2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Centroid-based Distance Loss Function for Lamina Segmentation in 3D Ultrasound Spine Volumes

Jason Wong, Solvin Sigurdson, Marek Reformat, *Member, IEEE*, and Edmond Lou, *Member, IEEE*

Abstract— **Ultrasound imaging of the spine to diagnose the severity of scoliosis is a recent development in the field, offering 3D information that does not require a complicated procedure of reconstruction, unlike with radiography. Determining the severity of scoliosis on ultrasound volumes requires labelling vertebral features called laminae. To increase accuracy and reduce time spent on this task, this paper reported a novel custom centroid-based distance loss function for lamina segmentation in 3D ultrasound volumes, using convolutional neural networks (CNN). A comparison between the custom and two standard loss functions was performed by fitting a CNN with each loss function. The results showed that the custom loss network performed the best in terms of minimization of the distances between the centroids in the ground truth and the centroids in the predicted segmentation. On average, the custom network improved on the total distance between predicted and true centroids by 33 voxels (22%) when compared with the second best performing network, which used the Dice loss. In general, this novel custom loss function allowed the network to detect two more laminae on average in the lumbar region of the spine that the other networks tended to miss.**

I. Introduction

Scoliosis is a three-dimensional spinal condition where the spine is characterized by lateral curvature coupled with vertebral rotation. Adolescent idiopathic scoliosis (AIS) is the most common form of this condition, affecting 1-3% of adolescents aged 10 to 16 years old. There is no known cause for it. Typically, radiography is employed to assess the severity of AIS, requiring a clinician to manually measure many parameters that describe the structural changes [1]. Due to the sheer number of patients that clinicians see per clinic, measuring these parameters is time consuming for clinicians. Automating the measurement process would minimize errors from human judgment and significantly reduce time spent on this task, freeing up clinicians' time to see more patients [2]. Additionally, taking radiographs exposes patients to ionizing radiation, which may increase the risk of cancer [3]. Consequently, the feasibility of ultrasound for scoliosis is currently being investigated and has been found to be comparable to radiographic measurements in terms of accuracy and reliability [4].

To facilitate measurement of scoliosis parameters on ultrasound scans, pairs of vertebral landmarks, called laminae, need to be marked on 17 vertebrae, meaning that 34 laminae must be identified per scan. All three views (coronal, sagittal, and axial) are crucial to identifying laminae positions. On the coronal projection of ultrasound volumes, the spinous process column, indicated by a dark curve in the middle, is used to estimate the locations of the pairs of laminae by identifying isolated bright regions. Each pair of laminae are generally equidistant from the spinous process and form a line when joined that is perpendicular to the dark spinous process column. The axial view is then used to confirm the precise locations by identifying two bright lines, as the lamina regions have a relatively flat surface. These flat surfaces reflect the ultrasound waves the strongest. Finally, the sagittal projection is used to confirm that all the laminae pairs follow a smooth curve as you move down the spine. An ultrasound scan, along with the different views, is illustrated in Fig. 1a. Once the laminae are identified, the center of lamina method can be used to assess the severity of scoliosis [5]. It should be emphasized that only the centroids of the labelled regions are relevant for measurement. This makes the task of labelling laminae quite unique in the medical segmentation field. If the centroids of the labelled laminae are correct, the size and shape of the labelled regions do not actually matter.

The convolutional neural network (CNN) is a type of neural network that is commonly used for image and volume segmentation [6]. A major design aspect of CNNs is the loss function, which is the error function that the algorithm tries to minimize when comparing the predictions of the network with the ground truths. The standard loss functions for segmentation are binary cross entropy and Dice loss [7]. However, custom loss functions tailored for specific applications have become more common to generate as precise a segmentation as possible. This is true particularly for medical segmentation, as these segmentations are often used in patient diagnoses [8, 9]. Due to the success of custom loss functions in medical image segmentation found in the literature and the unique nature of lamina segmentation where the centroids of the connected components take precedence, this paper reported on the development of a custom loss function for lamina segmentation in 3D ultrasound volumes using CNNs and evaluated its performance with respect to CNNs fit using Dice and binary cross entropy loss functions.

II. Data

A total of 70 ultrasound scans on children with AIS were acquired at the local scoliosis clinic. Ethics approval was granted from the local research health ethics board. All subjects signed written consents prior to participating in the study. The ultrasound volumes were processed using a completely automatic procedure. First, the top layer of voxels was cropped to remove the reflections from the skin, muscles, and fat that make the volumes noisy. The volumes were then narrowed to the region of interest by first identifying the dark

*Research supported by the Natural Sciences and Engineering Research Council of Canada, the Women and Children's Health Research Institute, and the Edmonton Orthopaedic Committee.

All authors are with the University of Alberta, Edmonton, Canada (corresponding author e-mail: elou@ualberta.ca)

spinous process column and then cropping around it. The effect of the processing steps is illustrated on the coronal projection in Fig. 1b, 1c, and 1d. The processed volumes were then labelled using a custom-built volume labelling graphical user interface (GUI). These 70 scans were split randomly into 50 training, 10 validation, and 10 test volumes.

The ultrasound volumes and labels were then pre-processed for input into the CNN. They were first scaled to the size 384x96x48, roughly one third the dimensions of the average ultrasound volume. The input volumes were scaled using bilinear interpolation, and the labels were scaled by determining the centroids of the connected components in the unscaled volume, mapping these centroid coordinates to the scaled-down size, and centering a 3x3x3 voxel cube around each centroid. This cube was used instead of just a single point to maximize the chances of the network predicting voxels that were at least close to the lamina and was found to work better in practice. The scaled volumes were then normalized to zero mean and unit variance. To increase diversity of the training set, a data augmentation method of randomly flipping along the sagittal plane was used. This has the effect of switching left and right on the coronal projection.

III. Methods

A. Loss functions

There are two common loss functions used for image segmentation: weighted binary cross entropy (WBCE) and Dice loss. WBCE is a voxel-wise loss function stemming from information theory that aims to minimize the difference between two probability distributions. Dice loss is defined as 1 minus the Dice coefficient [10]. This coefficient D measures the degree of overlap between the ground truth and prediction mask. This loss function is defined as:

$$\mathcal{L}_D = 1 - \frac{2\sum_{i=1}^{N_V} t_i p_i}{\sum_{i=1}^{N_V} t_i + \sum_{i=1}^{N_V} p_i} \tag{1}$$

where N is the number of voxels in the 384x96x48 volumetric space V, and t and p are the values of the voxels, indexed by i, in the ground truth and prediction, respectively.

A novel centroid-based loss function that encourages minimizing the distance between centroids of the connected components in the ground truth and prediction was developed. To take advantage of the available 3D information, these distances were computed on the coronal and sagittal projections. The projections were used instead of the entire volume to reduce training time. First, the raw prediction was turned into a binary array by thresholding the probabilities by 0.5, and the coronal and sagittal projections of both ground truth and prediction were calculated. Let the subscripts c and s denote the coronal and sagittal projections, respectively. The connected components for ground truth and prediction in both projections were then determined, and all centroids were calculated. Let the list of centroids of these connected components be denoted as Γ and Φ and a centroid in these lists as γ and φ for ground truth and prediction, respectively. We define a distance penalty as the mean of the minimum distances from all centroids in Γ to all centroids in Φ. The distance penalty d for each projection is defined as:

$$d_c = \frac{1}{N_{\Gamma_c}} \sum_{i=1}^{N_{\Gamma_c}} \left[\min_{j=1\ldots N_{\Phi_c}} \left\{ \left| \gamma_{c_i} - \varphi_{c_j} \right| \right\} \right] \tag{2}$$

$$d_s = \frac{1}{N_{\Gamma_s}} \sum_{i=1}^{N_{\Gamma_s}} \left[\min_{j=1\ldots N_{\Phi_s}} \left\{ \left| \gamma_{s_i} - \varphi_{s_j} \right| \right\} \right] \tag{3}$$

where N_X is the number of centroids in the list X. With these individual distance penalties defined, the distance loss is formulated as:

$$\mathcal{L}_d = d_c + d_s \tag{4}$$

The distance loss defined above only considers the minimum distances, meaning that all φ that are not closest to any γ do not contribute to the distance loss at all. Consequently, this could lead to the network tending to predict many false positives. To discourage this, a term penalizing the difference between the number of connected components in Γ and Φ was added. How far these false positives are from a γ does not matter for this application; therefore, distance was not involved in this penalty term. This penalty term is defined as:

$$\mathcal{L}_N = \max\left\{1, \left| N_{\Gamma_c} - N_{\Phi_c} \right| + \left| N_{\Gamma_s} - N_{\Phi_s} \right| \right\} \tag{5}$$

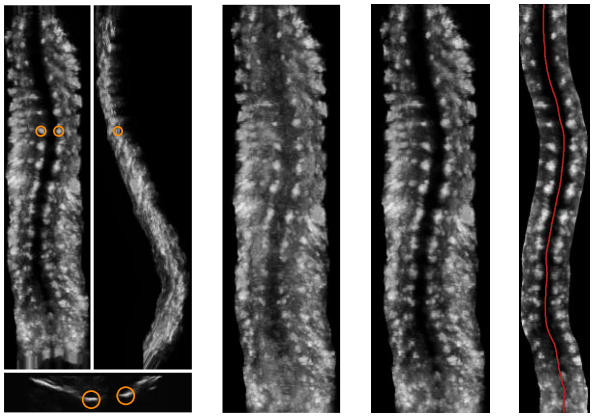
Finally, maximizing the amount of overlap is still desired, and so the last term added to the centroid loss is the Dice loss. Combining these terms, the final loss is defined as:

$$\mathcal{L}_c = \mathcal{L}_D \mathcal{L}_d \mathcal{L}_N \tag{6}$$

B. Network Architecture and Parameters

The CNN architecture used for segmentation in this paper was based on the U-net [11]. This architecture is very common in medical image segmentation tasks, as it has worked well in practice even with very little training data. For this experiment, there were three changes made to vary this architecture. First, one less pooling and upsampling stage was used to reduce the number of training parameters. Second, because the inputs were volumes, all operations (convolution, pooling, and upsampling) were replaced with their 3D equivalents. Finally, same padding was used, so the volumes remained the same size after each convolution. This was implemented because in some cases, laminae could exist on the edges of the volumes.

To improve the generalizability of the model, dropout and batch normalization were employed. Dropout was performed with a probability of 0.125 after each pooling and upsampling layer, and batch normalization was performed after each



(a) (b) (c) (d)

Fig. 1. (a) Different views of an ultrasound scan of an AIS patient's spine with a pair of laminae marked as orange circles on the coronal (left), sagittal (right), and axial B-mode (bottom) views, (b) original un-processed coronal projection, (c) top layer cropped coronal projection, (d) narrowed region of interest with dark spinous process column (red) of coronal projection.

convolutional layer. The leaky rectified linear unit with an alpha value of 0.01 and sigmoidal activation functions were used for the hidden layers and output layer, respectively. The Adam optimizer was employed with a learning rate of 10^{-4}. Due to the sheer size of this network, a batch size of 1 was used. The 3D U-net variant architecture is illustrated in Fig. 2.

C. Experiment

To evaluate whether the custom loss function performed better than the traditional loss function, three 3D U-net variants were fit – Dice loss (D-U-net), WBCE loss (B-U-net), and custom centroid-based distance loss (C-U-net). All networks were fit for 200 epochs, and model checkpoints were used to save the optimal model found during training with respect to the loss on the validation set. The minimum required number of epochs to train was determined by training each type of network for 1500 epochs and analyzing the validation loss curves during training. For all cases, the validation loss was lowest before 200 epochs. The results reported in this paper correspond to the optimal models taken at the lowest validation loss.

There are three metrics used to compare performance between the two networks. The first is the Dice coefficient D. The second metric d_{3D} is similar to the distance penalty d_c and d_s, but computed using the whole 3D volume instead:

$$d_{3D} = \frac{1}{N_{\Gamma_{3D}}} \sum_{i=1}^{N_{\Gamma_{3D}}} \left[\min_{j=1\ldots N_{\Phi_{3D}}} \left\{ \left| \gamma_{3D_i} - \varphi_{3D_j} \right| \right\} \right] \quad (7)$$

The third metric is the difference between the number of connected regions in the predicted segmentation and the number of connected regions in the ground truth Δ. Again, this is computed for the 3D volume instead of just the projections:

$$\Delta = \left| N_{\Gamma_{3D}} - N_{\Phi_{3D}} \right| \quad (8)$$

D. Implementation

All code for this experiment was programmed in Python with the networks being implemented using the TensorFlow library. The networks were trained on the supercomputing hub at the University of Alberta with an NVIDIA Tesla V100 16GB GPU and an Intel Xeon Gold 6138 dual processor.

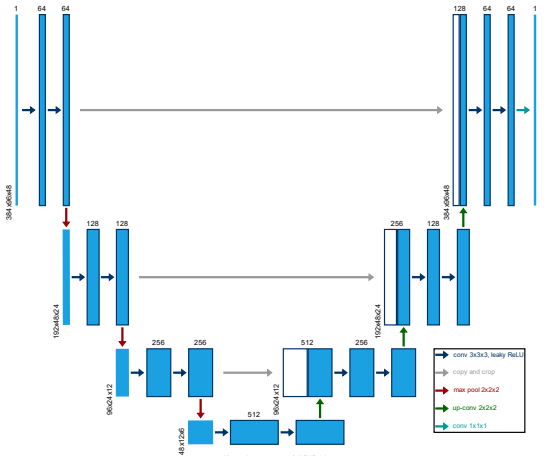


Fig. 2. U-net architecture diagram with blue boxes representing feature maps and white boxes representing copied feature maps. The number of feature maps is above the box and the size of the feature maps is to the left of each convolutional block.

IV. Results

The training times for the D-U-net, B-U-net, and C-U-net over 200 epochs were 253, 254, and 259 minutes, respectively. The D-U-net converged in 80 epochs, the B-U-net in 176, and the C-U-net in 95 with respect to the validation set loss.

Regarding the test set, the results for the networks' performance are reported in Table I. The C-U-net's centroid placements were on average a total of 33 voxels closer to the ground truths' centroids when compared with D-U-net and 129 voxels closer when compared with B-U-net. Box and whisker plots of the test set results, with the whiskers representing 1.5x the interquartile range, are shown in Fig. 3a.

One-tailed paired Student's t-tests were performed on the differences between the performance metrics in the test set to determine statistical significance. These tests were conducted between the C-U-net and D-U-net and the C-U-net and B-U-net. The C-U-net performance improvement was statistically significant ($p < 0.05$) for all three metrics when compared with B-U-net, while in the case of D-U-net, this occurred only for the d_{3D} metric.

Finally, the segmentations for one of the test volumes are illustrated in Fig. 3b and 3c. The segmentations for the B-U-net are not shown here because it performed significantly worse than the other two networks.

V. Discussion

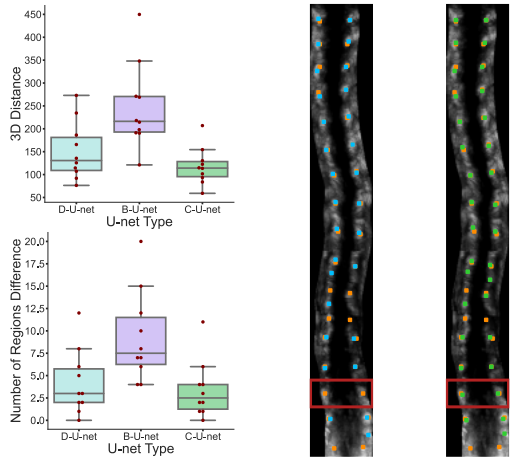
A. Analysis of Results

The B-U-net was outperformed by the other two U-nets overall. Its average metric values were all worse than the other two networks, and it underperformed in terms of d_{3D} for every individual test volume as well. This was expected since it has been found that metric-sensitive losses generally perform better than voxel-wise losses in medical image segmentation [12]. Between the D-U-net and C-U-net, the difference between D and Δ distributions is not statistically significant. However, the improvement in the d_{3D} performance was found to be statistically significant. This demonstrates that d_{3D} minimization was made without significantly impacting D, highlighting that D is not a good sole metric for evaluating performance in this application. The C-U-net improved upon the d_{3D} for eight out of the ten cases. The two cases where D-U-net outperformed C-U-net only resulted in an improvement of less than ten voxels. Two of the cases where the C-U-net outperformed the D-U-net saw significant improvement, with one case resulting in an improvement of over 100 voxels.

The U-net segmentations are typically more accurate in the thoracic region of the spine. This is because the thoracic region of the spine is closer to the skin, while the lumbar region has thicker muscle, which attenuates the ultrasound signals more.

TABLE I. Performance Metric Values for the Test Set

	D		d_{3D}		Δ	
	Mean	St. D	Mean	St. D	Mean	St. D
D-U-net	0.407	0.050	151.1	63.7	4.2	3.6
B-U-net	0.280	0.100	247.2	93.6	9.3	5.1
C-U-net	**0.412**	**0.050**	**118.1**	**40.8**	**3.4**	**3.2**



(a) (b) (c)

Fig. 3. (a) Box and whisker plots of d_{3D} (top) and Δ (bottom) for the test set under different U-nets, (b) D-U-net segmentation (blue) of a test volume overlaid with ground truth labels (orange), (c) C-U-net segmentation (green) of a test volume overlaid with ground truth labels (orange).

The C-U-net tends to handle the noisy signals better than the D-U-net. For example, in Fig. 3c, the C-U-net segments a full pair of laminae perfectly in the lumbar region that the D-U-net completely misses (red box in Fig. 3b and 3c), and this trend of performing better in the lumbar region was present in the other test volumes as well.

When evaluating the absolute quality of the lamina segmentations, we value avoiding false negatives the most, even at the expense of obtaining more false positives. For this application, these errors are defined as illustrated in Fig. 4. The reason for prioritizing false negatives is because it is much easier to eliminate false positives from contention rather than trying to identify false negatives with no thresholded voxels associated with it. Based on an inspection of the test volume segmentations, the D-U-net missed five laminae per scan on average, whereas the C-U-net missed only three. Most of these false negatives were in the lumbar region of the spine. Although missing laminae is not ideal, an experienced operator faces the same challenges in labelling the lumbar laminae. This means that an algorithm that eliminates false positives and identifies false negatives through post-processing can be designed if planned network improvements do not appreciably affect the segmentation quality.

B. Limitations

Hyperparameter optimization is one design procedure that was not fully explored in this experiment. The hyperparameter combination used was merely the best one out of various common combinations that were attempted. A hyperparameter search algorithm or grid search should be conducted to optimize the network; however, this comes at significant computational costs, as 3D CNNs are being trained. Another limitation of this study is the small number of volumes used in training and evaluation. However, as a pilot study, this study demonstrated that the proposed method has the potential to automatically segment the entire spine. Labelling more volumes or implementing more data augmentation methods will be the next step to fully validate the proposed method.

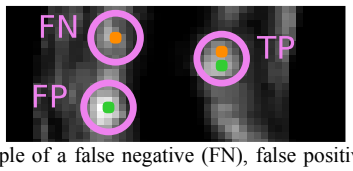


Fig. 4. An example of a false negative (FN), false positive (FP), and true positive (TP) with ground truth in orange and prediction in green.

VI. Conclusion

This study introduced a novel centroid-based distance loss function designed for lamina segmentation in 3D ultrasound scans of AIS spines. The U-net CNN architecture, commonly used in medical segmentation tasks, was used to verify the validity of the loss function, with a 3D U-net variant architecture being fit for three different loss functions: Dice (D-U-net), WBCE (B-U-net), and centroid-based distance loss (C-U-net). The C-U-net yielded the lowest distance between centroids from the ground truth and predicted segmentation in the test set, and this improvement was found to be statistically significant. The improvement of using the C-U-net was seen primarily in the lumbar region of the spine, where fewer false negatives were predicted. Overall, the final segmentations were sufficiently accurate to move towards post-processing including elimination of false positives and identification of false negatives.

References

[1] S. L. Weinstein *et al.*, "Adolescent idiopathic scoliosis," *The Lancet*, vol. 371, no. 9623, pp. 1527–1537, May 2008, doi: 10.1016/S0140-6736(08)60658-3.

[2] H. Sun *et al.*, "Direct Estimation of Spinal Cobb Angles by Structured Multi-output Regression," in *Information Processing in Medical Imaging*, 2017, pp. 529–540.

[3] M. Morin Doody *et al.*, "Breast Cancer Mortality After Diagnostic Radiography: Findings From the U.S. Scoliosis Cohort Study," *Spine*, vol. 25, no. 16, p. 2052, Aug. 2000.

[4] R. Zheng *et al.*, "Factors influencing spinal curvature measurements on ultrasound images for children with adolescent idiopathic scoliosis (AIS)," *PLoS One*, vol. 13, no. 6, p. e0198792, Jun. 2018, doi: 10.1371/journal.pone.0198792.

[5] W. Chen, L. H. Le, and E. H. M. Lou, "Ultrasound Imaging of Spinal Vertebrae to Study Scoliosis," *Open Journal of Acoustics*, vol. 2, no. 3, Art. no. 3, Sep. 2012, doi: 10.4236/oja.2012.23011.

[6] Y. LeCun, Fu Jie Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, Jun. 2004, vol. 2, pp. II-104 Vol.2, doi: 10.1109/CVPR.2004.1315150.

[7] C. H. Sudre *et al.*, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," *arXiv:1707.03237 [cs]*, vol. 10553, pp. 240–248, 2017, doi: 10.1007/978-3-319-67558-9_28.

[8] Q. Huang *et al.*, "Robust liver vessel extraction using 3D U-Net with variant dice loss function," *Computers in Biology and Medicine*, vol. 101, pp. 153–162, Oct. 2018, doi: 10.1016/j.compbiomed.2018.08.018.

[9] W. Zhu *et al.*, "AnatomyNet: Deep Learning for Fast and Fully Automated Whole-volume Segmentation of Head and Neck Anatomy," *Med. Phys.*, vol. 46, no. 2, pp. 576–589, Feb. 2019, doi: 10.1002/mp.13300.

[10] S. Jadon, "A survey of loss functions for semantic segmentation," *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–7, Oct. 2020, doi: 10.1109/CIBCB48159.2020.9277638.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[12] T. Eelbode *et al.*, "Optimization for Medical Image Segmentation: Theory and Practice when evaluating with Dice Score or Jaccard Index," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2020, doi: 10.1109/TMI.2020.3002417.