Investigating Convolution-Attention Model for Bone Scan Image Segmentation

Alfinata Yusuf Sitaba[1], Ema Rachmawati[2], Mahmud Dwi Sulistiyo[3]
[1,2,3]*School of Computing, Telkom University, Bandung, Indonesia*
[1]alfinata@student.telkomuniversity.ac.id, [2]emarachmawati@telkomuniversity.ac.id, [3]mahmuddwis@telkomuniversity.ac.id

***Abstract*—Bone scan image segmentation is a crucial step in the early detection of a tumor spreading across the human body. By dividing each bone region, the Bone Scan Index can be analyzed for further follow-up against cancer. Much research has been done in this field, including the uses of Computer Aided Diagnosis (CAD) in an application developed by EXINI, the development of Active Shape Model (ASM), and Constrained Local Model (CLM) as a further development of ASM. However, these models still rely on landmark points for their training phase instead of using masks for the annotations. A recent convolution model, DeepLabv3+, relies on a convolution mechanism to extract local features. A new approach using a pure transformer, Segmenter, can extract global features in a parallel process. A combination of convolution and attention, Dual Attention Network (DANet) uses a similar backbone from DeepLabv3+ and implements attention modules to capture long-range contextual information from the input image. In this paper, a bone scan image segmentation system using DANet will be proposed. All models are trained using the bone scan dataset divided into anterior and posterior groups. The annotation is composed of 12 different classes of bone regions. The results show that the convolution-attention approach of DANet outperformed existing models in both the anterior and posterior sections. A performance of 76.85% mIoU is achieved in the anterior section, and 80.99% mIoU is achieved in the posterior section.**

Keywords—bone scan image, segmentation, Dual Attention Network, convolution, attention, DeepLabv3+, Segmenter

I. INTRODUCTION

Cancer is one of the diseases with the highest mortality rate globally, ranked second in the US [1]. In Indonesia, about 2.2 billion rupiahs were spent treating 1.3 million cancer cases [2]. However, as technology advances, the risk of cancer death has been declining since 1991 by 32%, and about 3.5 million deaths have been avoided since 2019. This decline is caused by advances in surgical techniques, targeted therapy, and early detection. One of the early detection methods is doing a whole-body bone scan to monitor the spread of cancer in the body [1].

Bone Scan Index (BSI) is a scale used to calculate the spread of cancer in the bone. This index represents how many regions in the bone have been infected by the tumor and the spread of the metastasis [3]. But before calculating BSI, a segmentation of the bone scan needs to be done. EXINI has made software to do bone scan segmentation and calculate BSI [4]. In this case, some experiments on bone scan segmentation have shown promising results [5]–[7]. However, these models, such as the Constrained Local Model and Active Shape Model, are still using landmark points as annotations in their training phase and not using a segmentation mask.

One of the semantic segmentation models is DeepLabv3+ with its ResNet backbone which relies on its convolution mechanism to extract local features during the segmentation process [8]. Another newer approach, Vision Transformer, focuses on self-attention mechanisms to extract global context from an image [9], [10]. One of the models based on this Vision Transformer backbone, Segmenter, uses a pure transformer approach in the case and has a performance that rivals Deeplab [11]. Another model, DANet, combines attention and convolution to do the segmentation. Using the ResNet backbone as its convolutional part and two attention modules as the attention part, the model tries to extract local and global context from the image [12].

In this study, we proposed a bone scan image segmentation system based on the dual attention model. DANet was chosen as the baseline method because it combines the attention and convolution approach. A comparison is also made with Segmenter [11] and DeepLabv3+ [13] to analyze the strengths and weaknesses of the model. Several model variants, mainly focusing on the depths of the network, are also implemented. These models were chosen as the baseline methods for evaluation because each represents their transformer and convolution approach.

This article is organized into five sections. The first section describes this study's importance and overviews various methods related to the case. The second section explains the different methods and their experiment in more detail. In the third section, the proposed method is thoroughly described. The fourth section shows the experimental and analytical results. The fifth section concludes the experiment.

II. RELATED WORKS

To interpret bone scans qualitatively, Imbriaco et al. [3] proposed the Bone Scan Index (BSI), an index that represents the regions of bone infected by the tumor and the regional spread of the metastasis. BSI is designed to be calculated automatically. For automating the segmentation process, EXINI develops computer-aided diagnosis (CAD) tools to do bone scan image segmentation using the active shape model as the basis model. This experiment by Sadik et al. [4] used patients' data from January 1999 to June 2002. This software helps doctors spot metastasis. However, this model still focuses on one center, and yet from different centers with different cameras and protocols.

A similar experiment by Kikuchi et al. [14] focuses on different patients. The atlas being used is from Sweden and Japan. The experiment uses CAD, and the conclusion is that the atlas from Japan is better to be used for Japanese patients. Another experiment using a constrained local model (CLM) was done by Rachmawati et al. [5]. This model works by replicating variations in the annotations in the form of landmark points. Further research is done by developing an active shape model (ASM) with a cumulative error distribution of 0.0446. Both experiments use data from the Department of Nuclear Medicine and Molecular Theranostic, Faculty of Medicine, Universitas Padjajaran. However, these methods still use landmark points as their data annotation.

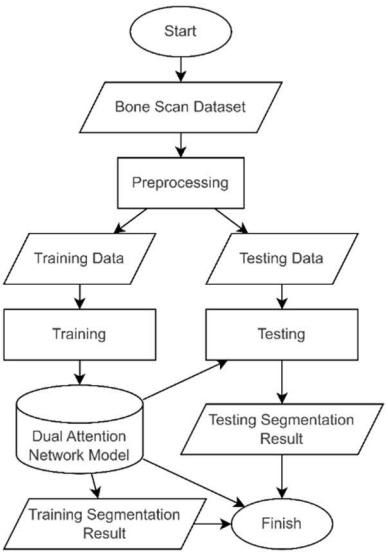

Fig. 1. Proposed system overview.

In semantic segmentation, the convolutional neural network with its convolution has been state-of-the-art for some time. Chen et al. [13] developed DeepLabv3+ relies on convolution for semantic segmentation. By combining the Xception model and atrous convolution, DeepLabv3+ performs well in the PASCAL VOC 2012 and Cityscapes datasets. DeepLabv3+ has its share of uses in the medical field [15], [16].

On the other side, a new method in natural language processing called transformer [9] started being integrated into computer vision [10]. Transformers can extract global context and can run in parallel. Strudel et al. [11] have researched this case and propose Segmenter, a model that uses encoder-decoder architecture based on pure transformer approach. This model outperforms the state-of-the-art methods in ADE20K, Pascal Context, and Cityscapes dataset. In evaluating the transformer-based semantic segmentation model, Segmenter performs well in pathological image segmentation [17].

An experiment combining both approaches by Fu et al. [12] proposes a dual-attention network. This method can capture long-range contextual information effectively and give precise segmentation results. DANet is shown to achieve outstanding performance consistently in PASCAL VOC 2012, Pascal Context, and COCO Stuff dataset. Several medical studies have implemented this method [18]–[20].

III. Our Proposed System

The system is built by using bone scan image dataset with its annotation. 80% of the dataset (60% training and 20% validation) is used for training the model, and 20% for the testing phase. The images and annotations used are first passed through the preprocessing step. Training images are used to train the dual attention network, and the testing images are used to evaluate the model's performance. The illustration is shown in Fig. 1.

A. Dual Attention Network

ResNet serves as the foundation of the semantic segmentation model known as Dual Attention Network (DANet). Additionally, it makes use of two attention modules to gather broad contextual data in the spatial and channel dimensions. The foundation is a pretrained ResNet using the dilated strategy. The final two ResNet blocks use dilated convolutions instead of downsampling to retain more information without using additional parameters. The collected features are then input into two attention modules running in parallel. Fig. 2 displays the illustration of the model.

1) Position Attention Module

The position attention module is introduced to model rich contextual associations between local features. This module improves the representation capability by capturing the larger scope of contextual information into local features. The illustration of this module is shown in Fig. 3. Given a local feature $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$, **A** is fed to convolutional layers to produce three feature maps **B**, **C**, and **D**, where $\{\mathbf{B}, \mathbf{C}, \mathbf{D}\} \in \mathbb{R}^{C\times H\times W}$. These feature maps are then reshaped into $\mathbb{R}^{C\times N}$ where $N = H \times W$ is the number of pixels. A matrix multiplication is done between the transpose of **C** and **B,** and then a softmax layer is applied to calculate the spatial attention map $\mathbf{S} \in \mathbb{R}^{N\times N}$ as shown in (1).

$$s_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)} \tag{1}$$

s_{ji} measures the i^{th} position's impact on the j^{th} position. More excellent correlation is contributed by closer feature representations between the two positions. Another matrix multiplication is done between **D** and the transpose of **S**, which then is reshaped to $\mathbb{R}^{C\times H\times W}$. Finally, the result is multiplied by a scale parameter α, and an element-wise sum operation is done with the features **A** to obtain the final output $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$ as shown in (2).

$$E_j = \alpha \sum_{i=1}^{N}(s_{ji}D_i) + A_j \tag{2}$$

α is initialized as 0 and steadily learns to assign more weight. As shown in Equation 2, from original features and all positions, a weighted sum of features at each position is represented as feature **E**. As a result, it has selectively aggregated the contexts according to the spatial attention map and has a contextual view globally. Intra-class compactness and semantic regularity have been improved due to the mutual gains achieved by similar semantic features. The position attention module is introduced to model rich contextual relationships between local features. This module enhances the representation capability by capturing a more extensive range of contextual information into local features.

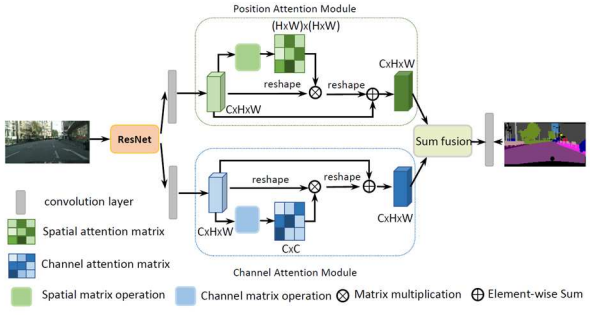

Fig. 2. Overview of the Dual Attention Network [12].

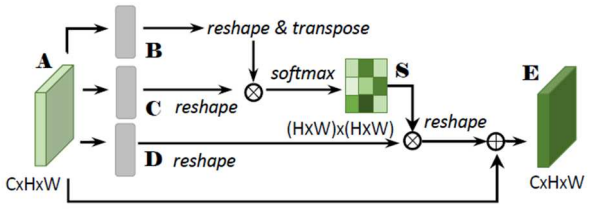

Fig. 3. Position attention module illustration [12].

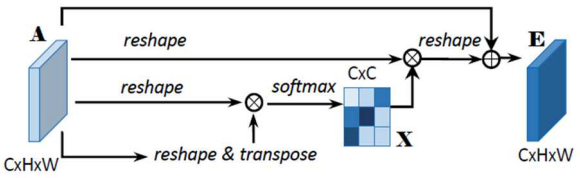

Fig. 4. Channel attention module illustration [12].

2) Channel Attention Module

Class-specific responses can be obtained for each channel map of high-level features, and different semantic responses can be connected to one another. By utilizing the interdependencies between each channel map, it is possible to improve the representation of a feature with specific semantics and the emphasis on interdependent feature maps. The objective of the channel attention module was to explicitly model the interdependencies between the channels. Fig. 4 provides an illustration of this module.

This channel attention map $\mathbf{X} \in \mathbb{R}^{C\times C}$ is calculated directly from the original features $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$. $\mathbf{A}$ is reshaped to $\mathbb{R}^{C\times N}$. Then a multiplication matrix is done between $\mathbf{A}$ and the transpose of $\mathbf{A}$. A softmax layer is applied to obtain the channel attention map $\mathbf{X} \in \mathbb{R}^{C\times C}$ as shown in (3).

$$x_{ji} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)} \tag{3}$$

x_{ji} measures the i^{th} channel's impact on the j^{th} channel. Another matrix multiplication is done between the transpose of $\mathbf{X}$ and $\mathbf{A}$ and reshaped to $\mathbb{R}^{C\times H\times W}$. The result is then multiplied by another scale parameter β to which is done an element-wise sum operation with $\mathbf{A}$ to obtain the final output $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$ as shown in (4).

$$E_j = \beta \sum_{i=1}^{C}(x_{ji}A_i) + A_j \tag{4}$$

β slowly learns a weight from 0. As shown in Equation 4, the final feature of each channel is a weighted sum of the features of all channels and original features. It also represents the long-range semantic relations between feature maps. This supports the discriminability between the features.

3) Attention Module Embedding

The features from the last two attention modules are aggregated to obtain rich long-range contextual information. The outputs of two attention modules are formed by a convolutional layer, which then an element-wise sum is applied to produce the merging of features. Another convolutional layer then follows it to make the final prediction map.

B. Evaluation Metric

This bone scan image segmentation system's performance is measured by intersection over union. IoU represents the ratio of the intersection over union between the ground truth and the prediction, as shown in (5).

$$IoU = \frac{True\ Positive}{True\ Positive + False\ Positive + False\ Negative} \tag{5}$$

The model's overall performance can be measured by averaging the IoU of each class (mIoU).

IV. Experimental Results and Analysis

This section explains the dataset, model training phase, experiment result, and analysis.

A. Dataset

The dataset is composed of 18 Indonesian and 19 non-Indonesian patients. Each patient has anterior and posterior bone scan images, resulting in 36 and 38 images. Indonesian images were obtained from the Department of Nuclear Medicine and Molecular Theranostic, Faculty of Medicine, Universitas Padjajaran. Non-Indonesian images were obtained by crawling from Google Images. All raw images have been converted into .PNG format and resized to 128 × 512 pixels [21]. The images have also been annotated into 12 classes (skull, cervical vertebrae, thoracic vertebrae, rib, scapula, humerus, lumbar vertebrae, sacrum, pelvis, and femur) according to the number of bone regions. However, the posterior section only has ten classes because the bone regions 'sternum' and 'collarbone' were not visible from behind. Samples of the raw images and the annotations are shown in Fig. 5.

B. Result and Analysis

We divide the dataset into anterior and posterior groups. Each group containing 37 images was divided into 24 images for training, 6 for validating, and 7 for testing. We used the tool MMSegmentation [22] for the training, testing, and inference of the model. For the configuration, we set the parameters as follows: batch size to 3, the image is resized to 256 × 1,024 pixels which are then cropped into a 512 × 512-pixel image, and training iteration to be 5,000. While training, evaluation was done every 500 iterations to monitor the process. This evaluation reports the model's performance in the IoU of each class and the mean IoU of all classes.

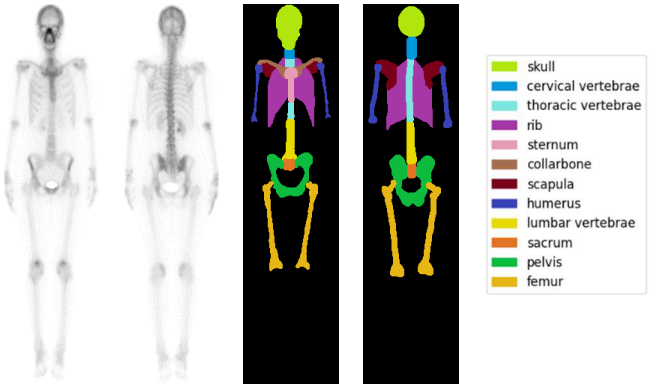

Fig. 5. Samples of raw bone scan images and their annotations.

TABLE I. DANET PERFORMANCE ON ANTERIOR SECTION

Method	depth	pam_channels	mIoU	
			val	*test*
DANet-50-16	50	16	75.1	73.11
DANet-50-32		32	76.22	73.58
DANet-50-64		64	75.05	74.33
DANet-101-16	101	16	76.76	73.48
DANet-101-32		32	**76.85**	73.53
DANet-101-64		64	75.08	**74.66**

TABLE II. DANET PERFORMANCE ON POSTERIOR SECTION

Method	depth	pam_channels	mIoU	
			val	*test*
DANet-50-16	50	16	79.82	80.38
DANet-50-32		32	80.14	80.67
DANet-50-64		64	79.87	80.19
DANet-101-16	101	16	80	80.68
DANet-101-32		32	80.82	80.53
DANet-101-64		64	**80.99**	**80.71**

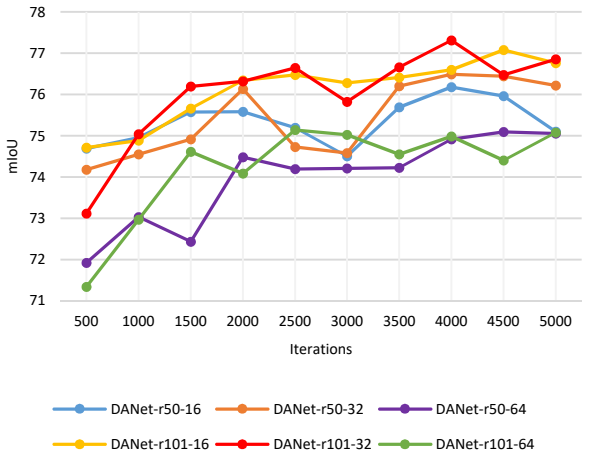

Fig. 6. mIoU graph of DANet variants in anterior section training phase.

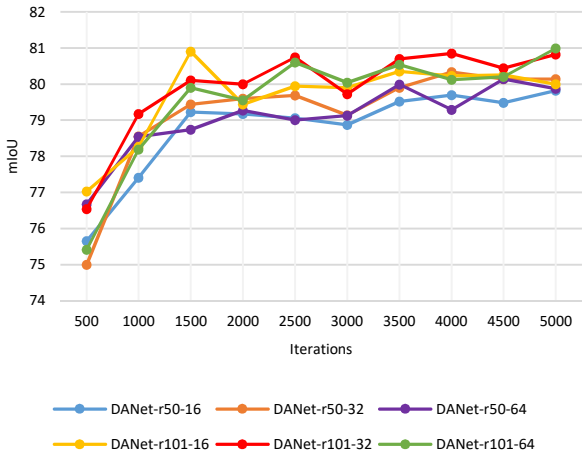

Fig. 7. mIoU graph of DANet variants in posterior section training phase.

After training, the model is tested using the seven testing images and then evaluated. The evaluation also reports the IoU of each class and the mean IoU. This process is done for all models to be compared and analyzed later. Two major DANet variants were made: DANet-50, which has a depth of 50, and DANet-101, which has a depth of 101. From each major variant, minor variants were made based on their position attention module channels, represented by pam_channels. As shown in Table I, the DANet model in the anterior section performs at mIoU in the 75% to 76% range. Increasing the depth parameter improves the performance slightly. Having the biggest number of positional attention module channels does not perform the best in this case. The value of 32 in pam_channels perform the best and is followed by 16 channels. This is indicated by the slower-growing performance of the 64 channels, as shown in Fig. 6.

In the posterior section, an average of 79% - 81% of mIoU is achieved, as shown in Table II. Increasing the depth also improves the performance slightly in this case. However, the bigger depth variant performs better with bigger positional attention module channels. Compared to the anterior section, all variants in the posterior section tend to follow a similar mIoU training phase curve, as shown in Fig. 7.

C. Comparison of DANet, Segmenter, and DeepLabv3+

In this section, we present the comparison of segmentation results using DANet, Segmenter [11], and DeepLabv3+ [13]. Segmenter used a purely self-attention mechanism in its architecture, while DeepLabv3+ focused on the convolutional part of its architecture. Three variants of Segmenter representing embedded dimensions and the number of self-attention heads and two variants of DeepLabv3+ consisting of 50 and 101 depths were used in this experiment.

For the anterior section, the DANet models perform on par and slightly better than DeepLabv3+ models, as shown in Table III. The segmenter model did not perform as well as the other two, with the best performance at the mIoU of 64.17%. While DeepLabv3+'s best model tops at mIoU of 75.82%, the DANet model still performs best at 76.85% with a margin of about 1.03%. For comparing each best variant from each model, a visual comparison of the raw image, ground truth, and segmentation results can be seen in Fig. 8. The individual IoU of each class can also be seen in Table IV. The model DANet-101-64 outperforms other models in 7 out of 12 classes. In the cervical vertebrae class, DANet's attention mechanisms can perform well, corresponding to Segmenter's similar performance and convolution's struggle shown in DeepLabv3+'s performance. The same thing can be said for convolution parts and against the attention mechanisms in classes such as the sternum, collarbone, scapula, and femur.

TABLE III. ANTERIOR SECTION MODEL COMPARISON

Method	mIoU	
	val	*test*
Seg-T	60.02	56.71
Seg-S	62.13	59.12
Seg-B	64.17	57.29
DANet-50-64	75.05	74.33
DANet-100-64	75.08	74.66
DANet-50-16	75.1	73.11
Deep-50	75.1	73.17
Deep-100	75.82	**74.96**
DANet-50-32	76.22	73.58
DANet-100-16	76.76	73.48
DANet-100-32	**76.85**	73.53

TABLE IV. STRONGEST PER-VARIANT ANTERIOR PERFORMANCE

Class	IoU		
	Seg-B	***Deep-101***	***DANet-101-32***
skull	93.52	94.37	**94.93**
cervical vertebrae	48.42	36.89	**49.46**
thoracic vertebrae	57.61	69.63	**70.74**
rib	73.63	79.74	**81.02**
sternum	62.35	**86.04**	80.71
collarbone	53.91	**78.91**	76.61
scapula	52.22	**72.55**	72.51
humerus	29.6	**74.82**	63.51
lumbar vertebrae	75.24	76.56	**78.4**
sacrum	37.37	55.3	**66.51**
pelvis	82.71	**87.38**	84.6
femur	57.73	80.78	**83.39**

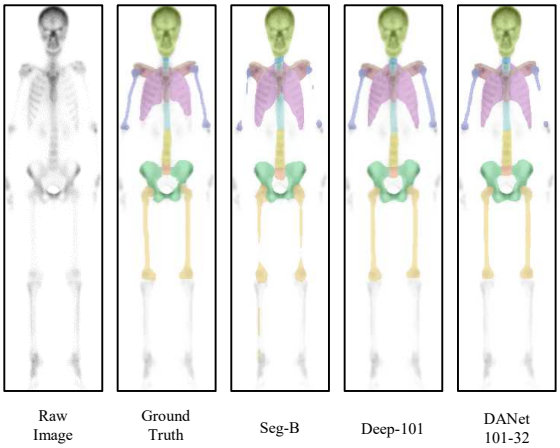

Fig. 8. Visual comparison between the anterior dataset's raw image, ground truth, and model segmentation.

By combining both approaches, DANet can benefit by outperforming both models in the remaining classes. In humerus class, DANet's attention module can be seen incorrectly segmenting like Segmenter at the right side of the image. The less clear right side of the image might be the cause, but DeepLabv3+ can still segment those regions.

In the posterior section, all DANet variants outperform even the best of DeepLabv3+'s model. DANet's smallest variant can be seen in Table V to perform better at the margin of 0.29%, and its best variant at mIoU of 80.99% at the margin of 1.46%. Segmenter, however, still does not perform as well as the other two, with the best variant still at the mIoU of 72.75%. We observed that DANet performs well in both anterior and posterior datasets.

An average of 75% mIoU is achieved in the anterior dataset with its variant. While the bigger depth improves the performance slightly, the middle value of positional attention module 32 gives the best performance at mIoU of 76.85%. An average mIoU is achieved at 79% for the posterior dataset between all variants. Unlike the anterior dataset, the biggest channel in the biggest depth gives the best performance at mIoU of 80.99%.

As seen in Table VI, DANet outperforms the other two in most of the classes. Most were small gaps between DeepLabv3+, but it shows both approaches' advantages. By using these combinations, DANet's performance can be seen in the class lumbar vertebrae and sacrum.

TABLE V. POSTERIOR SECTION MODEL COMPARISON

Method	mIoU	
	val	***test***
Seg-T	68.72	68.67
Seg-S	70.86	67.2
Seg-B	72.75	72.2
Deep-100	79.46	80.25
Deep-50	79.53	78.76
DANet-50-16	79.82	80.38
DANet-50-64	79.87	80.19
DANet-100-16	80	80.68
DANet-50-32	80.14	80.67
DANet-100-32	80.82	80.53
DANet-100-64	**80.99**	**80.71**

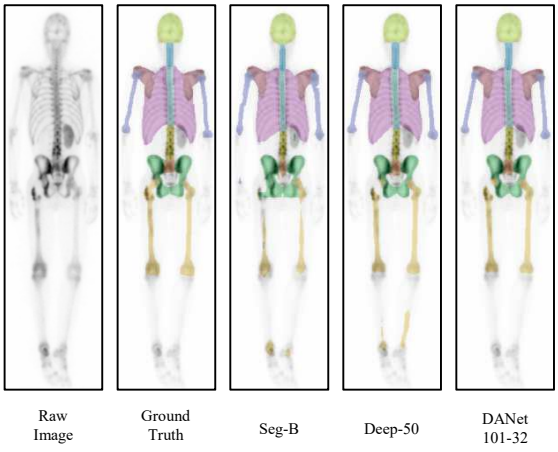

Fig. 9. Visual comparison between the posterior dataset's raw image, ground truth, and model segmentation.

As shown in Fig. 9, DANet also did not incorrectly segment femur parts in the bottom part of the image, unlike both Segmenter and DeepLabv3+. In the comparison study, DANet outperforms the other two models. Even DANet's most minor depth and positional attention module channel variant in the posterior dataset still performs better than DeepLabv3+'s biggest variant.

For the anterior dataset, DANet performs better at the margin of 12.68% against Segmenter and 1.03% against DeepLabv3+. In the posterior dataset, DANet performs better at the margin of 8.23% against Segmenter and 1.46% against DeepLabv3+. Looking deeper into the per-class IoU of each best variant of the three models, DANet outperforms the other two models in most classes.

TABLE VI. STRONGEST PER-VARIANT POSTERIOR PERFORMANCE

Class	IoU		
	Seg-B	***Deep-50***	***DANet-101-64***
skull	91.85	95.03	**95.91**
cervical vertebrae	65.31	82.44	**84.8**
thoracic vertebrae	61.71	65.89	**67.73**
rib	82.21	87.45	**88.9**
scapula	76.62	**87.05**	86.42
humerus	69.75	77.56	**80.09**
lumbar vertebrae	68.68	75.47	**84.1**
sacrum	53.82	67.37	**73.78**
pelvis	72.77	85.25	**85.7**
femur	44.07	66.54	**77.13**

In some classes, DANet is shown to take advantage of the combination of both approaches, giving a performance like one of the models in contrast to the other one. This indicates that DANet's model of combining the two methods gives significant value to performance. DeepLabv3+ performs best in the anterior dataset with the variant Deep-100 with mean intersection over union at 75.82. DANet achieves the best in the posterior dataset with the variant DANet-100 with mean intersection over union at 80.99. Segmenter struggles with the current dataset and model tuning compared to DeepLabv3+ and DANet.

V. CONCLUSION

Bone scan image segmentation plays a massive role in the early detection of countermeasures against cancer. Using the available datasets divided into anterior and posterior groups, we demonstrated the capabilities DANet, which combines the attention and convolution approach, and compared it with Segmenter and DeepLabv3+. In this study, the proposed system outperformed the other two models with the same batch size and iteration in the training phase. For further studies, the same or different approaches could be used to further develop a better model for the case. By incorporating a broader variety of approaches, we can deepen our knowledge and insight into the intricacies of attention mechanisms, convolutions, and other methods. This expanded range of approaches enables us to explore different perspectives, techniques, and assumptions, leading to a more nuanced and comprehensive understanding of these complex concepts.

ACKNOWLEDGMENT

The Department of Nuclear Medicine and Molecular Theranostic, Dr. Hasan Sadikin General Hospital, Faculty of Medicine, Universitas Padjajaran, Indonesia, was a great help to the authors in gathering data. The Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia provided funding for this project under the National Competitive Basic Research program with grant number 021/SP2H/RT-JAMAK/LL4/2023. Telkom University also provided support under research grant number 105/PNLT2/PPM/2023.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," CA Cancer J Clin, vol. 72, no. 1, pp. 7–33, 2022, doi: https://doi.org/10.3322/caac.21708.

[2] S. Gondhowiardjo et al., "Five-Year Cancer Epidemiology at the National Referral Hospital: Hospital-Based Cancer Registry Data in Indonesia," JCO Glob Oncol, no. 7, pp. 190–203, 2021, doi: 10.1200/GO.20.00155.

[3] M. Imbriaco et al., "A new parameter for measuring metastatic bone involvement by prostate cancer: the Bone Scan Index," Clin Cancer Res, vol. 4, no. 7, pp. 1765–1772, Jul. 1998.

[4] M. Sadik et al., "Computer-Assisted Interpretation of Planar Whole-Body Bone Scans," Journal of Nuclear Medicine, vol. 49, no. 12, pp. 1958–1965, 2008, doi: 10.2967/jnumed.108.055061.

[5] E. Rachmawati, J. Jondri, K. Ramadhani, A. H. 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, 2020, doi: 10.11591/eei.v9i6.2631.

[6] E. Rachmawati, F. R. Sumarna, Jondri, A. H. S. Kartamihardja, A. Achmad, and R. Shintawati, "Bone Scan Image Segmentation based on Active Shape Model for Cancer Metastasis Detection," in 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1–6. doi: 10.1109/ICoICT49345.2020.9166193.

[7] 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, pp. 389–400, Mar. 2020.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Trans Pattern Anal Mach Intell, vol. 40, no. 4, pp. 834–848, 2018, doi: 10.1109/TPAMI.2017.2699184.

[9] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017.

[10] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," CoRR, vol. abs/2010.11929, 2020.

[11] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," arXiv preprint arXiv:2105.05633, 2021.

[12] J. Fu et al., "Dual Attention Network for Scene Segmentation," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2019, pp. 3141–3149. doi: 10.1109/CVPR.2019.00326.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in ECCV, 2018.

[14] A. Kikuchi, M. Onoguchi, K. Horikoshi Hiroyuki and Sjöstrand, and L. Edenbrandt, "Automated segmentation of the skeleton in whole-body bone scans: influence of difference in atlas," Nucl Med Commun, vol. 33, no. 9, pp. 947–953, Sep. 2012.

[15] J. Wang and X. Liu, "Medical image recognition and segmentation of pathological slices of gastric cancer based on Deeplab v3+ neural network," Comput Methods Programs Biomed, vol. 207, p. 106210, Aug. 2021, doi: 10.1016/j.cmpb.2021.106210.

[16] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention Deeplabv3+: Multi-level Context Attention Mechanism for Skin Lesion Segmentation," 2020, pp. 251–266. doi: 10.1007/978-3-030-66415-2_16.

[17] C. Nguyen, Z. Asad, R. Deng, and Y. Huo, "Evaluating transformer-based semantic segmentation networks for pathological image segmentation," in Medical Imaging 2022: Image Processing, I. Išgum and O. Colliot, Eds., SPIE, Apr. 2022, p. 128. doi: 10.1117/12.2611177.

[18] Y. Li, H. Tang, W. Wang, X. Zhang, and H. Qu, "Dual attention network for unsupervised medical image registration based on VoxelMorph," Sci Rep, vol. 12, no. 1, p. 16250, Sep. 2022, doi: 10.1038/s41598-022-20589-7.

[19] L. Chen, T. Mao, and Q. Zhang, "Identifying cardiomegaly in chest x-rays using dual attention network," Applied Intelligence, vol. 52, no. 10, pp. 11058–11067, Aug. 2022, doi: 10.1007/s10489-021-02935-w.

[20] R. Hu, H. Yan, F. Nian, R. Mao, and T. Li, "Unsupervised computed tomography and cone-beam computed tomography image registration using a dual attention network," Quant Imaging Med Surg, vol. 12, no. 7, pp. 3705–3716, Jul. 2022, doi: 10.21037/qims-21-1194.

[21] D. B. Nugraha, E. Rachmawati, and M. D. Sulistiyo, "Semantic Segmentation of Whole-Body Bone Scan Image Using Btrfly-Net," in 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, Oct. 2022, pp. 264–269. doi: 10.1109/ICITEE56407.2022.9954073.

[22] MMsegmentation Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," 2020.