A Comprehensive Approach for COVID-19 CT Images Segmentation using Swin U-Net Transformer(Swin UNETR) and U-Net

Safia Mazoz
safia.mazoz@centrale-casablanca.ma

Ayoub Benrguig
ayoub.benrguig@centrale-casablanca.ma

Aymane Rahouti
aymane.rahouti@centrale-casablanca.ma

Reda Benkirane
reda.benkirane@centrale-casablanca.ma

Walid Hirouche
walid.hirouche@centrale-casablanca.ma

Oussama Es-Semyry
oussama.essemyry@centrale-casablanca.ma

[a]*Ecole Centrale Casablanca, Bouskoura*

Abstract

Accurate medical image segmentation is essential for advancing healthcare systems, especially for disease diagnosis and treatment planning. While the U-shaped architecture (**U-Net**) has shown Significant success in many segmentation tasks, it has limitations in modeling long-range dependencies due to the local nature of convolution operations. Transformers, designed for sequence-to-sequence prediction, offer global self-attention but may lack precise localization. In this article, we propose **Swin-UNeTR** as a strong alternative to **U-Net** for medical image segmentation. Our approach combines the strengths of Transformers for encoding global contexts and **U-Net** for recovering localized spatial information, achieving exceptional performance in multi-organ and pulmonary segmentation tasks. this work aims to reduce the manual workload of specialists, to facilitate the marking of lesions for optimal diagnosis, and to rationalize analysis time correct.

Keywords: medical image, segmentation, Transformers

1. Introduction

The segmentation of medical images involves the crucial process of identifying and delineating regions of interest(**ROIs**), such as lesions, organs, and tissues within various medical images. This crucial task plays a significant role in numerous clinical applications, spanning from the diagnosis of diseases like **COVID-19** to treatment planning and the monitoring of pathological progression [5],[26]. Traditionally, radiologists have performed manual segmentation by meticulously examining volume, shape, and location layer by layer. However, this approach proves time-consuming and challenging to adapt to the ever-expanding volume of medical data. Additionally, manual segmentation process suffers from inter-observer variabilities were two medical practitioners may

disagree between the exact locations of the **ROIs**. In such situations, there is an urgent need to embrace automated and robust medical image segmentation methods within clinical settings to enhance efficiency and facilitate scalability.

Recent advancements in the field of medical image analysis has been characterized by a notable increase in the adoption of deep learning models[11]. This surge has led to the emergence of various segmentation models, each designed and trained for specific tasks, commonly denoted as 'specialist' models. Despite the notable success of deep learning algorithms in achieving high performance in cross-validated diagnostic tasks such as **COVID-19** screening from X-ray and CT modalities, it is important to note that these accomplishments have predominantly relied on extensive amounts of training data. This includes substantial volumes of **COVID-19** positive images sourced from many cases. Using many positive cases to train a model poses problems, especially with a new disease. The challenge lies in the delay between when the disease becomes a public health problem and the availability of training data, due to rules such as the Health Insurance Portability and Accountability Act (**HIPAA**) and the institutional Review Board (**IRB protocols**) [31],[4]. This has led to research into how well a deep learning algorithm can detect disease with few positive samples [23],[19]. **COVID-19** has been successfully classified from CT scans using many positive cases[23],[19], but it needs to be shown that this works well even with fewer samples. This is important in prevention for a future pandemic where getting enough training examples might be difficult.
The present work aims is to automate the identification and segmentation of **COVID-19** lesion regions on chest computed tomography (CT) images. Our approach objective is to reduce the manual workload of specialists, to facilitate the marking of lesions for optimal detection of **COVID-19**, and to rationalize analysis time.
We have established a segmentation procedure to identify the pulmonary lesions and thus assist specialists in monitoring these alterations, thus contributing to the management of complications in the infected patient. To this end, a preprocessing phase was conducted on the CT scans to prepare the dataset for training purposes. Finally, the **swin UNETR** deep learning technique [14] was used to precisely segment the regions of lesions caused by **COVID-19**. The results of this segmentation are used to generate visualizations of lesions, thus optimizing the analysis of specialists while allowing monitoring of lesions in patients.

The major contributions of our methodology are as follows:

1° Evaluation of different segmentation models such as **U-Net**, **Basic U-Net**, and **Flexible U-Net**
2° Proposal and Evaluation of a Novel State-of-The-Art Architecture: **Swin UNETR** , aiming to enhance capabilities in **COVID-19** lesion identification and segmentation.

The remaining of this work is organized as follows : a review of related works is presented in **Section 2**, while **Section 3** provides detailed information on the tools and methods employed. **Section 4** outlines the experiences conducted and the results obtained, including comparisons between different models. A conclusion ends our article.

2. Related works

Since the introduction of **U-Net** [29], CNN-based networks have demonstrated exceptional results in various 2D and 3D medical image segmentation tasks [7], [43], [41], [8], [17]. For volume-wise segmentation, tri-planar architectures, also known as 2.5D methods [17], [20], [38], are often utilized. These involve combining three-view slices for each voxel. In contrast, 3D methods directly use the full volumetric image, represented by a sequence of **2D** slices or modalities. To capture downsampled features of images, multi-scan and multi-path models [13], [14], [2] have been employed. Additionally, hierarchical frameworks have been investigated to exploit **3D** context and address computational resource limitations, with some methods focusing on feature extraction at multiple scales [11]. Roth et al. [30] proposed a multi-scale framework for pancreas segmentation, offering varying resolution information. These methods form the basis of pioneering studies in 3D medical image segmentation at multiple levels.

Vision Transformers have recently become prominent in the field of computer vision. Highlighted by Dosovitskiy et al. [6], they have shown exceptional performance in image classification through extensive pre-training and fine-tuning. These transformers have also made significant strides in object detection, distinguishing themselves in various benchmarks [1], [44]. Recent developments include hierarchical vision transformers [21], [36], [3], [40], which progressively reduce feature resolution while employing sub-sampled attention for efficiency. Different from these approaches, the **UNETR** encoder maintains a consistent representation size across all transformer layers. For **3D** medical image segmentation, approaches like those of Xie et al. [39] and Wang et al. [37] have integrated transformers with **CNN** backbones, showcasing their potential in complex segmentation tasks like brain tumor segmentation. Our method uniquely leverages skip connections to directly link the transformer-encoded representation to the decoder.

3. Methodology

In recent years, Fully Convolutional Neural Networks (**FCNNs**), specifically the "U-shaped" architecture, have dominated 3D medical image segmentation. However, due to limited convolution layer kernel size, long-range information modeling is sub-optimal[28]. In contrast, inspired by the Transformer scaling successes in Natural Language Processing (**NLP**) , [35] originally proposed by Vaswani et al.,we reformulate the task of volumetric (**3D**) medical image segmentation as a sequence-to-sequence prediction problem. Hence, we propose Swin transformers [22, 21], that computes self-attention in an efficient shifted window partitioning scheme [9], inspired by successful **vision transformers**. Thus, we compare the performance of **ViT** models with "**U-shaped**" architectures in the task of Covid-19 infection segmentation.

*3.1. **Vision transformer models***

Recently proposed by Dosovitskiy et al. [6], vision transformers have become the de-facto standard for image recognition and segmentation. **ViT** excel in various image recognition benchmarks, outperforming existing models, like **CNN** [16] or classic ResNet-like architectures that were the state of the art back in 2020 [6]. Notably, the top-performing **ViT** model achieves an accuracy of 88.55% on **ImageNet**, 90.72% on **ImageNet-ReaL**, 94.55% on **CIFAR-100**, and 77.63% on the **VTAB** suite, which consists of 19 distinct tasks. [6]

3.1.1. *UNEt TRansformers : UNETR*

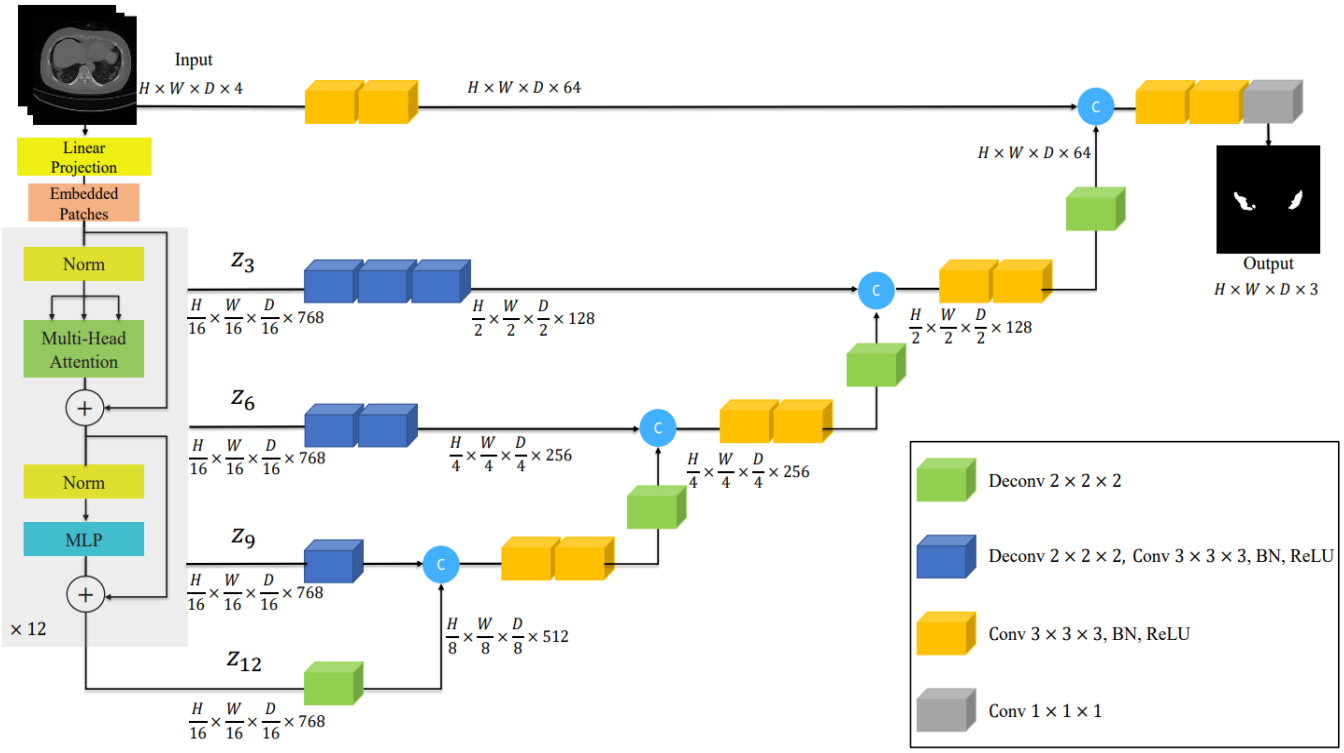

Figure 1: Overview of **UNETR**. The proposed model consists of a transformer encoder that directly utilizes **3D** patches and is connected to a CNN-based decoder via skip connection.

The **UNETR** model is configured with the following parameters:

Model	Input channels	Output channels	Image size	Dropout rate	Spatial dimensions
UNETR	3	3	224x224	0.5	2

Table 1: UNETR parameters implementation

This architecture is designed to effectively process **2D** images with three input channels, producing three output channels for segmentation. The input image size is set to (224, 224), and dropout regularization with a rate of 0.5 is applied to enhance model generalization. The spatial dimensions are tailored for **2D** image processing. As for the optimizer, we used Adam, an algorithm for optimizing stochastic objective functions using first-order gradients, relying on adaptive estimates of lower-order moments. [15]

Our loss function is a combination of soft dice loss [24, 10] and cross-entropy loss [42], it appears that a mild imbalance is well handled by these loss strategies designed for imbalanced datasets. The loss function is calculated voxel-wise as per the following expression.

$$L(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j}$$

Where I is the number of voxels; J is the number of classes; $Y_{i,j}$ and $G_{i,j}$ denote the probability

output and one-hot encoded ground truth for class j at voxel i, respectively.

We also considered a sigmoid activation when applied to the model's output, withing the dice losss function.

3.1.2. ***Swin UNEt TRansformers : Swin UNETR***

This model is designed for image processing [9] tasks involving 3-channel input images (commonly representing RGB colors) and producing 3-channel output. The expected size of input images is set to (224, 224), indicating a preference for square images of this dimension. The incorporation of a dropout rate of 0.5 suggests that during training, approximately half of the neurons will be randomly omitted at each update, serving as a regularization technique to prevent overfitting. Additionally, the model is configured for **2D** spatial data, as denoted by the parameter spatial dimensions set to 2. The configuration of the parameters is the same as the ones of the UNETR model.

As for this time, We used the soft Dice loss function [24] which is computed in a voxel-wise manner as :

$$L(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2}$$

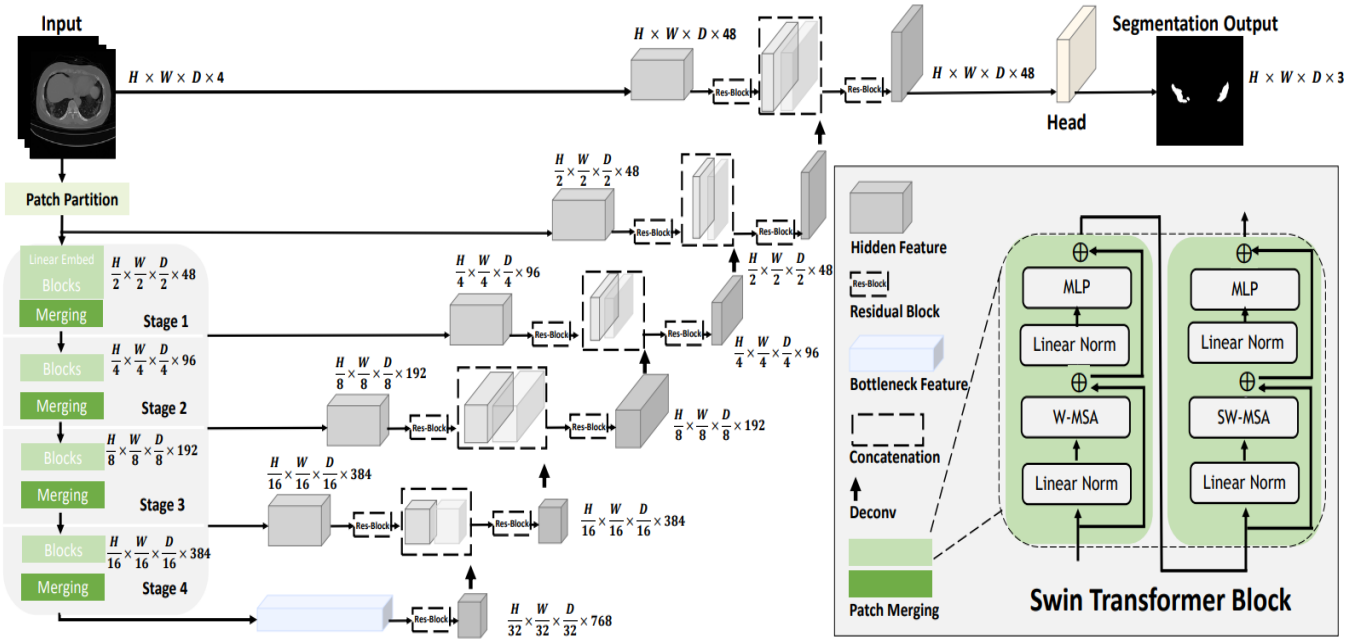

Figure 2: Overview of the Swin UNETR architecture

*3.2. **U-Net approach and its variants***

In our implementation, we leveraged the U-Net architecture [29], which is obviously widely used for image segmentation tasks. Henceforth, we explored variants provided by MONAI, a leading medical imaging library, where we employed the **Flexible U-Net**, which incorporates the **EfficientNet-B0** [34] backbone for improved efficiency and performance. Further, we used the **Basic U-Net** and **SegResNet** [18, 32], designed for **2D** spatial data.

4. Experiments

*4.1. **Dataset description***

The dataset used merges the **COVID-19** lesion **masks** and their corresponding **frames** of 3 public datasets with 2729 image and ground truth mask pairs. All different types of lesions are mapped to white color for consistency across datasets. The 3 public datasets are, MOSMED dataset collected in NIfTI format by Mozorov et al. [25]. The second dataset is CoronaCases [12]. The third one is Radiopaedia. [27]

*4.2. **Implementation Details***

We implement UNETR in PyTorch and MONAI. All our models were trained using a Kaggle notebooks with **P100 GPU**, 30GB in RAM and 16GB in **GPU** memory. Kaggle offered us a quota of 30 hours in the use of its **GPU** P100 and a quota of 12 hours in uninterrupted session.

The models were trained with the **batch size of 32**, using the Adam optimizer [15] with constant learning rate of 0.001 and $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for a null value of weight decay. For the specified batch size, the average training time of Swin UNETR and Flexible U-Net was 2 hours for 30 epochs. As per the others, it was around 8 minutes to 1 hours. Additionally, we applied data augmentation transforms of random rotation (180 degrees), random affine transformations (180 degrees), random horizontal flipping (applied with a probability of 1, i.e., 100% of the time), and random vertical flipping (also applied with a probability of 1). These augmentations aim to enhance the diversity of the training dataset by introducing variations in orientation and spatial transformations, improving generalization of the trained models.

5. Evaluation and Metrics

In the evaluation process, a comprehensive set of metrics was employed to assess the performance of the trained models. Following each epoch of training, key metrics were calculated and recorded, providing insights into the model's performance. The metrics include **Dice Loss**, **intersection over union (IOU)**, **generalized dice (GD)** [33], **recall**, **precision**, and **F1 score**. These metrics offer a good understanding of the model's ability to accurately segment **COVID-19** infection areas in CT images. The recorded values were logged using the **TensorBoard** writer, facilitating detailed analysis and comparison of different training runs. This rigorous evaluation approach not only ensures the reliability of the model but also serves as a basis for informed decisions on hyperparameter tuning and model improvement.

6. Results and Discussion

Model	Dice Loss	IOU	GD	F1 score	Precision	Recall	Training time (s.)
U-Net	0.45	0.427	0.561	0.563	0.581	0.629	480
Basic U-Net	0.302	0.571	0.699	0.7	0.691	**0.791**	2400
Flexible U-Net	0.272	0.602	0.729	0.731	0.772	0.749	3300
SegResNet	0.251	0.636	0.752	0.753	0.756	0.788	3300
UNETR	0.347	0.527	0.662	0.661	0695	0.707	5400
Swin UNETR	**0.243**	**0.64**	**0.762**	**0.763**	**0.812**	**0.792**	7200

Table 2: Evaluation metrics on local validation set

In the evaluation results presented in Table 2 for the segmentation task on our validation set. The Swin UNETR model is clearly the most dominant among the considered architectures. The metrics, including Dice coefficient, IOU, F1 score, Precision, Recall, collectively underscore the superiority of Swin UNETR, although the execution time of this model is the most expensive one even when trained on NVIDIA GPU P100, but the performance compensates this problem. In the Table 2 above, we observe that there is a great balance between **Recall** and **Precision** as far as most of the models displayed, which indicates that we are correctly classifying as many positive instances as possible. Hence, we are capturing as many True positive instances as possible, minimizing false negatives and false positives. Achieving this balance is crucial in the field of medical imaging, since it is indeed too dangerous to predict a positive instance as negative than to predict a negative instance as positive, thus the importance of the balance in recall and precision. Furthermore, high Generalized Dice score in Swin UNETR indicates its effectiveness in achieving a well-balanced segmentation across classes. This makes Swin UNETR a promising model for accurate and clinically relevant segmentation in the context of COVID-19 CT images. Although there is a slight difference of 0.1% in the Recall metric between Swin UNETR and **Basic U-Net**. We highlight also the competitive results of **SegResNet** model in terms of all the metrics which affirms its use for segmentation tasks, since it is Residual Network model refined for segmentation tasks.

7. Limitations and Challenges:

The implementation of our proposed solutions encountered several challenges, reflecting the intricacies of working on such a complex project.

7.1. Dataset Availability:

Firstly, the designated dataset for training and testing, Mosmed, containing **1100 CT images** in 3D (approximately **26,000 slices post-preprocessing**), was not directly available for download via a standard link. The only source we found to obtain the data was through a BitTorrent link. Conse- quently, we had to wait for a person with the complete dataset and a good upload speed to be online before initiating the download process.

7.2. Issues with Data Preprocessing Notebook:

We faced a big challenge with a poorly documented and error-prone notebook used for the project's data preprocessing. The notebook, initially meant for data preparation, contained deprecated modules that required code modifications. Unfortunately, these outdated components were better suited for testing data rather than the intended purpose of training data preprocessing. Adapting the code to address these issues consumed a considerable amount of time and significantly hindered our progress.

Our initial working environment on Google Colab presented additional challenges due to the limitations of the free version—restricted to **12GB RAM** and an insufficient GPU (T4) for efficient GAN model training, lasting a minimum of 9 hours. We also faced recurrent runtime disconnections. Consequently, we transitioned to Kaggle, which eased most of these challenges by providing **29GB of RAM**, a more powerful GPU (P100), and a stable runtime without disconnections. However, the inability to connect directly to Google Drive required us to download individual files using the Python *gdown* module.

8. Conclusion

This study represents a significant advancement in the automated segmentation of COVID-19 lesions in CT images, showcasing the potential of combining advanced deep learning architectures like Swin UNETR which is one of the novel architectures. The results affirm the efficiency of thi model which ranks among top performing approaches in the validation phase and demonstrates competitive performance in the testing phase as compared to the U-Net model and its variants in improving the accuracy and of medical image analysis, particularly in pandemic scenarios like COVID-19. Future research could explore the application of these methods to other medical imaging tasks like semantic segmentation of brain tumors using MRI images and the integration of such technologies into clinical workflows for enhanced diagnostic and treatment planning.

Acknowledgments

We would like to thank Dr. Oumayma Banouar for guiding and supervising us through this research project. We would like also to acknowledge her aid for funding our project and also providing us the necessary computational resources. We acknowledge also the full fledged datasets collected by renowned researchers ie. Mozorov et al. and others, which may help revolutionize the medical imaging field.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[2] Jianxu Chen, Lin Yang, Yizhe Zhang, Mark Alber, and Danny Z Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3044–3052, 2016.

[3] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *arXiv preprint arXiv:2104.13840*, volume 1, 2021.

[4] Joseph Paul Cohen, Paul Morrison, Long Dao, Kevin Roth, Tung Quang Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[5] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automatic liver segmentation from ct volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 149–157. Springer, 2016.

[8] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Transactions on Medical Imaging*, 37(8):1822–1834, 2018.

[9] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[11] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[12] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, et al. Covid-19 ct lung and infection segmentation dataset. 2020.

[13] Konstantinos Kamnitsas, Liang Chen, Christian Ledig, Daniel Rueckert, and Ben Glocker. Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. In *Ischemic Stroke Lesion Segmentation*, volume 13, page 46, 2015.

[14] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[17] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

[18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[19] Siqi Liu, Bogdan Georgescu, Zhoubing Xu, et al. 3d tomographic pattern synthesis for enhancing the quantification of covid-19. *arXiv preprint arXiv:2005.01903*, 2020.

[20] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 851–858. Springer, 2018.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[23] Mohamed Loey, Gunasekaran Manogaran, and Nour Eldeen M Khalifa. A deep transfer learning model with classical data augmentation and cgan to detect covid-19 from chest ct radiography digital images. *Neural Computing and Applications*, pages 1–13, 2020.

[24] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[25] Sergey P Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

[26] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

[27] Radiologists. Covid-19 ct segmentation dataset, 2020. Accessed 23 December, 2020.

[28] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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, 2015.

[30] Holger R Roth, Hirohisa Oda, Yuichiro Hayashi, Masahiro Oda, Natsuki Shimizu, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. Hierarchical 3d fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382*, 2017.

[31] Junaid Shuja, Eman Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. Covid19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51(3):1296–1325, 2021.

[32] Nahian Siddique, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: theory and applications. *arXiv preprint arXiv:2011.01118*, 2020.

[33] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[36] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[37] Wenxuan Wang, Chen Chen, Meng Ding, Jiangyun Li, Hong Yu, and Sen Zha. Transbts: Multimodal brain tumor segmentation using transformer. *arXiv preprint arXiv:2103.04430*, 2021.

[38] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3646–3655, 2020.

[39] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. *arXiv preprint arXiv:2103.03024*, 2021.

[40] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. *arXiv preprint arXiv:2104.06399*, 2021.

[41] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[42] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[43] Qikui Zhu, Bo Du, Baris Turkbey, Peter L Choyke, and Pingkun Yan. Deeply-supervised cnn for prostate segmentation. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 178–184. IEEE, 2017.

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *arXiv preprint arXiv:2010.04159*, 2020.