Fully automatic cervical vertebrae segmentation framework for X-ray images

S. M. Masudur Rahman Al Arif[*, a]

s.al-arif@city.ac.uk

Karen Knapp[b]

Greg Slabaugh[a]

[a]Department of Computer Science, City, (Please keep this comma, as my University has this in its official name. See the bottom of this page for your reference https://www.city.ac.uk/)University of London, London, UK

[b]University of Exeter Medical School, Exeter, UK

*Corresponding author.~~Tel.: +447960464707.~~

Abstract

The cervical spine is a highly flexible anatomy and therefore vulnerable to injuries. Unfortunately, a large number of injuries in lateral cervical X-ray images remain undiagnosed due to human errors. Computer-aided injury detection has the potential to reduce the risk of misdiagnosis. Towards building an automatic injury detection system, in this paper, we propose a deep learning-based fully automatic framework for segmentation of cervical vertebrae in X-ray images. The framework first localizes the spinal region in the image using a deep fully convolutional neural network. Then vertebrae centers are localized using a novel deep probabilistic spatial regression network. Finally, a novel shape-aware deep segmentation network is used to segment the vertebrae in the image. The framework can take an X-ray image and produce a vertebrae segmentation result without any manual intervention. Each block of the fully automatic framework has been trained on a set of 124 X-ray images and tested on another 172 images, all collected from real-life hospital emergency rooms. A Dice similarity coefficient of 0.84 and a shape error of 1.69 mm have been achieved.

Keywords: Segmentation; ~~Deep Learning~~Deep learning; FCN; UNet; Localization; Cervical vertebrae; X-ray

1 Introduction

The cervical spine consists of seven vertebrae, label~~l~~ed C1~~to~~–C7. These vertebrae support the head and protect the spinal column in the neck region. The cervical spine is a highly flexible anatomy, capable of flexion, extension, lateral flexion, and rotation [1]. Due to this wide range of motion, the cervical spine is particularly vulnerable to injury. According to [2], 43.9~~-~~–61.5% of the spinal cord injuries occur in the cervical region. Despite being a highly injurious anatomy, unfortunately, about 20% of the injuries in radiological exams remain unnoticed. And a significant proportion, 67%, of the of the patients with unnoticed cervical injuries suffer tragic extensions of their injuries later in life [3,4]. Recent developments in the fields of computer vision and artificial intelligence have the potential to reduce the number of missing injuries.

Towards building a fully automatic cervical spine injury detection system, in this paper, we propose an automatic segmentation framework for cervical vertebrae in X-ray images. Segmenting the vertebrae correctly is a crucial part for further analysis in an injury detection system. Previous work in vertebrae segmentation has largely been dominated by statistical shape model (SSM)-based approaches [5–12]. These methods record statistical information about the shape and/or the appearance of the vertebrae based on a training set. Then the mean shape is initialized either manually or semi-automatically near the actual vertebra and a search procedure is performed to converge the shape on the actual vertebral boundary. Recent literature utilizes random forest-based machine learning models in order to achieve the shape convergence [9–12].

However, to the best of our knowledge, a fully automatic method is absent from the literature. To fill this gap, in this work, we propose a fully automatic framework for vertebrae segmentation. Starting with a real-life emergency room X-ray image, the framework first locates the spine, then localizes the vertebral centers and finally, achieves segmentation. In other words, the fully automatic framework can be divided into three subtasks: global localization, center localization and vertebrae segmentation. Different specialized fully convolutional neural networks (FCN) are used to solve each of these tasks. The complete framework is shown in Fig. 1.

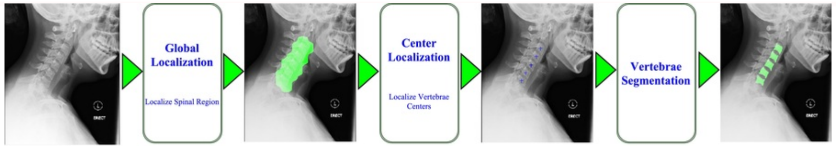

Fig. 1 Fully automatic cervical vertebrae segmentation framework.

alt-text: Fig. 1

Previous work in spine localization includes generalized Hough transform-based approaches [6,13] and more recent random forest-based approaches [14–16]. The state-of-the-art work on cervical ~~vertebrae~~ spine localization uses a sliding window technique to extract patches from the images [16]. A random forest classifier then decides which patches belong to the spinal area. Finally, a rectangular bounding box is generated to localize the spinal region. In contrast to these approaches, we approach the localization problem as a segmentation problem in a lower resolution. Given a set of high-resolution images and manually segmented vertebra~~e~~ ground truth, at a lower resolution, the ground truth becomes a single connected region. Then an FCN can be trained to predict this region. The proposed framework can produce localization map of arbitrary shape in a one-shot process and provide~~s~~ a localization result that models the cervical spine much better than a rectangular box like [16].

Once the spinal region has been localized, the next task is to determine the vertebra~~e~~l centers. Previous work in vertebral~~e~~ landmark localization involves patch-based regression techniques [10,17–19]. Based on the image patches, these methods use different machine learning m~~ethod~~odels to predict vectors pointing towards vertebral~~e~~ landmarks. Random regression forest [10], Hough forest [17,18] and deep fully connected neural network [19] have been used to learn the model. Contrary to these methods, we propose a novel FCN-based probabilistic spatial regressor to localize vertebral~~e~~ centers. Given an image patch, our novel network predicts a two-dimensional probability distribution for the localized centers over the patch space. A novel loss function has been introduced to adapt the FCN as a spatial probability predictor.

Finally, a novel shape-aware deep segmentation FCN is proposed for the vertebrae segmentation phase. Shape is an important characteristic of the vertebra. Previous work in vertebrae segmentation has largely been dominated by statistical shape model (SSM)-based approaches [5–12]. On the other hand, deep segmentation networks have been outperforming the state-of-the-art in different medical image modalities [20–22]. However, combining shape information in a deep segmentation network is not straightforward. In this paper, we provide a solution to this problem by introducing a novel shape-aware term in a segmentation loss function.

Achievements The proposed global localization algorithm has been able to outperform the previous state-of-the-art [16] by 17.1% in terms of sensitivity. The novel center localization framework has produced an average error of only 1.81 mm which i near human-level. A patch-level Dice similarity coefficient of 0.94 has been achieved by the proposed shape-aware segmentation framework. Finally, the fully automatic framework has been able to achieve a Dice similarity coefficient of 0.84 and a shap error of 1.69 mm. All these metrics are computed over a challenging dataset of 172 emergency room X-ray images.

Contributions We make several contributions in this work. First, we propose a deep segmentation network-based spine localization algorithm which outperforms the previous state-of-the-art by a large margin. Second, we propose a novel spatia probability prediction network which achieves human-level performance in localizing vertebral~~e~~ centers. Third, we introduce a shape-aware segmentation loss function which augments the capability of a deep segmentation network with shape informatio and achieves better performance than simple FCN and other traditional shape model-based approaches. The final and the most important contribution is the fully automatic framework which combines the global localization, center localization and vertebra segmentation in a single thread and provides a segmentation result for a real-life emergency room X-ray images without any manual input.

2 Data

A total of 296 lateral cervical spine X-ray images were collected from Royal Devon and Exeter Hospital in association with the University of Exeter. The age of the patients varied from 17 to 96. Different radiographic systems (Philips, Agfa, Kodak, GE) were used to produce the scans. Image resolution varied from 0.1 to 0.194 mm per pixel. Image size varied from 1000 to 5000 pixel~~s~~ with different zoom, crop, spine position and patient position. The images include examples of vertebrae with fractures, degenerative changes and bone implants. The data is anonymized and standard research protocols have been followed. The size, shape, orientation of the spine, image intensity, contrast, noise level all varied greatly in the dataset. For this work, 5 vertebrae C3–C7 are considered. C1 and C2 have an ambiguous appearance due to their overlap in lateral cervical radiographs, and our clinical experts were not able to provide ground truth segmentations for these vertebral bodies. For this reason, they are excluded in this study, similar to other cervical spine image analysis research [5,11,16,23]. Each vertebra from the images was manually annotated for the vertebral body boundaries and centers by expert radiographers. A few examples with the corresponding manual annotations are shown in Fig. 2.

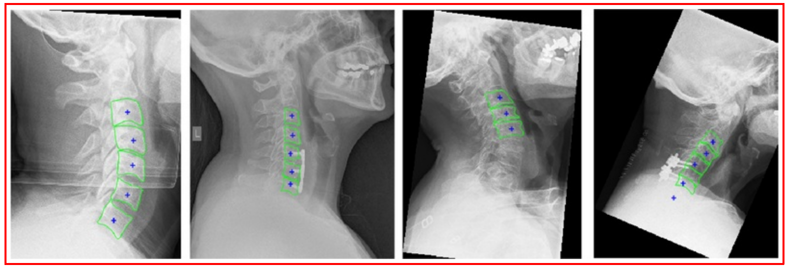

Fig. 2 X-Ray images and corresponding manual annotations~~,~~ of vertebral ~~C~~centers~~, blue plus~~ (+) ~~V~~vertebra~~e~~l (vertebral)boundary curves~~ (green)~~. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 2

The images were received in two sets. The first set contained 138 images. A random 90% or 124 images from this set is used as training dataset in this work. The remaining 10% or 14 images from this set was used for testing the algorithms. The second set of 158 images were received later into the study and added to the test dataset bringing the total number of test images to 172.

3 ~~Global Localization~~Global localization

The first subtask for our fully automatic framework is to locate the spinal region in an arbitrary X-ray image. We approached this problem as a segmentation problem at a lower resolution. In the lower resolution, the cervical vertebrae become a single connected spinal region. A deep fully convolutional network (FCN) is trained to predict this region.

3.1 Data

Based on the manual annotation of the vertebra~~e~~l boundaries, a binary ground truth can be created for each image in our dataset. To create the training and test dataset for the global localization algorithm, these images are converted into square images by padding an appropriate number of zeros in the smaller dimension and the square images are resized to a lower resolution using bicubic interpolation. This resolution can vary based on the available memory and size of the training networks. For our case, we chose this resolution to be 100×100 pixel. The binary vertebra~~e~~ ground truth images forms a single connected region in this resolution. However, our network predicts a segmentation mask of even smaller resolution, 25×25 pixel. The 100×100 pixel localization ground truth~~s~~ ~~are~~is converted to a 25×25 pixel mask using a max-pooling operation with a mask size of 4×4 and stride 4. Max-pooling was used over interpolation~~-~~-based methods to keep the localization mask sharp. Fig. 3 shows some of the localization ground truth overlayed on the image after transforming back to the original resolution.

3.2 Network

A fully convolutional network (FCN) is designed for the global localization task which takes an input image of resolution 100×100 and predicts a localization mask of the resolution 25×25. Our network has six convolutional layers and two max-pooling layers. Batch normalization and rectified linear unit (ReLU) layers are used after each convolution layers. The network diagram is shown in Fig. 4. The total number of parameters in the network is 1,152,450.

3.3 Training

In order to train any network with a large number parameters, 124 images are not enough. ~~In order t~~To increase the number of training data, we have augmented the images by rotating each image from 5° to 355° with a step of 5°. This results in a training set of ~~8,928~~8928 images. It also made the framework rotation invariant. Our choice for data augmentation was only limited to rigid transformations since non-rigid transformation will affect the natural appearance of the spine in the image.

Given a dataset of training image (x)-segmentation label (y) pairs, training a deep segmentation network means finding a set of parameters $\hat{W}$ that minimizes a loss function, L_t. The simplest form of the loss function for segmentation problem is the pixel-wise log loss or the cross-entropy loss.

$$\hat{W} = \arg\min_{W} \sum_{n=1}^{N} L_t\left(\left\{x^{(n)}, y^{(n)}\right\}; W\right) \quad \hat{W} = \arg\min_{W} \sum_{n=1}^{N} L_t\left(\left\{x^{(n)}, y^{(n)}\right\}; W\right), \tag{1}$$

where N is the number of training examples and $\{x^{(n)}, y^{(n)}\}$ represents n-th example in the training set with corresponding manual segmentation. The pixel-wise segmentation loss per image can be defined as:

$$\boxed{L_t(\{x,y\};\boldsymbol{W}) = -\sum_{i\epsilon\Omega_p}\sum_{j=1}^{L} y_i^j \log P\left(y_i^j = 1 \middle| x_i;\boldsymbol{W}\right)} \boxed{L_t(\{x,y\};\boldsymbol{W}) = -\sum_{i\epsilon\Omega_p}\sum_{j=1}^{2} y_i^j \log P\left(y_i^j = 1 \middle| x_i;\boldsymbol{W}\right),} \quad (2)$$

$$\boxed{P\left(y_i^j = 1 \middle| x_i;\boldsymbol{W}\right) = \frac{\exp\left(a_j\left(x_i\right)\right)}{\sum_{k=1}^{L}\exp\left(a_k\left(x_i\right)\right)}} \boxed{P\left(y_i^j = 1 \middle| x_i;\boldsymbol{W}\right) = \frac{\exp\left(a_j\left(x_i\right)\right)}{\sum_{k=1}^{L}\exp\left(a_k\left(x_i\right)\right)},} \quad (3)$$

where $a_j(x_i)$ is the output of the penultimate activation layer of the network for the pixel x_i, Ω_p represents the pixel space and P are the corresponding class probabilities.

The network is trained on a system with a NVIDIA Quadro M4000 GPU for 30 epochs with a batch-size of 10 images. The training took approximately 18 h~~ours~~. The weight optimization is performed by the RMSprop version of the stochastic gradient descent algorithm throughout this work [24].

3.4 ~~Inference and Metrics~~Inference and metrics

~~At test time~~During inference, a test image is padded with zeros to form a square, resized to 100 × 100 pixel~~s~~ and fed forward through the network to produce the localization map. The average time for the network to produce a localization map is less than 0.1 s~~ee~~. This map is compared with the corresponding localization ground truth. Pixel-level accuracy, Dice similarity coefficient (DSC), sensitivity and specificity are computed. These metrics demonstrate the performance of the trained networks at the lower resolution at which the network generates the prediction. From a practical point of view, the performance of the localization should also be computed at the original resolution with the manually segmented vertebra~~e~~ ground truth. In order to achieve this, the predicted localization map is transformed (resized and unpadded) back to the original image resolution, and sensitivity and specificity are computed by comparing them with the manually segmented vertebra~~e~~ ground truth.

3.5 Results

The median, mean and standard deviation of the metrics over 172 test images are reported in Table 1. At the lower resolution, we have been able to achieve an average pixel-level accuracy of 99%. In the original resolution, the algorithm has been able to produce an average sensitivity score of 0.96 when compared with the vertebra~~e~~ ground truth, which indicates 96% of the vertebra~~e~~l area has been covered by our predicted localization maps.

Table 1 (Please make the texts center-aligned. Also, please show the cell borderlines. Without the borders and with left-aligned text, the merged cells look really confusing in the pdf document.)Performance of global localization.

alt-text: Table 1

Resolution	25 × 25				Original	
	Pixel ~~A~~accuracy	DSC	Sensitivity	Specificity	Sensitivity	Specificity
Median	0.99	0.91	0.89	1.00	1.00	0.96
Mean	0.99	0.89	0.86	1.00	0.96	0.96
Std	0.01	0.10	0.13	0.00	0.11	0.01

The box-plot of these metrics are shown in Fig. 5. It can be seen that only a few outliers perform poorly. Most of these images have ~~clinical~~surgical implants and/or severe clinical conditions in the spinal region. A few of these hard cases are shown in Fig. 6. Fig. 6b~~,~~ and c shows (show) examples of images with clinical conditions where the localization algorithm performed well. Two of the outlier results are shown in Fig. 6e~~,~~ and f. Compared with the previous state-of-the-art in cervical vertebra localization, which uses a random forest-based algorithm and provides a rectangular bounding box [16], our algorithm produces a 17.1% improvement in average sensitivity with a clear qualitative improvement on the same training and test images. In terms of the time required for the algorithm to produce a result, our algorithm is more than 70 times faster than [16]. Our algorithm is capable of producing a localization result for any image under a second while the sliding window-based method of [16] requires 70~~to~~–180 s~~econds~~ depending on the image size.

4 ~~Center Localization~~Center localization

The next task for our fully automatic framework is to localize the vertebra~~e~~ centers in the already localized spinal region. Instead of the common practice of regressing vectors pointing towards the location of the center, we design our center localization framework to produce a probability map. We will use a novel fully convolutional network (FCN) to learn the model~~l~~ing. Given an image patch, the network learns to predict a probability distribution over the image space indicating where the centers are most probable. In contrary to the vector regression techniques, our method can predict multiple centers for a single patch.

4.1 Data

Our data comes with a large number of vertebrae with clinical conditions. Thus~~,~~ the geometrical center of the manually annotated shape is not robust for each vertebra and varies based on the extent of the vertebra~~e~~ conditions. So, our medical partners have provided us with manually clicked center points. Each vertebra has one manually clicked center. However, because the vertebral center is not attached to any visible landmark, human perception of the center also varies to some extent. This motivated us to convert the manually clicked centers into probabilistic distributions.

The probability distribution at a vertebral center (x_c, y_c) can be defined as a 2D anisotropic Gaussian distribution [25].

$$F(x,y) = \frac{1}{2\pi\sqrt{v_w v_h}} e^{-\frac{1}{2v_x v_y}\left(a_1(x-x_c)^2 - 2a_2(x-x_c)(y-y_c) + a_3(y-y_c)^2\right)} \quad F(x,y) = \frac{1}{2\pi\sqrt{v_w v_h}} e^{-\frac{1}{2v_w v_h}\left(a_1(x-x_c)^2 - 2a_2(x-x_c)(y-y_c) + a_3(y-y_c)^2\right)}, \tag{4}$$

(Eqn. (4).. the number (4) crosses the column margin the pdf document. Please ensure this stays inside column margin.)

where

$$a_1 = v_w \cos^2\theta + v_h \sin^2\theta \quad a_1 = v_w cos^2\theta + v_h sin^2\theta, \tag{5}$$

$$a_2 = (v_w - v_h)\cos\theta\sin\theta \quad a_2 = (v_w - v_h)\cos\theta\sin\theta, \tag{6}$$

$$a_3 = v_w \sin^2\theta + v_h \cos^2\theta \quad a_3 = v_w \sin^2\theta + v_h \cos^2\theta, \tag{7}$$

and

$$\theta = \frac{\theta_l + \theta_b + \theta_r + \theta_t}{4} \quad \theta = \frac{\theta_l + \theta_b + \theta_r + \theta_t}{4}, \tag{8}$$

$$v_w = \frac{\frac{w_t + w_b}{2} R}{k} \quad v_w = \frac{\frac{w_t + w_b}{2} R}{k}, \tag{9}$$

$$v_h = \frac{\frac{h_l + h_r}{2} R}{k} \quad v_h = \frac{\frac{h_l + h_r}{2} R}{k}, \tag{10}$$

where R is pixel spacing (in millimeter per pixel) of the image, $k = 60$ is a~~n empirical~~ constant chosen based on visual evaluation of the ground truth and θ_l, θ_b, θ_r, θ_t, w_t, w_b, h_l, h_r are computed from the manually annotated vertebra~~e~~ ~~c~~ground truth~~orners~~ and demonstrated in Fig. 7a.

The process is repeated for all the vertebral~~e~~ centers and a single probabilistic distribution defined over the image space is generated. A few images with overlayed probabilistic center distributions are shown in Fig. 8a.

To generate ~~a~~the training image patches and corresponding probability distributions, a grid of 9 uniformly spaced points were generated per vertebra and 3 points were generated in between two consecutive vertebrae. An example of these grid points is shown in Fig. 7b. From each of these grid points, patches were extracted with two scales (original vertebra~~e~~ size + 2 mm and 4 mm) and five orientations ($-20°$ to $20°$ with a step of $5°$ where $0°$ is the mean vertebral axis). All these extracted patches are then resized to 64×64 pixel~~s~~, the resolution at which the network will be trained. A total of ~~66,600~~66,600 patches were generated from our 124 training images. Fig. 8b shows how these distributions look at the patch-level.

4.2 Network

Here, our intention is to predict a two-dimensional probabilistic distribution for an input patch of 64×64 pixel~~s~~. We want our predicted distribution to have the same spatial resolution as the input patch. The FCN architecture used for the global localization framework predicts an output with a lower spatial resolution than the input. Thus~~,~~ it cannot be used here. DeConvNet [26] and UNet [20] are two fully convolutional neural networks that have been used for segmentation problems where the spatial resolution of the input image and output predictions are similar. Among the two networks, our initial experiments showed better performance with UNet architecture. Here, for the probabilistic spatial regressor-based center localization framework, we used a modified version of the UNet [20] architecture. UNet has a downsampling path and an upsampling path. Our downsampling path has nine convolutional layers. Each convolutional layer is followed by a batch normalization and rectified linear unit (ReLU). Three max-pooling layers in between the convolutional layers downsample the spatial dimension from 64×64 to 8×8. The upsampling path forms a mirrored version of the downsampling path. Upsampling is done by deconvolutional layers. The network shares information between the downsampling and upsampling path using concatenation. The network diagram is shown in Fig. 9. The number of filters in each layer can be tracked from the number of channels in the data blocks. The total number of parameters for the center localization UNet is 24,238,210.

4.3 Training

The softmax layer at the end of the network creates a probabilistic two-channel output, just like a binary segmentation problem. However, the ground truth here is a probabilistic map, not a binary segmentation map. Thus the standard segmentation log loss of Section 3.3 cannot be used. We formulate a novel loss function for training the network to predict a probabilistic map.

Loss function for probabilistic spatial regression To match the two-channel output of the final softmax layer, the ground truth probability (GT_p) is also converted to a softmax-like two channel distribution, P_{GT}.

$$\boxed{P_{GT_{i,channel=1}} = \frac{GT_{p_i} - min\left(GT_p\right)}{max\left(GT_p\right) - min\left(GT_p\right)}} \boxed{P_{GT_{i,channel=1}} = \frac{GT_{p_i} - min\left(GT_p\right)}{max\left(GT_p\right) - min\left(GT_p\right)},} \quad \textbf{(11}$$

$$\boxed{P_{GT_{i,channel=2}} = 1 - P_{GT_{i,channel=1}}} \boxed{P_{GT_{i,channel=2}} = 1 - P_{GT_{i,channel=1}},} \quad \textbf{(12}$$

where $i\epsilon\Omega_p$ is the pixel space. Notice that, $P_{GT_{channel\,=\,1}}$ is no longer a normalized probability distribution (i.e. doesn't integrate to unity), rather a stretched distribution where the maximum is unity and minimum is zero. This ensures that the softmax layer i able to produce a similar distribution, as it squashes the input activations to the range from 0 to 1.

Training our UNet would then mean finding an optimized set of parameters $\widehat{W}_o$ which minimizes a loss, L, between the predicted $\hat{y}^{(n)}$ and updated ground truth $P_{GT}^{(n)}$ over the training dataset.

$$\boxed{\widehat{W}_o = \arg\min_{W} \sum_{n=1}^{N} L\left(\left\{x^{(n)}, P_{GT}^{(n)}\right\}; W\right)} \boxed{\widehat{W}_o = \arg\min_{W} \sum_{n=1}^{N} L\left(\left\{x^{(n)}, P_{GT}^{(n)}\right\}; W\right),} \quad \textbf{(13)}$$

where N is the number of training examples and $\left\{x^{(n)}, P_{GT}^{(n)}\right\}$ represents n-th example in the training set with corresponding ground truth probability of the regression target. Since the target probabilities are spatially distributed over the pixel space, we can define a pixel-wise loss function per training sample as:

$$\boxed{L\left(\left\{x, P_{GT}\right\}; W\right) = \frac{1}{2\left|\Omega_p\right|} \sum_{i\epsilon\Omega_p} \sum_{j=1}^{2} w_i \left(\hat{y}_i^j - P_{GT_{i,channel=j}}\right)^2} \boxed{L\left(\left\{x, P_{GT}\right\}; W\right) = \frac{1}{2\left|\Omega_p\right|} \sum_{i\epsilon\Omega_p} \sum_{j=1}^{2} w_i \left(\hat{y}_i^j - P_{GT_{i,channel=j}}\right)^2,}$$

(14) (Eqn. (14).. the number (14) crosses the column margin and remains invisible in the pdf document. Please ensure this stays inside column margin.)

where

$$\boxed{w_i = \begin{cases} \frac{\left|\Omega_{p_\phi}\right|}{\left|\Omega_{p_o}\right|} & \text{if } i\epsilon\Omega_{p_\phi} \\ 1 & \text{otherwise} \end{cases}} \boxed{w_i = \begin{cases} \frac{\left|\Omega_{p_\phi}\right|}{\left|\Omega_{p_o}\right|} & \text{if } i\epsilon\Omega_{p_\phi}, \\ 1 & \text{otherwise} \end{cases}} \quad \textbf{(15)}$$

where Ω_p is the pixel space, Ω_{p_ϕ} is set of pixels where the ground truth probabilities are not zero and $\Omega_{p_o} = \Omega_p - \Omega_{p_\phi}$.

The term $\left(\hat{y}_i^j - P_{GT_{i,channel=j}}\right)$ measures the ~~distance~~difference between the prediction and the ground truth. This pixel-wise distance is weighted by w_i to solve the data imbalance problem. As most of the pixels in the output probability space have zero probabilities, without this weighting term the solution becomes biased towards the probability of the majority pixels. In our case, $< 5\%$ pixels have non-zero values, thus without the weighting term, the network converges to predict a flat distribution of zeros.

The network is trained on a system with a NVIDIA Pascal Titan X GPU for 30 epochs with a batch-size of 25 image patches. The training took approximately 72 h~~ours~~.

4.4 ~~Inference and Post-processing~~Inference and post-processing

At the test time, our localization algorithm provides an automatic region of interest. Using this automatic localization result, we create a grid of uniformly distributed points and from each point, multiple patches are generated with different scales and rotations. These patches are passed through the center localization network to generate patch-level probability maps. The network takes about 0.14 s~~econd~~ to generate a patch-level prediction. The patch size, orientation and position of these probability maps on the original are known from the patch creation process. These probability maps are then put back on the original image (Fig. 10a). The process includes resizing the 64 × 64 pixel patch to the original patch resolution and projecting it back on the original image using the known patch orientation and position. The probabilities on the original resolution are then thresholded to remove noise (Fig. 10b). The noise is defined as predictions with less than 30% of the maximum probability. For every remaining proposal for a possible vertebral center, the pixel location with the maximum probability is considered as a potential center (Fig. 10b). Further post-processing is performed by removing multiple centers in close proximity by keeping the most confident center in a radius of 10 mm (Fig. 10c). The radius is chosen based on the average size of the training vertebra~~e~~. Finally, we keep the maximum number of possible centers to five (C3~~-~~–C7) and delete less confident center proposals if more than five centers are detected (Fig. 10d).

4.5 ~~Experiments and Metrics~~Experiments and metrics

The center localization framework is tested on our 172 test images. At the patch~~-~~-level, the performance of the network is measured ~~a~~by comparing the predicted probability maps and ground truth maps using the Bhattacharyya coefficient [27].

After the post-processing step, the centers are localized on the original image. The predicted vertebra~~le~~ centers can be divided into three sets: true positive (TP), false positive (FP) and false negative (FN). The TP represents the set of vertebrae whose centers have been correctly detected. A correct detection is considered if the predicted center falls inside a vertebral body studied in this work i.e. C3~~-~~–C7. The FP represents the set of predicted centers which did not fall inside any of these vertebrae. Finally, the FN is the set of the studied vertebrae whose centers have not been detected. Based on the TP, FP and FN, we can report two metrics: true positive rate (TPR) and false discovery rate (FDR) [28]. We also report the Euclidean distance between the correctly detected centers and corresponding ground truth in mm as the distance error.

$$TPR = \frac{|TP|}{|TP| + |FN|} \times 100\% \quad TPR = \frac{|TP|}{|TP| + |FN|} \times 100\%.$$

$$FDR = \frac{|FP|}{|FP| + |TP|} \times 100\% \quad FDR = \frac{|FP|}{|FP| + |TP|} \times 100\%.$$

4.6 Results

The performance of the center localization algorithm is measured independent of the global localization results. For this independent study, the uniform grid needed for the patch creation is generated using the localization ground truth (Fig. 3) instead of the prediction of the spine localization framework as mentioned in Section 4.4. A Bhattacharyya coefficient (BC) of zero represents the worst result and one represents a perfect match between ground truth and prediction probability. Over all the test patches, an average BC of 0.58 has been achieved at the patch~~-~~-level. Some of the graphical results with corresponding BC are shown in Fig. 10e~~-~~–h. It can be seen that even with low BC (Fig. 10g and ~~,~~h), the results are similar. The histogram of the BC over all the patches is plotted in Fig. 11a, a BC of >0.5 was achieved for 71% of the test patches. A few qualitative results for the center localization at the patch~~-~~-level are shown in Fig. 11b.

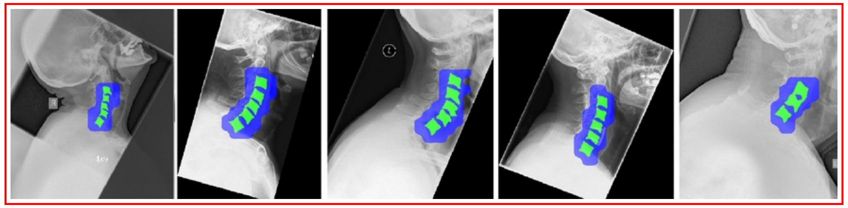

Fig. 3 Global localization ground truth: vertebrae are shown in green, blue overlay indicates the extra area covered by the localization ground truth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 3

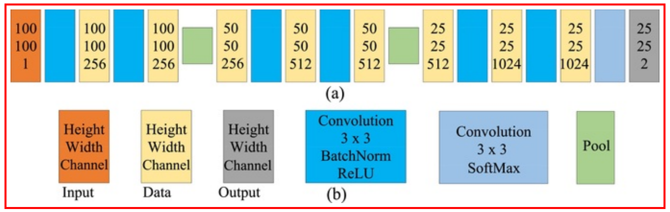

Fig. 4 Fully convolutional network for localization of spinal region (a) Network architecture, (b) Legends.

alt-text: Fig. 4

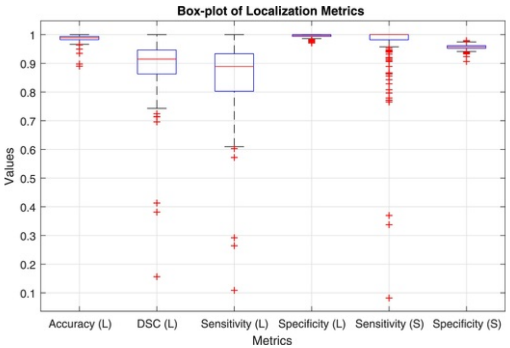

Fig. 5 Box-plot of global localization metrics. 'L' indicates the metrics computed at the lower resolution of 25×25. 'S' indicates the metrics computed at the original image resolution by comparing the prediction with the vertebra~~e~~ segmentation ground truth (green area in Fig. 3).

(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 5

After the post-processing phase, the centers are localized on the full resolution test image. Table 2 reports the true positive rate (TPR), false discovery rate (FDR) and distance error for the correctly detected centers in millimeters (mm).

Table 2 (Please center align the texts with respect the cells and show the border lines. The table becomes really confusing without the borders. Keep the percentages as before, not in the column header as the numbers in the last row are not %.)Performance of the center localization framework. The 'Manual' patch creation process uses localization ground truth and the results reported below are independent of the accuracy of the global localization framework. Results from the fully automatic procedure which uses the localized spine from the global localization framework are reported in the right under the 'Automatic' patch creation process.

alt-text: Table 2

Test patch creation	Manual (%)				Automatic (%)		
True positive rate (TPR)	93.73~~%~~				90.46~~%~~		
False discovery rate (FDR)	4.72~~%~~				10.89~~%~~		
	Median	Mean	Std		Median	Mean	Std
Distance error (mm)	1.63	1.81	0.95		1.54	1.69	0.92

Among 797 vertebrae from our 172 test images, 747 centers were detected with an average error of 1.81 mm. Number of false positive was 37, most of these false positives belong to neighbo~~u~~ring vertebrae C2 and T1. To compare the performance of the center localization algorithm with human performance, an expert radiographer was asked to click on the vertebra~~e~~l centers on ten random test images three times. These manually predicted centers are compared with the ground truth centers for those image. The average error was 1.92 mm which is higher than the average error of correctly detected centers by our algorithm. The performance curve is shown in Fig. 12.

It can be seen that the distance error is <3 mm for almost 90% of the correctly detected vertebral~~e~~ centers. The process is repeated by changing the uniform grid creation process in the beginning. In this case, the uniform grid for patch generation is done using the area predicted by our global localization algorithm (instead of the global localization ground truth), as discussed in Section 4.4. The metrics are reported on the right side of Table 2. It can be seen the TPR dropped from 93.73% to 90.46%, where the FDR is increased from 4.72% to 10.99%. This degradation is because of the incorrect global localization results, as shown in Fig. 6e~~,~~ and f. However, among the correctly detected centers, the distance error drops from 1.81 mm to 1.69 mm. The reason behind this is that much of the bad quality image areas have already been cut off by the global localization prediction. So the remaining image areas are of comparatively of good quality thus the center localization performs better on average on these image areas. Some graphical center localization results in the original resolution are shown in Fig. 13.

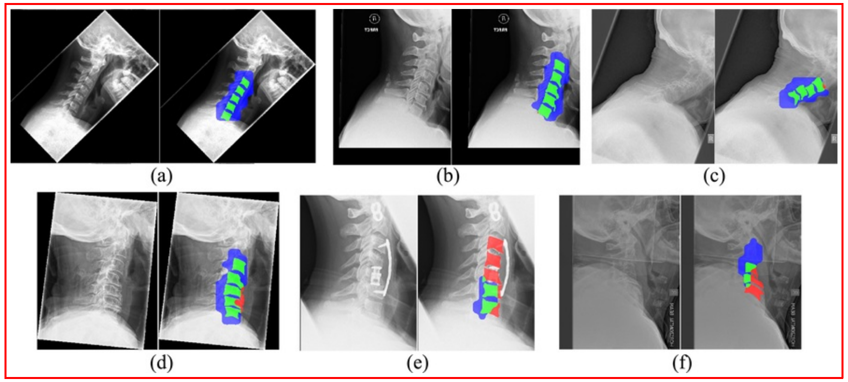

Fig. 6 Qualitative global localization results compared with vertebrae ground truths: true positive (green), false positive (blue), false negative (red), true negative (no overlay) (a) healthy subject (b) Osteophytes (c) Severe degeneration (d) Osteophytes (e) Implants (f) Severe degeneration and osteophytes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 6

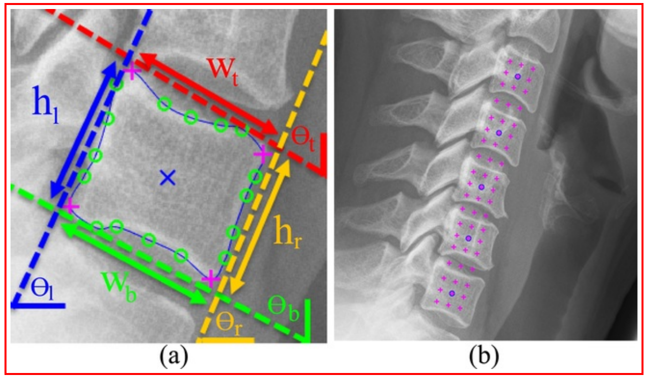

Fig. 7 (a) Probabilistic ground truth creation: manually clicked vertebral center (✕), manually annotated vertebral boundary (○) and corner (+) points (b) Grid points (+) for training patches.

alt-text: Fig. 7

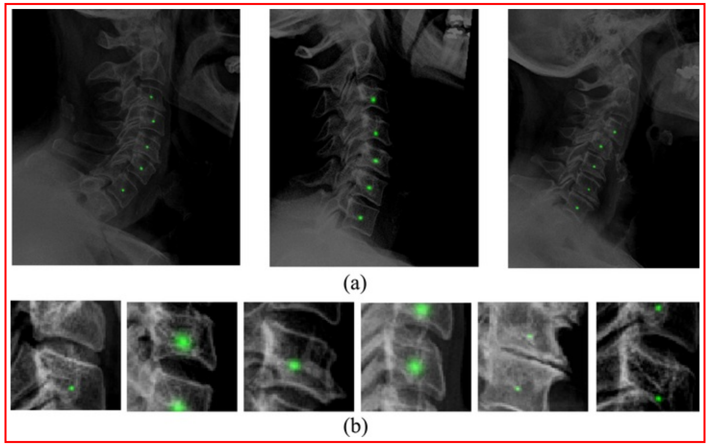

Fig. 8 (a) Probabilistic distribution for vertebrale centers defined over the image space. The intensity of the green overlay represents the probability of the manually clicked centers. (b) Patch-level ground truth for center localization framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 8

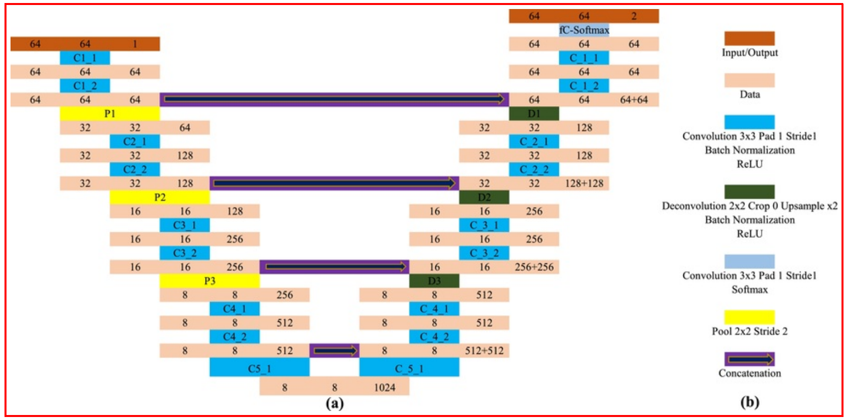

Fig. 9 UNet architecture: (a) Network diagram. (b) Legends.

alt-text: Fig. 9

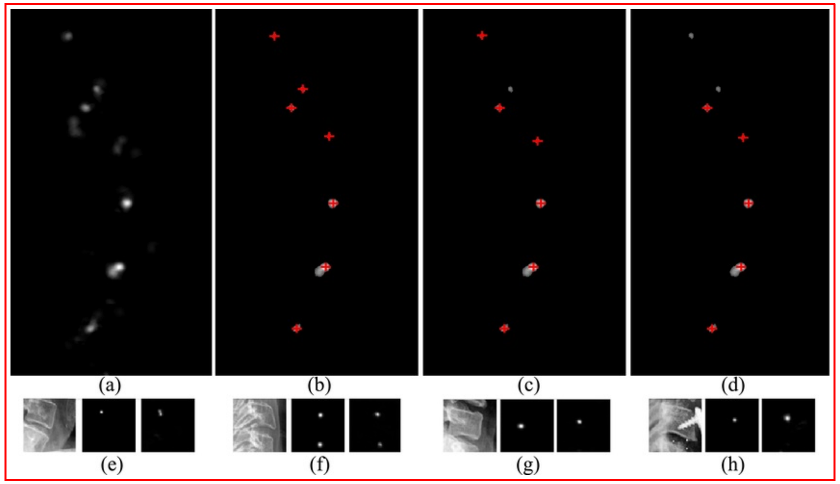

Fig. 10 (a)–(d): Center localization post-processing. (a) Probability map on the original image. (b) Thresholded map and potential centers (+). A1 (c) Filtered centers by after proximity analysis. (d) Five most confident centers. (e)–(h): Bhattacharyya coefficients between the ground truth (middle) and predicted (right) probability distributions with corresponding input image patch (left). (e) 0.8285. (f) 0.7153. (g) 0.3304. (h) 0.3715.

alt-text: Fig. 10

Annotations:

A1. it looks magenta, if so, please change it to red.

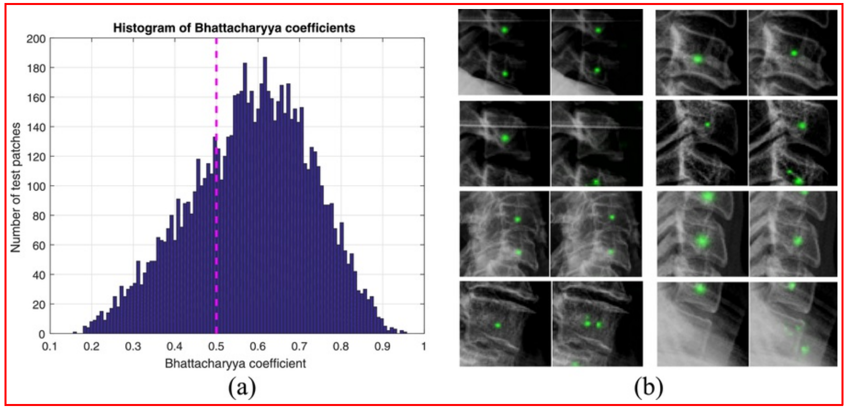

Fig. 11 (a) Histogram of Bhattacharyya coefficients. (b) Patch-level center localization results: ~~G~~ground truth (left) and ~~P~~prediction (right).

alt-text: Fig. 11

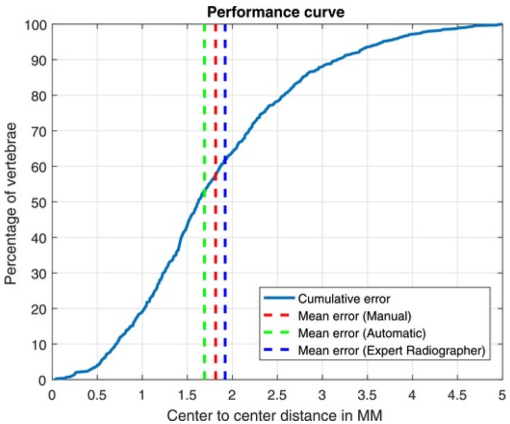

Fig. 12 Performance curve for center localization.

alt-text: Fig. 12

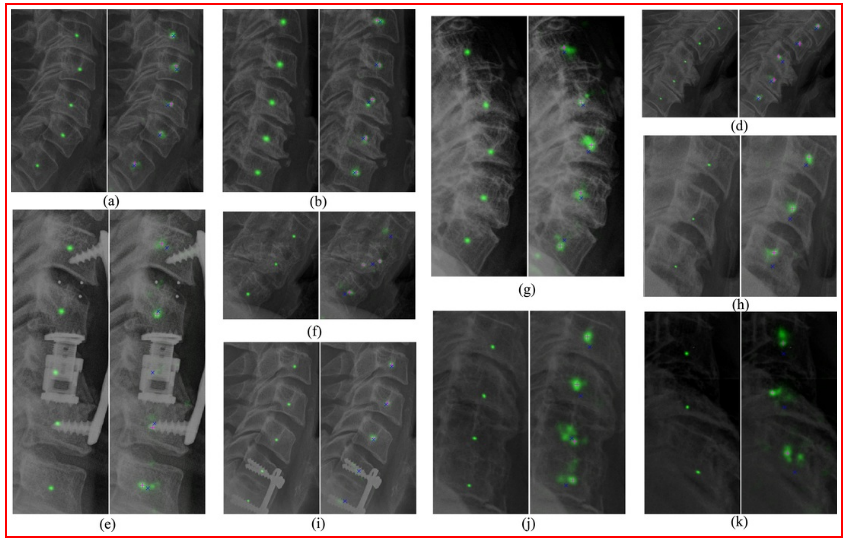

Fig. 13 Qualitative center localization results. For each pair ground truth distribution is shown on the left, prediction distributions are shown on the right. On the prediction image, the ground truth center is denoted as a blue cross (×) and predicted centers are denoted as magenta plus (+). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 13

5 ~~Vertebrae Segmentation~~Vertebrae segmentation

The final and the most important task in our fully automatic segmentation framework is to segment the vertebrae. We use the same UNet architecture with a segmentation loss function for this task. We also introduce a novel shape-aware term in segmentation loss function to predict the vertebral~~e~~ shape with better accuracy.

5.1 Data

To train and test our segmentation framework independent of the global and center localization phase, the manually clicked center points are used to extract the vertebra~~e~~ image patch and corresponding segmentation masks. These can be replaced by the predicted centers making the process fully automatic. From our 124 training images, we have only 586 training vertebrae. To augment the training data different patch sizes and rotation angles are considered. After data augmentation, there were ~~26,370~~26,370 vertebra~~e~~ training patches. All the patches were then resized to 64×64 pixel patches. The corresponding binary segmentation masks were created using the manually annotated vertebral~~e~~ boundary curves (green curves shown in Fig. 2). The pixels inside the boundary curves are considered as the foreground class and outside are considered as the background class [29]. A few training vertebral~~e~~ patches and corresponding overlayed segmentation masks are shown in Fig. 14. Note the differences in intensity, texture, and contrast, coupled with the possibility of surgical implants, making for a challenging problem on real-world data. Similarly, vertebral~~e~~ patches were also collected from the test images, a total of 797 vertebrae were extracted. No augmentation was performed for the test vertebrae.

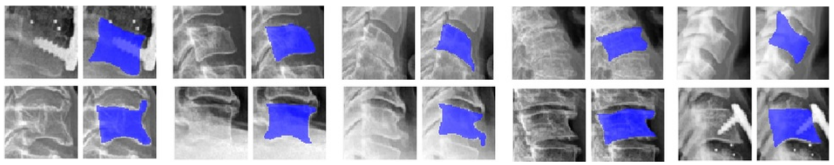

Fig. 14 Training vertebral~~e~~ patches and corresponding segmentation masks (blue overlay). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 14

5.2 Training

The same 24,238,210 parameter version of UNet is used for vertebrae segmentation. The network takes a single channel vertebral patch of spatial dimension 64 × 64 and predicts a binary mask of the same size.

Since the global localization network addressed in Section 3.3 also deals with a binary segmentation problem, the same loss function described in Eqs. (1)~~, 2 and~~–(3) can be used for training the segmentation network. However, this loss, L_t, does not~~n't~~ constrain the predicted masks to conform to possible vertebra-l shapes (vertebral shapes). Since vertebra~~e~~l shapes are known from the provided manual segmentation curves, we add a novel shape-aware term in the loss function to force the network to learn to penalize predicted areas outside the curve.

5.3 ~~Shape-aware Loss Term~~Shape-aware loss term

For training the deep segmentation network, we introduce a novel shape–based loss term, L_s. This term encourages the network to produce a prediction masks similar to the training vertebral shapes. This term can be defined as:

(I have modified the equation to take E_i out of the second summation and replace M with 2. After that the lower limit (j=1) and upper limit (2) of the second summation are not aligning with the summation in the pdf document. Please make sure they align correctly in the final document.)

$$L_s(\{x,y\};W) = -\sum_{i\in\hat{\Omega}_p}\sum_{j=1}^{M} y_i^j E_i \log P\left(y_i^j = 1 \middle| x_i; W\right) \quad L_s(\{x,y\};W) = -\sum_{i\in\hat{\Omega}_p} E_i \sum_{j=1}^{2} y_i^j \log P\left(y_i^j = 1 \middle| x_i; W\right),$$

$$E_i = D\left(\hat{C}, C_{GT}\right) \quad E_i = D\left(\hat{C}, C_{GT}\right), \tag{16}$$

where $\hat{C}$ is the curve surrounding the predicted regions and C_{GT} is ground truth curve. The function, $D(.)$, computes the average point to curve Euclidean distance between the predicted shape, $\hat{C}$ and the ground truth shape, C_{GT}. $\hat{C}$ is generated by locating the boundary pixels of the predicted mask. The redefined pixel space, $\hat{\Omega}_p$, contains the set of pixels where the prediction mask doesn't match the ground truth mask. These terms can also be explained using the toy example shown in Fig. 15. Given a ground truth mask (Fig. 15a) and a prediction mask (Fig. 15b), E_i is computed by measuring the average distance between the ground truth (green) curve and prediction (red) curve (Fig. 15c). Fig. 15d shows the redefined pixel space, $\hat{\Omega}_p$. This term is an additional penalty proportional to the Euclidean distance between predicted and ground truth curve to the pixels that do not match the ground truth segmentation mask. In the case when the predicted mask is a cluster of small regions, especially during the first few epochs in training, E_i becomes large because of the increase in the boundary perimeters from the disjoint predictions.

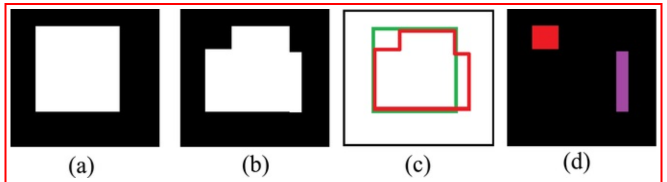

Fig. 15 Shape-aware loss: (a) Ground truth mask. (b) Prediction mask. (c) Ground truth shape, C_{GT} (green) and predicted shape, $\hat{C}$ (red). (d) Refined pixel space, $\hat{\Omega}_p$: False positive (purple) and false negative (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 15

Finally, the loss function of Eq. (1) can be extended as:

$$\hat{W} = \arg\min_{W} \sum_{n=1}^{N} \left(L_t\left(\left\{x^{(n)}, y^{(n)}\right\}; W\right) + L_s\left(\left\{x^{(n)}, y^{(n)}\right\}; W\right)\right) \quad \hat{W} = \arg\min_{W} \sum_{n=1}^{N} \left(L_t\left(\left\{x^{(n)}, y^{(n)}\right\}; W\right) + L_s\left(\left\{x^{(n)}, y^{(n)}\right\}; W\right)\right). \tag{17}$$

The contribution of each term in the total loss can be controlled by introducing a weight parameter in Eq. (17). However, in our case, the best performance was achieved when both terms contributed equally.

5.4 ~~Experiments and Metrics~~Experiments and metrics

We have two versions of the deep segmentation network: UNet and UNet-S. '-S' signifies the use of the updated shape-aware loss function of Eq. (17). Both segmentation networks are trained on a system with a NVIDIA Pascal Titan X GPU for 30 epochs with a batch-size of 25 image patches. Each network took approximately 28 h~~ours~~ to train. In order to compare with the deep segmentation network–based prediction results, three active shape model (ASM)–based shape prediction frameworks have been implemented. A simple maximum gradient–based image search–based ASM (ASM-G) [30], a Mahalanobis distance–based ASM (ASM-M) [5] and a random forest–based ASM (ASM-RF) [11].

The latter two have been used in cervical vertebrae segmentation in different datasets.

At test time, 797 vertebrae from 172 test images are extracted based on the manually clicked vertebral centers. These patches are sent through each of the networks in a forward pass to get the prediction masks. It takes about 0.13 s~~econd~~ to produce a patch-level prediction. These prediction masks are compared with the ground truth segmentation mask to compute pixel-wise accuracy (pA) and Dice similarity coefficients (DSC). For the ASM-based shape predictors, the predicted shape is converted to a prediction map to measure these metrics. These metrics are well suited to capture the number of correctly segmented pixels, but they fail to capture the differences in shape. In order to compare the shape of the predicted mask appropriately with the ground truth vertebral~~e~~ boundary, the predicted masks of the deep segmentation networks are converted into shapes by locating the boundary pixels. These shapes are then compared manually annotated vertebral boundary curves by measuring average point to curve Euclidean distance between them, similar to Eq. (16). A final metric, called fit failure [10], is also computed which measures the percentage of vertebrae having an average point to ground truth curve error of greater than 1 mm.

5.5 Results

Table 3 reports the average median, mean and standard deviation (std) of the metrics over the test dataset of 797 vertebrae for all the methods. Deep segmentation network-based methods clearly outperform the ASM based methods. Even the worst performing version of our framework, UNet, achieves a 2.9% increase in terms of the pixel-wise accuracy and an increase of 5.5% for the Dice similarity coefficient. Among the two versions of the deep networks, the use of the novel loss function improves the performance by 0.31% in terms of pixel-wise accuracy. In terms of the Dice similarity coefficient, the improvement is in the range of 0.6%. The differences are small quantitatively, but the improvements are statistically significant according to a paired t-test at a 5% significance-level. ~~Corresponding p-values between the two versions of the network are reported in Table 3.~~ Also, one would expect a larger pixel-wise accuracy and Dice similarity when there are many true positive pixels in the center of the segmentation result. Corresponding p-values between the two versions of the network are reported in Table 3. A bold font indicates the best performing metrics. Interestingly, among the ASM-based methods, the simplest version, ASM-G, performs better than the alternatives. Recent methods [5,11], have failed to perform robustly on our challenging dataset of the test vertebrae.

Table 3 (p-values should be in merged cells for UNet and UNet-S. Merged cells become confusing without cell borders. Please add borders if possible. The table should be equivalent to the table in the attached image. Also if possible, please center-align the texts both vertically and horizontally.)Average quantitative metrics for vertebrae segmentation.

alt-text: Table 3

	Pixel-wise accuracy (%)				Dice similarity coefficient			
	Median	Mean	Std	p-value	Median	Mean	Std	p-value
ASM-RF	95.09	90.77	8.98		0.881	0.774	0.220	
ASM-M	95.09	93.48	4.92		0.900	0.877	0.073	
ASM-G	95.34	93.75	4.48		0.906	0.883	0.066	
UNet	97.71	96.69	3.04	$< 10^{-12}$	0.952	0.938	0.048	$< 10^{-12}$
UNet-S	**97.92**	**97.01**	**2.79**		**0.957**	**0.944**	**0.044**	

Although statistically significant, the stability of the small improvement between UNet and UNet-S may be subjected to the fixed set of data used for the training and testing. In order to test the stability of the performance, two new sets of UNet and UNet-S were trained with randomly scrambled datasets. In both cases, UNet-S outperformed the UNet with statistical significance. The Dice similarity coefficients for these re-scrambled datasets are reported in Table 4.

Table 4 (p-values should be in merged cells for UNet and UNet-S. The table should be equivalent to the table in the attached image. Also if possible, please center-align the texts both vertically and horizontally.)Dice similarity coefficients for re-scrambled datasets.

alt-text: Table 4

	Re-scrambled Dataset 1	Re-scrambled Dataset 2

	Mean	Std	p-value	Mean	Std	p-value
UNet	0.9371	0.0412	$< 10^{-03}$	0.9433	0.0712	< .013
UNet-S	0.9411	0.0366		0.945	0.0692	

The average point to curve error for the methods are reported in Table 5. This measure is important as it captures the differences in the segmentation boundary which defines the shape. The deep segmentation framework, UNet, produced a 35% improvement over the ASM-based methods in terms of the mean values. The introduction of the novel loss term in the training further reduced the average error by 11% achieving the best error of 0.55 mm. The most significant improvement can be seen in the fit failure which denotes the percentage of the test vertebrae having an average error of higher than 1 mm. The novel shape-aware network, UNet-S, has achieved a drop of around 76% from the ASM-RF method. The cumulative distribution of the point to curve error is also plotted in the performance curve of Fig. 16. It can be seen that deep segmentation networks provide a large improvement among the deep networks, shape-aware UNet performs better.

Table 5 (p-values should be in merged cells for UNet and UNet-S. Merged cells become confusing without cell borders. Please add borders if possible. The table should be equivalent to the table in the attached image. Also if possible, please center-align the texts both vertically and horizontally.)Average quantitative metric for shape prediction.

alt-text: Table 5

	Average point to curve error in mm				Fit failure(%)
	Median	Mean	Std	~~p~~p-value	
ASM-RF	1.51	1.74	0.95		74.40
ASM-M	0.87	1.02	0.56		39.52
ASM-G	0.77	0.95	0.54		31.49
UNet	0.43	0.62	0.81	~~0~~.0062	9.41
UNet-S	**0.44**	**0.55**	**0.40**		**7.40**

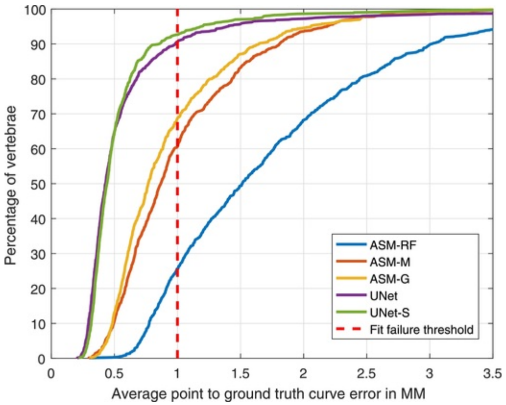

Fig. 16 Performance curve: ~~C~~cumulative distribution of point to curve errors.

alt-text: Fig. 16

The box-plots of the quantitative metrics are shown in Fig. 17. It can be seen that even the worst outlier for the shape-aware network, UNet-S, has a pixel-wise accuracy higher than 70%, signifying the regularizing capability of the novel term. Most of the outliers are caused by bone implants, fractured vertebrae or abnormal art~~e~~ifacts in the images. A few examples for qualitative assessment are shown in Fig. 18. Fig. 18a shows an easy example where all the methods perform well. Examples with surgical bone implants are shown in Fig. 18b and c. Fig. 18d and e show ~~shows~~ vertebrae with abrupt contrast change. Vertebrae with fracture and osteophytes are shown in Fig. 18f and g. Fig. 18g also shows how UNet-S has been able to capture the pattern of the vertebra~~e~~l fracture~~s pattern~~. Fig. 18h and i shows vertebrae with image art~~e~~ifacts. A complete failure case is shown in Fig. 18j. The shape-aware network, UNet-S, has produced better segmentation results than its counterpart, UNet. Qualitatively we conclude that the novel shape-aware term provides equivalent or improved results in nearly all cases.

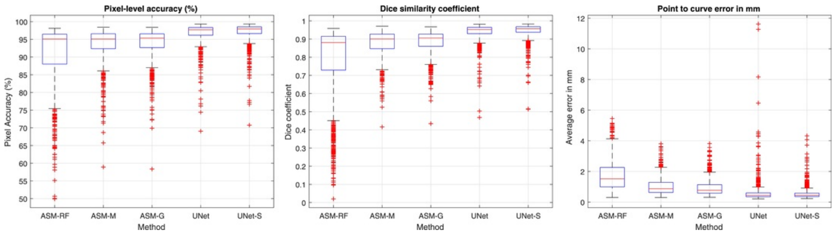

Fig. 17 Box-plots of the quantitative metrics: pixel-level accuracy (left), Dice similarity coefficients (middle) and point to manual segmentation curve error (right).

alt-text: Fig. 17

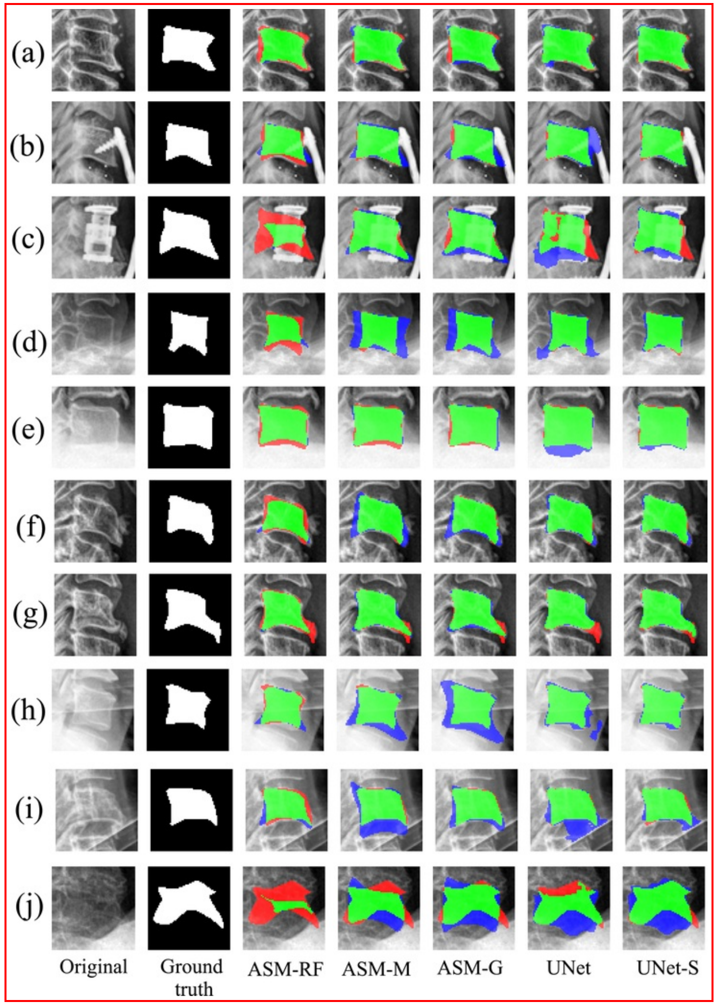

Fig. 18 Qualitative segmentation results: true positive (green), false positive (blue) and false negative (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 18

Analysis on harder cases Although the difference in performance between the UNet and UNet-S is stable and statistically significant, the improvement is subtle over the whole dataset of the test vertebrae. This is because the majority of th vertebrae are healthy and easier to segment. Therefore adding the shape-aware term does not improve the results by a large margin. However, on more challenging vertebrae a larger difference is observed. To show the usefulness of adding the shape-awar term in UNet-S, a selection of 52 vertebrae with severe clinical conditions are chosen. The average metrics for this subset of test vertebrae between UNet and UNet-S are reported in Table 6. An improvement of 1.2% and 0.02 have been achieved in terms c the pixel-wise accuracy and the Dice similarity coefficient, respectively. The difference over the whole dataset were only 0.31% and 0.006. The metric, point to curve error produces the most dramatic change. The novel shape-aware network, UNet-S, reduce the error by 25% for this subset of vertebrae with severe clinical conditions. Fig. 19 shows a few examples of these images.

Table 6 Comparison of UNet and UNet-S for vertebrae with clinical conditions.

alt-text: Table 6

	Average quantitative metrics

	Pixel-wise accuracy (%)	Dice coefficient	Point to curve error
UNet	94.01	0.91	0.84
UNet-S	95.21	0.93	0.63

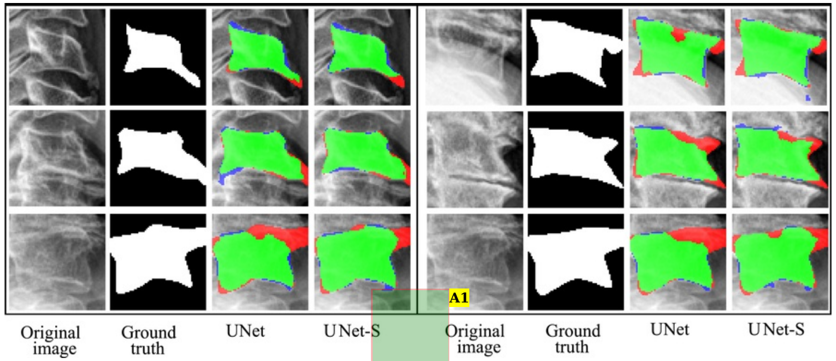

Fig. 19 Comparison of performance for vertebrae with severe clinical condition.

alt-text: Fig. 19

Annotations:

A1. will it be possible to replace the embedded text in this figure with normal text as provided in the original pdf version of this image? if not possible to put text accordingly please leave the figure as it is.

6 ~~Fully Automatic Segmentation Framework~~Fully automatic segmentation framework

Now, having the three subtasks i.e. global localization, center localization and vertebrae segmentation frameworks in place, a single fully automatic vertebrae segmentation framework can be formulated. Given a high resolution test image, the image can be zero-padded to form a square image and resized to 100×100 pixel. This image can be fed into the global localization FCN to predict the spinal region. The global localization algorithm localizes the spinal region at a lower resolution of 25×25 pixel, which can then be transformed back, i.e. resizing and unpadding, to the original image. The process is summarized in Fig. 20-1.

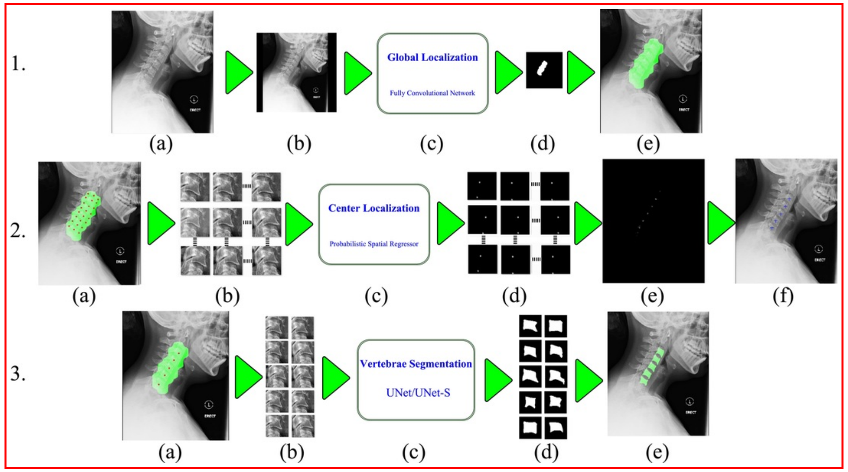

Fig. 20 1. Global localization ~~F~~framework (a) Full resolution X-ray image. (b) 100×100 pixel input image. (c) Global Localization FCN. (d) Network prediction at 25×25 pixel. (e) Localized spine in the original resolution. 2. Center localization framework (a) Grid points on localized spinal region. (b) Generated image patches. (c) Center localization network. (d) Patch-level probabilities. (e) Probabilistic center maps on original image space. (f) Localized centers after post-processing. 3. Vertebrae segmentation

framework (a) Localized spinal region and centers. (b) Extracted vertebra~~e~~l patches. (c) Segmentation network (UNet/UNet-S). (d) Patch-level segmentation results. (e) Segmented vertebrae on the original image.

alt-text: Fig. 20

Based on the global localization result, a uniformly spaced grid of points can be generated. From these points, image patches can be extracted with multiple scales and orientations. All the patches are then resized to 64 × 64 pixel and passed through the novel probabilistic spatial regressor network. Each patch generates a probability map of localized centers. These patch-level probabilities are then put back on the original image space. And centers are localized using the post-processing steps of Section 4.4. Fig. 20-2 depicts the center localization process.

The localized spinal region map from the global localization step and the localized centers from center localization step are used to determine the orientations and scales of each vertebra in the image. Based on this information, for each center proposal, multiple patches are extracted and resized to 64 × 64 pixel. These patches are then passed through ~~the~~ one of the vertebrae segmentation networks, UNet or UNet-S. The patch-level predictions are then put back on the original image space to create the final segmentation results. The process of the vertebrae segmentation is shown in Fig. 20-3.

Since none of the subtasks requires manual intervention and the input information required by the latter subtasks is provided by the result of the previous subtasks, a complete framework can be designed by cascading the subtasks sequentially. The complete framework is fully automatic and does not~~n't~~ require any human input to generate vertebrae segmentation of an X-ray image. To our knowledge, the proposed framework is the first in the literature that presents a fully automatic cervical spine segmentation method. The flowchart for the complete framework has been shown in Fig. 1. The runtime for the framework varies from 11 s~~econds~~ to one minute with an average time of 24 s~~econds~~ using unoptimized Matlab implementation on a system without GPUs. Most of this time is taken by the post-processing steps of the center localization and vertebrae segmentation subtasks where the patch-level predictions are transformed back to the original image space.

6.1 Results

The Dice similarity coefficient (DSC) and shape error for the final segmentation results are summarized in Table 7. The predicted shape is computed by locating the boundary pixels of the predicted final segmentation map. The predicted shapes are compared with the manually annotated shapes, illustrated by green curves in Fig. 2. The average error in millimeter (mm) is reported as the shape error. Both UNet and UNet-S have been tested as the final segmentation module. Both perform similarly in terms of the reported metrics. The mean Dice similarity coefficient is exactly the same at 0.84. The performance is lower than the Dice similarity coefficient of 0.944 reported in Table 3 because of the full automation and the accumulated errors from the global localization and center localization phase. Since most of the difficult vertebra~~e~~ samples do not get into the segmentation phase and difference in performance is not noticeable in terms of DSC. However, as the major difference between the networks is a shape-aware term, shape error have achieved a 0.35% relative improvement even after full automation. The histogram plots of these two metrics are shown in Fig. 21.

Table 7 Performance of fully automatic framework.

alt-text: Table 7

	Dice similarity coefficients		Shape error in mm	
	Mean	Std	Mean	Std
UNet	0.840	0.136	1.695	2.614
UNet-S	0.840	0.135	1.689	2.555

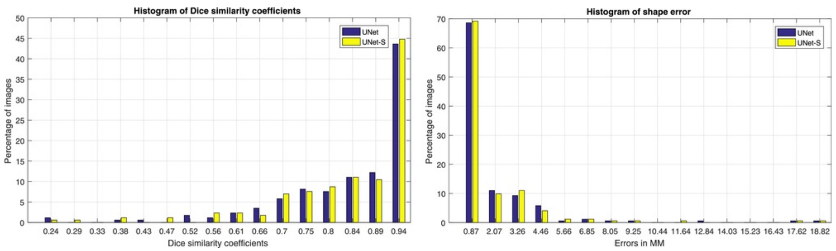

Fig. 21 Histogram plot of Dice similarity coefficients (~~top~~left) and shape error (~~bottom~~right) for the fully automatic framework with UNet and UNet-S.

alt-text: Fig. 21

Some qualitative results are shown in Fig. 22. It can be seen that even with severe clinical conditions (rows 3~~,~~ and 4) and image art~~e~~ifacts (row 5) the fully automatic algorithm has been able to produce accurate segmentation results. However, the algorithm does not~~n't~~ guarantee acceptable segmentation everywhere. Some less accurate results on difficult cases are shown in Fig. 23. Row 1 of Fig. 23 shows a case where the center localization framework failed to detect a vertebra~~e~~l center with osteophytes (C5) and detected a false center from vertebra~~e~~ C2. Thus the final segmentation results have a false positive in vertebra~~e~~ C2 and a false negative for C5. The second row shows a case where both global localization and center localization failed due to surgical implants in the lower vertebrae (C6~~-~~ and C7). A test case with severe osteoporosis and bone loss is shown in row 3. Even with such severe clinical condition the global localization and center localization algorithm were able to produce correct results for C3 and C4, however, the segmentation framework still suffered to segment those correctly. Another severe condition with bone loss, osteoporosis and vertebrae fusion is shown in row 4 of Fig. 23. Even with such ~~severe~~ conditions, global localization and center localization have been able to correctly detect four vertebral~~e~~ centers, but unfortunately, a false center has also been detected in the extended part of the C2. Interestingly, the segmentation framework also segmented a vertebra~~e~~-like structure in the extension where the top and bottom border followed the bone structure. However, the segmentation results for the actual vertebrae are incorrect because of the severity of the condition. Finally, in the last row, we have shown a complete failure due to the presence of large surgical implants. The global localization algorithm failed completely thus the following subtasks were not able to perform either.

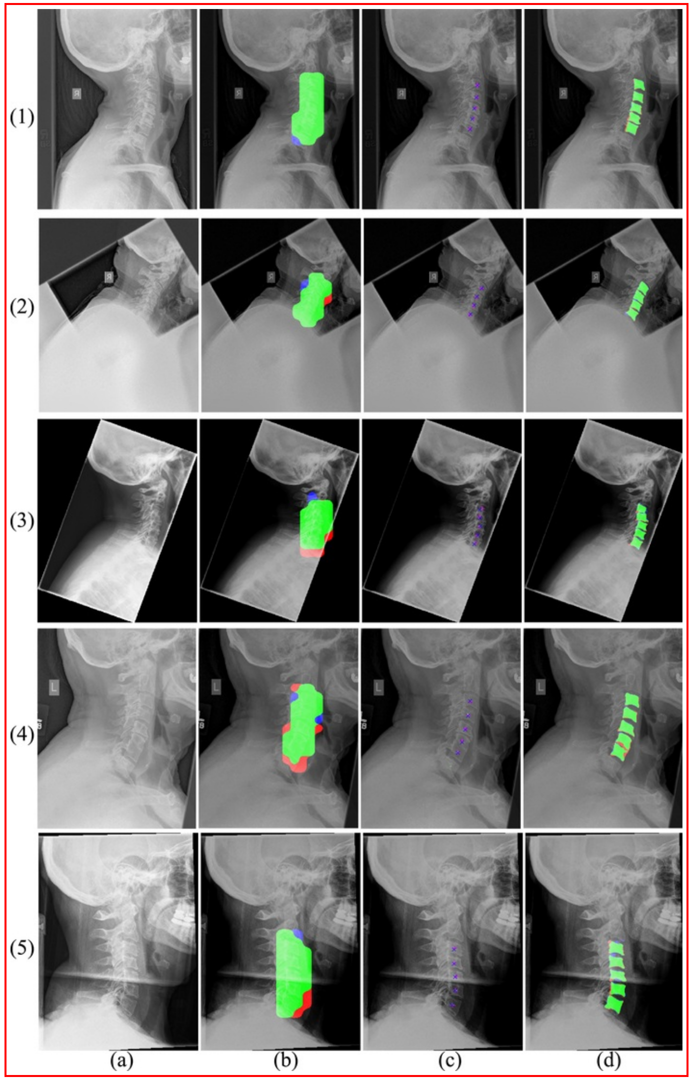

Fig. 22 Fully automatic framework results. True positive (green), false positive (blue) and false negative (red). Ground truth center (×) and predicted centers (+). (a) Original image. (b) Global localization. (c) Center localization. (d) Vertebrae segmentation. (For interpretation of th references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 22

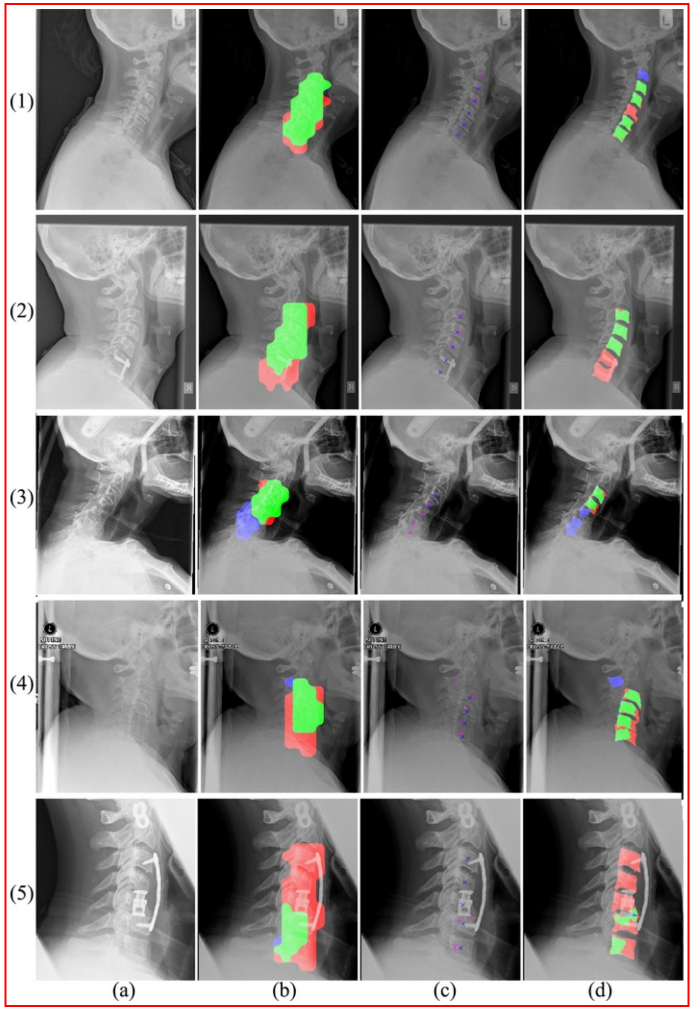

Fig. 23 Fully automatic framework results for challenging cases. True positive (green), false positive (blue) and false negative (red). Ground truth center (×) and predicted centers (+). (a) Original image, (b) Global localization, (c) Center localization, (d) Vertebrae segmentation.

(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

alt-text: Fig. 23

7 Conclusion

The cervical spine is one of the most important yet vulnerable anatomies of the human body. Despite advances in imaging technologies, a large number of cervical injuries remain unnoticed in the emergency room. Towards building a fully automatic injury detection system, in this paper, using the recent advances in deep learning technologies, we have proposed a fully automatic vertebrae segmentation framework for X-ray images. The complete process is divided into three subtasks: localization of the spine, localization of the vertebra~~e~~l centers and segmentation of the vertebrae. We have proposed a solution ~~to~~for each of these subtasks using deep learning concepts. First, we have proposed a novel approach of using fully convolutional segmentation network for solving a localization problem. Our global localization algorithm produced a sensitivity and a specificity of 0.96 in localizing the vertebrae in the X-ray images. Second, we have introduced a novel loss function for predicting a probabilistic map using a fully convolutional network for localizing image landmarks. Our center localization framework has been able to correctly detect 93.73% of vertebrae with an average error of 1.81 mm. Third, we have proposed a novel shape-aware loss term for vertebrae segmentation. The shape-aware segmentation has produced an average Dice similarity coefficient of 0.944

and an average point to curve error of 0.55 mm over a dataset full of real-life emergency room X-ray images, containing surgical implants, clinical conditions and image art~~e~~ifacts. Last but not the least, we have proposed a complete and fully automatic framework for vertebrae segmentation in X-ray images which has been able to produce a final Dice similarity coefficient of 0.84.

The current framework still has several limitations. The center localization framework can be further improved by removing outlier centers away from the vertebral curve. The current patch~~-~~-based center localization framework has the limitation of not knowing which center belongs to which vertebra. We are currently working on a vertebra detection framework, which will be able to determine which vertebrae are visible in the image. The shape-aware segmentation framework can further be improved to determine if a segmented vertebral~~e~~ shape is regular or injurious/fractured. The next step in our research is to build a complete injury detection system which will be able to help the emergency room physicians by highlighting spinal areas with high possibility of injuries. The proposed framework is general and can be extended to other views of the cervical spine, including odontoid peg and anteroposterior (AP) views.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] D. A. P. Singh, Range of Motion of Cervical Spine, (http://boneandspine.com/range-motion-cervical-spine/), Accessed: 2017-11-09.

[2] A. Singh, L. Tetreault, S. Kalsi-Ryan, A. Nouri and M.G. Fehlings, Global prevalence and incidence of traumatic spinal cord injury, ~~*Journal of Clinical Epidemiology*~~*J. Clin. Epidemiol.* **6**, 2014, 309–331.

[3] P. Platzer, N. Hauswirth, M. Jaindl, S. Chatwani, V. Vecsei and C. Gaebler, Delayed or missed diagnosis of cervical spine injuries, *J. Trauma Acute Care Surg.* **61** (1), 2006, 150–155.

[4] C. Morris and E. McCoy, Clearing the cervical spine in unconscious polytrauma victims, balancing risks and effective screening, *Anaesthesia* **59** (5), 2004, 464–482.

[5] M. Benjelloun, S. Mahmoudi and F. Lecron, A framework of vertebra segmentation using the active shape model-based approach, *J. Biomed. Imaging* **2011**, 2011, 9.

[6] M.A. Larhmam, S. Mahmoudi and M. Benjelloun, Semi-automatic detection of cervical vertebrae in X-ray images using generalized hough transform, Proceedings of the 2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA), 2012, IEEE, 396–401.

[7] M. Roberts, T.F. Cootes and J.E. Adams, Vertebral morphometry: semiautomatic determination of detailed shape from dual-energy X-ray absorptiometry images using active appearance models, ~~*Investigative Radiology*~~*Investig. Radiol.* **41** (12), 2006, 849–859.

[8] M. Roberts, E. Pacheco, R. Mohankumar, T. Cootes and J. Adams, Detection of vertebral fractures in DXA VFA images using statistical models of appearance and a semi-automatic segmentation, *Osteoporos. Int.* **21** (12), 2010, 2037–2046.

[9] M.G. Roberts, T.F. Cootes and J.E. Adams, Automatic location of vertebrae on DXA images using random forest regression, Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2012), 2012, Springer, 361–368.

[10] P. Bromiley, J. Adams and T. Cootes, Localisation of vertebrae on DXA images using constrained local models with random forest regression voting, Recent Advances in Computational Methods and Clinical Applications fo Spine Imaging, 2015, Springer, 159–171.

[11] S.M.M.R. Al-Arif, M. Gundry, K. Knapp and G. Slabaugh, Improving an active shape model with random classification forest for segmentation of cervical vertebrae, Proceedings of the 4th International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, CSI 2016, held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, **10182**, 2017, Springer, 3.Revised Selected Papers

[12] T.F. Cootes, Fully automatic localisation of vertebrae in ct images using random forest regression voting, Proceedings of the 4th International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, CSI 2016, held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, **10182**, 2017, Springer, 51.Revised Selected Papers

[13] A. Tezmol, H. Sari-Sarraf, S. Mitra, R. Long and A. Gururajan, Customized Hough transform for robust segmentation of cervical vertebrae from X-ray images, Proceedings of the Fifth IEEE Southwest Symposium on Image Analysis and Interpretation, 2002, 2002, IEEE, 224–228.

[14] B. Glocker, J. Feulner, A. Criminisi, D.R. Haynor and E. Konukoglu, Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans, Proceedings of the International Conference on Medical Image

Computing and Computer-Assisted Intervention (MICCAI 2012), 2012, Springer, 590–598.

[15] B. Glocker, D. Zikic, E. Konukoglu, D.R. Haynor and A. Criminisi, Vertebrae localization in pathological spine CT via dense classification from sparse annotations, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2013, Springer, 262–270.

[16] S.M.M.R. Al-Arif, M. Gundry, K. Knapp and G. Slabaugh, Global localization and orientation of the cervical spine in x-ray images, Proceedings of the 4th International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, CSI 2016, held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, **10182**, 2017, Springer, 64.Revised Selected Papers

[17] S.M.M.R. Al-Arif, M. Asad, K. Knapp, M. Gundry and G. Slabaugh, Hough forest-based corner detection for cervical spine radiographs, Proceedings of the 19th Conference on Medical Image Understanding and Analysis (MIUA), 2015, 183–188.

[18] S.M.M.R. Al-Arif, M. Asad, K. Knapp, M. Gundry and G. Slabaugh, Cervical vertebral corner detection using Haar-like features and modified Hough forest, Proceedings of the 2015 5th International Conference on Image Processing Theory, Tools and Applications (IPTA), 2015, IEEE.

[19] B. Aubert, C. Vazquez, T. Cresson, S. Parent and J. De Guise, Automatic spine and pelvis detection in frontal x-rays using deep neural networks for patch displacement learning, Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 2016, IEEE, 1426–1429.

[20] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, Springer, 234–241.

[21] A. BenTaieb and G. Hamarneh, Topology aware fully convolutional networks for histology gland segmentation, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, Springer, 460–468.

[22] H. Chen, X. Qi, J.-Z. Cheng and P.-A. Heng, Deep contextual networks for neuronal structure segmentation, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, AAAI Press, 1167–1173.

[23] S.A. Mahmoudi, F. Lecron, P. Manneback, M. Benjelloun and S. Mahmoudi, GPU-based segmentation of cervical vertebra in X-ray images, Proceedings of the 2010 IEEE International Conference on Cluster Computing Workshops and Posters (Cluster workshops), 2010, IEEE, 1–8.

[24] S. Ruder, An Overview of Gradient Descent Optimization Algorithms, arXiv preprint arXiv:1609.04747 (2016).

[25] G. Slabaugh, Q. Dinh and G. Unal, A variational approach to the evolution of radial basis functions for image segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07), 2007, IEEE, 1–8.

[26] H. Noh, S. Hong and B. Han, Learning deconvolution network for semantic segmentation, Proceedings of the IEEE International Conference on Computer Vision, 2015, 1520–1528.

[27] A. Bhattachayya, On a measure of divergence between two statistical population defined by their population distributions, *Bull. Calcutta Math. Soc.* **35** (99~~-~~–109), 1943, 28.

[28] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. stat. Soc. Ser. B (Methodol.)* 1995, 289–300.

[29] Convert Region of Interest Polygon to Region Mask, (https://www.mathworks.com/help/images/ref/poly2mask.html#f6-465457/), Accessed: 2017-09-18.

[30] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham, Active shape models-their training and application, ~~*Computer Vision and Image Understanding*~~*Comput. Vis. Image Underst.* **61** (1), 1995, 38–59.

Highlights

- A deep segmentation network based spine localization algorithm which outperforms the previous state-of-the-art by a large margin.
- A novel spatial probability prediction deep convolutional network which achieves human-level performance in localizing vertebrae centers.
- A novel shape-aware deep segmentation network for vertebrae segmentation.

- A first of its kind fully automatic framework which combines the global localization, center localization and vertebrae segmentation in a single thread and provides a segmentation result for a real-life emergency room X-ray images without any manual input.

Queries and Answers

Query: Please confirm that givennames and surnames have been identified correctly.
Answer: Yes

Query: Please validate affiliation “b”.
Answer: correct

Query: Please validate if corresponding author has been correctly identified.
Answer: Identified correctly.

Query: Figs. 2, 3, 5, 6, 8, 13, 14, 15, 18, 22 and 23 have been submitted as color images; however, the captions have been reworded to ensure that they are meaningful when your article is reproduced both in color and in black and white. Please check and correct if necessary.
Answer: I have corrected the captions where necessary. I have also provided original images in hope that the embedded texts in some of the images in the paper could be removed. In the original paper I used '.pdf' images thus the texts appear as text. here the images have been formated as png or tif image where the texts become embedded text, making the quality of article unprofessional. Whereeve possible, please, update the images with the provided replacement images and add the subfigure indexes appropriately.

Query: Please provide year information in Refs. [1] and [29].
Answer: These are web-links. They do not have a publication year. The date when these web-links have been checked are added in the references as Accessed: xxxx-xx-xx.

Query: Please complete and update Ref. [24].
Answer: This is an unpublished work only available as arXiv preprint. It has already been cited in 151 times.

Query: Please provide volume number in Ref. [28].
Answer: No volume number available