Semi-supervised Unpaired Medical Image Segmentation Through Task-affinity Consistency

Jingkun Chen, Jianguo Zhang, *Senior Member, IEEE*, Kurt Debattista and Jungong Han, *Member, IEEE*

***Abstract*— Deep learning-based semi-supervised learning (SSL) algorithms are promising in reducing the cost of manual annotation of clinicians by using unlabelled data, when developing medical image segmentation tools. However, to date, most existing semi-supervised learning (SSL) algorithms treat the labelled images and unlabelled images separately and ignore the explicit connection between them; this disregards essential shared information and thus hinders further performance improvements. To mine the *shared* information between the labelled and unlabelled images, we introduce a class-specific representation extraction approach, in which a *task-affinity* module is specifically designed for representation extraction. We further cast the representation into two different views of feature maps; one is focusing on low-level context, while the other concentrates on structural information. The two views of feature maps are incorporated into the task-affinity module, which then extracts the class-specific representations to aid the knowledge transfer from the labelled images to the unlabelled images. In particular, a task-affinity consistency loss between the labelled images and unlabelled images based on the multi-scale class-specific representations is formulated, leading to a significant performance improvement. Experimental results on three datasets show that our method consistently outperforms existing state-of-the-art methods. Our findings highlight the potential of consistency between class-specific knowledge for semi-supervised medical image segmentation. The code and models are to be made publicly available at https://github.com/jingkunchen/TAC.**

***Index Terms*— Semi-supervised, Segmentation, Contextual, Structural, Task-affinity, Consistency.**

I. INTRODUCTION

This work is supported in part by National Key Research and Development Program of China (2021YFF1200800), and Shenzhen Science, Technology and Innovation Commission Basic Research Project under Grant No. JCYJ20180507181527806.

J. Chen and J. Zhang are with the Department of Computer Science and Engineering, the Southern University of Science and Technology, Shenzhen, China. email: jingkun.chen@warwick.ac.uk, zhangjg@sustech.edu.cn.

J. Chen, K. Debattista and J. Han are with the Warwick Manufacturing Group, the University of Warwick, Coventry, UK. email: jingkun.chen, k.debattista@warwick.ac.uk, jungonghan77@gmail.com.

J. Han is with the Department of Computer Science, the Aberystwyth University, Aberystwyth, UK. email: jungonghan77@gmail.com.

J. Zhang is with the Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China. email: zhangjg@sustech.edu.cn.

J. Zhang is with the Peng Cheng Lab, Shenzhen, China. email: zhangjg@sustech.edu.cn.

Corresponding authors: Jianguo Zhang, Jungong Han and Kurt Debattista.

THE precise segmentation of organs, tissues, and lesions on medical images plays an important role in medical diagnosis and treatment. Deep neural networks based on the supervised learning paradigm have been firmly established as a powerful tool in image segmentation [39]. However, obtaining exhaustive annotations on medical images for learning remains a major challenge. With the development of deep learning, semi-supervised learning (SSL) has been increasingly employed in clinical research, given its comparable performance to fully supervised learning [24] by training on a combination of a small amount of labelled images and large amounts of unlabelled ones.

Previous SSL algorithms have successfully improved the medical image segmentation accuracy, where consistency regularization [30] and pseudo-labeling [35] based algorithms have arisen as the two main streams. For consistency regularization-based methods, the predictions for the individual image under different augmentations, dropout and stochastic perturbations are made consistent. Despite their promising performance, two augmented versions of the same image are often utilized as *paired* images to regularize a pair of outputs to be identical (shown in Fig. 1 (A)), where the explicit common characteristics of the organs in both the labelled and unlabelled images such as texture, structure, colour and intensity are not highlighted. There is no direct knowledge transfer between the *unpaired* instances, *i.e.,* a labelled image and an unlabelled image (shown in Fig. 1 (B)). For the traditional consistency learning methods, a popular approach is to generate two images by applying two different augmentations to the same image, and then force the network to produce similar outputs. When the generalisation ability of the model is weak, it is difficult to ensure that two different unlabelled augmented images can contribute to the optimisation of the model as it is possible that the two images will be far from each other in the feature space due to the weak generalisation power [40]. Alternatively, pseudo-labeling methods [13], [25] utilize the supervised pretrained networks to generate pseudo-labels for the unlabelled images to retrain the network. By providing more explicit supervision, these methods generally outperform consistency regularization methods if the pseudo-labels are good enough to provide correct supervision as the labelled ones. However, pseudo-labeling still treats unpaired images *independently*, *i.e.,* labelled images for supervised learning, unlabelled images for pseudo-labeling learning. Effectively, using unpaired data to improve segmentation capabilities is not straightforward because of two unsolved problems identified

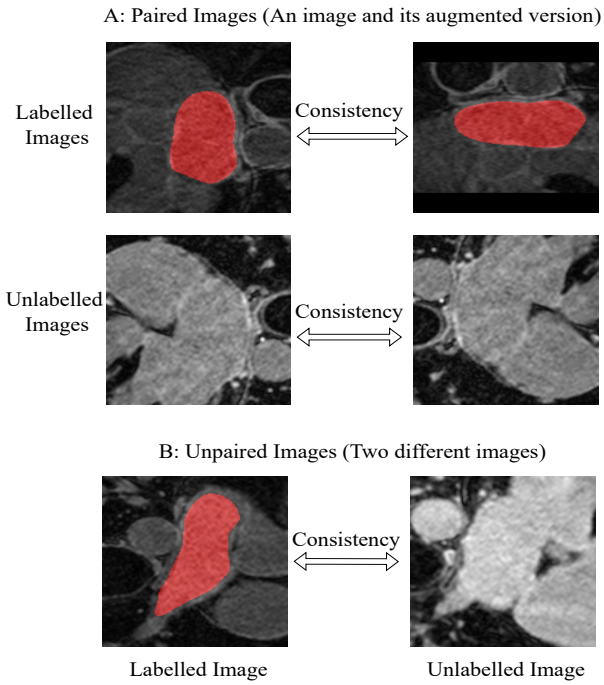

Fig. 1: The difference between paired images and unpaired images in semi-supervised learning. (A) is using paired images (an image and its augmented version). (B) is using unpaired images (two different images, one is a labelled image and one is an unlabelled image).

from the current SSL literature. The first issue is how to solve the problem that consistency regularization methods with two different augmentations from one unlabelled image may have a negative effect on optimising the network when the generalisation ability of the model is weak. The second unsolved problem is which kinds of representations can be used to build a bridge from the labelled images to the unlabelled images and how to use them.

Until now, the knowledge transfer between the unpaired images has not yet been widely explored in SSL for medical image segmentation. In segmentation tasks, there is a significant difference between the different kinds of feature maps from the original layers and the enhancement layers. Combining such complementary information on the same organ usually improves the medical image segmentation performance [50]. It is also possible to construct different descriptions of the same image in the feature layer with different enhancement modules to highlight different information, *i.e.,* constructing one kind of feature map that focuses on the contour of the target area and another feature map that focuses on the texture of the target area. Moreover, the different descriptive characteristics of different patients, such as the texture, colour, intensity, and structure, whether it is a labelled image or an unlabelled image, tend to have the same characteristics. For example, the appearance of the left atrium in images from different patients often presents similar texture, colour and intensity. In convolutional neural networks, we hope such texture and intensity information can be extracted from the feature maps as for an image, the shape, texture, colour, intensity and other information of the same organ of different patients should be roughly the same and can also be used as discriminative information for learning.

Inspired by the above, in this paper, we propose a novel SSL unpaired medical image segmentation framework, explicitly extracting the representations from labelled images to supervise feature extraction of unlabelled images. We construct two different types of feature maps to highlight different information about the target region, for example, structural affinity features focus on the structure of the target region and contextual affinity features focus on the context of the target region. In our approach, these two different feature maps are used together for consistency learning of both labelled and unlabelled data. The ground truth masks of the labelled images and the estimated masks of the unlabelled images are combined with the structural and contextual affinity feature maps to highlight target region representation at the feature level; we refer to this as *class-specific* knowledge in the following. To extract the class-specific knowledge from contextual affinity features and structural affinity features in the network, a new loss function, termed as task-affinity loss, is introduced. Task-affinity loss optimises the target area representations in the training process. We evaluate our method on three multi-class medical segmentation tasks, including 2D cardiac, 3D left atrium and 3D hippocampus segmentation tasks. On all three datasets, our proposed algorithm consistently outperforms the state-of-the-art methods compared. The main contributions of this paper are:

1) We proposed a novel semi-supervised segmentation framework by designing a class-specific knowledge extraction mechanism, which transfers class-specific knowledge from contextual and structural affinity features from labelled to unlabelled images.
2) A module is specifically introduced to model the consistency of class-specific knowledge between the unpaired images on the multi-scale contextual and structural affinity features in semi-supervised segmentation. To the best of our knowledge, this is the first attempt to enable direct knowledge transfer for *unpaired* images in SSL.
3) We demonstrate the wider applicability of our method to the different backbones of both 2D and 3D networks in multi-class segmentation tasks. Compared to existing methods, our method achieves the best results in DSC in all conducted experiments. In particular, by performing multiple independent evaluations, the performance of our method shows higher reliability with much lower variance on the three datasets.

II. RELATED WORK

SSL has been an active topic of research for the last decades. A large number of methods have been proposed before and after the emergence of deep learning. Compared to supervised learning methods, one of the key challenges of SSL is how to effectively utilize the unlabelled training images. Depending on the backbones and whether the adversarial training is adopted or not, existing semi-supervised learning methods can be roughly divided into CNN-based [36], graph-based [41], transformer-based [22] and GAN-based [46] [18] methods. However, in this paper, we focus on developing new

strategies of how the knowledge could be better transferred or regularised between the labelled and unlabelled sets. To this point, we broadly categorized existing SSL methods into pseudo-labeling and consistency regularization depending on the strategies of knowledge leveraging between the two sets. In this section, we follow this categorisation and review these methods highly relevant to ours.

A. Pseudo-Labeling

One earlier pseudo-labeling method for SSL was developed by Lee et al. [13], which uses a model to predict pseudo labels for a large number of unlabelled images to improve performance. Such a prediction is in turn used as a (pseudo) ground truth to fine-tune the network. This enables the involvement of unlabelled training data in the whole learning pipeline. However, as pseudo labels are usually noisy, a mechanism needs to be in place to select the most informative unlabelled training data for refining the networks. Therefore, the subsequent research has focused on strategies for selecting useful unlabelled images. A common practice of selection mechanisms is to filter out the wrong pixels by using the confidence of pseudo-label predictions, such as uncertainty measurements [20], [33], adversarial learning [37], [17], and selection of valuable pixels [36], [25]; however, these methods only rely on the confidence prediction from the softmax layer and are not informative enough to provide reliable guidance for unlabelled images in supervised learning. In our work, instead of focusing on designing optimisation strategies to improve the quality of pseudo labels, we take a different perspective that explores the shared class-specific information in features between the unpaired images, with the aim of transferring the knowledge from the labelled images to the unlabelled ones.

B. Consistency Regularization

In the context of deep neural networks, the conceptual idea of consistency regularization is from Bachman et al., [1], *i.e.,* the predictions should be robust to different perturbations of the input samples. The process adds data enhancement and stochastic perturbations to the input but does not permit changes in the predictions. In medical image segmentation, Cui et al. [6] introduced the mean-teacher framework by constraining the outputs of the same image under different noise perturbations. In [19], the geometric transformation consistency loss is used to optimise the network for skin lesions, optic disc, and liver tumors. Bortsova et al. [3] proposed a siamese segmentation network on chest x-ray images. They introduced a separate layer to minimize the distance of outputs from the different transformations at the output. Luo et al. [22] introduced a dual-task consistency regularization to minimize the distance of the signed distance maps and the directly estimated masks of the left atrium and pancreas segmentation.

In the context of semi-supervised medical image segmentation, although deep learning has achieved promising results, most of the methods advocate the use of paired images (two augmented versions of the same image) but pay little attention to the unpaired images (images from different patients). Our method extends the existing consistency regularization method but preserves the consistency between unpaired images. Alongside class-specific regions at different feature layers, our method allows the knowledge transfer from labelled images to unlabelled images, thereby improving the performance of semi-supervised learning, based on the unpaired consistency which is not yet explored before.

III. METHOD

The framework of our proposed unpaired semi-supervised learning method is shown in Fig. 2. The backbone of our framework is an encoder-decoder structure. As each specific type of region largely shares the same key information across different images (*e.g.* the atrium always has a thicker wall, a similar texture and a variable shape across subjects [9]), we hypothesize class-specific information should be similar across the regions of the same interest in the feature maps regardless whether it is labelled or not. Therefore, to make use of this class-specific information, we design a module called *task-affinity module* (as shown in Fig. 2) to transfer the class-specific knowledge from labelled images to unlabelled ones. We note that, unlike existing consistency-based methods, our method enables knowledge transfer between the unpaired images. The purpose of the task-affinity module is to transfer the class-specific knowledge between the same class regions in different layers across the unpaired images. Therefore, the class-specific knowledge extracted from the labelled images can be used to supervise the unlabelled images in the training process. Details of the proposed framework, task-affinity module, unpaired consistency regularization loss and the unpaired training strategies are described below.

A. Class-specific Representation

The key of our framework is to transfer the class-specific knowledge from the labelled images to the unlabelled images. In the segmentation task, recent studies [5], [16] have revealed that the feature maps in a network could reflect the saliency of the class of interest in a multi-class setting. However, the class agnostic representations are insensitive to class labels [45], in principle, it can't be used to build the consistency of features between the unpaired images. Different from it, we aim to extract transferable knowledge by building class-specific representations. We use the CAM-based methods [44], [47] to visualise the feature maps. Fig. 3 shows feature maps and the masks of the labelled images and unlabelled images. The salient area corresponding to the target class has the same characteristics both for labelled and unlabelled images. In Fig. 3 (b), a feature map can show regions with distinct characteristics, *e.g.* the area of higher luminance is the representation of the left atrium, while the low-luminance area represents the background. For labelled images, we leverage a ground truth mask to extract the correct salient region in the feature maps, which contain accurate class-specific knowledge. But for unlabelled images, we use the estimated mask to extract the salient region. The estimated mask may have false positive and false negative areas, which may point to the features of the other classes, so using the estimated masks to extract the region of interest in the feature maps may not

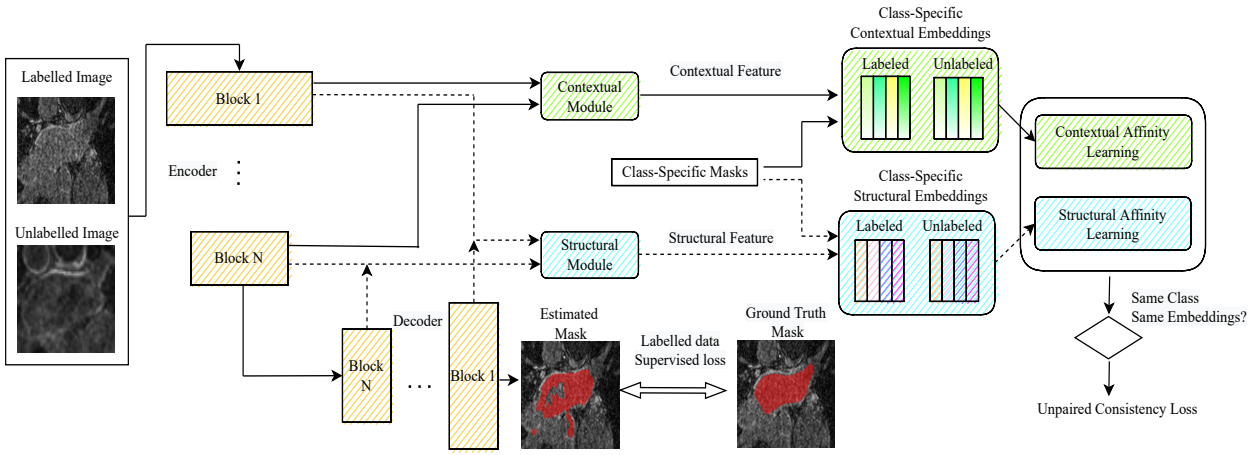

Fig. 2: The proposed semi-supervised task-affinity learning framework for segmentation. The modules in the green boxes and the solid lines with arrows are for contextual affinity learning; the modules in the blue boxes and the dashed lines with arrows are for structural affinity learning. Our framework consists of the task-affinity module with an encoder-decoder structure as the backbone. In the task-affinity module, the ground truth masks and estimated masks are used separately on the labelled images and unlabelled images to locate the class-specific regions from the contextual affinity feature and the structural affinity feature. Features from class-specific regions are then compressed into vectors (the different embeddings' colours indicate different classes) for unpaired images affinity consistency learning.

always be correct (as shown in the blue boxes in Fig. 3 (c)). By comparing Fig. 3 (b) with (c), the labelled images and unlabelled images have similar regions of interest with respect to the same class in the feature layer. According to the mask in (c), the labelled images can perfectly draw the target class region in the feature maps. However, the estimated masks of unlabelled images cannot always be good enough to show the region of interest in the feature maps. In the case when the accuracy of the estimated mask for the unlabelled image is poor for a particular class (e.g., left atrium), the features from this masked region will differ significantly from the features computed from the ground-truth masked region (e.g., mask for left atrium) for the labelled image. This large inconsistency will provide a strong supervision signal in the training process.

Since our input data (whether labelled or unlabelled) belongs to the same distribution, and predictions are made through the same network, the corresponding feature of the same class from the different images should be salient due to the similar contextual information (as shown in Fig. 3 (b)). The class-specific representations can be used with the ground truth masks of the labelled images to build a reference, which optimises the inaccurately estimated mask of the unlabelled images at the feature level. In this way, we consider this class-specific knowledge in the feature maps can be encoded as a bridge for the knowledge transfer between the unpaired images in semi-supervised learning. Therefore, through the guidance of class-specific knowledge from labelled images, the false positive and false negative regions of unlabelled images can potentially be eliminated.

To achieve this, the salient regions in feature maps are first localised using segmentation masks, *i.e.,* ground truth masks of the labelled images and the estimated masks of the unlabelled images to localise feature regions. In our case, we encode class-specific knowledge through the task-affinity module, and expand it in a layer-wise manner as follows:

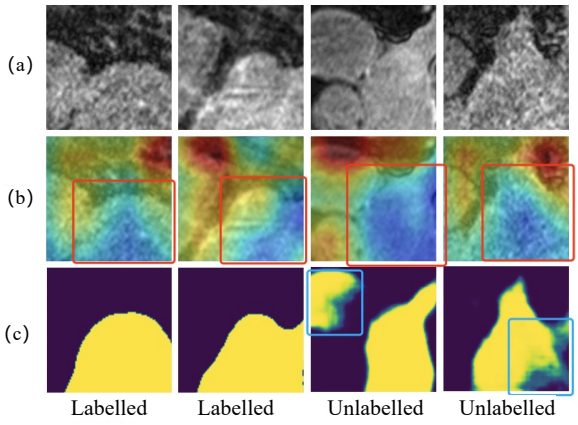

Fig. 3: Feature maps and the masks of labelled/unlabelled images. From left to right are the labelled and unlabelled images. From top to bottom are (a) original images, (b) feature maps and (c) ground truth masks of the labelled images and estimated mask of unlabelled images. In row (b), the class-specific region in the feature maps is located in the red boxes. The blue boxes in row (c) show the false positive and false negative masks of the unlabelled images. The area inside the blue box will be gradually refined as the training progresses.

$$T_i^c = \sum_{w,h,z} \left(x^i \odot M^c\right) / \sum_{w,h,z} M^c. \tag{1}$$

Eq. 1 is the i_{th} class-specific representation (T_i^c) generated by the task-affinity module, where x_i denotes the learned feature maps. M^c denotes the mask on the c_{th} class. For the labelled images, the class-specific knowledge can be extracted via the ground truth mask, while, for the unlabelled images, the estimated mask from the network is used for the extraction. $\odot$ is the Hadamard product, w, h, z are the width, length, and height of the feature map x^i, respectively (for the 2D tasks, we use w and h to calculate T_i^c). The class-specific representation

is then normalized by the area of the corresponding class.

We separate the labelled and unlabelled images and adopt different strategies to handle them. For the labelled images, the ground truth mask can provide correct supervision when calculating the T_i^c. However, for the unlabelled images, the estimated mask cannot always be good enough to extract the region of interest. With the training of the network, based on the correct T_i^c from the labelled images and their ground truth masks, the T_i^c generated by the estimated mask and the unlabelled images will gradually become consistent.

B. Structural and Contextual Affinity Modules

Contextual (*e.g.,* texture and intensities) and structural information (*e.g.,* shapes) are two distinct types of cues used to characterize regions of interest in medical images [14], [15]. For example, the left atrium in MR images shows thicker walls and smoother endocardial surface [10]. Recent work also demonstrates that segmentation of the region of interest can benefit from highlighting its structural information [7], [18]. They both consist of features at different layers, upon which the affinity features are constructed. Therefore, we introduce contextual and structural affinity modules to construct these two kinds of features.

To highlight the structural view of the features, we design a structural enhancement layer [2] and apply it to the encoder feature maps (Fig. 4) to construct a structural enhancement map, formulated as:

$$\alpha = \sigma\left(f_\psi\left(ReLU\left(f_{\bar{x}}\left(\bar{x}^i\right) + f_x\left(x^i\right)\right)\right)\right), \tag{2}$$

where, f_x, $f_{\bar{x}}$ and f_ψ are the $1 \times 1 \times 1$ convolution layers (1×1 convolution for 2D tasks) that offer a channel-wise pooling to retaining the salient features [42], [43], x^i and $\bar{x}^i$ are the i_{th} pair of the encoder and decoder feature maps. σ is the sigmoid activation function to restrict the range to [0,1] [44]. The output α is then applied to the encoder feature maps to highlight the region of interest $\hat{x}^i$ (structural affinity features in the following sections) as:

$$\hat{x}^i = \alpha \times x^i. \tag{3}$$

A detailed architecture of the structural enhancement layer is shown in Fig. 4. In the representation extraction process, the feature x^i generated by the encoder layers will be concatenated (along the feature channel dimensions) with the feature map $\bar{x}^i$ of the same size from the decoder layers, and then fed to a fusion module composed of a ReLU layer followed by a $1 \times 1 \times 1$ convolution layer, (1×1 for 2D tasks). The output of the fusion module is then fed to a sigmoid function, whose output is further applied to the encoder layers as $\hat{x}^i$ that suppresses the response in the irrelevant background area. Finally, it is merged into the decoder layer at the corresponding position through a skip connection. This structural enhancement module provides different feature representations and focuses on the structure of the target region (shown in Fig. 5 (c)).

Feature maps before and after applying this layer reflect two different characteristics of the class-specific regions at the feature level, therefore, resulting in two views of the feature

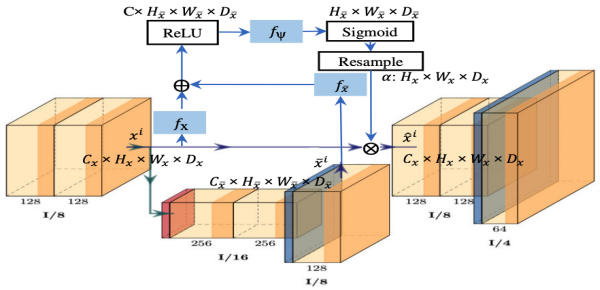

Fig. 4: The proposed structural enhancement layer. f_x, $f_{\bar{x}}$ and f_ψ are the $1 \times 1 \times 1$ convolution layers, $\oplus$ is the addition function, $\otimes$ is the multiplication formula. After applying this layer, the structural information is more prominent as shown in Fig. 5 (c).

maps. Fig. 5 shows examples illustrating the difference of such feature maps with respect to the same image. In Fig. 5 (c), it is evident that the contour of the target region is sharpened, thus providing a strong structural cue/view for that class. Fig. 5 (d) shows that feature maps that highlight the contextual texture-like patterns, resulting from the variations of texture and intensity in the original images. It has been previously demonstrated that such contextual information can promote robust and better segmentation results [12].

In our work, we assume that this class-specific information should be consistent at the feature level often with different scales resulting from a sequence of convolutions. More importantly, we show that this information can be used as a piece of prior knowledge to provide supervision for model learning across the unpaired images.

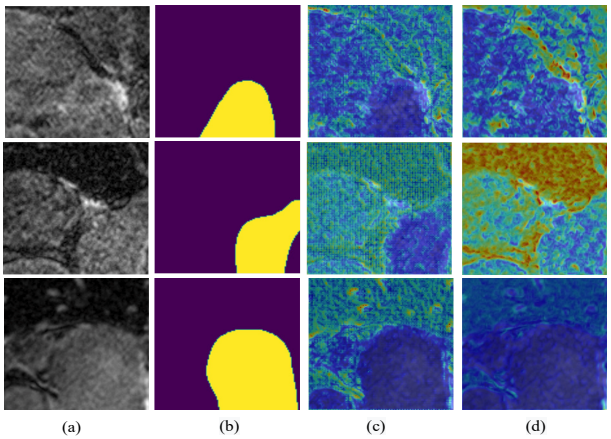

Fig. 5: From left to right are (a) original images, (b) ground truth labels, (c) feature maps emphasizing the structural information, and (d) feature maps focusing on the contextual information. In (c), the contour of the target area is more obvious. In (d), the texture of the target area is the major cue. In (c) and (d), different colours mean different values, similar colours reflect similar features, such as blue and red.

To make full use of the information from those two different views of features, we introduce two different types of representations: contextual representation (E) and structural representation (S). Let $x_k \in [x_1, \cdots, x_K]$ be the contextual affinity feature maps of K scales and $\hat{x}_k \in [\hat{x}_1, \cdots, \hat{x}_K]$ be the corresponding structural feature maps of those scales.

The corresponding region of the interest can be marked by the ground truth masks of the labelled images and estimated masks of the unlabelled images by resizing to the same size with the corresponding feature maps as $M_k^c \in [M_1^c, \cdots, M_K^c]$. Then multi-scale $E_k^c \in [E_1^c, \cdots, E_K^c]$ and multi-scale $S_k^c \in [S_1^c, \cdots, S_K^c]$ for the c_{th} class are computed in the same way as T by Eq. 4 and Eq. 5:

$$E_k^c = \sum_{w,h,z} (x_k \odot M_k^c) \,/\, \sum_{w,h,z} M_k^c, \tag{4}$$

$$S_k^c = \sum_{w,h,z} (\hat{x_k} \odot M_k^c) \,/\, \sum_{w,h,z} M_k^c. \tag{5}$$

Finally, the overall affinity features is defined as the summation of the multi-scale $[E_1^c, \cdots, E_K^c]$ and $[S_1^c, \cdots, S_K^c]$ for the all the classes, collecting two different types of information.

C. Unpaired Consistency Loss

Let E_l and E_u (S_l and S_u) denote the class-specific representations computed from Eq. 4 and 5 for contextual (structural) representations from labelled and unlabelled images respectively. Intuitively, for unlabelled images, their class-specific representations for both the structural and contextual views should be consistent with those from the labelled ones; *i.e.,* E_u (S_u) should be similar to E_l (S_l) at the feature level, coined as affinity consistency in our text. Fig. 6 shows the affinity learning between the labelled images and the unlabelled images. It is noted that both E_l and S_l are computed based on ground-truth masks, therefore providing a strong supervision signal for learning information from unlabelled images. Naturally, medical images are decoupled into a multi-scale representation in CNNs. Therefore, task-affinity consistency should be applied at multiple scales. Overall, we introduce and formulate a loss to incorporate this unpaired *multi-scale* consistency as follows:

$$\begin{aligned} \mathcal{R} = &\frac{\tau_1}{C} \sum_{c}^{C} \sum_{k}^{K} \|(E_{k,l}^c - E_{k,u}^c)\|_2^2 \\ &+ \frac{\tau_2}{C} \sum_{c}^{C} \sum_{k}^{K} \|(S_{k,l}^c - S_{k,u}^c)\|_2^2, \end{aligned} \tag{6}$$

where l denotes the labelled images, u represents the unlabelled images. K is the number of layers corresponding to different scales, where the affinity consistency is computed. For each layer i, we employ the L_2 distance to measure the similarity between the class-specific representations of the labelled images and the class-specific representations of unlabelled images on class c. The final consistency loss is accumulated over all the C target classes and is normalized by the total number of classes C. τ_1 and τ_2 are two hyperparameters to control the trade-off between the contextual affinity loss and the structural affinity loss.

D. Total Loss

Our final loss L is a combination of the supervised loss for the labelled images and the task-affinity consistency loss between the labelled and unlabelled images. It is formulated as follows:

$$\mathcal{L} = \sum_{x_l, y_l} \mathcal{L}_s(x_l, y_l) + \sum_{x_u} \sum_{x_l, y_l} \mathcal{R}((x_u, y_e), (x_l, y_l)), \tag{7}$$

where $\mathcal{L}_s$ denotes the supervised loss on labelled samples; Dice loss is used in our experiment. R denotes the consistency loss between the labelled samples and unlabelled samples. x_l is a labelled image with ground truth masks y_l, whilst x_u denotes an unlabelled image with the online estimated masks y_e. It is worth noting that our SSL method is designed for the features of the network; It could be applied to both the 2D and 3D networks, as will be demonstrated in the experimental section.

IV. EXPERIMENT

A. Datasets and Settings

Our method is evaluated on three different datasets, ranging from atrium segmentation to hippocampus segmentation. It involves both binary and multi-class segmentation tasks. The datasets are: the two-class 3D left atrium segmentation dataset [34], the three-class 3D hippocampus segmentation dataset [27] and the four-class 2D MS-CMR segmentation dataset [4], [38]. For each dataset, when doing semi-supervised learning, a subset of training images is randomly selected as labelled images with a *fixed* split ratio, the rest is treated as unlabelled. For a more robust estimation, this random split is repeated *eight* times, and average performances are reported together with standard deviations for each dataset to show both the performance gain and reliability. To the best of our knowledge, our protocol with repeated random split for a fixed split ratio is more rigorous than those evaluation protocols using a single run with a specific split adopted by most of the state-of-the-art methods [18], [22], and [36].

1. Two-class 3D left atrium segmentation.

This is a dataset presented at the MICCAI 2018 Atrial Segmentation Challenge [34], termed *3D-LASeg* in our experiments. Atrium segmentation is an important task for clinicians to assess the level of atrial fibrillation, a common type of arrhythmia in heart diseases. However, due to the low contrast between the atrial cavity and the surrounding background, it is challenging to segment the left atrial directly from 3D scans. The organizer of the challenge provided 100 3D Gadolinium-Enhanced Magnetic Resonance Imaging scans, as well as their corresponding 3D binary masks of the left atrial cavity. Note that this split is conducted across patient IDs such that no patient scans in the testing set are seen in the training. In our experiment, this dataset is repeated eight times for evaluation. Specifically, following the state-of-the-art methods [22], [18], [36], 80 3D scans are used for training, and 20 scans for testing. 10% (8 scans) and 20% (16 scans) are used as labelled images, and the rest of the training set is the unlabelled training images.

2. Three-class 3D hippocampus segmentation.

The dataset is provided by Vanderbilt University Medical Center [27] and contains a total of 260 adult 3D MRIs. The

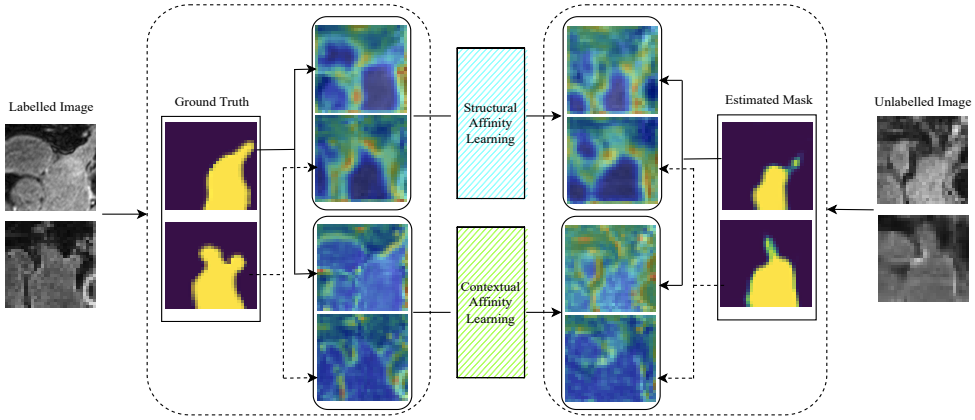

Fig. 6: The proposed affinity learning for the labelled images and unlabelled images, is applied to the contextual affinity features and structural affinity features respectively.

TABLE I: Average metrics of two class 3D left atrium segmentation task.

Method	Scans used		Metrics				Cost
	L	U	DSC (%) ↑	JSI (%)↑	95HD (mm) ↓	ASD (mm) ↓	Params(M)
VNet[23] (*sup*)	80	0	89.32±1.86	81.83±1.98	8.06±2.43	1.50±0.15	9.45
VNet (*inf*)	8	0	77.12±5.05	65.34±5.29	19.87±4.09	4.19±1.34	9.45
VNet (*inf*)	16	0	83.84±2.61	73.86±2.85	12.86±2.51	2.37±0.29	9.45
MT[29]	8	72	79.74±5.43	68.58±6.43	15.28±4.33	3.03±0.86	9.45
EM[32]	8	72	81.00±5.37	70.11±6.16	13.90±3.60	3.02±1.13	9.45
ICT[31]	8	72	80.29±5.95	69.60±6.87	14.81±3.91	2.94±0.83	9.45
UAMT[36]	8	72	83.91±3.29	73.31±4.28	14.74±3.27	4.10±0.81	9.45
SASSNET[18]	8	72	84.11±3.64	74.22±4.00	12.08±2.53	**2.70±0.56**	20.47
DTC[22]	8	72	83.90±3.42	73.63±3.99	12.65±3.18	2.90±0.89	9.45
Ours	8	72	**84.73±2.45**	**74.38±3.43**	**11.45±1.88**	2.72±0.69	9.47
MT[29]	16	64	82.32±3.84	71.81±4.74	14.06±3.15	2.39±0.36	9.45
EM[32]	16	64	83.72±2.83	73.26±3.49	12.03±2.36	2.42±0.29	9.45
ICT[31]	16	64	84.66±4.01	74.60±5.23	11.59±3.10	2.23±0.41	9.45
UAMT[36]	16	64	86.02±2.36	76.47±2.94	10.43±1.98	2.75±0.45	9.45
SASSNET[18]	16	64	86.45±2.76	77.04±3.47	10.52±2.33	2.21±0.21	20.47
DTC[22]	16	64	84.02±4.01	74.44±4.81	12.41±3.45	2.15±0.36	9.45
Ours	16	64	**87.75±1.49**	**78.60±2.18**	**9.45±1.57**	**2.04±0.30**	9.47

TABLE II: Average metrics of three class 3D hippocampus segmentation task.

Method	Scans used		Metrics				Cost
	L	U	DSC (%) ↑		95HD (mm) ↓		Params (M)
			body	head	body	head	
VNet[23] (*sup*)	210	0	81.18±1.21	82.02±1.43	1.98±0.19	2.07±0.16	9.45
VNet (*inf*)	10	0	73.89±4.77	73.59±1.66	2.65±0.32	2.97±0.69	9.45
MT[29]	10	200	67.84±6.72	72.68±5.72	4.60±2.41	3.77±1.40	9.45
EM[32]	10	200	66.59±13.92	71.64±4.60	5.00±3.87	4.37±3.53	9.45
ICT[31]	10	200	57.95±18.39	63.88±10.92	7.76±6.59	7.56±3.16	9.45
UAMT[36]	10	200	72.59±5.48	74.34±2.73	2.93±0.74	6.91±2.07	9.45
SASSNET[18]	10	200	70.20±3.85	72.06±3.81	2.86±0.35	2.87±0.35	20.47
DTC[22]	10	200	72.41±2.32	74.21±3.15	2.64±0.39	2.69±0.28	9.45
Ours	10	200	**75.22±1.81**	**76.55±1.38**	**2.61±0.31**	**2.52±0.23**	9.47

TABLE III: Average metrics of four class 2D MS-CMRSeg task.

Method	Scans used		Metrics						Cost
	L	U	DSC (%) ↑			95HD (mm) ↓			Params (M)
			RV	myo	LV	RV	myo	LV	
UNet[26] (*sup*)	35	0	90.38±0.97	78.03±1.03	73.30±4.70	4.78±1.87	4.30±1.27	9.68±3.25	2.47
UNet (*inf*)	7	0	83.44±4.11	66.85±3.58	59.92±2.94	14.34±6.70	11.00±4.84	19.75±2.36	2.47
MT[29]	7	28	84.91±3.49	67.94±4.59	65.70±5.29	11.36±6.43	9.49±3.94	**15.87±3.29**	2.47
EM[32]	7	28	84.63±1.91	67.96±3.77	64.39±4.80	11.64±6.44	9.13±3.65	17.80±3.21	2.47
ICT[31]	7	28	84.68±2.79	68.72±3.95	66.19±3.93	12.38±5.48	9.37±3.69	15.90±2.30	2.47
UAMT[36]	7	28	83.19±3.03	67.61±4.74	64.85±4.72	11.26±6.02	11.41±4.92	17.76±3.27	2.47
SASSNET[18]	7	28	85.19±2.57	70.15±4.35	64.99±4.04	12.14±7.56	9.22±4.23	16.87±5.32	5.23
DTC[22]	7	28	80.02±4.22	67.70±4.29	61.02±4.81	**7.30±1.97**	9.08±3.04	19.81±2.92	2.47
Ours	7	28	**88.31±1.44**	**72.68±2.11**	**68.91±3.09**	13.47±4.95	**7.15±4.57**	19.50±7.67	3.19

hippocampus head and body are delineated for each scan. All scans are captured as a 3D T1-weighted MPRAGE sequence with a Philips Achieva scanner. The voxel size of each scan is 1.0 mm^3. In our experiment, we use 210 images as the training set and the remaining 50 images as the testing set. As with the previous dataset, the 210/50 random split is conducted eight times for a robust performance assessment. In the 3D segmentation task, our method is further tested in a more challenging case in which 5% of training data is used for supervised learning, and the rest of the training images are treated as unlabelled.

3. Four-class 2D MS-CMR segmentation.
This dataset is the one used for the MS-CMRSeg 2019 [4], [38] contest and contains 45 patient cases with cardiomyopathy. Each case contains LGE MRI (Late-gadolinium enhance-MRI), T2-weight MRI, and bSSFP MRI (balanced Steady-State Free Procession-MRI) modalities; three-class segmentation masks are presented: left ventricle, myocardium, and right ventricle. The multi-organ segmentation of the LGE CMR image is more challenging because it is difficult to delineate the contour of the myocardium on the LGE image alone, and to get a more accurate boundary, it is usually necessary to rely on other modal information, *i.e.,* the bSSFP modality [21], [28]. In our experiment, we use only the LGE modality, a harder task than the contest setting. Note that the resolution of the 3D scan along the z-dimension is too small, as each LGE raw case has 10-16 slices. Therefore, a 2D segmentation task is conducted on this dataset, and segmentation is performed on each of the 2D slices. For a fair comparison, in testing, performances are reported on 3D cases. For this dataset, the ratio of the training and test split is set to 35/10 and such a random split is repeated eight times with performance averaged over all the splits. For the 35 cases in the training set in each split, 20% (7 cases) are used as labelled images and the remaining 80% (28 cases) are treated as unlabelled images. Again, there is no overlap in patient IDs between the training and testing sets.

B. Network and Implementation Details

To evaluate the performance of our method, 3D *VNet* [23] and 2D *UNet* [26] are employed as the backbones for 3D and 2D segmentation tasks respectively.

In each iteration, the training batch is a half-and-half combination of labelled and unlabelled images. It is worth noting that, in our framework, we do not use the memory bank as in [49] to store the feature representations as building a memory bank requires significant storage; instead, we directly extract the class-specific representations in real time in each training mini-batch and do not store those representations from the labelled cases, because using this batch-wise sampling method could achieve similar performance with the use of memory bank as in [43]. In this way, our strategy will reduce the costs, and thus a GPU with 12G of RAM is sufficient to run our model. The unpaired images are separately fed to the network after a z-score normalization. The SGD optimiser is used in the training process with a momentum of 0.9 and weight decay of 0.0001. The learning rate decays via 'poly' learning rate policy: $(1-epoch/epoch_{max})^{0.9}$ [11]. We apply three augmentations during training: randomly flipping, rotation, and cropping. For the cropping, the patches of size 112 $\times$ 112 $\times$ 80 are cropped for the left atrium task (the same setting as in [22], [18] and [36]), 224 $\times$ 224 for cardiac segmentation (the same setting as in [4] and [5]) and 32 $\times$ 32 $\times$ 64 for the hippocampus task. τ_1 and τ_2 are set to 0.5 and 1 consistently in all experiments for the two 3D tasks, and are set to 1 and 1 for the 2D task. The effects of those parameters are also tested in the ablation study. For the two-class 3D left atrium segmentation task, we follow exactly the same setting as in [32], [18], [22], and the same evaluation metrics: DSC, JSI, 95HD and ASD. For the three-class 3D hippocampus segmentation task and four-class 2D MS-CMRSeg task, due to the multi-class comparison, we show the performance in terms of the commonly used metrics DSC and 95HD.

C. Comparison with State-of-the-arts

Our method is compared to six state-of-the-art semi-supervised segmentation methods, including Mean Teacher [29] (MT), Adversarial Entropy Minimization [32] (EM), Interpolation Consistency Training [31] (ICT), Uncertainty-aware Self-ensembling Model [36] (UAMT), the Shape-aware approach [18] (SASSNET) and the Dual-task Consistency approach [22] (DTC). The published code of each method was used for the implementation in this experiment.

For each split, we also report performances of backbones in two *purely* supervised learning cases: 1) using all the scans with provided labels for training (denoted as *sup*); 2) only using the selected labelled images for training, *i.e.,* without using any unlabelled images for that split, denoted as *inf*. Ideally, the performances of *sup* and *inf* could serve as the *upper* bound and *lower* bound of the performances of all the methods compared, namely, MT, EM, ICT, UAMT, SASSNET, DTC and Ours, in each semi-supervised setting.

We would like to highlight that, to date, two-class medical segmentation tasks are the most commonly used to evaluate methods in semi-supervised settings; while multi-class (2+) segmentation tasks, although more challenging, are much *less* explored for semi-supervised learning. In our comprehensive study, we compared our method against others in a larger scale manner, from 2 to 4 classes medical image segmentation tasks.

1) Results of 2-class 3D left atrium segmentation task: Table I tabulates the results obtained by respectively using 20% (16/64 split) and 10% (8/72 split) of the 80 scans as labelled during training. It could be observed that our method outperforms all others in 7 out of 8 cases, and is the second-best in terms of ASD when using an 8/72 split (Table I). Specifically, in terms of the most commonly used metric, our method outperforms all others for both *10%* (8/72 split) and *20%* (16/64 split) settings, *e.g.* a 1.15% increase (87.75% vs. 86.45%) in DSC when comparing with SASSNET using 16/64 split; for the 8/72 split, performance increases from 84.11% to 84.73%. In particular, it is worth noting that the standard deviations of our method are the lowest, 2.45% when using 10% images as labelled and 1.49% by using 20% images as labelled. This clearly demonstrates that our method is

TABLE IV: The effect of contextual/structural module in left atrium segmentation task.

Left Atrium	Scans used		Metrics			
	labelled	Unlabelled	DSC (%) ↑	JSI (%) ↑	95HD (mm) ↓	ASD (mm) ↓
w/o structural module	8	72	83.71±3.00	73.06±4.03	12.73±2.11	3.03±0.80
w/o contextual module	8	72	83.45±2.99	72.91±3.88	12.67±2.77	**2.72±0.60**
contextual and structural modules	8	72	**84.73±2.45**	**74.38±3.43**	**11.45±1.88**	**2.72±0.69**
w/o structural module	16	64	86.32±2.36	77.00±2.90	10.65±2.43	2.24±0.42
w/o contextual module	16	64	85.27±4.70	75.80±5.63	11.86±4.13	2.05±0.36
contextual and structural Modules	16	64	**87.75±1.49**	**78.60±2.18**	**9.45±1.57**	**2.04±0.30**

TABLE V: The effect of contextual/structural module using *ten* scans in hippocampus segmentation task.

Hippocampus	Scans used		Metrics			
	L	U	DSC (%) ↑		95HD (mm) ↓	
			body	head	body	head
w/o structural module	10	200	68.49±9.58	**77.35±0.71**	4.24±2.83	2.71±0.35
w/o contextual module	10	200	68.55±7.98	76.01±1.96	4.15±2.29	2.80±0.46
Contextual and Structural Modules	10	200	**75.22±1.81**	76.55±1.38	**2.61±0.31**	**2.52±0.23**

TABLE VI: The effect of contextual/structural module using *seven* scans in MS-CMRSeg task.

Method	Scans used		Metrics					
	L	U	DSC (%) ↑			95HD (mm) ↓		
			RV	myo	LV	RV	myo	LV
w/o structural module	7	28	87.59±1.58	71.78±2.41	68.12±4.65	**13.37±4.80**	7.74±3.24	**13.80±5.34**
w/o contextual module	7	28	87.69±1.42	72.23±2.07	67.60±2.63	16.03±4.02	7.56±3.96	20.01±5.44
Contextual and Structural Modules	7	28	**88.31±1.44**	**72.68±2.11**	**68.91±3.09**	13.47±4.95	**7.15±4.57**	19.50±7.67

robust to different random splits at the same ratio, showing a better generalisation ability. In terms of JSI, 95HD and ASD, our method also achieves the best results, and the number of parameters is comparable to the backbone with a slight increase of 0.02M, which does not introduce many additional parameters. We also conducted Wilcoxon t-tests on DSC and 95HD, and all the p-values are less than 0.05, indicating our improvement is statistically significant.

2) Results of 3-class 3D hippocampus segmentation task: Our method is compared to the others on a three-class 3D hippocampus segmentation task, a task usually more challenging than two-class segmentation in semi-supervised learning. Performances for each of the segmentation targets (head and body) are reported in Table II with only 5% of training data labelled (10 cases). Again, for the sake of performance calibration, we also report the performances when all the training data are used as labelled (denoted as *sup*), and the performances when 5% are used as labelled during training *without* using any unlabelled data (denoted as *inf*).

Compared to the performance of VNet (*inf*), i.e. 73.59% by purely supervised learning, both DCT and UAMT successfully improve the performance by leveraging the unlabelled images. However, it is also interesting to note that other methods, including MT, EM and ICT, present a lower performance than the VNet (*inf*) baseline. In particular, the performance of ICT has dropped a lot and the standard deviation is large (10.92%) for the hippocampus task. This low performance of ICT might be attributed to the fact that the mean teacher needs to get an update from the students model. When there is a limited number of labelled images for training (as low as 5% in this case), the student may not get enough supervision, which will degrade the performance of the teacher module. Therefore a co-optimisation between teacher and student by simply sharing weights will not necessarily lead to a better model than the baseline. On the contrary, our method achieves the highest performance among all the methods compared in the semi-supervised segmentation with only 5% of the images used as labelled; specifically compared to the VNet (*inf*) baseline, it obtains an improvement of 2.96% for the head region, and 1.33% for the body region in terms of DSC. The p-values of Wilcoxon t-tests on DSC and 95HD are less than 0.05. This clearly demonstrates that the proposed task-affinity consistency is very effective in learning to transfer knowledge between the labelled and unlabelled images (*i.e.,* unpaired), even in semi-supervised *multi-class* segmentation settings.

3) Results of 4-class 2D MS-CMRSeg task: Our method is further tested and compared to other methods on a 4-class segmentation task in a 2D setting. Table III shows the results of all the methods, together with the two baselines using *purely* supervised learning. Due to the low resolution of z-axis, we use a 2D UNet [26] as backbone. 2D single slices for a patient were fed into the 2D network for segmentation, and those segmented slices were then rearranged into a 3D volume to evaluate the segmentation performance for that patient. Since we set the number of feature channels in the 2D network to 4 times that of the 3D network, therefore, we need to introduce more parameters when calculating the structural affinity features (2.47M to 3.19M). Compared to the other methods, our method performs the best in most of the cases, e.g. when comparing to the SASSNET, our method improves the performance by 3.12% (RV), 2.53% (myo), 3.92%(LV) in terms of DSC, with much lower variance, which indicates that our method can achieve better and more stable results than the state-of-art methods. We again conducted Wilcoxon t-tests on DSC and 95HD, with p-values being all less than 0.05, which indicates our improvement is statistically significant.

TABLE VII: The effect of hyper-parameters in mean and std of DSC of LA segmentation: τ_1 and τ_2.

τ_1	0.25				0.5				0.75				1			
τ_2	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1	0.25	0.5	0.75	1
mean	86.05	86.99	87.03	86.81	86.96	87.03	86.66	**87.75**	86.78	86.64	87.24	87.14	86.65	86.52	86.78	86.36
std	2.84	2.33	2.08	2.49	2.14	1.89	2.15	**1.49**	1.95	2.50	2.02	1.75	2.43	2.31	1.82	2.35

D. Ablation Studies

In this section, we further conduct a set of experiments to test the effects of different components of our framework including the structural and contextual modules, the influence of the hyperparameters such as the weights in the Eq.6 balancing the trade-off between the structural and contextual modules, as well as the effect of scales in the affinity learning.

1) The effect of structural affinity module: We test the effects of with and without using the structural affinity module, on four different cases following our state-of-the-art comparisons: 1) using 16 scans as labelled in the left atrium segmentation task; 2) using 8 scans only as labelled in the LA segmentation task; 3) using 10 scans as labelled in the hippocampus segmentation task; 4) using 7 scans as labelled on the MS-CMRSeg task. Results are shown in Table IV, V and VI. It could be seen that in the LA segmentation task (10% and 20% labelled cases), using the contextual module outperforms the model without the structural module. For example, the model performance in terms of DCS using the structural and contextual modules can improve by 1.43% and 1.02% (from 86.32% to 87.75% in LA segmentation with 16 scans as labelled, and 83.71% to 84.73% in LA segmentation with 8 scans as labelled) respectively. For the hippocampus segmentation and MS-CMRSeg task, the model with the structural module outperforms the model w/o the structural module in terms of the *mean* value of DSC, *i.e.,* with an improvement of 2.97% and 0.8% separately. In terms of the 95HD, the structural module is not a clear advantage on some region tasks. However, by taking all testing cases together, the structural module clearly improves the performance in a majority of the cases.

2) The effect of contextual affinity module: We also tested the effect of the contextual module, and results are included in Tables IV, V and VI. Similar conclusions can be drawn for the effect of the contextual module. Models including the contextual module outperform the models that do not use it in most of the cases. It could be seen that contextual-only (w/o structural) works slightly better than structural-only (w/o contextual), whilst structural-only works better in a few cases in 95HD. The combination of these two modules works best overall, which implicitly suggests the two modules are complementary.

3) The effect of the number of blocks in affinity learning: The different depths of blocks can explore different kinds of information, the higher layers learn more high-level discriminative representations, while earlier layers capture more general and low-level visual information [51], [52]. In our method, we explore the effect of the different number of blocks in affinity learning. As shown in Fig. 2, our affinity learning is applied to the features on four blocks, denoted as B1, B2, B3 and B4. Due to the sequential convolution in our method, different blocks extract features at different scales and the combination of those features presents a multi-scale representation of the input, with B1 corresponding to a coarser scale and B4 to a finer scale. To explore the effect of the number of blocks used in affinity learning, we systemically vary the number of blocks (the value of K in Eq.6) used in our affinity learning, by recursively adding one block to B4; thus presenting four cases for testing 1) no blocks, 2) B1, 3) B2, 4) B3, 5) B4, 6) B1+B2 (denoted as B1-2), 7) B1+B2+B3 (denoted as B1-3) and 8) B1+B2+B3+B4 (denoted as All). We test the performance of the model in all of those cases when taking 16 scans as labelled on the LA segmentation task. Box plots of the results over eight random splits are shown in Fig. 7. It could be observed that when only one block is used for affinity learning, the performance (the average of DSC over 8 runs) of the later blocks is generally higher than that of the earlier blocks. Adding more blocks in general could improve the results. In particular, our approach using all the blocks outperforms without using affinity learning significantly, *i.e.,* from 83.84% to 87.75%.

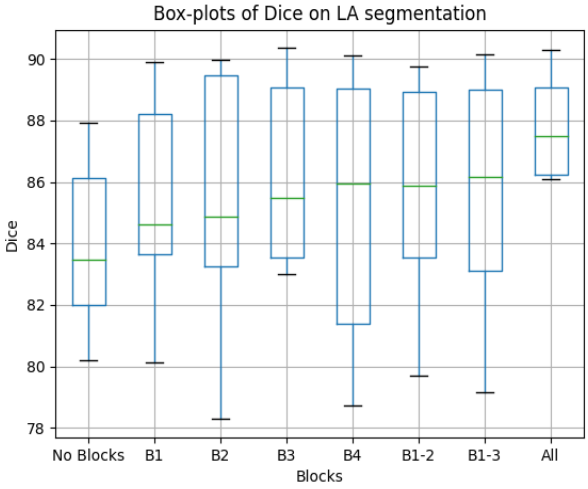

Fig. 7: The effect of number of blocks: from left to right: 1) No Blocks; 2) B1; 3) B2; 4) B3; 5) B4, 6) B1-2; 7) B1-3; 8) All.

4) The effect of weights balancing structural and contextual affinity modules: It is noted that the weights τ_1 and τ_2 in Eq. 6 are used to control the trade-off between the structural module and the contextual module in the consistency loss. To test their effects, we systematically vary the values of τ_1 and τ_2 within the range of [0, 1] with a step of 0.25, and conduct the experiments on the left atrium segmentation task with 16 scans used as labelled. Results averaged over eights random splits are listed in Table VII. It could be observed that, the performances with respect to different weight values of τ_1 and τ_2 remain relatively *stable*, and the model with $\tau_1 = 0.5$ and

$\tau_2 = 1$ achieves slightly better performance.

5) The performance of different ratios of the labelled data: We evaluated the LA dataset with 2.5%, 5%, 10%, 20%, 40%, 50% of the training data over eight independent runs. The results are shown in Fig. 8. It can be seen that our method improves the DSC in all settings (42.06%, 36.79%, 7.61%, 3.91%, 1.41%, 0.67% improvement with 2.5%, 5%, 10%, 20%, 40%, 50% of the training data). Thus, our task-affinity consistent approach demonstrates its potential as a semi-supervised segmentation algorithm in practical applications with both a small number of and a large number of labelled data.

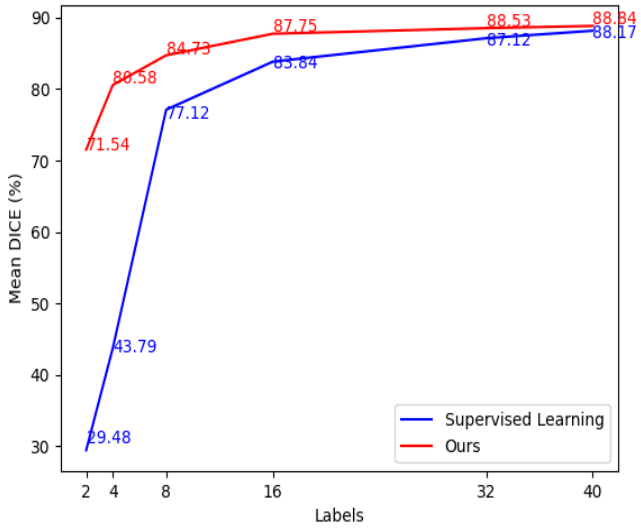

Fig. 8: Plot of the mean DSC w.r.t different ratio of the training set over the LA Dataset (2.5%, 5%, 10%, 20%, 40% and 50% of the labelled data for training), with the blue line representing our baseline: supervised learning, the red line representing our proposed method.

6) The effect of the enhancement backbone: We developed variants of other compared methods by inserting our enhancement module into their original version and conduct all the experiments by using the LA dataset with 20% labelled data. Our method is compared with those variants of state-of-the-art methods. The results are summarized in Tab. VIII. Our method outperforms all six methods with the variant backbone.

TABLE VIII: Enhancement backbone comparison with state-of-the-art methods on LA dataset with 20% labelled data.

Method	DSC (20% labelled cases)
VNet (*sup*)	89.316±1.863
VNet (*inf*)	83.00±3.57
MT	83.88±4.32
EM	82.91±5.11
ICT	84.31±4.59
UAMT	85.46±3.86
SASSNET	86.16±2.77
DTC	85.91±3.63
Ours	87.75±1.49

V. VISUALIZATION

Figure 9 visually compares the segmentation boundaries of different methods in the LA segmentation task. For clarity, the first row (A) compares our method with the DTC, SASSNET, and UAMT methods. The second row (B) shows the comparison between our method and ICT, EM, and MT methods. Both the ground truth mask and the segmented boundary of our method are imposed on the copies in both rows. The predicted boundaries of the atrium by our method are more accurate than others, *i.e.,* much closer to the ground truth masks.

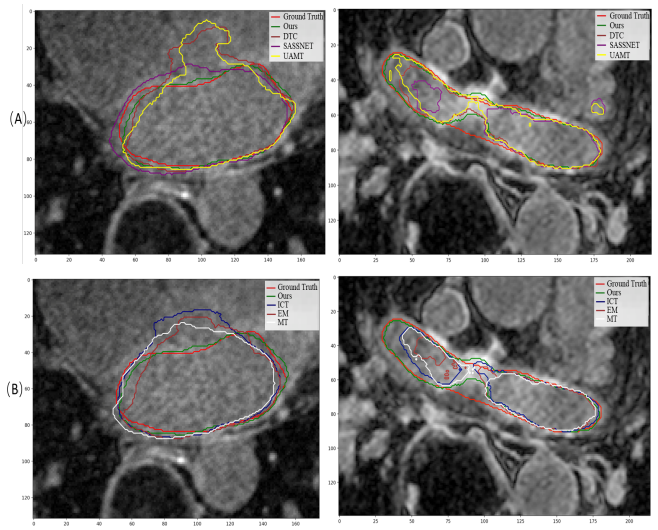

Fig. 9: Visual comparison of segmentation boundaries of different methods, marked in different colours. For clarity, we split the compared methods into two sets: 1) DTC, SASSNET, and UAMT shown in the first row; 2) ICT, EM, and MT compared in the second row. Each column shows the results on two copies of the same image, with our result (green) and ground truth (red) imposed on both copies of the same image. It could be clearly seen that our method produces a much better segmentation boundary than all others.

VI. CONCLUSION

In this paper, we have proposed a novel semi-supervised method based on task-affinity consistency regularization in the feature maps. The structural affinity module and contextual affinity module are introduced to separate the structural information and contextual information at the feature level. This separation method can be used to transfer the class-specific knowledge from the labelled images to the unlabelled images. Because our task affinity consistency regularization method is based on the feature maps. It becomes easier to apply to other mainstream segmentation networks, which indicates the usability and scalability of the method. At the same time, our method achieves the best on DSC and achieves superior results to other methods on most of the other metrics in LA (10% and 20% labelled setting), MS-CMRseg and HP segmentation tasks.

REFERENCES

[1] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in neural information processing systems*, vol. 27, pp. 3365-3373, 2014.

[2] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, Jan. 2015.

[3] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *MICCAI*, Oct. 2019, pp. 810-818.

[4] J. Chen, H. Li, J. Zhang, and B. Menze, "Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac MR images segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Oct. 2019, pp. 317-325.

[5] J. Chen, W. Li, H. Li, and J. Zhang, "Deep class-specific affinity-guided convolutional network for multimodal unpaired image segmentation," in *MICCAI*, Oct. 2020, pp. 187-196.

[6] W. Cui *et al.*, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *International Conference on Information Processing in Medical Imaging*, Jun. 2019, pp. 554-565.

[7] R. Fan, X. Jin, and C. C Wang, "Multiregion segmentation based on compact shape prior," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 3, pp. 1047-1058, 2014.

[8] H. U. Gerth, K. U. Juergens, U. Dirksen, J. Gerss, O. Schober, and C. Franzius, "Significant benefit of multimodal imaging: PET/CT compared with PET alone in staging and follow-up of patients with Ewing tumors," *Journal of Nuclear Medicine*, vol. 48, no. 12, pp. 1932-1939, 2007.

[9] R. A Gonzales *et al.*, "Automated left atrial time-resolved segmentation in MRI long-axis cine images using active contours," *BMC Medical Imaging*, vol. 21, no. 1, pp. 1-12, 2021.

[10] S. Y. Ho, J. A. Cabrera, and D. Sanchez-Quintana, "Left atrial anatomy revisited," *Circulation: Arrhythmia and Electrophysiology*, vol. 5, no. 1, pp. 220-228, 2012.

[11] F. Hutter, L. Kotthoff, and J. Vanschoren, Automated machine learning: methods, systems, challenges, Gewerbestrasse 11, 6330 Cham, Switzerland: Springer Nature Switzerland AG, 2019.

[12] A. Khosravanian, M. Rahmanimanesh, P. Keshavarzi, and S. Mozaffari, "Fuzzy local intensity clustering (FLIC) model for automatic medical image segmentation," *The Visual Computer*, vol. 37, pp. 1185-1206, 2021.

[13] D. H Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, Jun. 2013, vol. 3, no. 2, pp. 896.

[14] H. J. Lee, J. U. Kim, S. Lee, H. G. Kim, and Y. M. Ro, "Structure boundary preserving segmentation for medical image with ambiguous boundary," in *CVPR*, 2020, pp. 4817-4826.

[15] B. Lei *et al.*, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Medical Image Analysis*, vol. 64, pp. 101716, 2020.

[16] A. Levine, S. Singla, and S. Feizi, "Certifiably robust interpretation in deep learning," *arXiv preprint arXiv:1905.12105*, 2019.

[17] C. Li, and H. Liu, "Generative Adversarial Semi-Supervised Network For Medical Image Segmentation," in *ISBI*, Apr. 2021, pp. 303-306.

[18] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *MICCAI*, Oct. 2020, pp. 552-561.

[19] X. Li, L. Yu, H. Chen, C. W. Fu, L. Xing, and P. A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523-534, 2020.

[20] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *MICCAI*, Oct. 2020, pp. 614-623.

[21] Y. Lu, G. Wright, and P. E. Radau, "Automatic myocardium segmentation of LGE MRI by deformable models with prior shape data," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, pp. 1-2, 2013.

[22] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised Medical Image Segmentation through Dual-task Consistency," in *AAAI*, 2021, vol. 35, no. 10, pp. 8801-8809.

[23] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, Oct. 2016, pp. 565-571.

[24] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic Evaluation of Deep Semi-Supervised Learning Algorithms," *Advances in Neural Information Processing Systems*, vol. 31, pp. 3235-3246, 2018.

[25] M. N.Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning," in *ICLR*, Sep. 2020.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, Oct. 2015, pp. 234-241.

[27] OA. L. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[28] Q. Tao, S. R. Piers, H. J. Lamb, and R. J. van der Geest, "Automated left ventricle segmentation in late gadolinium-enhanced MRI for objective myocardial scar assessment," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 2, pp. 390-399, 2015.

[29] A. Tarvainen, and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, Dec. 2017, pp. 1195-1204.

[30] J. E. Van Engelen, and H. H Hoos, "A survey on semi-supervised learning. Machine Learning," *Machine Learning*, vol. 109, no. 2, pp. 373-440, 2020.

[31] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *IJCAI*, 2019, pp. 3635-3641.

[32] T. H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, 2019, pp. 2517-2526.

[33] Y. Xia *et al.*, "3d semi-supervised learning with uncertainty-aware multi-view co-training," in *WACV*, 2020, pp. 3646-3655.

[34] Z. Xiong *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, pp. 101832, 2021.

[35] X. Yang, Z. Song, I. King, and Z. Xu, "A Survey on Deep Semi-supervised Learning," *arXiv preprint arXiv:2103.00550*, 2021.

[36] L. Yu, S. Wang, X. Li, C. W. Fu, and P. A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *MICCAI*, Oct. 2019, pp. 605-613.

[37] Y. Zhou *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical image," in *CVPR*, 2019, pp. 2079-2088.

[38] X. Zhuang *et al.*, "Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge," *arXiv preprint arXiv:2006.12434*, 2020.

[39] X. Liu *et al.*, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, pp. 1224, 2021.

[40] L. Weng, "From gan to wgan," *arXiv preprint arXiv:1904.08994*, 2019.

[41] D. Mahapatra *et al.*, "Semi-supervised learning and graph cuts for consensus based medical image segmentation," *Pattern recognition*, vol. 63, pp. 700–709, 2017.

[42] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," *CVPR*, pp. 9729–9738, 2020.

[43] L. Liu *et al.*, "Bootstrapping Semantic Segmentation with Regional Contrast," *ICLR*, 2022.

[44] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[45] J. Xie *et al.*, "Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation," *arXiv preprint arXiv:2203.13505*, 2022.

[46] Springenberg, and T. Jost, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.

[47] J. Schlemper *et al.*, "Attention-gated networks for improving ultrasound scan plane detection," *arXiv preprint arXiv:1804.05338*, 2018.

[48] X. Hao *et al.*, "Multimodal magnetic resonance imaging: The coordinated use of multiple, mutually informative probes to understand brain structure and function," *Human brain mapping*, vol. 34, no. 2, pp. 253–271, 2013.

[49] H. Hu, J. Cui, and L. Wang, "Region-Aware Contrastive Learning for Semantic Segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16291–16301, 2021.

[50] F. Cheng *et al.*, "Learning directional feature maps for cardiac mri segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 108–117, 2020.

[51] L. Xu *et al.*, "Multi-class Token Transformer for Weakly Supervised Semantic Segmentation," *arXiv preprint arXiv:2203.02891*, 2022.

[52] L. Wang *et al.*, "Training deeper convolutional networks with deep supervision," *arXiv preprint arXiv:1505.02496*, 2015.