Multi-dimensional Fusion and Consistency for Semi-supervised Medical Image Segmentation

Yixing Lu[1], Zhaoxin Fan[2], Min Xu[2]

[1] University of Liverpool [2] MBZUAI

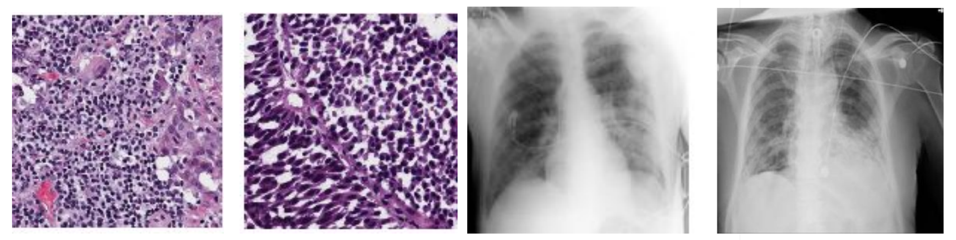
Arxiv Page

Motivation

For medical image segmentation:

1. Medical Data sources can also come from the language modality (e.g., medical text report) in addition to visual images.
2. Rising use of Vision Transformer (ViT) while Convolutional Neural Network (CNN) remains the state-of-the-art.
3. Urgent need to reduce the model's dependency on annotated medical image-mask pairs.

Example visual image input and text input

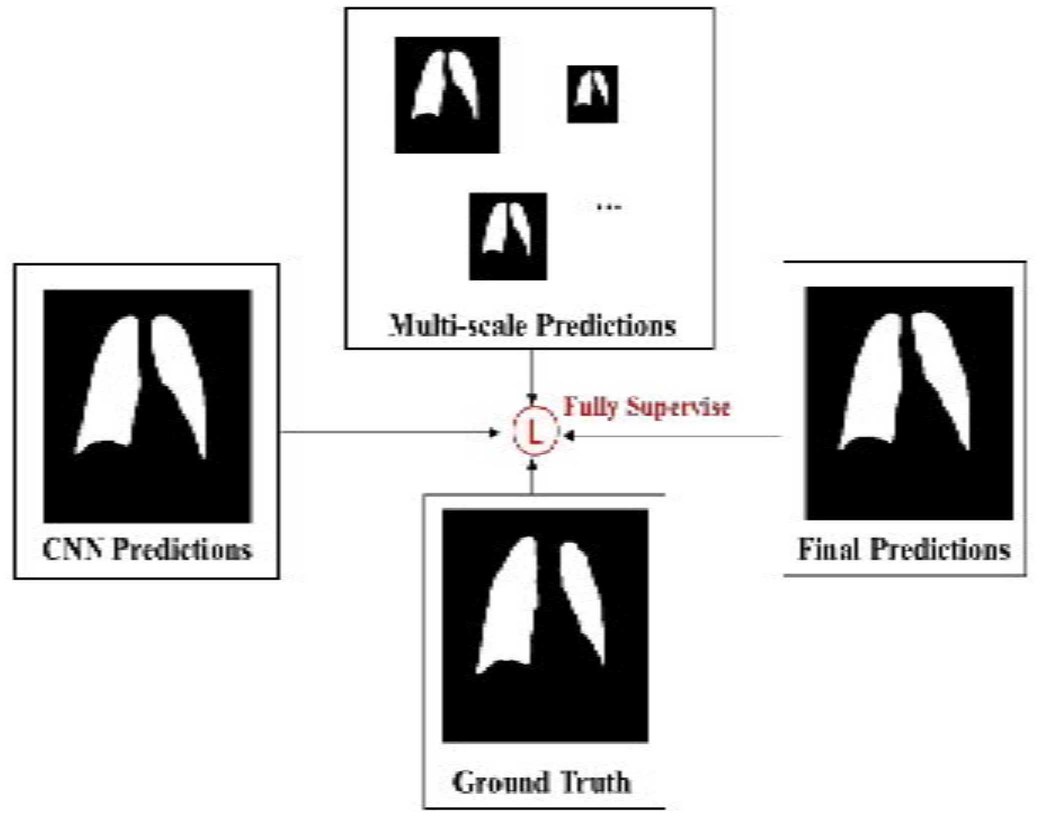


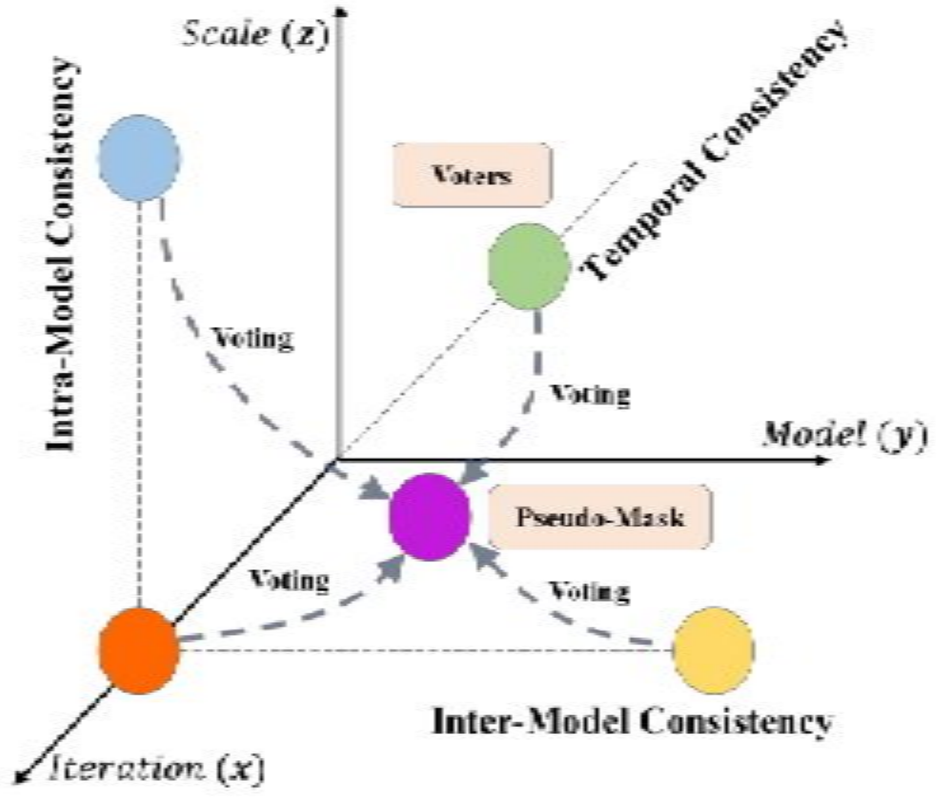
Bilateral pulmonary infection, two infected areas, all left lung and all right lung.
Unilateral pulmonary infection, one infected area, middle left lung.
The nuclei density in the left is high.
......

Contribution

We pioneer a semi-supervised framework that harnesses the power of textual information to support fused ViT-CNN networks for medical image segmentation:

- A novel Multi-scale Text-aware ViT-CNN Fusion methodology to boost segmentation accuracy.
- A Multi-Axis Consistency Learning module that capitalizes on consistency regularizations for semi-supervised learning.



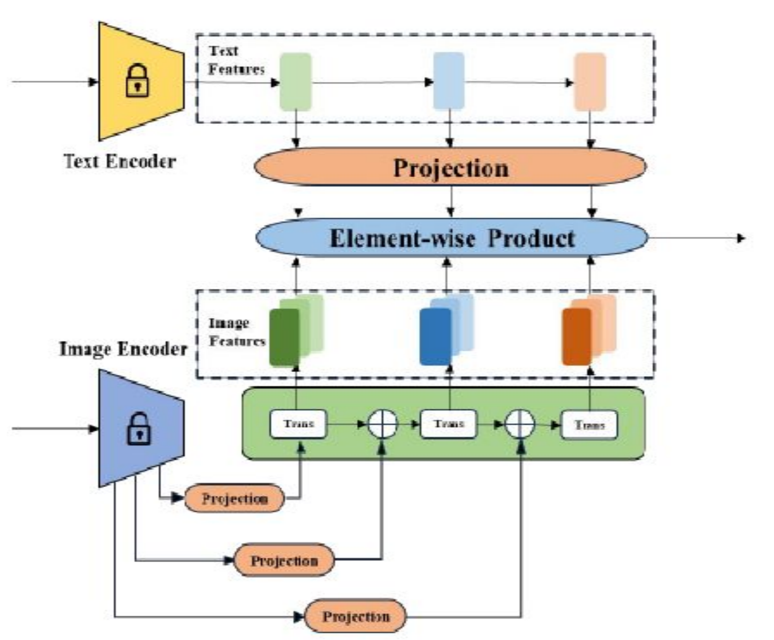


Proposed Framework

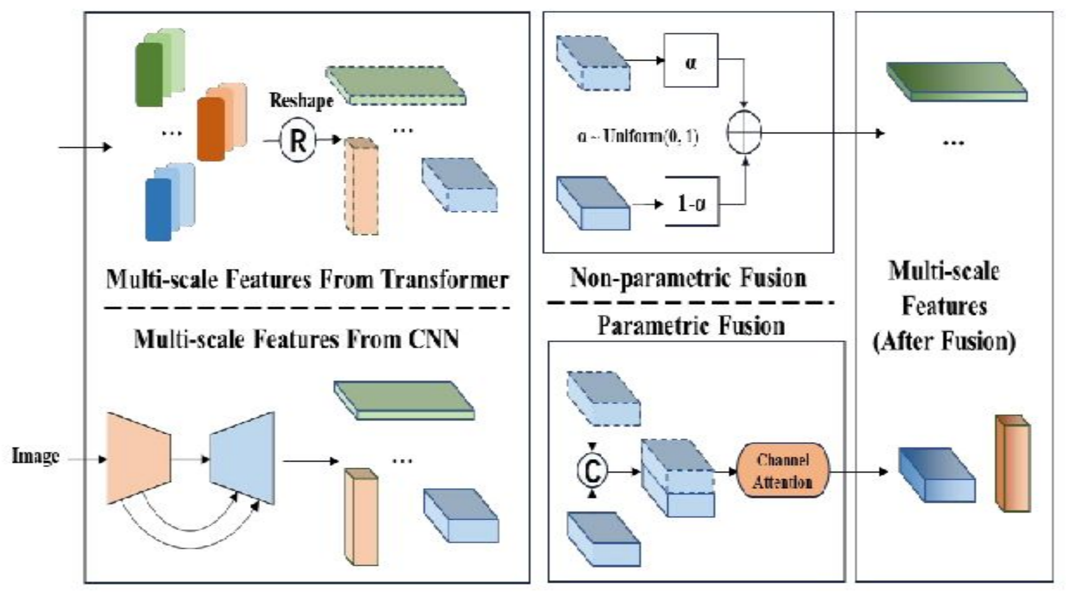
1. Multi-scale Text-aware ViT-CNN Fusion

We present an architectural design named Multi-scale Text aware ViT-CNN Fusion:

- Dense Vision-Language Alignment module: Fuse dense features from both data source modalities.
- Multi-scale ViT-CNN Fusion module: Fuse dense features from both model architectures.



a) Dense Vision-Language Alignment module

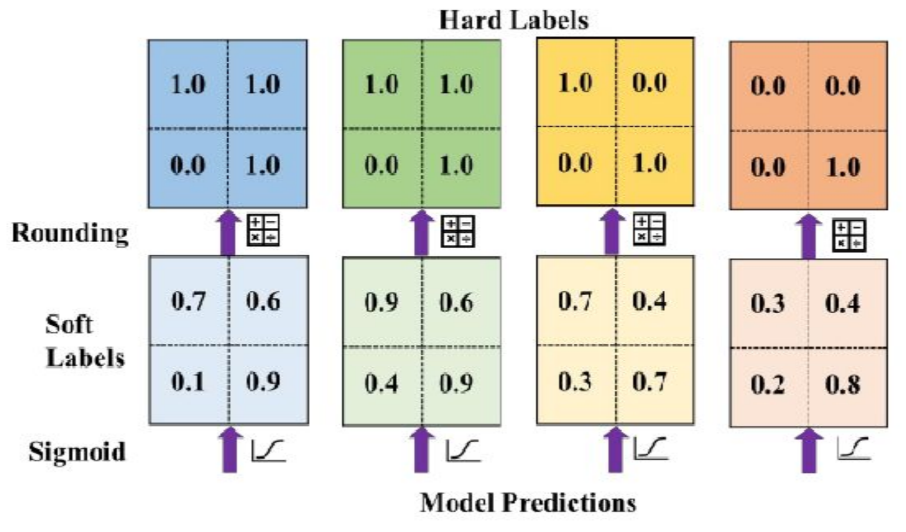


b) Multi-scale ViT-CNN Fusion module

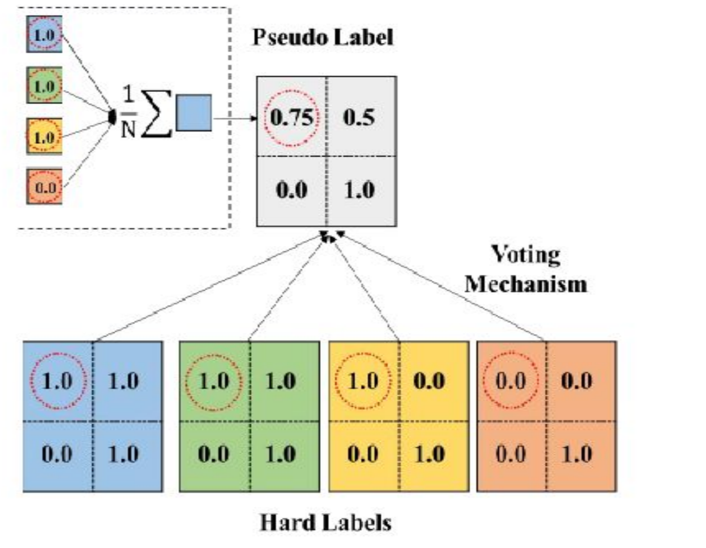
2. Multi-Axis Consistency Framework

We propose the Multi-Axis Consistency framework:

- Soft-Hard Label Generation Module: Post-process model predictions.
- Voting Mechanism: Aggregate predictions from different sources to generate pseudo-label for supervision.



a) Multi-Axis Consistency Soft-Hard Label Generation Module.



b) Multi-Axis Consistency Voting Mechanism

Q: How to learn?

Sol: Leveraging multi-dimensional predictions and consistency!

Experiments

- **Quantitative Results**

Method	MoNuSeg Dice (%)	MoNuSeg mIoU (%)	QaTa-COV19 Dice (%)	QaTa-COV19 mIoU (%)
Unet	76.45	62.86	79.02	69.46
Unet++	77.01	63.04	79.62	70.25
AttUnet	76.67	63.74	79.31	70.04
nnUnet	80.06	66.87	80.42	70.81
MedT	77.46	63.37	77.47	67.51
TransUnet	78.53	65.05	78.63	69.13
GTUnet	79.26	65.94	79.17	69.65
Swin-Unet	77.69	63.77	78.07	68.34
UCTransNet	79.87	66.68	79.15	69.60
Ours+PF	79.91	66.74	**82.29**	**72.87**
Ours+NPF	**80.60**	**67.66**	82.03	72.80

Fully-supervised setting

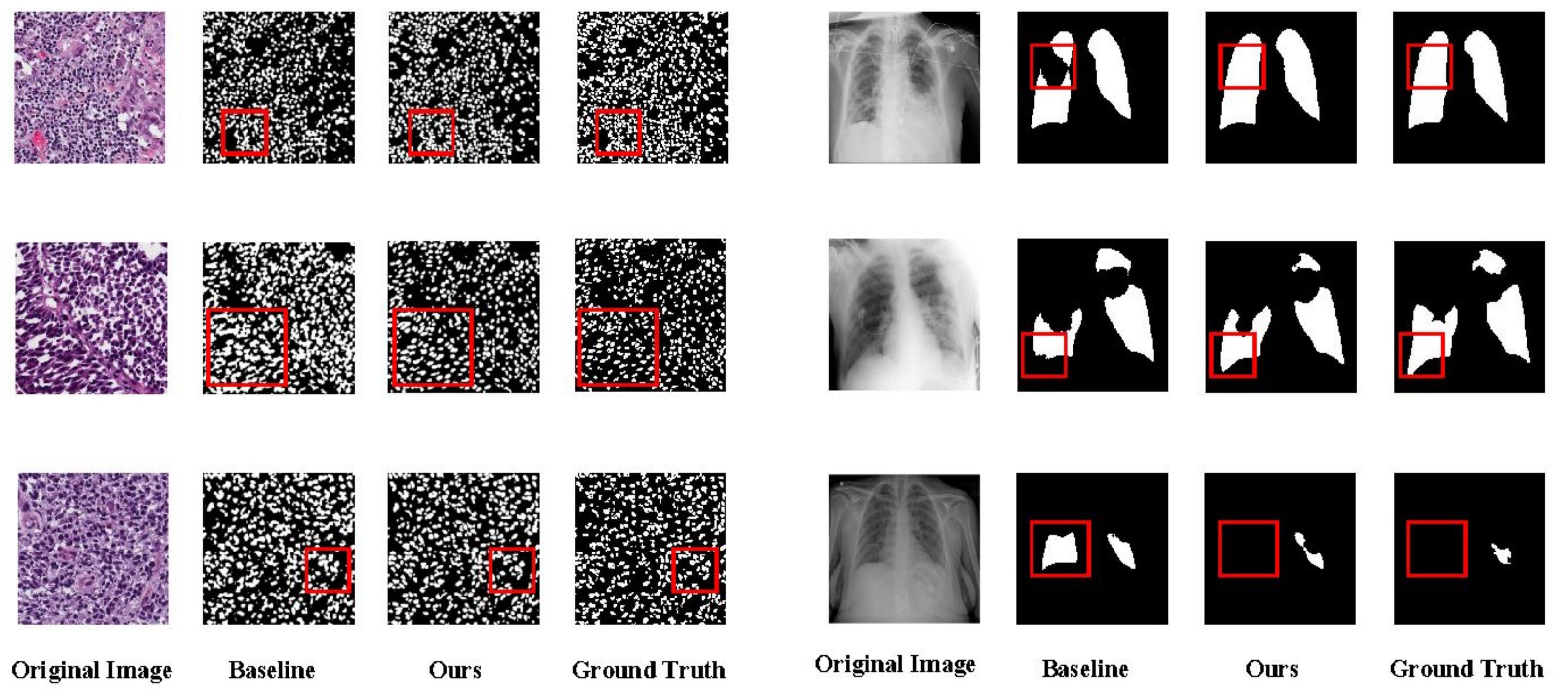
Please also see our paper for more experimental analysis (e.g., ablation study)!

Setting	Labels (%)	MoNuSeg Dice (%)	MoNuSeg mIoU (%)
PF	25	78.59	64.99
	50	78.85	65.36
	100	79.91	66.74
NPF	25	78.47	64.88
	50	79.26	65.94
	100	80.16	67.06

Semi-supervised setting

*** PF and NPF represents Parametric fusion and Non-Parametric Fusion, respectively.**

- **Qualitative Results**



*** See our paper for the baseline setting.**

Conclusion

In this paper, we propose a novel semi-supervised learning framework for medical image segmentation. In our work, a Text-aware ViT-CNN Fusion scheme is proposed to take advantages of both pretrained ViTs and CNNs as well as extracting both abstract features and medical domain specific features. Besides, a novel Multi-Axis Consistency framework is proposed to vote for pseudo label to encourage semi-supervised training. Experiments on several widely used datasets have demonstrated the effectiveness of our method.