Self-supervised 3D Anatomy Segmentation Using Self-distilled Masked Image Transformer (SMIT)

Jue Jiang, Neelam Tyagi, Kathryn Tringale, Christopher Crane and Harini Veeraraghavan

Memorial Sloan Kettering Cancer Center - New York, USA

Poster made by : Elycheva DRAY, Théo DI PIAZZA - Group 28

Memorial Sloan Kettering Cancer Center

Context, Motivation

Multi-organ 3D segmentation is a very important area of research in medical imaging because when it is done correctly and accurately, it can help for the delineation of certain anatomical structures to assist physicians in surgery, biopsies, and other clinical tests.

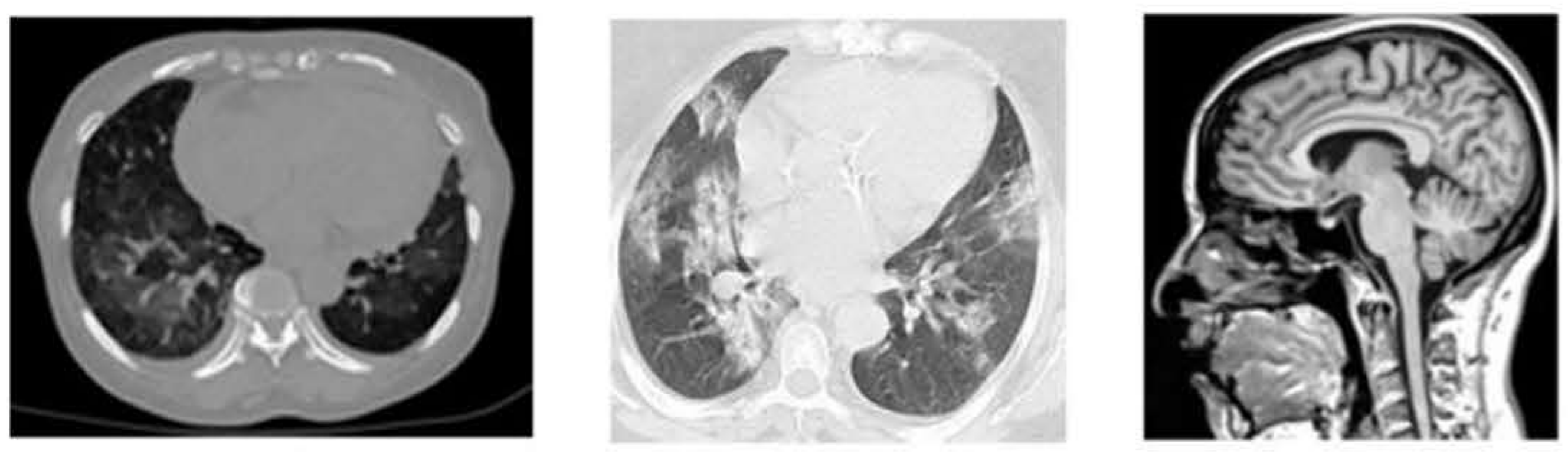

Fig.1 : 3 scans from CT/MRI datasets

Indeed, segmentation can assist in the diagnosis of diseases by isolating regions of interest. For example, it can aid in tumour detection, and also with treatment planning by providing information about structure of interest. In the state of the art, many supervised learning models have been developed for this segmentation task to predict for each image's pixel, the associated organ.

However, these methods require a large amount of labeled data and the labeling process is costly, time consuming and difficult to obtain for medical imaging. That's why unsupervised-learning methods are introduced. The objective of SMIT [1] to pre-train a backbone able to learn relevant visual information before being fine-tuned for organ segmentation.

Method, Contribution

The SMIT method has 2 main steps : the first step is to pre-train a backbone, and the second is to fine-tune this backbone to a segmentation task.

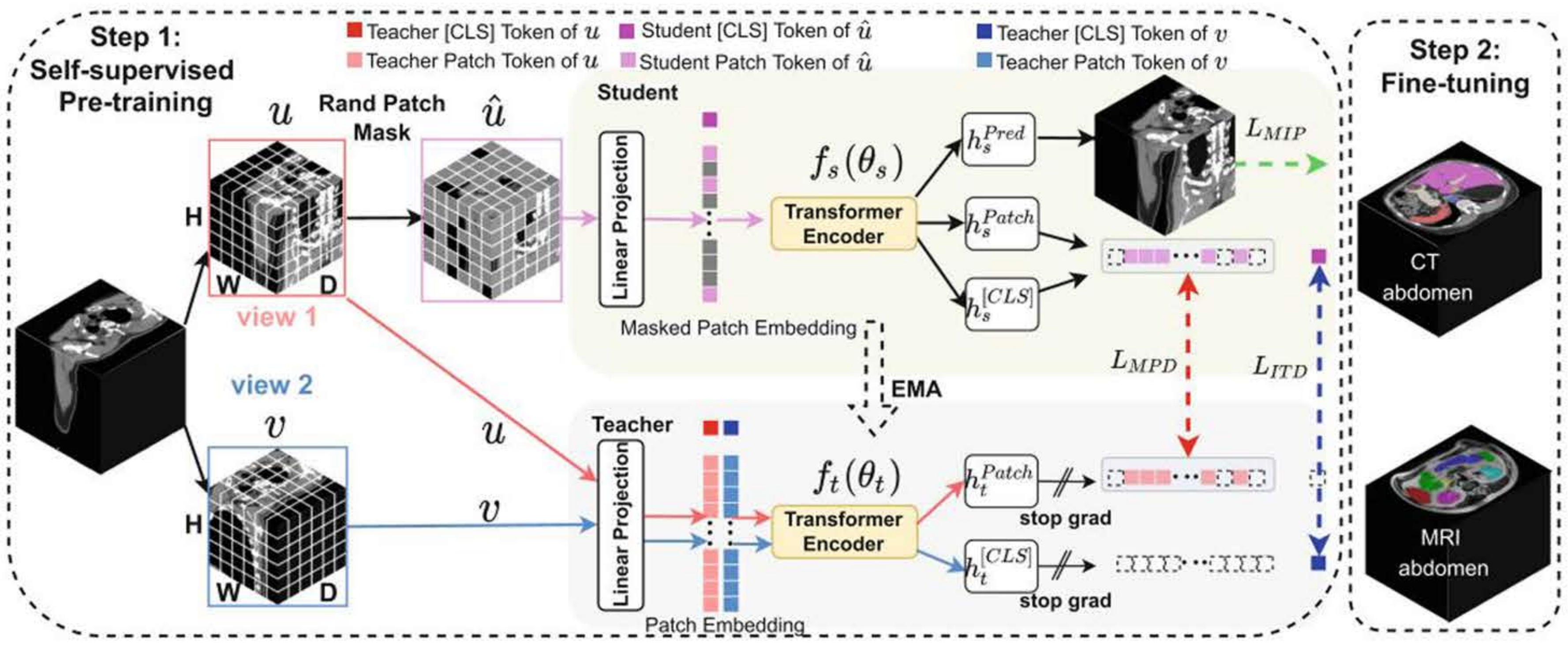

Fig.2 : Scheme of the SMIT method

Step 1 :

1. From a given 3D scan image : 2 different views are obtained : u and v.
2. From u, some patches are masked to obtain a corrupted view : û.
3. Teacher receives : u and v ; and predicts the patch tokens of masked patch of u : h_t^{patch} and the global image token of v : h_t^{cls} with the ViT Encoder.
4. Student receives : û ; to predict masked patches h_s^{pred}, the patch tokens of masked patch h_s^{patch}, the global image token h_s^{cls} , with the ViT Encoder.
5. The loss function is computed with these 3 terms :

$$\rightarrow L_{MIP} = \sum_{i}^{N} E\|m_i \cdot (h_s^{Pred}(f_s(\tilde{u}_i, \theta_s))) - u_i)\|_1$$

$$\rightarrow L_{MPD} = -\sum_{i=1}^{N} m_i \cdot P_t^{Patch}(u_i, \theta_t) log(P_s^{Patch}(\tilde{u}_i, \theta_s))$$

$$\rightarrow L_{ITD} = -\sum_{i=1}^{N} m_i \cdot P_t^{[CLS]}(v_i, \theta_t) log(P_s^{[CLS]}(\tilde{u}_i, \theta_s))$$

$$L_{total} = L_{MIP} + \lambda_{MPD} L_{MPD} + \lambda_{ITD} L_{ITD}$$

Step 2 :

Fine-Tune Student Network on segmentation task.
The metric of evaluation is Dice Similarity Coefficients.

$P_t^{[CLS]}$ $P_s^{[CLS]}$ *: patch token distributions for teacher and student networks*

f_s *: student visual tokenizer, ViT*

m_i *: masked portions*

Validation, Results

The backbone is pre-trained on CT train set.
For validation, it is fine-tuned on organs segmentation tasks and tested both on CT test set and MRI dataset. 5 Cross-Validation is used for MRI.

- Organs for MRI : abdominal organs at risk for pancreatic cancer radiation treatment.
- Organs for CT : all organs. 21 images used for training. Others for test.

Segmentation accuracy of SMIT is compared to other state of the art SSL methods :

	DINO	iBOT	SSIM	SMIT
CT - AVG	0.826	0.833	0.830	0.848
MRI - AVG	0.835	0.848	0.842	0.875
CT - STO	0.891	0.900	0.898	0.903
MRI - SB	0.729	0.744	0.759	0.775

Table.1 : CT and MRI segmentation accuracy comparisons to SSL methods with DSC metric. AVG for average, SB for Smal Bowel, STO for STOmach

SMIT produced more accurate segmentations than other methods for most organs. In particular for Small Bowel a difficult organ to segment due to the presence of closely packed bowel loops and for stomach whose appearance and size are highly variable.

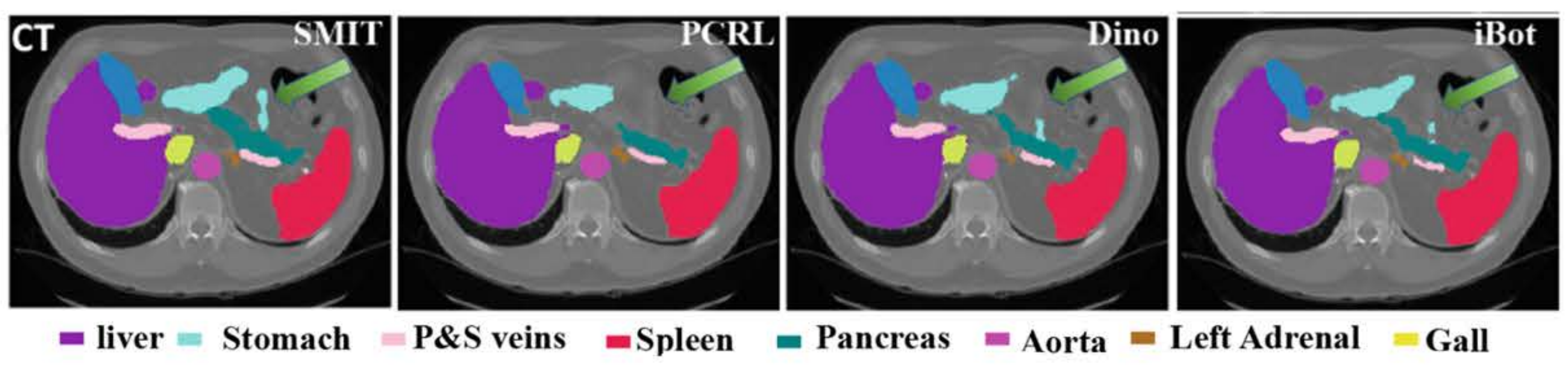

Fig 3. : Segmentation performance of different methods on CT abdomen organs.

In addition, an analysis of the pretexts tasks is carried out to show the importance and effectiveness of MIP for SMIT.
Regarding the strengths of the model, here are 3 additional points :

- On the BTCV dataset, SMIT is more accurate than all other SSL methods, regardless of the size of the dataset used for fine-tuning.
- It outperformed iBot, which uses MPD and ITD, which indicates effectiveness of MIP task for SSL.
- SMIT has very good segmentation accuracy, even with a large mask ratio.

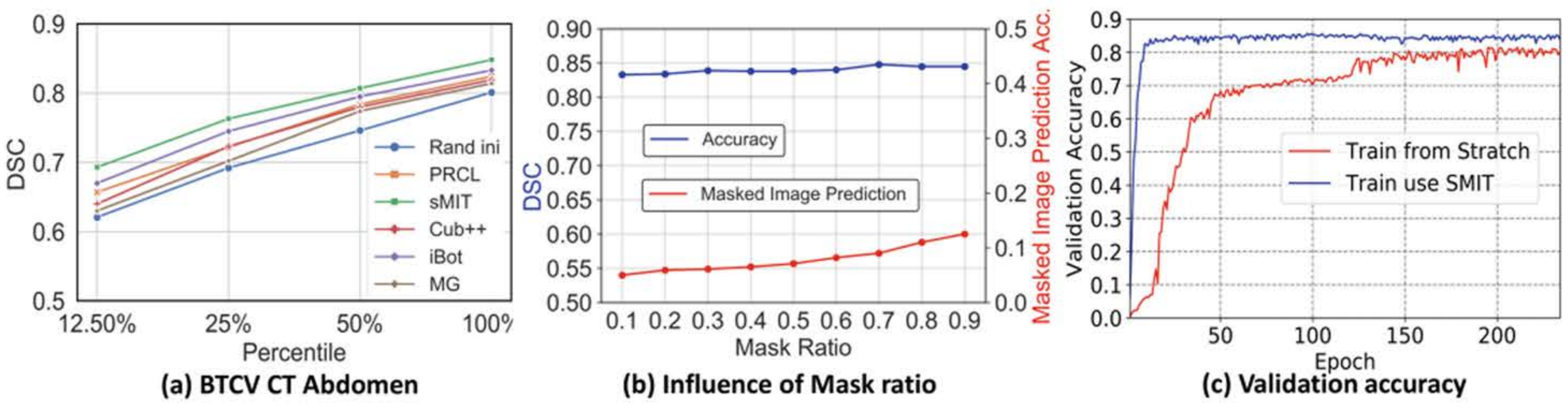

Fig.3 : (a) Impact of SSL task on fine-tuning sizes, (b) impact of mask ratio on masked image prediction and segmentation accuracy, (c) training convergence.

Conclusion

SMIT is a new Self-Supervised Learning method that combines the advantages of different methods: self-distilled learning, ViT architecture and especially masked dense pixel prediction, to allow a backbone to learn relevant visual representations. The paper showed the effectiveness of the method compared to the state of the art in pre-training for segmentation on CT and MRI datasets.

Bibliography

[1] J. Jiang, N. Tyagi, K. Tringale, C. Crane, H. Veeraraghavan : Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (SMIT) (2022)

[2] Y., Zhang, J., Shen, C., Xia, Y.: COTR: efficiently bridging CNN and transformer for 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention (2021)

[3] Zhou, J., et al.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022)