Medical Image Analysis 67 (2021) 101876

Contents lists available at ScienceDirect

Medical Image Analysis

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

Pairwise learning for medical image segmentation

Renzhen Wang[a], Shilei Cao[b], Kai Ma[b], Yefeng Zheng[b], Deyu Meng[a,c,*]

[a] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China
[b] Jarvis Lab, Tencent, Shenzhen, 518075, China
[c] Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau

ARTICLE INFO

Article history:
Received 1 March 2020
Revised 29 September 2020
Accepted 5 October 2020
Available online 17 October 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Medical image segmentation
Conjugate fully convolutional network
Pairwise segmentation
Proxy supervision

ABSTRACT

Fully convolutional networks (FCNs) trained with abundant labeled data have been proven to be a powerful and efficient solution for medical image segmentation. However, FCNs often fail to achieve satisfactory results due to the lack of labelled data and significant variability of appearance in medical imaging. To address this challenging issue, this paper proposes a conjugate fully convolutional network (CFCN) where pairwise samples are input for capturing a rich context representation and guide each other with a fusion module. To avoid the overfitting problem introduced by intra-class heterogeneity and boundary ambiguity with a small number of training samples, we propose to explicitly exploit the prior information from the label space, termed as proxy supervision. We further extend the CFCN to a compact conjugate fully convolutional network (C^2FCN), which just has one head for fitting the proxy supervision without incurring two additional branches of decoders fitting ground truth of the input pairs compared to CFCN. In the test phase, the segmentation probability is inferred by the learned logical relation implied in the proxy supervision. Quantitative evaluation on the Liver Tumor Segmentation (LiTS) and Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) datasets shows that the proposed framework achieves a significant performance improvement on both binary segmentation and multi-category segmentation, especially with a limited amount of training data. The source code is available at https://github.com/renzhenwang/pairwise_segmentation.

1. Introduction

Semantic segmentation is a classic computer vision task which aims at predicting semantic labels for each pixel of an image so as to partition the image into meaningful objects. In medical image analysis, semantic segmentation plays an important role in quantitative measurement of volume and shape, and preliminary preprocess of computer-aided detection pipelines (Litjens et al., 2017). Benefitting from the recent advancement of fully convolutional networks (FCNs) (Long et al., 2015), deep learning based medical image segmentation approaches (Ronneberger et al., 2015; Milletari et al., 2016) have attracted vast attention and achieved great success in many scenarios (Hesamian et al., 2019).

Despite achieving great success, current medical image segmentation still faces some challenges deserving broad attention. First, the success of deep segmentation models is mostly attributed to a large number of training data. Gathering pixel-level annotations of medical images, however, is very difficult because the manual delineation process is expertise-required and time-consuming. To address this issue, several remarkable techniques have been adopted, including data augmentation (Ronneberger et al., 2015; Pereira et al., 2016; Christ et al., 2016; Dong et al., 2017) and pre-training models (Tajbakhsh et al., 2016; Wu et al., 2017; Zhou et al., 2019; Chen et al., 2019). Data augmentation methods directly enlarge the amount of training data by adopting a set of affine/elastic transformations (Pereira et al., 2016; Ronneberger et al., 2015) and appearance adjustment (Christ et al., 2016; Dong et al., 2017). However, augmented samples have a strong correlation with the original ones and different augmentation methods usually yield unstable results in different segmentation scenarios. Pre-training tricks, subordinated to transfer learning, usually fine-tune the network trained on general images (Tajbakhsh et al., 2016; Wu et al., 2017) or medical images (Zhou et al., 2019; Chen et al., 2019). Although it has been proven that the segmentation performance can be significantly improved compared with random initialization, pre-training implies that the architecture of model has been completely or partially determined, which may take adverse effects when the source and target images suffer from a large domain shift.

Second, and more typically, the intra-class heterogeneity and boundary ambiguity of the target object are still big challenges in

* Corresponding author.
E-mail address: dymeng@mail.xjtu.edu.cn (D. Meng).

https://doi.org/10.1016/j.media.2020.101876
1361-8415/

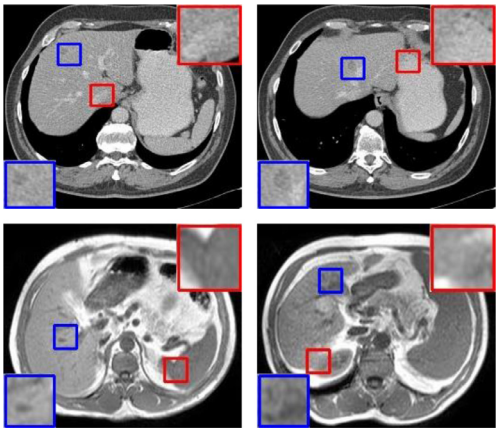

Fig. 1. Examples show the intra-class heterogeneity and boundary ambiguity in medical imaging, visualized with two abdomen CT images (first row) and two abdomen MR images (second row). Blue and red rectangles denote heterogeneous appearance and ambiguous boundary, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

medical image segmentation (Wang et al., 2019; Hesamian et al., 2019), as shown in Fig. 1. Actually, the anatomical structures or lesions are generally very heterogeneous in size, shape, and location from patient to patient. Even in one single target, local contexts are usually very different. Besides, the ambiguous boundary is a known inherent imaging challenge where in-between target organs and the neighboring tissues have a low contrast (Hesamian et al., 2019; Xie et al., 2017), which usually occurs during imaging, such as attenuation coefficient in Computed Tomography (CT) and relaxation time in Magnetic Resonance Imaging (MRI) (Dou et al., 2017; Kronman and Joskowicz, 2016). Increasing the quantity of training data may be the most direct way to alleviate this problem; however, acquiring large amount of manually annotated training data is not realistic. Data augmentation and pre-training tricks are also not ideal for solving this problem, since the former may magnify the intra-class inconsistency when enlarging the training pool and the latter only focuses on extracting general instead of specific features for the target task. Exploiting hard-to-recognize pixels, like weighted loss function (Ronneberger et al., 2015) and cascade network (Wu et al., 2017), seems to be a feasible strategy to address this challenge; however, its performance tends to be degenerated when there is a limited amount of training data.

To address the two aforementioned issues, a feasible solution is to combine data augmentation and weighted loss function, just like U-Net (Ronneberger et al., 2015) does. Aside from this ensemble strategy, a natural question is whether we can achieve it through a unified network architecture. We suspect that embedding the prior knowledge into deep models may be a feasible strategy, although current methods are mainly focusing on modeling manifolds of target objects (Chen et al., 2016; Ravishankar et al., 2017b; Araújo et al., 2019; Ravishankar et al., 2017a; Mosinska et al., 2018). Actually, all the objects in medical images/volumes, not only the ones to be segmented, usually lie in a low-dimensional manifold and modeling the intrinsic relations of them is of great significance for segmentation. For example, in liver segmentation, the relative positions of the surrounding organs are very important for locating the liver and helpful to the liver segmentation. Based on this prior information, we aim to address both the aforementioned issues with a unique pairwise learning framework.

The proposed pairwise segmentation framework is based on the paradigm that a pair of samples is taken as input and jointly segmented in the network, and one more additional output is involved to explicitly learn the prior information from the label space (dubbed proxy supervision). Concretely, we adopt a Siamese architecture, referenced as conjugate fully convolutional network (CFCN), which includes two identical parallel branches with each taking one sample and outputting the corresponding mask probability map. Benefitting from training on pairs, the training samples are quadratically augmented so as to alleviate the overfitting issue of the network compared to the traditional FCNs. To address intra-class heterogeneity and boundary ambiguity, CFCN introduces one more sub-network to fit the proxy supervision that is derived from a function of the ground truth of input pairs. Considering that medical image segmentation is a location-aware task where the relative position of the anatomical structure is very important for locating the target object, the proxy supervision utilizes logical operations, including *logic AND* and *logic XOR*, to establish the correlation of the two input samples at the same location. In this way, the *logic AND* can improve predictive confidence to eliminate intra-class inconsistency through comparison learning, and *logic XOR* can improve the exposure of pixels lying in the target boundary to encode the shape prior. Interestingly, with the formulated logical relation, we can remove both decoders in the Siamese structure to avoid directly fitting the ground truth segmentation, but just preserve the sub-network for fitting the proposed proxy supervision only (dubbed compact conjugate fully convolutional network, and abbreviated as C^2FCN). We can utilize the learned logical relation implied in proxy supervision to infer the segmentation probability of the target objects in the test phase. Under the premise of ensuring performance, our C^2FCN can use any off-the-shelf segmentation network to implement pairwise segmentation with a negligible number of additional parameters.

The main contributions of this paper are mainly three-fold:

- we propose a new pairwise segmentation framework to address medical image segmentation with limited training data, intra-class heterogeneity and boundary ambiguity in medical imaging using a unified network architecture.
- we propose a proxy supervision which explicitly encodes the prior information from the label space and acts as global constraint on the network in the training phase.
- we propose a new segmentation paradigm through C^2FCN with more concise architecture beyond CFCN, which learns the logical relation of the input pair in the training phase only, and infers the segmentation probability from the learned logical relation in the test phase.

This paper is extended from our preliminary work in (Wang et al., 2019), and the main extension includes:

- We extend the proposed CFCN to a general pairwise segmentation framework, for which we present concrete mathematical formulation. Moreover, we extend CFCN from binary segmentation to multi-category segmentation, which is demonstrated with additional multi-organ segmentation.
- We further propose a compact architecture C^2FCN, which is a slim version of CFCN as the number of parameters and the computational overhead are largely reduced during training, yet the segmentation performance is comparative or even better especially in alleviating overfitting issues against CFCN on all our segmentation experiments.

The rest of the paper is organized as follows. Section 2 describes related works. Section 3 presents the proposed pairwise segmentation framework as well as implementation details. Section 4 demonstrates experimental results and the final discussion and conclusion is summarized in Section 5.

2. Related work

2.1. Deep pairwise learning

Deep pairwise learning, also known as Siamese network, was firstly introduced by Bromley et al. (1994) in the signature verification application. Subsequently, pairwise neural network models were extensively applied in computer vision, including face verification (Chopra et al., 2005; Taigman et al., 2014), image matching (Zagoruyko and Komodakis, 2015; Singh and Lee, 2016), object tracking (Tao et al., 2016; Bertinetto et al., 2016), fine-grained classification (Zhang et al., 2018; Dubey et al., 2018), and visual co-segmentation (Li et al., 2018a; Chen et al., 2018a; Banerjee et al., 2019; Lu et al., 2019). Among these works, co-segmentation are closely related to our approach as they both take a pair of samples as input and predict pixel-level masks for segmenting meaningful objects. However, there are at least two major differences between our method and co-segmentation. First, our method falls into semantic segmentation where the task consists of the training and test phases, which requires annotated training data to learn the model in training, and segment the target in test images during the test phase. In other words, the target object could be segmented in test if and only if such category samples are annotated in training samples and the semantic concepts they stand for have been learned by the model in training. Comparatively, co-segmentation aims to discover objects commonly appearing in multiple images, where the input of the model is more than one image and the co-occurring objects are annotated in training, which enables the model to learn whether the input images have co-occurring objects and segment them. Second, methods of semantic segmentation cope with multiple known categories in training and test phases. In contrast, co-segmentation usually works on multiple images to predict whether the segmented pixels belong to a single yet unknown category.

Regardless of the application scenarios, the aforementioned works mainly fall into two categories, namely metric-learning methods (Bromley et al., 1994; Chopra et al., 2005; Taigman et al., 2014; Zagoruyko and Komodakis, 2015; Singh and Lee, 2016; Tao et al., 2016; Bertinetto et al., 2016; Zhang et al., 2018; Dubey et al., 2018; Banerjee et al., 2019) and feature interaction methods (Li et al., 2018a; Chen et al., 2018a; Lu et al., 2019). The former methods adopt such a strategy where the two streams of Siamese networks, respectively, extract the features of input pairs, and then use the features to compute a similarity metric or to learn a similarity metric with an additional network. For example, in (Bromley et al., 1994), two sub-networks were adopted to extract features from two signatures, and the two branches of features are then used to learn a metric to predict whether the input signatures are from the same class. The latter methods directly adopt a Siamese structure to capture the relationship between two feature streams, which utilizes the information shared across them and enables them to mutually learn from each other. For example, Li et al. (2018a) employed a mutual correlation layer to compute localized correlations for highlighting the common objects. Chen et al. (2018a) adopted a similar way where an attention module was used to select the semantically related features for image co-segmentation. Lu et al. (2019) proposed a global co-attention mechanism to model inherent correlation among video frames for video object segmentation. Although our method introduces an additional sub-networks (dubbed fusion net) except for two Siamese sub-networks, like metric learning methods (Zhang et al., 2018; Banerjee et al., 2019), our method mainly focuses on exploiting prior knowledge from the label space through a global operator, and the fusion net plays the role for bridging the feature space and the label space by fitting the global operator.

2.2. Deep prior knowledge modeling

There are some remarkable works for modeling the manifold or prior knowledge underlying the target objects. One group leverage the intrinsic relations among the same category of pixels to improve the performance of FCNs. For example, a dense conditional random field (CRF) was attached to the FCN as a postprocessing step (Chen et al., 2017) or jointly trained with the FCN (Zheng et al., 2015) to preserve the boundary of the target objects. Similarly, Liu et al. (2017) proposed a spatial propagation network to learn an affinity matrix for modeling dense, global pairwise relationships of an image, and Ke et al. (2018) proposed an adaptive affinity field to encode spatial structural information and geometric regularities through the label relations in the training process.

Another group of methods improve the segmentation performance of FCNs by explicitly or implicitly modeling high-order prior knowledge in-between different objects in medical images/volumes, such as shape and topological structures. Typically, Chen et al. (2016) took gland objects and contours as auxiliary information under a multi-task learning framework to boost the gland segmentation from histology images. To model the shape manifold space and correct topological incoherency of segmentation networks, a non-linear shape model pre-learned by convolutional autoencoder (CAE) (Ravishankar et al., 2017b) and a topology coherence model learned by variational autoencoder (Araújo et al., 2019) were respectively incorporated in an FCN. Ravishankar et al. (2017a) proposed a novel framework based on deep learning to jointly learn the foreground, background and shape to improve segmentation accuracy. BenTaieb and Hamarneh (2016) proposed a topology-aware loss to train the FCN for coding geometric and topological priors of containment and detachment on histology gland segmentation. A similar idea (Mosinska et al., 2018) was used to capture higher-order topological features of linear structures, where the topology-aware loss was constructed by the response of selected filters from a pre-trained VGG19 network (Simonyan and Zisserman, 2014).

Different from the two groups of methods, we encode the prior knowledge through a global function of ground truth of input pairs, and focus on both local context and global shape priors.

3. Methodology

Given a training dataset $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{N}$ of Nsamples, $\mathbf{x}_i \in \mathbb{R}^{H\times W}$is the i-th sample with height $H$and width W,and $\mathbf{y}_i \in \{0,1\}^{H\times W\times K}$is the associated label over Kclasses. Different from the traditional semantic segmentation aiming at training a model that predicts the target segmentation with a single input, the proposed pairwise segmentation framework takes sample pairs as input and synergistically segment them through the CFCN model f,which can be formally formulated as

$$f(\mathbf{x}_i,\mathbf{x}_j;\mathbf{W})=(\mathbf{y}_i,\mathbf{y}_j,\mathbf{y}_{ij}), \tag{1}$$

where $\mathbf{W}$is the parameters of f,and $\mathbf{y}_{ij}$is the proxy supervision derived from a user-designed function $g_{proxy}$to explicitly exploit the prior information from the mask $\mathbf{y}_i$and $\mathbf{y}_j$,i.e.,

$$\mathbf{y}_{ij}=g_{proxy}(\mathbf{y}_i,\mathbf{y}_j). \tag{2}$$

From the point of network optimization, proxy supervision can be regarded as a global constraint on the network to fit inherent prior knowledge explored by g_{proxy}.

3.1. Conjugate fully convolutional network

In the following sections, we present the details of the proposed pairwise segmentation network CFCN and introduce the proxy supervision $g_{proxy}$to model location correlation and shape prior for

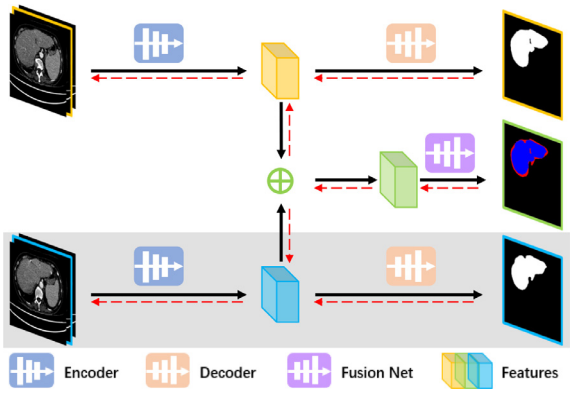

Fig. 2. The framework of the proposed conjugate fully convolutional network, where the two encoders share the same architecture and parameters, so do the two decoders. The black solid line arrows show the forward propagation, and the red dotted line arrows mark the gradient flow direction of the back propagation. Note that only one segmentation branch (the shaded part) is required for the inference phase. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

addressing the challenges arisen by intra-class heterogeneity and boundary ambiguity.

3.1.1. Network architecture

As illustrated in Fig. 2, the CFCN model consists of three parts: two conjugate sub-networks made up of an encoder and a decoder, with each for segmenting one single sample of an input pair, and a fusion net for learning the proxy supervision.

The two conjugate sub-networks employ two identical FCNs with encoder-decoder architecture, e.g., U-Net (Ronneberger et al., 2015) and DeepLabv3+ (Chen et al., 2018b). To capture the intrinsic relations of pairwise input and encode the manifold of target objects, the two sub-networks share the same weights in the encoder and decoder layers, which implies that the features captured by CFCN should be sufficient to represent the target object and robust for distinguishing background. In this paper, in order to reduce the number of parameters and improve computation efficiency, we further adopt a conjugate DeepLabv3+ with a ResNet-18 backbone, which contains four residual blocks with 18 layers in total.

The fusion net takes the element-wise sum of the features (lying at the same layers) of two conjugate sub-networks from one or more layers as input, in order to capture the location-aware representation under the proxy supervision. In this paper, the proposed fusion net exploits two streams of input, including low-level features from the first residual blocks of ResNet18 and high-level features from the decoder of two conjugate sub-networks. (For simplicity, we only show one stream of input to the fusion net in Fig. 2.) To adaptively learn global context information for fitting the proxy supervision, we specially design a fusion net. Inspired by Yu et al. (2018), we adopt a channel attention block to refine the low-level features, and further element-wisely add the resulting features to the high-level features. Then, the features are fed into a 3×3convolutional layer and bilinearly up-sampled to the same size as output corresponding to the proxy supervision maps.

3.1.2. Proxy supervision

As aforementioned, proxy supervision is a global constraint on the network, which can be employed to refine the network by designing g_{proxy}. Although the form of $g_{proxy}$is general, we hope that g_{proxy}explicitly models task-related prior knowledge. To this end, we study the challenging issues arisen by intra-class heterogeneity and boundary ambiguity with a handful of training samples on medical image segmentation.

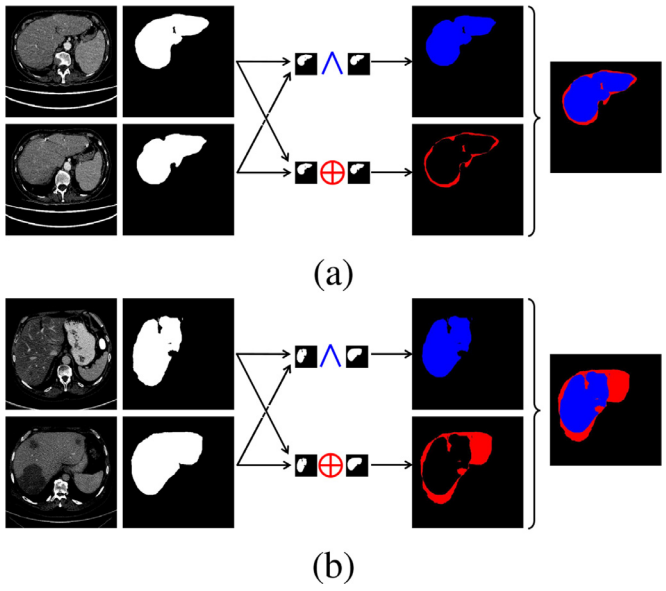

Fig. 3. The illustration of the proxy supervision implemented by logical operation between input pairs for addressing intra-class heterogeneity and boundary ambiguity. The logical operation $\wedge$and $\oplus$denote *logic AND* and *logic XOR*, which result in an intersection map marked in blue and a difference map marked in red, respectively. (a) Intra-patient pair, the proxy supervision between two slices from the same case; (b) Inter-patient pair, the proxy supervision between two slices from different cases. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As Fig. 1 shows, we revisit the aforementioned challenges in medical imaging: 1) Intra-class heterogeneity, i.e., the target objects in different images/volumes share the same semantic label but different appearances. For example, in liver segmentation, both individual anatomy variability and imaging device difference can result in intra-class heterogeneity. Especially with limited training data, the network is sensitive to intra-class inconsistency and prone to overfitting. 2) Boundary ambiguity, where in-between target organs and the neighboring tissues suffer from a low contrast, which usually occurs during imaging. Since the traditional FCN in this scenario is over-parameterized, it tends to perform well on the training dataset while poorly on the test dataset. Thus, how to trade-off between exactly modeling the manifold of target objects and robustly representing the individual difference is the key of the proxy supervision.

Inspired by the human-like comparison learning, we utilize logical operations, including *logic AND* and *logic XOR*, to establish the correlation of the two input samples at the same location for addressing the intra-class heterogeneity and boundary ambiguity. Formally, *logic AND* is to make the pixels, belonging to the target object at the same location, respond to one single mask map, i.e.,

$$\mathbf{y}_{ij}^{1} = \mathbf{y}_i \wedge \mathbf{y}_j, \tag{3}$$

where $\wedge$is the element-wise *logic AND* operation. *Logic XOR* is to make the pixels, only one belonging to the target object at the same location, respond to one single mask map, which can be achieved by element-wise *logic XOR* operation $\oplus$on groundtruth maps $\mathbf{y}_i$and $\mathbf{y}_j$.i.e.,

$$\mathbf{y}_{ij}^{2} = \mathbf{y}_i \oplus \mathbf{y}_j. \tag{4}$$

As shown in Fig. 3, the blue region denotes the proxy supervision $\mathbf{y}_{ij}^{1}$and the region masked in red denotes proxy supervision $\mathbf{y}_{ij}^{2}$. Specifically, Fig. 3(a), shows the proxy supervision between two slices from the same case (an intra-patient pair), and Fig. 3(b) shows the proxy supervision between two slices from different

cases (an inter-patient pair). In the training phase, a slice is paired with different intra/inter-patient slices, and each pixel's label is determined by itself and the other pixel from its paired slice (at the same location). In other word, its label change dynamically according to its paired pixels. During training, the network is trained under the guidance of the proxy supervision, which discriminates the following three scenarios for binary segmentation: the two pixels are from the target object simultaneously, there is only one from the target object, they are both from background. This indicates that fitting the aforementioned proxy supervision requires to discriminate the logical correlation of paired pixels through such a comparison learning, thus the predictive confidence will be improved and a misclassified pixel may be corrected by different paired pixels (a slice may be paired with different slices) to alleviate the intra-class heterogeneity.

As for boundary ambiguity, we know that accurately modeling manifolds of target objects is a potential solution to address it. Therefore, we propose to sample the intra-patient pairs with a small interval along the axial direction, where the exposure of pixels lying in the target boundary is improved and the difference of the masks helps encode the shape prior (see Fig. 3(a)). Specifically, we employ the proxy supervision $g_{proxy}$to model the shape prior in an end-to-end training manner, which ultimately assists the network to distinguish ambiguous boundary.

Taking liver segmentation as an example, we consider $\mathbf{y}_{ij}^1$and $\mathbf{y}_{ij}^2$as an integration, and then the overall proxy supervision $\mathbf{y}_{ij} = \{\mathbf{y}_{ij}^1, \mathbf{y}_{ij}^2\}$can be implemented by an additional 3-category segmentation task, where the target label map consists of three channels, including the intersection of background, the intersection of the target objects $\mathbf{y}_{ij}^1$and the difference of the target objects $\mathbf{y}_{ij}^2$,respectively. For multi-category segmentation, the extension is natural where the maps of the proxy supervision are twice the number of categories and the fitting of the proxy supervision is multi-label segmentation (the reason is that a pixel of the proxy supervision may have two labels simultaneously) instead of multi-class segmentation.

3.1.3. Training and inference

The objective function of pairwise segmentation can be formulated as

$$\min_{\mathbf{W}} \mathbf{E}(\mathbf{W}) = \sum_{i,j} \mathcal{L}\Big((\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_{ij}), f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W})\Big), \tag{5}$$

where $\mathcal{L}(\cdot,\cdot)$is the loss function to measure the error between the output of CFCN model $f$and the overall supervision constituted by $\mathbf{y}_i$,$\mathbf{y}_j$and $\mathbf{y}_{ij}$. Since the output $f(\mathbf{x}_i, \mathbf{x}_j)$consists of three branches of segmentation maps (denoted as $\mathbf{p}_i$,$\mathbf{p}_j$and $\mathbf{p}_{ij}$for convenience) corresponding to the two ground truth of paired inputs $\mathbf{y}_i$and $\mathbf{y}_j$,and the proxy supervision $\mathbf{y}_{ij}$,respectively, the CFCN can be viewed as a multi-task model. More specifically, the loss function $\mathcal{L}$can be rewritten as

$$\begin{aligned}\mathcal{L}\Big((\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_{ij}), f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W})\Big) = \mathcal{L}_1(\mathbf{y}_i, \mathbf{p}_i) + \mathcal{L}_2(\mathbf{y}_j, \mathbf{p}_j) \\ +\lambda \mathcal{L}_{proxy}(\mathbf{y}_{ij}, \mathbf{p}_{ij}),\end{aligned} \tag{6}$$

where $\mathcal{L}_1$,$\mathcal{L}_2$and $\mathcal{L}_{proxy}$refer to the pixel-level segmentation losses for measuring the error between $\mathbf{y}_k$and $\mathbf{p}_k, k = i, j, ij$,respectively, and $\lambda$is a user-preset weight used to balance the contribution of losses between input pairs and the proxy supervision. In this paper, we adopt the Dice loss for $\mathcal{L}_1$and $\mathcal{L}_2$,and the multi-class Dice Loss for $\mathcal{L}_{proxy}$,and preset $\lambda$to 2.

The proposed CFCN is trained in an end-to-end manner with all three branches updated simultaneously. As the input pairs are randomly selected during the training phase without any particular contextual order, either of the two conjugate branches should be self-sufficient for the segmentation. Therefore, only one subnetwork is required in the test phase (see Fig. 2) and the additional fusion branch is also removed.

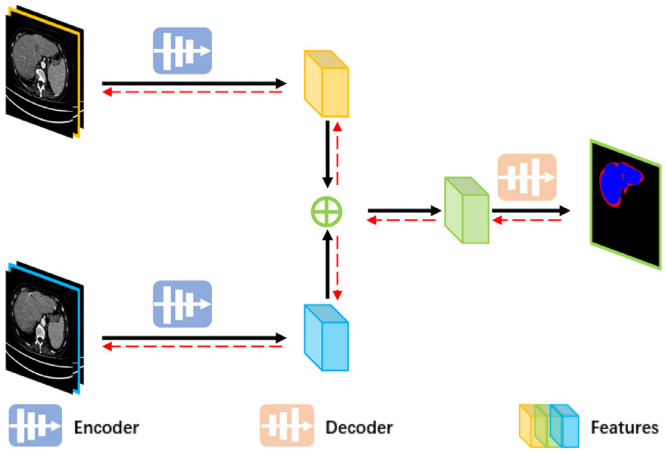

Fig. 4. The framework of the proposed compact conjugate fully convolutional network. The network has two identical encoders with shared parameters, and only one branch for fitting the proxy supervision. Note that a plain decoder, same as the baseline (e.g., DeepLabv3+ (Chen et al., 2018b) in this paper), is used in this branch for modifying the baseline as little as possible, which differs from CFCN that employs a tailored fusion net. The black solid and red dotted line arrows show the forward and backward propagation, respectively. In the test phase, the segmentation probability map is inferred according to Eq. (12). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Compact conjugate fully convolutional network

3.2.1. Architecture

Revisiting the pairwise segmentation framework and combining Eq. (1) and Eq. (2), we have

$$f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W}) = \Big(\mathbf{y}_i, \mathbf{y}_j, g_{proxy}(\mathbf{y}_i, \mathbf{y}_j)\Big), \tag{7}$$

where the image space and the label space are bridged by a pairwise segmentation network f,and the proxy supervision operator g_{proxy}acts as global constraints for exploiting the prior knowledge. Although CFCN model achieves significant performance improvement compared with the corresponding baseline, it also brings about the increase of model parameters. In fact, CFCN has large model redundancy because only part of the network parameters is used in the inference process, and the network is asymmetric, i.e., $f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W}) \neq f(\mathbf{x}_j, \mathbf{x}_i; \mathbf{W})$. A bold yet natural strategy to improve CFCN is to use any off-the-shelf network constrained by an appropriate proxy supervision to implement pairwise segmentation, which means the pairwise segmentation network just has one head for fitting the proxy supervision, where the two branches of decoders fitting ground truth of the input pairs are removed from the Siamese structure, as shown in Fig. 4. The formulation is

$$f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W}) = g_{proxy}(\mathbf{y}_i, \mathbf{y}_j), \tag{8}$$

where the two essential properties, i.e., reflexivity and symmetry are simultaneously strengthened.

Reflexivity: The reflexivity indicates that when the input pair consists of the same sample, the network can output the objects corresponding to its ground truth. Under this rule, we have

$$f(\mathbf{x}_i, \mathbf{x}_i; \mathbf{W}) = \mathbf{y}_i. \tag{9}$$

Symmetry: The symmetry signifies that the network is invariant to the ordering of pairwise inputs

$$f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{W}) = f(\mathbf{x}_j, \mathbf{x}_i; \mathbf{W}). \tag{10}$$

As our proxy supervision operator $g_{proxy}$is predefined, the symmetry rule just requires the network invariant to input permutation and a simple strategy is to aggregate the information of each input using a simple symmetric function (Zaheer et al., 2017; Qi et al., 2017). Specifically, we employ a fully parameter-shared encoder $f_{en}(\cdot;\mathbf{W}_1)$to capture features of the pairwise inputs, respectively, and aggregate the features with a permutation-invariant aggregation operation (such as element-wise *sum* operation and element-wise *max* operation), and then the resulting features are fed into the decoder $f_{de}(\cdot;\mathbf{W}_2)$to fit the proxy supervision. This can be formulated as

$$f(\mathbf{x}_i,\mathbf{x}_j;\mathbf{W}_1,\mathbf{W}_2)\approx f_{de}\Big(f_{en}(\mathbf{x}_i;\mathbf{W}_1)\odot f_{en}(\mathbf{x}_j;\mathbf{W}_1);\mathbf{W}_2\Big), \tag{11}$$

where $\odot$denotes a permutation-invariant aggregation operation.

The next cornerstone is how to predefine the proxy supervision to train the network to satisfy the reflexivity and the symmetry. According to the Eqs. (8)-(10), the proxy supervision g_{proxy}should be such a function that makes $\mathbf{y}_i$and $\mathbf{y}_j$satisfy the consistent relation, i.e., 1) reflexivity $g_{proxy}(\mathbf{y}_i,\mathbf{y}_i)=\mathbf{y}_i$; 2) symmetric $g_{proxy}(\mathbf{y}_i,\mathbf{y}_j)=g_{proxy}(\mathbf{y}_j,\mathbf{y}_i)$. Revisiting the proposed proxy supervision in Section 3.1.2, the *logic AND*operation obviously satisfies the reflexivity and symmetry; however, the *logic XOR* operation only satisfies the symmetry but not reflexivity ($\mathbf{y}_{ij}^2=\mathbf{y}_i\oplus\mathbf{y}_i=\mathbf{0}$). Actually, for the proxy supervision $\mathbf{y}_{ij}=\{\mathbf{y}_{ij}^1,\mathbf{y}_{ij}^2\}$proposed in Section 3.1.2, we have $g_{proxy}(\mathbf{y}_i,\mathbf{y}_i)=\{\mathbf{y}_i,\mathbf{0}\}$. According to Eq. (8), we can get a slice's ground truth when it is taken as input for both branches of the network f,i.e.,

$$f(\mathbf{x}_i,\mathbf{x}_i;\mathbf{W})=\{\mathbf{y}_i,\mathbf{0}\}. \tag{12}$$

Therefore, the proxy supervision proposed in Section 3.1.2 can be reused for training the one-head network $f$by minimizing the following objective function

$$\min_{\mathbf{W}}\mathbf{E}(\mathbf{W})=\sum_{i,j}\mathcal{L}\Big(\mathbf{y}_{ij},f(\mathbf{x}_i,\mathbf{x}_j;\mathbf{W_1},\mathbf{W_2})\Big), \tag{13}$$

and in the test phase, the segmentation probability maps are inferred according to Eq. (12). Concretely, we just input a pair of the same slice (i.e., $\mathbf{x}_i=\mathbf{x}_j$) and leverage the partial output maps fitting the *logic AND*proxy supervision to predict the segmentation maps. In a word, we need not to incur overhead to employ two decoders fitting each ground truth of the input pair explicitly.

3.2.2. Complexity analysis

As the proposed compact conjugate fully convolutional network can be equipped with any off-the-shelf network, we hereby analyze relative complexity between C^{2}FCN and its baseline. In fact, the parameters of C^{2}FCN have the same magnitude as its baseline, since the increase of the number of parameters is introduced by doubling the output channels, where the increased number of parameters is the same as that of the last convolutional layer of its baseline. In this paper, we take DeepLabv3+ (Chen et al., 2018b) with a ResNet-18 backbone as a baseline, and the resulting C^{2}FCN has a total of 16.60M parameters, which increases the number of parameters less than 0.01M compared with DeepLabv3+. More encouragingly, C^{2}FCN brings negligible computational overhead in the test phase, since the input pair is the same slice such that we only need to forward propagate it through the encoder and decoder once, and the additional computation is mainly determined by the increased number of parameters of the last convolutional layer, which will be negligible compared with the computational overhead that the network requires.

4. Results and discussion

In this section, we present experimental results of our proposed framework for 2D medical image segmentation. More specifically, we first substantiate the effectiveness of our proposed framework for the pathological liver segmentation. Second, we further experiment with multi-modal MR images for multi-organ segmentation. All experiments are implemented with the PyTorch framework (Paszke et al., 2017) running on two NVIDIA GTX 1080 Ti graphics cards.

4.1. Datasets and evaluation criteria

4.1.1. Datasets

For pathological liver segmentation, we evaluate our method on the public benchmark dataset of the Liver Tumor Segmentation Challenge Dataset[1] (LiTS) (Bilic et al., 2019), which consists of 201 contrast-enhanced abdominal CT volumes acquired by different scanners and protocols from multiple clinical sites. The dataset has a largely varying in-plane resolution from 0.55 mm to 1.0 mm and slice thickness from 0.45 mm to 6.0 mm. Since the challenge organizers only provided a subset of 131 volumes with manually labelled liver masks, we perform all our experiments on this subset.

As for multi-organ segmentation, we evaluate our method on Combined (CT-MR) Healthy Abdominal Organ Segmentation[2] (CHAOS) for segmentation of four abdominal organs, including liver, spleen, right kidney, and left kidney, from MRI images. In this paper, we use the second dataset (Abdominal MRI Dataset) of CHAOS to evaluate our method. This dataset was acquired by a 1.5T Philips MRI scanner with two different sequences, i.e. T1-DUAL (in-phase and out-phase) and T2-SPIR, each of which has 40 volumes (20 volumes of each sequence containing manually labelled ground truth) acquired to scan abdomen using different radio frequency pulse and gradient combinations. On average, each volume has 36 slices with a slice size of 256×256pixels. To substantiate the robustness and generalization capability of the proposed framework, the in-phase and out-phase of T1-DUAL are simply considered as two different modalities, together with the T2-SPIR, which constitute a three-modality dataset, termed as Sub-CHAOS in this paper. We perform all the multi-organ segmentation experiments on the Sub-CHAOS over 5-fold cross-validation.

4.1.2. Evaluation criteria

To quantify the segmentation accuracy of pathological liver segmentation, we follow the evaluation procedures of the LiTS challenge to compute the Dice coefficient. Dice coefficient measures the similarity of the two sets $V_{Seg}$and $V_{GT}$and is defined as

$$Dice=\frac{2|V_{GT}\bigcap V_{Seg}|}{|V_{GT}|+|V_{Seg}|}, \tag{14}$$

where $V_{Seg}$and V_{GT}denote the automatically segmented set of voxels and the manually annotated ground truth, respectively, and $|\cdot|$denotes the cardinality of a set (i.e., the total number of elements in the set). In this paper, we evaluate the segmentation performance with an average of Dice per volume score (Dice-per-case) and a global Dice score (Dice-global) which concatenates all test volumes into one long volume and computes the Dice coefficient on it. We additionally employ a distance-based evaluation metric, Average Symmetric Surface Distance (ASSD), to quantify the boundary dissimilarity of the automatic segmentation $V_{Seg}$from the ground truth V_{GT},which is defined as

$$ASSD=\frac{1}{|B_{GT}|+|B_{Seg}|}\left(\sum_{x\in B_{Seg}}d(x,B_{GT})+\sum_{x\in B_{GT}}d(x,B_{Seg})\right), \tag{15}$$

[1] https://competitions.codalab.org/competitions/17094.

[2] https://chaos.grand-challenge.org/.

where $B_{GT}$and $B_{Seg}$are the border voxel sets of $V_{Seg}$and V_{GT},respectively, and $d$is the Euclidean distance from one voxel to a voxel set.

As for evaluating multi-organ segmentation, we run all models with 5-fold cross-validation, and evaluate the performance adopting the average evaluation of Dice-per-case and ASSD scores for each organ and the average scores for all organs.

4.2. Experiments on pathological liver segmentation

Although there exist lots of sound methods for liver segmentation (Li et al., 2018b; Fang et al., 2020; Wang et al., 2020), automatic segmentation of pathological liver remains a challenge to deep FCNs as the presence of any pathology or abnormality may seriously distort the scanned texture, especially with small or moderate amount of training data.

4.2.1. Implementation details

We use the Adam Optimizer (Kingma and Ba, 2014) with a batch size of 16, a learning rate of 10^{-4},and a weight decay of $5 \times 10^{-4}$for a total of 40 epochs in training. As the voxel intensity of CT scans range from -1000 HU to over +3000 HU, we truncate the HU values of all volumes to the range of [-200, +250] HU to remove the irrelevant details, and then normalize them to [0,1] (Li et al., 2018b). For each of paired inputs, we adopt a 2.5D input with three adjacent slices, and the input pairs are sampled from one volume at intervals of 5, 9, and 13 slices along the axial direction in intra-patient pairs[3] and sampled from two random volumes with each slice paired twice in inter-patient pairs.[4]

4.2.2. Comparison experiment

We evaluate the proposed methods CFCN and C^{2}FCN on the LiTS dataset in comparison with two benchmark methods U-Net (Ronneberger et al., 2015) and DeepLabv3+ (Chen et al., 2018b), which are variants of FCN (Long et al., 2015) and have been applied to various scenarios of medical and natural image segmentation, and their effectiveness and generalization have been widely proven. Moreover, we compare the performance of CFCN and C^{2}FCN with a Siamese encoder-decoder structure ABDOCS (Chen et al., 2018a) to verify the effectiveness of the proposed structure.

U-Net (Ronneberger et al., 2015): The architecture of U-Net consists of a encoder to capture abstract features, a symmetric decoder to recover detailed location information, and a skip-connection between encoder and decoder to compensate for missing details of the pooling layers.

DeepLabv3+ (Chen et al., 2018b): The architecture integrates spatial pyramid pooling module named strous spatial pyramid pooling (ASPP) and encoder-decoder structure into a united FCN, which aids to encode multi-scale contextual information and capture sharper object boundaries by gradually recovering the spatial information.

ABDOCS (Chen et al., 2018a): The method adopts a semantic attention learner to spotlight feature channels that have high activation in all input images and suppress other irrelevant feature channels. With the attention learner, the relationship between a pair of images is captured with certain high-level features, and the information shared into these features is utilized to boost the segmentation performance. Since the network was originally used for co-segmentation, we use the segmentation loss to enable it applying to the segmentation task, and update its backbone as ResNet-18 for fair comparison.

Table 1 shows the average performance of all the comparison methods under 80% and 5% proportions of training samples (training ratio) on the LiTS dataset. It is easy to observe that the proposed CFCN and C^{2}FCN consistently outperform other methods with respect to all evaluation measures. Specifically, with a training ratio of 5%, CFCN achieves a Dice-global of 95.11%, which is 2.4% higher than standard FCN variants (Ronneberger et al., 2015; Chen et al., 2018b) and 1.9% than deep pairwise model (Chen et al., 2018a); The Dice-per-case is 95.01%, which results in about 1.8% improvement over standard FCN variants and 1.5% improvement over the deep pairwise model; The ASSD score is 1.82mm, outperforming that of standard FCN variants over 1.4mm and 0.9mm against deep pairwise model. Furthermore, we also try to train deep models with an extremely limited training set of one volume (with a training ratio of 1%). We randomly select one volume from the training set of LiTS and whose index is *volume-56*. Since in this setting the training data only have one volume, sampling the inter-patient pair is impossible. As a result, we shuffle the slices of the volume and simulate inter-patient pair with each slice paired twice (this is not implemented in our conference paper). As shown in Table 1, our CFCN outperforms the second best comparison method about 6.1%, 7.5%, and 3.0mm in Dice-per-case, Dice-global and ASSD, respectively. The results demonstrate that CFCN is of great potential in dealing with medical image segmentation under a limited number of training data. This is partly because our CFCN has a Siamese architecture where the input pairs augment the magnitude of training samples. More importantly, the CFCN model focuses on modeling the manifold of target objects and eliminating the effect of intra-class inconsistency, which is the main reason that CFCN is superior to the ABDOCS (Chen et al., 2018a).

Despite the significant performance improvement, an obvious drawback of CFCN is the increase of the number of parameters. In contrast, C^{2}FCN only has one-head output and reduces both the number of parameters and the inference time to the same level as DeepLabv3+ (Chen et al., 2018b) with an acceptable performance degradation (0.01% in Dice-global, 0.52% in Dice-per-case, and 0.26mm in ASSD with 5% training ratio) against CFCN. An interesting finding is that the performance of C^{2}FCN inclines to be superior to that of CFCN with an extremely limited amount of training data, i.e., the training ratio is 1%. This is mainly because the parameters of C^{2}FCN is fewer than CFCN, reducing the overfitting risk with such a small training set. Besides, CFCN involves an additional hyper-parameter in loss function (see Eq. (6)). Such a fixed value in training fails to make full use of the proxy supervision. By contrast, C^{2}FCN just has one-head output to fit the proxy supervision, which dominates the training of the network and reduces the computational overhead for tuning the hyper-parameter during training.

We show the qualitative segmentation results on LiTS dataset with a training ratio of 5% in Fig. 5, CFCN and C^{2}FCN are superior to the comparison methods in delineating the boundary and maintaining intra-class consistency of the pathological liver segmentation, which manifests the effectiveness of the proposed methods on modeling the manifold of target objects and eliminating the effect of intra-class inconsistency only with a small training set.

4.2.3. Ablation study

Effectiveness of the architecture components: To investigate the effect of each component of our CFCN model, we perform an ablation study on the LiTS dataset with 5% training ratio. We respectively study the ablation for the Siamese architecture, the

[3] Please refer to Fig. 2 for a more intuitive explanation. The proposed network contains two branches of input, and each one is taken for segmenting one slice, such as the upper one framed in yellow and the lower one framed in blue. Here, the two framed slices are intra-patient pairs sampled along the axial direction, and 2.5D input means we also take the adjacent slices of the two framed slices, respectively, for capturing rich context information.

[4] Sampling slice pairs from similar position of different volumes cannot be recommended, due to tedious data processing arisen by the alignment of volumes with a different number of slices and slice thickness.

Table 1
Quantitative comparison between the proposed approaches and the state-of-the-arts on the LiTS dataset (Bilic et al., 2019) with 80%, 5% and 1% proportions of training samples (training ratio), and the test set is always set as a fixed proportion of 20%.

Training Ratio	Evaluation metric	U-Net	DeepLabV3+	ABDOCS	CFCN(Ours)	C^2FCN(Ours)
80%	Dice-per-case*(%)	95.91/96.09	95.90/96.12	96.10/96.22	96.26/96.35	96.18/96.30
	Dice-global (%)	96.51	96.51	96.59	96.82	96.71
	ASSD (mm)	1.35	1.46	1.29	1.22	1.25
5%	Dice-per-case*(%)	92.01/93.25	92.19/92.65	93.20/93.53	94.76/95.01	94.29/94.49
	Dice-global (%)	92.66	92.71	93.18	95.11	95.10
	ASSD (mm)	3.81	3.25	2.73	1.82	2.08
1%	Dice-per-case*(%)	82.73/83.65	82.42/83.14	83.70/85.46	90.77/91.61	90.87/92.01
	Dice-global (%)	81.83	81.09	82.77	90.24	91.60
	ASSD (mm)	8.79	7.30	6.76	3.77	3.65

* Without/with postprocessing through a largest connected component labeling.

Table 2
The performance of ablation study for network architectures on the LiTS dataset (Bilic et al., 2019) with 5%training samples, and the test set remains fixed as Table 1.

Model	Siamese architecture	Fusion net	Multi-head output	Proxy supervision	Dice-per-case*(%)	Dice-global (%)
DeepLabv3+					92.19/92.65	92.71
SiamFCN	✓		✓		93.07/93.47	92.93
C^2FCN	✓			✓	94.29/94.49	95.10
C^2FCN+	✓	✓		✓	94.85/95.08	95.19
CFCN	✓	✓	✓	✓	94.76/95.01	95.11

* Without/with postprocessing through a largest connected component labeling.

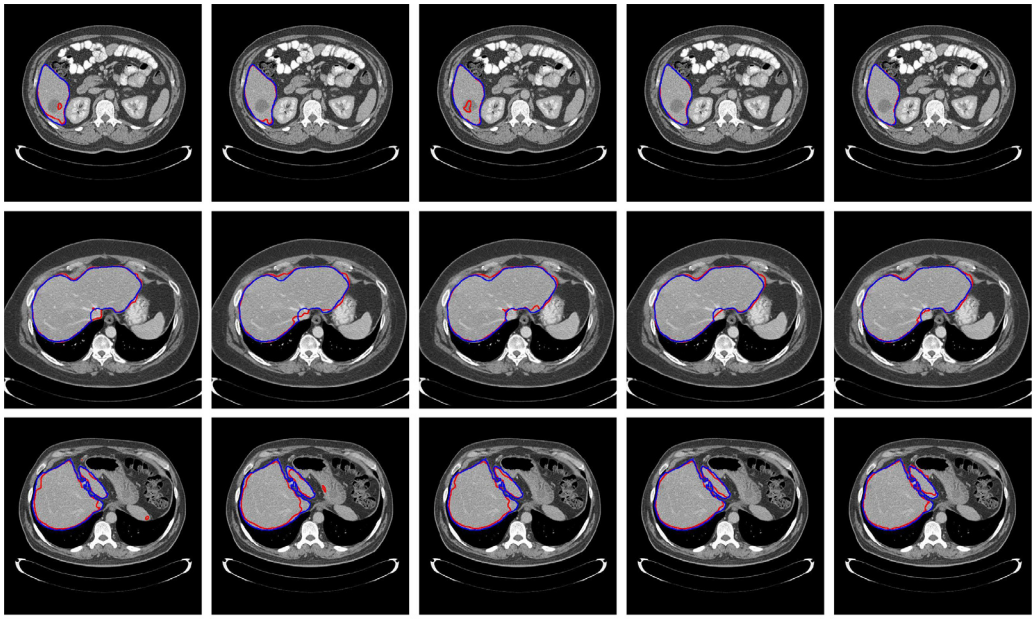

Fig. 5. Examplar segmentation results on the LiTS dataset with a training ratio of 5%. Here, blue and red lines manifest ground truth and automatic segmentation results of deep models, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

fusion net, multi-head output, and the proxy supervision. Specifically, if we assemble these model components from the baseline architecture of DeepLabv3+ (Chen et al., 2018b) in different ways, we will get four different models: 1) SiamFCN, of which the two branches are the same as DeepLabv3+ with shared weights, and for each branch the input of the decoder is compensated by features at the same layer of the other branch; 2) C^2FCN, which can be regarded as DeepLabv3+ assembling Siamese architecture and the proxy supervision, but the output is one-head; 3) C^2FCN+, replacing the decoder of CFCN with the fusion net, which can be regarded as a network removing the two segmentation branches of CFCN; 4) CFCN, in which the Siamese architecture, the fusion net, multi-head output and the proxy supervision are simultaneously included. It can be seen from Table 2that: 1) The four models achieve higher segmentation performance than DeepLabv3+ (Chen et al., 2018b), which indicates the Siamese architecture are helpful for the task as the pairwise learning is an efficient data augmentation method. 2) C^2FCN significantly outperforms SiamFCN, which shows the proposed proxy supervision plays an important role for the performance gains and the network can learn the logical relation implied in the proxy supervision. Actually, with a small training set, fitting the proposed proxy supervision requires to discriminate the logical correlation of pixels at the same location of input pairs, which improves predictive confidence to alleviate the intra-class heterogeneity. 3) C^2FCN+ inclines to achieve slight performance gains against C^2FCN, which shows the fusion

Table 3
The performance of ablation study for intra/inter-patient pairs on the LiTS dataset (Bilic et al., 2019) with 5%training samples, and the test set remains fixed as Table 1.

	Dice-per-case (%)*	Dice-global (%)
Identical pairs	89.98/93.07	90.12
Intra-patient pairs	93.20/94.69	93.14
Inter-patient pairs	94.31/94.53	94.78
Intra & inter-patient pairs	94.76/95.01	95.11

* Without/with postprocessing through a largest connected component labeling.

net can promote the fitting of the proxy supervision. 4) Compared with C^{2}FCN, CFCN mainly equips two segmentation branches to learn the ground truth for input pairs, respectively. Its performance is just comparative with C^{2}FCN and C^{2}FCN+, which further manifests that the proposed proxy supervision are sufficient to exploit the semantic information for the segmentation task.

Effectiveness of Intra-patient Pairs and Inter-patient Pairs: As aforementioned, CFCN and C^{2}FCN take paired slices as input, which quadratically augment the training samples. For the efficiency of network training, we sample the input pairs in two ways: intra-patient pairs and inter-patient pairs, as shown in Fig. 3. Concretely, the input pairs are sampled from one volume at intervals of 5, 9, and 13 slices in intra-patient pairs, and sampled from two random volumes with each slice paired twice in inter-patient pairs. The results on the LiTS dataset with 5% training samples are listed in Table 3. It can be observed that both intra-patient pairs and inter-patient pairs largely improve the segmentation performance compared with naive CFCN, which is trained using thesame slice taken as input pairs. Moreover, the two ways of sampling the input pairs are complementary, since it achieves more performance gain when intra-patient and inter-patient pairs are used together.

4.3. Experiments on multi-organ segmentation

Multi-organ segmentation Gibson et al. (2018); Peng et al. (2020) from different modalities is another challenging but valuable task for many clinical procedures.

4.3.1. Implementation details

We split the Sub-CHAOS data into 5-folds for cross-validation, and in each round the training set contains 16 cases and the test set remains four cases. For each patient's case, the image is clipped to the [2.0, 98.0]percentiles of the intensity values of the entire image and normalized to [0,1] in the end, and all slices in the training phase are resized to 256 $\times$256 yet kept the original size in the test phase. We train the segmentation networks using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4},a weight decay of $5 \times 10^{-4}$with mini-batch size of 20 for a total of 200 epochs. The input pairs are sampled from one volume at intervals of 3, 5, and 7 slices in intra-patient pairs and sampled from two random volumes with each slice paired twice in inter-patient pairs.

4.3.2. Results and discussion

Since we run all models with 5-fold cross-validation, we present the results adopting the average evaluation of Dice-per-case and ASSD scores for each organ and the average scores for all organs in Tabel 4. As shown, 1) compared with the baseline DeepLabv3+ (Chen et al., 2018b), our CFCN achieves a large performance improvement, i.e., average Dice-per-case score increase from 86.37% to 88.25% while ASSD decreases from 3.14mm to 2.83mm; 2) in terms of Dice-per-case and ASSD, the proposed C^{2}FCN achieves average scores of 88.62% and 2.47mm, respectively, which significantly outperform all comparison methods; 3) in contrast to left kidney segmentation, the overall performance of C^{2}FCN is better than that of CFCN. It is probably because the proxy supervision acts as global constraints in the training phase of CFCN, and its contribution is affected by the hyper-parameter λin Eq. (6). Such a single parameter fails to adequately imbalance the divergence of different categories in multi-category segmentation compared with binary segmentation. As a result, the performance gain of CFCN is not as significant as that of C^{2}FCN, where the proxy supervision is treated as a complete supervision.

5. Discussion and conclusion

In this paper, we proposed a new framework for medical image segmentation with a limited number of training samples. Specifically, we focused on addressing the challenging issues arisen by intra-class heterogeneity and boundary ambiguity. Extending our preliminary work (Wang et al., 2019), we improved our framework from two-fold: First, we extended the proposed CFCN to a general pairwise learning framework in Section 3, where the proxy supervision acted as a global constraint on the network to fit inherent prior knowledge from the label space. Except for binary segmentation on the LiTS dataset (Bilic et al., 2019), we further extended CFCN to multi-category segmentation for multi-organ segmentation on benchmark dataset CHAOS in Section 4.3, and the results demonstrated that our CFCN could achieve state-of-the-art results among all comparison methods.

Second, we extended the CFCN to a compact architecture C^{2}FCN in Section 3.2, which can equip with any off-the-shelf segmentation networks with a negligible number of additional parameters and computational overhead in the test phase. Specifically, we adopted DeepLabv3+ (Chen et al., 2018b) as baseline, the resulting C^{2}FCN achieved competitive results on the LiTS dataset

Table 4
Quantitative comparison between the proposed approaches and the state-of-the-art on the Sub-CHAOS dataset. Dice-per-case scores (±std) are reported in percentage over 5-fold cross-validation.

Method	Liver	Right Kidney	Left Kidney	Spleen	Mean
Dice-per-case (%)					
U-Net	90.70±3.68	86.06±4.92	85.43±4.99	81.14±7.82	85.83
DeepLabv3+	90.46±2.84	86.91±1.94	86.00±4.86	82.12±7.12	86.37
ABDOCS	90.90±2.78	88.77±1.30	86.97±4.78	84.33±4.58	87.74
CFCN (ours)	91.76±2.28	89.38±1.79	88.45±4.19	83.41±6.13	88.25
C^{2}FCN (ours)	92.09±1.52	89.38±2.40	88.17±4.17	84.82±4.66	88.62
ASSD (mm)					
U-Net	4.18±2.19	3.46±1.16	3.53±1.25	7.79±7.41	4.74
DeepLabv3+	3.42±1.13	2.49±0.78	2.95±1.17	3.71±1.63	3.14
ABDOCS	2.95±0.80	1.89±0.33	2.09±0.75	3.73±1.48	2.67
CFCN (ours)	2.99±1.17	1.78±0.41	2.06±0.79	4.49±3.34	2.83
C^{2}FCN (ours)	2.75±0.68	1.73±0.44	2.20±1.21	3.19±1.94	2.47

(see Table 1) and superior results on the Sub-CHAOS dataset (see Table 4) compared with CFCN. However, the number of parameters and computational overhead of C^2FCN were largely reduced during training against those of CFCN. More importantly, the proposed C^2FCN learned the logical relation implied in proxy supervision with only one head in the training phase, and the segmentation probability was inferred by the learned logical relation in the test phase. Compared with CFCN, there was no need to incur overhead of two decoders to explicitly fit the ground truth of the input pair in C^2FCN, which manifested that learning the general relation through deep models is feasible and potential.

It is worth mentioning that we employed the proposed pairwise segmentation framework to address the challenges arisen by intra-class heterogeneity and boundary ambiguity in this paper. More prior information, however, can be exploited by explicitly designing the function g_{proxy}in Eq. (2). Another research line in the future is to employ the proposed CFCN and C^2FCN to more segmentation scenarios, especially the automatic segmentation of a tube-like structure, such as a blood vessel or small bowel, which often folds in the 3D space and the shape prior is difficult to model by the intra-patient adjacent slices. Besides, extending our method to 3D models will be a direction of our future research.

In conclusion, we proposed a new pairwise segmentation framework in this paper for medical image segmentation with a limited number of training samples. To address intra-class heterogeneity and boundary ambiguity in medical imaging, we proposed a proxy supervision to explicitly encode the prior information from the label space, which acted as a global constraint on the network in the training phase. Particularly, we proposed a new segmentation paradigm through C^2FCN, where the network aimed to learn the logical relation between the input pair rather than to directly fit the ground truth of the targets as conventional FCNs did, and the segmentation probability of a test sample is inferred by the learned logical relation. The experimental results demonstrated that the proposed pairwise segmentation could significantly improve segmentation accuracy with a limited amount of training data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Renzhen Wang: Conceptualization, Methodology, Software, Investigation, Resources, Writing - original draft. **Shilei Cao:** Methodology, Investigation, Software, Validation. **Kai Ma:** Investigation, Supervision, Writing - review & editing. **Yefeng Zheng:** Resources, Supervision, Writing - review & editing. **Deyu Meng:** Methodology, Resources, Supervision, Project administration, Writing - review & editing.

Acknowledgement

This work was supported by National Key R&D Program of China (2018YFB1004300), the China NSFC (11690011, 61721002, U1811461), the Key Area Research and Development Program of Guangdong Province, China (2018B010111001), and the Science and Technology Program of Shenzhen, China (No. ZDSYS201802021814180).

References

Araújo, R.J., Cardoso, J.S., Oliveira, H.P., 2019. A deep learning design for improving topology coherence in blood vessel segmentation. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention. Springer, pp. 93–101.

Banerjee, S., Hati, A., Chaudhuri, S., Velmurugan, R., 2019. CoSegNet: image co-segmentation using a conditional Siamese convolutional network. In: Proc. Int'l Joint Conf. on Artificial Intelligence. AAAI Press, pp. 673–679.

BenTaieb, A., Hamarneh, G., 2016. Topology aware fully convolutional networks for histology gland segmentation. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention. Springer, pp. 460–468.

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H., 2016. Fully-convolutional Siamese networks for object tracking. In: Proc. European Conf. Computer Vision. Springer, pp. 850–865.

Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.-W., Han, X., Heng, P.-A., Hesser, J., et al., 2019. The liver tumor segmentation benchmark (LiTs). arXiv preprint arXiv:1901.04056.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a "Siamese" time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744.

Chen, H., Huang, Y., Nakayama, H., 2018. Semantic aware attention based deep object co-segmentation. In: Asian Conf. Computer Vision. Springer, pp. 435–450.

Chen, H., Qi, X., Yu, L., Heng, P.-A., 2016. DCAN: Deep contour-aware networks for accurate gland segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2487–2496.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Analysis Machine Intell. 40 (4), 834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. European Conf. Computer Vision, pp. 801–818.

Chen, S., Ma, K., Zheng, Y., 2019. Med3D: transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625.

Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1, pp. 539–546.

Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., et al., 2016. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention. Springer, pp. 415–423.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Annual Conf. Medical Image Understanding and Analysis. Springer, pp. 506–517.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Imag. Analysis 41, 40–54.

Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R., Naik, N., 2018. Pairwise confusion for fine-grained visual classification. In: Proc. European Conf. Computer Vision, pp. 70–86.

Fang, X., Du, B., Xu, S., Wood, B.J., Yan, P., 2020. Unified multi-scale feature abstraction for medical image segmentation. In: Medical Imaging 2020: Image Processing, 11313. International Society for Optics and Photonics, p. 1131319.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE Trans Med Imaging 37 (8), 1822–1834.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: achievements and challenges. J. of Digital Imaging 1–15.

Ke, T.-W., Hwang, J.-J., Liu, Z., Yu, S.X., 2018. Adaptive affinity fields for semantic segmentation. In: Proc. European Conf. Computer Vision, pp. 587–602.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kronman, A., Joskowicz, L., 2016. A geometric method for the detection and correction of segmentation leaks of anatomical structures in volumetric medical images. Int. J. of Computer Assisted Radiology and Surg. 11 (3), 369–380.

Li, W., Jafari, O.H., Rother, C., 2018. Deep object co-segmentation. In: Asian Conf. Computer Vision. Springer, pp. 638–653.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A., 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging 37 (12), 2663–2674.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Imag. Analysis 42, 60–88.

Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., Kautz, J., 2017. Learning affinity via spatial propagation networks. In: Advances in Neural Information Processing Systems, pp. 1520–1530.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 3431–3440.

Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F., 2019. See more, know more: unsupervised video object segmentation with co-attention Siamese networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 3623–3632.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth Int'l Conf. on 3D Vision, pp. 565–571.

Mosinska, A., Marquez-Neila, P., Koziński, M., Fua, P., 2018. Beyond the pixel-wise loss for topology-aware delineation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 3136–3145.

Paszke, A., Gross, S., Chintala, S., et al., 2017. Automatic differentiation in PyTorch. In: Advances in Neural Information Processing Systems Workshop Autodiff, pp. 1–4.
Peng, Z., Fang, X., Yan, P., Shan, H., Liu, T., Pei, X., Wang, G., Liu, B., Kalra, M.K., Xu, X.G., 2020. A method of rapid quantification of patient-specific organ doses for ct using deep-learning-based multi-organ segmentation and gpu-accelerated monte carlo dose computing. Med Phys.
Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging 35 (5), 1240–1251.
Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 652–660.
Ravishankar, H., Thiruvenkadam, S., Venkataramani, R., Vaidya, V., 2017. Joint deep learning of foreground, background and shape for robust contextual segmentation. In: Int'l Conf. Information Processing in Medical Imaging. Springer, pp. 622–632.
Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., et al., 2017. Learning and incorporating shape models for semantic segmentation. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention, pp. 203–211.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention, pp. 234–241.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Singh, K.K., Lee, Y.J., 2016. End-to-end localization and ranking for relative attributes. In: Proc. European Conf. Computer Vision. Springer, pp. 753–769.
Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. DeepFace: Closing the gap to human-level performance in face verification. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1701–1708.
Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312.
Tao, R., Gavves, E., Smeulders, A.W., 2016. Siamese instance search for tracking. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1420–1429.
Wang, R., Cao, S., Ma, K., Meng, D., Zheng, Y., 2019. Pairwise semantic segmentation via conjugate fully convolutional network. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention. Springer, pp. 157–165.
Wang, S., Cao, S., Chai, Z., Wei, D., Ma, K., Wang, L., Zheng, Y., 2020. Conquering data variations in resolution: aslice-aware multi-branch decoder network. IEEE Trans Med Imaging.
Wu, L., Xin, Y., Li, S., Wang, T., Heng, P.-A., Ni, D., 2017. Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation. In: Proc. IEEE Int'l Sym. Biomedical Imaging. IEEE, pp. 663–666.
Xie, Q., Zeng, D., Zhao, Q., Meng, D., Xu, Z., Liang, Z., Ma, J., 2017. Robust low-dose ct sinogram preprocessing via exploiting noise-generating mechanism. IEEE Trans Med Imaging 36 (12), 2487–2498.
Yu, C., Wang, J., Peng, C., et al., 2018. Learning a discriminative feature network for semantic segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1857–1866.
Zagoruyko, S., Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 4353–4361.
Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J., 2017. Deep sets. In: Advances in Neural Information Processing Systems, pp. 3391–3401.
Zhang, J., Xie, Y., Wu, Q., Xia, Y., 2018. Skin lesion classification in dermoscopy images using synergic deep learning. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention. Springer, pp. 12–20.
Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks. In: Proc. Int'l Conf. Computer Vision, pp. 1529–1537.
Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: generic autodidactic models for 3D medical image analysis. In: Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention. Springer, pp. 384–393.