Multi-ConDoS: Multimodal Contrastive Domain Sharing Generative Adversarial Networks for Self-Supervised Medical Image Segmentation

Jiaojiao Zhang, Shuo Zhang, Xiaoqian Shen, Thomas Lukasiewicz, and Zhenghua Xu, *Member, IEEE*

***Abstract*— Existing self-supervised medical image segmentation usually encounters the domain shift problem (i.e., the input distribution of pre-training is different from that of fine-tuning) and/or the multimodality problem (i.e., it is based on single-modal data only and cannot utilize the fruitful multimodal information of medical images). To solve these problems, in this work, we propose multimodal contrastive domain sharing (Multi-ConDoS) generative adversarial networks to achieve effective multimodal contrastive self-supervised medical image segmentation. Compared to the existing self-supervised approaches, Multi-ConDoS has the following three advantages: (i) it utilizes multimodal medical images to learn more comprehensive object features via multimodal contrastive learning; (ii) domain translation is achieved by integrating the cyclic learning strategy of CycleGAN and the cross-domain translation loss of Pix2Pix; (iii) novel domain sharing layers are introduced to learn not only domain-specific but also domain-sharing information from the multimodal medical images. Extensive experiments on two publicly multimodal medical image segmentation datasets show that, with only 5% (resp., 10%) of labeled data, Multi-ConDoS not only greatly outperforms the state-of-the-art self-supervised and semi-supervised medical image segmentation baselines with the same ratio of labeled data, but also achieves similar (sometimes even better) performances as fully supervised segmentation methods with 50% (resp., 100%) of labeled data, which thus proves that our work can achieve superior segmentation performances with very low labeling workload. Furthermore, ablation studies prove that the above three improvements are all effective and essential for Multi-ConDoS to achieve this very superior performance.**

***Index Terms*— Self-supervised learning, multi-modal medical image segmentation, contrastive learning, domain translation, domain sharing.**

Manuscript received 22 May 2023; accepted 24 June 2023. Date of publication 28 June 2023; date of current version 2 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62276089 and Grant 61906063; in part by the Natural Science Foundation of Hebei Province, China, under Grant F2021202064; in part by the Key Research and Development Project of Hainan Province, China, under Grant ZDYF2022SHFZ015; and in part by the AXA Research Fund. *(Corresponding author: Zhenghua Xu.)*

Jiaojiao Zhang, Shuo Zhang, and Zhenghua Xu are with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment and the Tianjin Key Laboratory of Bioelectromagnetic Technology and Intelligent Health, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin 300130, China (e-mail: 2195468073@qq.com; szhang5566@gmail.com; zhenghua.xu@hebut.edu.cn).

Xiaoqian Shen is with the Department of Computer Science, King Abdullah University of Science and Technology, Jedda 23955, Saudi Arabia (e-mail: 2695203287@qq.com).

Thomas Lukasiewicz is with the Institute of Logic and Computation, TU Wien, 1040 Vienna, Austria, and also with the Department of Computer Science, University of Oxford, OX1 3AZ Oxford, U.K. (e-mail: thomas.lukasiewicz@cs.ox.ac.uk).

Digital Object Identifier 10.1109/TMI.2023.3290356

I. Introduction

SUPERVISED deep learning has achieved some great successes in medical image segmentation tasks [1], [2], [3], [4], where a large amount of labeled data are presented. However, although a huge number of medical images are generated in daily clinical practice, annotating them is a highly professional task that can only be done by radiologists with extensive clinical experience. Due to the limited number, time, and annotating efficiency of professional radiologists, obtaining a large medical image dataset with accurate annotations is usually very difficult, which thus limits the use of supervised-learning-based segmentation approaches in daily clinical practice.

Consequently, recent research has witnessed several efforts in self-supervised medical image segmentation [5], [6], [7], [8], [9], where self-supervised learning is first conducted on a large amount of unlabeled data to learn the general features of medical images, then, by using the resulting model as the pre-trained model, fully supervised learning is further conducted on a small number of labeled data for fine-tuning. Therefore, the performance of self-supervised medical image segmentation relies heavily on the quality of self-supervised pre-training.

The recovery paradigm is a state-of-the-art self-supervised learning approach, where some strategies, e.g., rotation [10] and jigsaw puzzle [11], [12] are first applied to modify unlabeled images; then by using the resulting modified images as input and the original images as ground-truth, self-supervised learning is conducted to learn general image features via image recovery. However, existing recovery-paradigm-based self-supervised learning methods have the following two shortcomings. (i) *Domain shift problem*: Due to the use of artificial strategy modified images as the inputs of the pre-training network, in recovery paradigm-based methods, the input distributions of the upstream pre-training networks are usually different from the input distributions of the downstream segmentation networks, where original images instead of modified

images are used as inputs; consequently, the general features learned in the pre-trained models may not be applicable in the segmentation networks, which inevitably makes the fine-tuning more difficult and thus limits the segmentation performances. (ii) *Multimodality problem*: existing recovery-paradigm-based methods are mostly based on single-modal data and lack the capability to exploit multimodal information of medical images. Compared with single images, multimodality images are helpful to extract features from different views and bring complementary information, making the network have a better ability of segmentation.

Contrastive learning [13], [14], [15], [16], [17], [18], [19], [20] is another self-supervised pre-training approach, which uses a contrastive strategy to minimize the distance of pairs of similar images (that are usually obtained from the same original image using different data augmentation strategies) in the latent space while maximizing the distance of pairs of dissimilar ones. By applying the same set of data augmentation strategies in both upstream and downstream tasks, contrastive learning-based methods can have similar input distributions in both the upstream pre-training and the downstream segmentation networks; this thus overcomes the domain shift problem and makes their pre-trained models more applicable in the downstream segmentation networks than recovery paradigm-based methods. Consequently, contrastive learning based methods have demonstrated good performances in medical image segmentation tasks [9], [14]. However, the multimodality problem still exists.

In recent years, Multi-alignment [21] and ContIG [20] are proposed to utilize the multimodal information of medical images by multimodal contrastive semantic alignment. Specifically, Multi-alignment [21] first encodes images as spatial feature maps and then estimates the local similarities between images of different modalities to achieve multi-modal alignment by multimodal contrastive learning; consequently, the annotations of one modality are transferred to another modality as pseudo-labels to achieve unpaired multi-modal medical image segmentation using solely one set of annotations. ConIG [20] utilizes contrastive learning to align fundus images with multiple genetic modalities and then discover the cross-modal semantic relationships between images and genetic data.

Nevertheless, our studies show that solely using multi-modal contrastive learning still cannot fully utilize the fruitful multimodal information of medical images to achieve satisfactory performances. Therefore, in this paper, we propose a new multimodal contrastive self-supervised medical image segmentation method, called **Multi**modal **Con**trastive **Do**main **S**haring (**Multi-ConDoS**) generative adversarial networks, where pairwise multimodal medical images are adopted for self-supervised pre-training. Since these pairwise multimodal medical images are different imaging results of the same objects (e.g., lesions and/or organs inside human bodies), they should contain the semantic information of different aspects of the same object, which thus can be used together after proper registrations, to complement each other to learn more comprehensive features of the corresponding objects via domain translation. Generally, Multi-ConDoS are learned by a two-stage learning procedure. First, the novel *domain sharing generative adversarial networks (DSGANs)* are used to conduct multimodal contrastive self-supervised pre-training using a large number of unlabeled multimodal medical images. Then, the resulting modules are used to construct U-Net models for the downstream segmentation tasks, which are fine-tuned in a fully supervised way using solely a small amount of labeled medical images.

Specifically, different from Multi-alignment [21] and ContIG [20] that solely rely on multi-modal contrastive learning, the proposed Multi-ConDoS integrates domain translation and domain sharing techniques with multimodal contrastive learning to utilize the important multimodal information more comprehensively. First of all, DSGANs aim to utilize domain translation to learn the complementary mutual information of multimodal medical images for self-supervised pre-training. As we know, CycleGAN [22] and its variants [23] are widely-adopted domain translation approaches, which, however, are usually used for unpaired. Therefore, we propose to construct DSGAN by combining CycleGAN with the classic paired image translation model, Pix2Pix [24], where the GAN loss and the L1-based cross domain translation loss work together to better learn pixel-wise detailed information and features from paired images and reduce blurring.

Furthermore, in the process of domain translation, the networks have to learn not only domain-specific information representation, but also domain-sharing information representation; this is because, besides domain-specific features, the pairwise multimodal medical images also share a lot of general features (i.e., common features exist in both modalities), which are important learning objectives for self-supervised pre-training. However, the encoders in CycleGANs are mainly designed to learn the specific features of different domains; therefore, to effectively learn both specific and general features, a domain sharing technique is proposed in DSGAN by upgrading the traditional framework of CycleGAN with additional *domain sharing layers*. Finally, to further enhance DSGANs' feature learning capability, a *multimodal contrastive loss* is also used to maximize the similarities between features generated by multimodal images within the same pairs while minimizing the similarities of those within different pairs.

The contributions of this work can be summarized as follows. (i) We identify the shortcomings of the existing self-supervised medical image segmentation approaches, and propose a multimodal contrastive self-supervised medical image segmentation method, Multi-ConDoS, which utilizes a novel domain-sharing generative adversarial networks (DSGANs) to learn more comprehensive object features for self-supervised pre-training from multimodal medical images. (ii) There exist three advancements in DSGAN: First, DSGAN is a fusion of CycleGAN and the classic paired image translation model, Pix2Pix, so it can utilize both the cyclic learning strategy of CycleGAN and the cross domain translation loss of Pix2Pix to achieve better domain translation capability. Second, novel domain sharing layers are introduced to help DSGAN learn not only domain-specific but also domain-sharing information. Third, the multimodal contrastive loss is also used to better learn multimodal features. (iii) Extensive

experiments are conducted on two public multimodal medical image segmentation datasets. The experimental results show that, with only 5% (resp., 10%) of labeled data, Multi-ConDoS not only greatly outperforms the state-of-the-art self-supervised and semi-supervised medical image segmentation baselines with the same ratio of labeled data, but also achieves similar (sometimes even better) performances as fully supervised segmentation methods with 50% (resp., 100%) of labeled data, which thus proves that our work can achieve superior segmentation performances with very low labeling workload. In addition, ablation studies prove that the three improvements (i.e., the fusion of CycleGAN and Pix2Pix for domain translation, domain-sharing layers, and multimodal contrastive loss) are all effective and essential for Multi-ConDoS to achieve the very superior performance.

II. Related Work

A. Self-Supervised Learning (SSL)

SSL uses many unlabeled data to learn the general structural and anatomical representation. Then, the learned representation is applied to downstream tasks with a small amount of data to fine-tune [10], [11], [12], [25], [26]. For example, Rotation [10] encourages the model to learn visual representations by simply predicting the angle by which the input image is rotated. Jigsaw [11] derives a jigsaw puzzle grid from an input image and solves it to learn both a feature mapping of object parts and their correct spatial arrangement. Multimodality jigsaw (M-Jigsaw) [12], where puzzle pieces come from different modality images, is proposed to facilitate rich representation learning by confusing images at the data level. Different from the above recovery paradigm-based SSL methods, Multi-ConDoS keeps the inputs of the upstream and downstream network the same and avoids the domain shift problem. Then, we choose the above three recovery paradigm-based SSL methods (i.e., Rotation, Jigsaw, and M-Jigsaw), which have been successfully applied in medical image analysis, as our self-supervised learning baselines.

B. Contrastive Learning

Contrastive learning [13], [14], [15], [16], [17] enforces positive samples closer and negative samples further away in the latent space. Such methods are usually achieved by applying a contrastive loss [27]. Exemplar [28] trains with a triplet loss to achieve this goal. So, Exemplar can be considered as a simplified version of contrastive SSL. Aiming at different ways of collecting negative samples for positive samples during training, different architectures are developed from contrastive learning. SimCLR [14] trains an encoder to generate pairwise positive representations for different views of an input image, maximizing the similarity of the positive representations of the input image while minimizing the similarity to the representations of views from other images in the same batch. Different from SimCLR, BYOL [18] does not rely on negative samples, it replaces the contrastive loss with MSE loss and trains the online network to predict the target network representation of the same image under a different augmented view, achieving superior performance. Furthermore, SwAV [19] is proposed to utilize a clustering algorithm to cluster similar features together, i.e. the goal is not just to make a pair of samples close to each other, but also to ensure all features that are similar to each other are clustered together. Recently, contrastive learning has been successfully applied in medical image analysis [9], [29], [30] and other related fields [20]. For example, CPC [30], [31], which utilizes the idea of contrastive learning in the latent space, predicts the embedding of the next or adjacent sample. G-L [9] proposes to combine a local contrastive strategy (i.e., using 3D inter-layer location information for contrastive learning) with a global contrastive strategy to learn beneficial information representations. However, these designs are based on single-modal medical image data without considering the multimodality of medical images. Therefore, Multi-alignment [21] is then proposed to learn segmentation models for unpaired multi-modal medical images with solely a single annotation set, where a contrastive learning framework is proposed for multimodal image matching, and the segmentation results of one modality are transferred to the other modalities as pseudo-labels. Similarly, ContIG [20] also proposes to use multimodal contrastive learning to align the fundus image and several other genetic modalities in the feature space. Different from the existing methods, Multi-ConDoS is not only based on multimodal contrastive learning but also utilizes domain translation and domain sharing techniques to learn complementary mutual information of multimodal data, so it can utilize the important multimodal information more comprehensively. Finally, we choose the above contrastive learning SSL methods (i.e., Exemplar, SimCLR, BYOL, SwAV, CPC, G-L, ContIG) as the self-supervised contrastive learning baselines. Multi-alignment is not used as the baseline, because the inputs in our task are paired images with the same annotations so the very key pseudo-labels generation functionality of Multi-alignment is not applicable.

C. Semi-Supervised Learning

Semi-supervised learning is another paradigm to solve the problem of lacking of sufficient segmentation labels in medical image segmentation tasks [32]. Instead of learning in a two-stage way as self-supervised learning, semi-supervised models are learned using both a large amount of unlabeled data and a small amount of labeled data learning in a one-stage way. Semi-supervised learning has been successfully applied in the field of medical image segmentation [33], [34], [35]. For example, MT [33] proposes to use the averaging model weights method to construct a teacher model, and encourages the teacher model and the student model to have consistent predictions on the input data under different disturbances to achieve semi-supervised learning. SASS [35] proposes a shape-aware semi-supervised segmentation method, which implements geometric shape constraints through a signed distance map of object surfaces to improve the use of unlabeled data. DTML [34] proposes a dual-task mutual learning semi-supervised method to explore the useful knowledge of unlabeled data by generating target segmentation maps and regressing the signed distance maps. However, these methods

still are based on the single-modal state, without considering multimodality. Finally, since self-supervised learning and semi-supervised learning are two different types of approaches for the same problem, we choose the above three methods (i.e., MT, SASS, and DTML) as the semi-supervised learning baselines in our experiments to comprehensively demonstrate the superiority of the proposed Multi-ConDoS in solving the lacking of sufficient segmentation labels more.

III. Problem Setting

A. Fully-Supervised Medical Image Segmentation

Given a dataset D containing a large number (n) of medical image samples (denoted x) that are *all* associated with pixel-level segmentation masks (denoted $label$), i.e., $D = \{x_i, label_i\}_{i=1}^{n}$, fully-supervised medical image segmentation aims to learn a segmentation model that is capable of achieving accurate medical image segmentation performances using a large amount of well-labeled data.

B. Self-Supervised Medical Image Segmentation

Different from the fully-supervised settings, self-supervised medical image segmentation aims to achieve similar or even better segmentation performances than fully supervised solutions using only *a small ratio* of annotations, which will greatly boost the application of intelligent medical image segmentation systems in clinical practices because obtaining pixel-level segmentation masks for medical images not only requires a lot of professional knowledge but also is very time-consuming. The specific learning procedure of self-supervised medical image segmentation is as follows. Given a dataset, D_s with a large number (n) of medical image samples but only a small ratio (r) of them are associated with pixel-level segmentation masks, i.e., $D_s = \{\{x_i\}_{i=1}^{n}, \{label_j\}_{j=1}^{m}\}$ and $r = (m/n) \times 100\%$, self-supervised medical image segmentation first solely utilizes a large number of medical images (without using any label) to learn a pre-trained model (Net_{pre}) that contains general vision features of the medical images by self-supervised learning methods. Then a medical image segmentation model Net_{Seg} can be further obtained by fine-tuning Net_{pre} using the small number of medical images $\{x_j\}_{j=1}^{m}$ that contain pixel-level segmentation masks $\{label_j\}_{j=1}^{m}$.

C. Multimodal Self-Supervised Medical Image Segmentation

Since multimodal medical images contain fruitful and complementary information, in this work, we further apply multimodal medical images in the self-supervised medical image segmentation tasks with the aim of achieving accurate multimodal self-supervised medical image segmentation using a small ratio of annotations. Given a dataset D_m with a large number (n) of *pre-registered and pre-aligned multimodal medical image pairs* (denoted $x = \{x^a, x^b\}$) and a small ratio (r) of pixel-level segmentation masks, i.e., $D_m = \{\{x_i^a, x_i^b\}_{i=1}^{n}, \{label_j\}_{j=1}^{m}\}$ and $r = (m/n) \times 100\%$, multimodal self-supervised medical image segmentation aims to first solely utilize the paired medical image samples $\{x_i^a, x_i^b\}$ (without using any label) to learn *multiple pre-trained models* Net_{pre}^{a} and Net_{pre}^{b} (each of which corresponds to a modality and contains general vision features of the corresponding modality) by multimodal self-supervised learning methods. Then, the small portion of annotated multimodal medical images $\{x_j^a, x_j^b\}_{j=1}^{m}$ are further utilized together with their segmentation masks $\{label_j\}_{j=1}^{m}$ to fine-tune the corresponding pre-train models, i.e., annotated medical images $\{x_j^a, label_j\}_{j=1}^{m}$ (resp., $\{x_j^b, label_j\}_{j=1}^{m}$) are used to fine-tune Net_{pre}^{a} (resp., Net_{pre}^{b}); finally, *multiple medical image segmentation models* (i.e., Net_{seg}^{a} and Net_{seg}^{b}) are obtained, each of which is used for the specific segmentation task of the corresponding medical image modality.

IV. Methodology

Fig. 1 illustrates the overall structure of the proposed multimodal contrastive domain sharing (Multi-ConDoS) self-supervised medical image segmentation approach. Multi-ConDoS mainly consists of two processing steps. First, domain-sharing generative adversarial networks (DSGANs) are used to conduct multimodal contrastive self-supervised pre-training using a large number of unlabeled multimodal medical images. The resulting modules are then used to construct the classic U-Net models, which are trained in a fully supervised way with a small amount of labeled medical images to achieve the downstream segmentation tasks. Generally, DSGAN is a fusion of CycleGAN and the classic paired image translation model, Pix2Pix [24], with additional improvements (i.e., multimodal contrastive learning and shared layers). So, we can see DSGAN as either a paired translation extension of CycleGAN or a cyclic extension of Pix2Pix. The reasons for integrating CycleGAN with Pix2Pix for domain translation instead of solely using CycleGAN or Pix2Pix are as follows. First, comparing to the classic paired image translation models (e.g., Pix2Pix), CycleGAN's cyclic training strategy is very beneficial for fully and comprehensively learning modal feature information: Pix2Pix only learns the unidirectional mapping relationship of pairwise multimodal images, while CycleGAN's cyclic training strategy can learn the one-to-one bidirectional mapping relationship of pairwise multimodal images, which helps the generator network learn a more accurate latent representation space. Second, CycleGAN can simultaneously learn the feature information of two domains and conduct cross domain generation in two directions, which is the structural basis for setting the shared layer (SL) and introducing contrastive loss. Third, by introducing the cross domain translation loss L_T from Pix2Pix into DSGAN, similar to the paired image translation models (e.g., Pix2Pix), DSGAN can also learn the pixel-wise detailed information and features from paired images like. Consequently, combining CycleGAN with Pix2Pix makes DSGAN has the advantages of both models.

Specifically, in the pre-training step, DSGANs utilize a domain-sharing generator (DSG) to first take the original unlabeled medical images X (resp., Y) as inputs to generate images in the other domain, so we call this image generation process *image translation* and the resulting

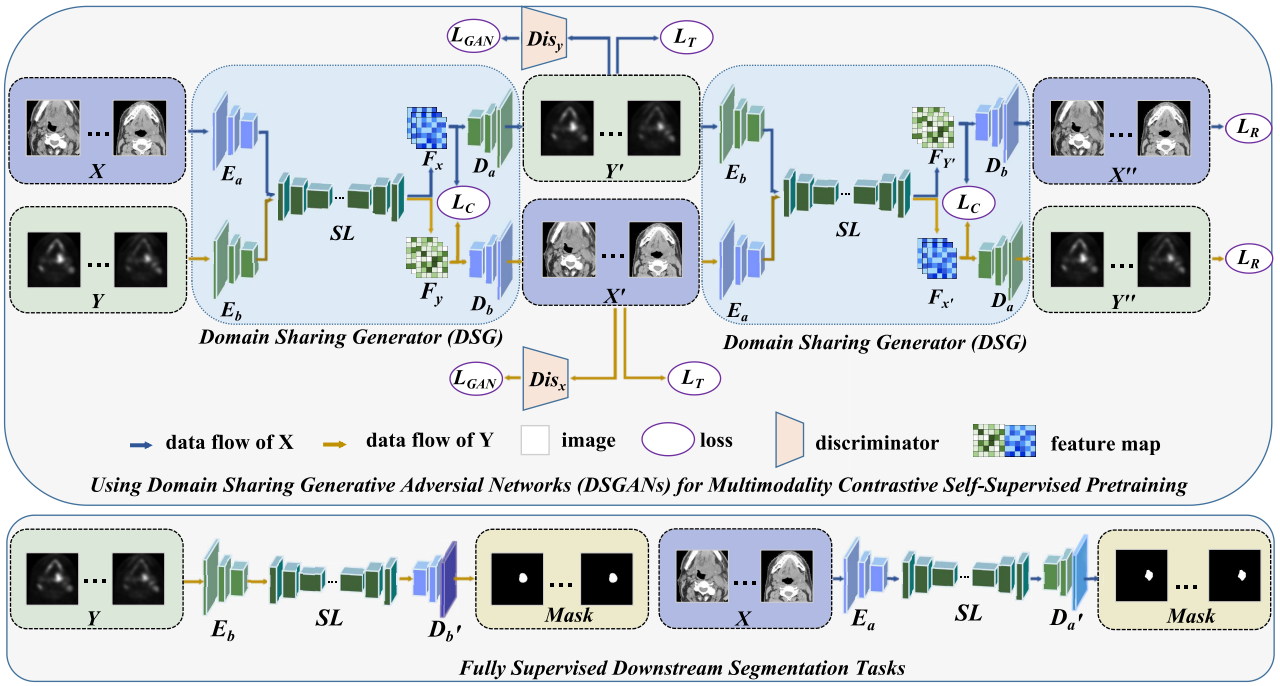


Fig. 1. Overview of multimodal contrastive domain-sharing (Multi-ConDoS) self-supervised medical image segmentation, where DSGANs are used to learn a bidirectional mapping of multimodal data using shared layers (SL) and contrastive losses (L_C), and the resulting modules are then used as pre-trained modules in the downstream segmentation task.

images Y' (resp., X') *translation images*. Then, similarly to CycleGANs [22], Y' and X' are further used as inputs of DSG to generate images X'' and Y'', respectively. Since X'' (resp., Y'') is generated from Y' (resp., X'), which is obtained from X (resp., Y), X'' (resp., Y'') can be seen as the *reconstructive images* of X (resp., Y). As shown in Fig. 1, the structure of DSG is similar to the generator of CycleGANs but uses shared layers (SL) to better capture the general features that commonly exist in both domains. Furthermore, two discriminators Dis_x and Dis_y are used to discriminate between the translated image X' (resp., Y') and the original input image X (resp., Y) to encourage the domain-sharing generator (DSG) to generate images that are more similar to the realistic original input images. Finally, the resulting modules of DSG are used in the fully supervised downstream segmentation tasks as pre-trained modules.

A. Domain-Sharing Generative Adversarial Networks

Generative adversarial networks (GANs) [22], [36], [37], [38], [39], [40] have an excellent performance in image-to-image translation. CycleGANs [22] use two separate generators to learn the bidirectional mapping of two domains by performing two cross-domain translations. Inspired by this, we propose Domain Sharing Generative Adversarial Networks (DSGANs), which learn a representation through bidirectional cross-domain learning and multimodal contrastive learning. Specifically, cross-domain learning can learn modality-specific knowledge, while the shared-layer (SL) and multimodal contrastive learning are designed to learn general knowledge of both modalities. As illustrated in Fig. 1, our framework consists of a generation module DSG: $\{G_a\,(E_a, SL, D_a)$: $X \Rightarrow Y$, $G_b\,(E_b, SL, D_b)$: $Y \Rightarrow X\}$ and two domain-specific discriminators (i.e., Dis_x and Dis_y).

The *discriminators* Dis_x and Dis_y aim to discriminate between real and translated images in the domains X and Y, thus facilitating the generator to produce more realistic images. The *generator* DSG aims to generate images that are as close to reality as possible, which contains two encoders $\{E_a, E_b\}$, a shared layers module (SL), and two decoders $\{D_a, D_b\}$. The two encoders (i.e., E_a and E_b) extract features of images from different modalities, and send the features of the input images into SL, so that the contents of the two domains are mapped to the same latent space. The contents encoded by SL for the two domains are then fed into their respective decoders (i.e., D_a and D_b).

In the process of cross-domain generation, E_a and E_b are domain-specific encoders, and they tend to learn domain-specific information. In fact, the extracted features actually contain domain-specific and domain-sharing information. To better learn the domain-sharing features and realize the mutual complement of multimodal information, we develop shared layers (SL) and multimodal contrastive loss. SL can receive the representations of two domains and map the features extracted from E_a and E_b to the same latent space. The intuitions of sharing layers are as follows: To achieve good domain translation, the networks have to learn not only domain-specific information, but also domain-sharing information (i.e., common features exist in both modalities). Consequently, with the design of sharing some layers, the proposed Multi-ConDoS will learn both domain-specific information (in E_a and E_b) and domain-sharing information

(in SL), i.e., although the design of sharing layers may result in less domain-specific features, it can help gain much more domain-sharing features. We believe this is not harmful but beneficial for domain translation, because, with the help of domain-sharing features, the network can comprehensively model not only the differences but also the similarities between different modalities. This is believed to be better than modeling solely the domain-specific features, as the excessive domain-specific information may make the model overly focus on the domain-specific characteristics and may lose the generalization to some extent, i.e., resulting in a kind of "domain-overfitting" problem. Therefore, we believe sacrificing an acceptable extent domain-specific information in exchange for better modeling of domain-sharing information in the proposed Multi-ConDoS will avoid the potential "domain-overfitting" problem and enhance the model's generalization, thus its domain translation performances will be increased.

B. Multimodal Contrastive Loss

However, sharing the same latent space does not mean that SL encodes the consistency information of the paired image features of the two domains. Therefore, the contrastive loss is used to minimize (respectively, maximize) the distance between paired (respectively, unpaired) images to highlight the important domain-sharing information. Our multimodal contrastive loss is based on the contrastive loss [14] that is shown to achieve a state-of-the-art performance in many cases. We begin with a general form of the contrastive loss as given in Equation (1), and then introduce our *multimodal contrastive loss*.

$$L_{i,j} = -\log \frac{e^{\mathrm{sim}(\hat{x}_i,\hat{y}_j)/t}}{e^{\mathrm{sim}(\hat{x}_i,\hat{y}_j)/t} + \sum_{\tilde{x}\in\Lambda^-} e^{\mathrm{sim}(\hat{x}_i,\tilde{x})/t}},$$
$$\hat{x}_i = g\left(f\left(x_i\right)\right), \hat{y}_j = g\left(f\left(y_j\right)\right), \tag{1}$$

where, for a mini-batch $\{z_i\}_{i=1}^N$, $\{x_i, y_i\}_{i=1}^N$ are different views of the same image, as a similar pair, and Λ^- is a set that is not similar to x_i, as dissimilar pairs. sim is defined as cosine similarity, g and f are non-linear mappings, and t denotes the temperature parameter. Minimizing the loss l increases the similarity between similar pairs, while increasing the dissimilarity between dissimilar pairs. The *multimodal contrastive loss* is defined as follows:

$$l(x_i, y_i; \theta; t) = -\log \frac{e^{\mathrm{sim}\left(F_{x_i}(\theta), F_{y_i}(\theta)\right)/t}}{e^{\mathrm{sim}\left(F_{x_i}(\theta), F_{y_i}(\theta)\right)/t} + \sum_{F_z\in\Gamma^-} e^{\mathrm{sim}\left(F_{x_i}(\theta), F_z(\theta)\right)/t}}, \tag{2}$$

$$L_C(G;\theta;t) = \frac{1}{\left|\Gamma^+\right|} \sum_{\forall(x_i,y_i)\in\Gamma^+} \left[l\left(x_i, y_i; \theta; t\right) + l\left(y_i, x_i; \theta; t\right)\right] + \frac{1}{\left|\Gamma'^+\right|} \sum_{\forall(x_i',y_i')\in\Gamma'^+} \left[l\left(x_i', y_i'; \theta; t\right) + l\left(y_i', x_i'; \theta; t\right)\right], \tag{3}$$

where Γ^+ and Γ'^+ are sets of all similar pairs. We form a batch by randomly sampling N pairs of images $\{x_i, y_i\}_{i=1}^N$. In our model, we use the encoders $\{E_a, E_b\}$ and SL to extract the feature representations $\{Fx_i, Fy_i\}$ and $\{Fx_i\prime, Fy_i\prime\}$ of pairs of different modality views $\{x_i, y_i\}$ and $\{x_i\prime, y_i\prime\}$, respectively. Multimodal medical images are characterized by the use of different ways to capture the contents of the same tissue area, so that different modality views corresponding to the same area can be considered similar. We used the feature representations of different modality views extracted for contrastive loss calculation. Here, $\{Fx_i, Fy_i\}_{i=1}^N$ and $\{Fx_i\prime, Fy_i\prime\}_{i=1}^N$ are similar pairs, while $\left\{Fx_i, Fy_j\right\}_{i\neq j}$ and $\left\{Fx_i\prime, Fy_j\prime\right\}_{i\neq j}$ are dissimilar pairs, where $Fx_i \in F_X$, $Fy_i \in F_Y$, $Fx_i\prime \in F_X'$, and $Fy_i\prime \in F_Y'$. The optimization goal of contrastive loss is to make pairs of modal view features as close as possible in the latent space, while unpaired modal view features as far as possible.

Since the multimodal contrastive loss is based on cosine similarity, which is usually applied to one-dimensional feature vectors, so, in Multi-ConDoS, the resulting two-dimensional feature maps are flattened to one-dimensional feature vectors from the dimension $dim = 1$ before they are used to compute the contrastive loss. Actually, the operations of converting two-dimensional feature maps into one-dimensional feature vectors to make them applicable for the cosine similarity based contrastive loss are widely observed in the existing contrastive learning based image processing works, e.g., SimCLR [14], ContIG [20], and Multi-alignment [21], where non-linear projection heads are usually adopted for the conversion. However, we note that using non-linear projection heads may result in the loss of detailed information, which is harmful to the pixel-wise segmentation tasks. Therefore, flatten operations are directly used in Multi-ConDoS for the dimension conversion to avoid information loss.

C. Other Losses

Besides the multimodal contrastive loss, several other losses are also used in DSGANs for bidirectional cross-domain learning. Since the goal of the discriminators, Dis_X and Dis_Y, is to tell the differences between real and translation images, while the generator, DSG, aims to generate realistic images, an adversarial *GAN loss* $\mathcal{L}_{\mathrm{GAN}}$ is defined as:

$$\mathcal{L}_{\mathrm{GAN}}\left(G, Dis_x, Dis_y\right) = \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\log Dis_y(y)\right] + \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\log\left(1 - Dis_y(G_a(x))\right)\right] + \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\log Dis_x(x)\right] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\log\left(1 - Dis_x(G_b(y))\right)\right]. \tag{4}$$

In addition, integrating the GAN loss with L1 loss is reported to be beneficial for reducing blurring [24] and helping the model learn pixel-wise detailed information and features from paired images; therefore, similar to Pix2Pix [24], a L1-based *translation loss* is further used to minimize differences between the input and translation images. Formally,

$$\mathcal{L}_{\mathrm{T}}(G) = \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\|G_a(x) - y\|_1\right] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\|G_b(y) - x\|_1\right]. \tag{5}$$

Finally, a *reconstructive loss* is applied to minimize the distance between the reconstructed image X'' (resp., Y'') and the input image X (resp., Y). Formally,

$$\mathcal{L}_{\mathrm{R}}(G) = \mathbb{E}_{x\sim p_{\mathrm{data}}(x)}\left[\|G_b(G_a(x)) - x\|_1\right] + \mathbb{E}_{y\sim p_{\mathrm{data}}(y)}\left[\|G_a(G_b(y)) - y\|_1\right]. \quad (6)$$

Consequently, the *complete learning loss* of DSGANs is:

$$\mathcal{L}\left(G, Dis_x, Dis_y\right) = \mathcal{L}_{\mathrm{GAN}}\left(G, Dis_x, Dis_y, X, Y\right) + \gamma\mathcal{L}_{\mathrm{T}}(G) + \beta\mathcal{L}_{\mathrm{R}}(G) + \lambda\mathcal{L}_{\mathrm{C}}(G, \theta, t), \quad (7)$$

where γ, β, and λ are the coefficients of $\mathcal{L}_{\mathrm{T}}(G)$, $\mathcal{L}_{\mathrm{R}}(G)$, and $\mathcal{L}_{\mathrm{C}}(G, \theta, t)$, respectively.

The learning objective of DSGANs is as follows:

$$G^* = \arg\min_{G}\max_{Dis_x, Dis_y}\mathcal{L}\left(G, Dis_x, Dis_y\right). \quad (8)$$

D. Multi-ConDoS Self-Supervised Medical Image Segmentation

DSGANs first uses unlabeled bimodal data in the training set to learn the beneficial representations of multimodal images. To learn feature representations comprehensively, we preserve all slices that contain the corresponding organs (i.e., only discard medical images that are all black). Since the last layer of the network is highly task-related, we then transfer the weights of E_a (resp., E_b), SL, and the first three layers of D_a (resp., D_b), to the downstream task $X \Rightarrow label$ (resp., $Y \Rightarrow label$), where the resulting new decoder (with the first three layers pre-trained and the last layer random initialized) is denoted as D'_a (resp., D'_b). Then, we fine-tune two entire segmentation networks separately using a portion (5% and 10%) of labeled data, where only the slices that contains the target objects (i.e., tumors) and the corresponding labels are used.

V. Experiments and Results

Datasets: For the evaluation of the proposed approach, we use two publicly available multimodal datasets. (i) The *Hecktor* dataset [41], [42] was released by the Hecktor challenge hosted at MICCAI 2020 for head and neck tumor segmentation. It contains 201 3D head and neck CT-PET scans. (ii) The *BraTS2018* dataset [43], [44], [45] was released by the BraTS'18 challenge hosted at MICCAI 2018 for segmenting brain tumor, including WT (whole tumor), ET (enhancing tumor), and TC (tumor core). All BraTS scans include four MRI modalities: T1, T1CE, T2, and FLAIR volumes, in this work, we divide this dataset into two parts: the multimodal volumes of T1CE and T2 are first paired and used together as the multimodal inputs of Multi-ConDoS, then those of FLAIR and T1 are grouped and used together. The training set for the entire dataset included magnetic resonance imaging (MRI) scans of different qualities for high-grade gliomas (HGG) and low-grade gliomas (LGG). We only use 210 3D MRI HGG in our experiments. In the self-supervised pre-training stage, all slices containing the corresponding organs are preserved (i.e., only the images that are all black will be discarded); in the downstream stage, since the labels of the segmenting objects (i.e,, tumors in our work) are required for fully supervised fine-tuning, only the slices that contain the target objects (i.e., tumors) and the corresponding labels are preserved.

Pre-Processing: We apply the following pre-processing steps: (i) re-sampling of all volumes and corresponding labels to a fixed pixel size $1.0 \times 1.0 \times 1.0\ mm^3$ using nearest-neighbour interpolation, intensity normalization of each 3D volume, clip the value into the range [1, 99] and normalize the image with mean and std for region nonzero. (ii) all 2D images and corresponding 2D labels are obtained from the z-axis of 3D volumes. The dimension d for each dataset are: (a) Hecktor: $d = 144 \times 144$, and (b) BraTS2018: $d = 160 \times 160$. Note that the medical images on both Hecktor (with CT and PET modalities) and BraTS2018 (T1, T1ce, T2 and FLair modalities) datasets are natively registered when they are generated from the imaging equipments; as for other multimodal medical images, the pre-processing of image registration should be conducted before sending them into Multi-ConDoS for pre-training to ensure the proper matching of the paired images.

Experimental Setup: In our experiments, the multimodal medical image datasets Hecktor and BraTS2018 are separately divided into two subsets, i.e., the training set and the testing set. We first use all the images in the training set to learn the upstream pre-training model without using any annotations; then only a small ratio (i.e., 5% and 10%) of labeled images are used together with their labels to fine-tune the resulting pre-trained model and obtain the final segmentation model. Finally, the testing set is used to evaluate the performances of the segmentation models.

Specifically, the detailed setting information for the two datasets is as follows. **Hecktor** contains 201 pairs of CT and PET volumes belonging to 201 patients, where 180 pairs (i.e., 90%) of CT-PET volumes (25923 paired image slices in total) are divided into the training set, and 21 pairs (i.e., 10%) of CT-PET volumes (3026 paired image slices in total) are in the testing set. Similarly, **BraTS2018** contains 210 cases of multimodal MRI volumes (i.e., T1CE, T2, FLAIR, and T1 modalities), which are divided into 168 training cases (80%) and 42 testing cases (20%), with respectively 26040 and 6510 images for each modality. In the pre-training stage, all images in the training set are used without annotations, while only 5% or 10% of annotated images in the training set are used together with their labels for fine-tuning.

Implementation Details: Our Multi-ConDoS is implemented based on Torch 1.6.0 and CUDA-10.1. All experiments are done on 8 GeForce RTX 2080 GPUs. For self-supervised learning, the Adam [46] optimizer is used, with a learning rate of 0.0002. The temperature parameter t is set to 0.1. Since the values of different losses have different orders of magnitudes (i.e., L_T is 10^{-2}, L_R is 10^{-1}, L_C is 10^0), to avoid the dominant effect, we use the coefficient to erase their magnitude differences and make them have relatively close importance. So, the weights of the translation, reconstructive, and multimodality contrastive loss (i.e., L_T, L_R, and L_C) are 10, 1, and 0.1, respectively. The batch size is 48 for

TABLE I
NETWORK ARCHITECTURE OF ONE BRANCH OF THE DOMAIN SHARING GENERATOR (DSG) WITH DIFFERENT SETTINGS OF SL

Net Layer		Output Shape	SL_{1u}^{1d}	SL_{2u}^{2d}	SL_{3u}^{3d}
Input layer	Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	(f, h, w)	not share	not share	not share
Downsampling layer1	MaxPool2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 2, h/2, w/2)$	not share	not share	not share
Downsampling layer2	MaxPool2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 4, h/4, w/4)$	not share	not share	share
Downsampling layer3	MaxPool2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 8, h/8, w/8)$	not share	share	share
Downsampling layer4	MaxPool2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 16, h/16, w/16)$	share	share	share
Upsampling layer1	ConvTranspose2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 8, h/8, w/8)$	share	share	share
Upsampling layer2	ConvTranspose2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 4, h/4, w/4)$	not share	share	share
Upsampling layer3	ConvTranspose2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	$(f \times 2, h/2, w/2)$	not share	not share	share
Upsampling layer4	ConvTranspose2d Conv2d, BatchNorm2d, ReLU Conv2d, BatchNorm2d, ReLU	(f, h, w)	not share	not share	not share
Output layer	Conv2d	$(1, h, w)$	not share	not share	not share

Hecktor and 36 for BraTS2018. And a total of 200 epochs is trained.

For transfer learning, the U-Net [47] for downstream segmentation tasks is also trained with the Adam optimizer. The initial learning rate is 0.0002, the weight decay is 0.0001, and the learning rate strategy is warmup-cosine-lr. When using 5% labels, the batch size is set to 31 for Hecktor and 56 for BraTS2018. When using 10% labels, the batch size is set to 90 for Hecktor and 70 for BraTS2018. To make the network reach convergence, 70 epochs are trained for Hecktor and BraTS2018.

Network Architecture: We use the U-Net, which includes four downsampling and upsampling modules, as a segmentation backbone network for all methods. Each downsampling module consists of two 3×3 convolutions and a 2×2 max-pooling with stride 2, while each upsampling module consists of two 3×3 convolutions and a 2×2 transposed convolution with stride 2. This U-Net network also serves as the backbone network of DSG. The detailed network structure of DSG with different shared layer strategies is shown in Table I, where f, h, and w are the number of filters, the height of the input images, and the width of the input images, respectively.

Evaluation: Two widely used evaluation metrics for medical image segmentations are used, i.e.,: Dice similarity coefficient (DSC) and sensitivity (Sen). The definitions are as follows: $\text{DSC} = 2 * TP/(FP + 2 * TP + FN)$ and $\text{Sen} = TP/(TP + FN)$, where TP, FP, TN, and FN are True Positive, False Positive, True Negative, and False Negative, respectively. The value range of DSC and Sen is [0, 1], and the higher the value of these indicators, the better. All assessment metrics were calculated using each patient's 3D scan and then averaged as final results.

Baselines: To evaluate the performance of Multi-ConDoS, randomly initialized U-Nets without self-supervised pre-training, i.e., fully supervised learning from scratch (denoted **Fully Supervised**), using 5% and 10% annotations are selected as our original baselines.

Several state-of-the-art self-supervised learning methods applied in the field of medical image segmentation are chosen as the *self-supervised learning baselines* in our experiments, namely, **Rotation** [10], **Jigsaw** [11], **M-Jigsaw** [12], **Exemplar** [28], **CPC** [30], **SimCLR** [14], **BYOL** [18], **SwAV** [19], **G-L** [9], and **ContIG** [20]. We evaluate the quality of representations learned from different self-supervised methods in the following way. Transfer the model weights derived from different self-supervised methods to several downstream tasks, then fine-tune with different ratios of annotated data, and finally measured their impact in Section IV-A.

Since semi-supervised learning is another paradigm to address the scarcity of medical labels, to demonstrate the superiority of our method, we not only compare our methods with the self-supervised SOTA methods but also

TABLE II
RESULTS OF THE PROPOSED MULTI-CONDOS AND THE BASELINES ON **HECKTOR(CT, PET)** AND **BRATS2018(T1CE, T2, FLAIR, T1)** DATASETS

Ratios	Methods	Hecktor				BraTS2018							
		CT		PET		T1CE		T2		FLAIR		T1	
		DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen
5%	**Fully Supervised Learning from Scratch with Partial Labeled Data**												
	Fully Supervised	0.1740	0.2337	0.5452	0.7044	0.3234	0.2703	0.3952	0.4371	0.4035	0.4837	0.1743	0.2702
	Self-Supervised Learning Baselines												
	Rotation [10]	0.1542	0.1918	0.5525	0.7382	0.4239	0.3837	0.3718	0.3704	0.4156	0.6323	0.2218	0.3804
	Jigsaw [11]	0.2119	0.2599	0.5145	0.6366	0.3564	0.3138	0.3988	0.4055	0.4341	0.6042	0.1967	0.2779
	M-Jigsaw [12]	0.1886	0.2324	0.5814	0.7083	0.4083	0.3485	0.3702	0.3770	0.4463	0.6233	0.1997	0.2830
	Exemplar [28]	0.1872	0.2394	0.5924	0.6646	0.3525	0.3666	0.3850	0.4028	0.4325	0.6137	0.1711	0.1980
	CPC [30]	0.2102	0.2520	0.5784	0.6557	0.4452	0.3808	0.3842	0.4003	0.3421	0.5469	0.2152	0.3955
	SimCLR [14]	0.2201	0.3113	0.5720	0.7143	0.5403	0.5737	0.4610	**0.5870**	0.4342	0.5396	0.2908	0.4376
	BYOL [18]	0.1967	0.2555	0.5780	0.6879	0.5245	0.5590	0.4487	0.5603	0.4561	0.6297	0.2811	0.4545
	SwAV [19]	0.2186	0.3793	0.5517	0.7091	0.3252	0.4321	0.4160	0.5564	0.4346	0.6052	0.2277	0.4466
	G-L [9]	0.1754	0.1719	0.5615	0.6993	0.3195	0.4130	0.4412	0.5281	0.4355	0.5603	0.2137	0.2917
	ContIG [20]	0.1755	0.3208	0.5458	0.7452	0.3112	0.3597	0.4217	0.5389	0.4261	0.5586	0.1859	0.2524
	Semi-Supervised Learning Baselines												
	MT [33]	0.2340	0.3295	0.5959	0.6957	0.5317	0.4976	0.4491	0.5718	0.4238	0.4880	0.2489	0.3241
	DTML [34]	0.2857	0.3219	0.5819	0.7040	0.4889	0.4730	0.4159	0.4646	0.3868	0.3992	0.2223	0.2080
	SASS [35]	0.2604	0.3119	0.5923	0.7079	0.4841	0.4108	0.4197	0.4524	0.4380	0.4980	0.2061	0.2112
	The Proposed Solution												
	Multi-ConDoS	**0.3025**	**0.4293**	**0.6430**	**0.7488**	**0.5752**	**0.5747**	**0.4730**	0.5469	**0.4569**	**0.6340**	**0.3161**	**0.5335**
10%	**Fully Supervised Learning from Scratch with Partial Labeled Data**												
	Fully Supervised	0.2541	0.2875	0.5769	0.7067	0.4451	0.3696	0.4288	0.4716	0.4489	0.6225	0.2476	0.3287
	Self-Supervised Learning Baselines												
	Rotation [10]	0.2447	0.3026	0.6102	0.7339	0.4940	0.4321	0.4261	0.4485	0.4365	0.6138	0.2901	0.4336
	Jigsaw [11]	0.2818	0.3598	0.5860	0.6643	0.5416	0.4879	0.4680	0.5387	0.4520	0.5981	0.2914	0.3570
	M-Jigsaw [12]	0.2916	0.3396	0.6111	0.7440	0.5037	0.4364	0.4707	0.5053	0.4504	0.6149	0.2993	0.3810
	Exemplar [28]	0.2797	0.3237	0.5960	0.7195	0.4914	0.4579	0.4366	0.4834	0.4678	**0.6742**	0.2472	0.3967
	CPC [30]	0.2672	0.2880	0.5952	0.6695	0.5429	0.4937	0.4882	0.5540	0.3997	0.5481	0.3026	0.3828
	SimCLR [14]	0.2951	0.3895	0.5779	0.7773	0.6024	0.6217	0.4846	0.5729	0.4720	0.5217	0.3551	0.4868
	BYOL [18]	0.3013	0.3930	0.5891	0.7927	0.5991	0.6323	0.4850	0.5710	0.4763	0.6329	0.3458	0.3535
	SwAV [19]	0.2550	0.3340	0.5771	0.7804	0.4515	0.5646	0.4587	**0.5794**	0.4543	0.6184	0.2914	0.4694
	G-L [9]	0.2784	0.2945	0.5923	0.7304	0.4575	0.5045	0.4645	0.5718	0.4504	0.6691	0.2685	0.4086
	ContIG [20]	0.2572	0.3742	0.5728	0.7405	0.4464	0.4472	0.4346	0.5446	0.4509	0.6072	0.2793	0.3899
	Semi-Supervised Learning Baselines												
	MT [33]	0.2802	0.2467	0.6347	0.7067	0.6021	0.5580	0.4689	0.5510	0.4372	0.4897	0.3497	0.4496
	DTML [34]	0.3129	0.3669	0.6253	0.7251	**0.6423**	0.6191	0.4897	0.5022	0.4067	0.4404	0.3129	0.2930
	SASS [35]	0.2924	0.4586	0.6251	0.6993	0.6413	0.6034	0.4813	0.4736	0.4382	0.5242	0.3373	0.3481
	The Proposed Solution												
	Multi-ConDoS	**0.3442**	**0.4970**	**0.6488**	**0.8012**	0.6246	**0.6436**	**0.5045**	0.5566	**0.4805**	0.6281	**0.3749**	**0.6149**
50%	Fully Supervised	0.3114	0.3574	0.6252	0.6582	0.5941	0.5324	0.5094	0.5328	0.4830	0.5664	0.3581	0.4564
100%	Fully Supervised	0.3927	0.4736	0.6475	0.7307	0.7453	0.7204	0.5556	0.5575	0.5164	0.6561	0.4207	0.4923

with several state-of-the-art semi-supervised learning methods applied in medical image segmentation, namely, **MT** [33], **DTML** [34], and **SASS** [35], which are called the *semi-supervised learning baselines*. Finally, we also show the fully supervised results using large ratios (50% and 100%) of annotations.

All baselines are implemented and run using similar procedures and settings as those in their original papers, where all self-supervised methods (including our Multi-ConDoS method and the self-supervised baselines) are learned in a two-stage learning ways, i.e., first pre-trained using only unlabeled data and then fine-tuned using solely labeled data, while all the semi-supervised baselines (MT, DTML, and SASS) are directly trained using both labeled and unlabeled data in a one-stage learning way; and additional parameter adjustments are made to our best efforts.

A. Main Results

To investigate the effectiveness of Multi-ConDoS, we conduct experiments on two datasets and compare the performance of Multi-ConDoS to three state-of-the-art baselines: Fully Supervised Baseline (i.e., Fully Supervised), Self-Supervised Baselines (i.e., Rotation, Jigsaw, M-Jigsaw, Exemplar, CPC, SimCLR, BYOL, SwAV, G-L, and ContIG), and Semi-Supervised Baselines (i.e., MT, DTML, SASS). For a fair comparison, we use the same backbone network (U-Net) with 5% and 10% annotations across all methods. The experimental results are shown in Table II. A visualization

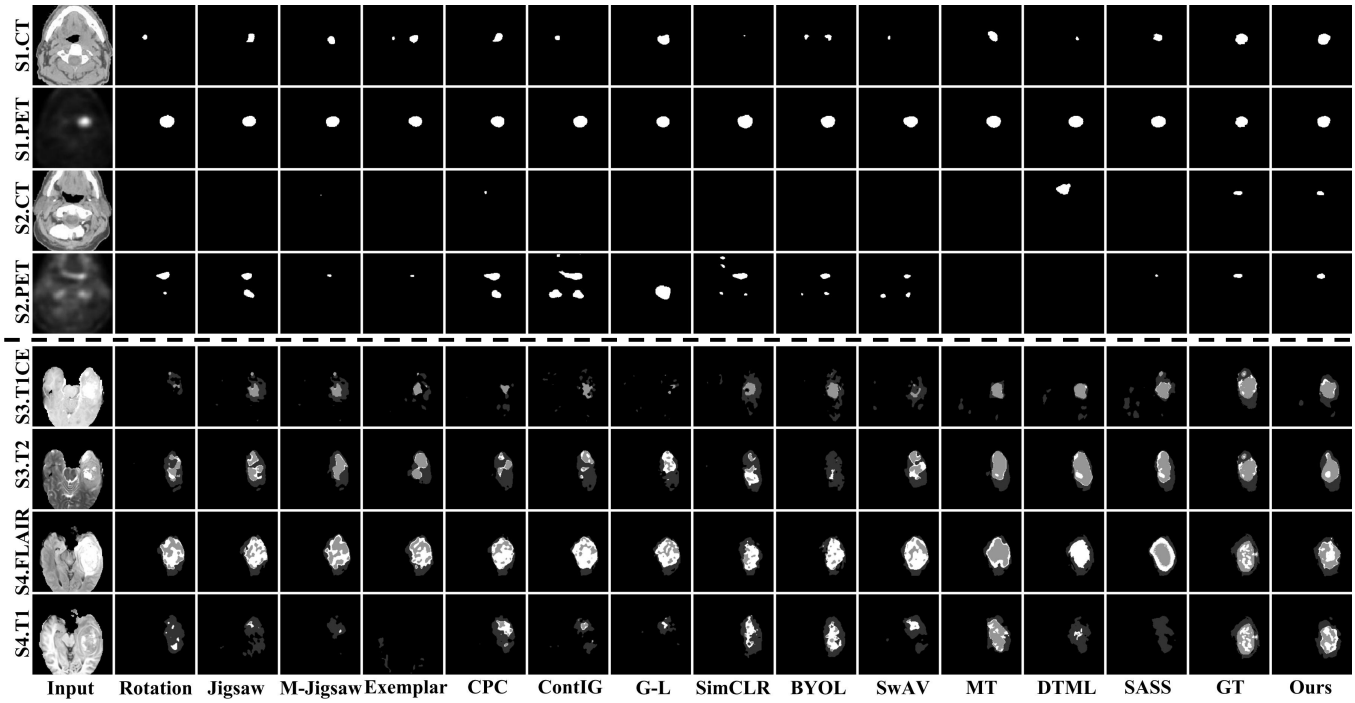


Fig. 2. Visualized segmentation results on the Hecktor and BraTS2018 datasets with 10**%** labeled data. In the first four rows, S1 and S2 are two paired samples in the Hecktor dataset, where the head and neck tumor (only one type of tumor) is marked with white in the segmentation results. In the last four rows, S3 and S4 are two paired samples in the BraTS2018 dataset, where WT (whole tumor), ET (enhancing tumor), and TC (tumor core) are marked with light gray, medium gray and white, respectively, in segmentation results. As observed, our segmentation results are closer to ground truth and better than other methods *in both two modalities*.

of the segmentation results of Multi-ConDoS and the self-supervised and semi-supervised baselines on the Hecktor and BraTS2018 databases is shown in Fig. 2.

As shown in Table II, Multi-ConDoS generally outperforms all the state-of-the-art self-supervised and semi-supervised medical image segmentation baselines in terms of all evaluation metrics and two small label-ratio settings (i.e., 5% and 10%) on both datasets, and its performances are similar to (and sometimes even better than) the fully supervised solution using much higher ratios (i.e., 50% and 100%) of label data. This observation proves that the proposed Multi-ConDoS can achieve superior medical image segmentation performances using only a small number of annotations, which thus greatly reduces the labeling workload of applying intelligent medical image segmentation systems in clinical practices. The detailed analysis is as follows.

1) Compare With Fully Supervised Learning From Scratch: As shown in Table II, both self-supervised learning and semi-supervised learning methods (including the proposed Multi-ConDoS) achieve much better segmentation performances than the fully supervised baseline with the same ratio of labeled data in also cases on both datasets. This is because besides of the small amount of labeled data, self-supervised and semi-supervised methods can mine additional useful information from the large amount of unlabeled data.

Furthermore, Multi-ConDoS generally outperforms the baseline model that is fully supervised learning from scratch by a large margin with 5% and 10% annotations. The performance of Multi-ConDoS with 5% annotations can be close to or even better than that of the fully supervised method with 50% annotations. As for 10% annotations, Multi-ConDoS not only outperforms the fully supervised method with 50% annotations but also outperforms the fully supervised method with 100% annotations in the PET modality on Hecktor. This is because Multi-ConDoS can use the complementary mutual information of multimodality data for self-supervised pre-training (i.e., using the additional information in one modality to enhance the feature learning of another), while this can not be achieved by the fully supervised work. This argument is also supported by the observation in Table II that the similar the multimodal data, the smaller the improvement gaps between Multi-ConDoS and the fully supervised baseline (e.g., the T1CE and T2 modalities on BraTS2018 are more similar than the CT and PET modalities on Hecktor, so the improvement gaps on BraTS2018 is generally smaller).

2) Compare With Self-Supervised Learning Baselines: Then, we further compare our Multi-ConDoS with the state-of-the-art self-supervised methods, i.e., Rotation, Jigsaw, M-Jigsaw, Exemplar, CPC, SimCLR, BYOL, SwAV, G-L, and ContIG. We can see that our Multi-ConDoS also significantly outperforms these methods in almost all cases on both datasets. Consequently, This proves our argument that by integrating domain translation and domain sharing techniques with multi-modal contrastive learning to achieve mutual complementation of modal information, Multi-ConDoS is capable to learn more comprehensive and fruitful information and features from the unlabeled multimodal data than the SOTA self-supervised contrastive learning baselines, and thus achieves better medical image segmentation performances.

In addition, we have observed that our method's Sen results may not always the best in some scenarios. This can be attributed to the challenge of distinguishing between lesions and the background, particularly at the edges where they appear similar. To address this issue, we plan to explore attention mechanisms [48] in the future to enable the model to focus more on foreground lesions.

3) Compare With Semi-Supervised Learning Baselines: Since semi-supervised learning is another paradigm to address the problem of lacking labeled data, to demonstrate the superiority of our method, we also compare our work with the SOTA semi-supervised baselines. The results in Table II shows that our method generally outperforms semi-supervised baselines in all evaluation cases on both datasets. This is because, compared with single-modality-based semi-supervised methods, our method adequately conducts the mutual complementation of multimodal beneficial information to achieve superior performance.

In summary, Table II exhibits that Multi-ConDoS generally achieves better segmentation performances than the fully supervised, self-supervised, and semi-supervised baselines on all modalities of two datasets in terms of DSC and Sen metrics. This is because (i) Multi-ConDoS keeps the same inputs of pre-training and segmentation networks, avoiding the domain shift problem, (ii) Multi-ConDoS utilizes multimodal medical images to learn more comprehensive object features via contrastive domain translation, and (iii) Multi-ConDoS utilizes novel domain sharing generative adversarial networks (DSGANs) to achieve a contrastive domain translation and to learn both specific and general features more effectively using domain-sharing layers and multimodal contrastive loss.

4) Analysis of Visualized Segmentation Results: Moreover, all the above findings are also well supported by the visualized results in Fig. 2, where Multi-ConDoS achieves obviously better (i.e., more similar to the ground-truth) segmentation results than all the self- and semi-supervised medical image segmentation methods. Please note that, although there are some algorithms with similar results to ours in a specific modality on a certain dataset, as our work is based on multi-modal segmentation and our goal is to get better segmentation results in both modalities, we find that there is not any existing methods having segmentation results that are similar to ours on both modalities, e.g., the result of G-L (resp., M-Jigsaw) has a good segmentation in S1.PET (resp., S1.PET), the segmentation result is slightly worse in S1.CT (resp., S1.CT).

Specifically, the segmentation results of the CT and PET modality of Hecktor at the first and second rows of Fig. 2 show that: (i) the segmentation results of M-Jigsaw are better than Jigsaw, (ii) the segmentation results of M-Jigsaw and Multi-ConDoS are better than others, and (iii) the segmentation performance of Multi-ConDoS is better than the self- and semi-supervised learning methods baselines.

Similarly, from the segmentation results of T1CE, T2, FLAIR and T1 modality of BraTS2018 at the third to sixth rows of Fig. 2, we can see that the segmentation results of the proposed Multi-ConDoS is best among all methods. Therefore, these visualized observations clearly demonstrate again that by the proposed domain-sharing generative adversarial network and multimodal contrastive learning, Multi-ConDoS remedies the drawbacks of the existing self-supervised medical image segmentation methods, and achieves a much better performance in medical image segmentation tasks with a small amount of annotations.

B. Ablation Studies

Since there are three improvements, i.e., domain translation, multimodal contrastive learning and domain sharing layers, ablation studies are conducted to verify the effectiveness of these advanced components. Specifically, we have implemented six intermediate models for self-supervised pre-training: (i) Con-Only means the self-supervised pre-training using only multimodal contrastive learning loss; (ii) P-GAN denotes the pre-training using the classic paired image translation model, Pix2Pix [24]; (iii) Cycle-GAN means the pre-training using the vanilla CycleGAN [22] to achieve domain translation; (iv) Cycle-P-GAN means the domain translation based pre-training is achieved by the combination of vanilla CycleGAN and Pix2Pix (i.e., the cyclic extension of Pix2Pix or the CycleGAN variant with pairwise images and L_T loss); (v) Multi-DoS denotes the pre-training using a simplified Multi-ConDoS without multimodal contrastive loss (i.e., extending Cycle-P-GAN based domain translation with domain sharing layers); (vi) Multi-Con is the pre-training using a simplified Multi-ConDoS, where the originally shared layers are trained separately but not shared, (i.e., extending Cycle-P-GAN based domain translation with multimodal contrastive loss). The ablation studies are conducted on Hecktor and BraTS2018 datasets using 5% and 10% ratios of annotations, and the corresponding results are shown in Table III. On the other hand, to study the effectiveness of different losses in our method, we also conduct loss ablation experiments on two datasets with 5% labeled data as shown in Table V.

1) Effectiveness of Self-Supervised Pre-Training: We first compare the results using Con-Only, P-GAN, and Cycle-GAN for pre-training with those of the fully supervised baselines that are training from scratch. The results show that Con-Only, P-GAN and CycleGAN all greatly outperform the fully supervised baseline in medical image segmentation tasks. This thus proves that multimodal contrastive learning, Pix2Pix-based domain translation, and CycleGAN-based domain translation are all effective for self-supervised pre-training, which will enhance the performances of self-supervised medical image segmentation using only a small number of labeled data.

2) Effectiveness of Combining P-GAN and Cycle-GAN for Domain Translation: Then, we further evaluate the effectiveness of three potential domain translation solutions in Multi-ConDoS, i.e., P-GAN, Cycle-GAN, and Cycle-P-GAN. As shown in Table III, the results of pre-training using Cycle-P-GAN for domain translation are generally better than that of solely using P-GAN or Cycle-GAN. This is because (i) comparing to the classic paired image translation model, Pix2Pix, CycleGAN's cyclic training strategy is very beneficial for fully and comprehensively learning modal feature information: Pix2Pix only learns the unidirectional mapping relationship of pairwise multimodal images, while CycleGAN's cyclic train-

TABLE III
RESULTS OF OUR ABLATION STUDIES ON THE HECKTOR AND BRATS2018 DATASETS

Ratios	Methods	Hecktor				BraTS2018							
		CT		PET		T1CE		T2		FLAIR		T1	
		DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen
5%	Fully Supervised	0.1740	0.2337	0.5452	0.7044	0.3234	0.2703	0.3952	0.4371	0.4035	0.4837	0.1743	0.2702
	Con-Only	0.1943	0.2859	0.5493	0.6606	0.3623	0.4108	0.4072	0.4599	0.4105	0.5611	0.1779	0.3263
	P-GAN	0.2132	0.2930	0.5465	0.6814	0.4575	0.4428	0.4117	0.4731	0.4167	0.5748	0.2255	0.3330
	Cycle-GAN	0.2254	0.2801	0.5598	0.7805	0.4502	0.4462	0.4144	0.4620	0.4238	0.6141	0.2341	0.3377
	Cycle-P-GAN	0.2279	0.2733	0.5630	**0.7936**	0.4788	0.4339	0.4128	0.4679	0.4259	0.6167	0.2363	0.3699
	Multi-DoS	0.2503	0.4277	0.5986	0.7852	0.4946	0.4648	0.4248	0.4993	0.4401	0.6035	0.2557	0.3792
	Multi-Con	0.2977	0.4219	0.6214	0.7681	0.5621	0.5640	0.4302	0.5405	0.4237	0.6153	0.2999	0.4301
	Multi-ConDoS (OURS)	**0.3025**	**0.4293**	**0.6430**	0.7488	**0.5752**	**0.5775**	**0.4730**	**0.5469**	**0.4569**	**0.6340**	**0.3161**	**0.5335**
10%	Fully Supervised	0.2541	0.2875	0.5769	0.7067	0.4451	0.3696	0.4288	0.4716	0.4489	0.6225	0.2798	0.3934
	Con-Only	0.2634	0.3648	0.5862	0.7773	0.5146	0.4993	0.3994	0.4432	0.4504	0.6238	0.2802	0.3518
	P-GAN	0.2602	0.3354	0.5953	0.7804	0.5296	0.5087	0.4194	0.4363	0.4514	0.6154	0.2806	0.3422
	Cycle-GAN	0.2607	0.3222	0.5964	0.8075	0.5235	0.5103	0.4222	0.4617	0.4527	0.6329	0.2832	0.3478
	Cycle-P-GAN	0.2715	0.2829	0.5936	**0.8158**	0.5393	0.5010	0.4213	0.4618	0.4561	0.6369	0.3166	0.4254
	Multi-DoS	0.2819	0.4849	0.6261	0.8036	0.5558	0.5247	0.4504	0.5051	0.4622	0.6230	0.3183	**0.4901**
	Multi-Con	0.3315	0.4580	0.6269	0.7783	0.5807	0.6264	0.4616	0.5241	0.4730	**0.6653**	0.3436	0.4759
	Multi-ConDoS (OURS)	**0.3442**	**0.4970**	**0.6488**	0.8012	**0.6246**	**0.6436**	**0.5045**	**0.5566**	**0.4805**	0.6281	**0.3586**	0.4871

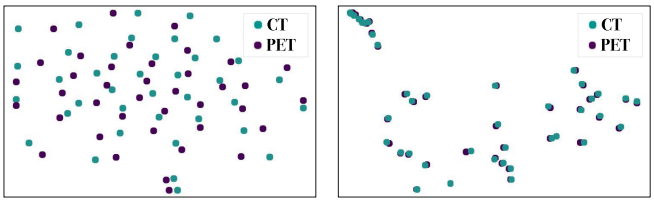
TABLE IV
QUANTITATIVE SIMILARITY RESULTS OF TRANSLATION IMAGES (I.E., X' AND Y') AND RECONSTRUCTION IMAGES (I.E., X'' AND Y'') W.R.T. REAL IMAGES (X AND Y) IN TERMS OF FID AND SSIM ON HECKTOR AND BRATS2018

Modalities	Methods	Trans.X'		Trans.Y'		Rec.X''		Rec.Y''	
		FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑	FID↓	SSIM↑
Hecktor (CT-PET)	P-GAN	79.2516	0.5976	64.4343	0.9909	-	-	-	-
	Cycle-GAN	81.0649	0.5904	80.9264	0.9898	119.4201	0.5508	64.4062	0.9925
	Cycle-P-GAN	76.4773	0.5966	62.5630	0.9925	112.6794	0.5611	60.2657	0.9925
	Multi-DoS	75.8879	0.5968	61.2264	0.9925	107.8398	0.5616	59.7154	0.9925
	Multi-Con	75.1041	**0.5972**	60.4564	0.9925	109.5743	0.5624	59.4457	**0.9926**
	Multi-ConDoS (OURS)	**74.8872**	0.5971	**60.3395**	**0.9926**	**107.0693**	**0.5662**	**58.3923**	0.9925
BraTS2018 (T1CE-T2)	P-GAN	74.4959	0.6263	87.9765	0.5789	-	-	-	-
	Cycle-GAN	86.7709	0.6165	98.2861	0.5612	114.0237	0.5519	71.1314	0.6556
	Cycle-P-GAN	74.4878	0.6373	89.9744	0.5739	52.1376	0.6288	52.1073	0.6613
	Multi-DoS	73.6858	0.6446	88.3821	0.5863	51.0662	0.6177	50.8883	**0.6698**
	Multi-Con	73.4416	0.6421	87.6225	0.5843	50.0329	0.6223	52.0842	0.6602
	Multi-ConDoS (OURS)	**73.1257**	**0.6616**	**82.8712**	**0.5987**	**47.4418**	**0.6441**	**50.1753**	0.6620
BraTS2018 (FLAIR-T1)	P-GAN	39.7750	0.7592	41.3167	0.7654	-	-	-	-
	Cycle-GAN	40.3078	0.7088	47.0860	0.7606	34.0731	0.7219	46.1486	0.7652
	Cycle-P-GAN	33.1999	0.7900	43.1002	0.7734	32.3369	0.7915	43.4657	0.7752
	Multi-DoS	32.0383	0.8029	**39.8608**	**0.7961**	31.3009	0.8061	42.0886	0.7854
	Multi-Con	28.9621	0.7974	44.8223	0.7757	28.7323	0.8009	45.2007	0.7745
	Multi-ConDoS (OURS)	**27.9031**	**0.8052**	41.7130	0.7845	**27.5386**	**0.8076**	**41.6759**	**0.7922**

ing strategy can learn the one-to-one bidirectional mapping relationship of pairwise multimodal images, which helps the generator network learn a more accurate latent representation space; (ii) by integrating Pix2Pix with CycleGAN, similar to the paired image translation models, the resulting Cycle-P-GAN can also learn the pixel-wise detailed information and features. Therefore, this proves that integrating P-GAN and Cycle-GAN in Multi-ConDoS for domain translation is sound and effective for the model to learn the features of multimodal medical images more comprehensively.

3) Effectiveness of Domain Sharing Layers: Furthermore, as shown in Table III, Multi-Dos (extending Cycle-P-GAN with domain sharing layers) is generally better than Cycle-P-GAN, and Multi-ConDoS is better than Multi-Con. This is because the domain sharing layers are incorporated into Multi-Dos and Multi-ConDoS make the models' capable of learning not only domain-specific information but also domain-sharing features from multimodal medical images. The effectiveness of using domain sharing layers is thus demonstrated.

4) Effectiveness of Multimodal Contrastive Loss: Finally, in Table III, the results of Con-Only, Multi-Con, and Multi-ConDoS are generally better than those of the fully supervised baseline, Cycle-P-GAN, and Multi-DoS, respectively. This is because the multimodal contrastive loss can help the deep models learn valuable mutual consistency information



(a) Multi-DoS (b) Multi-ConDoS

Fig. 3. Visualization of the learned feature embeddings of Multi-DoS and Multi-ConDoS on the Hecktor dataset. The closer the CT (green circle) and the corresponding PET (purple circle) are embedded, the better the learned modality-shared features.

of multimodal medical images. Therefore, it well demonstrates the effectiveness of using multimodal contrastive loss in Multi-ConDoS.

In addition, the effectiveness of multimodal contrastive loss can be visualized in Fig. 3, where the learned embedding of the paired modal (CT, PET) features are shown. We randomly select 40 pairs of images from the Hecktor dataset. The features of these pairs of images are mapped into the same latent space through the shared layers (SL), and the feature representation is obtained. Then, the feature dimension is reduced to 2 by t-SNE [49]. The closer the CT (green circle) and the corresponding PET (purple circle) embedding, the better modality-shared features are learned. We observe that, compared to Multi-DoS, the paired modal feature embedding in the Multi-ConDoS latent space is closer. Therefore, the use of the multimodal contrastive loss effectively enhances the learning of modality-shared features.

5) Analysis of Cross-Domain Image Translation Capability: To further demonstrate that all three proposed improvements (i.e., combining Pix2Pix with Cycle-GAN, domain-sharing layers, and multimodal contrastive loss) are effective and essential for Multi-ConDoS to achieve superior cross-domain image translation, additional experiments are further conducted to evaluate and compare the similarities between the synthesized translation images (resp., reconstruction images), i.e., X' and Y' (resp., X'' and Y''), generated by the proposed Multi-ConDoS and five GAN-based intermediate models (i.e., P-GAN, Cycle-GAN, Cycle-P-GAN, Multi-Dos, and Multi-Con), and the corresponding real images, i.e., X and Y.

Two widely used image quality evaluation metrics, Fréchet Inception Distance (FID) [50] and Structural Similarity Index (SSIM) [51], are used to measure the quantitative similarity values, where FID measures the statistical distribution distances between real and generated images in the feature space, and SSIM estimates the structural similarities between the real and generated images. Consequently, with the increase of similarities between the real and generated images, the FID values decrease and the SSIM values increase.

As shown in Table IV, we have the following observations. First, by comparing the results of P-GAN, Cycle-GAN, and Cycle-P-GAN, we find that the results of Cycle-P-GAN constantly outperforms those of P-GAN and Cycle-GAN; this finding further proves our conclusion that integrating P-GAN and Cycle-GAN in Multi-ConDoS for domain translation is sound and effective for the model to learn the features of multimodal medical images more comprehensively, and capable to enhance the model's cross-domain image translation capability. Second, we also notice that, by importing the design of sharing layers, Multi-DoS generally outperforms Cycle-P-GAN, while the results of Multi-ConDoS are generally better than those of Multi-Con; this finding greatly supports our argument that although using the shared layers may lose some domain-specific information, this will not weaken but enhance the model's capability in cross-domain image translation; this is because the shared layers sacrifice an acceptable extent domain-specific information in exchange for better modeling of domain-sharing information to help the model comprehensively learn not only the differences but also the similarities between different modalities, which thus avoid the potential "domain-overfitting" problem and enhance the model's generalization. Third, we also observe that the results of Multi-ConDoS (resp., Multi-Con) are generally better than those of Multi-ConDoS (resp., Cycle-P-GAN); this observation asserts that the proposed multimodal contrastive loss can also boost the generation models' cross-domain generation ability because it is helpful for the models to better learn the differences and similarities between different domains. Finally, it is obvious that the proposed Multi-ConDoS generally outperforms all the GAN-based intermediate generation models in terms of both evaluation metrics on both datasets; this demonstrates our conclusion that all three proposed improvements (i.e., Combining Pix2Pix with Cycle-GAN, domain-sharing layers, and multimodal contrastive loss) are effective and essential for Multi-ConDoS to achieve superior cross-domain image translation.

Fig. 4 shows the visualized examples of Multi-ConDos and five GAN-based intermediate in cross-domain image translation, where the examples on Hecktor (CT-PET) (in Fig. 4 (a)) and BraTS2018 (T1CE-T2) (in Fig. 4 (b)) are medical images with lesions, while those on BraTS2018 (FLAIR-T1) (in Fig. 4 (c)) are images without lesion. Generally, we can have the observations similar to Table IV in Fig. 4. For example, in Fig. 4 (a) and (b), compared with Cycle-P-GAN (resp., Multi-Con), Multi-DoS (resp., Multi-ConDoS) Can generate medical images whose lesion areas are more similar to those of real images; furthermore, in Fig. 4 (c), comparing to the real images, the medical images generated by Multi-DoS (resp., Multi-ConDoS) have more accurate texture and boundary details than those generated by Cycle-P-GAN (resp., Multi-Con); these thus proves the effectiveness of sharing layers in obtaining better cross-domain image translation. Besides these, the effectiveness of integrating Pix2Pix with Cycle-GAN, and multimodal contrastive loss can also be clearly observed and demonstrated in Fig. 4.

6) Effectiveness of Different Losses: To study the influence of different losses in our method, we have additionally conducted loss-based ablation studies on Hecktor and BraTS2018 datasets with 5% labeled data, and the corresponding results are shown in Table V.

Specifically, Multi-ConDoS mainly consists of four types of loss functions, i.e., multi-modal contrastive loss (L_C), translation loss (L_T), adversarial GAN loss (L_{GAN}), and reconstructive loss (L_R); L_C here is different from the

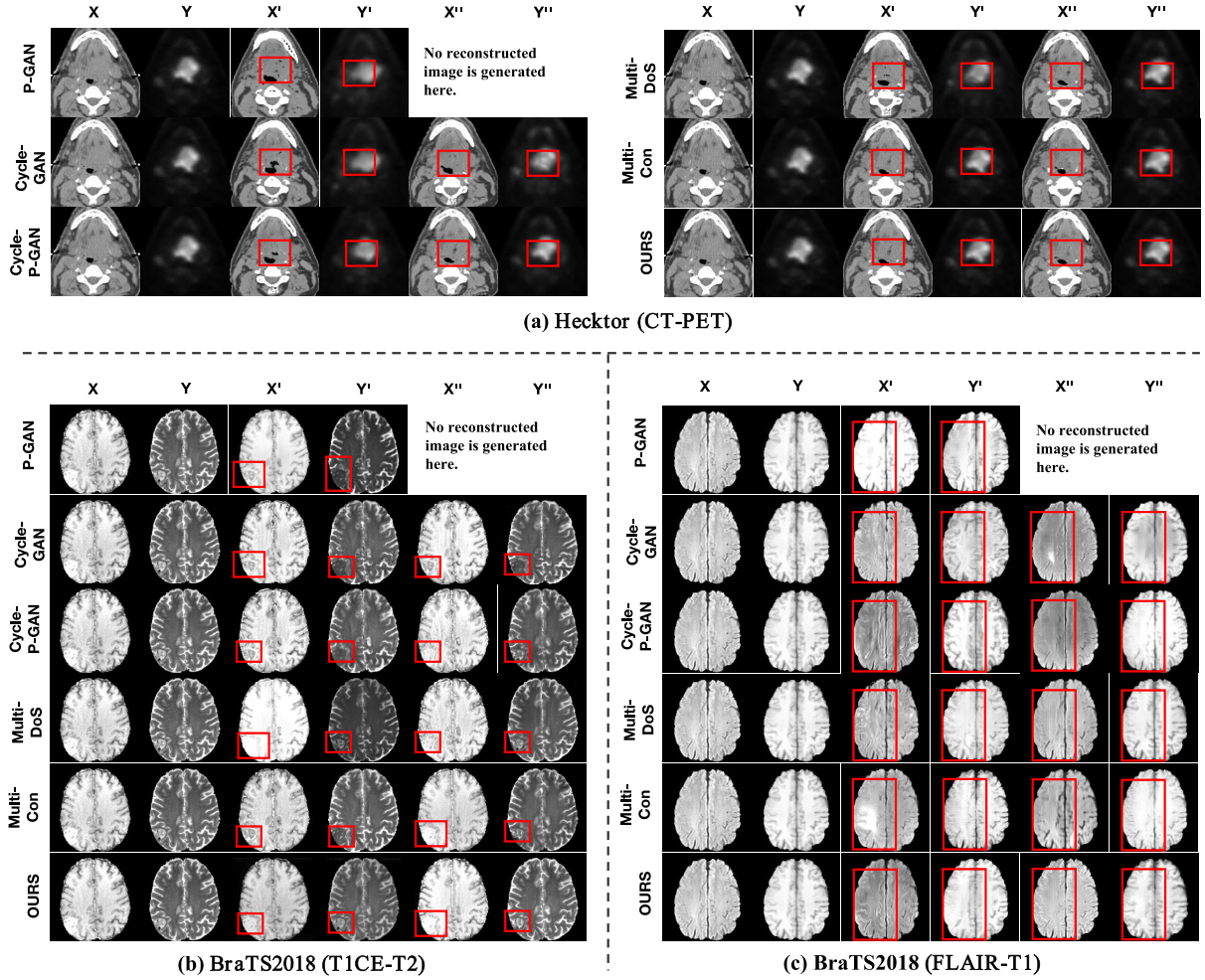


Fig. 4. Visualized cross-domain image generation results of Multi-ConDoS and five GAN-based intermediate models on Hecktor and BraTS2018, where translation images X' (resp., Y') are generated from real images Y (resp., X), and we then use X' (resp., Y') to generate reconstruction images Y'' (resp., X''). As P-GAN is a unidirectional instead of cyclic generative model, it generates only translation images but no reconstruction images. The areas included in red boxes are regions with relatively significant differences between images (resp., X', Y', X'' and Y'') generated by different methods, so we box these areas to make it easier to distinguish the different image generation capabilities of different models.

Con-Only in Table III because the model structure is kept unchanged in the loss-based ablation studies, i.e., there are shared layers for all cases in Table V (including L_C) but not for Con-Only. Therefore, we first apply only a loss function for the pre-training of deep models using unlabeled data; as shown in Table V, the results of using L_C, L_T or L_{GAN} for model pre-training are much higher than those of training from scratch, e.g., using L_C, L_T or L_{GAN} can respectively improve the DSC score of T1ce by 0.049, 0.0366, or 0.05. The reasons for improvements are as follows: L_C, L_T, and L_{GAN} help the networks learn the consistency information between different modalities, the specific details of the images, and the domain features of different modalities, respectively. Please note that L_R can not be added solely because L_T or L_{GAN} is needed to constrain the cross-domain generation of DSGAN and bring the domain-transfer capability to the deep models.

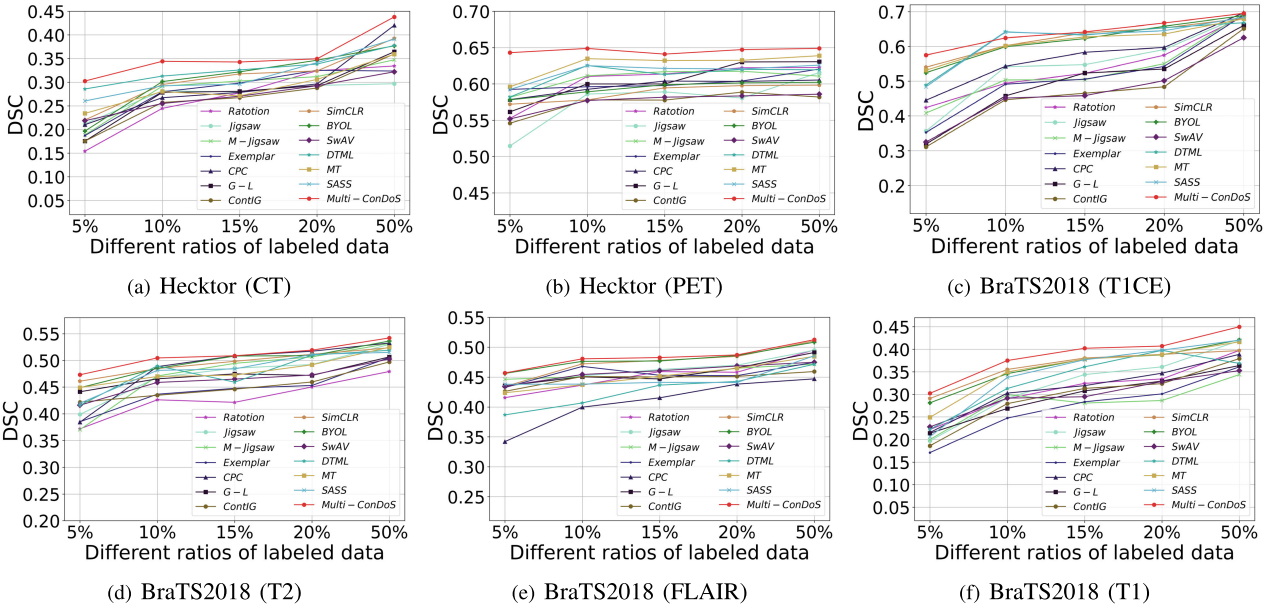
Then, additional experiments are further conducted to investigate the medical image segmentation performances of using the combinations of two loss functions (similarly, it is meaningless to only use L_C+L_R because L_T or L_{GAN} is needed for domain-transfer). We can observe in Table V that the cases of using two loss functions together generally outperform those of using the corresponding two losses solely in terms of all modalities on both datasets, e.g., the segmentation results of using $L_{GAN}+L_C$ are generally better than those of solely using L_{GAN} or L_C. This proves that combining different types of loss functions can enhance the models' segmentation performances.

Third, we study the segmentation performances of all possible combinations of using three types of loss functions, whose results are generally better than those of two-losses combinations, e.g., $L_{GAN}+L_R+L_C$ generally outperforms $L_{GAN}+L_C$ because L_R can enhance the networks' feature learning capability by enforcing the reconstructed image to retain as much information of the original image as possible.

Finally, we also notice that the segmentation performances of three-losses combinations are all worse than those of our Multi-ConDoS where all four types of loss functions are used. Consequently, all the above findings demonstrate that the four types of loss functions used in Multi-ConDoS are all effective and essential for the model to achieve the superior

TABLE V
RESULTS OF LOSS-BASED ABLATION STUDIES ON THE HECKTOR AND BRATS2018 DATASETS WITH 5% LABELED DATA

Losses	Hecktor				BraTS2018							
	CT		PET		T1CE		T2		FLAIR		T1	
	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen
Training from scratch	0.1740	0.2337	0.5452	0.7044	0.3234	0.2703	0.3952	0.4371	0.4035	0.4837	0.1743	0.2702
L_C	0.2177	0.3356	0.5713	0.7097	0.3724	0.4287	0.4137	0.5180	0.4140	0.5863	0.2299	0.3565
L_T	0.2098	0.2802	0.5544	0.7112	0.3600	0.4432	0.4122	0.5116	0.4269	0.5771	0.2221	0.3328
L_{GAN}	0.2132	0.2930	0.5465	0.6814	0.3734	0.4220	0.4037	0.5134	0.4216	0.5683	0.2255	0.3330
$L_T + L_R$	0.2280	0.3264	0.5672	0.7135	0.4027	0.4378	0.4211	0.5204	0.4302	0.5671	0.2252	0.3407
$L_T + L_C$	0.2237	0.3434	0.5716	0.7138	0.4125	0.4543	0.4344	0.5392	0.4307	0.5861	0.2557	0.3792
$L_{GAN} + L_R$	0.2200	0.3584	0.5822	0.7281	0.4097	0.4237	0.4340	0.5213	0.4377	0.5812	0.2363	0.3699
$L_{GAN} + L_C$	0.2256	0.3655	0.5693	0.7149	0.4168	0.4800	0.4359	0.5253	0.4370	0.5950	0.2848	0.3769
$L_{GAN} + L_T$	0.2553	0.3294	0.5711	0.7256	0.4279	0.4536	0.4336	0.5049	0.4246	0.6111	0.2498	0.3706
$L_T + L_R + L_C$	0.2337	0.4076	0.5737	0.7188	0.4317	0.5390	0.4445	0.5447	0.4303	0.6159	0.3005	0.4398
$L_{GAN} + L_R + L_C$	0.2268	0.4018	0.5836	0.7239	0.5057	0.5501	0.4533	0.5454	0.4433	**0.6415**	0.3059	0.4717
$L_{GAN} + L_T + L_R$	0.2458	0.4157	0.5850	0.7241	0.5003	0.5297	0.4478	0.5442	0.4401	0.6035	0.2991	0.4209
$L_{GAN} + L_T + L_C$	0.2977	0.4219	0.6117	0.7361	0.5117	0.5553	0.4712	0.5293	0.4373	0.6115	0.3082	0.4245
$L_{GAN} + L_T + L_R + L_C$ (OURS)	**0.3025**	**0.4293**	**0.6430**	**0.7488**	**0.5752**	**0.5775**	**0.4730**	**0.5469**	**0.4569**	0.6340	**0.3161**	**0.5335**



(a) Hecktor (CT) (b) Hecktor (PET) (c) BraTS2018 (T1CE)

(d) BraTS2018 (T2) (e) BraTS2018 (FLAIR) (f) BraTS2018 (T1)

Fig. 5. Variation curves of the segmentation results of our Multi-ConDoS and the self-supervised and semi-supervised baselines on different ratios (5%, 10%, 15%, 20%, and 50%) of labeled data.

performances in self-supervised medical image segmentation tasks.

C. Additional Experiments

1) Effectiveness of The Ratios of Labeled Data: Additional experiments are conducted with more ratios of labeled data to investigate the effect of the ratios of labeled data on the segmentation performances of Multi-ConDoS and the self-supervised and semi-supervised baselines on Hecktor (CT, PET) and BraTS2018 (T1CE, T2, FLAIR, T1). The results are shown in Fig. 5, containing five different ratios (5%, 10%, 15%, 20%, and 50%) of labeled data.

We have the following observations: (i) The segmentation performances of Multi-ConDoS and the baselines generally increase with the rise of the ratios of the labeled data, this is because the increase of the ratios of labeled data will gradually bring more and more strong supervision information to the downstream fine-tuning, which thus enhances the performances of all models. (ii) Multi-ConDoS generally outperforms all the baselines in terms of all cases, which proves that the self-supervised pre-training using Multi-ConDoS is so good that, even with more and more strong supervision information, it can still bring lots of valuable information to boost the final segmentation performances. (iii) The improvement gaps between Multi-ConDoS and the baselines gradually reduce with the increase of the ratios, which proves that the increase of the ratios gradually reduces the importance of self-supervised pre-training using unlabeled data (so the advantage of using Multi-ConDoS is gradually weakened).

2) Effectiveness in Additional Evaluation Metrics: Three additional evaluation metrics, i.e., 95% Hausdorff Distance (HD95) [52], boundary IoU (BIoU) [53], and positive predictive value (PPV) [54], are used to measure the performances of Multi-ConDoS and the self-supervised and semi-supervised

TABLE VI
RESULTS OF MULTI-CONDOS AND THE BASELINES IN TERMS OF BIOU AND HD95 WITH 10% LABELED DATA

Methods	Hecktor				BraTS2018							
	CT		PET		T1CE		T2		FLAIR		T1	
	BIoU	HD95	BIoU	HD95	BIoU	HD95	BIoU	HD95	BIoU	HD95	BIoU	HD95
Fully Supervised Learning from Scratch with Partial Labeled Data												
Fully Supervised	0.1224	64.0604	0.3199	41.8627	0.1643	50.9107	0.1510	43.9483	0.1724	30.4918	0.0042	59.5963
Self-Supervised Learning Baselines												
Rotation [10]	0.1382	58.8788	0.3465	53.5804	0.1720	36.8101	0.1719	36.8101	0.1861	30.6972	0.0564	44.1426
Jigsaw [11]	0.1558	42.5337	0.3313	59.5274	0.2062	38.3538	0.2061	38.3538	0.2105	30.7971	0.0834	43.4044
M-Jigsaw [12]	0.1727	41.6987	0.3474	51.6693	0.2235	35.8980	0.2235	35.8980	0.1949	32.7322	0.0925	30.4919
Exemplar [28]	0.1600	44.0192	0.3401	37.3441	0.1965	45.3473	0.1965	45.3473	0.2055	31.4018	0.0604	37.9360
CPC [30]	0.1795	42.5362	0.2998	40.9564	0.1926	49.7371	0.1510	43.9483	0.2081	28.8120	0.0734	37.4265
SimCLR [14]	0.1504	**37.4345**	0.3581	66.4539	0.2758	37.6094	0.2020	38.0898	0.2203	28.5869	**0.1247**	46.2637
BYOL [18]	0.1501	47.2453	0.3397	55.0181	0.2901	36.1710	0.1820	42.4852	0.2075	27.9066	0.1201	31.8026
SwAV [19]	0.1648	59.5009	0.3673	50.8417	0.2287	56.0846	0.1889	44.0658	0.2036	38.6023	0.1035	61.8490
G-L [9]	0.1389	61.0124	0.3380	20.8051	0.2328	49.7115	0.1782	34.8908	0.2122	38.9321	0.0891	57.4882
ContIG [20]	0.1472	58.2753	0.3283	27.4594	0.2120	47.4633	0.1846	49.3227	0.2182	34.8885	0.0964	51.4963
Semi-Supervised Learning Baselines												
MT [33]	0.1653	55.4624	0.3253	39.0457	0.2092	42.7724	0.2092	42.7724	0.1814	32.9597	0.0709	31.0762
DTML [34]	0.1665	48.1990	0.3307	44.4936	**0.3449**	36.2121	0.1943	37.1866	0.1828	28.3425	0.0861	34.3141
SASS [35]	0.1663	46.7244	0.3539	38.5084	0.3086	35.7435	0.1976	35.9732	0.1989	27.9396	0.0624	37.0571
The Proposed Solution												
Multi-ConDoS	**0.1738**	40.9652	**0.4277**	**25.8142**	0.3110	**33.8717**	**0.2134**	**34.3922**	**0.2321**	**26.7961**	0.0962	**30.8364**

TABLE VII
THE PPV RESULTS OF OUR METHOD AND THE SELF- AND SEMI- SUPERVISED BASELINES ON HECKTOR AND BRATS2018 WITH 10% LABELED DATA

Methods	Hecktor		BraTS2018			
	CT	PET	T1CE	T2	FLAIR	T1
Fully Supervised Learning from Scratch						
Fully Supervised	0.2548	0.4371	0.4136	0.4300	0.4081	0.2428
Self-Supervised Learning Baselines						
Rotation [10]	0.2263	0.4824	0.6599	0.4593	0.4153	0.2526
Jigsaw [11]	0.3007	0.6282	0.6731	0.4792	0.4171	0.3182
M-Jigsaw [12]	0.3125	0.5345	0.6279	0.5042	0.4241	0.2947
Exemplar [28]	0.3152	0.5854	0.6014	0.4367	0.4003	0.2456
CPC [30]	0.3237	0.5365	0.5167	0.4097	0.4260	0.3076
SimCLR [14]	0.3114	0.5235	0.6545	0.4853	0.5039	0.3459
BYOL [18]	0.3106	0.5400	0.6429	0.4293	0.4564	0.3058
SwAV [19]	0.2669	0.5386	0.4405	0.4333	0.4404	0.2562
G-L [9]	0.3184	0.6009	0.5004	0.4738	0.4052	0.2452
ContIG [20]	0.2637	0.5441	0.5388	0.3929	0.4285	0.2760
Semi-Supervised Learning Baselines						
MT [33]	0.2467	0.5297	0.5611	0.4645	0.4944	0.3465
DTML [34]	0.3422	0.5622	0.6613	0.5384	0.4985	0.4400
SASS [35]	0.2856	0.5574	0.6555	0.4580	0.4880	0.3947
OURS	**0.3500**	**0.6408**	**0.6846**	**0.5438**	**0.5214**	**0.4485**

baselines. We choose these three evaluation metrics because (i) Hausdorff Distance (HD) is a widely used distance-based metric, here we use HD95 (the variant of HD) to eliminate the impact of a very small subset of the outliers; (ii) Boundary IoU (BIoU) is a widely used boundary-based metric; and (iii) PPV, similar to DSC and Sen, is a widely used overlap-based metric. Consequently, using diverse types of evaluation metrics can help us evaluate the models' segmentation performances more comprehensively and ensure the superior segmentation performances of our proposed Multi-ConDoS.

The segmentation results of Multi-ConDoS and the baselines with 10% labeled data in terms of HD95 and BIoU are shown in Table VI, while those in PPV are shown in Table VII. It can be observed from the tables that the relative segmentation performances of Multi-ConDoS and the baselines in terms of HD95, BIoU, and PPV are very similar to their DSC-based and Sen-based relative segmentation performances as shown in Table II. Consequently, we have the following findings: (i) The self-supervised and semi-supervised baselines generally outperform the fully supervised baseline method, this is because besides the limited amount of labeled data, self-supervised and semi-supervised segmentation methods can learn additional useful information from a large amount of unlabeled data. (ii) Multi-ConDoS significantly outperforms the state-of-the-art self-supervised and semi-supervised baselines in terms of all metrics (HD95, BIoU, and PPV), comprehensively proving the superior performances of Multi-ConDoS in medical image segmentation tasks.

3) Analysis of Layer-Transfer Strategies: Additional experimental studies with 10% labeled data on Hecktor and BraTS2018 datasets have been conducted to verify the soundness of transferring the weights of the encoder, shared layers, and the first three layers of the decoder, to the downstream task. As shown in Table VIII, we first find that by, comparing to training from scratch, transferring the encoder and shared layers significantly enhance the performances in the downstream segmentation tasks; This is because the imaging features learned by encoders (i.e., E_a and E_b) and share layers in pre-training are also very important in the downstream segmentation tasks. This thus proves that transferring the encoder and shared layers is reasonable. Then, we also observe that gradually transferring the first three layers of the decoder can also gradually improve the performances of the downstream

TABLE VIII

RESULTS OF USING DIFFERENT LAYER-TRANSFER STRATEGIES IN MULTI-CONDOS ON HECKTOR AND BRATS2018 DATASETS WITH 10% LABELED DATA. ENCODER_SL DENOTES THE LAYERS OF ENCODER AND SL ARE TRANSFERRED TO DOWNSTREAM NETWORK. ENCODER_SL_D1 (RESP., ENCODER_SL_D2 AND ENCODER_SL_D3) DENOTES THE LAYERS OF ENCODER AND SL AND THE FIRST ONE (RESP., TWO AND THREE) LAYERS OF DECODER ARE TRANSFERRED TO DOWNSTREAM NETWORK

Transfer Layers	Hecktor				BraTS2018							
	CT		PET		T1CE		T2		FLAIR		T1	
	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen	DSC	Sen
Training from scratch	0.2541	0.2875	0.5769	0.7067	0.4451	0.3696	0.4288	0.4716	0.4489	0.6225	0.2476	0.3287
Encoder_SL	0.3103	0.4915	0.6126	0.7579	0.5909	0.6268	0.5003	0.5806	0.4622	0.6230	0.3442	0.6070
Encoder_SL_D1	0.3217	0.4752	0.6277	0.7711	0.6032	0.6260	0.5016	0.5542	0.4730	**0.6653**	0.3626	0.6100
Encoder_SL_D2	0.3392	0.4969	0.6378	0.7752	0.6165	**0.6550**	0.5019	**0.6031**	0.4789	0.6156	0.3689	0.6079
Encoder_SL_D3 (OURS)	**0.3442**	**0.4970**	**0.6488**	**0.8012**	**0.6246**	0.6436	**0.5045**	0.5566	**0.4805**	0.6281	**0.3749**	**0.6149**

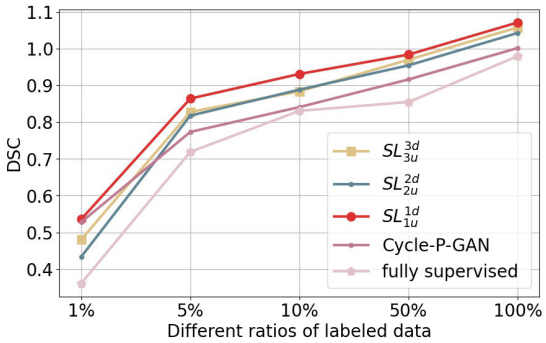
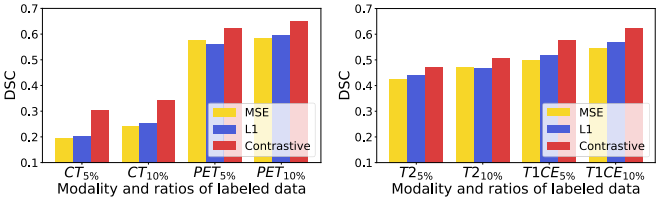


Fig. 6. Comparison of the different *SL* strategies performances on Hecktor. We show the sum of the segmentation results of the two modalities to better observe the effect of different *SL* strategies on the overall segmentation results of the two modalities.



(a) Hecktor (b) BraTS2018

Fig. 7. Comparison of the performances of different losses applied to multimodality features on Hecktor and BraTS2018.

segmentation tasks. This finding asserts that, although not as important as those learned by the encoder and shared layers, the features learned in the first three layers of decoders are still useful for the downstream learning tasks (i.e., at least better than random initialization).

4) Analysis of Shared Layers Strategies: To study the capacity of the SL on the performance of the Multi-ConDoS, we first study the strategy of the SL. Moreover, we have described different SL strategies in Table I in detail. To study the impact of different SL strategies, we conduct experiments on the Hecktor dataset with different ratios (1%, 5%, 10%, 50%, and 100%) labeled data. Then, Fig. 6 shows the experimental results of DSG using different SL strategies on Hecktor. In Fig. 6, fully supervised means that we train the network from scratch; CycleGAN represents that we use two completely separate generators for pair-wise training; SL_{1u}^{1d} represents that we share one downsampling layer and one upsampling layer of the middle part of the two generators; SL_{2u}^{2d} and SL_{3u}^{3d} are by analogy.

Please note that in order to better observe the effect of different SL strategies on the overall segmentation results of the two modalities, the DSC in Fig. 6 is the sum of the segmentation results of the two modalities.

From Fig. 6, we can observe that the three strategies of SL_{1u}^{1d}, SL_{2u}^{2d}, and SL_{3u}^{3d} all promote a more remarkable improvement in model performance than training from scratch (fully supervised) and CycleGAN in all ratios, and the performance of SL_{1u}^{1d} improves the most. This is because when too many layers are shared, the number of layers for learning modal-specific knowledge will decrease, which will damage the ability of the network to learn specific knowledge of different modalities. When using the SL_{1u}^{1d} strategy, both specific and general knowledge of different modalities has been well learned.

5) Analysis of Multimodal Contrastive Loss: To study the superiority of multimodality contrastive loss, we conducted experiments on two datasets and compared the performance of multimodality contrastive loss with the two most commonly used losses: L1 and MSE loss. For a fair comparison, the other settings are all kept consistent. The results are shown in Fig. 7. In Fig. 7, $CT_{5\%}$ means that we use 5% annotations of Hecktor (CT). Others are by analogy.

From Fig. 7, we can see that when using the MSE and L1 loss for the multimodality features extracted by the SL, the performance of downstream segmentation tasks is very similar. However, the performance is best when using multimodality contrastive loss. This is because the MSE and the L1 loss can only minimize the difference between positive sample pairs, but cannot maximize the difference between negative sample pairs like the multimodality contrastive loss. This means that the usage of the MSE and the L1 loss can only encourage the network to learn consistent representations, while the multimodality contrastive loss encourages the network to learn consistent representations and also encourages the network to learn discriminative representations, which is very important for the network to learn useful semantic information.

6) Analysis of Four Modal Selections on BraTS2018: There exist four modalities in the BraTS2018 database, therefore, to construct segmentation models for all four modalities, we need to divide the database into two sets of paired images, and apply the proposed Multi-ConDoS on them separately. Therefore, in order to enhance the practical values of Multi-ConDoS, in this section, we further investigate the best

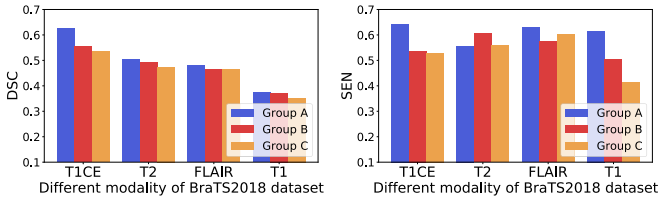


Fig. 8. Results of all four modalities of Multi-ConDoS on BraTS2018 using three different grouping strategies, i.e., Group A <T1CE-T2, FLAIR-T1>, Group B<T1CE-T1, FLAIR-T2>, and Group C <T1-T2, FLAIR-T1CE>.

grouping strategy for BraTS2018 and provide the results of the four modalities according to the best group strategy.

Specifically, there exist three kinds of potential grouping choices: namely, Group A: <T1CE-T2, FLAIR-T1>, Group B: <T1CE-FLAIR, T2-T1>, and Group C: <T1CE-T1, FLAIR-T2>. Experiments of Multi-ConDoS on BraTS2018 are conducted by applying different grouping strategies, whose results are shown in Fig. 8. It shows that the grouping strategy A (denoted Group A) consistently achieves the best results for all four modalities in terms of both DSC and Sen. This may be because the visual differences between the paired images in Group A (i.e., T1CE vs T2 and FLAIR vs T1) are much greater than those in Group B and Group C. This is consistent with our argument that the greater the difference between multimodal data, the higher Multi-ConDoS can improve. Therefore, the results of all four modalities in our work are obtained based on the grouping strategy <T1CE-T2, FLAIR-T1>. Furthermore, we believe that, when Multi-ConDoS is applied in clinical practices that involve more than two modalities, using the grouping strategy that can maximize the differences between the paired modalities is most possible to obtain the best segmentation performances.

VI. Discussion and Future Work

In this section, we further discuss three key points regarding the application of the proposed Multi-ConDoS in real-world clinical practices to justify its good practical values.

A. Multiple Segmentation Models in Multi-ConDoS

Theoretically, when the multi-modal images are well-registered, they should have the same segmentation mask (i.e., the same ground truth). This semantic consistency between multimodal images is exactly the theoretical basis on which we can utilize contrastive learning to enhance the feature learning capability of Multi-ConDoS. However, although the segmentation results of the two modalities are theoretically the same after registration, we believe it is still beneficial to construct and fine-tune multiple segmentation models for multimodal medical images in the downstream segmentation tasks, because this can effectively enhance the performances, applicability and flexibility of Multi-ConDoS in real-world clinical practices, and so as to enhance its practical value.

First, since the images of different modalities are generated using different imaging mechanisms, the information and features contained in the paired images with different modalities are very different. For example, as shown in Fig. 2 (visualized results) of the revised manuscript, CT (generated by the structural imaging technique) can clearly show details of the organs and tissues inside the human's body, while PET (generated by the functional imaging technique) only exhibits (abnormal) changes in metabolic processes and physiological activities. Therefore, although the segmentation results of the two modalities should theoretically be the same (i.e., the same ground truth), since it is almost impossible for deep models to predict with 100% accuracy in practice, the multimodal images with different information and features will result in different practical segmentation results, which usually contain complementary information (i.e., making different mistakes in predictions); consequently, constructing and fine-tuning multiple segmentation models for multimodal medical images in the downstream segmentation tasks will provide multiple possible predictions for one segmentation target, by adding some post-processing operations to comprehensively consider these possible predictions, it can provide the doctors and patients more accurate final predictions in clinical practices, i.e., enhance the practical segmentation performances.

Second, patients usually cannot afford to obtain multimodal medical images, so constructing and fine-tuning multiple segmentation models for multimodal medical images can significantly enhance the applicability and flexibility of Multi-ConDoS in practices: patients only need to get medical images of any modality, the corresponding segmentation model can be immediately used to automatically predict the segmentation results, which thus avoid unnecessary time and money consumption (i.e., patients do not have to obtain or wait for medical images of a specific modality, so save valuable time and money for the diagnosis and treatment of the diseases).

Finally, since we only use a small number of labeled data to fine-tune the models in the downstream segmentation tasks, the cost of obtaining multiple segmentation models for multimodal medical images is actually very low. Considering the significant improvements of the segmentation performances, applicability and flexibility in practices, such a limited extra cost is affordable and reasonable.

B. Pre-Training Time-Cost of Using Multi-ConDoS

As presented in Section V-B, DSGANs contain multiple modules and various loss terms, all of which are essential and effective for our Multi-ConDoS method to achieve the superior performances. However, this will inevitably increase the model's learning time. Therefore, in this work, we further compare the pre-training time-cost of Multi-ConDoS with the state-of-the-art self-supervised contrastive learning baselines to reveal its comparative complexity to the baselines and to see whether or not its application value in clinical practices will be undermined due to the complexity.

As shown in Table IX, the pre-training time-cost of Multi-ConDoS is much lower than that of G-L, close to that of BYOL, and higher than those of SimCLR, SwAV and ContIG. We note that, even compared to the most efficient baselines (i.e., ContIG on Hecktor and SwAV on BraTS2018), the time-cost of Multi-ConDoS is still acceptable: only about twice of the best one. Considering the rapid increase in the

TABLE IX
THE TIME-COSTS (IN HOURS) OF MULTI-CONDOS AND THE SOTA SELF-SUPERVISED CONTRASTIVE LEARNING BASELINES IN THE PRETRAINING STAGE

Methods	Hecktor	BraTS2018
SimCLR [14]	9.06	19.48
BYOL [18]	15.34	34.08
SwAV [19]	11.80	18.08
G-L [9]	81.40	183.68
ContIG [20]	8.06	18.72
Multi-ConDoS	16.21	36.60

computing capability of today's computing devices, and the fact that the improvement of prediction accuracy is of great significance for medical image analysis in clinical practices (i.e., even a little improvement in accuracy will be able to save more patients' lives), the importance of accuracy is actually much higher than that of efficiency for practical medical image analysis, so we believe sacrificing a limited amount of additional computational costs in exchange for a significant increase in the segmentation performances is acceptable, worthy and valuable for the practical medical image segmentation tasks. Therefore, we believe the complexity of Multi-ConDoS is acceptable and will not weaken its clinical value.

C. Medical Image Registration

In our work, images of different modalities need to be registered before they are used in Multi-ConDoS; however, we believe this requirement will not seriously affect the application of the proposed Multi-ConDoS in medical practices. The reasons are as follows. On one hand, there exist some multi-modal medical images that are natively registered, e.g., the multi-modal MR images and the CT-PET images (i.e., the ones used in our experimental studies) have been already registered when they are generated from the imaging equipments. On the other hand, medical image registration has been studying for quite a long time, there have existed many combinations of multi-modal medical images that can be easily registered using the existing registration solutions (e.g., MR to CT registration [55], [56], registration of fluoroscopic X-ray to CT [57], PET to MRI registration [58], and preoperative magnetic resonance (MR) to intraoperative ultrasound registration [59]) or the existing medical image registration tools (such as SimpleITK [60] and SimpleElastix [61]). Therefore, even if not all, our proposed multi-modal self-supervised segmentation solution can be applied to many clinical segmentation tasks by registering the corresponding multi-modal medical images before using them as inputs. More importantly, registration operations can bring additional benefits: after applying registration, we can use the same segmentation masks for both modalities, which thus reduces half of the annotation time cost. Consequently, we believe the need of image registration will not hinder the application of our work in real-world clinical practices; and an interesting future work is to conduct an in-depth survey to find out all the possible combinations of different multimodal medical images that can be properly registered and the corresponding registration methods to guide the application of our Multi-ConDoS in daily clinical practices.

VII. CONCLUSION

In this work, we proposed a multimodal contrastive domain-sharing (Multi-ConDoS) approach, which is the first multimodal contrastive-based self-supervised medical image segmentation approach that not only overcomes the domain shift problem but also takes the advantage of the fruitful multimodal information of medical images. In addition, the novel domain-sharing generative adversarial networks (DSGANs) were further proposed for the contrastive domain translation, which can learn both specific and general feature representations more effectively using domain-sharing layers and multimodal contrastive loss. Extensive experiments on two public multimodal medical image datasets were conducted; the experimental results demonstrated that the proposed Multi-ConDoS can achieve superior medical image segmentation performances using only a small number of annotations, which thus greatly reduced the labeling workload of using intelligent medical image segmentation systems in the real-world scenarios. The effectiveness and necessity of all three advanced components in the proposed Multi-ConDoS were also proved by ablation studies.

REFERENCES

[1] K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.
[2] L. Wang, B. Wang, and Z. Xu, "Tumor segmentation based on deeply supervised multi-scale U-Net," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 746–749.
[3] Z. Xu et al., "ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution," *Neurocomputing*, vol. 500, pp. 177–190, Aug. 2022.
[4] D. Yuan, Z. Xu, B. Tian, H. Wang, Y. Zhan, and T. Lukasiewicz, "μ-Net: Medical image segmentation using efficient and effective deep supervision," *Comput. Biol. Med.*, vol. 160, Jun. 2023, Art. no. 106963.
[5] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's cube," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 420–428.
[6] W. Bai et al., "Self-supervised learning for cardiac MR image segmentation by anatomical position prediction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*, 2019, pp. 541–549.
[7] N. Tajbakhsh et al., "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1251–1255.
[8] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.
[9] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.
[10] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.
[11] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 69–84.
[12] A. Taleb, C. Lipperit, T. Klein, and M. Nabi, "Multimodal self-supervised learning for medical image analysis," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2021, pp. 661–673.
[13] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.
[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.
[15] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," 2020, *arXiv:2006.10029*.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[17] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[18] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[19] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 9912–9924.

[20] A. Taleb, M. Kirchler, R. Monti, and C. Lippert, "ContIG: Self-supervised multimodal contrastive learning for medical imaging with genetics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20876–20889.

[21] R. Windsor, A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised multi-modal alignment for whole body medical imaging," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2021, pp. 90–101.

[22] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[23] Z. Xu, C. Qi, and G. Xu, "Semi-supervised attention-guided CycleGAN for data augmentation on medical images," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 563–568.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*.

[25] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 577–593.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[27] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1735–1742.

[28] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Proc. NeurIPS*, vol. 27, 2014, pp. 766–774.

[29] J. Iwasawa, Y. Hirano, and Y. Sugawara, "Label-efficient multi-task segmentation using contrastive learning," 2020, *arXiv:2009.11160*.

[30] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[31] O. J. Hénaff et al., "Data-efficient image recognition with contrastive predictive coding," 2019, *arXiv:1905.09272*.

[32] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102656.

[33] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[34] Y. Zhang and J. Zhang, "Dual-task mutual learning for semi-supervised medical image segmentation," 2021, *arXiv:2103.04708*.

[35] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2020, pp. 552–561.

[36] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[37] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," 2017, *arXiv:1703.00848*.

[38] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 172–189.

[39] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 35–51.

[40] H.-Y. Lee et al., "DRIT++: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128, nos. 10 and 11, pp. 2402–2417, Nov. 2020.

[41] Y. Yuan, "Automatic head and neck tumor segmentation in PET/CT with scale attention network," in *Proc. 3D Head Neck Tumor Segmentation PET/CT Challenge*, 2020, pp. 44–52.

[42] V. Andrearczyk et al., "Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans," in *Proc. Med. Imag. With Deep Learn.*, 2020, pp. 33–43.

[43] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[44] S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, pp. 1–13, Sep. 2017.

[45] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2015, pp. 234–241.

[48] Z. Xu, T. Li, Y. Liu, Y. Zhan, J. Chen, and T. Lukasiewicz, "PAC-Net: Multi-pathway FPN with position attention guided connections and vertex distance IoU for 3D medical image detection," *Frontiers Bioeng. Biotechnol.*, vol. 11, Feb. 2023, Art. no. 1049555.

[49] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[50] D. C. Dowson and B. V. Landau, "The Fréchet distance between multivariate normal distributions," *J. Multivariate Anal.*, vol. 12, no. 3, pp. 450–455, Sep. 1982.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[52] D. Karimi and S. E. Salcudean, "Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, Feb. 2020.

[53] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15329–15337.

[54] N. Chinchor and B. Sundheim, "MUC-5 evaluation metrics," in *Proc. 5th Conf. Message Understand. (MUC)*, 1993, pp. 1–10.

[55] H. A. Mohammed and M. A. Hassan, "The image registration techniques for medical imaging (MRI-CT)," *Amer. J. Biomed. Eng.*, vol. 6, no. 2, pp. 53–58, 2016.

[56] S. Roy, A. Carass, A. Jog, J. L. Prince, and J. Lee, "MR to CT registration of brains using image synthesis," *Proc. SPIE*, vol. 9034, Mar. 2014, Art. no. 903419.

[57] H. Livyatan, Z. Yaniv, and L. Joskowicz, "Gradient-based 2-D/3-D rigid registration of fluoroscopic X-ray to CT," *IEEE Trans. Med. Imag.*, vol. 22, no. 11, pp. 1395–1406, Nov. 2003.

[58] Z. Y. Shan, S. J. Mateja, W. E. Reddick, J. O. Glass, and B. L. Shulkin, "Retrospective evaluation of PET-MRI registration algorithms," *J. Digit. Imag.*, vol. 24, no. 3, pp. 485–493, Jun. 2011.

[59] I. Machado et al., "Deformable MRI-ultrasound registration using correlation-based attribute matching for brain shift correction: Accuracy and generality in multi-site data," *NeuroImage*, vol. 202, Nov. 2019, Art. no. 116094.

[60] R. Beare, B. Lowekamp, and Z. Yaniv, "Image segmentation, registration and characterization in R with SimpleITK," *J. Stat. Softw.*, vol. 86, no. 8, pp. 1–35, 2018.

[61] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A user-friendly, multi-lingual library for medical image registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 574–582.