Whole-Body Bone Scan Segmentation Using SegFormer

Rafif Fausta Kusuma Syam[1], Ema Rachmawati[2], Mahmud Dwi Sulistiyo[3]
[1,2,3]*School of Computing, Telkom University, Bandung, Indonesia*
[1]rafiffaustaks@student.telkomuniversity.ac.id, [2]emarachmawati@telkomuniversity.ac.id, [3]mahmuddwis@telkomuniversity.ac.id

***Abstract*— Semantic segmentation can help in everyday life, especially in the medical field, to help detect cancer metastasis at an early stage. In semantic segmentation, CNN-based approaches have been known to dominate the semantic segmentation field, such as FCN and DeepLabv3+. The success of the Transformer approach in the Nature Language Processing (NLP) area triggered many researchers to use the Transformer approach in solving semantic segmentation problems, so the Vision Transformer (ViT) was born. ViT accepts images as patches to produce local and global attention, unlike the convolution approach. One ViT-inspired model, SegFormer, combines a Hierarchical Transformer encoder component to generate low-resolution fine features that focus on capturing small detailed, fine-grained information such as edges, corners, and local patterns, and high-resolution coarse features that focus on capturing more general and global characteristics of a scene, and a lightweight All-MLP decoder to combine those multi-level features to produce the final semantic segmentation mask. This allows SegFormer to capture local and global contextual information from an image. This paper proposed the SegFormer model to perform semantic segmentation of bone scan images into 12 classes based on bone regions. As a result, by comparing FCN and DeepLabv3+ convolution approaches, SegFormer outperforms both models with the highest mIoU value of 77.86%.**

Keywords—bone scan image, semantic segmentation, Transformer, convolution, SegFormer, DeepLabv3+, FCN

I. INTRODUCTION

One of the leading causes of death globally is cancer [1]. In 2020, based on data from the Global Cancer Incidence, Mortality, and Prevalence (GLOBOCAN) produced by the International Agency for Research on Cancer (IARC), it was estimated that there are around 10 million people who died due to cancer [2]. In addition, there were an estimated 19.3 million new cancer cases in the same year. This indicates that cancer is still a severe problem that must be addressed.

Cancer results from the abnormal development of different types of body cells. If not treated quickly, cancer can spread through the circulatory system to other body parts. This process is called metastasis and is very dangerous if it applies to the bones because it can increase the risk of death [3]. Early bone scan detection can be performed to minimize the spread of cancer. The result of the scan is an anterior and posterior image of the whole-body bone scan, which can be used to detect the location of the metastasis using the Bone Scan Index (BSI) measuring method [4]. However, semantic segmentation techniques can improve cancer metastasis location determination performance.

Semantic segmentation techniques have been widely applied in various fields [5]–[8]. Even so, convolution-based approaches still dominate in solving semantic segmentation problems. Starting from the success of FCN in performing semantic segmentation [8], it triggered further development of deep convolutional neural networks such as by exploiting the contextual information [9], [10] and the boundary information [11], [12]. This has led to the development of many convolution-based models for semantic segmentation, thus dominating the field. One is DeepLabv3+ [13], which delivers state-of-the-art performance on PASCAL VOC 2012 and CityScapes datasets. In the medical area, there is also BtrflyNet, which is used to build an automated BSI calculation system, and the results are clinically acceptable [8]. However, the global context, essential for tasks like semantic segmentation, has been challenging for CNN architectures to capture [14]. Besides convolution approaches, the success of Transformers in Natural Language Processing (NLP) triggered researchers to develop a Transformer-based architecture for dealing with semantic segmentation tasks leading to the emergence of the Vision Transformer (ViT) [15]. In contrast to convolution-based architecture, Vision Transformer architecture utilizes the encoder's self-attention mechanism, which is used to extract both local and global contextual information from an image. On the other hand, SegFormer has shown outstanding performance in segmenting ADE20K and CityScapes datasets by combining a Hierarchical Transformer encoder with an All-MLP decoder to capture local and global contextual information, beating the performance of convolution-based state-of-the-art methods such as PSPNet, DeepLabv3+, and FCN [16].

In this study, we proposed a system that can perform semantic segmentation on whole-body bone scan images using SegFormer. Comparisons were also made by comparing the system built using SegFormer against FCN and DeepLabv3+ to evaluate the system's performance. We successfully demonstrated that SegFormer outperforms FCN and DeepLabv3+ in semantic segmentation of bone scan images with the highest mIoU of 77.86%.

This article is organized into five sections. The first section provides the problem faced in this study and an overview of the methods offered to solve it. The second section provides an in-depth description of the experiments and related methods. The third section describes the proposed system. The fourth section shows the experiments along with their analysis results. The fifth section explains the conclusions drawn from the analysis of the experiment.

II. RELATED WORK

The success of convolutional networks in image classification led Long et al. [13] to propose a model that utilizes ConvNets architecture for semantic segmentation tasks, namely Fully Convolutional Network (FCN). The idea is proposed to replace fully connected layers with convolutional layers to generate spatial output maps, use skip connection between layers to fuse the lower-resolution maps with high-level semantic information and use a deconvolutional layer for upsampling, which is effective for learning dense prediction. As a result, FCN successfully

outperforms several traditional models in performing semantic segmentation on PASCAL VOC, NYUDv2, and SIFT Flow datasets. Since then, FCN has become fundamental for semantic segmentation. It triggered many researchers to utilize and improve deep convolutional neural networks (CNN).

In response, DeepLabv3+ was proposed by Chen et al. [13]. It uses DeepLabv3 as an encoder to extract detailed contextual data and a simple, efficient decoder module to improve segmentation outcomes along object boundaries. By applying depth-wise separable convolution and Xception to its architecture, which reduces computational cost while maintaining performance, DeepLabv3+ can achieve mIoU scores on PASCAL VOC 2012 and CityScapes datasets of 89% and 82.1%, respectively, which establishes state-of-the-art performance on both datasets.

Due to the popularity of deep convolutional neural network (CNN) in the field of semantic segmentation, Shimizu et al. [17] conducted research measurement of Bone Scan Index (BSI) automatically using CNN for whole-body bone scintigrams. The system designed uses BtrflyNet to perform bone segmentation and hot spot extraction from bone lesions. The dataset used is a dataset of prostate cancer patients aged from prostate cancer patients aged 52 to 95 with a total of 246 images. The results of the study show that the combination of BtrflyNet with deep supervision (DSV) for bone segmentation and BtrflyNet with the use of residual blocks for hot spot extraction achieved a high correlation value of 93.37% between the automatically performed BSI and the original BSI value. The results prove that the model is clinically well acceptable.

On the other side, the success of Transformers in Natural Language Processing (NLP) has influenced the development of Transformer-based models, which led to the emergence of Vision Transformers. In contrast to convolution-based architecture, the Vision Transformer architecture utilizes the encoder's self-attention mechanism to capture local and global contextual information from an image [15]. An image is treated as a sequence of tokens and fed to multiple Transformer layers where each token represents semantic information of the image [18].

Due to the emergence of Vision Transformer, Xie et al. [16] proposed a simple and efficient architecture for semantic segmentation, namely SegFormer. The architecture combines a Hierarchical Transformer encoder to generate high-resolution coarse and low-resolution fine features and a Lightweight All-MLP decoder to combine those multi-level features to produce the final semantic segmentation mask. It allows SegFormer to capture local and global contextual information from an image. By pre-training on ImageNet-1K, the test results on non-real-time datasets ADE20K, CityScapes, and COCO, resulted in the most significant mIoU values of 51.8%, 84%, and 46.7%, respectively. SegFormer is proven to beat several state-of-the-art models on ADE20K, CityScapes, and COCO datasets, such as PSPNet, Deeplabv3+, and FCN.

III. Bone Scan Segmentation Using SegFormer

This study's proposed system has three primary process stages: preprocessing, training, and testing. In the preprocessing stage, the entire dataset is divided into three parts for training, validation, and testing purposes. Next, it enters the training stage, where the SegFormer model training is carried out on the training data. Finally, after the training process is complete, the testing process is carried out to see the performance of the trained model in performing semantic segmentation on new image variants. An overview of the system can be seen in Fig. 1.

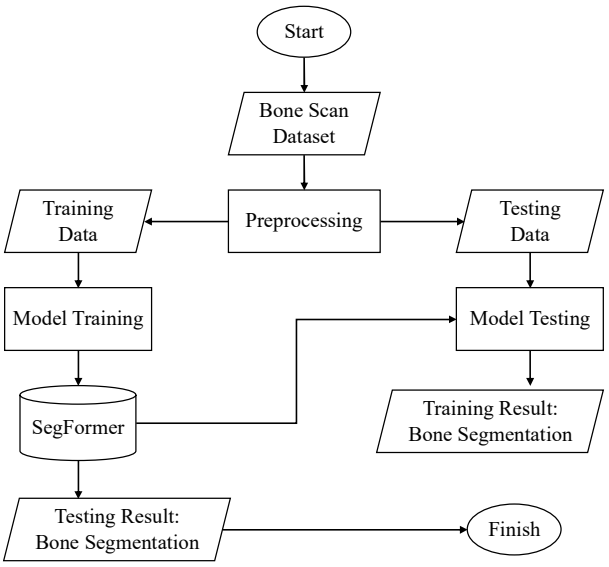


Fig. 1. Overview of the system's processing pipeline.

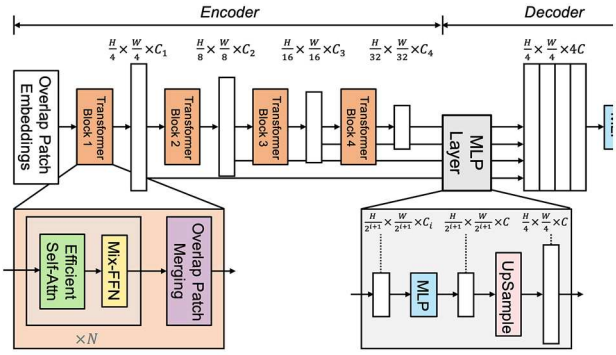


Fig. 2. Overview of the Transformer Architecture (captured from [16]).

A. SegFormer

SegFormer is a semantic segmentation model based on the Transformer architecture. The SegFormer architecture comprises two primary parts: a Hierarchical Transformer encoder to produce high-resolution coarse and low-resolution fine features and a Lightweight All-MLP decoder combining these multi-level features to make the final semantic segmentation mask. An overview of the SegFormer architecture can be seen in Fig. 2.

1) Hierarchical Transformer Encoder

SegFormer uses an encoder called Mix Transformer, an improvisation of Vision Transformer that focuses on performing semantic segmentation tasks. This encoder has four necessary modules: Hierarchical Feature Representation, Overlapped Patch Merging, Efficient Self-Attention, and Mix-FFN.

a) Hierarchical Feature Representation

In this module, patch merging is used to create a hierarchical feature map F_i with a resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ from an image with the dimensions $H \times W \times 3$, where $i \in$

{1,2,3,4}, and C_{i+1} is greater than C_i. Semantic segmentation performance can be improved by utilizing the high-resolution coarse features and low-resolution fine features offered by these multi-level features.

b) Overlapping Patch Merging

This module seeks to enhance the non-overlapping image or feature patch merging method utilized in the Vision Transformer architecture. In semantic segmentation, the patch merging process can be used to merge a feature patch of size $2 \times 2 \times C_i$ into a vector of size $1 \times 1 \times C_{i+1}$ to obtain a hierarchical feature map, where C_i refers to the input channels in feature patch and C_{i+1} refers to the output channels after merging process. Meanwhile, this process cannot maintain local continuity around the patch. Therefore, overlapping patch merging is used instead.

c) Efficient Self-Attention

In Transformer, the original self-attention layer causes computational bottlenecks in the encoder. Each of the head's Q, K, V, where Q represents a query, K represents a key, and V represents a value, has the exact dimensions as $N \times C$ in the original multi-head self-attention process, where $N = H \times W$ is the length of the sequences. The softmax activation function was also applied to the scaled dot product as shown in (1).

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V \quad (1)$$

H represents the height of the input data, W represents the width of the input data, K represents the sequence to be reduced, K^T represents the transpose of the sequence, d_{head} represents the dimensionality of the heads, C represents the input data channels, and N represents the total elements of the input data. Thus, the process has $O(N^2)$ complexity. In this module, the sequence reduction process reduces the computation by utilizing the reduction ratio R to reduce the sequence length, as shown in (2).

$$\hat{K} = Reshape\left(\frac{N}{R}, C \cdot R\right)(K)$$

$$K = Linear(C \cdot R, C)\left(\hat{K}\right) \quad (2)$$

The $Reshape\left(\frac{N}{R}, C \cdot R\right)(K)$ function refers to reshaping K into the $\frac{N}{R} \times (C \cdot R)$. The $Linear(C_{in}, C_{out})(\cdot)$ function refers to a linear layer that accepts the input dimension tensor C_{in} and produces the output dimension tensor C_{out}. As a result, K has a new dimension of $\frac{N}{R} \times C$ which reduces the computation from $O(N^2)$ to $O(\frac{N^2}{R})$.

d) Mix-FFN

The purpose of this module is to uncover location information by considering the effects of zero padding and using 3×3 convolution mixed with Multi-Layer Perceptron (MLP) on each Feed Forward Network (FFN as shown in (3).

$$x_{out} = MLP\left(GELU\left(Conv_{3\times3}(MLP(x_{in}))\right)\right) + x_{in} \quad (3)$$

x_{in} is the feature of the self-attention module. The sequence of operations starts from the MLP layer, followed by 3×3 convolution, and ends with Gaussian Error Linear Unit (GELU) activation function. As a result, Mix-FFN can generate spatial coordinates or positions of pixels within an image. This information can be used to figure out the relation between spatial regions.

2) Lightweight All-MLP Decoder

SegFormer implements only the MLP layer in the decoder to avoid computation due to interrelationships with other components. Another reason for using an MLP layer is to combine the high local and non-local attention produced by the encoder part. This allows the MLP layer to have a strong representation. There are four stages in the All-MLP decoder. First, the multi-level features generated from the MiT encoder are passed through the MLP layer to combine the channel dimensions. Second, the features are blended after being up sampled to 1/4. Third, the merged features are combined using an MLP layer. Finally, the segmentation mask is predicted using a second MLP layer with a $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution, as shown in (4).

$$\hat{F}_i = Linear(C_{in}, C)(F_i)$$

$$\hat{F}_i = Upsample\left(\frac{W}{4} \times \frac{W}{4}\right)(\hat{F}_i)$$

$$F = Linear(4C, C)(Concat(\hat{F}_i))$$

$$M = Linear(C, N_{cls})(F) \quad (4)$$

F_i refers to the feature map at index i, $Linear(C_{in}, C_{out})(\cdot)$ function refers to a linear layer with C_{in} as input vector and C_{out} as output vector, $Upsample\left(\frac{W}{4} \times \frac{W}{4}\right)\left(\hat{F}_i\right)$ function refers the function that up sampled the feature maps, $Concat\left(\hat{F}_i\right)$ refers to the function that merge the feature maps, and M refers to the prediction mask.

B. Evaluation Metric

The evaluation metric used in this study is Intersection Over Union (IoU) or the Jaccard Index, which measures the model's accuracy in segmenting a class through a comparative analysis between ground truth and predicted maps, as shown in (5).

$$IoU = J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

A refers to the ground truth, and B refers to the predicted maps [19]. In addition, mean Intersection over Union (mIoU) is also used to calculate the average value of all IoU class results, representing a model's performance in performing the entire segmentation.

IV. Experimental Results Analysis

A. Dataset

The dataset used is a bone scan image consisting of two parts, namely 18 data from Indonesian patients and 19 data from non-Indonesian patients. Since each patient has posterior and anterior bone scan images, the dataset increases to 36 and 38, resulting in 74 images. The 74 images are then split into three sections: training (37 images), validation (12 images), and testing (14 images). The number of posterior and anterior images is equally distributed for each section. The Indonesian

data came from the Faculty of Nuclear Medicine at Padjadjaran University, whereas the non-Indonesian data came from a crawl of Google Images [20]. All the datasets have been converted into .PNG files with image size standardized to 128 × 512 pixels, where 128 is the width, and 512 is the length. For labelling purposes, the anterior dataset has also been annotated into 12 classes: skull, cervical vertebrae, thoracic vertebrae, ribs, shoulder blade, humerus bone, lumbar vertebrae, sacrum bone, pelvis, and femur. Meanwhile, the posterior dataset is only annotated into ten classes because the sternum and collarbone cannot be seen from the back. The raw image sample can be seen in Fig. 3, and the example of the annotated image can be seen in Fig. 4.

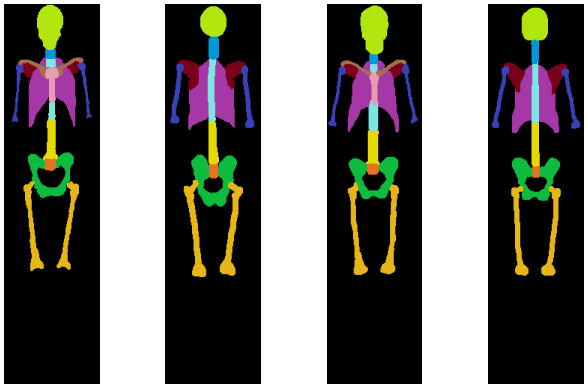
B. Experimental Settings

Optimization. In training the model, the AdamW optimizer was used as an optimizer. The learning rate value used was 10^{-3} while the weight decay value used was 5×10^{-4}.

Augmentation. Before training the model, Random Resize and Random Crop were applied in the training pipeline. The images were resized randomly into 256 × 1,024 pixels and then cropped randomly into 1,024 × 1,024 pixels. As for validating and testing the pipeline, only Random Resize was used with the same scale in the training pipeline. This method ensures that the images can be fed into those models before training, validating, and testing.

Indonesian patient Non-Indonesian patient

Fig. 3. Samples of raw anterior and posterior bone scan images. The left is a bone scan image from an Indonesian patient; the right is from a non-Indonesian patient.



Indonesian patient Non-Indonesian patient

Fig. 4. Samples of annotated anterior and posterior bone scan images. The left side is an annotated bone scan image from an Indonesian patient; the right is from a non-Indonesian patient.

Iteration and Batch. The maximum iteration for the training process is 6,000 iterations with a batch size of 3. The goal of selecting this value is to train the model effectively without incurring significant computational costs. During the training process, a validation interval was also performed for every 600 iterations to monitor and evaluate the performance of a model during training.

TABLE I. PERFORMANCE RESULTS OF SEGFORMER VARIANTS BY VARYING THE ENCODER SIZE

Method	Encoder	mIoU	
		Val	Test
SegFormer	MiT-B0	74.28	75.60
	MiT-B1	**75.63**	**76.95**

TABLE II. PERFORMANCE RESULTS OF SEGFORMER VARIANTS BY VARYING THE DECODER HEAD CHANNELS

Method	Encoder	Decoder Head Channels	mIoU	
			Val	Test
SegFormer	MiT-B0	256	74.28	75.60
		512	75.59	76.51
	MiT-B1	256	75.63	76.95
		512	**76.32**	**77.86**

C. Experimental Results

In the experimental stage, two experiments were conducted. For the first experiment, the model is divided into two variants from the Mix Transformer (MiT) encoder size aspect into B0 and B1 variants for training and testing. The difference between these two variants lies in the size of the embedding patch used, where variant B1 has larger dimensions than variant B0. The results, which can be seen in Table I, show that both variants achieved a mIoU value of more than equal to 74%. Variant B1 beat variant B0 with a thin score in the validation and testing results with a value of 75.63% and 76.95%. This indicates that increasing the encoder size can improve the performance of the SegFormer model. However, both variants have a slight underfit phenomenon, as seen from the more significant test mIoU result compared to the validation mIoU result.

As for the second experiment, the channels in the decoder head were increased from 256 to 512 channels for both variants. As shown in Table II, there is an improvement in the validation and testing of mIoU value results for both variants. The B0 variant has an increase of 1.31% for validation and 0.91% for test results. The B1 variant has an increase of 0.69% for validation and 0.91% for test results. However, the B1 performance still outperforms the B0 variant of the mIoU value with values of 76.32% and 77.86% for both validation and testing. This indicates that changing the number of channels in the decoder head can improve the performance of the SegFormer model. In addition, increasing the number of channels in the decoder head results in a more stable curve than previous variants, which can be seen in Fig. 5. Despite that, this does not fix the slight underfit phenomenon.

D. Comparison of SegFormer, FCN, and DeepLabv3+

To further evaluate the performance of the SegFormer variant results, we compared them with several convolution-based models, namely FCN [8] and DeepLabv3+ [13]. We divided these convolution models into two variants by

changing the ResNet encoder depth. The two variants are R18, which uses 18 layers, and R50, which uses 50 layers stacked on top of each other in the network. The results can be seen in Table III.

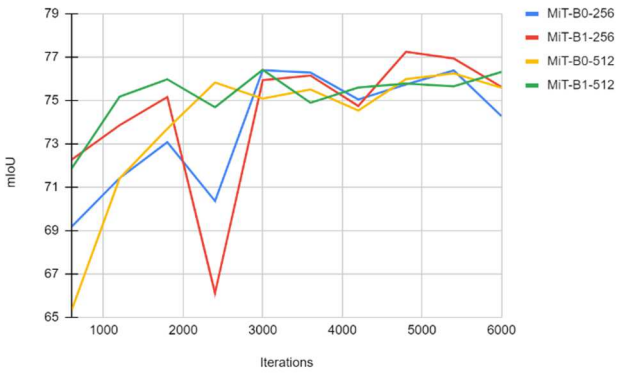


Fig. 5. mIoU graph of validation results for SegFormer variants during the training process.

TABLE III. MODEL COMPARISON ON TESTING DATA

Method	mIoU
DLv3+ (R18)	71.83
DLv3+ (R50)	59.71
FCN (R18)	75.48
FCN (R50)	73.13
SegFormer-B0-256	75.60
SegFormer-B1-256	76.95
SegFormer-B0-512	76.51
SegFormer-B1-512	**77.86**

All SegFormer variants surpassed the highest mIoU of all DeepLabv3+ and FCN variants. For DeepLabv3+, variant R18 achieved the highest mIoU with a value of 71.83%. Then, for FCN, the highest mIoU was achieved by variant R18 with a value of 75.48% which has a slight difference compared to SegFormer variant B0 with 256 decoder head channels with a difference value of 0.12%. As for SegFormer variants, the highest mIoU value is achieved by variant B1 with a decoder head using 512 channels with a value of 77.86%. For further evaluation, we compare the results of the IoU value for each class obtained from the segmentation results conducted on the testing data for the three best variants of each model. The comparison can be seen in Table IV.

SegFormer-B1-512 outperformed DeepLabv3+ (R18) and FCN (R18) variants by achieving the most significant value of IoU for ten classes out of 12. SegFormer achieved higher IoU scores than the convolutional model focusing only on the local context by applying a combination of local and global attention mechanisms to its decoder. This indicates that combining local and global information improves the spatial understanding of a pixel concerning its neighbors, which enhances the accuracy of segmentation. Based on the sample results of the anterior segmentation mask, shown in Fig. 6, SegFormer-B1-512 can perform segmentation well. DeepLabv3+ (R18) produces errors in segmentation where the background is segmented as ribs. Likewise, for FCN (R18), some parts should not be segmented but instead as a femur. As for the posterior segmentation mask sample results that can be seen in Fig. 7, FCN (R18) and SegFormer-B1-512 managed to segment well. As for Deeplabv3+ (R18) still produces the same error, namely the presence of a segmented background as a rib.

TABLE IV. COMPARISON OF STRONGEST VARIANT SEGMENTATION PERFORMANCE

Class	IoU		
	DLv3+ (R18)	FCN (R18)	SegFormer-B0-512
skull	90.07	89.75	**91.79**
cervical vertebrae	62.72	70.09	**71.10**
thoracic vertebrae	73.92	72.88	**76.65**
rib	71.72	84.58	**85.72**
sternum	76.78	77.56	**80.84**
collarbone	53.17	68.74	**69.33**
scapula	72.70	76.41	**76.84**
humerus	68.12	**74.82**	73.38
lumbar vertebrae	73.48	74.92	**79.17**
sacrum	65.50	**68.08**	66.80
pelvis	80.90	83.25	**83.58**
femur	72.96	64.67	**79.16**

Fig. 6. Visual comparison between raw image, ground truth, and predicted segmentation results for an anterior image.

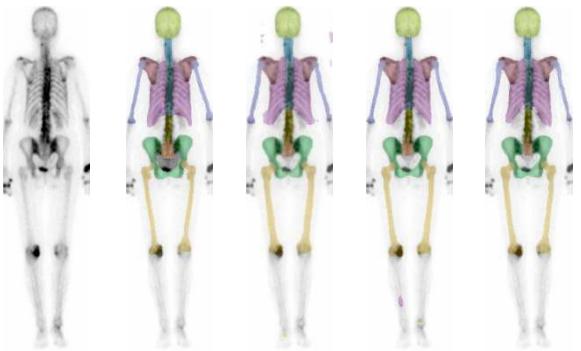


Fig. 7. Visual comparison between raw image, ground truth, and predicted segmentation results for a posterior image.

V. Conclusion

In this study, we present the whole-body bone scan segmentation using SegFormer. In contrast to convolution approaches that only use local attention, by combining the local and global attention mechanisms, SegFormer can generate a powerful representation of information that can improve the accuracy in performing semantic segmentation. To prove this, we compare the SegFormer model against several convolution-based models, such as FCN and DeepLabv3+. As a consequence, SegFormer outperforms both convolutional models. This indicates that the SegFormer approach can perform semantic segmentation well on bone scan images. For further research, modifications can be made by changing the size of the encoder and enlarging the channel in the decoder section to improve the performance.

Acknowledgment

The authors were greatly assisted in obtaining data by the Department of Nuclear Medicine and Molecular Theranostic, Dr. Hasan Sadikin General Hospital, Faculty of Medicine, Universitas Padjajaran, Indonesia. The Republic of Indonesia's Ministry of Education, Culture, Research, and Technology financed this study under the National Competitive Basic Research program, with grant number 021/SP2H/RT-JAMAK/LL4/2023. Telkom University additionally contributed funding under research project number 105/PNLT2/PPM/2023.

References

[1] E. Fasterius, M. Uhlén, and C. Al-Khalili Szigyarto, "Single-cell RNA-seq variant analysis for exploration of genetic heterogeneity in cancer," *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-45934-1.

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, 2021, doi: 10.3322/caac.21660.

[3] J.-F. Huang *et al.*, "Incidence of patients with bone metastases at diagnosis of solid tumors in adults: a large population-based study," *Ann Transl Med*, vol. 8, no. 7, 2020, doi: 10.21037/atm.2020.03.55.

[4] M. Imbriaco *et al.*, "A new parameter for measuring metastatic bone involvement by prostate cancer: The bone scan index," *Clinical Cancer Research*, vol. 4, no. 7, 1998.

[5] D. B. Nugraha, E. Rachmawati, and M. D. Sulistiyo, "Semantic Segmentation of Whole-Body Bone Scan Image Using Btrfly-Net," in *ICITEE 2022 - Proceedings of the 14th International Conference on Information Technology and Electrical Engineering*, 2022. doi: 10.1109/ICITEE56407.2022.9954073.

[6] B. A. Wicaksono, M. D. Sulistiyo, E. Rachmawati, T. A. Nugraha, and N. M. Z. Bin Hashim, "Identifying Image of the Correct Use of Face Mask Using Semantic Segmentation Technique," in *ICACNIS 2022 - 2022 International Conference on Advanced Creative Networks and Intelligent Systems: Blockchain Technology, Intelligent Systems, and the Applications for Human Life, Proceeding*, 2022. doi: 10.1109/ICACNIS57039.2022.10055202.

[7] H. Abdussyukur, M. D. Sulistiyo, E. Rachmawati, M. M. Arief, G. Kosala, and Adiwijaya, "Semantic Segmentation for Identifying Road Surface Damages Using Lightweight Encoder-Decoder Network," in *ICACNIS 2022 - 2022 International Conference on Advanced Creative Networks and Intelligent Systems: Blockchain Technology, Intelligent Systems, and the Applications for Human Life, Proceeding*, 2022. doi: 10.1109/ICACNIS57039.2022.10056030.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 4, 2017, doi: 10.1109/TPAMI.2016.2572683.

[9] H. DIng, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/ICCV.2019.00692.

[10] M. Zhen *et al.*, "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.01368.

[11] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang, "Context prior for scene segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.01243.

[12] H. Zhang *et al.*, "Context Encoding for Semantic Segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00747.

[13] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018. doi: 10.1007/978-3-030-01234-2_49.

[14] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 7242–7252. doi: 10.1109/ICCV48922.2021.00717.

[15] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput Surv*, vol. 54, no. 10s, 2022, doi: 10.1145/3505244.

[16] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," May 2021, [Online]. Available: http://arxiv.org/abs/2105.15203

[17] A. Shimizu *et al.*, "Automated measurement of bone scan index from a whole-body bone scintigram," *Int J Comput Assist Radiol Surg*, vol. 15, no. 3, 2020, doi: 10.1007/s11548-019-02105-x.

[18] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans Pattern Anal Mach Intell*, 2022, doi: 10.1109/TPAMI.2022.3152247.

[19] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 7, 2022, doi: 10.1109/TPAMI.2021.3059968.

[20] E. Rachmawati, Jondri, K. N. Ramadhani, A. H. S. Kartamihardja, A. Achmad, and R. Shintawati, "Automatic whole-body bone scan image segmentation based on constrained local model," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 6, pp. 2526–2537, Dec. 2020, doi: 10.11591/eei.v9i6.2631.