Multi-scale Hierarchical Vision Transformer with Cascaded Attention Decoding for Medical Image Segmentation

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Transformers have shown great success in medical image segmentation. However, transformers may exhibit a limited generalization ability due to the underlying single-scale backbone network. In this paper, we address this issue by introducing a Multi-scale hiERarchical vIsion Transformer (MERIT) backbone network, which improves the generalizability of the model by performing self-attention at multiple scales. We also incorporate an attention-based decoder, namely Cascaded Attention Decoding (CASCADE), for further refinement of multi-stage features generated by MERIT. Finally, we present a simple, yet effective multi-stage feature mixing augmentation method for better model training. Our experiments on two widely used medical image segmentation benchmarks (i.e., Synapse Multi-organ segmentation, and ACDC Segmentation) demonstrate superior performance over state-of-the-art methods. Our MERIT and multi-stage feature mixing augmentation can be used with other downstream medical image and semantic segmentation tasks.

Keywords: Vision Transformer, Attention, Multi-scale, feature-mixing augmentation.