Investigating Convolution-Attention Model for

Bone Scan Image Segmentation

Alfinata Yusuf Sitaba

School of Computing

Telkom University

Bandung, Indonesia

alfinata@student.telkomuniversity.ac.id

Ema Rachmawati

School of Computing

Telkom University

Bandung, Indonesia

emarachmawati@telkomuniversity.ac.id

Mahmud Dwi Sulistyo

School of Computing

Telkom University

Bandung, Indonesia

mahmuddwis@telkomuniversity.ac.id

*Abstract*—Bone scan image segmentation is a crucial step in the process of early detection of a tumor spreading across the human body. By dividing each bone region, an analysis of Bone Scan Index can be done for further follow-up against the cancer. Much research has been done in this field, including the uses of Computer Aided Diagnosis (CAD) in an application developed by EXINI, 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 its training phase, instead of using masks for the annotations. A recent convolution model, DeepLabv3+, relies on convolution mechanism to extract local features. A new approach using pure transformer, Segmenter, can extract global features in parallel process. A combination of convolution and attention, Dual Attention Network (DANet), uses 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 anterior and posterior section. 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

# Introduction

Cancer is one of the diseases with the highest mortality rate globally, ranked second in the US [[1](#_R._L._Siegel,)]. In Indonesia, about 2.2 billion rupiahs were spent on treating 1.3 million cancer cases [[2](#_S._Gondhowiardjo_et)]. 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 methods of early detection is by doing whole body bone scan to monitor the spread of cancer in the body [[1](#_R._L._Siegel,)].

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 has been infected by tumor and the spread of the metastasis [[3](#_M._Imbriaco_et)]. But before calculating BSI, a segmentation of the bone scan needs to be done. EXINI has made a software to do both bone scan segmentation and calculate BSI [[4](#_M._Sadik_et)]. In this case, some experiments about bone scan segmentation have shown good results [[5](#_E._Rachmawati,_J.), [6](#_E._Rachmawati,_F.), [7](#_Shimizu_et_al.,)]. However, these models such as Constrained Local Model and Active Shape Model are still using landmark points as annotations in its 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](#_L.-C._Chen,_G.)]. Another newer approach, Vision Transformer, focuses on self-attention mechanisms to extract global context from an image [[9](#_Vaswani_et_al.,), [10](#_Dosovitskiy_et_al.,)]. 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](#_R._Strudel,_R.)]. 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 both local and global context from the image [[12](#_Fu,_Jun,_et)].

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 and its recent uses which is detailed in Section II. A comparison is also made with Segmenter [11] and DeepLabv3+ [8] to analyze the strengths and weaknesses of the model. Several variants of the model 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 and each also has their own uses which is detailed in Section II.

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

# Related Works

To interpret bone scan qualitatively, Imbriaco et al. [[3](#_M._Imbriaco_et)] proposed the Bone Scan Index (BSI), an index that represents the regions of bone infected by 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 basis model. This experiment by Sadik et al. [[4](#_M._Sadik_et)] used data of patients 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. [13] 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 to Japanese patients. Another experiment using 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 active shape model (ASM) which reaches 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.

In the field of semantic segmentation, convolutional neural network with its convolution has been state-of-the-art for some time. Chen et al. [[14](#_L.-C._Chen,_Y.)] developed DeepLabv3+ relies on convolution for semantic segmentation. By combining Xception model and atrous convolution, DeepLabv3+ performs well in the PASCAL VOC 2012 and Cityscapes dataset. DeepLabv3+ has its share of uses in the medical field [[15](#_Wang,_Jing,_and), [16](#_Azad,_Reza,_et)].

On the other side, a new method in natural language processing called transformer [[9](#_Vaswani_et_al.,)] started being integrated to computer vision [[10](#_Dosovitskiy_et_al.,)]. Transformers can extract global context and can run in parallel. Strudel et al. [[11](#_R._Strudel,_R.)] has done research in this case and proposes Segmenter, a model which uses encoder-decoder architecture and is based on pure transformer approach. This model outperforms the state-of-the-art methods in ADE20K, Pascal Context, and Cityscapes dataset. In an evaluation of transformer-based semantic segmentation model, Segmenter performs well in pathological image segmentation [[17](#_Nguyen,_Cam,_et)].

An experiment combining both approaches by Fu et al. [[12](#_Fu,_Jun,_et)] proposes 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](#_Li,_Yong-xin,_et), [19](#_Chen,_Lifang,_Tengfei), [20](#_Hu,_Rui,_et)].

# 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% of the dataset is used 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 performance of the model. The illustration is shown in Fig. 1.

A diagram of a software development process

Description automatically generated

1. Proposed system overview.

## Dual Attention Network

Dual Attention Network (DANet) is a semantic segmentation model which uses ResNet as its backbone but also uses two attention modules to capture long-range contextual information in spatial and channel dimension respectively. A pretrained ResNet with the dilated strategy is employed as the backbone. The downsampling operation is removed and dilated convolutions are employed in the last two ResNet blocks to retain more details without adding extra parameters. The features extracted are then fed into two parallel attention modules. The model illustration is shown in Fig. 2.

A picture containing text, diagram, screenshot, plan

Description automatically generated

1. Overview of the Dual Attention Network (captured from [[12](#_Fu,_Jun,_et)]).

### Position Attention Module

Position attention module is introduced to model rich contextual association between local features. This module improves the representation capability by capturing larger scope of contextual information into local features. The illustration of this module is shown in Fig. 3. Given a local feature **A** ∊ ℝ*C*✕*H*✕*W*, **A** is fed to a convolutional layers to produce three feature maps **B**, **C**, and **D**, where {**B**, **C**, **D**} ∊ ℝ*C*✕*H*✕*W*. These feature maps are then reshaped into ℝ*C*✕*N* where *N = H ✕ 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 **S** ∊ ℝ*N✕N* :



where sji measures the ith position’s impact on jth position. Greater 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 ℝ*C*✕*H*✕*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 **E** ∊ ℝ*C*✕*H*✕*W* as follows:

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 the feature **E**. As a result, it has selectively aggregates the contexts according to the spatial attention map and has contextual view globally. Intra-class compactness and semantic regularity has been improved due to the mutual gains achieved by the similar semantic features. Position attention module is introduced to model rich contextual relationships between local features. This module enhances the representation capability by capturing larger range of contextual information into local features.

### Channel Attention Module

Class specific response can be obtained by each channel map of high-level features, and various semantic responses can be associated with each other. To improve in representing a feature of particular semantics and the emphasis on interdependent feature maps, it can be done by taking advantage of the interdependencies between each channel maps. The channel attention module is built to model the interdependencies between the channels explicitly. The illustration of this module is shown in Fig. 4.

A picture containing text, diagram, screenshot, font

Description automatically generated

1. Position attention module illustration (captured from [[12](#_Fu,_Jun,_et)]).

A picture containing text, screenshot, line, diagram

Description automatically generated

1. Channel attention module illustration (captured from [[12](#_Fu,_Jun,_et)]).

This channel attention map **X** ∊ ℝ*C*✕*C* is calculated directly from the original features **A** ∊ ℝ*C* ✕*H*✕*W*. **A** is reshaped to ℝ*C*✕*N*, and then a multiplication matrix is done between **A** and the transpose of **A**. A softmax layer is applied to obtain the channel attention map **X** ∊ ℝ*C*✕*C* :

3

where xji­ measures the i­th channel’s impact on the jth channel. Another matrix multiplication is done between the transpose of **X** and **A** and reshaped to ℝ*C*✕*H*✕*W*. The result is then multiplied by another scale parameter β to which is done an element-wise sum operation with **A** to obtain the final output **E** ∊ ℝ*C*✕*H*✕*W* :

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.

### 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. It is then followed by another convolutional layer to produce the final prediction map.

## 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 the following equation:

5

and a model’s overall performance can be measured by averaging the IoU of each class (mIoU).

# Experimental Results and Analysis

In this section, we give the explanation about the dataset, model training phase, experiment result, and analysis.

## Dataset

The dataset is composed of 18 Indonesian and 19 non-Indonesian patients. Each patient has anterior and posterior bone scan images, resulting in a total of 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 x 512 pixels [22]. 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, because the bone regions ‘sternum’ and ‘collarbone’ were not visible from behind, posterior section only have 10 classes. Samples of the raw images and the annotations are shown in Fig. 5.

1. Samples of raw bone scan images and its annotaions.

## Result and Analysis

We divide the dataset into anterior and posterior groups. Each group containing 37 images divided into 24 images for training, 6 for validating, and 7 for testing. We used the tool MMSegmentation [[21](Open-mmlab/mmsegmentat#_Open-Mmlab,_)] for the training, testing, and inference of the model. For the configuration, we set the parameters such as follows: batch size to 3, the image is resized to 256 x 1024 pixels which then cropped into a 512 x 512-pixel image, and training iteration to be 5000. While training, evaluation was done every 500 iterations to monitor the process. This evaluation reports the performance of the model in IoU of each class and mean IoU of all classes.

1. 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** |

1. mIoU graph of DANet variants in anterior section training phase.

After the training process, the model is tested using the 7 testing images and then evaluated. The evaluation also reports the IoU of each class and mean IoU. This process is done to each of 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 anterior section performs at mIoU at range 75% to 76%. Increasing the depth parameter improves the performance slightly. Having the biggest number of positional attention module channels does not perform at the best in this case. The value of 32 in pam\_channels perform the best and followed by 16 channels. This is indicated by slower growing performance of the 64 channels as shown in Fig. 6.

1. 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** |

In the case of 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 anterior section, all variants in the posterior section tend to follow the similar mIoU training phase curve as shown in Fig. 7.

## Comparison of DANet, Segmenter, and DeepLabv3+

In this section, we present the comparison of segmentation results using DANet, Segmenter [11], and DeepLabv3+ [14]. Segmenter used purely self-attention mechanism in its architecture, while DeepLabv3+ focused on convolutional part in 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.

1. mIoU graph of DANet variants in posterior section training phase.
2. 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 |

For anterior section, the DANet models perform on par and slightly better than DeepLabv3+ models as shown in Table III. 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%, 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 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 cervical vertebrae class, DANet’s attention mechanisms can be seen to perform well corresponding to Segmenter’s similar performance and convolution’s struggle is shown in DeepLabv3+’s performance. The same thing can be said for convolution parts and against the attention mechanisms in classes such as sternum, collarbone, scapula, and femur. By combining both approaches, DANet can gain the benefit by outperforming both models in the remaining classes. At 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.

1. 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** |

1. Visual comparison between the raw image, ground truth, and model segmentation in anterior dataset.
2. 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** |

In the case of posterior section, all DANet variants outperform even the best of DeepLabv3+’s model. DANet’s smallest variant can be seen at 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 dataset.

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%. For the posterior dataset, an average mIoU is achieved at 79% between all variants. Unlike the anterior dataset, the biggest channel in the biggest depth gives the best performance at mIoU of 80.99%.

1. Visual comparison between the raw image, ground truth, and model segmentation in posterior dataset.

As seen in Table VI, DANet outperforms the other two in most of the classes. Most of it were small gaps between DeepLabv3+, but it shows the advantage of using both approaches. By using these combinations, DANet’s performance can be seen in the class lumbar vertebrae and sacrum. 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. In posterior dataset, even DANet’s smallest depth and positional attention module channel variant still performs better than DeepLabv3+’s biggest variant.

1. 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** |

For 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 also outperforms the other two models in most of the classes. In some classes, DANet is shown to take advantage of the combination of both approaches, giving performance like one of the models in contrast to the other one. This shows that DANet’s model of combining the two approaches gives significant value to performance. DeepLabv3+ is shown to perform the best in the anterior dataset with the variant Deep-100 with mean intersection over union at 75.82. DANet is shown to perform the best in the posterior dataset with the variant DANet-100 with mean intersection over union at 80.99. Segmenter is shown to struggle with the current dataset and model tuning compared to DeepLabv3+ and DANet.

# Conclusion

Bone scan image segmentation plays a huge role in early detection in 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, in which we also compared it with Segmenter and DeepLabv3+. In this study, the proposed system generally 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.

##### References

###### Siegel, Rebecca L., Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. "Cancer statistics, 2022." CA: a cancer journal for clinicians 72, no. 1 (2022): 7-33.

###### Gondhowiardjo, Soehartati, Nadia Christina, Ngakan PD Ganapati, Salik Hawariy, Fahmi Radityamurti, Vito F. Jayalie, Steven Octavianus et al. "Five-year cancer epidemiology at the national referral hospital: hospital-based cancer registry data in Indonesia." JCO global oncology 5, no. 1 (2021): 190-203.

###### Imbriaco, Massimo, Steven M. Larson, Henry W. Yeung, Osama R. Mawlawi, Yusuf Erdi, Ennpadam S. Venkatraman, and Howard I. Scher. "A new parameter for measuring metastatic bone involvement by prostate cancer: the Bone Scan Index." Clinical cancer research: an official journal of the American Association for Cancer Research 4, no. 7 (1998): 1765-1772.

###### Sadik, May, Iman Hamadeh, Pierre Nordblom, Madis Suurkula, Peter Höglund, Mattias Ohlsson, and Lars Edenbrandt. "Computer-assisted interpretation of planar whole-body bone scans." Journal of Nuclear Medicine 49, no. 12 (2008): 1958-1965.

###### Rachmawati, Ema, Jondri Jondri, Kurniawan Nur Ramadhani, Achmad Hussein Sundawa Kartamihardja, Arifudin Achmad, and Rini Shintawati. "Automatic whole-body bone scan image segmentation based on constrained local model." Bulletin of Electrical Engineering and Informatics 9, no. 6 (2020): 2526-2537.

###### Rachmawati, Ema, Fazrian Ramadlan Sumarna, Achmad Hussein Sundawa Kartamihardja, Arifudin Achmad, and Rini 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), pp. 1-6. IEEE, 2020.

###### Shimizu, Akinobu, Hayato Wakabayashi, Takumi Kanamori, Atsushi Saito, Kazuhiro Nishikawa, Hiromitsu Daisaki, Shigeaki Higashiyama, and Joji Kawabe. "Automated measurement of bone scan index from a whole-body bone scintigram." International Journal of Computer Assisted Radiology and Surgery 15 (2020): 389-400.

###### Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40, no. 4 (2017): 834-848.

###### Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

###### Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

###### Strudel, Robin, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. "Segmenter: Transformer for semantic segmentation." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 7262-7272. 2021.

###### Fu, Jun, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. "Dual attention network for scene segmentation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3146-3154. 2019.

###### Kikuchi, Akihiro, Masahisa Onoguchi, Hiroyuki Horikoshi, Karl Sjöstrand, and Lars Edenbrandt. "Automated segmentation of the skeleton in whole-body bone scans: influence of difference in atlas." Nuclear medicine communications 33, no. 9 (2012): 947-953.

###### Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In Proceedings of the European conference on computer vision (ECCV), pp. 801-818. 2018.

###### Wang, Jing, and Xiuping Liu. "Medical image recognition and segmentation of pathological slices of gastric cancer based on Deeplab v3+ neural network." Computer Methods and Programs in Biomedicine 207 (2021): 106210.

###### Azad, Reza, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation." In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pp. 251-266. Springer International Publishing, 2020.

###### Nguyen, Cam, Zuhayr Asad, Ruining Deng, and Yuankai Huo. "Evaluating transformer-based semantic segmentation networks for pathological image segmentation." In Medical Imaging 2022: Image Processing, vol. 12032, pp. 942-947. SPIE, 2022.

###### Li, Yong-xin, Hui Tang, Wei Wang, Xiu-feng Zhang, and Hang Qu. "Dual attention network for unsupervised medical image registration based on VoxelMorph." Scientific Reports 12, no. 1 (2022): 16250.

###### Chen, Lifang, Tengfei Mao, and Qian Zhang. "Identifying cardiomegaly in chest x-rays using dual attention network." Applied Intelligence 52, no. 10 (2022): 11058-11067.

###### Hu, Rui, Hui Yan, Fudong Nian, Ronghu Mao, and Teng Li. "Unsupervised computed tomography and cone-beam computed tomography image registration using a dual attention network." Quantitative Imaging in Medicine and Surgery 12, no. 7 (2022): 3705.

###### Open-Mmlab, “Open-mmlab/mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark.,” GitHub, https://github.com/open-mmlab/mmsegmentation (accessed Jun. 4, 2023).

###### Nugraha, Dimas Bayu, Ema Rachmawati, and Mahmud Dwi Sulistiyo. "Semantic Segmentation of Whole-Body Bone Scan Image Using Btrfly-Net." In 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 264-269. IEEE, 2022.