 × 512 pixels. In the horizontal view, each pixel corresponds to a length of 0.54 mm to 0.98 mm (0.68 mm on average). The MR images in the multi-modal abdominal dataset are elicited from the Combined Healthy Abdominal Organ Segmentation challenge data of ISBI 2019 [58]. Each of the MR T1 in-phase, the MR T1 out-phase, and the MR T2w in the dataset has 20 cases of 3D sequences, and each case contains 25 to 50 slices of scanned images (36 slices on average). The resolution of 256 × 256 pixels. In the horizontal view, each pixel corresponds to a length of 1.36 mm to 1.89 mm (1.61 mm on average).

The multi-site prostate MR dataset has one prostate label and six different sites which are {RUNMC, BMC, HCRUDB, UCL, BIDMC, HK}, and the corresponding cases are {30, 30, 19, 13, 12, 12} (33 slices on average). It is elicited from the Multi-site Dataset for Prostate MRI Segmentation [60]. The resolution of 384 × 384 pixels. In the horizontal view, each pixel corresponds to a length of 0.25 mm to 0.79 mm (0.58 mm on average).

4.2. Evaluation metrics

We use standard evaluation metrics of medical image segmentation: Dice coefficient and the Average Symmetric Surface Distance (ASSD) to evaluate the performance of our model and competing methods. Dice coefficient (%) calculates the similarity between the prediction map and ground truth. A higher Dice value indicates better segmentation performance. ASSD (mm) calculates the average symmetric distances between the surface of the prediction mask and ground truth. A lower ASSD value indicates better segmentation performance. We use 2D images to train our model due to the limitation of computing power. After training and testing, we concatenate the 2D results into 3D sequences and calculate the evaluation metrics based on the 3D sequences.

4.3. Implementation details

We split our data into training, validation, and test sets, and the corresponding ratio of the abdominal dataset is (3 : 1 : 2) and the ratio of the prostate dataset is (6 : 1 : 2). We evaluate the performance of our model in 6 different ratios of labeled data, {10%, 30%, 50%, 70%, 90%, 100%} in semi-supervised multi-modal medical image segmentation. Also we test our method's performance on different modality sets in the 10% labeled data ratio scenario.

We use linear interpolation to change the image size so that each pixel matches the actual distance of 1.0 mm^3. To exclude the extreme values and irrelevant regions, all images are trimmed, centrally cropped, and modified to the size of 256 × 256 pixels. In addition, to exclude the extreme values, the values of the CT images are trimmed according to the range of Hounsfield Unit (HU) values of $[-500, 500]$. And the MR image pixel values are trimmed by the percentile of [0.5%, 99.5%]. Due to the limited amount of data set, we split our data into training and test sets in the ratios of (1 : 1) and (2 : 1) for the abdominal dataset and the prostate dataset respectively. The codes for data preprocessing, training, and evaluation are sourced from the open source project $SSL4MIS$.[1] We set hyperparameters {λ_{adv}, λ_{cls}, λ_{rec}, λ_{seg}, λ_{gp}, λ_{max}, λ_{nce}} as $\{1, 1, 10, 10, 10, 10, 1\}$ empirically. We set λ_{cons} to be zero within the first 1000 iterations (I_{pre}), in case the segmentation consistency loss would exacerbate the instability in the early stage of training.

[1] https://github.com/HiLab-git/SSL4MIS/

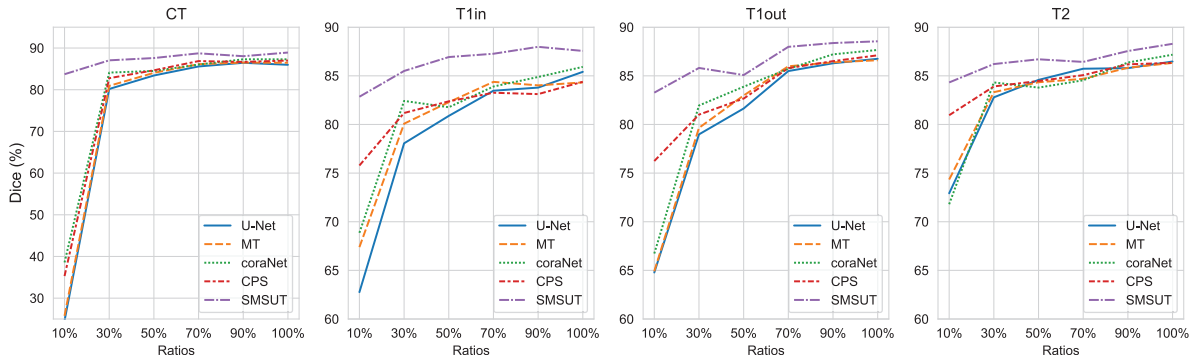

Fig. 2. Abdominal dataset Segmentation results on each modality, with ratio setting from 10% to 100%. The horizontal coordinate is the ratio and the vertical coordinate is the Dice value(%).

Table 1
Comparison models based on modality and supervision type.

Models	Modality	Supervision type
U-Net [34]	uni-modal	fully supervised
MT [4]	uni-modal	semi-supervised
coraNet [51]	uni-modal	semi-supervised
CPS [50]	uni-modal	semi-supervised
DualS [61]	multi-modal	fully supervised
KD [62]	multi-modal	fully supervised
MASS [28]	multi-modal	semi-supervised
M3L [32]	multi-modal	semi-supervised
SMSUT(ours)	multi-modal	semi-supervised

4.4. Comparison methods

We compare our model with eight segmentation methods. Table 1 categorizes the comparison models by modality and supervision type.

Uni-modal segmentation methods use the dataset as uni-modal data, while fully supervised methods train only on labeled data. Mean Teacher [4] is a classical semi-supervised method, which constrains the consistent results of the student model and the teacher model. CoraNet [51] combines the MT model with co-training segmentation networks that focus on different tasks. And CPS [50] co-trains two parallel networks with different parameter initialization. Due to the superior performance of U-Net in medical image segmentation, we adopt the U-Net as the backbone of all the uni-modal semi-supervised methods.

Multi-modal segmentation methods are designed based on cross-modal strategy. DualS [61] employs a dual-stream encoder–decoder architecture for unpaired image multi-modal segmentation. KD [62] utilizes knowledge distillation to constrain outputs from multiple modalities. MASS [28] leverages multiple atlases' labels and cross-modal consistency for multi-modal semi-supervised segmentation. The comparison results of DualS, KD, and MASS are directly sourced from the referenced articles. These methods utilize the same abdominal dataset as ours, which is constructed from the CHAOS dataset (MRI) [58] and the Multi-Atlas dataset (CT) [59]. M3L [32], on the other hand, is a multi-modal semi-supervised segmentation method that has demonstrated successful performance in natural images. It achieves this by integrating multi-modal information using Linear Fusion and employing a masked mean teacher strategy.

5. Results

5.1. Sensitivity analysis on labeled data ratio

Table 2 shows the quantitative segmentation performances on the multi-modal medical images under different ratios of labeled data. Table 5 shows the quantitative segmentation performances on the multi-site medical images under different ratios of labeled data. As these two tables show, our model SMSUT obtains consistently higher accuracy in terms of Dice and ASSD over the baseline, U-Net, across different ratio settings of the labeled data. Our method outperforms other comparison approaches in every ratio setting of the labeled data. Also, our model achieves better performance even with fewer labeled data, which indicates that our network can exploit more information within data. Our unified translation module can provide a thorough investigation of the multi-modal images and our consistency constraint can provide extra semantic information. The results show that our network can reduce the annotation workload of physicians, and become more applicable in real-world scenarios.

5.2. Sensitivity analysis on modalities

Table 4 demonstrates that increasing the number of training modalities enhances the performance of our methods. This suggests that our model effectively integrates complementary information from each modality, resulting in improved segmentation results. Each modality possesses distinct features that aid in analyzing the segmentation of other modalities. In the strict semi-supervised scenario with only 10% labeled data, our method achieves the highest average performance among all models when trained with all four modalities.

When considering the complementary information from CT and T1in images, T2 images achieve an impressive Dice value of 87.49%, surpassing the performance achieved by training with all four modalities. This phenomenon can be attributed to the modality ratio when applying the multi-modal segmentation model. In the setting of {CT, T1in, T2}, T2 achieves optimal results by leveraging the complementary information from both CT and T1in. However, in the more complex setting of {CT, T1in, T1in, T2}, the model must strike a balance among these four modality images, which may lead to a reduction in T2 segmentation performance. Table 4 reveals that our method, when using the {CT, T2} setting, exhibits relatively lower performance compared to the MASS model [28]. The discrepancy can be attributed to differences in dataset size: MASS utilizes 70% of the dataset, while our approach uses only 50%. Dataset size significantly impacts training performance. Importantly, our model demonstrates flexibility and can achieve higher performance when incorporating additional modalities. Compared to other cross-modal semi-supervised models, our approach consistently outperforms in experiments. While the M3L [32] model utilizes a transformer architecture to integrate multi-modal information, it requires a large dataset for effective training due to its extensive parameters. In contrast, our method leverages unified translation for multi-modal segmentation, enabling successful training even with limited data.

The results of Table 3, Table 4 and Fig. 2 show that our model significantly surpasses the other semi-supervised methods among all the modalities (CT, MR T1 in-phase, MR T1 out-phase, and MR T2) in the Dice metric. In the 10% ratio scenario, the few labeled CT

Table 2
Model performance comparison on the abdominal dataset. % of *Dice* and mm of *ASSD* (Average Symmetric Surface Distance) are omitted, the results are the means of multi-label and multi-modal results, and after ± is the standard deviation. *MT* means the teacher model [4], *CPS* means the Cross Pseudo model [50], *SMSUT* is our model, Semi-supervised Multi-modal Segmentation network with Unified Translation. The leftmost column shows the ratios of labeled data. The best results among models are in **bold**.

Ratio	U-Net [34]		MT [4]		coraNet [51]		CPS [50]		SMSUT(ours)	
	Dice	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*
10%	56.26	7.55	58.10	7.81	61.49	5.15	67.07	3.97	**83.54**	**1.51**
	±0.13	±0.50	±0.64	±1.01	±1.07	±0.37	±0.36	±0.37	±0.38	±0.10
30%	80.00	1.84	81.00	1.79	83.20	1.67	82.22	1.67	**86.14**	**1.22**
	±0.34	±0.09	±0.60	±0.07	±0.60	±0.10	±0.15	±0.08	±0.26	±0.07
50%	82.63	1.66	83.42	1.63	83.48	1.60	83.54	1.69	**86.58**	**1.19**
	±0.17	±0.029	±0.17	±0.076	±0.50	±0.01	±0.09	±0.071	±0.43	±0.05
70%	85.07	1.41	85.30	1.43	85.05	1.41	85.25	1.42	**87.60**	**1.20**
	±0.08	±0.09	±0.08	±0.055	±0.24	±0.01	±0.22	±0.049	±0.19	±0.04
90%	85.58	1.38	85.69	1.52	86.44	1.28	85.62	1.49	**87.99**	**1.09**
	±0.28	±0.05	±0.06	±0.07	±0.21	±0.11	±0.12	±0.09	±0.04	±0.12
100%	86.15	1.30	85.93	1.50	87.01	1.45	86.21	1.50	**88.33**	**1.05**
	±0.18	±0.10	±0.07	±0.09	±0.26	±0.26	±0.15	±0.07	±0.12	±0.03

Table 3
Multi-modal Performance on The Abdominal Dataset with ratio 10% and 100%.

Ratio	Models	CT		T1in		T1out		T2		Average	
		Dice	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*
10%	U-Net [34]	24.57 ±0.32	18.95 ±1.68	62.76 ±0.85	4.65 ±0.58	64.78 ±0.64	3.50 ±0.42	72.92 ±0.88	3.11 ±0.31	56.26 ±0.13	7.55 ±0.50
	MT [4]	25.76 ±1.94	20.46 ±5.17	67.39 ±1.79	3.78 ±1.18	64.91 ±1.52	3.69 ±0.14	74.34 ±2.39	3.30 ±0.29	58.10 ±0.64	7.81 ±1.01
	coraNet [51]	38.57 ±4.73	9.65 ±1.18	68.86 ±2.79	3.25 ±0.86	66.73 ±2.44	3.63 ±1.59	71.80 ±1.31	4.08 ±0.57	61.49 ±1.07	5.15 ±0.37
	CPS [50]	35.30 ±1.54	9.11 ±2.08	75.78 ±0.12	2.31 ±0.32	76.24 ±0.32	2.09 ±0.12	80.95 ±0.16	2.35 ±0.34	67.07 ±0.36	3.97 ±0.37
	SMSUT (ours)	**83.72** ±1.32	**1.56** ±0.03	**82.85** ±0.47	**1.48** ±0.17	**83.27** ±0.82	**1.37** ±0.14	**84.32** ±0.17	**1.63** ±0.10	**83.54** ±0.38	**1.51** ±0.10
100%	U-Net [34]	85.98 ±0.47	1.23 ±0.07	85.41 ±0.78	1.11 ±0.03	86.75 ±0.18	1.03 ±0.02	86.47 ±0.57	1.82 ±0.44	86.15 ±0.18	1.30 ±0.10
	MT [4]	86.47 ±0.26	1.30 ±0.03	84.32 ±0.42	1.25 ±0.09	86.60 ±0.44	1.08 ±0.01	86.33 ±0.20	2.38 ±0.25	85.93 ±0.07	1.50 ±0.09
	coraNet [51]	87.29 ±0.30	1.08 ±0.13	85.92 ±0.24	1.06 ±0.05	87.66 ±0.28	0.93 ±0.04	87.17 ±0.46	2.71 ±1.20	87.01 ±0.26	1.45 ±0.26
	CPS [50]	87.00 ±0.26	1.22 ±0.11	84.39 ±0.56	1.47 ±0.45	87.11 ±0.21	1.04 ±0.03	86.34 ±0.42	2.28 ±0.37	86.21 ±0.15	1.50 ±0.07
	SMSUT (ours)	**88.92** ±0.20	**0.96** ±0.07	**87.56** ±0.19	**1.03** ±0.04	**88.55** ±0.03	**0.92** ±0.01	**88.29** ±0.45	**1.28** ±0.17	**88.33** ±0.12	**1.05** ±0.03

images lead to bad performance in other models, while our network achieves a relatively high performance due to utilizing the unified translation method to provide extra information from other modal data and employing enough consistency constraint from feature level and pixel level. In the 100% ratio scenario, our model still proves that our unique structure and training strategy can serve better results. This suggests that SMSUT, by leveraging the unified translation module, can fully exploit both the complementary and the shared information of the different modalities to improve the segmentation performance in the multi-modal scenario.

5.3. The multi-site prostate segmentation task

In general, the images of the same modality collected by different institutions can have great variability in terms of imaging parameters, size of shooting area, brightness setting, etc. The multi-site problem can also be addressed by semi-supervised multi-modal methods. We can readily transfer our network to address multi-site problems due to its flexibility and portability. Table 5 shows the average performance of the 6 different sites on the prostate MRI dataset. For this task, all the models show better performance. This is due to the relatively simple segmentation object and the uniform modality of the prostate dataset. Our model still achieves the best performance for every ratio of the labeled data. It suggests that under the semi-supervised multi-site segmentation scenario, SMSUT can effectively leverage the complementary information from various site images to improve the segmentation performance. This experiment shows that our network is adaptable to any number of modalities and, therefore has great development prospects to be applied in real clinical scenes.

5.4. Ablation experiments

We investigate the impacts of the losses on segmentation performance separately, in the extreme condition of 10% labeled ratio. As Fig. 3 and Table 6 illustrate, the losses from the unified translation module provide more semantic accuracy and restraints to the model training from the multi-modal data which compensates for the lack of label information and brings significant improvements in segmentation performance. Without the adversarial loss (L_{adv}) and the classification loss (L_{cls}), the segmentation prediction would make large-area mistakes and the contour of the results would be discontinuous. The outcomes obtained from the model without the adversarial losses demonstrate a notable enhancement attributed to the unified translation module, particularly in scenarios involving limited annotation.

Furthermore, Fig. 3 and Table 6 show that the contrastive learning loss (L_{nce}) and the segmentation consistency loss (L_{cons}) constrain the deformation in the feature level and pixel level, which makes the model more capable of capturing the details of the segmented objects and makes the segmentation results closer to the real results locally. The results derived from the model lacking the segmentation consistency loss underscore the significance of segmentation consistency loss within our framework.

Table 4
Experiments of sensitivity analysis on modalities and model comparison experiments on the abdominal dataset. The labeled data ratio is set to 10%. Due to the design of multi-modal segmentation models based on cross-modal segmentation, we select the comparison modality settings that include only two modalities: CT and MRI. The first two column is the performance of fully supervised multi-modal segmentation methods, and the rest results are the performance of semi-supervised multi-modal segmentation methods. *DualS* means the DualStream model [61], *KD* means the unpaired multi-modal model with Knowledge Distillation [62]. *M3L* means the Multi-modal teacher for Masked Modality Learning model [32]. * denotes results directly sourced from the referenced articles without replication in this study.

Model	CT	T1in	T1out	T2	CT		T1in		T1out		T2		Average	
					Dice	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*
DualS*[61]	✓			✓	74.7 ±6.1	–	–	–	–	–	77.5 ±4.5	–	76.1	—
KD*[62]	✓			✓	76.6 ±5.2	–	–	–	–	–	78.3 ±4.2	–	77.5	—
MASS*[28]	✓			✓	81.3 ±4.6	–	–	–	–	–	82.1 ±3.4	–	81.7	—
M3L [32]	✓	✓			73.16 ±0.76	2.54 ±0.23	63.64 ±5.76	6.58 ±1.70	–	–	–	–	68.40 ±2.50	4.56 ±0.96
	✓		✓		69.47 ±2.46	3.52 ±0.11	–	–	69.91 ±0.21	2.97 ±0.27	–	–	69.69 ±1.20	3.24 ±0.15
	✓			✓	76.08 ±11.03	3.08 ±0.92	–	–	–	–	67.41 ±8.83	3.39 ±1.28	71.74 ±1.10	3.24 ±0.17
SMSUT (ours)	✓	✓			82.78 ±0.66	2.43 ±0.18	80.71 ±0.41	1.40 ±0.10	–	–	–	–	81.75 ±0.12	1.92 ±0.14
	✓		✓		83.21 ±0.30	1.49 ±0.42	–	–	80.77 ±0.96	1.64 ±0.23	–	–	81.99 ±0.64	1.57 ±0.32
	✓			✓	83.46 ±4.62	1.88 ±0.67	–	–	–	–	79.00 ±5.59	2.43 ±0.18	81.23 ±5.11	2.16 ±0.40
	✓	✓	✓		81.56 ±0.47	1.58 ±0.03	82.30 ±2.02	**1.37** ±0.28	81.50 ±1.82	1.56 ±0.38	–	–	81.79 ±1.44	1.51 ±0.23
	✓	✓		✓	83.14 ±1.88	**1.38** ±0.31	78.14 ±3.01	1.83 ±0.12	–	–	**87.49** ±0.72	**1.17** ±0.26	82.92 ±1.87	**1.46** ±0.23
	✓		✓	✓	82.18 ±2.40	1.70 ±0.18	–	–	77.11 ±4.55	1.76 ±0.32	87.16 ±0.97	1.49 ±0.07	82.15 ±2.64	1.65 ±0.19
	✓	✓	✓	✓	**83.72** ±1.32	1.56 ±0.03	**82.85** ±0.47	1.48 ±0.17	**83.27** ±0.82	**1.37** ±0.14	84.32 ±0.17	1.63 ±0.10	**83.54** ±0.38	1.51 ±0.10

Table 5
Model performance comparison on the prostate dataset.

Ratio	U-Net [34]		MT [4]		coraNet [51]		CPS [50]		SMSUT(ours)	
	Dice	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*
10%	74.06 ±0.01	2.32 ±0.46	75.83 ±0.01	1.94 ±0.05	75.81 ±0.02	2.25 ±0.19	80.98 ±0.01	1.58 ±0.05	**83.22** ±0.01	**1.35** ±0.05
30%	84.04 ±0.01	1.32 ±0.16	84.88 ±0.01	1.23 ±0.07	83.70 ±0.01	1.35 ±0.09	85.27 ±0.01	1.17 ±0.05	**86.52** ±0.01	**1.12** ±0.06
50%	86.75 ±0.01	1.22 ±0.11	86.86 ±0.01	1.05 ±0.02	86.01 ±0.01	1.14 ±0.06	86.89 ±0.01	1.08 ±0.04	**88.11** ±0.01	**1.02** ±0.06
70%	87.35 ±0.01	1.03 ±0.03	87.55 ±0.01	1.03 ±0.01	87.06 ±0.01	1.05 ±0.02	87.87 ±0.01	0.99 ±0.02	**88.82** ±0.01	**0.92** ±0.01
90%	88.00 ±0.01	0.97 ±0.02	87.85 ±0.01	1.02 ±0.02	87.50 ±0.01	1.02 ±0.02	87.71 ±0.01	1.00 ±0.02	**88.73** ±0.01	**0.90** ±0.04
100%	87.50 ±0.01	1.01 ±0.01	87.73 ±0.01	1.01 ±0.02	88.20 ±0.01	0.94 ±0.02	87.61 ±0.01	1.02 ±0.03	**89.22** ±0.01	**0.90** ±0.02

These findings highlight the substantial guidance provided by such losses, particularly in settings with limited annotation. These outcomes collectively affirm our network's adeptness at effectively leveraging semi-supervised multi-modal information to facilitate the process of multi-modal information extraction, thus bolstering the semi-supervised segmentation task.

5.5. Visualization

Due to the insufficient annotation of medical images in the real-world scenario, we choose ratio setting 10% of the labeled data to present the visualization performance. The results imply that training on a small amount of annotated data leads to the boundary of objects being difficult to classify clearly. As Fig. 4 shows, it is apparent that SMSUT outperforms the other comparative models in the localization and the segmentation boundaries. The unified translation results show that our model successfully changes the modality of the image and retains the content information distinctly. The translated images on the left columns show that our model extracts the different properties of each modality which can generate the four modality images successfully. By utilizing this learned information from our unified translation module and our consistency constraints, our network provides better performance on the limited-label multi-modal task.

6. Discussion

In medical images, precise object boundaries are essential for diagnosis and treatment planning. However, the segmentation models' limitation is the relatively low accuracy in delineating the contours of objects. As shown in Fig. 4, our model in the 10% labeled ratio setting can achieve a better performance in the results, but it still needs to enhance the ability to capture the fine details and irregularities in object boundaries. To address this limitation, our future work should focus on the development of methods that prioritize the accurate delineation of object contours. For example, add loss functions that explicitly penalize errors near object boundaries and incorporate multi-scale features in the segmentation network to help capture both local and global context.

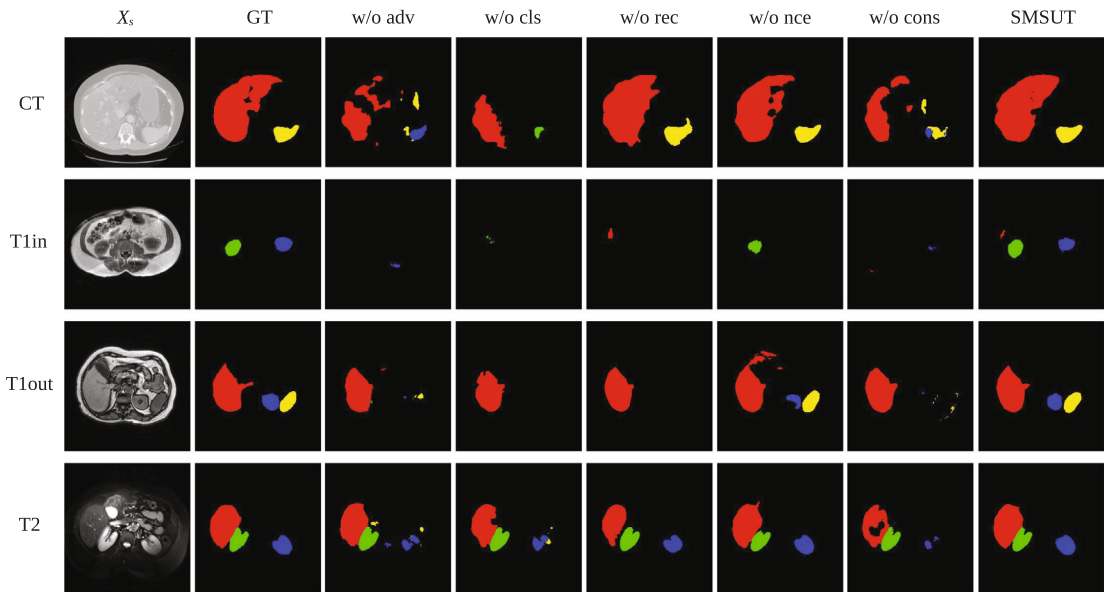

Fig. 3. Ablation experiment results in abdominal dataset, with labeled data ratio setting to 10%. The figure from left to right contains source images X_s, the ground truth (GT), the segmentation results of ablated models without L_{adv} (w/o adv), without L_{cls} (w/o cls), without L_{rec} (w/o rec), without L_{nce} (w/o nce), without L_{cons} (w/o cons), and SMSUT. The segmentation labels of the abdominal dataset denote four organs: red is for the liver, green is for the right kidney, blue is for the left kidney, and yellow is for the pancreas.

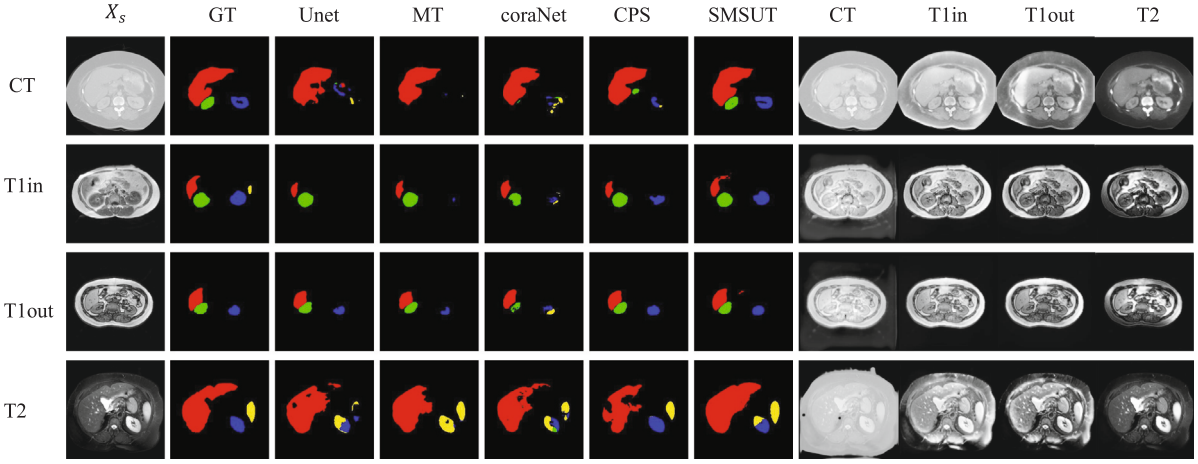

Fig. 4. Segmentation results and unified translation results in the abdominal dataset, with ratio setting of 10% of the labeled data. The figure contains source images X_s, the ground truth (GT), the segmentation results of Comparison methods and SMSUT, and the translated images (CT, T1in, T1out, T2). The segmentation labels represent four organs, {red, green, blue, yellow} denote {liver, right kidney, left kidney, pancreas}.

Table 6
Ablation experiments on the abdominal dataset with the labeled data ratio setting to 10%.

Models	CT		T1in		T1out		T2		Average	
	Dice	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*	*Dice*	*ASSD*
w/o L_{adv}	49.10 ±3.32	5.94 ±1.36	73.88 ±1.49	2.03 ±0.22	68.76 ±3.95	2.26 ±0.26	80.24 ±2.40	2.53 ±0.19	67.99 ±1.13	3.19 ±0.27
w/o L_{cls}	80.77 ±0.30	1.99 ±0.02	82.33 ±0.73	1.63 ±0.15	79.75 ±0.04	1.64 ±0.24	82.14 ±0.74	1.91 ±0.35	81.25 ±0.31	1.79 ±0.15
w/o L_{rec}	78.54 ±3.10	1.95 ±0.72	79.68 ±0.01	1.69 ±1.12	81.05 ±2.31	1.58 ±0.90	**84.45** ±0.22	**1.49** ±0.37	80.93 ±1.30	1.68 ±0.78
w/o L_{nce}	81.87 ±1.55	1.67 ±0.10	80.40 ±0.93	1.67 ±0.12	82.33 ±0.99	1.52 ±0.18	83.43 ±0.45	2.20 ±0.34	82.01 ±0.45	1.77 ±0.15
w/o L_{cons}	35.42 ±1.21	10.88 ±2.31	64.58 ±2.20	3.30 ±0.25	62.05 ±1.54	3.46 ±0.17	77.07 ±1.16	2.91 ±0.10	59.78 ±1.14	5.14 ±0.60
SMSUT (ours)	**83.72** ±1.32	**1.56** ±0.03	**82.85** ±0.47	**1.48** ±0.17	**83.27** ±0.82	**1.37** ±0.14	84.32 ±0.17	1.63 ±0.10	**83.54** ±0.38	**1.51** ±0.10

The radiological medical images are often volumetric (e.g., CT or MRI scans). Exploring the integration of three-dimensional perspectives into the segmentation network can bring accuracy and continuity prediction between slices. Our future work would consider adding context

across multiple slices so that our model can generate more continuous and anatomically consistent segmentation. The Transformer architecture is currently at the forefront of segmentation and other related tasks. Recent research has demonstrated its efficacy, particularly due to the multi-head cross and self-attention mechanisms, which facilitate the integration of multi-modal information. Our ongoing work also focuses on exploring how to leverage this powerful method effectively in scenarios with limited training data.

7. Conclusion

We introduce a novel idea of the segmentation network with unified translation in real-world semi-supervised multi-modal scenarios. By unified translation of multi-modal images and using consistency constraints at the pixel level and feature level, we successfully leverage the differential information between different modalities to achieve better semi-supervised learning. In particular, due to our network structure and training procedure, our model can be easily transferred to any other semi-supervised multi-modal segmentation tasks. Experiments show that our method can effectively handle multi-modal and multi-site segmentation and outperform other semi-supervised methods. Our future work will focus on increasing the accuracy of segmenting the contours of the objects and integrating the 3D perspectives into our network.

CRediT authorship contribution statement

Huajun Sun: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jia Wei:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision. **Wenguang Yuan:** Conceptualization, Methodology, Writing – review & editing. **Rui Li:** Writing – review & editing.

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 article.

Acknowledgments

This work is supported in part by the Guangdong Provincial Natural Science Foundation, China (2023A1515011431), the Guangzhou Science and Technology Planning Project (202201010092), the National Natural Science Foundation of China (72074105), National Science Foundation-1850492 and National Science Foundation-2045804.

References

[1] Zhongpai Gao, Guangtao Zhai, Chunjia Hu, Xiongkuo Min, Dual-view medical image visualization based on spatial-temporal psychovisual modulation, in: 2014 IEEE International Conference on Image Processing, ICIP, IEEE, 2014, pp. 2168–2170.

[2] Xiangli Yang, Zixing Song, Irwin King, Zenglin Xu, A survey on deep semi-supervised learning, IEEE Trans. Knowl. Data Eng. (2022).

[3] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M. Matthews, Daniel Rueckert, Semi-supervised learning for network-based cardiac MR image segmentation, in: Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20, Springer, 2017, pp. 253–260.

[4] Antti Tarvainen, Harri Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, Gustavo Carneiro, Perturbed and strict mean teachers for semi-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4258–4267.

[6] Haiming Xu, Lingqiao Liu, Qiuchen Bian, Zhen Yang, Semi-supervised semantic segmentation with prototype-based consistency regularization, Adv. Neural Inf. Process. Syst. 35 (2022) 26007–26020.

[7] Zishun Feng, Dong Nie, Li Wang, Dinggang Shen, Semi-supervised learning for pelvic MR image segmentation based on multi-task residual fully convolutional networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, IEEE, 2018, pp. 885–888.

[8] Xiangde Luo, Jieneng Chen, Tao Song, Guotai Wang, Semi-supervised medical image segmentation through dual-task consistency, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (no. 10) 2021, pp. 8801–8809.

[9] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P. Hughes, Danny Z. Chen, Deep adversarial networks for biomedical image segmentation utilizing unannotated images, in: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20, Springer, 2017, pp. 408–416.

[10] F. Lafitte, M. Boukobza, J.P. Guichard, C. Hoeffel, D. Reizine, O. Ille, F. Woimant, J.J. Merland, MRI and MRA for diagnosis and follow-up of cerebral venous thrombosis (CVT), Clin. Radiol. 52 (9) (1997) 672–679.

[11] Leon Axel, Ricardo Otazo, Accelerated MRI for the assessment of cardiac function, Brit. J. Radiol. 89 (1063) (2016) 20150655.

[12] B. Rajalingam, R. Priya, R. Scholar, Review of multimodality medical image fusion using combined transform techniques for clinical application, Int. J. Sci. Res. Comput. Sci. Appl. Manag. Stud. 7 (3) (2018) 1–8.

[13] Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, Amir H. Gandomi, A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, Comput. Biol. Med. 144 (2022) 105253.

[14] Massimo Salvi, Hui Wen Loh, Silvia Seoni, Prabal Datta Barua, Salvador García, Filippo Molinari, U. Rajendra Acharya, Multi-modality approaches for medical support systems: A systematic review of the last decade, Inf. Fusion 103 (C) (2024).

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, vol. 27, 2014.

[16] Chengjia Wang, Guang Yang, Giorgos Papanastasiou, Sotirios A. Tsaftaris, David E. Newby, Calum Gray, Gillian Macnaught, Tom J. MacGillivray, DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis, Inf. Fusion 67 (2021) 147–160.

[17] Siyi Xun, Dengwang Li, Hui Zhu, Min Chen, Jianbo Wang, Jie Li, Meirong Chen, Bing Wu, Hua Zhang, Xiangfei Chai, et al., Generative adversarial networks in medical image segmentation: A review, Comput. Biol. Med. 140 (2022) 105063.

[18] Jie Yang, Ye Zhu, Chaoqun Wang, Zhen Li, Ruimao Zhang, Toward unpaired multi-modal medical image segmentation via learning structured semantic consistency, 2022.

[19] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, Yefeng Zheng, Mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 107–117.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[22] Michael L. Goris, et al., Medical image acquisition and processing: Clinical validation, Open J. Med. Imaging 4 (04) (2014) 205.

[23] Lei Zhu, Kaiyuan Yang, Meihui Zhang, Ling Ling Chan, Teck Khim Ng, Beng Chin Ooi, Semi-supervised unpaired multi-modal learning for label-efficient medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer, 2021, pp. 394–404.

[24] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S. Mageras, Joseph O. Deasy, Harini Veeraraghavan, Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, Springer, 2018, pp. 777–785.

[25] Cheng Chen, Qi Dou, Hao Chen, Pheng-Ann Heng, Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest X-ray segmentation, in: Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9, Springer, 2018, pp. 143–151.

[26] Jun Chen, Heye Zhang, Raad Mohiaddin, Tom Wong, David Firmin, Jennifer Keegan, Guang Yang, Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data, IEEE Trans. Med. Imaging 41 (2) (2021) 420–433.

[27] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E. Newby, Rohan Dharmakumar, Sotirios A. Tsaftaris, Disentangle, align and fuse for multimodal and semi-supervised image segmentation, IEEE Trans. Med. Imaging 40 (3) (2020) 781–792.
[28] Xiaoyu Chen, Hong-Yu Zhou, Feng Liu, Jiansen Guo, Liansheng Wang, Yizhou Yu, MASS: Modality-collaborative semi-supervised segmentation by exploiting cross-modal consistency from unpaired CT and MRI images, Med. Image Anal. 80 (2022) 102506.
[29] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, Trevor Darrell, Cycada: Cycle-consistent adversarial domain adaptation, in: International Conference on Machine Learning, Pmlr, 2018, pp. 1989–1998.
[30] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, Mingkui Tan, Collaborative unsupervised domain adaptation for medical image diagnosis, IEEE Trans. Image Process. 29 (2020) 7834–7844.
[31] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, Zhenghua Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.
[32] Harsh Maheshwari, Yen-Cheng Liu, Zsolt Kira, Missing modality robustness in semi-supervised multi-modal semantic segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 1020–1030.
[33] Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, Qianjin Feng, SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation, IEEE Trans. Med. Imaging 40 (1) (2020) 262–273.
[34] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.
[35] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-Unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.
[36] Jing Zhang, Qiuge Qin, Qi Ye, Tong Ruan, ST-Unet: Swin transformer boosted U-Net with cross-layer feature enhancement for medical image segmentation, Comput. Biol. Med. 153 (2023) 106516.
[37] Fares Bougourzi, Cosimo Distante, Fadi Dornaika, Abdelmalik Taleb-Ahmed, PDAtt-Unet: Pyramid dual-decoder attention unet for COVID-19 infection segmentation from CT-scans, Med. Image Anal. 86 (2023) 102797.
[38] Xuping Huang, Junxi Chen, Mingzhi Chen, Lingna Chen, Yaping Wan, TDD-UNet:Transformer with double decoder UNet for COVID-19 lesions segmentation, Comput. Biol. Med. 151 (2022) 106306.
[39] Chufu Deng, Zhiguang Chen, Ruixuan Wang, Wanqi Su, Yili Qu, Modality-shared MRI image translation based on conditional GAN, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2021, pp. 1286–1290.
[40] Xiaoting Han, Lei Qi, Qian Yu, Ziqi Zhou, Yefeng Zheng, Yinghuan Shi, Yang Gao, Deep symmetric adaptation network for cross-modality medical image segmentation, IEEE Trans. Med. Imaging 41 (1) (2022) 121–132.
[41] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
[42] Taesung Park, Alexei A. Efros, Richard Zhang, Jun-Yan Zhu, Contrastive learning for unpaired image-to-image translation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, 2020, pp. 319–345.
[43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[44] Mehdi Mirza, Simon Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.
[45] Euijin Jung, Miguel Luna, Sang Hyun Park, Conditional GAN with 3D discriminator for MRI generation of Alzheimer's disease progression, Pattern Recognit. 133 (2023) 109061.
[46] Jonathan David Ziegler, Sajanth Subramaniam, Michela Azzarito, Orla Doyle, Peter Krusche, Thibaud Coroller, Multi-modal conditional GAN: Data synthesis in the medical domain, in: NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.
[47] Mohammad H. Jafari, Hany Girgis, Amir H. Abdi, Zhibin Liao, Mehran Pesteie, Robert Rohling, Ken Gin, Terasa Tsang, Purang Abolmaesumi, Semi-supervised learning for cardiac left ventricle segmentation using conditional deep generative models as prior, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019, pp. 649–652.
[48] Kevinminh Ta, Shawn S. Ahn, Allen Lu, John C. Stendahl, Albert J. Sinusas, James S. Duncan, A semi-supervised joint learning approach to left ventricular segmentation and motion tracking in echocardiography, in: 2020 IEEE 17th International Symposium on Biomedical Imaging, ISBI, IEEE, 2020, pp. 1734–1737.
[49] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, Pheng-Ann Heng, Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer, 2019, pp. 605–613.
[50] Xiaokang Chen, Yuhui Yuan, Gang Zeng, Jingdong Wang, Semi-supervised semantic segmentation with cross pseudo supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2613–2622.
[51] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, Yang Gao, Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation, IEEE Trans. Med. Imaging 41 (3) (2022) 608–620.
[52] Yassine Ouali, Céline Hudelot, Myriam Tami, Semi-supervised semantic segmentation with cross-consistency training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12674–12684.
[53] Hritam Basak, Zhaozheng Yin, Pseudo-label guided contrastive learning for semi-supervised medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19786–19797.
[54] Yue Zhang, Shun Miao, Tommaso Mansi, Rui Liao, Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II, Springer, 2018, pp. 599–607.
[55] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[56] Taesung Park, Alexei A. Efros, Richard Zhang, Jun-Yan Zhu, Contrastive learning for unpaired image-to-image translation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, 2020, pp. 319–345.
[57] Samuli Laine, Timo Aila, Temporal ensembling for semi-supervised learning, 2016, arXiv preprint arXiv:1610.02242.
[58] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, N. Sinem Gezer, CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, Zenodo, 2019.
[59] Bennett Landman, Zhoubing Xu, J. Igelsias, Martin Styner, T. Langerak, Arno Klein, Miccai multi-Atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge, vol. 5, 2015, p. 12.
[60] Quande Liu, Qi Dou, Pheng Ann Heng, Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains, in: International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI, 2020.
[61] Vanya V. Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker, Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, 2018, pp. 547–556.
[62] Qi Dou, Quande Liu, Pheng Ann Heng, Ben Glocker, Unpaired multi-modal segmentation via knowledge distillation, IEEE Trans. Med. Imaging 39 (7) (2020) 2415–2425.