A review on image segmentation techniques for biomedical images

Shourya Pratap Singh

Introduction

The advent of convolutional neural networks made it possible for neural nets to handle high-dimensional data. They can work well with image recognition and pattern identification tasks. CNNs are made of convolutional layers that use a technique known as a sliding kernel to extract important features from images, pooling layers that reduce the resolution of these feature maps to achieve translational and deformational invariance, and non-linear layers that introduce non-linearities into the model to capture complex patterns. They are further applied for object detection, to localise each object in an image and categorise it from the input image. [7][10]

Building on top of CNNs and object detection, is object segmentation, which clearly demarcates the object boundaries and classifies every unit into an object class or background. It can be mathematically formulated to be a functional mapping of an input space onto a label space, which then allows for better interpretation tasks. [7][8]

Image segmentation is the specific task of associating pixels in an image with their respective object class labels. Segmentation algorithms are developed to detect objects within images and to obtain pixel level understanding of the images; therefore, the representation of an image into simple and easy forms increases the effectiveness of
pattern recognition. This is extended to 3D image segmentation and video segmentation tasks, wherein voxel and point clouds for the former, and frames for the latter are individually annotated to obtain a deeper understanding of each instance and to classify into different objects. Image segmentation is basically a classification problem of semantic labels. [7][8][10]

The scope of our discussion extends to a very specific subset of image segmentation, wherein we study the various biomedical image segmentation techniques which have gained prominence recently. Computer aided medical image analysis is fast becoming significantly important in helping medical practitioners support their decisions through information obtained from machine learning models. Biomedical image segmentation is also important since it helps overcome limitations of medical images such as noise and instrument errors. This enables medical decision making to become increasingly accurate with the aid of technology. [6][10] We will also try to compare the results achieved by these various algorithms by testing them against the baseline established by the Medical Segmentation Decathlon data [11].

Medical Image preprocessing

Before we delve into the technicalities of the machine learning approaches, it will serve us well to understand the kind of data we would be dealing with and what would be the kind of the key preprocessing steps recommended for medical images before they are fed to the models. The first part is to remove artifacts - eliminating the unwanted and inconsistent distortions and signals which would otherwise interfere with the learning process. Furthermore, in biomedical image processing we need to deal with volumetric representations of the organs, however while sampling several slices in the brain we ought to take into consideration that there would exist

differences between each slice, which would be even more significant when considering the first and last slides. To work out with such differences, we employ slice timing correction (STC) post normalisation of images. Various other preprocessing techniques may also be applied depending on the kind of task at hand, however, these are predominantly widespread for preprocessing.

Convolutional Neural Network approach

ConvNets and deep ConvNets have been successfully developed to solve classification and segmentation problems over the years, leading to a number of studies of deep learning based medical image analysis over the years. We can see the statistics of CNN-based medical image processing papers retrieved from PubMed in Fig. 1

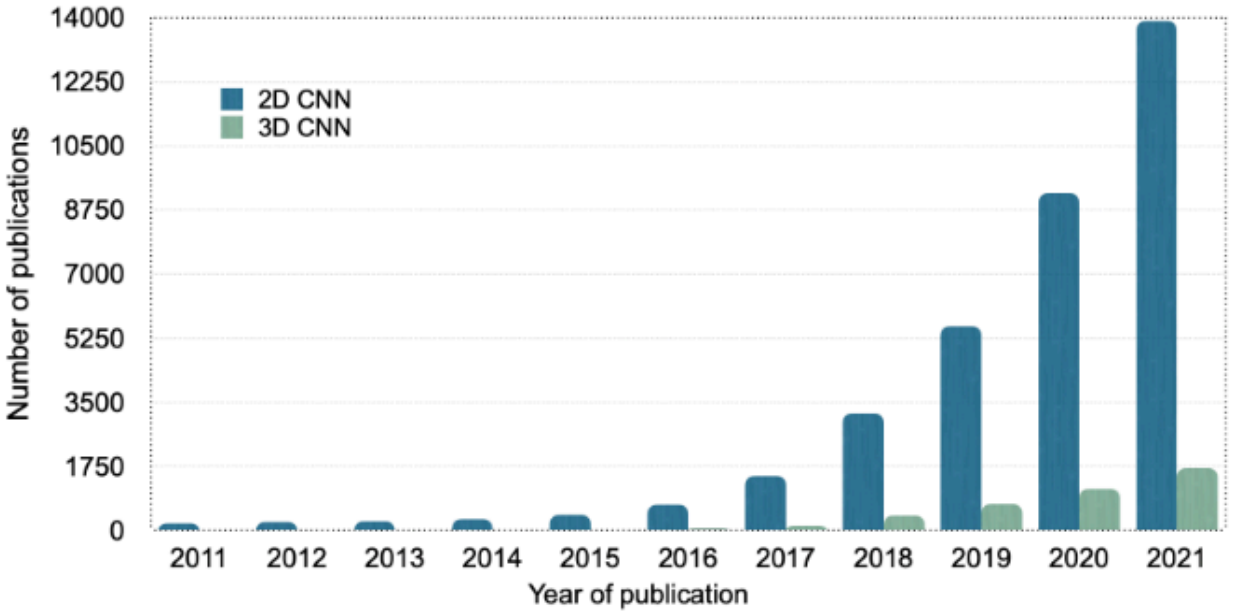

Fig. 1 - a comparison of the number of studies done using 2D and 3D CNNs for segmentation

Segmentation is primarily used to detect abnormalities and estimate the true extent of the organ or lesion, and is performed on 2 dimensional as well as 3 dimensional scans of the organs.
When performing 3 dimensional segmentation utilising CNNs, a set of kernels learn various characteristics and their values are updated based on the value of the cost function. The convoluted output learnt from each kernel is a feature map that gets passed onto the next layers. When we have a set of K kernel, the feature map in the nth layer can be mathematically represent as follows:

$$X_k^n = \sigma(w_k^n * X^{n-1} + b_k^n),\ K=\{1,2,3\ldots,K\}$$

where w_k^n is the k^{th} kernel in the n^{th} layer and b_k^n is the k^{th} bias in the n^{th} layer. X_k^n is the feature map we obtain in the n^{th} layer. [9] We have talked about convolution and pooling layers before, but I wish to define their use case here. A convolution layer uses small multi-dimensional kernels to extract spatial features from image locations. The pooling layer is designed to pass relevant features to the subsequent layers by downsampling the feature space and expanding the receptive field for multi-scale feature extraction, and providing translational invariance to small shifts. When used in segmentation, the convolutional and pooling layers determine the spatial relationship of pixels throughout the network, without using a fully connected layer. Additionally to

overcome the effect of the pooling layers, adequate upsampling layers are added. In this manner, the probability of each pixel is estimated and a segmentation mask is created.
Fabian et al [1][13] deploy a U-Net architecture which consists of a 2D U-Net, a 3D U-Net and a U-Net Cascade. The 2D and 3D U-Nets generate full resolution segmentations, the cascade generates low resolution segmentations which are subsequently refined by itself. [1][14] A U-Net architecture has been envisioned below [14]:

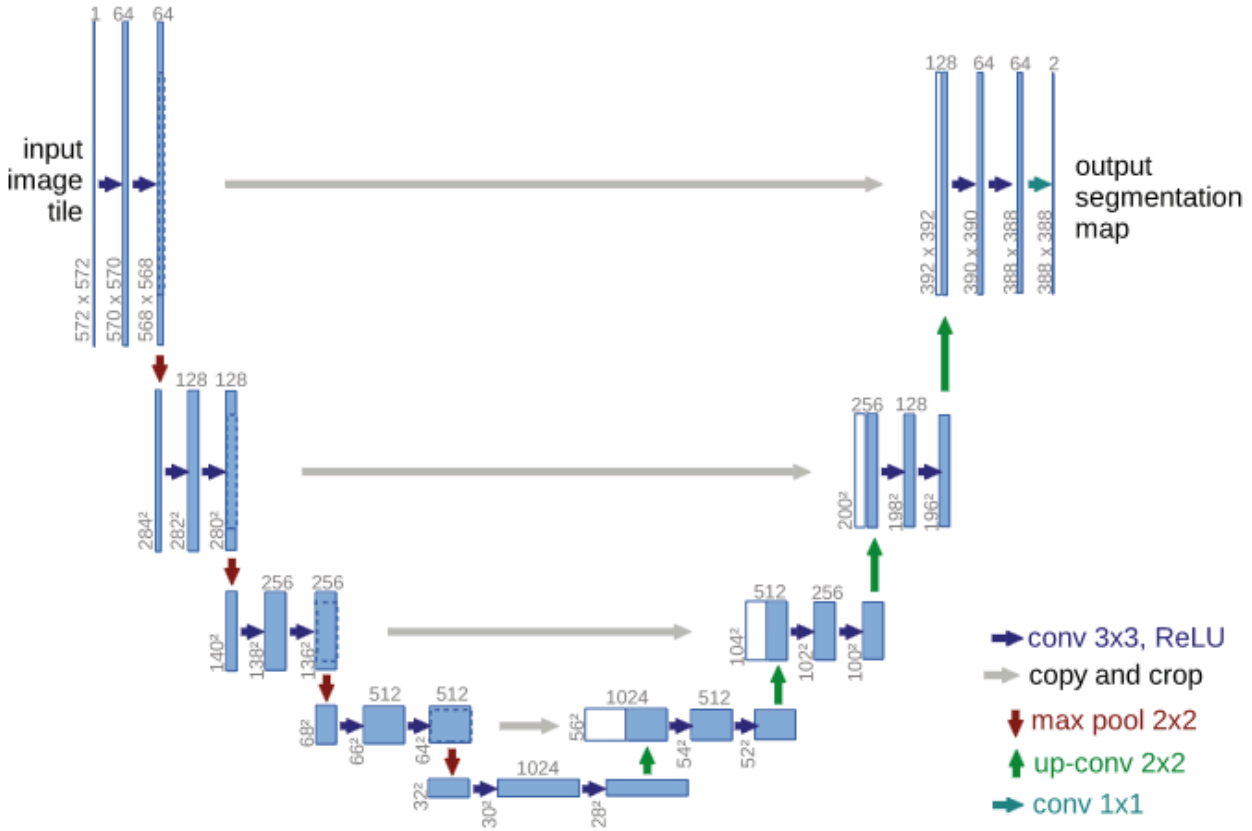

Fig. 2: U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

The U-Net is composed of a contracting path and of an expansive path, wherein the contracting path repeatedly performs 3x3 convolutions followed by rectified linear unit and 2x2 max pooling operations with stride 2 for downsampling [14]. Everytime there is downsampling, we double the number of feature channels, concatenate with the corresponding feature maps and perform 2 3x3 convolutions followed by a ReLU. At the end, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes, culminating with 23 convolutional layers [14].
Within nnUNet, a 2D U-Net is unable to aggregate information along z-axis and take it into consideration, however if the 3D dataset is anisotropic, it can be used to fulfil the shortcomings encountered by conventional 3D segmentation methods. A 3D U-Net trains using the image patches. This may limit the field of view and impede in collecting sufficient contextual data correctly, however few of these shortcomings are overcome by using a 3D cascade U-Net. First the 3D U-Net is trained on downsampled images and the results are upsampled to the original voxel spacing and passed as additional one hot encoded input to the next 3D U-Net, which is

trained on patches at full resolution. This allows nnUNet to handle a wide variety of target structures and image properties and can be utilised in a diverse range of tasks. [13]
When tested on our baseline dataset, it produces a dice score above 80 for heart, hippocampus scans and provide greater than 70 dice scores on (parts of/fully) of liver, prostate, pancreas and colon of the medical segmentation decathlon [1].
The DeepMedic team [6][15] worked by using a deep CNN to segment and detect brain metastases on MRI images, using multi scale CNN with 2 convolutional pathways and 11 convolutional layers. It consisted of a backbone, atrous convolution and classification layers. The schematic of this can be found here [6]:

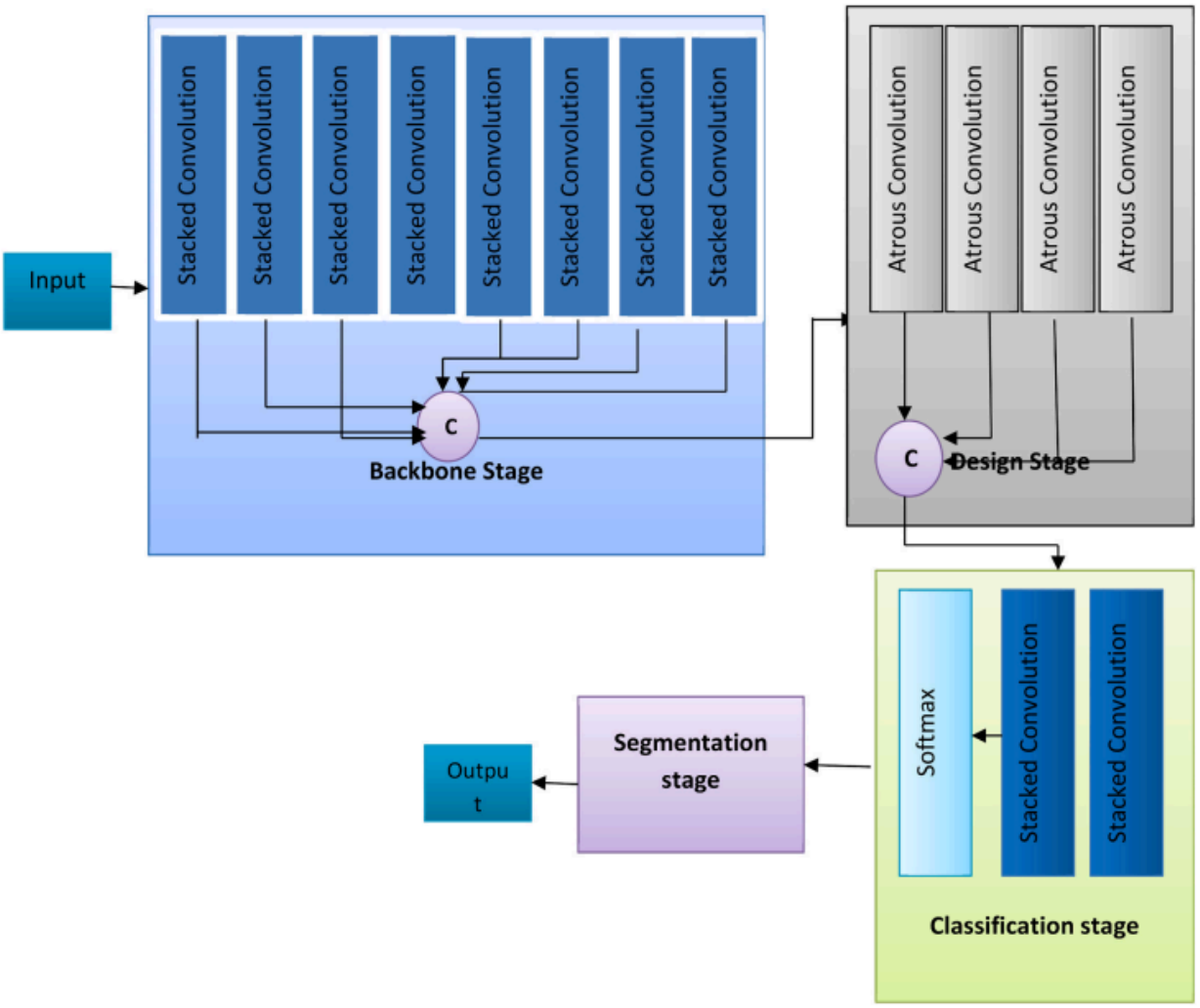

Fig 3: DenseAFPNet

This model could also not address the vanishing gradient issue, and there had to be the use of an additional 3D learning and interference approach to mitigate this concern, such as using cascaded networks in case of nnUNet or through attention pooling etc. When it is tested on the brain and liver datasets of the Medical Segmentation Decathlon, it reaches and average DICE score of 0.91 [6][14].

Transformer Based Approaches

Chen et al [4] and Perera et al [3] were the initial pioneers in developing a transformer based architecture to segment 3D scans of various organs. TransUNet incorporates the transformer as an encoder, performing image sequentialisation and thereafter patch embedding. It then uses a

block to estimate coarse candidates, wherein they treat the task as a mask classification problem, and aim to assign the corresponding organ label to each region when provided with N organ queries, and K segmentation classes. Then a transformer as a decoder is used to refine organ queries and enhance coarse prediction, and thereafter perform coarse-to-fine attention refinement. When the TransUNet is tested on the Medical Segmentation Decathlon HepaticVessel task, it applies a 5 fold cross validation to evaluate the methods on the dataset and achieves an average dice score of greater than 67, more than 1.5 points greater than nnUNet which was the best performing algorithm in MSD [1][4].

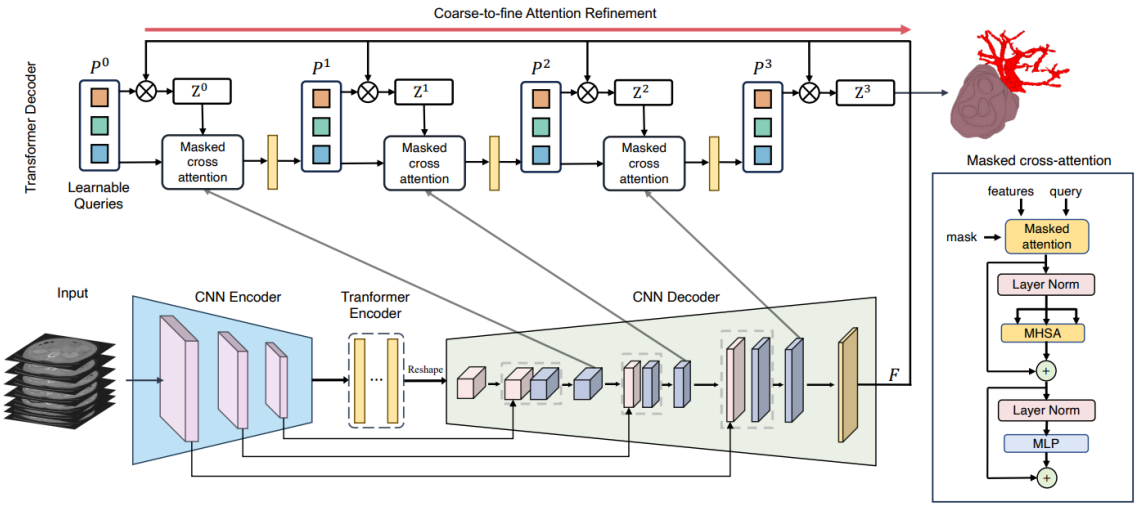

Fig. 4: Overview of 3D TransUNet.

On the other hand, SegFormer3D [3] is a volumetric hierarchical vision transformer that extends to 3D medical image segmentation tasks. Unlike normal ViTs it encodes feature maps at different scale of the input volumes following the Pyramid Vision Transformer, enabling it to capture a variety of features from the input scans, with varying levels of details. The self-attention used with SegFormer compresses the embedded sequence to a fixed ratio, reducing complexity without sacrificing performance. They also work with overlapping patch embedding module to preserve local continuity of input voxels, preventing accuracy loss by using a positional free encoding. The finally generated segment mask is through an all - multilayer perceptron decoder, which when tested on the liver, spleen, stomach and kidney scans performed exceptionally to provide an average dice score between 81 and 96, and was on an average 5 percentage points better than the previously discussed TransUNet [3][4].

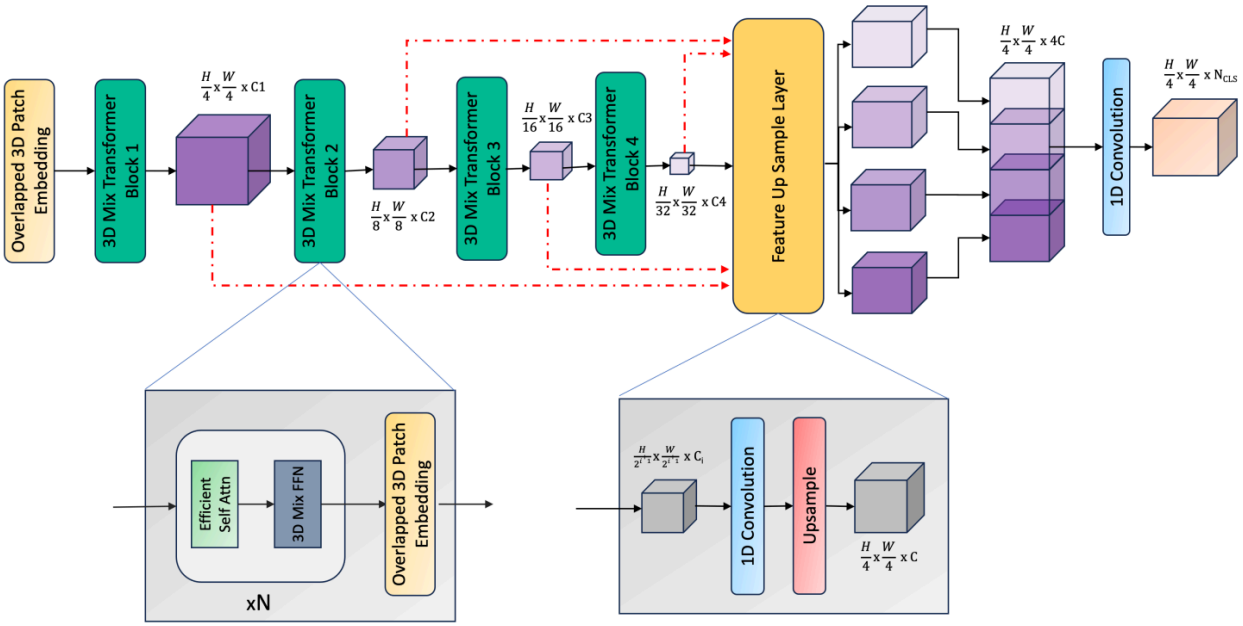

Fig. 5: Segformer3D Overview

For a given CT/MRI scan, SegFormer3D does inference in a sliding window manner, leveraging the aggregation of all patches, P they assign a probability vector to a voxel in a given position, followed by an argmax to obtain hard prediction.

The TransHRNet team [5] however, used a parallel transformer approach wherein they would deploy a CNN encoder block to obtain high resolution, medium resolution and low resolution features through the CNN based feature extraction module, which captured spatial and depth information, gradually expanding the receptive fields to contain rich information. This was followed by using a layered parallel transformer block (named EffTrans) for feature enhancement used to learn the long distance dependency in a global space and connecting the different resolution streams in parallel to repeatedly perform information interactions across the resolution streams, which helped it to learn a large number of features with effective information about them. This culminated in a decoder module which provided a high quality segmentation result with upsampling and convolutional operations.
The feature extraction was performed to extract local features, and mathematically the size of l-th features f_l is $D/2^l \times H/2^{l+1} \times W/2^{l+1} \times C$, C indicates the number of channels. This was furthered by using a feature enhancement layer through the efftrans decoder, helping it learn through a large number of feature information interactions. The image to sequence process in a transformer is the same as in NLP, using seq2seq patch embeddings as inputs, after collapsing the spatial and depth dimensions of the features in a layer into one dimension. This is followed by a feature fusion block which can capture different resolution features, using parallel multi-resolution Transformers and repeated multi-resolution fusion across parallel streams. Each stream has an equal number of EffTrans to learn long range dependency globally and repeatedly fuses multi-resolution features to exchange information across multi-resolution representations. We can view in Fig 6 a TransHRNet [5]:

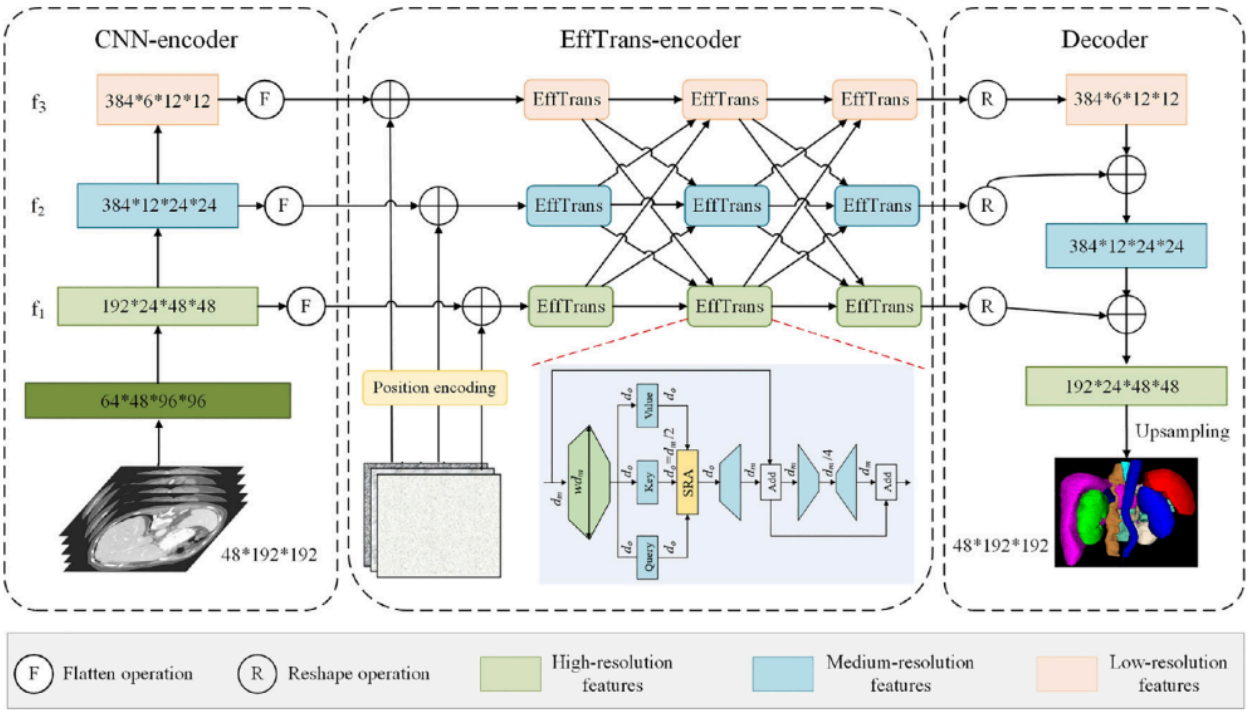

Fig 6: The overall architecture of the proposed TransHRNet

Here, I would like to explain what an EffTrans actually means. Based on feature fusion, to obtain long range dependency using standard transformers, it becomes complicated to handle high dimensional multi scale features in both time and space, due to its quadratic model complexity. For this reason, a parameter-efficient EffTrans is used with an attention based architecture that is to be scaled according to the width and depth. The EffTrans uses a DeLighT transformation to reduce the dimension of the input. It does so by first mapping a d_m dimensional space into a higher dimensional space and then reducing to a d_o vector using Group Linear Transformations to produce local representations deriving the input from a specific part over these two phases. Also, information exchanges occur between different groups in the group linear transformation using feature shuffling to generate global representations. It then adapts a spatial reduction attention layer to reduce the resource cost to learn high resolution feature maps. It receives a query Q, key K and value V as input and outputs a refined feature. The overall EffTrans layer is explained in Fig 7 [5]. The network culminates in a decoder part which reshapes the output sequence of each stream in the feature fusion block to the size of each scale, and then use progressive transpose convolution to upsample the features to the original 3D image space and obtain the final results through a 3D residual block. Skip connections integrate the encoder features with the decoder counterparts by concatenation for generating richer spatial details in the results. The overall training a binary cross entropy loss and dice loss between the prediction ground truth is used. The TransHRNet performs exceptionally on the spleen, liver and stomach scans achieving the best or second best dice scores, and an average dice score of more than 86 across scans of different organs. The overall average results according to average dice scores obtained by using SRA with different attention heads and reduction ratio settings are provided in Fig 8 [5].

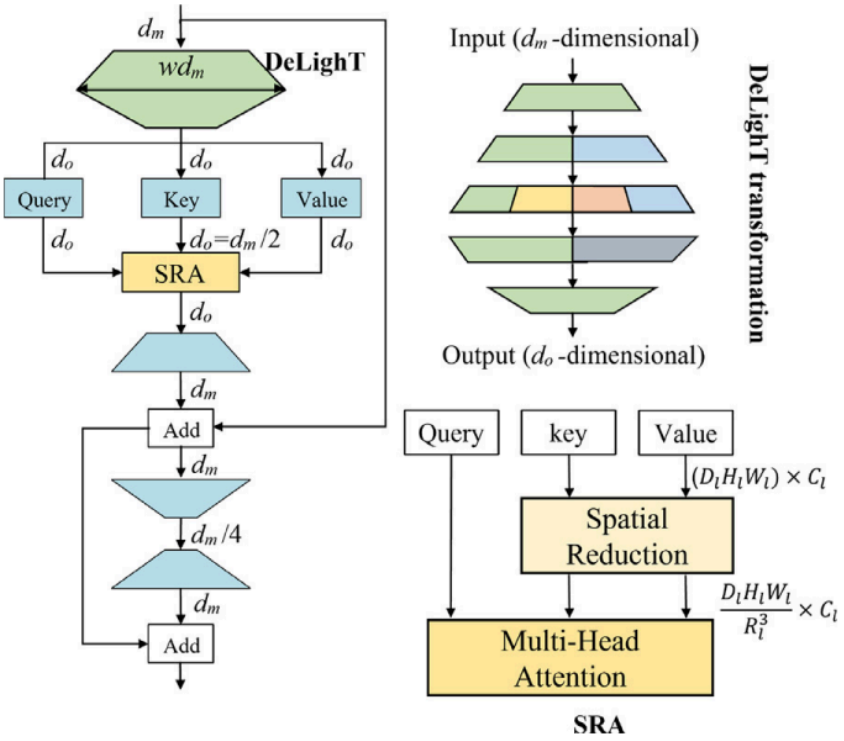

Fig 7: EffTrans Block

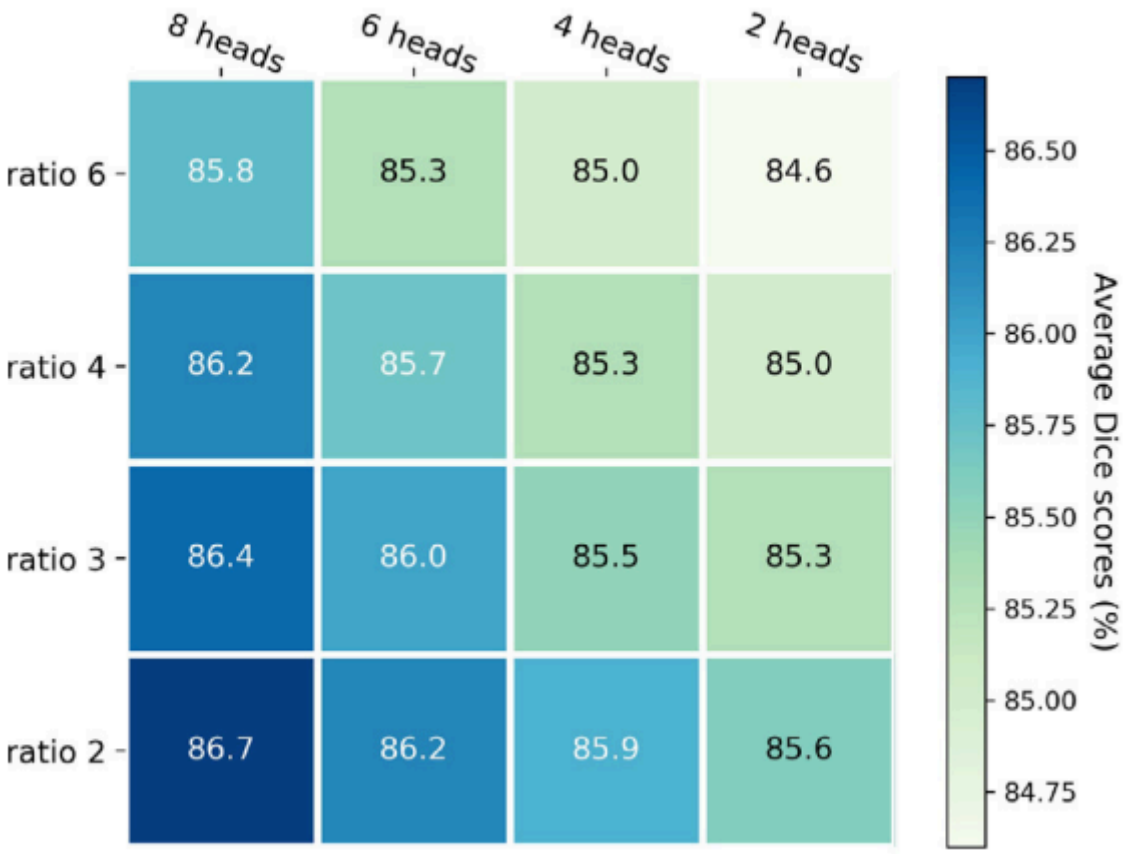

Fig 8: The average Dice scores of organ segmentation performance are obtained by using Spatial-Reduction Attention (SRA) with different attention heads and reduction ratio settings.

Conclusion

The various methods discussed here, although not encompassing all techniques currently being used, form a sufficient list of segmentation algorithms with their own pros and cons. The CNN based algorithms are more suited for tasks wherein the scan sizes are not large and the information obtained from an image is not inconsiderately large, otherwise they shall suffer from a vanishing gradient problem. Convolutional Neural Networks (CNNs) have been a dominant force in medical image segmentation. They excel at capturing spatial relationships within images through convolutional layers. Architectures like U-Net address challenges in medical image segmentation by effectively combining high-resolution features with contextual information through skip connections. However, CNNs can struggle with very large or complex datasets. Transformer-based approaches are a recent advancement that utilise transformers, known for their ability to model long-range dependencies. Architectures like TransUNet and SegFormer3D have shown promising results. Notably, TransHRNet leverages parallel processing of features at different resolutions, improving its segmentation accuracy. While Transformers offer strong performance, their computational cost can be high. Future work on these methods could focus on improving their efficiency and interpretability, making them even more viable for real-world medical applications. A future in AutoML wherein these models can be used to generate a pipeline which provides with the perfectly annotated segmentations, automating many hyperparameter tuning and model selection processes and improve generalizability to help create models which can be used across scans of different organs, allowing people without the technical expertise to also work with these models to help them in their medical decision making.

Papers referred

1. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation - Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
2. 3D reconstruction of individual anatomy from medical image data: Segmentation and geometry processing - Zachow, Stefan & Zilske, Michael & Hege, Hans-Christian. (2007). 3D Reconstruction of Individual Anatomy From Medical Image Data: Segmentation and Geometry Processing.
3. SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation - Perera, S., Navard, P., & Yilmaz, A. (2024). SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation. arXiv:2404.10156.
4. 3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers - Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M., Xing, L., Lu, L., Yuille, A. L., & Zhou, Y. (2023). 3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers. arXiv preprint arXiv:2310.07781.
5. 3D Medical image segmentation using parallel transformers - Yan, Q., Liu, S., Xu, S., Dong, C., Li, Z., Shi, J. Q., Zhang, Y., & Dai, D. (2023). 3D Medical image segmentation using parallel transformers. *Pattern Recognition, 138*, 109432. ISSN 0031-3203. https://doi.org/10.1016/j.patcog.2023.109432.
6. Reviewing 3D convolutional neural network approaches for medical image segmentation - Ilesanmi, A. E., Ilesanmi, T. O., & Ajayi, B. O. (2024). Reviewing 3D convolutional neural

network approaches for medical image segmentation. *Heliyon, 10*(6), e27398. ISSN 2405-8440. https://doi.org/10.1016/j.heliyon.2024.e27398.
7. A Comprehensive Review of Modern Object Segmentation Approaches - Wang, Y., Ahsan, U., Li, H., & Hagen, M. (2022). A Comprehensive Review of Modern Object Segmentation Approaches. *Foundations and Trends® in Computer Graphics and Vision, 13*(2–3), 111–283. ISSN 1572-2759. http://dx.doi.org/10.1561/0600000097.
8. A Comprehensive Review of Image Segmentation Techniques - Abdulateef, Salwa & Salman, Mohanad. (2021). A Comprehensive Review of Image Segmentation Techniques. Iraqi Journal for Electrical and Electronic Engineering. 17. 166-175. 10.37917/ijeee.17.2.18.
9. Medical Image Segmentation with 3D Convolutional Neural Networks: A Survey - Niyas, S., Pawan, S. J., Anand Kumar, M., & Rajan, J. (2022). Medical Image Segmentation with 3D Convolutional Neural Networks: A Survey. arXiv:2108.08467.
10. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). Image Segmentation Using Deep Learning: A Survey. *arXiv preprint arXiv:2001.05566*.
11. Amber L. Simpson et al. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv:1902.09063v1 [cs.CV].
12. Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020). 3D Deep Learning on Medical Images: A Review [arXiv:2004.00218]. arXiv preprint arXiv:2004.00218.
13. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., & Maier-Hein, K. H. (2018). nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation [arXiv:1809.10486]. arXiv preprint arXiv:1809.10486.
14. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234-241). Springer International Publishing. https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/
15. GitHub - deepmedic/deepmedic: Efficient Multi-Scale 3D Convolutional Neural Network for Segmentation of 3D Medical Scans