Self-supervised regularization for abdominal multi-organ segmentation

Zhengshan Huang[1], Yu Guo[1]

[1]School of Precision Instrument and Opto-electronics Engineering, Tianjin University, Tianjin, China.

Abstract

Automated segmentation of abdominal multi-organ from 3D computed tomography images (CTs) is necessary for organ quantification, surgical planning, and disease diagnosis. Manual delineation practices require anatomical knowledge, are expensive, time consuming and can be inaccurate due to human error. Here, we describe a semi-supervised efficient context-aware segmentation network for abdominal multi-organ segmentation from 3D CTs based on encoder-decoder architecture. For learning more useful feature information from unlabeled cases, a variational auto-encoder branch is added to reconstruct the input image itself in order to regularize the shared encoder and impose additional constraints on its layers. And for the purpose of consuming less source, an efficient context-aware segmentation backbone network is used in this paper.

Keywords

Semi-supervised Learning; Abdominal Multi-organ Segmentation; Efficient Context-Aware Network

1 Introduction

In this paper, we focus on semi-supervised abdominal multi-organ segmentation from CT scans. Here we obtain 50 labeled cases and 2000 unlabeled cases. As shown in Figure 1, the main difficulties stem from four aspects: 1) The variations in field-of-views, shape and size of different organs. 2) The abnormalities, like lesion-affected organ, may lead to segmentation failure. 3) The diversity of data source in term of multi-center, multi-phase and multi-vendor cases. 4) The limited GPU memory size and high computation cost. 5) The effectiveness of extracting organ feature information from unlabeled cases.

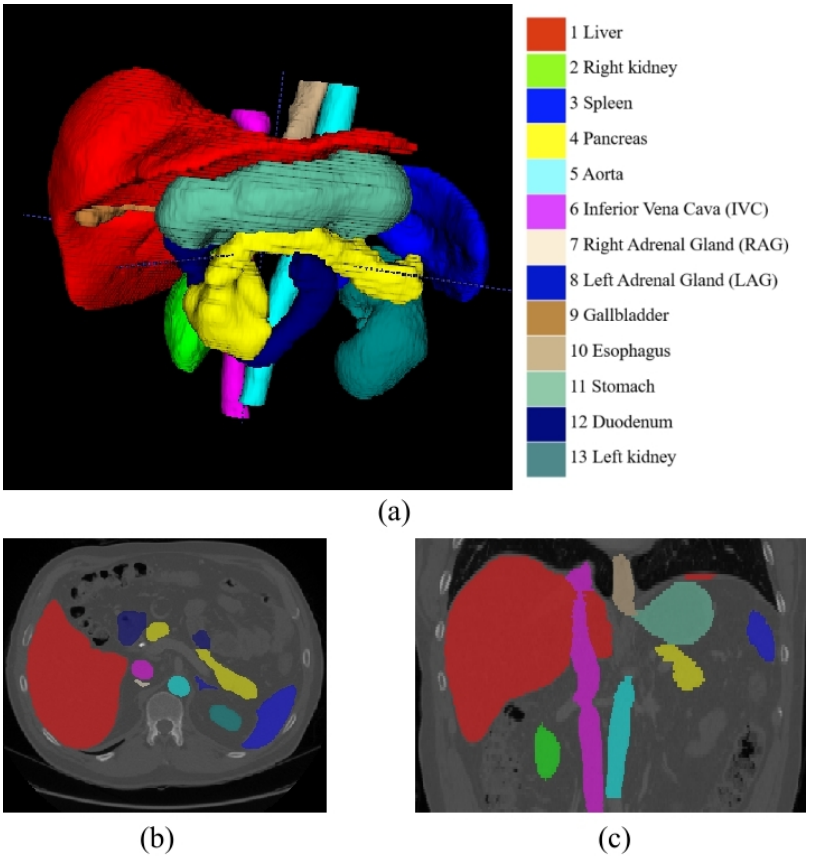


Figure 1. 13 abdominal organs of a person. (a) 3D organ segmentation display. (b-c) Transverse and coronal segmentation illustrations.

A common solution [1] is to develop a sliding-window method, which can balance the GPU memory usage. Usually, this method need to sample sub-volumes overlap with each other to improve the segmentation accuracy, while leading to more computation cost. Meanwhile, sub-volumes sampled from entire CT volume inevitably lose some 3D context, which is important for distinguishing multi-organ with respect to background.

2 Methods

As mentioned in Figure 2, this end-to-end semi-supervised network is composed of segmentation flow and reconstruction flow. A detail description of the method is as follows.

2.1 Preprocessing

The baseline method includes the following preprocessing steps:

- Reorientation image to target direction.
- Resampling image to fixed size. [160, 160, 160]
- Intensity normalization: First, the image is clipped to the range [-325, 325]. Then a z-score normalization is applied based on the mean and standard deviation of the intensity values.

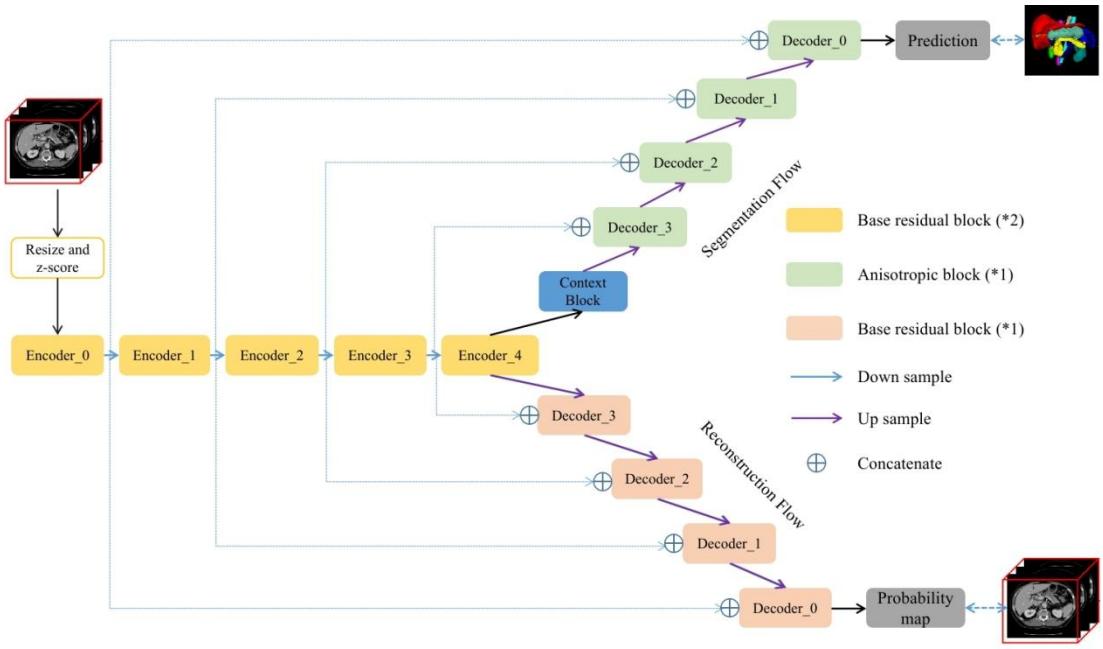
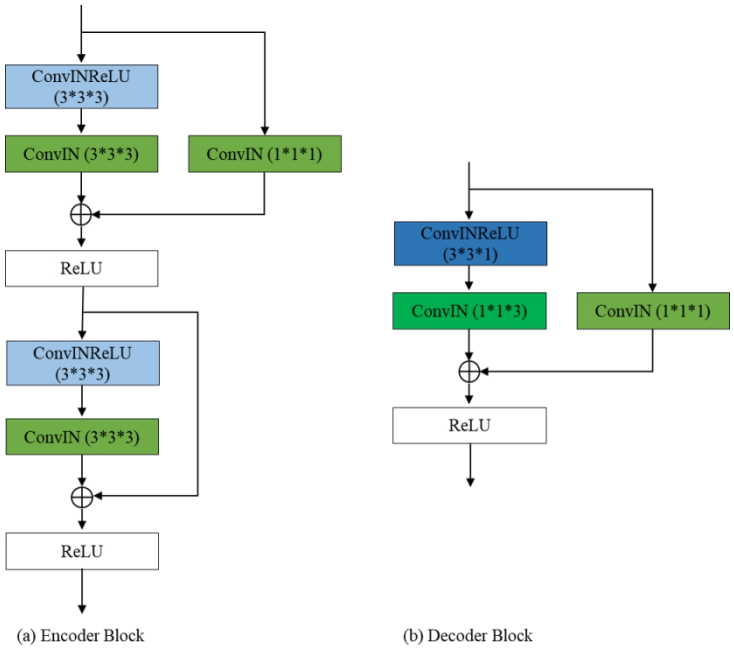


Figure 2. A schematic diagram of end-to-end semi-supervised segmentation network.



(a) Encoder Block (b) Decoder Block

Figure 3. Illustration of the encoder block and decoder block.

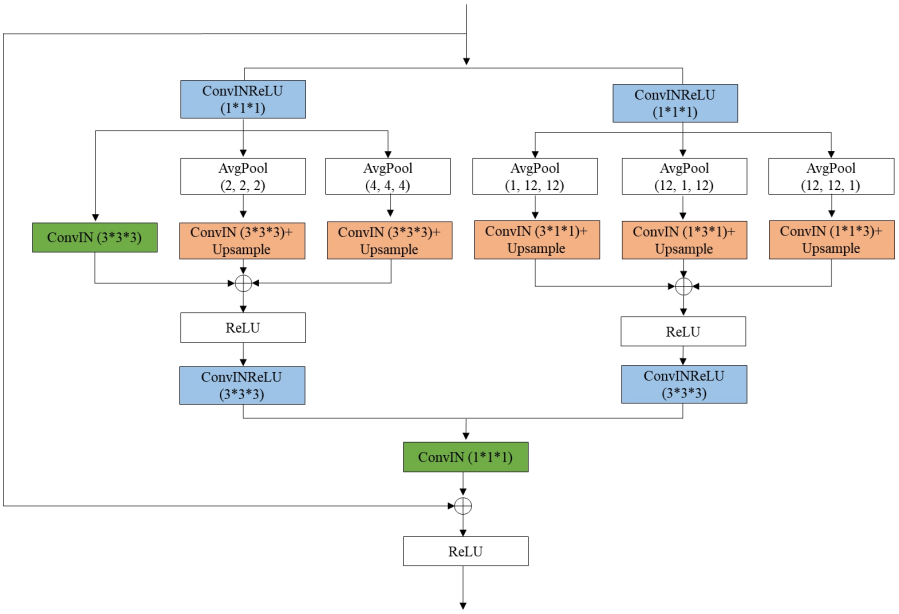


Figure 4. Illustration of the context block.

2.2 Proposed Method

The proposed framework consists of four major parts: the feature encoder module, the context extractor module, the feature decoder module for segmentation and the feature decoder module for reconstruction, as shown in Figure 2.

As depicted in Figure 3, the encoder module is composed of two residual convolution blocks, and the decoder module with one residual convolution block. As to decoder module, we separate a standard 3D convolution with kernel size 3×3×3 into a 3×3×1 intraslice convolution and a 1×1×3 inter-slice convolution. The residual convolution block is implemented as follows: conv-instnorm-ReLU-conv-instnorm-ReLU (where the addition of the residual takes place before the last ReLU activation). We adopt 3D-based mixed pyramid pooling (Figure 4) to extract contextual feature, which is composed of the standard spatial pooling and the anisotropic strip pooling. The standard spatial pooling employs two average pooling with the stride of 2×2×2 and 4×4×4. The anisotropic strip pooling with three different-direction receptive fields: 1×N×N, N×1×N and N×N×1, where N is the size of feature map in last encoder module.

3 Dataset and Evaluation Metrics

3.1 Dataset

There are a small number of labeled cases (50) and a large number of unlabeled cases (2000) in the training set, 50 visible cases for validation, and 200 hidden cases for testing. The segmentation targets include 13 organs: liver, spleen, pancreas, right kidney, left kidney, stomach,

gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum.

3.2 Evaluation Metrics

- Dice Similarity Coefficient (DSC)
- Normalized Surface Distance (NSD)
- Running time
- Maximum used GPU memory (when the inference is stable)

4 Results and discussion

Due to various reasons, our team did not have enough time to prepare for the challenge, so we achieved worse verification results.

Acknowledgements

The authors claim that any pre-trained models and addition datasets were not used in this experiment.

The proposed solution is fully automatic without any manual intervention.

[1] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.