COMPARISON OF RIB FRACTURE DETECTION AND SEGMENTATION PERFORMANCE BETWEEN 3D AND 2D DEEP LEARNING NETWORK ARCHITECTURES

Dieko de Graaf[1], Jochem van Oorschot[1], Guilly Kolkman[1], Alex Labro[1], Aries van der Werf[2]

[1]Master Artificial Intelligence, University of Amsterdam, Amsterdam, The Netherlands

[2]Master Artificial Intelligence, Vrije Universiteit, Amsterdam, The Netherlands

ABSTRACT

The most common type of injuries following chest trauma are rib fractures. 3D Computed tomography scans are used by radiologists and clinicians to diagnose the presence and type of fracture. Deep Learning segmentation models such as convolutional neural networks (CNN) and Transformers have been proven to approach medical experts in their ability to diagnose rib fractures. However, high-resolution 3D images require large storage databases and adequate computing power. We attempt a 2D approach to rib fracture segmentation, with the purpose of studying the potential of 2D rib fracture diagnoses. In contrast to 3D images, the storage and computation of 2D images can be less costly. We compare the performance of a 2D UNet CNN and a 2D VitSeg transformer model to the state-of-the-art performance of the 3D FracNet segmentation model.

1. INTRODUCTION

The most often occurring form of trauma seen in hospitals in the United States is thoracic trauma, comprising 10%-15% of all trauma-related hospital admissions. Moreover, it is responsible for approximately 35% of all trauma-related deaths, making it one of the most common reasons for trauma-related mortality [1]. Rib fractures are the most frequently occurring injuries following thoracic trauma. Additionally, rib fractures are highly clinically relevant, because they are causes of long-term morbidity of pain, mortality risk, and an overall significant decrease in quality of life [2].

It has become standard practice to perform whole upper-body computed tomography (CT) scans [3] on subjects to diagnose the presence of rib fractures. This provides a challenge to radiologists as they have to analyze full upper-body 3D images of their patients to determine potential rib fractures [4]. Analyzing full upper-body CT scans is a very time and labor-intensive task, which makes it more likely for (small) fractures to be missed. Especially in cases of multiple injuries or minor injuries, clinicians are more likely to not detect them on whole upper-body CT scans [5]. Moreover, the variability in the assessment of CT scans between radiologists and expert clinicians can be high. The application of artificial intelligence (AI) technologies, and specifically that of deep learning (DL) and convolutional neural networks (CNNs), has been shown to hold much promise to assist medical experts with overcoming the aforementioned challenges [6].

In recent years more and more studies are being performed to assess the effectiveness of DL applications for medical computer vision tasks. CNNs offer significant advantages in the context of medical rib fracture segmentation. Their hierarchical feature extraction capabilities enable the precise delineation of fractured rib structures. Their scalability allows for the analysis of a large volume of medical imaging data, which is vital in cases of multiple rib fractures and when dealing with a large number of patients. By consistently providing early fracture detection, CNNs can aid in more timely interventions and improved patient outcomes, ultimately contributing to cost savings and optimized healthcare delivery [7]. Multiple recent studies have documented that CNN-based model architectures can outperform or at least match the rib fracture detection performance of expert clinicians [8]. Moreover, in all cases the models can cover full upper-body CT scans significantly faster than a human could.

In addition to CNNs, the introduction of attention mechanisms has given rise to the so-called vision transformer (ViT) [9]. ViTs have emerged as powerful tools for medical image segmentation, offering a range of advantages. ViTs excel in understanding global contextual information in images, making them well-suited for comprehending the relationships between structures in medical images. Their attention mechanisms allow them to focus on relevant regions while diminishing noise and distractions. They can capture the spatial relationships between distant pixels, which is a highly beneficial trait for segmenting complex medical structures [10]. Moreover, ViTs are generally more interpretable than CNNs, due to the ability to visualize the attention maps that the network uses, which adds to their utility in diverse medical imaging scenarios.

One of the more recent state of the art models in the field of rib fracture detection and segmentation is FracNet [11]. FracNet is a network architecture based on the original U-Net auto-encoder model, which makes use of CNNs in multiple

image encoding and decoding layers [12]. FracNet has been proven to have significantly increased performance over medical experts, in both fracture detection and in rib fracture image segmentation. Moreover, the model significantly reduced the time for CT scan inference compared to the human benchmark. Yet, medical experts still showcased fewer false positive predictions compared to FracNet. Moreover, similarly to radiologists, FracNet uses the memory heavy full upper-body 3D CT scans for training and inference.

We propose a study to compare the performance of the 3D image based FracNet to simpler and less memory intensive 2D deep learning models. For the implementation of rib fracture detection and segmentation models in the hospital, it can be of value to know the potential benefits and performance differences between 3D and 2D-based models. The trade-off between lower memory requirements, due to smaller image storage size requirements, and the inference performance is certainly of interest for future model deployment. If 2D models can approach the performance of the state of the art, these smaller models could be deployed on smaller devices with lower memory capabilities. To gather insight into this trade-off, we compare the performance of the 3D FracNet with a 2D U-Net and a 2D ViT segmentation model.

2. METHOD

2.1. Data Description

To train and test our models we have gained access to the dataset of the original MICCAI 2020 RibFrac Challenge [13]. The RibFrac Challenge was held to evaluate the capabilities of submitted models for both rib fracture detection, including the segmentation of rib fracture images, and the classification of rib fractures into 4 separate types. For the purpose of our research, we focused purely on the rib fracture detection and segmentation, foregoing the classification of the rib fractures.

The dataset consists of 660 full upper-body 3D CT scans with around 5000 rib fractures in total. This dataset was split into a training set of 420 CT scans, all containing (multiple) rib fractures, 80 validation scans of which 20 did not contain any rib fractures, and 160 scans in the final test set. For the training and validation sets the full annotations for the 3D CT scans are included, which makes it possible to evaluate the detection and segmentation performance. The final test set annotations were not released. Test set evaluation is only possible through the online platform of the MICCAI 2020 RibFrac Challenge. All annotations were provided by expert radiologists.

2.2. Pre-processing

2.2.1. FracNet

A novel idea introduced in FracNet is part of the pre-processing pipeline for the CT scans. Firstly, the bone regions are extracted from the images by filtering the intensity of the voxels of the 3D input image at the bone level. This makes the ribs and spine clearly stand out in white on the image, while other elements of the body are set to the black color of the background. Secondly, it is important to note that the 3D datasets consist of very sparse rib fracture data. Slices in the 3D axes (axial, coronal, sagittal) of the CT scan will only hold a very limited amount of rib fracture, especially if full upper-body images are used. Most of the annotations will be negative for fractures, which leads to an excessively sparse data set in 3D space. For the purpose of creating a less sparse and more balanced data set during model training, FracNet directly centers a 96x96x96 window around the bone fracture(s), by looking up annotations during training. From this window, smaller patches of 64x64x64 are sampled. Negative samples are extracted from the mirrored rib from which positive patches were samples. In our implementation, the pre-processing pipeline sampled 4 positive and 4 negative samples for each full body CT scan. Resulting in a slightly less sparse and more balanced subset for each image.

2.2.2. 2D approach

The pre-processing for 2D architectures is similar to the pre-processing for 3D. For each fracture the centroid is calculated and used as a basis to create 64x64 patches. However, the fractures in the dataset are in 3D. Therefore the z-axis is used with a top-down approach. The first slice is at the top of the fracture in the z-axis and moves down towards the bottom of the fracture. Each slice is then 96x96 based on the centroid and a random 64x64 patch is taken from that slice to create the dataset. The negative dataset is the mirrored version of the 2D 64x64 slice if there are no fractures on the mirrored slice. This ensures the order of the fractures stays intact for testing purposes. The 2D pre-processing method is used for the UNet approach and ViT architectures.

2.2.3. ViT

ViT's strength comes from its ability to embed image patches with a linear projection turning them into tokens. These tokens are passed into a transformer encoder in sequence followed up by an MLP for class prediction, similar to how a sentence can be encoded. In our implementation for the rib fractures the patches are replaced by slices of a rib cage from a top-down view as explained in 2D approach. ViT is especially suited for this task because it takes in a sequence, which in this case becomes a list of slices of the rib cage from top to bottom. This particular task however proves to be calling for the model due to the sparsity of the data. We have explored methods to address this issue, such as only selecting slices with fracture data above a certain threshold. However, we have not extensively pursued this option too much, as it is an intermediate step.

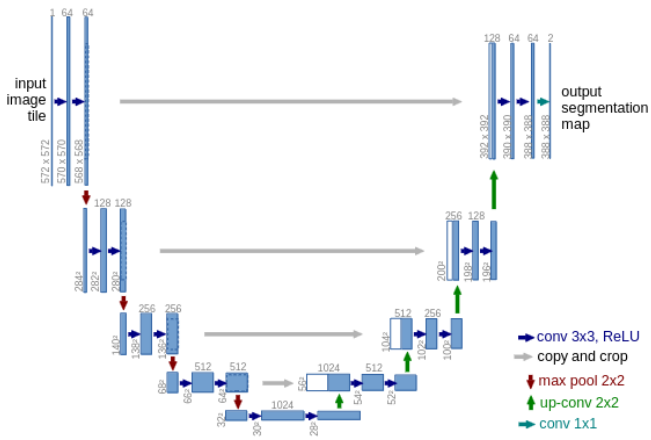

Fig. 1: 2D UNet architecture.

Classification is not the main goal of this paper as it is not enough to say a person has a fracture. Therefore this model will serve as the foundation for ViT Segmentation, which will extract more detail from the fractures such as size and location.

2.2.4. ViT Segmentation

Because our data was quite different from the regular model input it required an overhaul of the first few layers such as the input, attention, and output layer which became an infeasible amount of work. Due to the limited timeframe, we decided to model the data after the images expected by the model. This meant creating a 128x128 image from our 64x64 patches. This can be done by stitching four patches and their corresponding masks together in the x and y direction resulting in the required size. Per fracture, the four best, defined as containing the most amount of fracture pixels, were selected to present the model with as much positive data as possible. This approach helped in combatting the sparsity of the data as well. ViT segmentation also uses considerably more attention layers which should allow the model to learn the more common areas for fractures as well as what areas, such as the spine and the corners of the image outside of the body, can be avoided. It is important to note that we use 2600 samples here to speed up training times and avoid large chunks of data in the pre-loading phase.

2.3. Architecture

2.3.1. 2D UNet

The UNet architecture[12] is a 2D segmentation architecture that is widely used for segmenting medical imaging. It consists of a CNN-based encoder-decoder architecture, with residual connections between the encoding and decoding paths as shown in figure 1.

The encoding path is made up of 4 modules of two 3x3 convolutional layers followed by a ReLU and a 2x2 max pooling layer, where the convolutions in the encoding path each double the number of feature channels. The decoding path consists of similar modules, however during decoding each convolution halves the number of feature channels. The decoding path also includes concatenating the encoded feature maps to the decoding input at every module. The output layer is a convolutional layer that maps the feature maps to the number of classes.

Like the original paper, we used a cross-entropy loss with a high (0.99) momentum. The model was run for 100 epochs on a batch size of 16 with the learning rate initialized at 0.0001, which was reduced by 0.1 if the model did not improve after 5 epochs. Inputs consisted of preprocessed grayscale 64x64 images, outputs consisted of segmentation maps of the same size where 0 is background and 1 is a fracture.

2.3.2. FracNet

The network architecture of FracNet is based on a 3D U-Net framework, visualized in Figure 2. The encoder part of the network consists of 3 stages of 3D convolutions, batch normalization, ReLU non-linearity, and max-pooling layers. Gradually reducing the resolution of the feature maps before they are again restored in the decoding layers. A final layer with 1x1x1 convolutions and sigmoid non-linearity is used to create pixel-wise probabilities for fractures. Similarly to the original FracNet implementation, a decision threshold of 0.1 was used to binarize the fracture predictions.

Additionally, a post-processing step is applied after a patch has passed through FracNet, with the purpose of reducing the number of false positive fraction predictions. Positive predictions of fractures that are smaller than 200 voxels are automatically removed. Fractures that are larger than 200 connected voxels are considered as valid predictions. The final fracture-wise probability is computed by averaging the pixel-wise probabilities of all connected voxels in the fracture.

During model inference the full upper-body 3D scans are also divided up into patches of 64x64x64, similarly to the training pre-processing, before they are sent through the network. Once all patches have travelled through FracNet all the valid fracture predictions are outputted for the full 3D image. The detected predictions can then be used to test the accuracy of the detection and segmentation alignment for each CT scan.

2.3.3. ViTSeg

ViTSeg differs primarily from the classical ViT architecture in the input and output layers. The input takes in additional positional encoding to accommodate for segmentation tasks

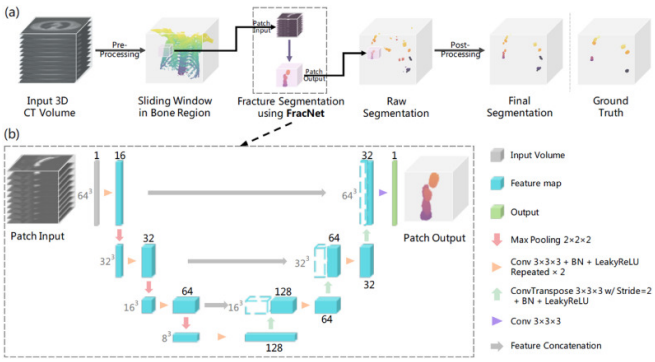

Fig. 2: (a) Showing the FracNet pipeline, (b) Showing the U-Net based architecture of FracNet [11]

that require more pixel data. Similarly, the output layer is modified to return per-pixel class probabilities which can be further processed into a segmentation mask. The MLP head for classification has been substituted by an MLP for pixel-level classification. It then goes through multiple attention layers, as seen in 3. The model uses 768 embedding dimensions for the PatchEmbedding layer, 12 encoder blocks with 8 attention heads, has a 20% dropout rate, and uses a cross-entropy loss.

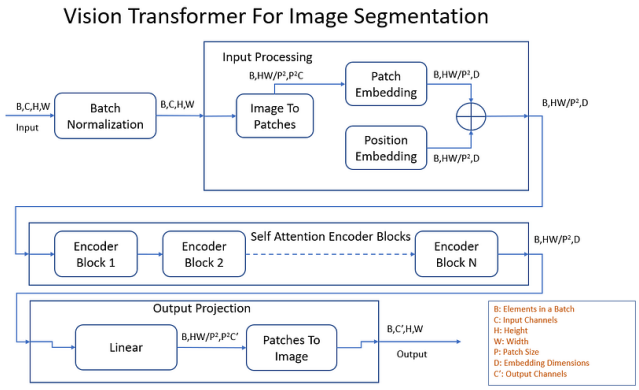

Fig. 3: ViTSeg architecture in detail. Source: Dhruv Matani

2.4. Metrics

We will evaluate the proposed models based on their performance on a classification task, which is also the main experiment of this technical report.

The different models are evaluated and trained by making use of a variety of metrics. The classification task consists of predicting whether a pixel in a 64x64 pre-processed input image belongs to a fracture, which can be referred to as pixel-wise binary classification. The F1-score and sensitivity are used in order to quantify the classification quality of the models. The sensitivity, also known as recall or true positive rate, is of high importance in the fracture detection task due to the significant implications of failing to find a fracture. The F1-score combines the recall together with the precision of the model. This gives an indication of the model being able to retrieve the fracture in an image while correctly differentiating between the intact and fractured parts of the rib.

The models learn to localize a fracture by including the Dice-coefficient (DSC) as an optimization during training. The Dice-coefficient is a metric for comparing the predicted fracture region to a mask of the ground truth region. The coefficient increases as the overlap between the fracture regions in the two masks increases. Training the models based on the segmentation score naturally improves the binary pixel-wise classification. Besides the Dice-coefficient, the Intersection-over-Union (IoU) is calculated. It positively correlates with the Dice-coefficient and is included in the analysis as existing fracture segmentation architectures report IoU scores.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \quad \text{and} \quad IoU = \frac{|X \cap Y|}{|X \cup Y|}$$

.

3. RESULTS

The original ViT results are not mentioned here as they are considered intermediate results for ViTSeg and their inclusion could lead to confusion due to their lack of comparability with the other results.

Some things to mention in the comparison between ViTSeg and FracNet. The comparison is not entirely fair due to Fracnet performing its validations on an entire rib cage. FracNet on the other hand uses individual images which explains the recall being lower. As a result, the lower recall in ViTSeg can be explained by this difference. Additionally, fractures typically have very little data at the top, which poses a more significant challenge for models that only focus on this area compared to models that consider the entire image.

The Dice of FracNet is higher than ViTSeg by 0.6, while the F1 score of ViTSeg is notably higher than FracNet by 0.21. This difference is likely due to the true positive rate of the models. Again, we have to consider the impact of validating over the entire rib cage contrary to the individual samples. Validating over the rib cage tends to result in a higher number of false positives, which could explain this difference.

ViTSeg's precision is significantly higher than that of FracNet and this difference suggests that ViTSeg maintains a lower false positive rate, indicating a more cautious approach when predicting positive classes.

4. CONCLUSION

In this technical report, we compared the performance of a 3D image segmentation model with a set of 2D medical image segmentation models with the intention of creating a model that could perform similarly to the winner of the MICCAI

	Dice	IoU	Recall	Precision	F1
UNet	0	0	0	0	0
VitSeg	0.63	0.45	0.56	0.72	0.63
FracNet	0.69	0.48	0.85	0.28	0.42

Table 1: Validation performance

2020 RibFrac challenge. The 2D approaches of UNet and VitSeg showed very different results, while the 3D approach of FracNet showed results consistent with the original implementation. UNet was not able to produce any meaningful output due to the difficulty of this task in 2D, leading to an IoU and Dice of 0. VitSeg on the other hand showed promising results, performing comparably to the winner of the RibFrac challenge with a Dice and IoU of 0.63 and 0.45 respectively. Our results demonstrate the potential of Transformer-based architectures within the field of medical imaging as well as the importance of adapting models accordingly depending on the task. Due to the limited time and resources, we were not able to thoroughly investigate how to improve the performance of our 2D models, which led to a low model performance for UNet.

5. DISCUSSION

Our UNet implementation drastically underperformed, which is something we did not expect. Unet is a relatively small architecture in comparison to Fracnet and ViT Segmentation. This resulted in the architecture not being able to parse the size of the dataset without crashing. Unet is originally proposed in a 2d, less sparse environment, with adequate results. In contrast, the RibFrac dataset is a high-sparsity dataset wherein a substantial amount of pixels are not annotated with a label. Thus, a lot of pixels are labeled with a zero value.

The loss used for our Unet implementation failed to take the amount of zero-labeled pixels into account. This resulted in the prediction being zero values, which were not penalized within the loss function. A future solution could have been a more in-depth look into the sparsity of the RibFrac dataset and options to solve said sparsity.

Our Vitseg results use the same metrics as FracNet. However, it is evaluated on data from the pre-processing. Sampling from the dataset with an equal amount of 2D positive and negative samples. On the contrary, FracNet evaluates on complete CT scans with a sliding window approach to sample from the upper body scan. Thus, our evaluation is on different datasets, 2D and 3D, which makes the comparison between architectures unreliable. It does show, despite our metrics being comparatively lower, our computational efficiency surpasses the 3D method.

Table 1 shows that our ViTSeg implementation has a low recall while achieving a high precision. High precision indicates that the predictions include a low number of False Positives. Combining this result with the fact that the model has a low sensitivity, indicates that the ViTSeg model is conservative in attributing a pixel to the fracture segmentation.

As stated in section 3, the segmentation quality is similar to that of FracNet. This is in contrast with our expectations since the 2D input images contain fracture regions of lower complexity compared to a 3D fracture region. Therefore, the ViTSeg model underperforms in its segmentation ability.

The low recall in combination with the underperforming segmentation scores show that the model is able to locate the fracture but fails to identify a significant number of pixels belonging to the fracture class. This has several possible explanations. The model being trained on a reduced number of training images, potentially comes at the cost of its ability to recognize the fractures' shapes. Additionally, the combination of a high precision and a low recall corresponds to an excessive preference for predicting the background class. To mediate these discrepancies, our approach should be fine-tuned such that the models and their parameters are optimized with the goal of achieving a higher recall.

6. REFERENCES

[1] J. Peek, Y. Ochen, N. Saillant, R. H. H. Groenwold, L. P. H. Leenen, T. Uribe-Leitz, R. M. Houwert, and M. Heng, "Traumatic rib fractures: a marker of severe injury. A nationwide study using the National Trauma Data Bank," *Trauma Surgery & Acute Care Open*, vol. 5, no. 1, p. e000441, Jun. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7292040/

[2] S. Marasco, G. Lee, R. Summerhayes, M. Fitzgerald, and M. Bailey, "Quality of life after major trauma with multiple rib fractures," *Injury*, vol. 46, no. 1, pp. 61–65, Jan. 2015.

[3] "Computed Tomography (CT)." [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct

[4] A. Blum, R. Gillet, A. Urbaneja, and P. Gondim Teixeira, "Automatic detection of rib fractures: Are we there yet?" *EBioMedicine*, vol. 63, p. 103158, Dec. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7718439/

[5] N. Banaste, B. Caurier, F. Bratan, J.-F. Bergerot, V. Thomson, and I. Millet, "Whole-Body CT in Patients with Multiple Traumas: Factors Leading to Missed Injury," *Radiology*, vol. 289, no. 2, pp. 374–383, Nov. 2018.

[6] P. H. S. Kalmet, S. Sanduleanu, S. Primakov, G. Wu, A. Jochems, T. Refaee, A. Ibrahim, L. v. Hulst, P. Lambin, and M. Poeze, "Deep learning in fracture detection:

a narrative review," *Acta Orthopaedica*, vol. 91, no. 2, p. 215–220, Jan. 2020.

[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, p. 221–248, Jun. 2017.

[8] L. H. M. Dankelman, S. Schilstra, F. F. A. IJpma, Doornberg *et al.*, "Artificial intelligence fracture recognition on computed tomography: review of literature and recommendations," *European Journal of Trauma and Emergency Surgery*, vol. 49, no. 2, p. 681–691, Apr. 2023.

[9] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," no. arXiv:2012.15840, Jul. 2021, arXiv:2012.15840 [cs]. [Online]. Available: http://arxiv.org/abs/2012.15840

[10] B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, and Y. Liu, "Segvit: Semantic segmentation with plain vision transformers," no. arXiv:2210.05844, Dec. 2022, arXiv:2210.05844 [cs]. [Online]. Available: http://arxiv.org/abs/2210.05844

[11] L. Jin, J. Yang, K. Kuang, B. Ni, Y. Gao, Y. Sun, P. Gao, W. Ma, M. Tan, H. Kang, J. Chen, and M. Li, "Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet," *eBioMedicine*, vol. 62, p. 103106, Dec. 2020.

[12] O. Ronneberger, "Invited talk: U-net convolutional networks for biomedical image segmentation," in *Bildverarbeitung für die Medizin 2017 - Algorithmen - Systeme - Anwendungen. Proceedings des Workshops vom 12. bis 14. März 2017 in Heidelberg*, ser. Informatik Aktuell, K. H. Maier-Hein, T. M. Deserno, H. Handels, and T. Tolxdorff, Eds. Springer, 2017, p. 3. [Online]. Available: https://doi.org/10.1007/978-3-662-54345-0_3

[13] B. N. . M. L. Jiancheng Yang, Liang Jin, "Ribfrac dataset: A benchmark for rib fracture detection, segmentation and classification (training set part 1) (v1.0) [data set]," *Zenodo*. [Online]. Available: https://zenodo.org/records/3893508