Swin-Unet: UNet-like Pure Transformer for Medical Image Segmentation

Tommaso Gattari

January 2023

1 Abstract

This study presents a paper that concerns the utilization and the potential of Swin Transformer-based to be applied in the vision domain, in particular in the medical domain comparing the results with the various convolutional neural network (CNN). The proposed method is a novel approach known as Swin-Unet, which is a pure Transformer-based U-shaped Encoder-Decoder architecture for medical image segmentation. The method incorporates skip-connections and employs a hierarchical Swin Transformer with shifted windows as the *encoder* and a symmetric Swin Transformer-based *decoder* with a patch expanding layer for up-sampling. Experimental evaluations on multi-organ datasets that the proposed method outperforms existing methods that utilize full convolution or a combination of transformer and convolution. The codes and trained models used in this study are publicly available on GitHub [1].

2 Introduction

For many years existing medical image segmentation methods mainly rely on Convolutional Neural Networks. The typical use of convolution networks is in classification tasks like Resnet where we input an image and we get some class labels as output for what is actually in that image. But in many visual tasks, especially medical image processing the output should include localization which is important for Semantic Segmentation but it is too expensive for medical image purposes. Now is possible to build a fully CNN U-shaped that solves this problem Fig. 1 [5]. U-net has achieved great success in a variety of medical image applications. Here there are many examples: Res-UNet, U-Net++, and UNet 3+. But even if these FCNN-based methods in medical segmentation have a strong ability of learning features, they cannot still fully meet the strict requirements of medical applications for segmentation accuracy. Since the intrinsic locality of convolution operation, it is difficult for CNN-based approaches to learn explicit global and long-range semantic information

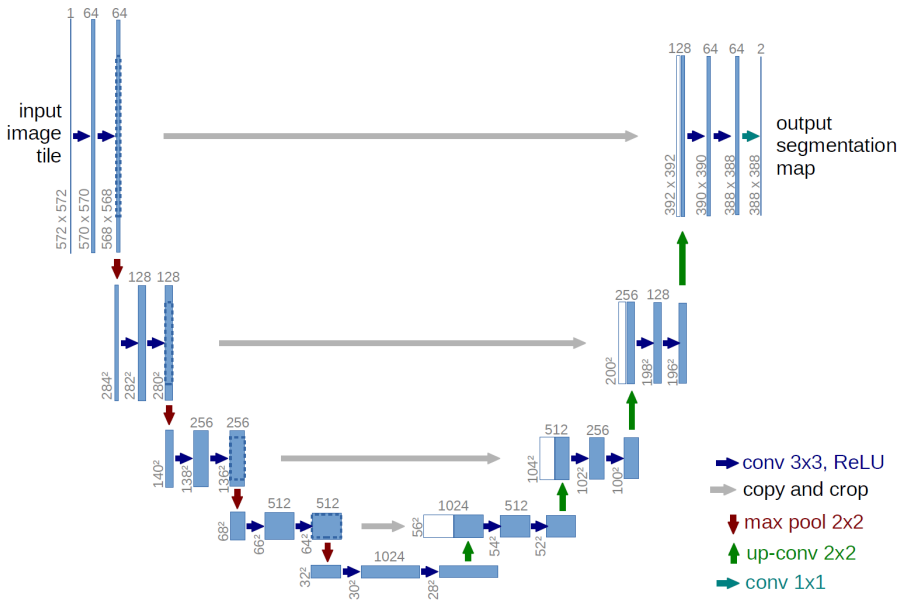


Figure 1: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

interaction. Researchers have been trying to apply the Transforme[4] architecture, which has been successful in natural language processing, to the field of computer vision. The Vision Transformer [3] (ViT) was proposed as a way to perform image recognition tasks. It achieved comparable performance to CNN-based methods by using 2D image patches with positional embeddings as inputs and pre-training on large datasets. A Hierarchical Swin Transformer [6] was developed, which achieved state-of-the-art performance on image classification, object detection, and semantic segmentation tasks. The success of ViT, DeiT, and Swin Transformer in image recognition tasks suggests that Transformer has the potential to be applied in the vision domain. For many years existing medical image segmentation methods mainly rely on Convolutional Neural Networks. The typical use of convolution networks is in classification tasks like Resnet where we input an image and we get some class labels as output for what is actually in that image. But in many visual tasks, especially medical image processing the output should include localization which is important for Semantic Segmentation but it is too expensive for medical image purposes. Now is possible to build a fully CNN U-shaped that solves this problem Fig. 1 [5]. U-net has achieved great success in a variety of medical image applications. Here there are many examples: Res-UNet, U-Net++, and UNet 3+.

But even if these FCNN-based methods in medical segmentation have a strong ability of learning features, they cannot still fully meet the strict requirements of medical applications for segmentation accuracy. Since the intrinsic locality of convolution operation, it is difficult for CNN-based approaches to learn explicit global and long-range semantic information interaction. The authors propose a new architecture called Swin-Unet, which leverages the power of the Transformer architecture for 2D medical image segmentation. The Swin-Unet is a pure Transformer-based U-shaped architecture that consists of an encoder, bottleneck, decoder, and skip connections. The encoder, bottleneck, and decoder are all built based on the Swin Transformer block, the input medical images are split into non-overlapping image patches and each patch is treated as a token and fed into the Transformer-based encoder to learn deep feature representations. The extracted context features are then up-sampled by the decoder with patch expanding layer, and fused with the multi-scale features from the encoder via skip connections, so as to restore the spatial resolution of the feature maps and further perform segmentation prediction. The authors show that the proposed method has excellent segmentation accuracy and robust generalization ability through experiments on multi-organ and cardiac segmentation datasets. The authors contribute to the field by building a symmetric Encoder-Decoder architecture with skip connections, developing a patch expanding layer to achieve up-sampling and feature dimension increase, and finding that skip connection is also effective for Transformer.

3 Method

3.1 Architecture overview

The authors propose a new architecture called Swin-Unet, which consists of an encoder, bottleneck, decoder, and skip connections. The basic unit of Swin-Unet is the Swin Transformer block, which was previously introduced in [6]. To generate sequence embeddings, the medical images are split into non-overlapping patches with a patch size of $4x4$, and each patch is transformed through a linear embedding layer to an arbitrary dimension represented as C. The transformed patch tokens then pass through several Swin Transformer blocks and patch merging layers to generate the hierarchical feature representations, while the patch merging layer is responsible for down-sampling and increasing dimension, and the Swin Transformer block is responsible for feature representation learning. The decoder is designed to be symmetric and transformer-based and is composed of a Swin Transformer block and patch-expanding layer. The extracted context features are fused with multi-scale features from the encoder via skip connections to complement the loss of spatial information caused by down-sampling. The patch expanding layer reshapes feature maps of adjacent dimensions into a large feature map with $2x$ up-sampling of the resolution, and the last patch expanding layer is used to perform $4x$ up-sampling to restore the resolution of the feature maps to the input resolution. A linear projection

layer is applied to these up-sampled features to output the pixel-level segmentation predictions. Different from the conventional multi-head self-attention

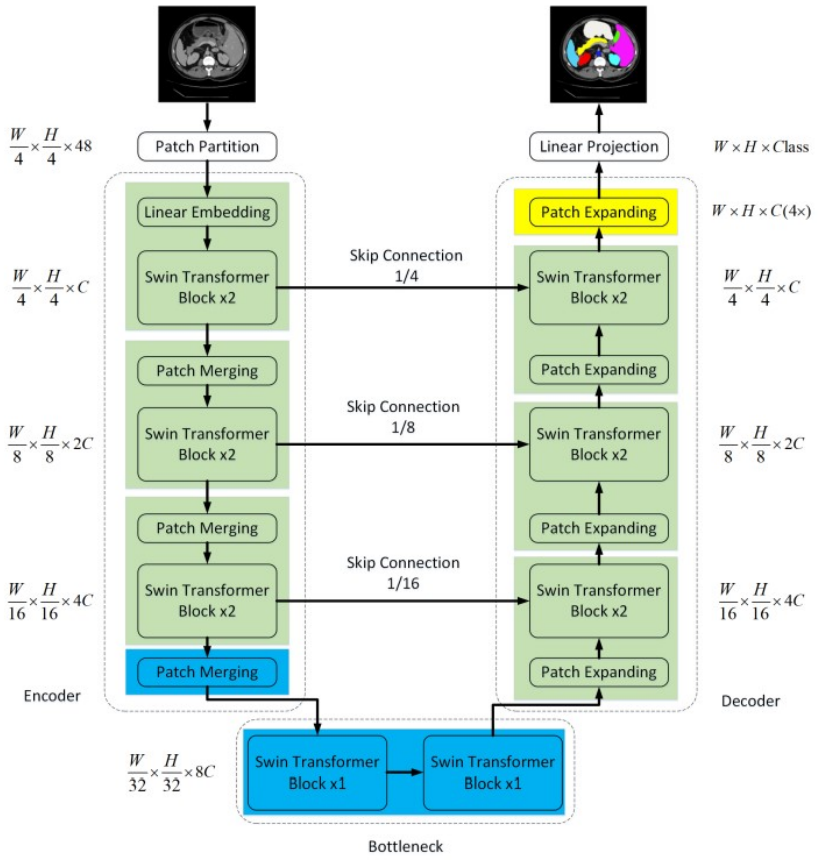


Figure 2: The architecture of Swin-Unet, is composed of an encoder, bottleneck, decoder, and skip connections. Encoder, bottleneck, and decoder are all constructed based on Swin transformer block

(MSA) module, the Swin transformer block is constructed based on shifted windows. In Figure. 2, two consecutive Swin transformer blocks are presented. Each Swin transformer block is composed of LayerNorm (LN) layer, multi-head self-attention module, residual connection, and 2-layer MLP with GELU nonlinearity. The window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-head self-attention (SW-MSA) modules are applied in the two successive transformer blocks, respectively.

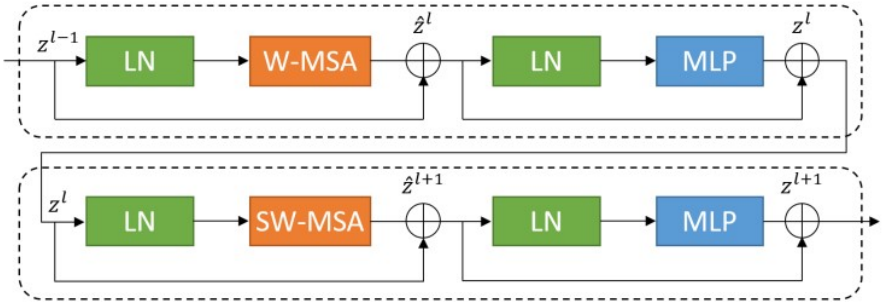


Figure 3: Swin transformer block

3.2 Encoder

The encoder in the model takes tokenized inputs with the resolution of $\frac{H}{4}x\frac{W}{4}$ and feeds them into two consecutive Swin Transformer blocks for representation learning. The patch merging layer then reduces the number of tokens by $2x$ (downsampling) and increases the feature dimension to $2x$ the original dimension. This process is repeated three times in the encoder.

3.3 Patch merging layer

The patch merging layer takes the input patches and divides them into 4 parts, and then concatenates them together. This reduces the feature resolution by $2x$. The concatenation operation also increases the feature dimension by $4x$, so a linear layer is applied to the concatenated features to unify the feature dimension to $2x$ the original dimension.

3.4 Bottleneck

The bottleneck in the model uses only two successive Swin Transformer blocks to learn deep feature representation, as Transformer is too deep to be converged. The feature dimension and resolution are kept unchanged in the bottleneck.

3.5 Decoder

The decoder in the model is symmetric to the encoder and is also built based on Swin Transformer blocks. Unlike the patch merging layer used in the encoder, the decoder uses the patch expanding layer to up-sample the extracted deep features. The patch expanding layer reshapes the feature maps of adjacent dimensions into a higher resolution feature map ($2x$ up-sampling) and reduces the feature dimension to half of the original dimension.

3.6 Patch expanding layer

The first patch expanding layer in the decoder starts by applying a linear layer on the input features $\frac{W}{32}x\frac{H}{32}x8C$ to increase the feature dimension to $2x$ the original dimension $\frac{W}{32}x\frac{H}{32}x16C$. Then, it uses a rearranged operation to expand the resolution of the input features to $2x$ the input resolution and reduce the feature dimension to a quarter of the input dimension $\frac{W}{32}x\frac{H}{32}x8C \rightarrow \frac{W}{32}x\frac{H}{32}x4C$. This process is used to perform up-sampling.

3.7 Skip connection

The model uses skip connections similar to U-Net, to fuse the multi-scale features from the encoder with the up-sampled features. The shallow features and the deep features are concatenated together to reduce the loss of spatial information caused by downsampling. A linear layer is applied on the concatenated features, to maintain the dimension of the concatenated features the same as the dimension of the upsampled features.

4 Experiments

4.1 Datasets

Synapse[2] is a multi-organ segmentation dataset and it is divided into three different folders: training, testing, and labels the type of these folders is nii.gz that is an open file format[1] commonly used to store brain imaging data obtained using Magnetic Resonance Imaging methods. In order to unzip files inside these folders there is a script "nii2gz" that creates a folder for each of them uploading .png images. After this, the images are converted to NumPy format, clip the images within [-125, 275], normalize each 3D image to [0, 1], and extract 2D slices from 3D volume for training cases while keeping the 3D volume in h5 format for testing cases. After this preprocessing process, we have the results in Fig. 4 Fig. 4 represents as follows:

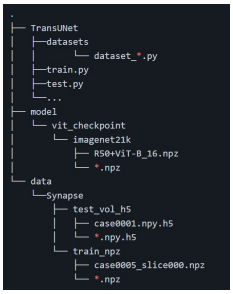


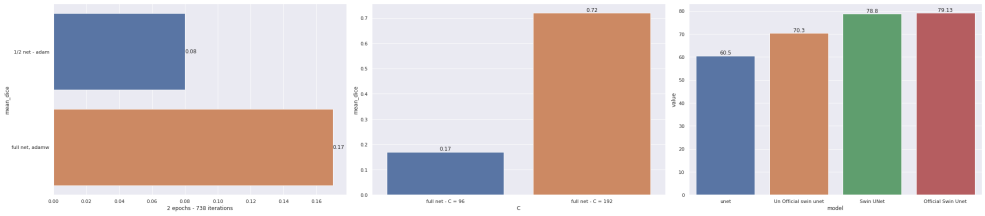
Figure 4: The directory structure of the whole project

4.2 Experiment results on Synapse dataset

In order to compare different results it has been used the mean dice value that is a measure of similarity between two sets, typically used in the field of image segmentation.

$(\frac{2x|X \cap Y|}{|X|+|Y|})$

To compute Dice, you would first need to determine the intersection of the two sets, and then divide that number by the sum of the sizes of the sets. The result will be a number between 0 and 1, with a value of 1 indicating that the sets are identical and a value of 0 indicating that there is no overlap.



4.3 Conclusion

The paper introduces a new model called Swin-Unet, which is a pure transformer-based U-shaped encoder-decoder for medical image segmentation. The model leverages the power of the Transformer by using the Swin Transformer block as the basic unit for feature representation and long-range semantic information interactive learning. The model was tested on multi-organ and cardiac segmentation tasks and it showed excellent performance and generalization ability.

References

[1] The codes for the work "swin-unet: Unet-like pure transformer for medical image segmentation.

[2] *dataset Synapse.*

[3] Alexey Dosovitskiy [..]. "an image is worth 16x16 words: transformers for image recognition at scale". 2021.

[4] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. "attention is all you need". 2017.

[5] P.Fischer O. Ronneberger and T. Brox. *"U-net: Convolutional networks for biomedical image segmentation.* MICCAI, 2015.

[6] Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Baining Guo Microsoft Research Asia Ze Liu, Yutong Lin. "swin transformer: Hierarchical vision transformer using shifted windows". 2021.