U-net Design Comparison on Medical Image Segmentation Tasks

Arthur Boschet[1] **Clément Detry**[2] **Frederic Gagne**[3]

[1,2,3]Université de Montréal

Abstract

Image segmentation of medical images is a crucial application in computer vision, as it can replace clinicians in performing time-consuming and critical tasks. One of the key architectures in this field is the U-net, which serves as the basis for many newer architectures such as the Swin-UNETR that utilizes Swin transformers blocks to achieve state-of-the-art performance. Other researchers have proposed various implementations using different convolutional blocks or alterations to the shortcuts path between the encoder and decoder. However, comparing the efficacy of different architectural choices can pose a challenge, as each research team typically has its own distinct preprocessing steps and pipeline optimizations to improve their results. Thus, the aim of this study is to compare proposed architectural modifications against a U-net baseline under a fixed set of hyper-parameters, with the goal of providing meaningful comparisons on the Medical Image Segmentation Decathlon datasets Simpson et al. (2019). After performing a 5-fold cross-validation analysis on the heart and lung datasets, we found that the most promising design modification was adding convolutional layers to the shortcut path, despite not reaching statistical significance. We then tested this modified model, along with the U-net baseline and Swin-UNETR, on all datasets and found that it had the best overall performance with an average rank of 1.59 across all tasks. Transformers-based models were not able to perform as well as convolution-based ones. It is assumed that preprocessing and data augmentation may play a crucial role in the performance presented in those papers, and further studies including these factors could help close the gap between our study and the literature.

1 Introduction

Image segmentation is a technique in computer vision that involves dividing an image into multiple regions or segments based on specific criteria. In medical imaging, image segmentation is crucial for analyzing MRI or CT scans to identify organs or track the size and location of tumors. However, the process of manually segmenting medical images is time-consuming and can be prone to error. That's why there has been growing interest in developing automatic image segmentation algorithms that can accurately and efficiently analyze medical images. By automating the segmentation process, clinicians can save time and resources, allowing them to focus on other aspects of patient care (Simpson et al., 2019).

The U-net architecture is widely used for medical image segmentation and has been the subject of many iterations proposed by the research community (Ronneberger et al., 2015) .However, comparing the results of different U-net implementations can be challenging since there are many variables that can impact performance and accuracy such as varying data processing strategies. To address this issue, the goal of this work is to implement several versions of the U-net architecture using a consistent framework. By controlling the size of the model and training hyper-parameters, we can more accurately compare the performance

MAIS 2023, April 28th 2023, Montréal

of different U-net models and evaluate the impact of different design choices. By analyzing the results of our experiments, we hope to gain insights into which U-net variations are most effective for different medical imaging tasks in the Medical segmentation decathlon datasets. Our study specifically focuses on exploring modifications at three different scales. Firstly, we investigate changes at the level of the convolutional blocks. As such, we compare two Resblock iterations and a ConvNext block to the original U-net implementation. Next, we experiment with different improvement for the shortcut connections. One approach involves using convolutional blocks between the encoder and decoder, while the second approach employs a self-attention mechanism with the downstream representation of the encoder as the query. Finally, we explore architectural variations with simplified versions of the decoder (Half-Unet) and an encoder based on the Swin-UNETR that utilizes transformers.

We have chosen to primarily focus our work on the heart and lung datasets due to their relatively manageable size and the distinct nature of the tasks involved. Additionally, we aim to train the best performing model on all MSD datasets and submit our result to the ongoing competition.

2 Litterature review

The U-net architecture is called "U-net" because of its U-shaped design. The network consists of a contracting path, which is used to capture the context of the image, and an expanding path, which is used to obtain a segmentation map of the image. The contracting path starts with a series of convolutional layers, which are used to extract features from the input image. The output of each convolutional layer is passed through a rectified linear unit (ReLU) activation function, which introduces non-linearity into the model. The number of filters in each convolutional layer is gradually increased, while the size of the image is reduced through pooling layers. This process is repeated several times to capture higher-level features of the image. At every steps, a shortcut paths connect with the corresponding dimension of the expending path (decoder) to be concatenated or added with encoded informations. The expanding path is used to reconstruct the segmentation map of the image. This path consists of a sequence of up-convolutions and valid convolutions. Additionally, the output of an up-convolutional layer in the expanding path is concatenated with the output of the corresponding layer in the contracting path. This process is repeated several times, with each up-convolutional layer followed by a convolutional layer. The number of filters in each convolutional layer is gradually decreased, while the size of the image is gradually increased (Ronneberger et al., 2015). Recent research in medical image segmentation still relies on the U architecture. One notable example is the Swin-UNETR, which takes inspiration from transformer-based computer vision models to improve long-range information retention. The Swin transformer blocks replace the convolutional blocks in the encoder of traditional U-nets. These transformer blocks utilize patch embedding before applying self-attention to a window of patches, followed by a shifted window self-attention mechanism. Layer normalization is applied between each step, and an MLP layer is used between each attention application. A convolutional block is used on the skip connection between the encoder and decoder, while the decoder itself continues to use convolutional blocks like the classic U-net. In addition, those convolutional blocks are modified to included a residual connection around two convolutions. The researchers evaluated their model on a brain tumor segmentation dataset and achieved state-of-the-art performance (Hatamizadeh et al., 2022).

Other researchers have also explored the use of transformers to improve performance on various segmentation tasks. One such model is the U-net Transformer, which uses self-attention between the convolutional layers and cross-attention on the skip connection,

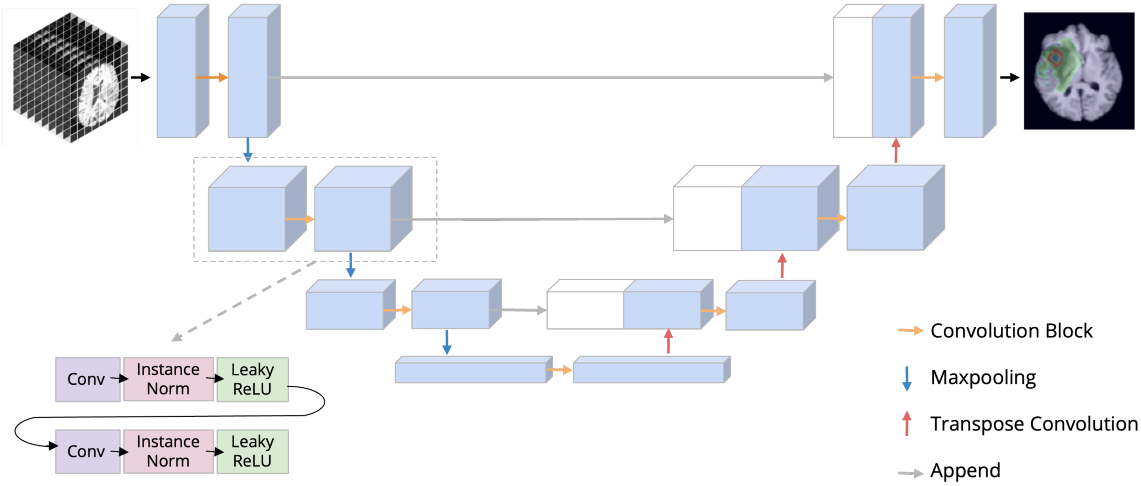


Figure 1: The U-net architecture is presented in this figure. The left side illustrates the contracting path of the encoder leading up to the bottleneck, while the right side displays the upsampling and convolution process of the decoder, along with the concatenation of the skip connections from the encoder

instead of replacing the convolutional blocks like the Swin-UNETR. This approach has shown improvements over the original U-net architecture in various segmentation tasks (Petit et al., 2021). The U-net architecture can also be utilized as a preprocessing tool for segmentation tasks, as demonstrated in the work on SNEMI3D by Kisuk Lee's team. In their study, a modified U-net model is employed to predict affinities between neighboring voxels. Unlike the original U-net, their model uses convolutional blocks consisting of a sequence of three convolutions with a residual connection between the last two (Lee et al., 2017). Also, the U-net model has been optimized for efficiency without sacrificing performance through recent research. One example is the Half-U-net, which proposes a simplified version of the encoder and decoder without compromising performance but with a significant increase in FLOPs. In this iteration, the decoder does not contain any convolution and only upsamples the shortcuts from the encoders before merging them using summation (Lu et al., 2022).

3 Methods

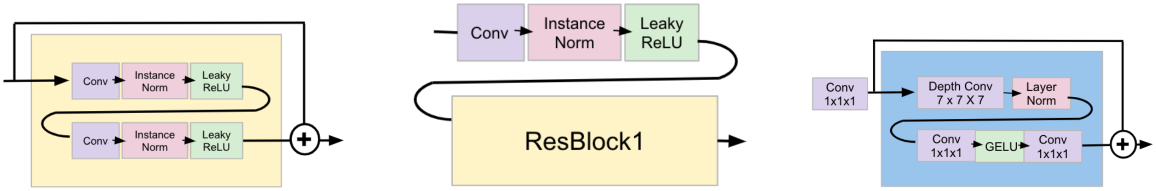
Two baseline models will be implemented for comparison with other experiments: Swin-UNETR, which represents the current state-of-the-art, and the original U-net. As all other experiments will involve only minor alterations relative to this baseline implementation, the original U-net will serve as a suitable comparison point. Our baseline implementation is based on the work done by Ronneberger and his team in 2015 (Ronneberger et al., 2015). We will modify the original architecture at three levels: convolution blocks, shortcut connections between the encoder and decoder, and architectural variations in the decoder.

The Medical Image Decathlon datasets comprise ten tasks that focus on different organs, such as the brain, pancreas, liver, and heart. The tasks vary across these organs, with some identifying different types of tumors, while others focus on partitioning organs. The data consists of 3D or 4D CT scans or MRIs, which are computationally heavy and challenging to work with without appropriate hardware. The fourth dimension corresponds to different types of scans. For this reason, we first trained and evaluated our experiments on the heart and lung datasets via 5-fold cross-validation over 100 epochs. Only the highest-performing model, the U-net and Swin-UNETR baselines will be trained and evaluated on all datasets in the official competition. We chose the heart and lung datasets as their sizes are manageable, and their tasks vary in difficulty. The heart dataset involves partitioning the heart, while the lung dataset focuses on detecting small tumors.

For all experiments, we perform a hyperparameter sweep using the Bayesian search algorithm to tune the learning rate. We utilize the Adam optimizer for all implementations. Each experiment on the lung and heart datasets is trained on 5-fold cross-validation to ensure robustness in our results. We keep hyperparameters, such as the depth of neural networks, constant across all experiments. Furthermore, we do not apply any preprocessing to the original images, besides normalization and resizing. The loss function used for training these models is a weighted sum of the cross-entropy and dice losses.

To identify the optimal number of epochs for the best model and two baselines across all datasets after the best model is selected, we evaluate their performance on a validation set using a patience parameter of 30 and a maximum of 1000 epochs. We select the number of epochs that yield the best results to train each model on the entire dataset.

In this study, we explore three types of convolutional blocks. Two of these blocks are conventional residual blocks. The first block, ResBlock1, simply adds a residual connection around the two basic convolutional blocks of the baseline U-net model. The second block, Resblock2, is based on the U-net architecture used by Kisuk Lee's team in the SNEMI3D Connectomics Challenge. It adds a convolutional block before the same structure as the first residual block, resulting in a total of three convolutional blocks per residual block (Lee et al., 2017). Finally, we also investigate a third alternative derived from the well-known ConvNext architecture (Liu et al., 2022). These blocks are described in figure 2.



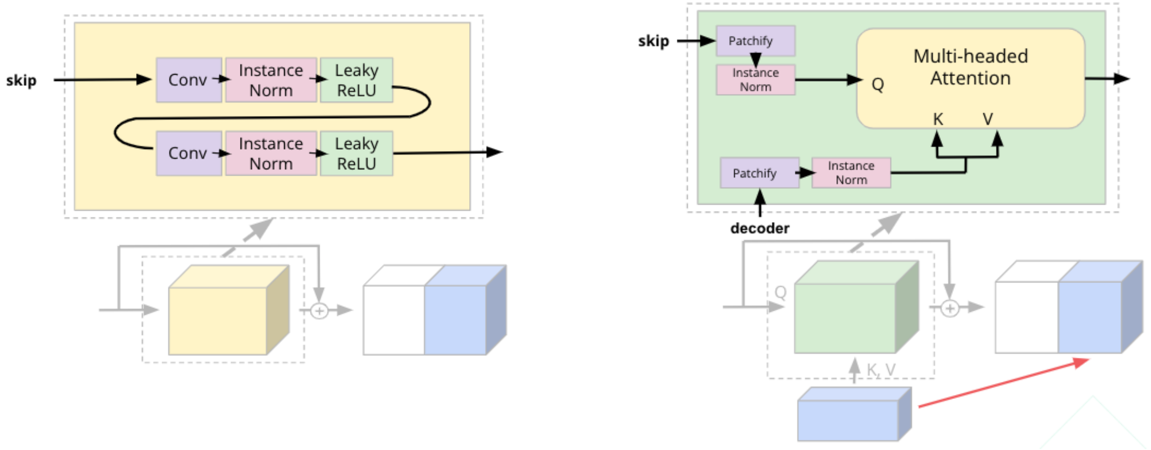
(a) Residual Block 1

(b) Residual Block 2

(c) ConvNext Block

Figure 2: The studied blocks are illustrated in this figure. The first set of blocks introduces a residual connection around the convolutions in the block to enhance the information flow. The second set adds a sequence of convolutions in front of the previous block for improved feature extraction. For both the first and second residual blocks, $3 \times 3 \times 3$ convolutions are used. Finally, the third set comprises ConvNext blocks, similar to the ones used in modern computer vision convolutional models.

In our study, we investigate two modifications to the encoder-decoder connection, as previously mentioned. The first type involves incorporating convolutional blocks into the skip connections, as demonstrated in the Swin-UNETR decoder (Hatamizadeh et al., 2022). The second type entails introducing a cross-attention module inspired by Oliver Petit's team's work on U-net Transformers (Petit et al., 2021). However, our approach differs from their implementation as we condition the attention mechanism using the downstream representation as the query. Additionally, we attempt an iteration of Residual Block 1 on the shortcut connection, inspired by the Swin-UNETR decoder, which has a convolution block dedicated solely to processing the residual connection (Hatamizadeh et al., 2022). This architectural change is referred to as the Conv-Skip model.



(a) Residual Block

(b) Attention Mechanism

Figure 3: Changes made to the skip connections between the encoder and decoder are demonstrated in this image. The first modification involves the addition of 3 × 3 × 3 convolutions to the pathway, while the second one incorporates a cross-attention mechanism, with the downstream layer acting as the query.

Finally, we explore an alternative version of the decoder called the Half-Unet, which is a simplified variation of the original U-net architecture (Lu et al., 2022). The Half-Unet also uses summation with the skip-connections from the encoder, as opposed to concatenation used in the previous U-net implementation. Our models are evaluated using the dice coefficient metric, commonly used in segmentation tasks, which takes into account both recall and precision. We also evaluate the best model and both the U-net and Swin-UNETR baselines by computing their average rank across all the test sets. A score closer to 1 indicates better performance across a wide variety of tasks. In addition to performance evaluation, we analyze

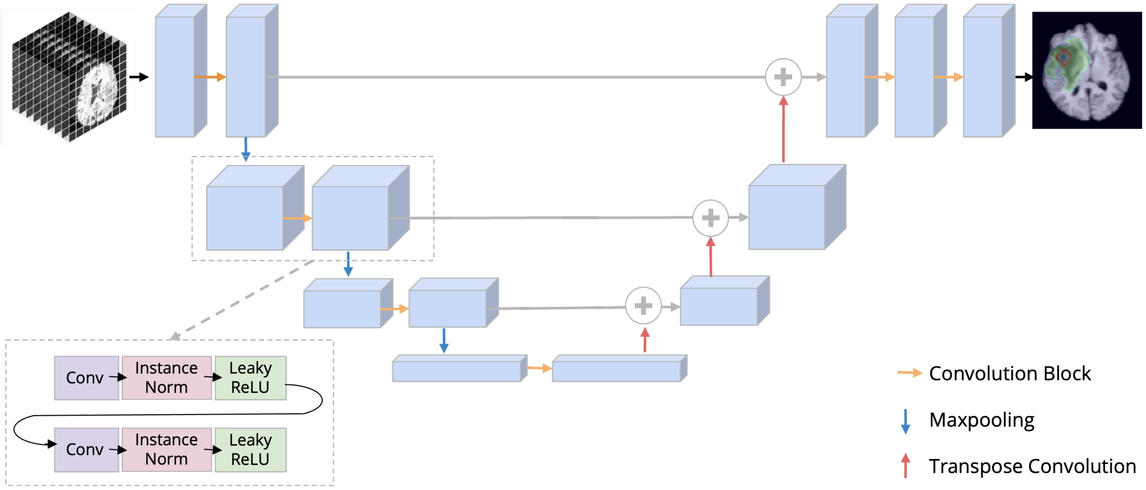


Figure 4: Half-Unet decoder is depicted in the following figure. The decoder has undergone significant changes, with all convolutions being removed and only upsampling remaining. Additionally, the concatenation operation has been replaced with summation to manage the sizes without convolutions

the computational efficiency of our models using the floating-point operations per second (FLOPs) metric.

4 Results

As previously stated, each model, including the baselines, was evaluated using a 5-fold approach on the heart and lung dataset. To account for the complexity of each fold's dataset, the scores of each fold were compared to those of the U-net baseline for the same fold. The results on the Resblock1 implementation produces highly variable results across the folds in both datasets. Conversely, the ConvNext block exhibits greater stability, but its average performance is lower than that of all other U-net iterations. As a result, experiments involving these implementations were terminated. Figures 5 and Table 1 present the results of each remaining model on the heart and lung dataset.

Dataset	Model	Val. Dice	Val. FWD	Val. p-value	Train Dice	Train FWD	Train p-value
Heart	**U-net**	0.81±0.03	NA.		0.84±0.009	NA.	
	Swin-UNETR	0.8±0.05	-0.0026±0.032	0.0	0.89±0.01	0.046±0.018	0.0
	Half-UNet	0.74±0.06	-0.072±0.044	0.0	0.84±0.04	-0.007±0.036	0.0
	Resblock-1	0.51±0.5	-0.3±0.43	0.1	0.53±0.5	-0.32±0.48	0.16
	Resblock-2	0.79±0.04	-0.012±0.044	0.35	0.83±0.03	-0.011±0.039	0.0
	ConvNext	0.46±0.08	-0.35±0.075	0.0	0.53±0.05	-0.32±0.05	0.0
	Trans-Unet	0.77±0.05	-0.034±0.037	0.0	0.86±0.01	0.02±0.0053	0.0
	Conv-Skip	0.81±0.03	-0.0015±0.022	0.18	0.84±0.01	-0.006±0.017	0.93
	Conv-Skip-Resblock-2	**0.82±0.02**	0.01±0.01	0.22	0.85±0.01	0.0074±0.0073	0.89
Lung	**Unet**	0.34±0.07	NA.		0.45±0.03	NA.	
	Swin-UNETR	0.28±0.05	-0.062±0.063	0.0	0.64±0.02	0.19±0.018	0.0
	Half-Unet	0.074±0.03	-0.27±0.056	0.0	0.19±0.06	-0.26±0.055	0.0
	Resblock-2	0.31±0.08	-0.03±0.07	0.67	0.42±0.09	-0.028±0.098	0.81
	Trans-Unet	0.14±0.1	-0.2±0.093	0.0	0.34±0.2	-0.11±0.21	0.66
	Conv-Skip	**0.35±0.09**	0.0093±0.052	0.82	0.5±0.02	0.054±0.02	0.0002
	Conv-Skip-Resblock-2	0.27±0.2	-0.069±0.13	0.03	0.35±0.2	-0.1±0.17	0.0

Table 1: This table presents the outcomes of the 5-fold cross-validation study on different models trained on both heart and lung datasets. The presented results are based on foreground dice scores, which refer to the heart class for the heart dataset and the tumor class for the lung dataset. FWD corresponds to the fold-wise difference with the U-net baseline, and the p-values indicate the probability of observing the 5-fold dice score under the U-net score distribution.

The Swin-UNETR and Trans-Unet models have shown better performance than the baseline on the training set, with dice scores of 0.89 and 0.86 respectively. However, their performance on the validation sets has been poor, with scores of only 0.8 and 0.77. Similarly, the Half-Unet model has also shown a decrease in performance on the validation set, with a dice score dropping from 0.84 on the training set to 0.74 on the validation set. These three models have also shown an increased variance in performance across folds on the validation sets. In contrast, the models that incorporate convolutions on their skip connections have shown improvement in both training and validation sets. The Conv-Skip model has an average dice score of 0.81, which is as good as the U-net model. Meanwhile, the Conv-skip model using Resblock 2 slightly outperforms them, with a dice score of 0.82. The Resblock 2 convolution blocks have also demonstrated good performance and exhibited greater stability across folds than the first iteration of the residual block. The segmentation task for the Lung dataset is more challenging than the previous dataset, as it involves identifying small tumors instead of segmenting the whole organ, and the dataset is larger. It was expected that models with higher capacity, such as those incorporating transformers, would demonstrate superior performance on this task compared to the previous one.

The segmentation task on the lung dataset is more challenging, resulting in lower performance for all models. However, the relative performance of the models is similar to

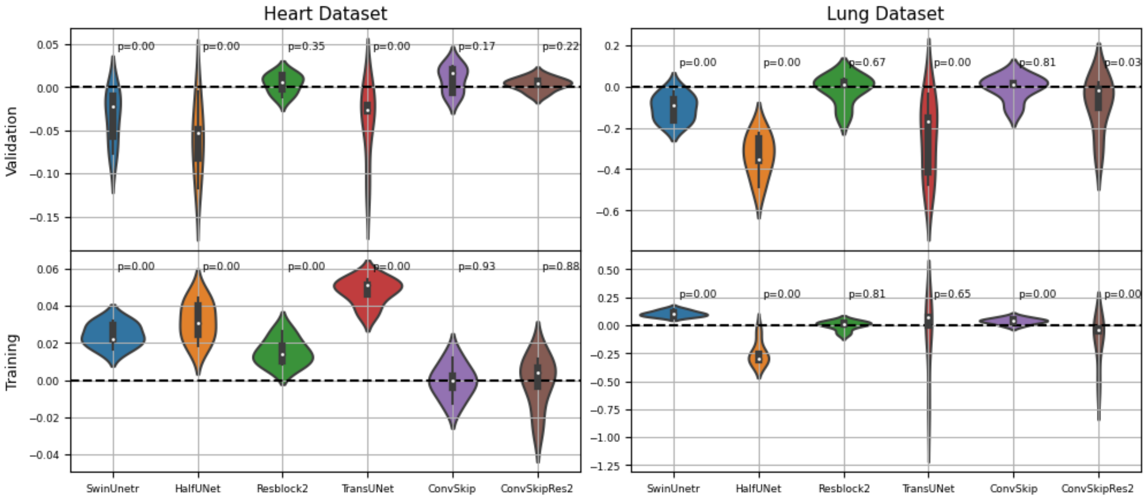


Figure 5: Foreground dice fold-wise difference with the U-net baseline. To evaluate the performance of various models, we calculated the foreground Dice score difference compared to the U-net baseline using a fold-wise approach. Specifically, we subtracted the training and validation foreground Dice scores of each fold of the U-net baseline model from the corresponding folds of the models under investigation, including Swin-UNETR, Half-Unet, U-net with Resblock 2, Trans-Unet, Unet-Conv-Skip, and Unet-Conv-Skip-Res2. We utilized a bootstrapping methodology to estimate the mean distribution used to compute the two-tailed p-values.

Dataset	Region	U-net	Swin-UNETR	Conv-Skip
BrainTumour	L1	0.643	0.627	**0.652**
	L2	0.432	0.398	**0.436**
	L3	**0.665**	0.628	0.655
Heart	L1	**0.837**	0.819	0.810
Liver	L1	**0.939**	0.929	0.938
	L2	**0.522**	0.376	0.498
Hippocampus	L1	0.873	0.870	**0.885**
	L2	0.881	0.873	**0.884**
Prostate	L1	**0.630**	0.508	0.621
	L2	0.850	0.797	**0.851**
Lung	L1	**0.465**	0.275	0.427
Pancreas	L1	0.618	0.501	**0.674**
	L2	0.083	0.088	**0.236**
HepaticVessel	L1	**0.457**	0.423	0.407
	L2	0.376	0.310	**0.452**
Spleen	L1	0.757	0.877	**0.913**
Colon	L1	**0.209**	0.044	0.121
Average Rank		1.647	2.765	**1.588**

Table 2: The performance of three models, U-net, Swin-UNETR, and Conv-skip, on all test sets in the Medical Segmentation Decathlon is presented. The metric used to evaluate the models is the dice score, and the results are tabulated. The average rank of each model is reported to enable a comprehensive comparison of their performance.

that on the heart dataset. Swin-UNETR performs better than the baseline on the training set, with a dice score of 0.64, but does not generalize well to the validation sets, with a score of only 0.28. Meanwhile, Trans-Unet performs the worst on both training and validation sets, with an average score of 0.34 and 0.14, respectively, and exhibits an increased variance in performance across folds. On the other hand, Conv-skip shows improvements over the baseline with a validation score of 0.82, while ResBlock 2 scores below the baseline with a dice score of 0.31. However, ResBlock 2 on the shortcut connection shows even lower performance than the baseline on average, with a dice score of 0.27. Additionally, the variance across folds on the validation set is greater for this task, even for the better-performing models, which could be attributed to the increased complexity of the lung segmentation task.The Half-Unet model's performance is below the baseline for both datasets and overall very low for the validation sets, which could be due to the model's decreased capacity resulting from the simplification of the decoder. It is important to consider the trade-off between efficiency and performance when evaluating whether the current simplification is worth the cost. It would also be helpful to assess the cost of performance increases for more promising models. Through analyzing this data, we can determine whether the potential benefits outweigh the costs.

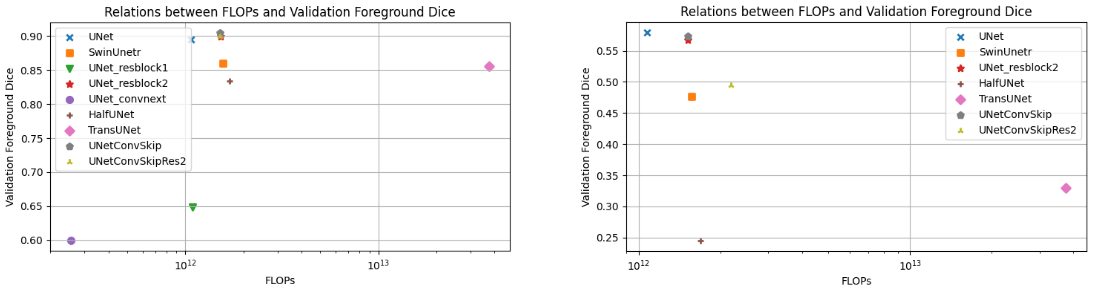


Figure 6: Heart validation dataset

Figure 7: Lung validation dataset

Figure 8: Dice score for the validation sets of both tasks is plotted against the FLOPs metric to illustrate the relationship between performance and computational efficiency for all studied models

FLOPs is a widely used metric in computational efficiency measurement in computer science. The higher the FLOPs, the more computationally expensive the code is. Figure 6 indicates that the two most efficient models had poor performance and were terminated. While the Trans-Unet implementation is the most computationally expensive among the models, it does not perform as well as some of the other models. The Half-Unet model has a similar FLOPs count to the other models in the experiment, but it does not perform as well.After conducting a more detailed examination, it has become evident that the convolution operation may not be the most expensive process in the decoder. Rather, it seems that the upsampling operation is more resource-intensive than initially estimated. Furthermore, the exclusion of the convolution operation appears to have a negative impact on performance. Resblock2 and Conv-Skip models have similar performance to each other in terms of Dice score and FLOPs efficiency. On average, they perform slightly better than the baseline U-net on the heart task but slightly worse on the lung task. However, they have a higher FLOPs cost than U-net. It is still unclear if the highest computational cost is worth it in this case. U-net-Conv-Skip2 performs similarly to the other Conv-Skip models on the heart task but does not perform as well on more challenging tasks like lung tumors. Finally, while Swin-UNETR has a similar FLOPs cost to the other models, its performance is not as good.

The Conv-Skip model was selected to train on all datasets due to its comparable validation performance to the baseline on the heart datasets and slightly better performance on the lung datasets. The model's performance also appears to be more stable than the other models. The increase in FLOPs is reasonable when compared to the other non-baseline models. However, the transformer-based model did not improve the performance of the standard U-net. One possible explanation for this is the lack of data augmentation or preprocessing in the study. Medical imaging datasets are often small and expensive to obtain labeled examples, and it is well-known that transformer models perform better with large datasets.

Moreover, the fixed depth of the model used in the study may limit its performance since the aim of transformers is to enhance long-range information retention, which may require deeper implementations. It is unlikely that these implementations will achieve state-of-the-art scores in the competition, as no data preprocessing or ensembling methods have been employed. The primary objective is to compare the best-performing model with the baseline on the test sets for all tasks.

Table 2 displays the test set results for the competition across all datasets. The Swin-UNETR implementation exhibited inferior performance compared to other models across all tasks. The optimal performance between U-net and Conv-Skip varied not only from task to task but also among different classes within a task. For example, in the HepaticVessel dataset, U-net outperformed Conv-Skip on the L1 region with a score of 0.457, whereas Conv-Skip performed better on L2 with a score of 0.452 compared to U-net's score of 0.407 and 0.376, respectively. Overall, Conv-Skip appeared to perform better on most classes of the BrainTumour, Hippocampus, Pancreas, and Spleen, while U-net performed better on the Heart, Liver, and Colon datasets. However, the performance of both models on other datasets varied based on the class, and it remains unclear what specific dataset characteristics caused one model to outperform the other. On average, Conv-Skip appeared to perform better on more classes and had a higher average rank with 1.588 than the baseline U-net with 1.647.

5 Conclusion

In this study, it was found that only a limited number of models performed better than the standard U-net implementation. The Residual block and ConvNext block were removed from the study because of their subpar performance. Simplifying the decoder in Half-Unet did not contribute significantly to efficiency gains and resulted in reduced overall performance. The state-of-the-art transformers-based models, including Swin-UNETR and Trans-Unet, were not able to match the performance of the U-net.The Conv-Skip model demonstrated the best performance, and adding a convolutional layer improved the performance of the residual block, making it as effective as, if not better than, the U-net baseline and the Conv-Skip model on most datasets. However, the performance of the best implementation, Conv-Skip, was not statistically significant enough to claim superiority over the classic U-net implementation. To enhance the study further, it would be worthwhile to consider incorporating pretreatment techniques and utilizing larger models in future research. Specifically, investigating the effects of data augmentation would be of great interest, particularly in determining whether models that utilize attention mechanisms and have greater capacity can outperform those based on convolution. However, implementing these modifications may require additional computational resources, such as larger GPUs or clusters, to handle larger datasets and models. With these enhancements to the methodology, the study could achieve results that are more comparable to those presented in the original paper and closer to the current state-of-the-art.

6 Team contribution

Arthur Boschet : Developed U-net baseline model, convolutional and attention blocks. Performed 5-fold cross-validation experiments for various models, analyzed results, worked on the final presentation and report.
Clément Detry : Exploration of the datasets, implementation of the pytorch data loaders and train function, setting up of the wandb project structure and visualisations, management of

the models experiments on the cluster.
Frederic Gagne : Exploration of the datasets, implementation of different encoders (such as patchify and imageskip) and decoders (including Half-Unet), analysis of results, work on presentations, redaction of the final report, and completion of mid-session assignments.

References

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D. (2022). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer.

Lee, K., Zung, J., Li, P., Jain, V., and Seung, H. S. (2017). Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.

Lu, H., She, Y., Tie, J., and Xu, S. (2022). Half-unet: A simplified u-net architecture for medical image segmentation. *Frontiers in Neuroinformatics*, 16.

Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., and Soler, L. (2021). U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., et al. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.