Hybrid Vision Transformer for ToS prediction using Myocardium Masks

Rugved S Chavan QXK6FB@VIRGINIA.EDU
Rishabh Jain TPE3SJ@VIRGINIA.EDU
Naeem Patel NVJ3YG@VIRGINIA.EDU
Ishita Gupta RSY7TR@VIRGINIA.EDU

* Team 11: All authors contributed equally.

Abstract

The landscape of cardiac image analysis is evolving, with traditional convolutional neural network (CNN) methods being supplemented by the emerging capabilities of Transformer architectures. In this research, we present a novel application of the Vision Transformer (ViT) model for predicting Target Organ Status (TOS) values from sequences of myocardium mask image patches. Our approach, which adapts the ViT model originally pioneered for natural language processing tasks, is tailored to handle the spatial and temporal complexities inherent in cardiac sequences. We introduce ViT+Dense and ViT+LSTM architectures, enhancing the predictive performance of the model by integrating temporal information across image sequences. The ViT+LSTM model demonstrated superior performance with a Mean Absolute Error (MAE) of 8.1491 on test data, outperforming the ViT+Dense and 3DCNN models, which achieved MAEs of 8.213 and 8.3106, respectively. These results not only highlight the efficacy of the ViT+LSTM model in capturing complex temporal dependencies but also underscore the transformative potential of Transformer-based models in cardiac imaging and medical diagnostics at large.

Keywords: Vision Transformers, 3D CNNS, Myocardium Masks, Machine Learning for Image Analysis

1. Introduction (Rishabh Jain)

The emergence of machine learning techniques in 2D temporal image analysis has revolutionized the understanding of dynamic visual data. With machine learning on the rise, particularly deep learning architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the field witnessed a paradigm shift. These models, equipped with their ability to learn hierarchical representations and temporal dependencies, enabled more sophisticated analysis of sequences of 2D images over time. The adoption of these models has paved the way for precise and nuanced insights derived from time-varying visual data, advancing the understanding of temporal information within 2D image sequences.

The advent of deep learning, particularly convolutional neural networks (CNNs) and their variants, has revolutionized medical imaging analysis. These deep learning architectures excel at automatically learning hierarchical representations from large volumes of imaging data, allowing for more accurate and efficient detection, segmentation, classification, and even generation of medical images. Their ability to discern intricate patterns and features within medical images has led to remarkable progress in various diagnostic areas, including but not limited to, detecting diseases such as cancer, identifying anomalies in radiology scans, segmenting organs or tissues, and predicting Target Organ Status (ToS) in cardiac images.

Transformers, initially popularized by (Vaswani et al., 2017) in 2017, have emerged as a pivotal model in NLP, commonly pre-trained on extensive text corpora before fine-tuning on specific datasets (Devlin et al., 2019). Despite their dominance in NLP, the adoption of Transformers in computer

vision remains limited, with convolutional architectures maintaining their popularity (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2020).

Due to the successes of Transformer scaling in NLP and the combination attempts of CNN-like architectures with self-attention mechanisms (Qiao et al., 2019), our research endeavors to explore the application of Vision Transformers directly to cardiac images, focusing on predicting ToS values using Myocardium Masks. The Vision Transformer adapts the Transformer architecture by segmenting the image into patches, treating them as analogous to tokens in NLP, and subsequently feeding these patches' linear embeddings as input to the Transformer model. In this project, we aim to predict the ToS values of cardiac images using Vision Transformers and compare their performance with state-of-the-art CNN models. This research may result in the adoption of Vision Transformers in medical imaging, replacing current neural network-based approaches.

This paper is divided into six main sections - Introduction, Background, Methodology, Results, Discussion, and Conclusion. The Background delves into this domain's foundational knowledge and previous work. The Methodology outlines the specific approach taken, detailing the segmentation, modeling, and training methodologies adopted. Following this, the Experiment section showcases our experiments' empirical outcomes and performance evaluations. The Discussion section critically analyzes these results, highlighting their implications, limitations, and potential for further exploration. Finally, the Conclusion encapsulates the key findings, affirming the significance of our research and outlining future directions in the field of cardiac image analysis leveraging Transformer architectures.

2. Background (Naeem Patel)

In recent years, the field of medical imaging, especially in the domain of cardiac imaging, has undergone a paradigm shift, driven by significant advancements in deep learning techniques. This section provides a comprehensive overview of the evolution of both Three-dimensional Convolutional Neural Networks (3D CNNs) and Vision Transformers (ViTs) in medical imaging, highlighting their contributions and applications. Additionally, it delves into the transformative impact of deep learning in the specific context of cardiac imaging, elucidating key research contributions and challenges faced by the integration of Target Organ Status (TOS) values derived from myocardium mask image patches.

2.1. Deep Learning in Cardiac Imaging

Deep learning has made remarkable strides in cardiac imaging, revolutionizing the analysis of myocardial tissue characteristics. (Avendi et al., 2016) laid foundational work with a combined deep-learning and deformable-model approach for fully automatic segmentation of the left ventricle in cardiac MRI. This breakthrough paved the way for subsequent tissue-specific analyses, influencing the trajectory of research in cardiac imaging. In disease detection, (Rafi and Woong Ko, 2022) presented HeartNet, a deep learning model for the automatic detection of hypertrophic cardiomyopathy, showcasing the potential of deep learning in identifying specific myocardial pathologies. The study by (Attia et al., 2019) introduced an artificial intelligence-enabled electrocardiogram for screening cardiac contractile dysfunction, emphasizing the integration of deep learning into real-time diagnostics. Advancements in the estimation of ejection fraction, a crucial aspect of cardiac functional analysis, were achieved with the work of (Ghorbani et al., 2020), proposing a fully automated echocardiogram interpretation model using deep learning. Additionally, (Bai et al., 2018) presented an automated cardiovascular magnetic resonance image analysis using fully convolutional networks, contributing to the accurate assessment of myocardial function. For real-time imaging and diagnosis, (Madani et al., 2018) introduced a fast and accurate view classification of echocardiograms using deep learning, showcasing the feasibility of real-time deep learning applications in cardiac imaging. However, challenges persist, including the need for large annotated datasets, interpretability, and addressing algorithmic bias.

2.2. Evolution of 3D CNNs in Medical Imaging

The inception of 3D CNNs in the 2010s set the stage for a transformative journey in medical imaging. Early works, as referenced by (Litjens G, 2017), explored the extension of 2D CNN architectures to handle volumetric data, laying the foundation for subsequent breakthroughs. The advent of powerful Graphics Processing Units (GPUs) has been instrumental in training larger and more sophisticated 3D CNN models, facilitating the design of custom architectures tailored specifically for medical imaging tasks (Özgün Çiçek et al., 2016). The integration of transfer learning strategies, such as pre-training on non-medical datasets like ImageNet, has become commonplace, enabling improved performance even with limited medical data (Shin et al., 2016). Attention mechanisms, inspired by their success in computer vision, have been seamlessly incorporated into 3D CNNs to capture relevant spatial and temporal dependencies within medical volumes (Chen et al., 2018). Applications of 3D CNNs in medical imaging are diverse, spanning tasks such as image segmentation (Milletari et al., 2016), disease detection and classification, image registration, reconstruction, drug discovery, real-time imaging analysis, and personalized medicine. Ongoing research, as highlighted in recent papers (Zhang et al., 2021), continues to refine and expand the capabilities of 3D CNNs, making substantial contributions to the fields of diagnosis, treatment planning, and healthcare delivery.

2.3. Evolution of Vision Transformers (ViTs) in Computer Vision

The landscape of Vision Transformers has witnessed significant evolution, spurred by seminal works and subsequent advancements. The foundational (Vaswani et al., 2017) introduced the Transformer architecture, initially demonstrating its efficacy in natural language processing and subsequently paving the way for applications in computer vision. (Dosovitskiy et al., 2010) highlighted the applicability of transformers to image recognition tasks, while "An Image is Worth 16x16 Words" (Dosovitskiy et al., 2021) scaled ViTs to achieve state-of-the-art performance. Innovative ViT variants, such as Swin Transformer (Liu et al., 2021), have addressed computational efficiency and scalability concerns by utilizing shifted windows for hierarchical feature representation. Other works, like (Touvron et al., 2021) explored hybrid data training to enhance ViT performance by integrating diverse data sources. Meanwhile, (Chu et al., 2021) delved into spatial attention mechanisms, refining ViT design for improved performance. (Caron et al., 2021) and (Yuan et al., 2021) showcasing the potential of ViTs in transfer learning scenarios. This collective body of work underscores the rapid evolution of Vision Transformers, encompassing architectural enhancements, scaling considerations, innovative attention mechanisms, and effective pre-training strategies.

3. Methodology (Rugved Chavan)

This section offers a systematic exploration, initially focusing on 3D Convolutional Neural Networks (CNNs). Subsequently, we will examine the functionality of the Transformer Model, emphasizing its Attention Mechanism. This will be followed by an analysis of how the Transformer is implemented in Vision Transformers (ViT). Furthermore, we will delve into the architecture of our novel hybrid model, which integrates LSTM and 3D CNN with ViT.

3.1. Dataset

The dataset integral to our research comprises two key components: Myocardium Masks and Target Organ Status (ToS) ground truth data as shown in figure 1. The Myocardium Masks image dataset consists of data from 128 patients. For each patient, there is a sequence of 25 frames, with each frame having a resolution of 80x80 pixels. This forms a data structure with a shape of (128, 80, 80, 25), indicating that the dataset includes 128 patients, each represented by a series of 25 images with a

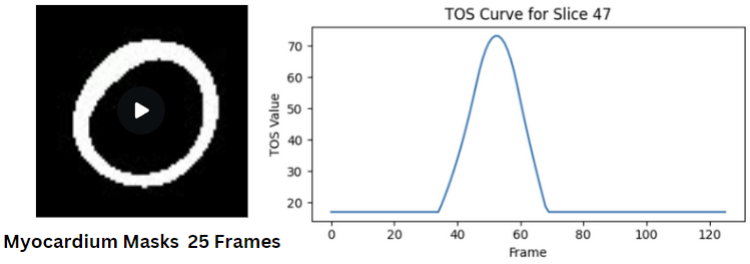

Figure 1: 3D CNN Architecture

resolution of 80x80 pixels. These images provide a comprehensive view of myocardial structure and function over time, essential for analyzing cardiac health and pathology.

Complementing the image data, the ToS ground truth component consists of a one-dimensional array containing 126 data points for each patient. This results in a data shape of (128, 126), representing the 128 patients and their corresponding 126 ToS values. The ToS data is crucial for our study as it provides the target values for the predictive modeling, enabling the evaluation of the model's accuracy in predicting clinically relevant outcomes.

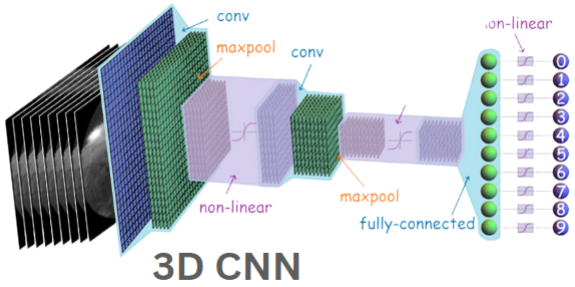

Figure 2: 3D CNN Architecture

3.2. 3D CNNS

Our study utilizes a 3D Convolutional Neural Network (3D CNN) to analyze myocardium masks and predict Target Organ Status (ToS) values as shown in 2. The 3D CNN is adept at processing volumetric data, extracting spatial-temporal features crucial in medical image analysis. Our model architecture begins with a 3D convolutional layer of 64 filters, followed by subsequent layers with increased filter sizes (128 and 256), interspersed with 3D max pooling for dimensionality reduction and feature abstraction. A dropout layer with a 0.3 rate is incorporated to mitigate overfitting. The network concludes with a global average pooling layer, maintaining essential features, and a dense output layer tailored to the 126 ToS values.

Implemented in TensorFlow and Keras, the model is optimized using Adam and evaluated using mean squared error. It is trained on a dataset comprising myocardium mask images from 128 patients, each represented by a sequence of 25 frames with a resolution of 80x80 pixels. The training process spans 60 epochs with a batch size of 30, including a validation split of 10% and a test split of 10%. This methodology capitalizes on the 3D CNN's capacity to interpret complex spatial-temporal relationships within the medical images, aiming for precise and reliable ToS prediction.

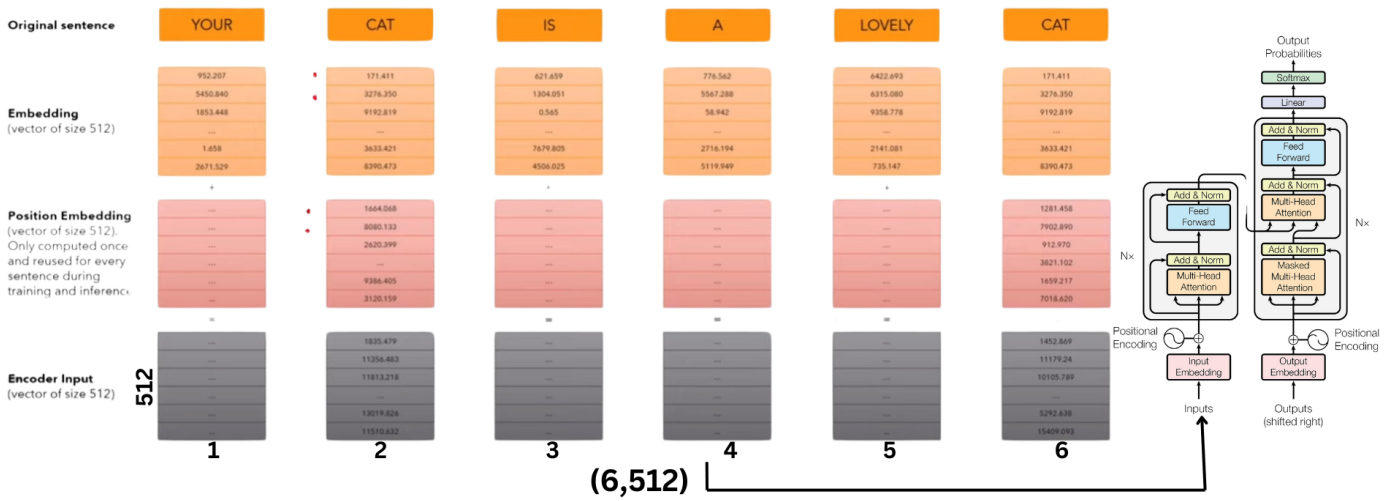

Figure 3: Transformers

3.3. Transformers

The Transformer architecture is divided into two main components: the Encoder Block on the left and the Decoder Block on the right, as illustrated in Figure 3. To understand how the Transformer operates, let's consider an example. Suppose we have a sentence with a length of six words: "Your cat is a lovely cat". This text undergoes a process known as word embedding, which converts each word into a numerical vector of a specific size, typically 512 dimensions. This conversion facilitates the comprehension of the text by the model.

After converting the words into embeddings, these are passed through sinusoidal functions to generate Positional Embeddings. Positional Embeddings are crucial as they provide information about the order or position of each word in the sentence. The equation for generating these embeddings using sinusoidal functions is as follows eq 1 & eq 2:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{1}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2}$$

where PE is the positional encoding, pos is the position of the word in the sentence, i is the dimension, and d_{model} is the dimension of the word embeddings.

The Positional Embedding and the word Embedding are then combined to create the Encoder Input. This combination ensures that the Encoder Input contains both the semantic information from the word embedding and the positional information from the positional embedding, providing a comprehensive representation of the original sentence for the Transformer model.

In the Transformer model, as illustrated in Figure 3, the Encoder Input is replicated into three matrices: Q (query), K (key), and V (values). It is important to note that Q, K, and V are identical copies of the Encoder Input. Based on our example sentence "Your cat is a lovely cat", the dimensions of these matrices are Q (6,512), K (6,512), and V (6,512).

Firstly, we transpose matrix K, resulting in K_{t} with dimensions (512,6). We then perform matrix multiplication between Q (6,512) and K_{t} (512,6), yielding a correlation matrix of dimensions (6,6). This correlation matrix encapsulates the relationship between each word in the sentence, as depicted in Figure 4.

Subsequently, this correlation matrix is multiplied with V (6,512) to generate the self-attention matrix, which retains the original dimensions of (6,512). This process can be encapsulated in the self-attention equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where d_{k} is the dimension of the key vectors. The softmax function is applied to the result of the matrix multiplication, normalizing the values.

The resulting attention embeddings now encompass not only the information of the individual words and their positions but also the contextual relationships among all the words. Following this, normalization and a feed-forward network is applied for further processing. A similar self-attention mechanism is employed in the Decoder block, where the Encoder and Decoder are interconnected via an attention mechanism. In this setup, K and V are derived from the Encoder, and Q comes from the Decoder. Another feed-forward layer is then applied to produce the desired output. Further explanation of Multihead and Masked Attention can be obtained from the original Attention paper (Vaswani et al., 2017).

This self-attention mechanism is pivotal in the Transformer architecture, enabling the model to capture complex relationships and dependencies in the input data, leading to more nuanced and contextually aware outputs.

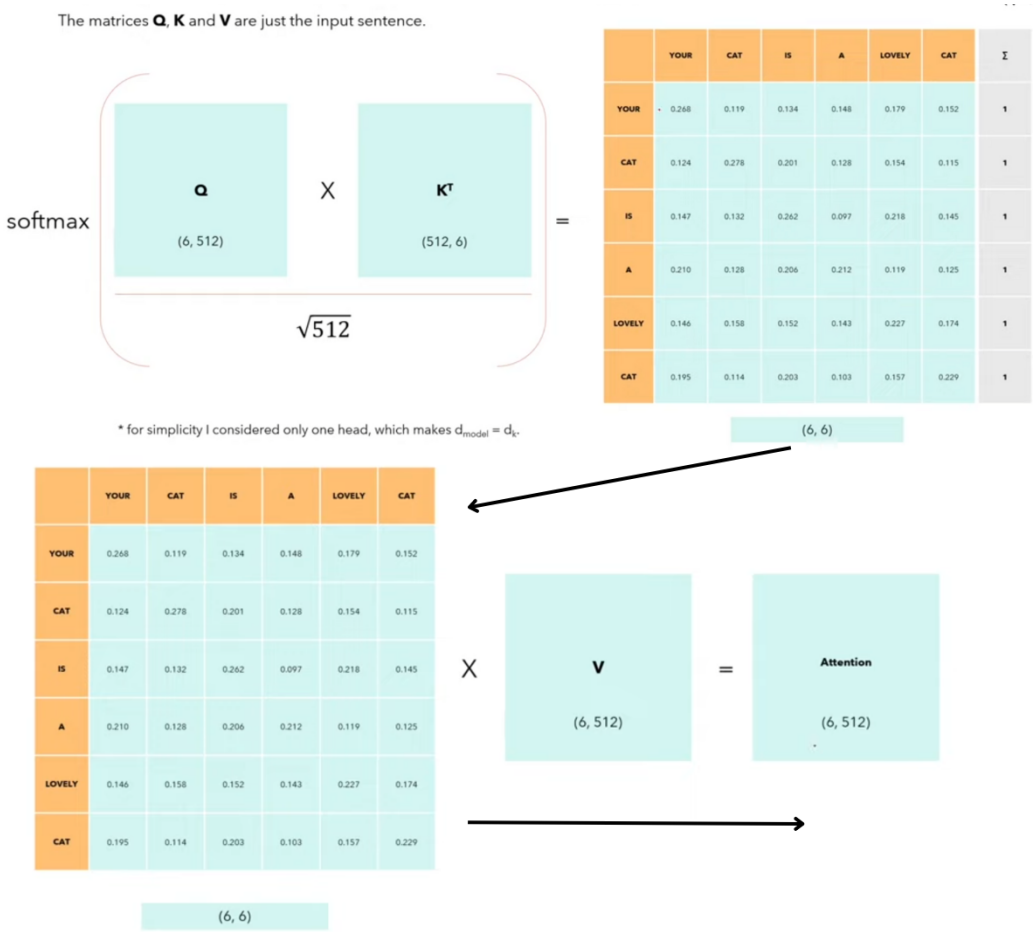

Figure 4: Attention Mechanism Visualized

3.4. VIT

The Vision Transformer (ViT) as shown in Figure 5 primarily aims to transform images into input embeddings of size 512. In their innovative approach, the authors of paper (Dosovitskiy et al., 2010) address this by dividing the image into multiple patches of size 16x16, each containing 3 color channels. These (16, 16, 3) patches are then flattened into 1D arrays, resulting in a shape of (768, 1). A crucial step involves the creation of learnable parameter vectors with dimensions (512, 768), which are initially randomly initialized. The Linear projection of these flattened patches is achieved by multiplying the learnable parameter matrix (512, 768) with the flattened patch array (768, 1), resulting in an output shape of (512, 1). This process effectively generates the input embeddings needed to feed into the transformer model.

Following this embedding process, the ViT applies the same principles of the Transformer model, utilizing an attention mechanism as discussed in the Transformer section. The architecture culminates in a Multilayer Perceptron (MLP) network, which is the final layer responsible for predicting the output class. This model showcases the adaptability of the Transformer architecture in handling image data, leveraging its capabilities for effective image classification or regression tasks.

However, this ViT architecture loses the capacity to handle temporal dependencies, i.e., processing multiple frames simultaneously. As a result, its performance on video datasets might not be as satisfying as that of 3D CNNs. So we have created new architecture ViT + Dense.

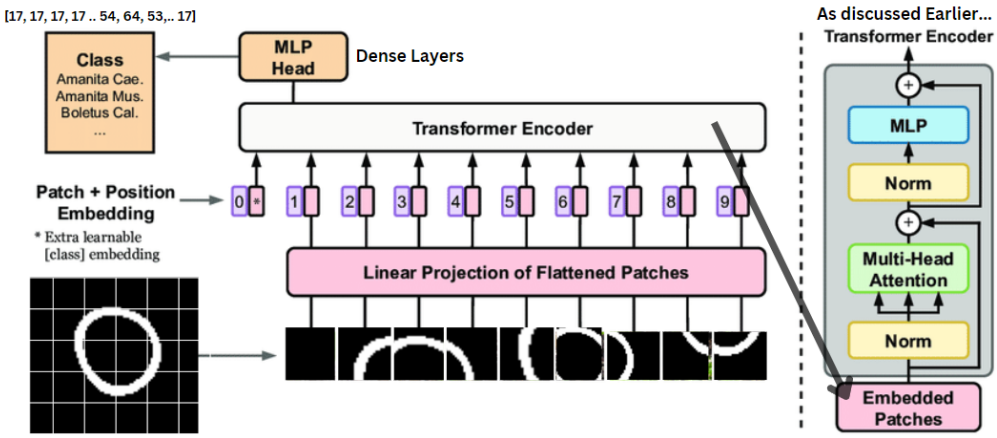

Figure 5: Vision Transformers ViT

3.5. ViT+Dense

The ViT+Dense architecture is an enhancement of the standard Vision Transformer (ViT), specifically designed to process sequential frame data effectively. This new architecture combines the spatial feature extraction capabilities of ViT with a Dense layer to capture temporal relationships across frames.

In our implementation for the myocardium mask dataset, each frame is first processed through the ViT model, generating individual frame embeddings. These embeddings are then aggregated and fed into a GlobalAveragePooling1D layer, which simplifies the data while retaining crucial temporal information. The final Dense layer maps these processed embeddings to our target output, the 126 TOS values.

This approach allows ViT+Dense to effectively analyze both the spatial details of individual frames and the temporal dynamics across a sequence, making it particularly suitable for complex sequential

data to perform image regression. However, the dense layer was not that capable of maintaining the temporal dependency so we built the ViT+LSTM model.

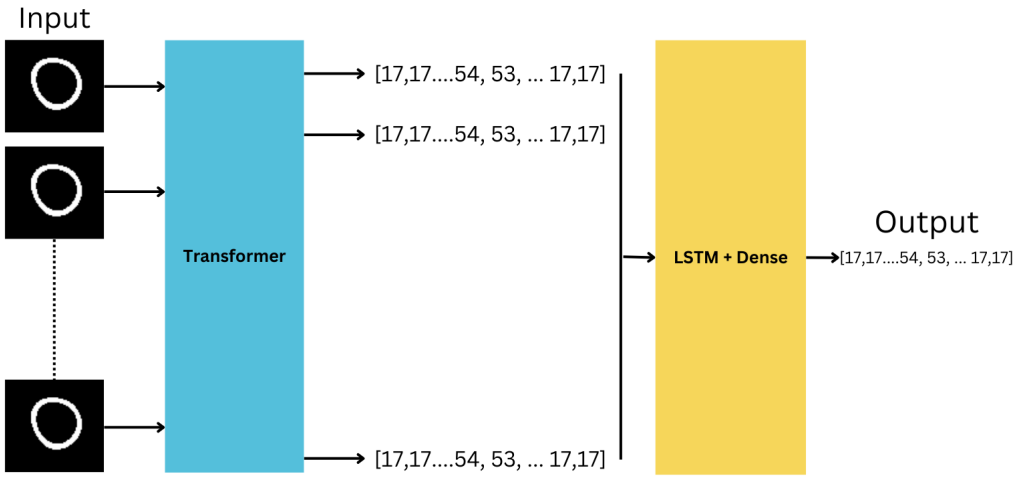

Figure 6: Vision Transformer + LSTM

3.6. Hybrid ViT+LSTM

Building on the ViT+Dense architecture, we developed the ViT+LSTM model to address the need for more sophisticated handling of temporal dependencies in sequential image data. While ViT+Dense was effective in analyzing spatial details and simple temporal dynamics, it could not fully capture complex temporal relationships. ViT+LSTM integrates the Vision Transformer (ViT) with Long Short-Term Memory (LSTM) networks, combining ViT's spatial feature extraction with LSTM's prowess in sequence modeling.

In this architecture as shown in Figure 6, each frame of the myocardium mask sequence is first processed through the ViT layer. This step involves reshaping and resizing the input frames to suit the ViT model, followed by extracting high-level spatial features. These features are then sequenced and fed into LSTM layers, specifically designed to capture and model temporal dependencies and dynamics across frames.

The ViT+LSTM model consists of two LSTM layers with 128 units each. Dropout layers are included for regularization, reducing the risk of overfitting. This sequential processing through LSTMs allows the model to learn from not only the individual frame's features but also the temporal relationships inherent in the sequence of frames. The output from the LSTM layers is passed through Dense layers, culminating in a final regression output that predicts the 126 ToS values.

By leveraging both ViT for its efficient image processing and LSTM for its sequential data handling capabilities, ViT+LSTM offers a powerful solution for tasks requiring intricate understanding of both spatial and temporal aspects of image sequences. This makes it particularly suitable for complex medical imaging tasks, like those in our dataset, where understanding the progression and changes across frames is vital.

In our implementation, we have observed that ViT+LSTM provides a more nuanced understanding of the temporal dynamics in the myocardium mask sequences, resulting in more accurate and reliable predictions of TOS values. This model, therefore, represents a significant step forward in medical image analysis, particularly for datasets involving sequential frames.

3.7. Hybrid 3D CNN and Vision Transformer Model

To effectively capture both spatial and temporal information from medical imaging data, we have developed a novel hybrid model that combines the strengths of 3D Convolutional Neural Networks (3D CNN) and Vision Transformers (ViT). The 3D CNN component is specifically designed to decompose each frame into 16x16 vectors. This decomposition process plays a pivotal role in generating a comprehensive latent vector, encompassing the spatial features extracted from all frames.

Subsequently, this latent vector is fed into the Vision Transformer. The ViT, equipped with positional embeddings, is adept at handling the temporal dependencies inherent in the sequence of frames. Its attention mechanism further enhances the model's capability, allowing it to selectively focus on specific frames that are crucial for accurate analysis. This strategic focus is particularly beneficial in the context of predicting the ToS values, as it ensures that the most relevant temporal features are emphasized during the prediction process.

By integrating the spatial processing power of 3D CNNs with the temporal analysis strengths of ViT, our hybrid model promises to significantly improve the accuracy of ToS predictions in cardiac imaging, addressing both the spatial and temporal complexities of the data.

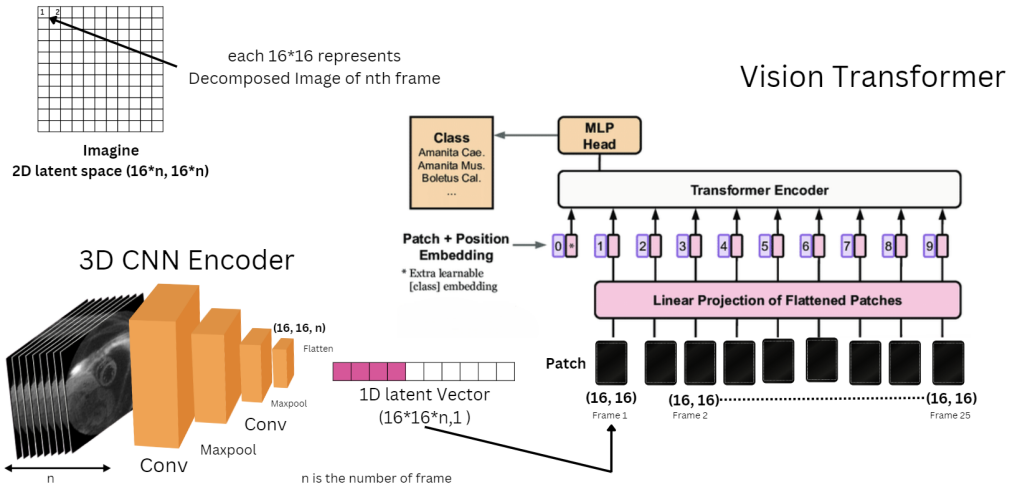

Figure 7: Vision Transformers ViT

Note: We are just discussing the architecture framework of the Hybrid 3D CNN + ViT model. This is not part of the project, as it may require more time for experiments. We thought we would share our approach, so we have not shown any results for this model yet. Thanks.

4. Experiment (Rugved Chavan)

Our experimental analysis involved evaluating three distinct models: a 3D Convolutional Neural Network (3DCNN), a Vision Transformer coupled with a Dense network (ViT+Dense), and a Vision Transformer integrated with a Long Short-Term Memory network (ViT+LSTM). Each model was tasked with predicting Target Organ Status (TOS) values from myocardium mask sequences.

Table 1 presents the Mean Absolute Error (MAE) on test data for each model. The ViT+LSTM model demonstrated superior performance with the lowest MAE of 8.1491, suggesting its enhanced ability to model the spatial-temporal dynamics of the cardiac sequences. The ViT+Dense model followed closely with an MAE of 8.213, while the 3D-CNN model registered an MAE of 8.3106.

The learning curves reveal distinct behaviors for each model as shown in figure 8. The 3D CNN (first graph) showed a rapid early reduction in both training and validation MAE, stabilizing quickly.

Model	MAE on test data
3DCNN	8.3106
ViT + Dense	8.213
ViT + LSTM	8.1491

Table 1: Comparison of model performance based on MAE on test data.

The ViT+Dense model (second graph) exhibited a continuous decline in MAE, with the validation curve closely tracking the training curve, indicating good generalization. Lastly, the ViT+LSTM model (third graph) not only achieved the best MAE but also displayed a significant reduction in validation loss, underscoring its capability to capture complex temporal dependencies more effectively than its counterparts.

These results highlight the potential of Transformer-based architectures, particularly when combined with LSTM networks, in providing robust analytical capabilities for cardiac image analysis, surpassing traditional 3D CNNs in predictive performance.

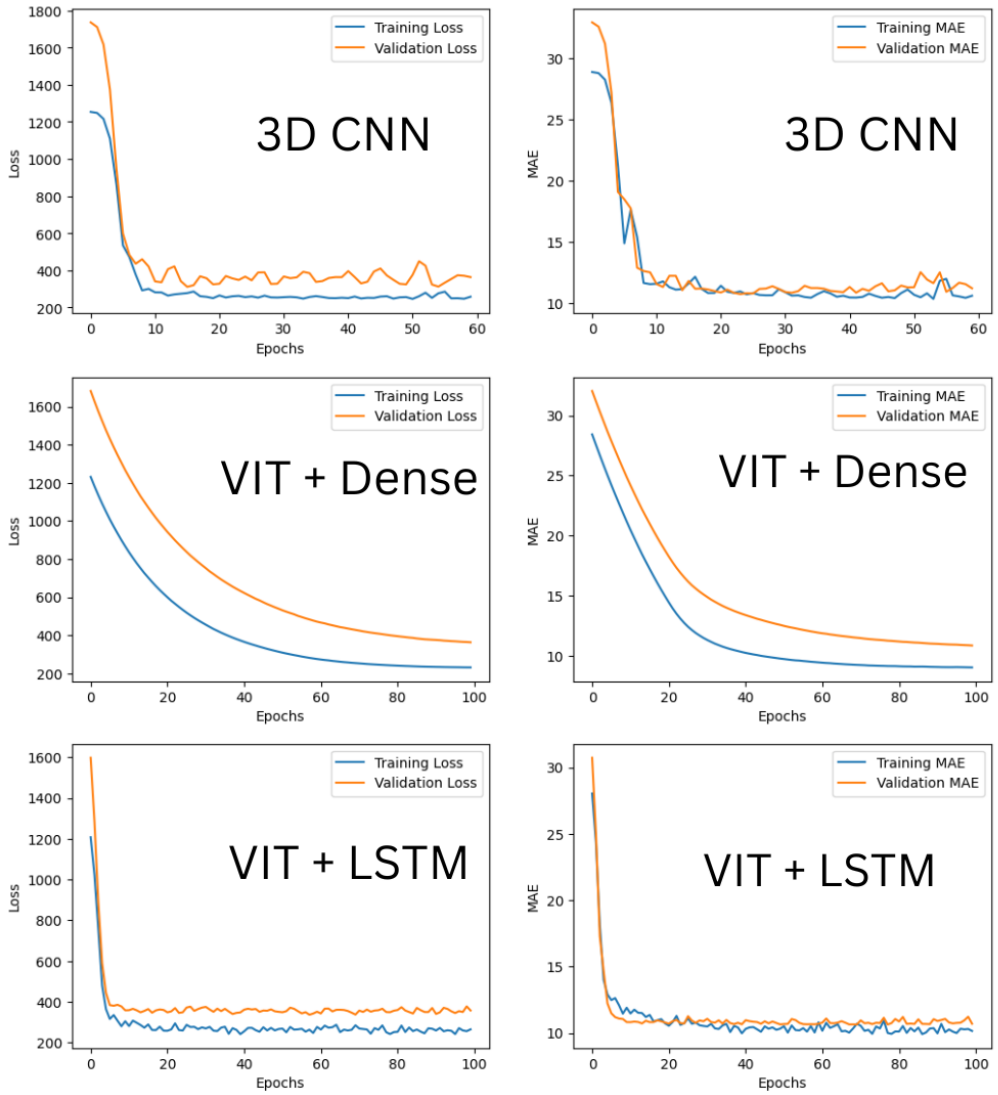

Figure 8: Vision Transformers ViT

5. Discussion (Ishita Gupta)

The results from the three models tested—3D CNN, ViT+Dense, and ViT+LSTM—indicate clear distinctions in the prediction of Target Organ Status (TOS) values in myocardium mask sequences. The ViT+LSTM model demonstrated the lowest MAE on test data, followed by ViT+Dense and 3D CNN, suggesting an improved ability to capture the intricate spatial-temporal features within the cardiac imaging data.

In the comparative graphs as shown in Figure 9, the ViT+LSTM and ViT+Dense models display a closer alignment with the actual TOS values than the 3D CNN, as evidenced by the overlap between the predicted and actual data points. The improved performance of the ViT-based models could be attributed to their ability to better leverage spatial dependencies within the image patches. Furthermore, the integration of LSTM with ViT appears to have provided an advantage in capturing temporal relationships, which is a critical aspect of myocardium mask sequence analysis.

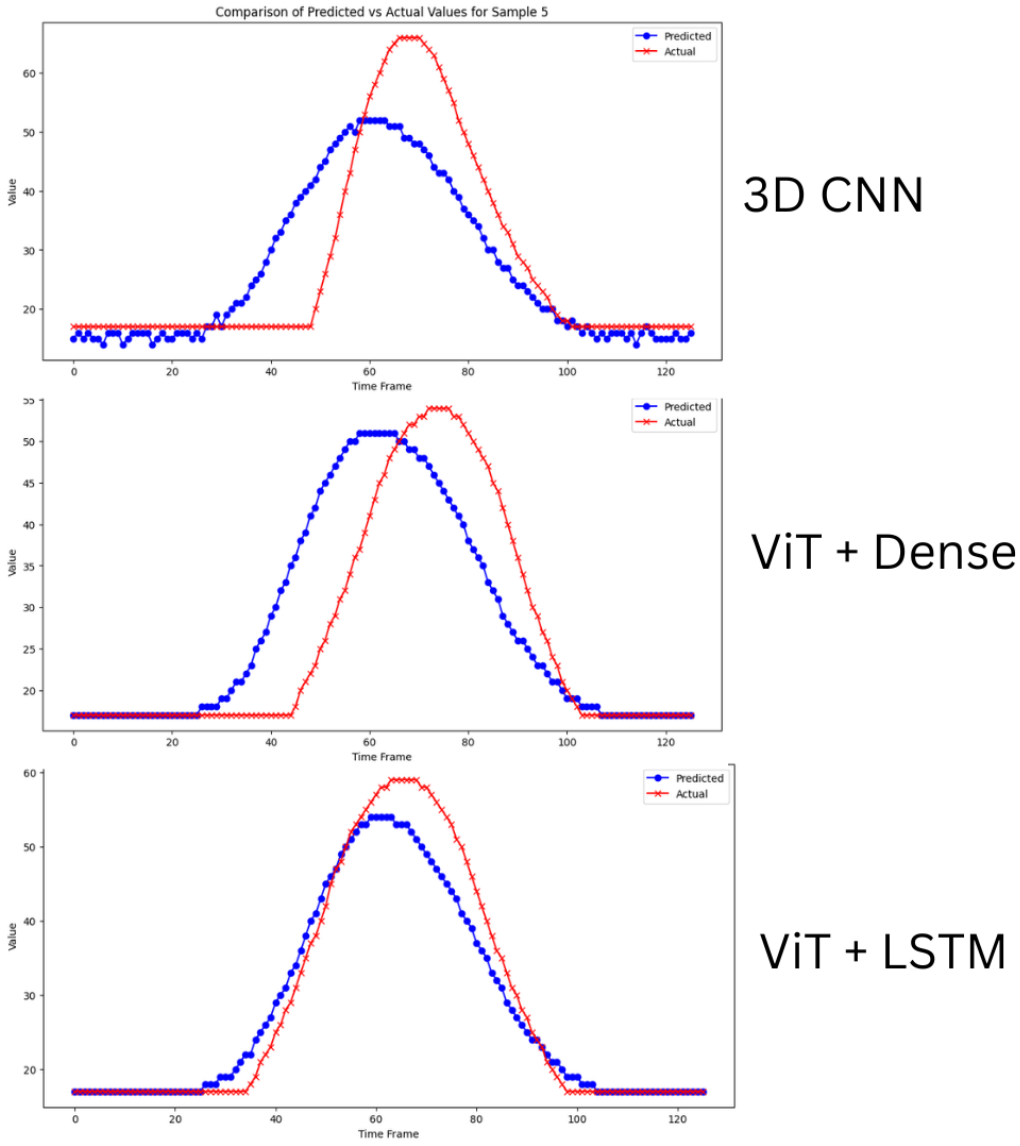

Figure 9: Model comparison with the ground truth

The third graph in the Figure 9 shows that the ViT+LSTM model closely matches the actual TOS values across the sequence, which indicates its strong temporal modeling capabilities. This outcome

supports the hypothesis that for complex sequential image data, the combination of Transformer and LSTM networks can effectively enhance predictive accuracy.

These findings strengthen the argument for adopting advanced Transformer architectures in medical image analysis, particularly when temporal dynamics play a crucial role, as is the case in cardiac cycle imaging. The implications of this research could extend to the development of more nuanced models that are capable of handling a wide array of medical imaging tasks, potentially leading to better diagnostic tools and patient outcomes.

6. Conclusion (Ishita Gupta)

This study successfully demonstrates the potential of Vision Transformer (ViT) architectures in the field of cardiac image analysis. We have shown that by adapting ViT models, which are traditionally used in natural language processing, we can achieve notable improvements in predicting ToS values from myocardium mask sequences. Our ViT+LSTM model outperformed the traditional 3D CNN approach, yielding the lowest Mean Absolute Error (MAE) of 8.1491 on the test data, signifying a substantial advancement in the accuracy of cardiac imaging predictions.

The efficacy of ViT models, particularly when combined with LSTM networks, suggests a promising direction for future research. The integration of LSTM appears to be crucial in capturing the temporal dynamics of the cardiac sequences, which is an area where standard CNNs have limitations. The dedication and hard work invested in fine-tuning these models are evident in the close alignment of the predicted TOS values with the actual values, especially in the ViT+LSTM architecture.

For future work, we aim to explore the Hybrid 3DCNN + VIT with integration of additional context-aware layers and attention mechanisms that may further enhance the model's sensitivity to the subtle nuances of temporal sequences in cardiac images. The implementation of such improvements could lead to even more accurate diagnostic tools, helping clinicians to provide timely and precise treatment for cardiac patients.

Note: Although other models like Video Vision Transformer (ViViT) or 3D Vision Transformers may give a better performance by yielding lower Mean Absolute Error (MAE) values, our Team was tasked with implementing the raw architecture of ViT and making modifications at a fundamental level.

References

Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Bradley J Erickson, Suraj Kapa, and Paul A Friedman. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, Sep 7 2019. doi: 10.1016/S0140-6736(19)31721-0. Epub 2019 Aug 1.

Mohammad Reza Avendi, Arash Kheradvar, and Hamid Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical Image Analysis*, 30:108–119, May 2016. doi: 10.1016/j.media.2016.01.005. Epub 2016 Feb 6.

Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Valentina Carapella, Young Jin Kim, Hideaki Suzuki, Bernhard Kainz, Paul M. Matthews, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, Ben Glocker, and Daniel Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional

networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65, Sep 14 2018. ISSN 1532-429X. doi: 10.1186/s12968-018-0471-x. URL https://doi.org/10.1186/s12968-018-0471-x.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, and Julien Mairal. Data-efficient image transformer. *arXiv preprint arXiv:2121.02141*, 2021.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers, 04 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

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 2020. *arXiv preprint arXiv:2010.11929*, 2010.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H. Chen, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Deep learning interpretation of echocardiograms. *npj Digital Medicine*, 3(1):10, Jan 24 2020. ISSN 2398-6352. doi: 10.1038/s41746-019-0216-8. URL https://doi.org/10.1038/s41746-019-0216-8. PMID: 31993542.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Y LeCun, B Boser, J Denker, D Henderson, R Howard, W Hubbard, and L Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

Bejnordi BE Setio AAA Ciompi F Ghafoorian M van der Laak JAWM van Ginneken B Sánchez CI Litjens G, Kooi T. A survey on deep learning in medical image analysis. 2017.

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, 2021.

Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine*, 1(1):6, Mar 21 2018. ISSN 2398-6352. doi: 10.1038/s41746-017-0013-1. URL https://doi.org/10.1038/s41746-017-0013-1.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

Taki Hasan Rafi and Young Woong Ko. Heartnet: Self multihead attention mechanism via convolutional network with adversarial data synthesis for ecg-based arrhythmia classification. *IEEE Access*, 10:100501–100512, 2022. doi: 10.1109/ACCESS.2022.3206431.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation, 2016.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, October 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Lu Yuan, Yujun Chen, Tao Wang, Weihao Yu, and Yugen Chen. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2106.06729*, 2021.

Jie Zhang, Bowen Zheng, Ang Gao, Xin Feng, Dong Liang, and Xiaojing Long. A 3d densely connected convolution neural network with connection-wise attention mechanism for alzheimer's disease classification. *Magnetic Resonance Imaging*, 78:119–126, 2021. ISSN 0730-725X. doi: https://doi.org/10.1016/j.mri.2021.02.001. URL https://www.sciencedirect.com/science/article/pii/S0730725X21000138.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016.