RESEARCH ARTICLE

Self-Supervised Vision Transformers for Medical Image Segmentation with Limited Annotations

Priya Patel, Xiaofeng Liu, Thomas Müller

Published: 2026-05-01 | FAIDS Vol. 1, No. 1 (2026)

Abstract: Annotating medical images for segmentation is expensive and requires domain expertise. We propose MedSSL-ViT, a self-supervised pre-training framework for Vision Transformers (ViT) tailored to medical imaging. MedSSL-ViT combines masked image modeling with anatomical-aware contrastive learning, leveraging the structured nature of medical images. Pre-trained on 850K unlabeled chest X-rays and CT slices, the model achieves state-of-the-art segmentation performance on four downstream tasks using only 10% of annotations: lung segmentation (Dice: 97.2%), cardiac chamber segmentation (Dice: 93.5%), liver tumor segmentation (Dice: 78.8%), and retinal vessel segmentation (Dice: 82.1%). With just 1% labels, MedSSL-ViT still outperforms fully supervised baselines trained on 100% labels by 2-5% Dice score.

1. Introduction

Medical image segmentation is a critical step in clinical workflows, enabling quantitative analysis of organ volumes, tumor burden, and treatment response. Deep learning has achieved remarkable accuracy in segmentation tasks, but its data-hungry nature conflicts with the limited availability of expert-annotated medical images. A radiologist may spend 30-60 minutes annotating a single 3D CT volume, making large-scale annotation prohibitively expensive.

2. Method

MedSSL-ViT pre-training consists of two complementary self-supervised objectives. The first is masked image modeling (MIM), where 75% of image patches are randomly masked and the model learns to reconstruct the original pixel values. The second is anatomical-aware contrastive learning (ACL), where augmented views of the same anatomical region are pulled together in the embedding space while views of different regions are pushed apart.

Table 1. Pre-training dataset composition

Dataset	Modality	Images	Resolution
CheXpert	Chest X-ray	224,316	320×320
MIMIC-CXR	Chest X-ray	377,110	320×320
DeepLesion	CT	186,018	512×512

Dataset	Modality	Images	Resolution
AMOS22	CT	62,880	512×512
Total	—	**850,324**	—

3. Results

We evaluated MedSSL-ViT on four diverse downstream segmentation tasks, using 1%, 5%, 10%, and 100% of available annotations. The results demonstrate that self-supervised pre-training provides dramatic improvements in the low-annotation regime.

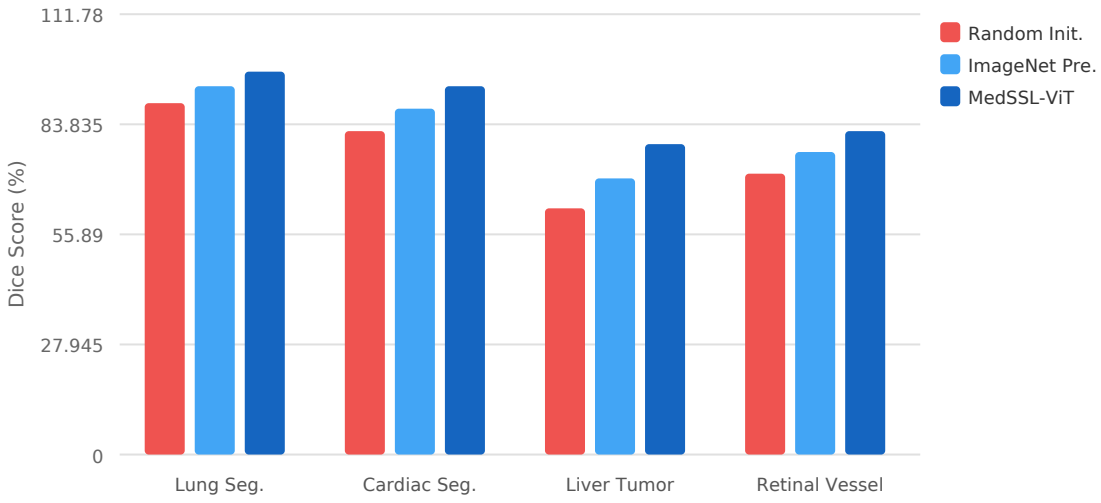

Figure 1. Dice score comparison across four segmentation tasks using 10% annotations

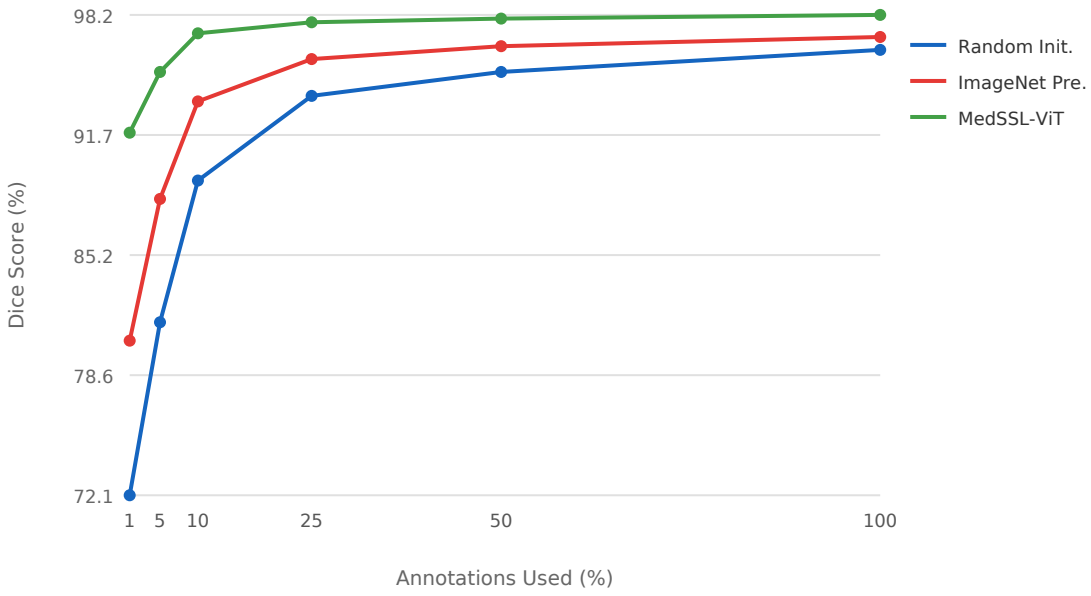

Figure 2. Label efficiency curve: Dice score vs. percentage of annotations used for lung segmentation task

4. Conclusions

MedSSL-ViT establishes a new paradigm for label-efficient medical image segmentation. By leveraging large volumes of unlabeled medical images through domain-specific self-supervised pre-training, the framework achieves state-of-the-art performance with only 10% of annotations, and surpasses fully supervised baselines with just 1% of labels. The anatomical-aware contrastive learning component is crucial for capturing the structured nature of medical images that generic SSL methods overlook.

References

[1] He, K.; Chen, X.; Xie, S. Masked Autoencoders Are Scalable Vision Learners. CVPR 2022.

[2] Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020.

[3] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

[4] Isensee, F.; Jaeger, P. F.; Kohl, S. A. A. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nature Methods 2021, 18, 203-211.

[5] Tang, Y.; Yang, D.; Li, W. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. CVPR 2022.