Review of Deep Learning Models for Spine Segmentation

Review of Deep Learning Models for Spine Segmentation

Neng Zhou*
Jiangnan University
Wuxi, Jiangsu Province, China

Hairu Wen*
China University of Mining and technology(Beijing)
Beijing, China

Yi Wang
Jiangnan University
Wuxi, Jiangsu Province, China

Yang Liu
Beijing Wuzi University
Beijing, China

Longfei Zhou†
Massachusetts Institute of Technology
MA, USA
longfei@mit.edu

ABSTRACT

Medical image segmentation has been a long-standing challenge due to the limitation in labeled datasets and the existence of noise and artifacts. In recent years, deep learning has shown its capability in achieving successive progress in this field, making its automatic segmentation performance gradually catch up with that of manual segmentation. In this paper, we select twelve state-of-the-art models and compare their performance in the spine MRI segmentation task. We divide them into two categories. One of them is the U-Net family, including U-Net, Attention U-Net, ResUNet++, TransUNet, and MiniSeg. The architectures of these models often ultimately include the encoder-decoder structure, and their innovation generally lies in the way of better fusing low-level and high-level information. Models in the other category, named Models Using Backbone often use ResNet, Res2Net, or other pre-trained models on ImageNet as the backbone to extract information. These models pay more attention capturing multi-scale and rich contextual information. All models are trained and tested on the open-source spine MRI dataset with 20 labels and no pre-training. Through the comparison, the models using backbone exceed U-Net family, and DeepLabv3+ works best. We suppose it is also necessary to extract multi-scale information in a multi-label medical segmentation task.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

Deep Learning, Automatic Segmentation, Magnetic Resonance Imaging, Spine Segmentation

ACM Reference Format:
Neng Zhou, Hairu Wen, Yi Wang, Yang Liu, and Longfei Zhou. 2022. Review of Deep Learning Models for Spine Segmentation. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3512527.3531356

*Both authors contributed equally to this research.
†Corresponding author

ICMR '22, June 27–30, 2022, Newark, NJ, USA.

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531356

1 INTRODUCTION

Image segmentation is a pixel-level image classification method popular in computer vision tasks, converting input images into areas of interest with highlights. Traditional segmentation methods like edge detection [1], threshold-based segmentation [2], region growth [3], activity-based contour-based strategies [4], and cluster-based segmentation methods [5] were not that efficient due to the cumbersome process and the noise effect of medical imaging equipment. The widely used way in labeled data processing like semantic segmentation is deep learning. In 2017, Litjens, Geert et al. [6] reviewed deep learning pertinent concepts, and summarized over 300 contributions to medical image analysis, most of which appeared within one year.

The study of vertebral medical imaging is significant to assist clinical diagnosis and solve vertebral problems. Spine segmentation is one of its main tasks. It's challenging due to the complex shape and structure of the vertebrae, similar design, pathology, and spatial interrelationship between the vertebrae and ribs [7], as well as the unevenness of the sample size and blur boundary of different classes. However, many challenging problems have been solved with the help of various equipment and methods in recent years. For instance, Adela Arpith and Rangarajan Lalitha [8] developed a system for computer-aided diagnosis to help in the detection, labeling, and segmentation of lumbar compression fractures. Leena Silvoster and MathuSoothana S.Kumar Retnaswami [9] proposed an automatic strategy to extract the 3D segmentation of the standard disc as well as degenerated lumbar intervertebral discs (IVDs) from T2-weighted Turbo Spin Echo MRI of the spine. The challenges faced include partial volume effects, intensity inhomogeneity and gray level overlap of different soft tissues. Al Kafri Ala S. et al [10] experimented with semantic segmentation of a dataset containing MRI studies of 515 patients with symptomatic back pains using SegNet. Suzani A et al. [7] used deep learning models to automatically locate, identify, and segment vertebrae in MR images and made improvements over the previous semi-automated approach [11]. Li S et al[12] used a deep learning model to locate, implement segmentation, and avoid overfitting by allowing each task to correct each other. Jiawei Huang et al. [13] developed a deep learning based program called Spine Explorer for automated segmentation and

quantification of the vertebrae and intervertebral discs on lumbar spine MRIs.

Despite the studies described earlier, there are few unified evaluations and performance comparisons of different methods on lumbar MRI image segmentation tasks. So we write this paper to compare the performance of several well-known deep learning neural network models on the lumbar MRI image, and to analyze the possible causes of the differences in the results. We make two main contributions as follows.

(1) Use a multi-class dataset, that is, a lumbar spine MRI dataset with a total of 20 classes, to evaluate the performance of the models.

(2) Twelve networks, U-Net, PSPNet, DeepLabv3+, DenseASPP, Attention U-Net, DANet, EMANet, Inf-Net, MiniSeg, PSANet, ResUNet++, and TransUNet, are compared in the medical segmentation task on the same dataset of lumbar MR image.

The rest of this paper organize as follows. Section 2 and 3 give models used in the experiment. Section 4 is the experimental part including, the dataset, preprocessing, evaluation metrics, model result, and discussion. Section 5 is the conclusion.

2 U-NET FAMILY

U-Net [14] is one of the most famous models in medical image segmentation due to its highly symmetric and U-shaped architecture. It's proven to be an efficient model even when the target object is small relative to the background, and many researchers are making progress to improve the performance. In this work, we selected five classic variants and compared their performance.

2.1 U-net

Due to the noise and artifacts in medical datasets, common CNN models are often difficult to obtain accurate results, especially on tiny objects. Based on these problems, U-Net [14] model as shown in Figure 1 was designed to be composed of encoder and decoder. The encoder structure used 3 X 3 convolutions and max-pooling to extract more detailed information, while the decoder structure used several 3 X 3 convolutions and up-convolutions to restore the data. U-Net also creatively designed skip-connections to link low-level and high-level features, making the model concentrate on low-resolution and high-resolution feature maps. Nowadays, U-Net has become one of the most robust structures in medical image segmentation and has also promoted the emergence of various other variants.

2.2 Variants of U-Net

(1) Attention U-Net: The siginificant divergence of the target organs between different patients brought out multi-stage cascaded CNNs. These models implementing segmentation within an extracted region of interest (ROI), which may lead to more parameters and unnecessary consumption. To solve these problems, Attention U-Net [15] first applied the soft attention technique to a medical image task by using the proposed Attention gates as shown in Figure 2. It fused the low-level and high-level information by using Relu, 1x1 convolution and Sigmoid to better focus in the regions needing to be segmented while limiting the irrelevant background regions.

(2) ResUNet++ [16] which was inspired by ResUNet connected previously proposed squeeze and excitation Unit (SE unit), Atrous

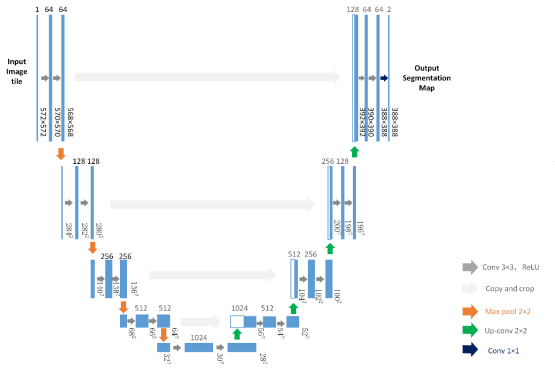

Figure 1: The U-Net architecture shown in [14].

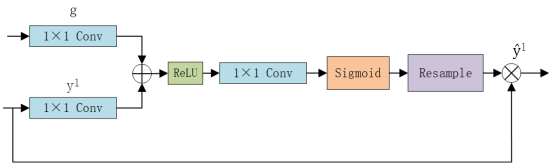

Figure 2: The attention gate in Attention U-Net [15].

Spatial Pyramidal Pooling (ASPP) and attention unit in a precise way. The model is shown in Figure 3. Like U-Net, the model consists of an encoder and decoder, each of which has three corresponding blocks. Both SE unit and attention unit are used to strengthen the ability of the model to capture representative features and suppress the irrelevant features, especially the background features. ASPP links the encoder and decoder like a bridge. With the hierarchical style pooling modules in the ASPP, different scales of extracted information are fused, allowing the model to capture the more contextual information simultaneously.

(3) TransUNet: J. Chen et al. [17] proposed TransUNet which takes advantage of both Transformer and U-Net. The Transformer is a widely used and solid model in natural language processing (NLP), whose highly efficient performance was verified in machine translation tasks. It first proposed the attention mechanism and replaced the classical CNN with the attention and feed-forward module. The Transformer was later used in computer vision tasks and achieved state-of-the-art performance. The architecture is shown in Figure 4. In the encoder, the input image goes through CNN to extract low-level information before being sent to the Transformer layer to generate global contextual information. A cascaded upsampler (CUP) is introduced in the decoder to upsample the feature map to the full resolution step by step. Skip connection is used to fuse low-level information and high-level information. We select TransUNet because it is the first Transformer-based medical image segmentation model that achieved state-of-the-art on synapse multi-organ CT dataset and we want to see its generalization ability on other medical datasets such as the MRI dataset in our paper.

(4) MiniSeg[18] is proposed to meet the demand for efficiency and lightweight with only 83K parameters. The labeled medical dataset is often limited, thus a model with millions of parameters to train may cause overfitting, not to mention the computational

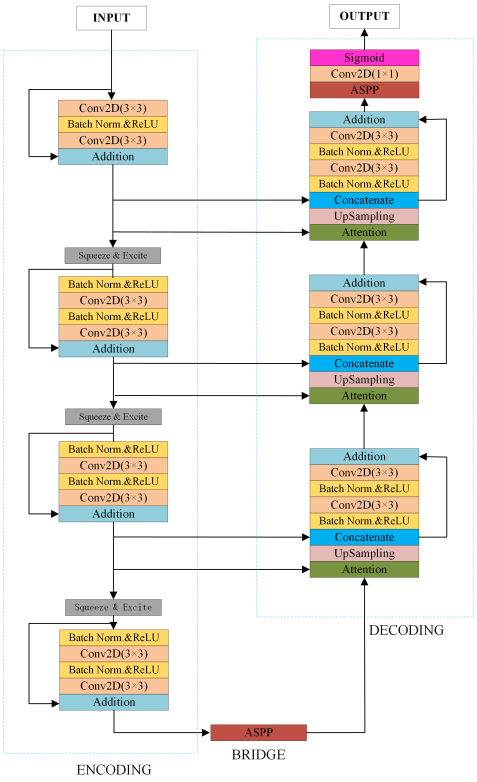

Figure 3: The architecture of ResUNet++ [16].

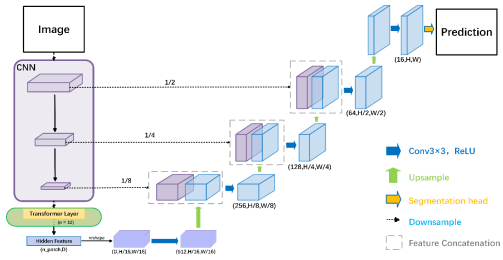

Figure 4: The architecture of TransUNet [17].

resources spent. Encouragingly, MiniSeg achieved the best performance and is most efficient among state-of-the-art medical image segmentation models on several COVID-19 datasets. In the whole model architecture Figure 5, the encoder-decoder structure is introduced to first extract the essential information and then upsampled to complete resolution. Its most remarkable innovation is the design of the Attentive Hierarchical Spatial Pyramid (AHSP) Figure 6module. Multi-scale features learned by dilated convolutions of AHSP were merged through the attention block for more focus on the critical part. Feature Fusion Module (FFM) module is proposed to fuse low-level information and high-level information.

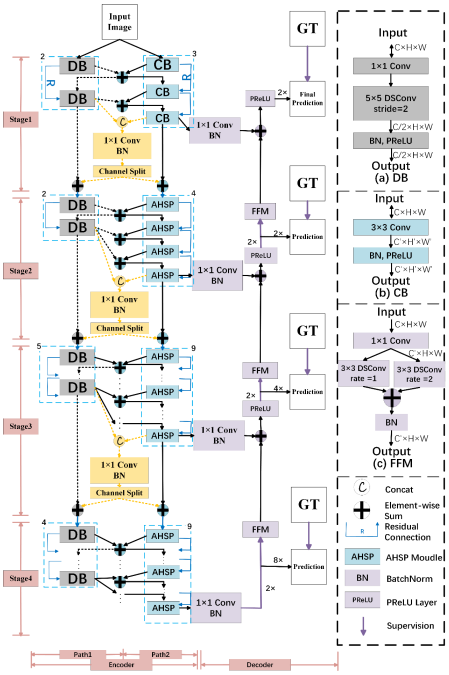

Figure 5: The architecture of MiniSeg [18].

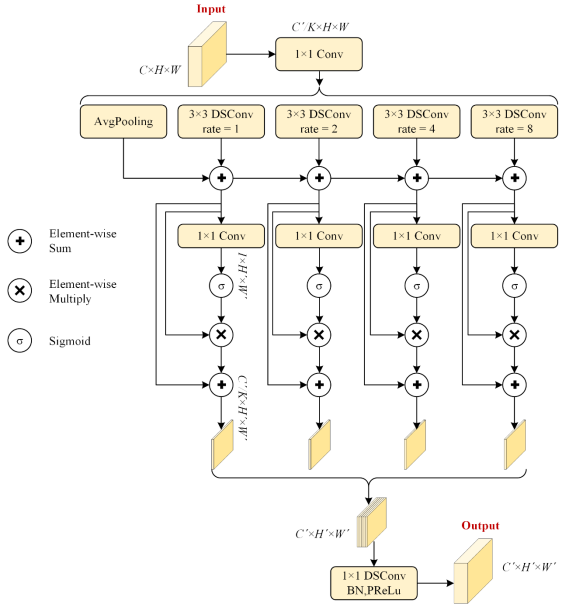

Figure 6: The Attentive Hierarchical Spatial Pyramid (AHSP) module [18].

3 MODELS USING BACKBONE

A backbone network is always used in computer vision, especially segmentation tasks where every pixel counts, to extract image features as they are proved effective on the ImageNet dataset [19].

Different backbone networks are proposed to meet with the demand of different downstream tasks. This section will introduce two backbone networks commonly used in segmentation tasks and seven famous models using these two backbones. These models focus more on achieving richer contextual and multi-scale information that might be helpful for segmentation of the medical MRI dataset.

3.1 Backbone

(1) ResNet proposed by K. He et al. [20] consists of multiple layers of so-called bottleneck blocks Figure 7. Each block is a stack of three layers, including 1 X 1, 3 X 3, 1 X 1 convolution layers, and has a short cut directly links the input and the output. The shortcut permits linear transformation for the information of every trained neural network, which means the performance will not degrade as the number of layers increases or forbade researchers from building deeper neural networks.

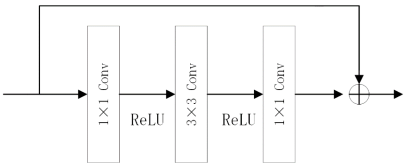

Figure 7: The bottleneck block in ResNet [20].

(2) Res2Net: Res2Net [21] was proposed to optimize the multi-scale feature representations, which are important in various computer vision tasks. Res2Net module Figure 8 replaces the original 3x3 filter in ResNet with a group of them. Each is connected in a hierarchical residual-like style. The Res2Net module split the original feature map into several subsets with the same spatial size but divided channels. With more shortcuts in the module, Res2Net can have a wider receptive field size and learn more features from different scales. Though this may cause more time to train the model, Res2Net does outperform ResNet [20], ResNeXt [22] and DLA [23] on the ImageNet dataset.

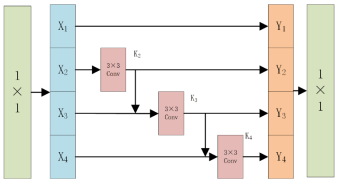

Figure 8: The Res2Net module [21].

3.2 Models

(1) PSPNet: Pyramid scene parsing network (PSPNet) [24] was proposed for complex-scene parsing issues, including wrong segmentation results under similar appearance, confusion of different labels for a single object, and poor performance of relatively small things. PSPNet designed pyramid pooling module Figure 9 inspired by global average pooling commonly used in image classification. Instead of directly applying global average pooling, which might cause information loss, the pyramid pooling module uses different sizes of convolution kernels to pool the global contextual information extracted by CNN in different sub-regions before being flattened and concatenated. This feature map in different scales is finally upsampled to the original size to restore the extracted data and get the prediction. PSPNet used the dilated ResNet50 network as its backbone in [24] while we use ResNet101 in this paper.

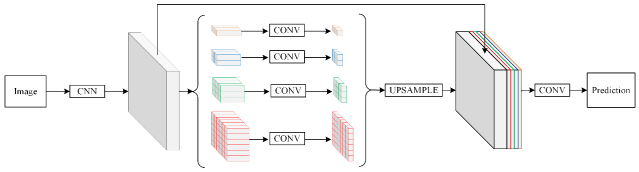

Figure 9: The architecture of PSPNet [24].

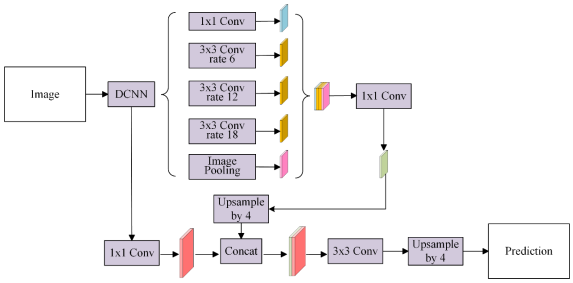

Figure 10: The architecture of DeepLabv3+ [25].

(2) DeepLabv3+: DeepLabv3+ [25] employed DeepLabv3 as the encoder and added another practical decoder module to capture sharper boundaries in semantic segmentation. As in Figure 10, the encoder, that is, DeepLabv3, applied the Atrous Spatial Pyramid Pooling (ASPP) module consisting of several parallel atrous convolutions with different rates. This module enabled modifiable density of the encoder features, thus can capture multi-scale contextual information. In the decoder, the extracted information is upsampled twice by the factor of 4 instead of 16 in DeepLabv3 to recover more detailed information. DeepLabv3+ used a modified Xception which performs better on ImageNet dataset as backbone in [25] while we use ResNet101 in this paper.

(3) DenseASPP: The atrous convolutions in Atrous Spatial Pyramid Pooling (ASPP) module is well-known for enlarging the size of the receptive field without increasing the parameters. But it still has the problem of missing enormous information, especially when dilated rate rises. M. Yang et al. [26] proposed DenseASPP apply dense connection to the ASPP module to solve this problem. As shown in Figure 11, the extracted feature map is sent to the DenseASPP module with a series of dilated convolutions of increasing dilated size. Each output is fed to the subsequent ones, thereby avoiding the problem of degradation in ASPP. DenseASPP achieved state-of-the-art compared to DeepLabv2-CRF, PSPNet, GCN, and so on, using DenseNet161(wider) as its backbone. We use DenseNet169 as its backbone in this paper.

(4) DANet: Dual Attention Network (DANet) proposed by J. Fu et al. [27] can capture richer contextual dependencies by using the self-attention mechanism. To capture more long-range contextual

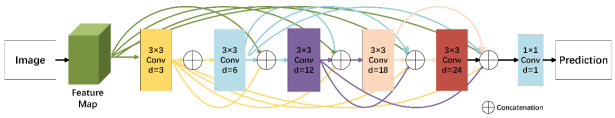

Figure 11: The architecture of DenseASPP [26].

information, DANet combined the position attention module and channel attention module Figure 12together . The position attention module makes sure similar features at different distances make the same contribution to the final feature maps. The channel attention module introduced a self-attention mechanism as aggregated features in the channel dimension. We used ResNet101 as its backbone as in [27]. As DANet achieved state-of-the-art on three popular datasets, we want to confirm whether these two attention modules perform well on the medical MRI dataset.

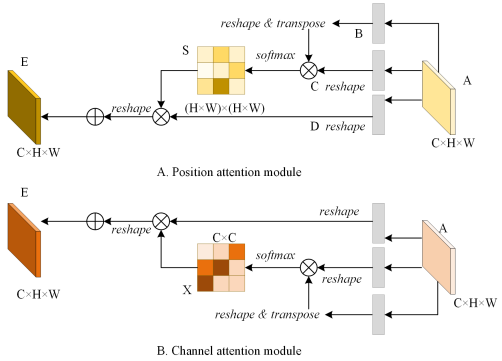

Figure 12: The Dual Attention module (DA) [27].

(5) EMANet: Inspired by expectation-maximization (EM) algorithm, Li et al. [28] first introduced EM iterations into attention mechanism in Expectation-Maximization Attention Networks (EMANet). AE is used to calculate the attention map expectation while AM maximizes the data likelihood. Thus, EMA Unit Figure 13 the key component, can obtain a maximum likelihood estimate of parameters (base) by running the AE and AM steps alternatively. Instead of computing the attention map upon the whole image, the EMA Unit can build the attention map upon a much more compact set of bases, therefore making the EMA Unit more light-weighted and easy to be embedded into an existing neural network.

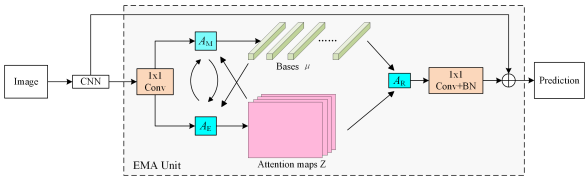

Figure 13: The EMA Unit [28].

(6) Inf-Net: To better learn the boundary between different classes explicitly, D. Fan et al. [29] proposed Inf-Net, using a parallel partial decoder (PPD) Figure 14 and reverse attention module (RA) Figure 15. Inspired by a two-step procedure in clinical practice, framing and labeling, Inf-Net used PPD to generate a coarse global feature map by aggregating high-level information and then RA module as a fine labeler gradually erasing the infected regions. Inf-Net learned more edge information by the edge attention module with a single convolution layer to generate an edge map directly from ground truth. We used Res2Net101 as backbone as in [29].

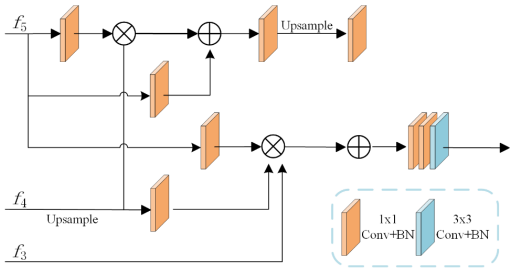

Figure 14: The parallel partial decoder (PPD) [29].

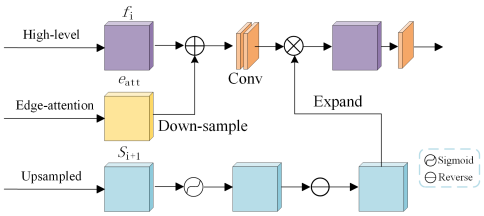

Figure 15: The reverse attention module (RA) [29].

(7) PSANet: Zhao et al. [30] proposed a point-wise spatial attention network (PSANet) to capture long-range dependency. PSA module Figure 16 split the input into two branches, collect and distribute. Each of them is used to aggregate contextual information under the guidance of the attention maps, which connect every position with all others. The PSA module harvest long-range dependency using an attention map that is more self-adaptive than non-local methods. We used Res2Net101 as its backbone as in [30].

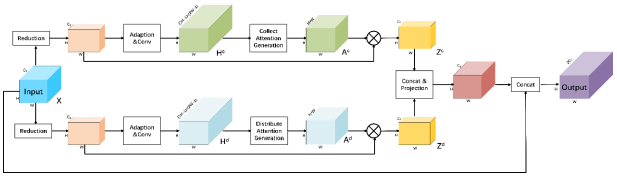

Figure 16: The PSA module [30].

4 EXPERIMENT

4.1 Dataset and Preprocess

The experiment dataset is a T2-weighted MR image sequence of 200 patients. It contains 200 T2-weighted MR images, which are finely annotated. All MR images consist of 20 labels, including thoracic,

lumbar, sacral vertebrae, and background. All images are in the format of NIFTI (*.nii.gz) and have not been preprocessed for bias field correction or intensity normalization. This dataset can be used for medical image segmentation in all images of the T2-weighted MR sequence given by each patient.

In this paper, we slice all of the NIFTI inputs from 3D form to 2D form in the depth dimension with 20 labels Figure 17. Since the third dimension (depth dimension) of the original NIFTI input is 12-15, each 3D information is converted into 12-15 2D tensors accordingly. As the authentic MR images have a high resolution, we keep most of the images' resolution for as many details as possible. For consistency, all MR images are resampled to 880x880.

We use TorchIO [31] to load all the experimental MR images and make the following augmentation transforms: random noise, random blur, random motion, and ghost. Thus, all the models we compared in this paper are robust to noise and artifacts so we can better compare the differences between them based on the model structures.

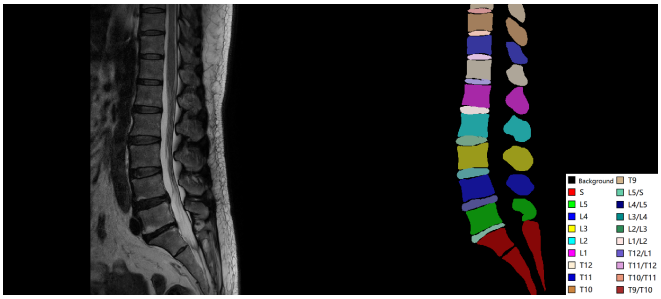

Figure 17: The sliced 2D MRI from original 3D MRI dataset. Left: the original image. Right: corresponding labels, where S is sacral, L is lumbar, T is thoracic and T9/T10 is the disc between T9 and T10.

4.2 Loss

Dice loss Is proposed by Milletari et al. [32] and its formula is defined as

$$DiceLoss(A,B) = (1 - \frac{2\times \mid A \cap B \mid}{A+B}) \times 100\% \quad (1)$$

Where A is the prediction and B is the ground-truth.

In our experiments, usage of the Dice loss has the highest score in both Dice and IOU, compared to the cross-entropy loss and weighted cross-entropy loss. In addition, Dice loss is much more suitable for uneven samples. Therefore, we employed Dice loss for all the models during the training.

4.3 Evaluation Metrics

(1)Dice coefficient: Dice is one of the most essential evaluation metrics in medical image segmentation. It describes the overlap degree between the prediction and ground-truth from 0 to 1. Similar to IOU, a higher dice score indicates a better segmentation performance. The formula for Dice is defined as:

$$Dice(A,B) = \frac{2\times \mid A \cap B \mid}{A+B} \times 100\% \quad (2)$$

Table 1: Comparison between twelve models

Method	Dice	mIOU	FN_rate	FP_rate	Recall
U-Net	79.87	74.02	34.04	1.21	83.90
PSPNet	87.38	81.79	22.19	2.69	**92.24**
DeepLabv3+	**87.55**	**81.80**	18.45	**0.11**	90.81
DenseASPP	71.44	67.13	57.42	0.15	74.67
Attention U-Net	75.55	70.00	36.33	0.18	78.98
DANet	84.74	77.35	30.82	0.25	88.95
EMANet	87.32	81.63	**15.88**	0.13	90.47
Inf-Net	81.24	75.42	42.28	0.37	90.92
MiniSeg	85.69	79.25	42.29	2.40	90.63
PSANet	87.04	81.18	27.45	0.16	89.94
ResUNet++	83.14	77.10	18.89	0.46	86.78
TransUNet	76.57	70.53	40.62	0.31	79.56

Where A is the segmentation results and B is the ground-truth labels.

(2)IOU: Intersection Over Union (IOU) is a commonly used metric in object detection. IOU precisely evaluates the overlap degree between the predicted bounding box and the actual bounding box, which makes sense for medical image segmentation. The formula is defined as:

$$IOU(A,B) = \mid \frac{A \cap B}{A \cup B} \mid \times 100\% \quad (3)$$

Where A is the prediction map and B is the ground-truth.

(3)Other metrics we used are defined as follows:

$$FalseNegativeRate(FNR) = \frac{FN}{FN+TP} \quad (4)$$

$$FalsePositiveRate(FPR) = \frac{FP}{FP+TN} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

4.4 Implementation Details

The input layer of all networks is replaced with a new one of the same size as the input spinal slice(i.e.,1x880x880). The number of channels output by the last layer of all networks was set to 20 to match the number of classes. During the training, all networks used Adam optimizer and initial learning rate was set to 5e-4. The dataset was divided into training set and validation set in the proportion of 8:2, and the training set is augmented and shuffled. All the models are trained on an Nvidia RTX3090 GPU to compare the performance between different network structures.

4.5 Results

Table.1 indicates the overall performance on validation dataset between twelve models. Networks could segment different part of the spine with Dice in the range from 71.44% to 87.55% and mIOU in the range from 67.13% to 81.8%. DeepLabv3+ achieved the best scores in Dice (87.55%), mIOU (81.8%) and FP_rate (0.11%). Likewise, PSPNet also made the second-best segmentation in Dice (87.38%), mIOU (81.79%) and then goes EMANet in Dice (87.32%), mIOU (81.63%).

We also listed the performance of these twelve models in 19 classes, excluding the background. From Table.2, the relationship between mDice and the performance on 19 types respectively is

Table 2: Dice coefficient in 19 classes between twelve models

Method	Label																			mDice
	S	L5	L4	L3	L2	L1	T12	T11	T10	T9	L5/S	L4/L5	L3/L4	L2/L3	L1/L2	T12/L1	T11/T12	T10/T11	T9/T10	
U-Net	**88.80**	83.93	72.20	70.92	71.44	71.77	73.32	68.87	85.21	**94.67**	16.77	80.00	75.83	78.06	80.12	84.37	82.83	79.88	96.45	76.59
PSPNet	87.82	**87.72**	85.51	85.58	83.08	81.63	78.48	74.97	85.21	**94.67**	84.76	85.20	87.91	**88.13**	89.13	91.47	87.06	79.88	96.45	86.03
DeepLabv3+	87.75	87.10	85.17	**86.07**	**84.02**	**83.52**	80.02	**76.58**	85.21	**94.67**	**85.60**	85.50	87.39	87.70	**89.39**	**91.62**	**90.47**	79.88	96.45	**86.53**
DenseASPP	87.34	87.31	84.67	84.26	82.53	81.77	**80.35**	74.07	85.21	**94.67**	16.77	14.20	85.59	87.14	26.04	32.35	41.22	79.88	96.45	69.57
Attention U-Net	87.97	83.56	70.29	67.55	63.11	69.77	70.44	9.86	85.21	**94.67**	84.85	77.54	71.13	71.67	71.26	81.96	41.22	79.88	96.45	72.55
DANet	81.55	82.42	80.22	79.70	75.94	74.18	72.24	67.93	85.21	**94.67**	82.74	82.54	86.43	84.89	87.07	86.96	86.44	91.60	96.45	83.12
EMANet	87.35	87.52	**86.13**	85.78	83.54	82.79	79.09	74.15	85.21	**94.67**	84.16	83.56	87.33	87.52	89.23	90.61	86.11	79.88	96.45	85.85
Inf-Net	87.80	87.25	85.01	83.62	81.28	83.01	79.23	71.76	85.21	0.97	85.03	**86.07**	**88.07**	87.51	84.85	90.94	87.25	89.85	96.45	81.11
MiniSeg	86.29	86.05	84.03	83.10	79.42	77.79	74.88	70.85	**88.83**	**94.67**	83.66	84.65	85.72	86.67	87.66	88.49	85.32	**92.59**	**97.30**	85.16
PSANet	87.41	87.56	86.04	85.70	81.55	80.88	78.84	75.07	85.21	**94.67**	84.80	84.06	87.15	86.38	87.94	89.75	87.01	79.88	96.45	85.60
ResUnet++	87.37	85.59	78.46	74.35	72.45	77.11	76.07	70.55	85.21	**94.67**	82.11	82.32	79.27	77.08	81.79	88.85	85.68	79.88	96.45	81.85
TransUNet	86.15	79.69	71.91	74.00	71.02	68.32	65.32	59.27	85.21	**94.67**	16.77	73.52	75.00	78.56	81.33	81.53	80.65	79.88	96.45	74.70

Table 3: Comparison between the twelve models in terms of parameters, FLOPs and memory

Method	Backbone	#Param	FLOPs	Memory
U-Net	-	13M	366.85G	4.30GB
PSPNet	ResNet101	27M	470.54G	2.54GB
DeepLabv3+	ResNet101	59M	261.38G	3.09GB
DenseASPP	DenseNet169	18M	124.4G	4.43GB
Attention U-Net	-	34M	787.19G	7.06GB
DANet	ResNet101	66M	139.01G	2.53GB
EMANet	ResNet101	53M	681.19G	7.67GB
Inf-Net	Res2Net101	50M	192.55G	3.66GB
MiniSeg	-	83K	1.5G	5.41GB
PSANet	ResNet101	78M	482.13G	8.19GB
ResUNet++	-	14M	836.62G	8.88GB
TransUNet	-	110M	514.03G	4.78GB

clear, so we can see if mDice is lagged by typically one or two categories. DeepLabv3+ also achieves the best performance with the highest scores on 9 labels. Scores on several courses such as T10, T9, T10/T11, and T9/T10 are almost the same, and we suppose it is caused by the unevenness of the dataset, as the number of these four classes is 53, 9, 68, and 14 respectively (200 MR images in total). We assume models like Inf-Net get the lowest score on class T9 (0.97%) because of not learning essential features for segmentation. Surprisingly, even if the sample sizes of these four classes (T10, T9, T10/T11, T9/T10) are small, all the twelve models seem to achieve acceptable performance in these classes.

The comparison between these twelve models in terms of parameters, FLOPs, and memory are in Table.3. MiniSeg has a pretty small number of parameters (83K) and FLOPs (1.5G), while its performance just slightly lagged behind DeepLabv3+. By contrast, TransUNet which combines Transformer as part of its encoder has the most significant number of parameters and took the longest time to train, according to our experiment.

To explicitly show the segmentation result, we select five MR images for each of the twelve models in Figure 18. The best image in Figure 18 is a slice in the middle of the original 3D medical image. As the MR image has a clear vision of every label and boundary, the segmentation is perfect and accurate. Regard the worst picture in Figure 18, it has the lowest score (from 25% to 40%) among all other images. Class L2, L1, T12 and T11 are mislabeled in almost every model and full of vague boundaries. The other three images (c, d, e) are selected randomly and implicate the general performance of each model to some extent. All parts of the spine are clearly distinguished from the background, although some classes are mislabeled and even mixed with other courses.

In addition, all twelve models are evaluated on the validation dataset during the training progress for every metric. From Figure 19, the convergence speeds of PSPNet, DeepLabv3+, DANet, EMANet, PSANet and ResUNet++ are faster than other models. They converged at about 40 epochs while other models achieved their own best performance at about 100 to 120 epochs. The false-negative rate is quite low, and many models score below 10%. This is because the number of background pixels is much more than all 19 labels. When TP is much larger than FN, the false-negative rate becomes very low according to equ.4. Similarly, almost every model reached a false-positive rate of 0.2% or below because TN far exceeds FP according to equ.5.

4.6 Analysis and discussion

Segmenting the target organ or different parts of the tissues from the medical images helps clinicians make diagnoses. Magnetic resonance imaging (MRI), along with computed tomography (CT) and positron emission tomography (PET), is the most popular methods used in diagnosing spinal disorders. With the development of artificial intelligence, auto segmentation in the medical image is gradually catching up with and even surpassing manual segmentation, which needs prior knowledge and may be subjective. The limitation of the labeled datasets is a challenge in medical image segmentation. Most of models are evaluated on two-classes dataset [14] [15] [16] [18]. This paper uses a multi-class one, a spine MRI dataset with 20 classes including background.

We first apply the commonly used U-Net and its variants. The performance of U-Net is not good enough with Dice at 79.87% and mIOU at 74.02%. The performance even went worse as more modules were added to U-Net Table.1, However, we found the dataset used is similar to the semantic segmentation with 20 classes and fixed shapes. Thus, we review some state-of-the-art models and apply them to this dataset. These models often use the network of ImageNet such as ResNet to extract features and then add modules to better capture multi-scale and long-rang contextual information.

DeepLabv3+ outperforms all other models in this paper and achieved 87.55%, 81.8% in Dice and mIOU. This indicates that these semantic segmentation models have potential for multi-class medical datasets. For FFigure 18, we speculate there might be a unique pattern in this patient's lumbar and thoracic, leading to the worst

segmentation performance. This may be pathological information while other patients in this dataset do not carry it. In random1 from Figure 18, U-Net labeled L2 as both L2 and L3 and Attention U-Net labeled L4 as L4 and L5. This may be unacceptable in medical segmentation, as the cost of misdiagnosis is very high. We suppose that these adjacent parts have the similar appearances that mislead the model to a wrong prediction.

5 LIMITATIONS AND FUTURE WORK

Firstly, some large models like MedT [33], Unetr [34], Swin-Unet [35] are not included in the review due to our computational resource limitations. We would verify their performance on the spine MRI dataset in future work. Secondly, we slice the 3D MRI images into 2D slices as the input of the models so some contextual information between slices may be lost. Some 3D segmentation models like V-Net [32], nnU-Net [36] would be taken into consideration in future studies for 3D segmentation.

6 CONCLUSION

As one of the main tasks of computer vision, medical image segmentation is of great help to clinical and differential diagnosis in modern medicine. More and more challenging problems have been solved with the help of advanced equipment and methods in recent years. Deep learning is creatively introduced to automatically locate, identify, segment, and improve the previous semi-automated approach. Despite breakthroughs in recent years, medical image segmentation still has considerable challenges, such as unevenness of the sample size and blur boundaries of different classes. Therefore, more improvements are needed to help improve the segmentation performance.

This paper focuses on medical image segmentation from spine MR images. We reviewed several state-of-the-art models and divided them into two categories. One is U-Net Family and the other is Models Using Backbone. This study showed that the performance of Models Using Backbone exceeds those U-Net variants and DeepLabv3+ achieves the best performance both on Dice and mIOU among all twelve models. We suppose that multi-scale and long-range contextual information also count for medical image segmentation.

REFERENCES

[1] A. Kulkarni, R. Shevgaonkar, and S. Sahasrabudhe, "Edge detection using scale space knowledge," in *Proceedings of TENCON'93. IEEE Region 10 International Conference on Computers, Communications and Automation*, vol. 2. IEEE, 1993, pp. 986–990.

[2] J. K. Leader, B. Zheng, R. M. Rogers, F. C. Sciurba, A. Perez, B. E. Chapman, S. Patel, C. R. Fuhrman, and D. Gur, "Automated lung segmentation in x-ray computed tomography: development and evaluation of a heuristic threshold-based scheme1," *Academic radiology*, vol. 10, no. 11, pp. 1224–1236, 2003.

[3] H. Gao, L. Dou, W. Chen, and G. Xie, "The applications of image segmentation techniques in medical ct images," in *Proceedings of the 30th Chinese Control Conference*. IEEE, 2011, pp. 3296–3299.

[4] D. Gaweł, P. Główka, T. Kotwicki, and M. Nowak, "Automatic spine tissue segmentation from mri data based on cascade of boosted classifiers and active appearance model," *BioMed research international*, vol. 2018, 2018.

[5] Y. Tian, T. Guan, C. Wang, L. Li, and W. Liu, "Interactive foreground segmentation method using mean shift and graph cuts," *Sensor Review*, 2009.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] A. Suzani, A. Rasoulian, A. Seitel, S. Fels, R. N. Rohling, and P. Abolmaesumi, "Deep learning for automatic localization, identification, and segmentation of vertebral bodies in volumetric mr images," in *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9415. International Society for Optics and Photonics, 2015, p. 941514.

[8] A. Arpitha and L. Rangarajan, "Computational techniques to segment and classify lumbar compression fractures," *La radiologia medica*, vol. 125, no. 6, pp. 551–560, 2020.

[9] L. Silvoster and M. Retnaswami, "Spine mri segmentation and efficient detection of lumbar intervertebral disc degeneration," *IET Image Processing*, vol. 14, 2020.

[10] A. S. Al-Kafri, S. Sudirman, A. Hussain, D. Al-Jumeily, F. Natalia, H. Meidia, N. Afriliana, W. Al-Rashdan, M. Bashtawi, and M. Al-Jumaily, "Boundary delineation of mri images for lumbar spinal stenosis detection through semantic segmentation using deep neural networks," *IEEE Access*, vol. 7, pp. 43 487–43 501, 2019.

[11] A. Suzani, A. Rasoulian, S. Fels, R. N. Rohling, and P. Abolmaesumi, "Semi-automatic segmentation of vertebral bodies in volumetric mr images using a statistical shape+ pose model," in *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9036. International Society for Optics and Photonics, 2014, p. 90360P.

[12] R. Zhang, X. Xiao, Z. Liu, Y. Li, and S. Li, "Mrln: Multi-task relational learning network for mri vertebral localization, identification, and segmentation," *IEEE journal of biomedical and health informatics*, vol. 24, no. 10, pp. 2902–2911, 2020.

[13] J. Huang, H. Shen, J. Wu, X. Hu, Z. Zhu, X. Lv, Y. Liu, and Y. Wang, "Spine explorer: a deep learning based fully automated program for efficient and reliable quantifications of the vertebrae and discs on sagittal lumbar spine mr images," *The Spine Journal*, vol. 20, no. 4, pp. 590–599, 2020.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[16] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[18] Y. Qiu, Y. Liu, S. Li, and J. Xu, "Miniseg: An extremely minimum network for efficient covid-19 segmentation," *arXiv preprint arXiv:2004.09750*, 2020.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[22] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[23] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[26] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3684–3692.

[27] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[28] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9167–9176.

[29] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[30] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 267–283.

[31] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, p. 106236, 2021.

[32] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[33] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *arXiv preprint arXiv:2102.10662*, 2021.

[34] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," *arXiv preprint arXiv:2103.10504*, 2021.

[35] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[36] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128*, 2019.

A VISUAL COMPARISON

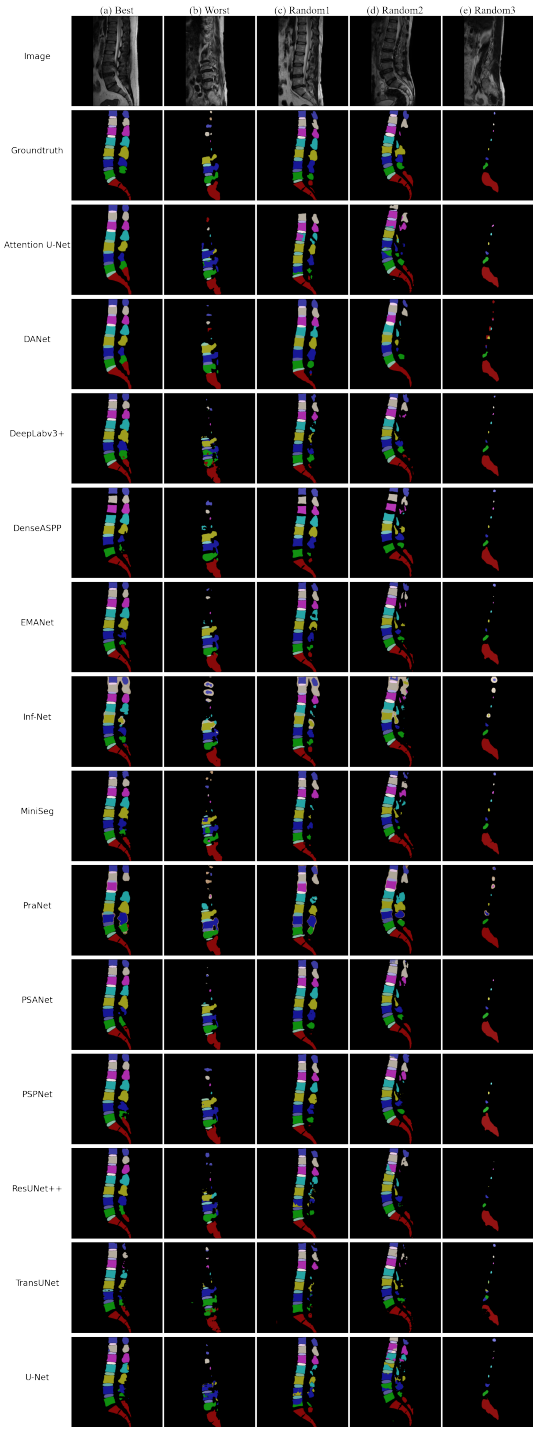

Figure 18: Visual comparison between the twelve models. The first image is the best, the second image is the worst and the other are selected randomly.

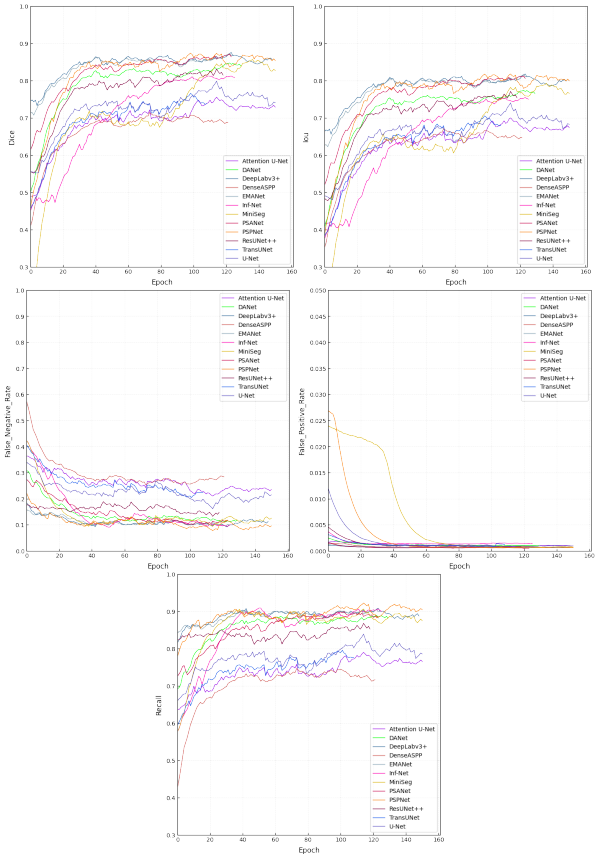

Figure 19: Visual comparison of the evaluation metrics for the twelve models.