Combinatorial CNN-Transformer Learning with Manifold Constraints for Semi-supervised Medical Image Segmentation

Huimin Huang[1], Yawen Huang[2 †], Shiao Xie[1], Lanfen Lin[1*], Ruofeng Tong[1,3], Yen-Wei Chen[4*], Yuexiang Li[5], Yefeng Zheng[2]

[1] Zhejiang University
[2] Jarvis Research Center, Tencent YouTu Lab
[3] Zhejiang Lab
[4] Ritsumeikan University
[5] Medical AI Research Group, Guangxi Medical University

Abstract

Semi-supervised learning (SSL), as one of the dominant methods, aims at leveraging the unlabeled data to deal with the annotation dilemma of supervised learning, which has attracted much attentions in the medical image segmentation. Most of the existing approaches leverage a unitary network by convolutional neural networks (CNNs) with compulsory consistency of the predictions through small perturbations applied to inputs or models. The penalties of such a learning paradigm are that (1) CNN-based models place severe limitations on global learning; (2) rich and diverse class-level distributions are inhibited. In this paper, we present a novel CNN-Transformer learning framework in the manifold space for semi-supervised medical image segmentation. First, at intra-student level, we propose a novel class-wise consistency loss to facilitate the learning of both discriminative and compact target feature representations. Then, at inter-student level, we align the CNN and Transformer features using a prototype-based optimal transport method. Extensive experiments show that our method outperforms previous state-of-the-art methods on three public medical image segmentation benchmarks.

1 Introduction

Medical image segmentation, in the pursuit of integrating boundary detection, region formation and agglomeration for analyzing tissue structures, is a long-standing fundamental task. Recently, convolutional neural networks (CNNs) (Ronneberger, Fischer, and Brox 2015; Zhou et al. 2019b) have achieved remarkable success benefiting from the large-scale annotated dataset. However, collecting pixel-level annotations is expensive and time-consuming, especially for medical images. Semi-supervised learning (SSL) attracts high attention by using both labeled data and large amount of unlabeled data to relieve the pressure of sufficient labeling.

Learning such a paradigm also remains being exposed to ongoing adverse representational conditions, for example, the lack of global attention makes imprecise performance. Specifically, CNN-based SSL methods fail to model the explicit long-range relations beyond local regions (shown in Fig. 1 (c) and (d)), since the receptive field of a network's units is severely limited. To increase model diversity, previous approaches focused on different perturbations (French et al. 2019), varying network structures (Luo et al. 2021a), or various initializations (Ke et al. 2019). However, capturing complementary information is arduous in the later stage of training, as an indirect result of counterproductive decoupling two feature extractors (Zheng et al. 2022). In addition, the existing SSL methods directly leverage pixel-wise predictions from CNNs, which ignore rich class-level dependencies, resulting in very limited capability for accurate segmentation, especially for organs with similar contextual information or/and surrounding position, *e.g.*, a part of pancreas is incorrectly segmented as liver in Fig. 1 (f) and (g).

Recently, Transformers have made remarkable achievements toward establishing long-range dependencies alternative to CNNs and achieving excellent performance in multifarious visual tasks (Dosovitskiy et al. 2020). Although global relations can be well captured by Transformers, the lack of inductive biases and the receptive field of convolutional kernels all lead to less effective learning, let alone the quadratic computational complexity. Fortunately, the above problems can be solved by combining CNNs and Transformers for fully associating local features with global cues. Enlightened by the success of joint learning, in this paper, we explore the essence of CNNs and Transformers for building both local invariant translation and global long-range dependency in semi-supervised medical image segmentation.

Plenty of works have been presented to learn features in general Euclidean space, yet seldom considering the topological structures of data, which is crucial on manifold. In fact, directly operating in Euclidean space is challenging, considering: **(i) Intra-student problem:** Both CNNs and Transformers face the difficulty of inconsistency in the same category and the confused semantics among categories, which all lead to inseparable representations. **(ii) Inter-student problem:** the inner feature and output paradigm of Transformers is heterogeneous from CNNs, which leads to different class distributions. How to learn the complementarity of two-style features and train the Transformer with few annotations remains an open question.

*Corresponding Authors: Lanfen Lin (llf@zju.edu.cn), Yen-Wei Chen (chen@is.ritsumei.ac.jp).
†Huimin Huang and Yawen Huang are co-first authors, and this work is done during the internship at Tencent YouTu Lab.

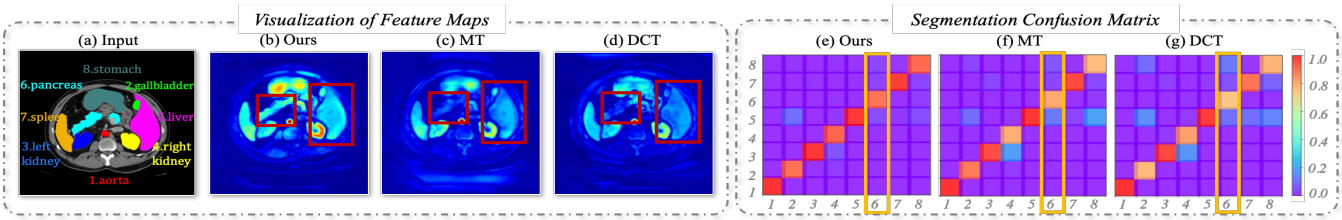

Figure 1: Left: Visualization of feature maps. The CNN-based SSL methods, *i.e.,* MT (in (c)) and DCT (in (d)), cannot capture long-term relations and thus fail to attend on objects beyond local regions (*e.g.,* unactivated pancreas and liver). Right: Segmentation confusion matrix, where the diagonal should be brighter (intra-class compactness), while the rest should be darker (inter-class discrepancy). As observed in yellow boxes, MT (in (f)) and DCT (in (g)) tend to mis-classify pancreas as liver (*e.g.,* brighter of pancreas-to-liver) and under-segment pancreas (*e.g.,* darker of pancreas-to-pancreas). Our M-CnT (in (b)) can attend objects in long-range scenarios, and achieve accurate localization with better compactness (in (e)).

To address the above issues, we define the **M**anifold constraints in combinatorial **C**NN-**T**ransformer learning (termed as M-CnT) for semi-supervised medical image segmentation. The proposed method can be learned synergistically with CNN and Transformer to adaptively learn manifolds of varying structure across samples. There have been some methods (Konstantinidis et al. 2022; Huang et al. 2017) employing manifolds to prove that features in different manifolds carry special statistical and geometrical properties, which bring complementary discriminative power for various tasks. Recent advances in manifold learning reveal two properties (Konstantinidis et al. 2022): (1) allowing similar features to appear closer to each other, while dissimilar features move further apart; (2) providing superiors manifold metric to measure the discrepancy with different statistical and geometrical properties. Inspired by these attempts, M-CnT incorporates more compact and discriminative embeddings in the manifold space considering that: **(1) Intra-student class-wise consistency**: to strengthen the discriminative power of both CNN and Transformer, the intra-class samples are required to be compact, while the inter-class samples are separable. **(2) Inter-student knowledge transfer**: CNN and Transformer have distinct inner feature flow forms, in which their features with complementary class-wise distribution create a potential opportunity for collaboration. In that, the inter-student discrepancy is represented as the distance between two submanifolds, which is then minimized based on the defined manifold metric (*e.g.*, symmetric positive definite metric (Konstantinidis et al. 2022)).

Based on these considerations, we learn an implicit consistency regularization with complementary information for producing more stable pseudo labels to overcome the above deficiencies. As observed in Fig. 1 (b), our M-CnT can recognize objects in varying sizes and long-range scenarios (*e.g.*, large liver and tiny pancreas), owing to local-global cues and class-specific characteristics. M-CnT learns features with better discriminability among different classes (darker non-diagonal), leading to more accurate results in Fig. 1 (e). Our main contributions are summarized as follows: (**I**) We analyze the intra- and inter-student problems raised by the CNN-based SSL methods for semi-supervised medical image segmentation, and propose a novel scheme, named M-CnT, to fully capitalize the unlabeled data in the manifold space. (**II**) We introduce an intra-student class-wise consistency to construct more compact class-wise representations and reduce inter-class dependencies. (**III**) An inter-student knowledge transfer loss is explored to reinforce class-wise statistics in the manifold space and thus strengthen the discrimination of inner features. (**IV**) Extensive experiments are performed on three public datasets, resulting in new state-of-the-art results on different scenarios.

2 Related Work

Semi-supervised Medical image Segmentation. Recent efforts in semi-supervised segmentation have been focused on incorporating unlabeled data into CNNs, which can be largely categorized into four groups: self-training (Bai et al. 2017; Ouali, Hudelot, and Tami 2020a), co-training (Qiao et al. 2018; Zhou et al. 2019a), deep adversarial learning (Zhang et al. 2017; Zheng et al. 2019) and self-ensembling (Π-model (Li et al. 2018) and Mean-Teacher (MT) model (A.Tarvainen and H.Valpola 2017)). For example, Qiao *et al.* (Qiao et al. 2018) achieved Deep Co-Training (DCT) by learning two classifiers on two views. Zhang *et al.* (Zhang et al. 2017) designed a Deep Adversarial Network (DAN), which enforced the segmentation of unlabeled data to be similar to the labelled ones. Yu *et al.* (Yu et al. 2019) proposed a Uncertainty-Aware Mean-Teacher (UA-MT) which enables the student to learn from the reliable targets.

Class-wise Losses in Segmentation. Recent works have shown the advantages of utilizing class-wise statistics in processing intra-class consistency and inter-class distance. Under a fully supervised setting, Li *et al.* (Li et al. 2022) designed an inter-class loss based on Euclidean Distance (ED), which distinguished similar pixels from different categories; while Wang (Wang et al. 2020) proposed the intra-class feature variation (IFVD), which alleviated the difference of class distributions between the student and teacher. In the semi-supervised learning, Alonso *et al.* (Alonso et al. 2021) considered intra-class compactness by using pixel-level contrastive learning (PLCL) to create a better class separation.

Manifold Learning. Recently, manifolds learning has been widely employed in vision tasks (Konstantinidis et al. 2022; Luo et al. 2020; Huang et al. 2017), largely attributes to

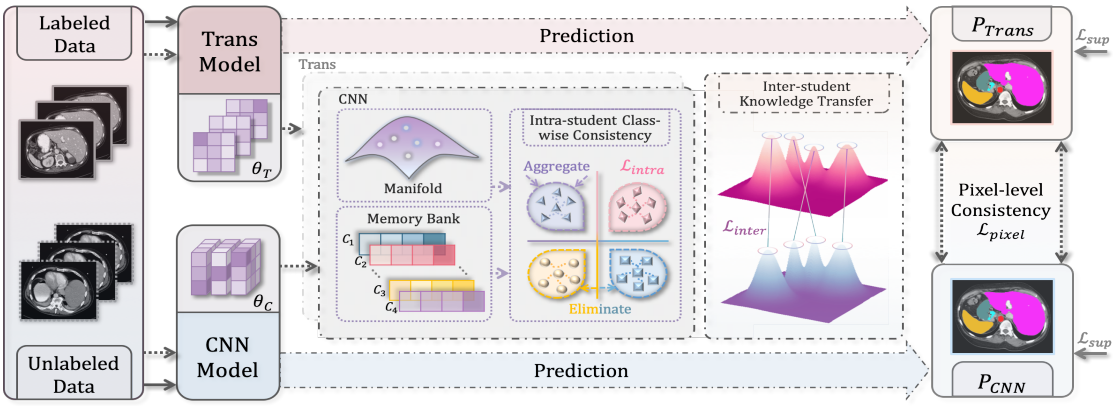

Figure 2: An overview of our M-CnT. It takes both labeled and unlabeled images as input, and obtains two-style predictions (P_{CNN} and P_{Trans}) after feature extraction by a parallel CNN-Transformer architecture. For the labeled image I_l, we apply the commonly-used supervised loss $\mathcal{L}_{sup}$ to update model parameters. As for the unlabeled image I_u, the model is optimized with the pixel-level consistency loss $\mathcal{L}_{pixel}$ and two manifold constrains ($\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$) to explore class statistics.

the discriminative power, which carries statistical and geometrical properties. For example, Konstantinidis *et al.* (Konstantinidis et al. 2022) proposed a multi-manifold attention (MMA) that employed three different manifolds to learning a highly descriptive attention map. Hao *et al.* (Jia et al. 2021) explored the rich patch-level information in manifold space and thus designed a fine-grained distillation loss. Inspired by these attempts, we design an effective manifold constraint to learn their separable representations.

3 Method

In the semi-supervised learning, the training set consists of two parts: a labeled set $D_l = \{(I_l, Y_l)\}_{l=1}^{N}$ and an unlabeled set $D_u = \{(I_u)\}_{u=N+1}^{N+M}$, where Y_l is the ground truth. The overview of our M-CnT is shown in Fig. 2, consisting of two manifold constraints upon the dual-student scheme with CNN-Transformer architecture (Sec. 3.1): (**i**) The intra-student class-wise consistency $\mathcal{L}_{intra}$ is explored to simultaneously perform intra-class aggregation and inter-class elimination (Sec. 3.2); (**ii**) The inter-student knowledge transfer loss is implemented to align class-wise knowledge between two-style students (Sec. 3.3). By combining these two constraints (Sec. 3.4), our M-CnT is able to explore global capability and class-level features from unlabeled data.

3.1 Architecture of CNN-Transformer

Our M-CnT exploits the complementarity of two students: the convolutional student $f(I;\theta_C)$ and Transformer student $f(I;\theta_T)$. Each component will be introduced in detail.

Convolutional Student. The conventional U-Net, which adopts the feature pyramid structure, is employed as backbone to extract local-level context information. Specifically, there are five stages with stacking convolutional blocks in the encoder of U-Net for feature embedding. When the encoder goes deeper, the resolution of feature maps gradually decreases to $\{1/1, 1/2, 1/4, 1/8, 1/16\}$ for the five stages; while the number of channel dimensions accordingly increases to $\{16, 32, 64, 128, 256\}$.

Transformer Student. Different from the convolutional student embedding local cues, Transformer aims to incorporate global representations. In particular, we employ Swin-UNet (Cao et al. 2021) as the backbone for the branch of Transformer student, which is designed in a hierarchical manner with spatial reduction and window-based attention for reducing computational complexity. There are also five stages in the encoder of Swin-UNet. The resolution of feature maps gradually decreases to $\{1/4, 1/8, 1/16, 1/32, 1/32\}$ for five stages, respectively; while the number of channels accordingly increases to $\{96, 192, 384, 768, 768\}$.

Pixel-level Consistency Loss $\mathcal{L}_{pixel}$. Given the unlabeled images I_u, we can obtain the pixel-wise confidence maps (P_{CNN} and P_{Trans}), from the CNN and Transformer branches, and the pseudo label can be calculated via: $\widehat{Y} = \arg\max(P)$. Instead of directly averaging predictions from both students as the pseudo label, we consider the uncertainty in the CNN and Transformer, which can effectively eliminate the unreliable pseudo label. Specifically, we calculate the entropy of pixel-wise probability distribution in both P_{CNN} and P_{Trans} to obtain the pseudo labels of higher quality as: $\mathcal{H}(P)^{h,w} = -\sum_{c=0}^{C-1} P^{h,w}(c) \log P^{h,w}(c)$, where $P^{h,w}$ is the prediction of P_{CNN}/P_{Trans} in the position of (h, w), and C is the number of classes. As the prediction with lower uncertainty is more accurate, we select more confident part from either CNN or Transformer to form the reliable pseudo label $\widehat{Y}_u^{h,w}$ for unlabeled I_u:

$$\widehat{Y}_u^{h,w} = \mathcal{S}^{h,w} \odot \widehat{Y}_{CNN}^{h,w} + (\mathbf{1} - \mathcal{S}^{h,w}) \odot \widehat{Y}_{Trans}^{h,w}, \quad (1)$$

$$\mathcal{S}^{h,w} = \mathbf{1}\left\{\mathcal{H}(P)_{CNN}^{h,w} < \mathcal{H}(P)_{Trans}^{h,w}\right\}, \quad (2)$$

where $\mathcal{S}^{h,w}$ records the position of pixels with lower uncertainty from CNN; while $(\mathbf{1} - \mathcal{S}^{h,w})$ can select the confident pixels from Transformer. Then, we utilize the reliable

pseudo label $\widehat{Y}_u^{h,w}$ to supervise the output of two students by using the Dice loss, which can be denoted as:

$$\mathcal{L}_{pixel} = \frac{1}{2}\left(\mathcal{L}_{dice}(P_{CNN}, \widehat{Y}_u) + \mathcal{L}_{dice}(P_{Trans}, \widehat{Y}_u)\right). \tag{3}$$

Manifold Layers. Enlightened by the successful application of manifold learning paradigm, we aim to apply the manifold constraints to fully exploit the crucial class-wise statistics embedded in the unlabeled data. The representations in the manifold space are built over the final decoder stage before the segmentation head. Specifically, the input feature X is firstly transformed to the same resolution of input image (size of $H \times W$), *i.e.*, $X \in \mathbb{R}^{H \times W \times D}$. Then, it is rearranged into a sequence with size of $HW \times D$. Inspired by the previous work (Luo et al. 2020), we employ the fully-connected layers on the rearranged sequence, which map them into the manifold spaces $\left\{\mathcal{M}^i \in \mathbb{R}^d | i = 1, 2, \ldots, HW\right\}$. To this end, we can obtain CNN-based $\mathcal{M}_{CNN}$ and Transformer-based $\mathcal{M}_{Trans}$ with the same size of $HW \times d$.

3.2 Intra-student Class-wise Consistency

In this section, we investigate the way to generate the separable features in the manifold space for each student. Specifically, we define an **intra-class aggregation loss** to achieve more compact features and an **inter-class elimination loss** to further generate relatively larger margin between classes.

Intra-class Aggregation Loss. In our approach, a prototype-based strategy is employed, which connects the class-specific prototype of unlabeled data (*i.e.*, class center) with the high-quality features extracted from labeled data of the same class. Such that, the indistinguishable unlabeled prototypes can explicitly learn from the representative features from labeled data, in a class-aware manner.

- **Class-specific Prototypes $\mathcal{G}$.** The prototype of c-th class can be calculated by averaging manifold of the class:

$$\mathcal{G}_c = \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \mathcal{M}^i, \tag{4}$$

  where $\mathcal{M}^i$ is the i-th feature in manifold $\mathcal{M}$; $\mathcal{S}_c$ is the set of features having the same label of c, according to the pseudo label $\widehat{Y}_u^{h,w}$ and $|\mathcal{S}_c|$ represents the size of the set.
- **Memory Bank $\mathcal{B}$.** We employ a memory bank $\mathcal{B}$ to collect the high-quality features from the labeled data. Inspired by (Alonso et al. 2021), for each class, we utilize the attention modules to obtain the ranking score for the candidate features (*i.e.,* the ones have the correct predictions to the ground truth). The attention module comprises of two sequential linear functions and a Sigmoid function, which correspondingly yields a score in the range of $[0, 1]$ for each feature. Then, the top-K highest-scoring features are selected and added to the memory bank; hence, the memory bank has the size of $C \times K \times d$.

With the prototypes and memory bank, we force the class-wise prototypes $\mathcal{G}$ to approach their corresponding high-quality class-specific representations in the memory bank $\mathcal{B}$, aiming to shrink the intra-class distribution. Concretely, we utilize the cosine similarity to compute the distance between $\mathcal{G}_c$ and $\mathcal{B}_c^k$, and the loss $L_{aggregate}$ can be defined as:

$$\mathcal{L}_{aggregate} = \frac{1}{C}\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}\left(1 - \frac{\langle \mathcal{G}_c, \mathcal{B}_c^k \rangle}{\|\mathcal{G}_c\|_2 \cdot \|\mathcal{B}_c^k\|_2}\right). \tag{5}$$

Here, $\mathcal{L}_{aggregate}$ can align the ambiguous unlabeled class centers to the confident labeled high-quality features, and make each class distribution more compact, resulting in a good separation of various classes in the latent space.

Inter-class Elimination Loss. Another efficient way to achieve the separable boundary between classes is to maximize the distance between any two different clusters in the latent space (Luo et al. 2020). Hence, for the class-specific prototypes $\mathcal{G}$, we reduce the inter-class dependency by maximizing the dissimilarities among prototypes:

$$\mathcal{L}_{eliminate} = \frac{2}{C(C-1)}\sum_{c<j}\frac{\langle \mathcal{G}_c, \mathcal{G}_j \rangle}{\|\mathcal{G}_c\|_2 \cdot \|\mathcal{G}_j\|_2}. \tag{6}$$

By combining both intra-class and inter-class losses, our intra-student class-wise consistency $\mathcal{L}_{inra}$ can enforce network to separate features belonging to different classes from each other, which is applied on both convolution and Transformer branches to improve their generalization capacity:

$$\mathcal{L}_{inra} = \frac{1}{2}\sum_{i \in \{CNN, Trans\}}\left(\mathcal{L}_{aggregate}^i + \mathcal{L}_{eliminate}^i\right). \tag{7}$$

3.3 Inter-student Knowledge Transfer

In this section, we focus on the distinctive inner features embedded by the convolutional and Transformer student branches, which follow a heterogeneous class-wise distribution. An inter-student knowledge transfer loss is proposed to investigate the complementary class-level statistics in the manifold space. Inspired by the Symmetric Positive Definite (SPD) metric in the manifold space, we utilize it to measure the class-wise discrepancy of two student branches. Based on the SPD representation setting (Konstantinidis et al. 2022), it is composed of square matrices $\mathbf{M}$ of size $d \times d$, which can be denoted as:

$$S_{++}^d = \left\{\mathbf{M} \in \mathbb{R}^{d \times d} : \mathbf{u}^T\mathbf{M}\mathbf{u} > 0 \; \forall \; \mathbf{u} \in \mathbb{R}^d - \{0_d\}\right\}. \tag{8}$$

The SPD matrices forming a manifold have a necessary-and-sufficient condition that if the matrix is regarded as the point, it should be symmetrical and have positive eigenvalues. Intuitively, covariance matrices are ubiquitous in any statistical related field, which are SPD with capable to gain structure properties in data. Recently, covariance matrices have attracted attentions for computer vision and machine learning tasks to support conditional independences and model image textures. Particularly, in the context of several Transformer-based deep networks (Konstantinidis et al. 2022), the introduction of covariance matrices in the processing of multi-head attention exhibits superior performance, leading to the enhancement of discrimination for their learned feature representations. To measure the distance between the points in an SPD manifold, following (Konstantinidis et al. 2022), we leverage the Frobenius norm

Method	ACDC (3%)		ACDC (10%)		ACDC (15%)		ISIC (3%)		ISIC (10%)		ISIC (15%)	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
MT	0.566	34.5	0.810	14.4	0.831	6.2	0.728	37.4	0.734	34.0	0.759	32.3
UA-MT	0.610	25.8	0.815	14.4	0.845	8.4	0.730	38.6	0.734	33.2	0.752	27.2
EM	0.602	24.1	0.791	14.5	0.838	12.0	0.723	36.3	0.727	39.3	0.766	25.8
DCT	0.582	26.4	0.804	13.8	0.854	7.9	0.729	40.6	0.760	35.7	0.777	31.4
CCT	0.586	27.9	0.816	13.1	0.837	7.8	0.677	42.2	0.723	31.7	0.765	27.9
CPS	0.603	25.5	0.833	11.0	0.850	8.0	0.686	44.4	0.743	35.7	0.771	28.4
ICT	0.581	22.8	0.811	11.4	0.854	6.5	0.732	37.2	0.753	34.6	0.785	28.9
DAN	0.528	32.6	0.795	14.6	0.841	8.4	0.695	39.5	0.724	30.4	0.755	26.5
URPC	0.567	31.4	0.829	10.6	0.841	4.8	0.703	39.3	0.758	32.8	0.752	28.6
CTCT	0.704	12.4	0.864	8.6	0.875	4.3	0.713	43.2	0.760	37.3	0.778	27.3
SSNet	0.705	17.4	0.853	10.6	0.881	4.3	0.728	40.8	0.758	32.8	0.789	26.2
ICT-Med	0.563	22.6	0.837	13.1	0.849	8.2	0.714	39.2	0.749	33.1	0.774	33.2
Ours	**0.753**	**10.7**	**0.884**	**4.4**	**0.899**	**3.5**	**0.779**	**32.1**	**0.811**	**24.4**	**0.829**	**19.8**

Table 1: Comparison with SOTA methods on ACDC and ISIC datasets under different ratios of labeled data.

instead of log-based measurement, considering that Frobenius norm is not restricted by the values of the elements in covariance matrices. Given the two class-wise manifolds $\widetilde{\mathcal{M}}_{CNN} = [\mathcal{G}^0_{CNN}, \mathcal{G}^1_{CNN}, \ldots, \mathcal{G}^{C-1}_{CNN}]$ and $\widetilde{\mathcal{M}}_{Trans} = [\mathcal{G}^0_{Trans}, \mathcal{G}^1_{Trans}, \ldots, \mathcal{G}^{C-1}_{Trans}]$, the covariance matrices of these manifolds are initially computed as:

$$cov(\widetilde{\mathcal{M}}) = \mathbf{E}[(\widetilde{\mathcal{M}} - \mathbf{E}[\widetilde{\mathcal{M}}])(\widetilde{\mathcal{M}} - \mathbf{E}[\widetilde{\mathcal{M}}])^T] \quad (9)$$

The covariance matrices $cov(\widetilde{\mathcal{M}}_{CNN})$, $cov(\widetilde{\mathcal{M}}_{Trans})$ can be regarded as points in SPD manifold. Then, we can calculate the inter-student knowledge transfer loss with Frobenius distance between these matrices:

$$\mathcal{L}_{inter} = \frac{1}{\sqrt{d}} \left\| cov(\widetilde{\mathcal{M}}_{CNN}) - cov(\widetilde{\mathcal{M}}_{Trans}) \right\|^2_F. \quad (10)$$

3.4 Optimization Objective

The overview of our M-CnT is presented in Fig. 2, which is optimized using the following loss:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_p \mathcal{L}_{pixel} + \lambda_m(\mathcal{L}_{intra} + \mathcal{L}_{inter}), \quad (11)$$

where $\mathcal{L}_{sup}$ is supervised loss on labeled data (following (Chen et al. 2021a), we use the combination of cross entropy loss and Dice loss as the supervision, which is applied to both convolutional and Transformer students); $\mathcal{L}_{pixel}$ refers to the pixel-level consistency loss in Eq. (3); $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ are two manifold constrained losses for intra-student class-wise consistency (Eq. (5) and Eq. (6)) and inter-student knowledge transfer (Eq. (10)); $\mathcal{L}_{pixel}$, $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ are applied to unlabeled data; λ_p and λ_m are loss weights to balance the relationship between losses (specifically, we choose $\lambda_p = 1$ and λ_m as Gaussian warming up function $\lambda(t) = 0.1 \times e^{(-5(1-t/t_{max})^2)}$, where t was the current training step and t_{max} was the maximum training step).

4 Experiments and Results

4.1 Datasets

We evaluate the effectiveness of our M-CnT on three public datasets, *i.e.*, Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al. 2018), International Skin Imaging Collaboration (ISIC) (Codella et al. 2018), and Synapse (a multi-organ dataset) (Landman et al. 2015). **(1)** ACDC dataset contains 100 magnetic resonance imaging (MRI) scans of three organs. Following (Chen et al. 2021a), we adopt 70, 10 and 20 cases for training, validation and testing. Consistent to the semi-supervised setting defined by (Luo et al. 2021a), we evaluate with 3%, 10%, and additional 15% partitions of labeled data. **(2)** ISIC dataset is a skin lesion segmentation dataset including 2,594 dermoscopy images, with 1,838 training images and 756 validation images. Under a semi-supervised setting, 3%, 10%, and 15% partitions of training data are provided with ground truth, while the rest training images are unlabeled. **(3)** Synapse dataset consists of 30 computed tomography (CT) scans annotated with eight abdominal organs. We adopt 18 and 12 cases for training and testing (Chen et al. 2021a). There are three partitions of training data, *i.e.,* 15%, 30% and 50%, are labeled data.

4.2 Implementation Details

All models are trained with the Stochastic Gradient Descent (SGD) optimizer, where the initial learning rate is 0.01, momentum is 0.9 and weight decay is 10^{-4}. The network converges after 30,000 iterations of training. An exception is made for the first 1,000 iterations, where λ_m is set to 0, which prevents the model collapse caused by the initialized prototypes. The batch size is 16, consisting of eight labeled images and eight unlabeled images. We randomly crop a patch with size of 224×224 as the input. We perform the standard data augmentation to avoid overfitting, including randomly flipping and rotating. The size of memory bank $\|\mathcal{B}_c\|$ for each class is set to 32. For inference, we average the predictions from two students as final results.

Evaluation Metrics. Following existing work (Chen et al. 2021a), the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) are utilized for quantitative comparisons. All ablation studies (in Sec. 4.4) are conducted on ACDC and ISIC datasets with 3% labeled data.

4.3 Comparison with the State-of-the-Arts

We compare our M-CnT with 12 semi-supervised methods, including: MT, UA-MT, EM (Vu, Jain, and Bucher 2019),

Method	15%		30%		50%	
	DSC	HD	DSC	HD	DSC	HD
MT	0.497	69.4	0.611	63.8	0.703	56.4
UA-MT	0.513	93.4	0.578	63.9	0.713	56.5
EM	0.495	72.7	0.597	63.8	0.706	61.9
DCT	0.510	77.0	0.606	64.2	0.708	54.4
CCT	0.402	75.9	0.576	69.9	0.687	71.5
CPS	0.479	66.2	0.607	69.0	0.704	50.8
ICT	0.527	70.5	0.627	59.6	0.719	39.9
DAN	0.470	93.3	0.583	73.3	0.675	72.1
URPC	0.489	69.6	0.597	66.0	0.722	42.4
CTCT	0.604	45.4	0.687	44.3	0.743	43.9
SSNet	0.581	47.3	0.668	34.9	0.750	31.8
ICT-Med	0.515	62.0	0.612	59.1	0.705	54.0
Ours	**0.653**	**32.6**	**0.714**	**31.2**	**0.772**	**24.6**

Table 2: Comparison with SOTAs methods on Synapse.

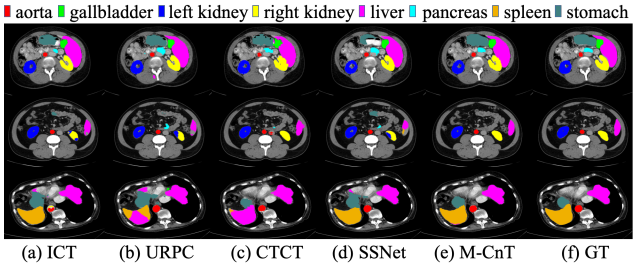

Figure 3: Exemplar segmentation results on Synapse.

DCT, CCT (Ouali, Hudelot, and Tami 2020b), CPS (Chen et al. 2021b), ICT (Verma et al. 2019), DAN, URPC (Luo et al. 2021b), CTCT (Luo et al. 2021a), SSNet (Wu et al. 2022), and ICT-Med (Basak et al. 2022).

ACDC Dataset. As shown in Table 1, M-CnT surpasses the existing methods under all settings and achieves a new SOTA. It also achieves the remarkable improvements, compared to the second best results on three partitions (DSC: +4.8%, +2.0%, +1.8%; HD: -1.7mm, -4.2mm, and -0.8mm). In particular, compared with CTCT, which uses the same CNN-Transformer architecture, our M-CnT outperforms CTCT by a notable margin, especially under the challenging setting of 3% labeled data (DSC: +4.9%, HD: -1.7mm).

ISIC Dataset. Table 1 also reports the comparison results on ISIC dataset. It can be observed that our method achieves the best segmentation performance on all settings, with the improvements of DSC: +4.7%, +5.1%, +4.0% and HD: -4.2mm, -6.0mm, and -6.0mm over the runner-up.

Synapse Dataset. Table 2 lists the comparison results on a more challenging Synapse dataset with nine categories. It is worthwhile to mention that Synapse has the limited training data (18 cases) of complex anatomical contrasts, anfractuous boundaries and heterogeneous textures. The improvements of M-CnT over the SOTA are +4.9%, +2.7%, +2.2% of DSC and -12.8mm, -3.7mm, and -7.2mm of HD on different settings, which substantiate the fine robustness of M-CnT.

Visual Comparison. Fig. 3 illustrates some segmentation results of top-five methods on Synapse with 50% labeled data, where M-CnT is able to segment both tiny objects with complex boundaries (*e.g.,* gallbladder) and large objects with fine structures (*e.g.,* stomach).

#	$\mathcal{L}_{sup}$	$\mathcal{L}_{pixel}$	$\mathcal{L}_{intra}$		$\mathcal{L}_{inter}$	ACDC	ISIC
			$\mathcal{L}_a$	$\mathcal{L}_e$			
1	✓					0.510	0.677
2	✓	✓				0.715	0.734
3	✓	✓	✓			0.726	0.749
4	✓	✓		✓		0.738	0.759
5	✓	✓	✓	✓		0.743	0.768
6	✓	✓			✓	0.748	0.755
7	✓	✓	✓	✓	✓	**0.753**	**0.779**

Table 3: Ablation study of different losses included in Eq. 11 on both ACDC and ISIC datasets with 3% labeled. $\mathcal{L}_a$ and $\mathcal{L}_e$ are two abbreviations of $\mathcal{L}_{aggregate}$ and $\mathcal{L}_{eliminate}$.

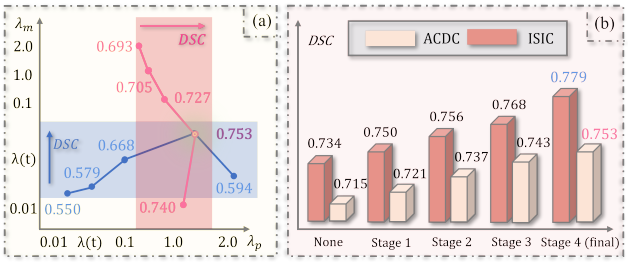

Figure 4: (a) Performance of M-CvT w.r.t. λ_p and λ_m. (b) Impact of the insertion location of manifold constraints at different stages of U-Net and Swin-UNet decoders.

4.4 Hyper-Parameters

Impact of λ_p and λ_m. As mentioned in Eq. 11, λ_p and λ_m are two coefficients that control the overall optimization objective. Hence, we evaluate the impact of these two parameters and the Gaussian warming up function $\lambda(t)$ on 3% labeled ACDC. As shown in Fig. 4 (a), we first evaluate the effect of λ_m with λ_p=1.0 (in pink). When the value is low (*i.e.*, λ_m=0.01), our manifold constraints start to make positive impacts on the optimization. However, high values are detrimental. The underlying reason may be that the overall losses are overwhelmed by the unlabeled data with high uncertainty, which hinders the training process. The best performance is achieved when choosing $\lambda(t)$ as λ_m, which can adaptively increase the weight according to the training iterations. Further, we freeze λ_m=$\lambda(t)$ and utilize different value of λ_p (in blue). Overall, λ_p=1 achieves the best results, which is chosen in the following experiments.

Effectiveness of Manifold Constraints on Each Stage. We explore the influence of integrating manifold constraints ($\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$) at different stages with 3% labeled data on ACDC and ISIC datasets. As seen in Fig. 4 (b), our manifold constraints can consistently increase the accuracy on various stages and datasets, and the best results (ACDC: +3.8%, ISIC: +4.5%) are achieved when the constraints are applied to the final stage.

Methods	ACDC	ISIC
baseline	0.715	0.734
+ ED	0.720	0.742
+ IFVD	0.727	0.754
+ PLCL	0.732	0.739
+ ours	**0.753**	**0.779**

Table 4: Comparison of different class-wise losses.

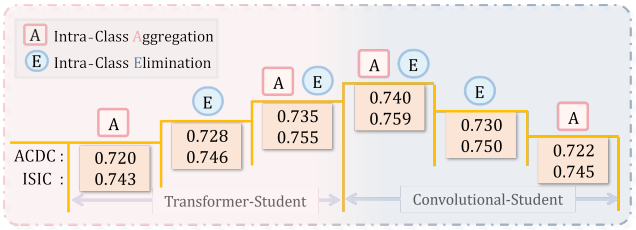

Figure 5: Integration of intra-class and inter-class losses into each student on ACDC and ISIC with 3% labeled data.

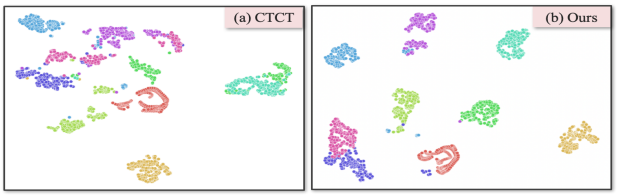

Figure 6: t-SNE visualization of deep feature representations extracted from (a) CTCT and (b) our M-CnT, which are trained with 30% labeled data on Synapse.

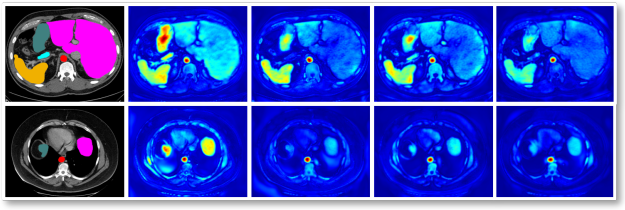

Figure 7: Feature maps of four methods on Synapse.

4.5 Ablation Study

Loss Impact. As shown in Table 3, a significant improvement (ACDC: +20.5%, ISIC: +5.7%) is achieved when utilizing the unlabeled data with our pixel-level consistency $\mathcal{L}_{pixel}$ (#2). Compared with CTCT learning from cross-teaching, our $\mathcal{L}_{pixel}$ considers the confidence of CNN and Transformer, which results in an improvement of (ACDC: +1.1%, ISIC: +2.1%). Regarding to the intra-student consistency $\mathcal{L}_{intra}$, our intra-class loss $\mathcal{L}_{aggregate}$ (*i.e.,* $\mathcal{L}_a$) encourages the network to compact the class cluster of each class in the latent space (#3), yielding improvements of (ACDC: +1.1%, ISIC: +1.5%); while our inter-class loss $\mathcal{L}_{eliminate}$ (*i.e.,* $\mathcal{L}_e$) maximizes the distance among classes (#4), achieving improvements of (ACDC: +2.3%, ISIC: +2.5%). Benefiting from both $\mathcal{L}_a$ and $\mathcal{L}_e$, our $\mathcal{L}_{intra}$ (#5) can generate a total improvement of (ACDC: +2.8%, ISIC: +3.4%). Moreover, $\mathcal{L}_{inter}$ (#6) can align two-style class distributions and obtain the better performance. Overall, M-CnT (#7) achieves the best performance of (ACDC: 75.3%, ISIC: 77.9%) with distinctive features.

Effectiveness of Intra- and inter- class Losses on Each Student. As shown in Fig. 5, both intra-class ($\mathcal{L}_{aggregate}$) and inter-class losse ($\mathcal{L}_{eliminate}$) improve the performance of each student, and the best result can be achieved via combining the two losses with compactness and discrepancy.

Comparison of Class-wise Losses. To verify the capability of manifold constraints for class reasoning, we replace it with several class-wise losses introduced in Sec. 2, including ED (Li et al. 2022), IFVD (Wang et al. 2020) and PLCL (Alonso et al. 2021) . As seen in Table 4, our method achieves a higher accuracy by learning class-wise distributions in both intra- and inter-student manners.

4.6 Interpretation of M-CnT

Distribution of Deeply Learned Features. Compared with the runner-up CTCT, our M-CnT is capable to generatie more compact and well separated feature embeddings, as shown in Fig. 6. It reveals that our manifold constraints can improve the discriminativity of learned features, which is crucial for semi-supervised segmentation.

Visualization of Feature Maps. We further visualize the feature maps of top four methods trained with 50% labeled data in Fig. 7. Our M-CnT inherits the advantages of both retaining local features and capturing global dependency, *e.g.,* full coverage of large-sized liver (in the first row) and small-sized stomach (in the second row).

5 Conclusion

We proposed a novel manifold constrained combinatorial CNN-Transformer learning algorithm, namely M-CnT, for semi-supervised medical image segmentation. M-CnT follows a typical dual-student scheme with CNN-Transformer architecture, where the complementary class-wise characteristics were explored by the presented manifold conditions, *i.e.*, intra-student class-wise consistency and inter-student knowledge transfer losses. The former one helps to generate compact and discriminative representations in each student branch, while the latter one transfers class-wise knowledge cross students. Our method has been evaluated on three public datasets and outperformed other SSL approaches.

Acknowledgments

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China (No. LZ22F020012), Major Technological Innovation Project of Hangzhou (No. 2022AIZD0147), the National Key Research and Development Project (No. 2022YFC2504605), Major Scientific Research Project of Zhejiang Lab (No. 2020ND8AD01), and Japanese Ministry for Education, Science, Culture and Sports (No. 20KK0234, No. 21H03470 and No. 20K21821).

References

Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; and Murillo, A. C. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8219–8228.

A.Tarvainen; and H.Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Bai, W.; Oktay, O.; Sinclair, M.; Suzuki, H.; Rajchl, M.; Tarroni, G.; Glocker, B.; King, A.; Matthews, P. M.; and Rueckert, D. 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 253–260. Springer.

Basak, H.; Bhattacharya, R.; Hussain, R.; and Chatterjee, A. 2022. An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation. *arXiv preprint arXiv:2202.00677*.

Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.; Cetin, I.; Lekadir, K.; Camara, O.; and Ballester, M. 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11): 2514–2525.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2021. Swin-UNet: UNet-like pure Transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.

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

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021b. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Codella, N.; Gutman, D.; Celebi, M.; Helba, B.; Marchetti, M.; Dusza, S.; Kalloo, A.; Liopyris, K.; Mishra, N.; and Kittler, H. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *IEEE International Symposium on Biomedical Imaging*, 168–172. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; and Gelly, S. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; and Finlayson, G. 2019. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*.

Huang, Z.; Wang, R.; Li, X.; Liu, W.; Shan, S.; Van Gool, L.; and Chen, X. 2017. Geometry-aware similarity learning on SPD manifolds for visual recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10): 2513–2523.

Jia, D.; Han, K.; Wang, Y.; Tang, Y.; Guo, J.; Zhang, C.; and Tao, D. 2021. Efficient vision Transformers via fine-grained manifold distillation. *arXiv preprint arXiv:2107.01378*.

Ke, Z.; Wang, D.; Yan, Q.; Ren, J.; and Lau, R. W. 2019. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6728–6736.

Konstantinidis, D.; Papastratis, I.; Dimitropoulos, K.; and Daras, P. 2022. Multi-manifold attention for vision Transformers. *arXiv preprint arXiv:2207.08569*.

Landman, B.; Xu, Z.; Igelsias, J. E.; Styner, M.; Langerak, T.; and Klein, A. 2015. Segmentation outside the cranial vault challenge. *Synapse*.

Li, J.; Yu, H.; Chen, C.; Ding, M.; and Zha, S. 2022. Category guided attention network for brain tumor segmentation in MRI. *Physics in Medicine & Biology*, 67(8): 085014.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; and Heng, P.-A. 2018. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*.

Luo, X.; Hu, M.; T.Song; Wang, G.; and Zhang, S. 2021a. Semi-supervised medical image segmentation via cross teaching between CNN and Transformer. *arXiv preprint arXiv:2112.04894*.

Luo, X.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Chen, N.; Wang, G.; and Zhang, S. 2021b. Efficient Semi-supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 318–329.

Luo, Y.-W.; Ren, C.-X.; Ge, P.; Huang, K.-K.; and Yu, Y.-F. 2020. Unsupervised domain adaptation via discriminative manifold embedding and alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5029–5036.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020a. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020b. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; and Yuille, A. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision*, 135–152.

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

Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Bengio, Y.; and Lopez-Paz, D. 2019. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*.

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

Wang, Y.; Zhou, W.; Jiang, T.; Bai, X.; and Xu, Y. 2020. Intra-class feature variation distillation for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 346–362. Springer.

Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; and Cai, J. 2022. Exploring smoothness and class-separation for semi-supervised medical image segmentation. *arXiv preprint arXiv:2203.01324*.

Yu, L.; Wang, S.; Li, S.; Fu, C.; and Heng, P. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 605–613. Springer.

Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D.; and Chen, D. 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 408–416.

Zheng, H.; Lin, L.; Hu, H.; Zhang, Q.; Chen, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Tong, R.; and Wu, J. 2019. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 148–156. Springer.

Zheng, X.; Luo, Y.; Wang, H.; Fu, C.; and Wang, L. 2022. Transformer-CNN Cohort: Semi-supervised Semantic Segmentation by the Best of Both Students. *arXiv preprint arXiv:2209.02178*.

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

Zhou, Z.; Siddiquee, M.; Tajbakhsh, N.; and Liang, J. 2019b. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6): 1856–1867.