Ultrasound Segmentation of Cervical Muscle during head motion: a Dataset and a Benchmark using Deconvolutional Neural Networks

Ryan Cunningham, *Member, IEEE*, María B. Sánchez, and Ian D. Loram, *Member, IEEE*

***Abstract*— Objectives: To automate online segmentation of cervical muscles from transverse ultrasound (US) images of the human neck during functional head movement. To extend ground-truth labelling methodology beyond dependence upon MRI imaging of static head positions required for application to participants with involuntary movement disorders. Method: We collected sustained sequences (> 3 minutes) of US images of human posterior cervical neck muscles at 25 fps from 28 healthy adults, performing visually-guided pitch and yaw head motions. We sampled 1,100 frames (approx. 40 per participant) spanning the experimental range of head motion. We manually labelled all 1,100 US images and trained deconvolutional neural networks (DCNN) with a spatial SoftMax regression layer to classify every pixel in the full resolution (525x491) US images, as one of 14 classes (10 muscles, ligamentum nuchae, vertebra, skin, background). We investigated 'MaxOut' and Exponential Linear unit (ELU) transfer functions and compared with our previous benchmark (analytical shape modelling). Results: These DCNNs showed higher Jaccard Index (53.2%) and lower Hausdorff Distance (5.7 mm) than the previous benchmark (40.5%, 6.2 mm). SoftMax Confidence corresponded with correct classification. 'MaxOut' outperformed ELU marginally. Conclusion: The DCNN architecture accommodates challenging images and imperfect manual labels. The SoftMax layer gives user feedback of likely correct classification. The 'MaxOut' transfer function benefits from near-linear operation, compatibility with deconvolution operations and the dropout regulariser. Significance: This methodology for labelling and training segmentation networks is applicable for dynamic segmentation of moving muscles and for participants with involuntary movement disorders who cannot remain still.**

***Index Terms*—cervical muscles, deep learning, dystonia, head movement, movement disorders, segmentation, ultrasound.**

I. Introduction

Online segmentation of ultrasound (US) images of cervical muscles is required for visualisation, analysis of deep, muscle structure and function, and monitoring of patient specific treatment protocols [1]–[6]. Personalised muscle diagnosis is needed for neck/back pain/injury, work related disorders, myopathies and neuropathies. Specifically for cervical dystonia, online segmentation of neck muscles is required for automated diagnosis, for improving diagnosis and monitoring the delivery and effectiveness of botulinum toxin injections into individual deep muscles[1], [6].

Image segmentation can be very challenging [6]–[14], particularly medical image segmentation [15], [16], [25]–[28], [17]–[24]. Recently, deep learning methods have solved complicated segmentation tasks [29]–[34], in medical imaging [35]–[42], and US [43]–[46], but not in skeletal muscle US, though there are some applications to muscle US [47]–[49]. The lack of publicly available labelled data, benchmark methods and results hinder the development of this domain. A methodology for obtaining participant specific labelled data, suitable for patients with involuntary movement disorders should not require subjects to be still. This study uses direct manual annotation of US images to produce ground truth labels and tests whether the combination of US image quality, direct manual annotation of US images, and deep learning is sufficient to train a system for robust, accurate segmentation of cervical muscles from transverse US images of the human neck, during functional movement. We contribute a labelled dataset of 1,100 US images of the human posterior cervical muscles, with a description of the process used to create the labels (Fig. 1). We present reproducible, deep learning methods and a state-of-the-art benchmark for full resolution segmentation of neck muscles during head movement.

A. Background

Previously, cross-domain magnetic resonance imaging (MRI)-to-US registration was used to obtain accurate ground-truth labels of cervical US images [6]. MRI images provide higher quality tissue contrast enabling more accurate annotation of muscle boundaries which can subsequently be registered to linked US images. However, use of MRI for annotation limits application to static head, and to participants who can keep their head still during MRI acquisition (typically > 5 minutes). Participants with pain, and movement disorders such as dystonia, which cause unavoidable body movement and contractions in muscles of interest are excluded. Though more difficult, direct annotation of US images for ground truth allows a system to be trained for functional movement and for participants with involuntary movement disorders. US is also more available than MRI.

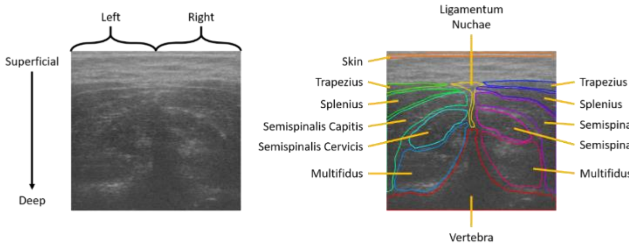

Figure 1. Axial cervical neck US image and segment categories. Left: Typical transverse cervical "C4-6", US image, depth 5 cm, width 5.9 cm. Right: Manually annotated boundaries of 13 segments added. Image left indicates anatomical left.

Research supported by Manchester Metropolitan University (Dept. of Healthcare Science and Dalton Research Institute) and The Dystonia Society.

US images highlight reflecting tissue boundaries and provide texture indicating tissue content (Fig. 1). All muscles have similar composition-texture and appear dark (hypoechoic). Collagenous tissue between and within muscles appears bright (echogenic). In US, the boundary between muscles is not always visible. Assuming a constant probe position, as the head moves, the shape of muscle boundaries changes and sometimes muscles are no longer present within the image. Further, the quality of the US image deteriorates with depth (Videos 1, 2, Supp. Material).

Understanding the boundaries of muscles in the US image requires training. For training purposes, two annotators directly labelled 86 US images from 86 participants. For 77 of these participants we collected linked MRI images from which annotators labelled MRI images, then registered MRI muscle boundaries to the US images, and then annotated US images directly using registered MRI and a 3D atlas [50] as a reference [6] (Fig A. Supp. Material). Once annotation was consistent and agreed between labellers, we proceeded to direct labelling of the current dataset. With this article, we publish a fully labelled dataset of 1,100 cervical muscle US images, from 28 healthy participants during functional head movement. For 20 of these participants, the annotators had MRI annotations, registered to linked US images of the participant in the MRI posture, available for reference.

To provide a deep learning method and state of the art benchmark to automate online labelling of US images, we use a popular neural network architecture for semantic segmentation applications, namely the deconvolutional neural network (DCNN), (Fig. 2) [30], [51], [52]. Here the term deconvolution does not describe the linear algebra being used, but describes the aim and function of a DCNN to reverse the effects of convolution and reconstruct full-resolution data/labels. The DCNN shows invariance to local transformations (through max-pooling and recovery of pooled vectors using un-pooling) and implements an up-sampling route immediately after the convolutional and down-sampling part of the network.

Within the DCNN architecture, we investigate two different transfer functions. We investigate the recently introduced Exponential Linear Unit (ELU), because it alleviates the ‘dying ReLU’ problem (where unit function derivatives are zero over the training set), is linear in the positive part, has smooth nonlinearities in the negative part, and has demonstrated computational efficiency and superior performance over the popular alternative batch normalization method [53]. Second, we investigate the MaxOut unit, since it too alleviates ‘dying ReLU’, is theoretically capable of approximating any transfer function, and it has unit derivatives almost everywhere. We predict these units are naturally applicable to DCNNs since the popular dropout regulariser is more suited at test time to the linear response of each MaxOut subunit. With respect to deconvolutions, we predict MaxOut units have more flexibility than ELU because they can encode two different positions (with respect to units with only 2 components) with a single MaxOut unit, therefore theoretically increasing the model capacity within the decoder architecture.

II. Methods

A. Data Collection

Using a probe attached to the posterior neck, US images were recorded from 28 healthy adults (19 male, mean 31.4, s.d. 8.9) during head movement tasks (Fig. B Supp. Material). The tasks defined a range of pitch and yaw head rotations from which to sample axial images of the cervical muscles. These experiments, performed in the Cognitive Motor Function laboratory, Dept. of Healthcare Science, were approved by the Research Ethics Committee of the Faculty of Science and Engineering, Manchester Metropolitan University. Participants gave written, informed consent to these experiments, which conformed to the standards set by the latest revision of the Declaration of Helsinki.

Participants were strapped at the chest to a vertical support. A projector displayed a moving target on a screen 1 *m* in front of the participants. Participants were instructed to allow their head to turn within a comfortable range to follow the target with the tip of their nose. “Horizontal”, “vertical” and “combined” trials contained respectively horizontal motion of the target, vertical motion of the target and a combination of independent horizontal and vertical components similar to [49]. These independent components included sines of multiple frequencies leading to transiently correlated and un-correlated intervals. A Vicon MX motion capture system, with 10 infrared cameras recorded the 3D spatial location of 9 retroreflective markers, which were placed on the head (bilaterally on the zygomatic arch and inferior orbit, unilaterally on frontal bone) and thorax (manubrium, right clavicle, right and left acromion). A T-shaped US probe (7.5

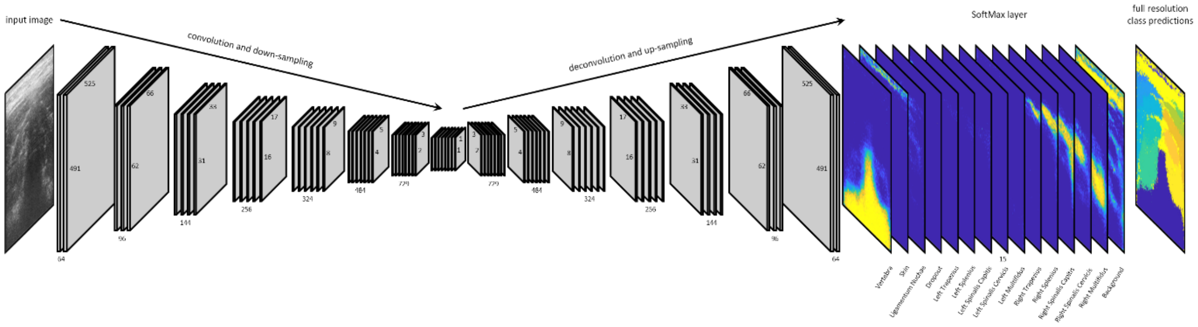

Figure 2. Neural network architecture. A DCNN is shown with a spatial SoftMax classification layer. On the far left, the raw US image is input to the encoder part of the DCNN, then a series of convolution filters are applied (first layer uses a stride of 2 x 2), immediately followed by 4 x 4 spatial MaxPooling. This pattern of convolution and pooling (2 x 2) is repeated a further 6 times, increasing the number of convolution filters with each repetition, and with the last pooling layer being 2 x 3 MaxPooling. This results in the entire image being encoded by a 1x1 (spatial) feature representation consisting of 729 real valued nonlinear units/neurons. The decoder part of the DCNN then unpacks the representation, layer by layer (mirroring the encoder), using the indices of each pooled unit response (MaxUnPooling) to reconstruct the position-accurate full spatial resolution of the feature representation with an appended depth of 15 (the number of segment categories). The SoftMax function is then executed in a depth-wise manner separately for each pixel, over the appended 15 units. Finally, on the far right we take the index of the maximum over the appended 15 units, for each pixel, to reveal the predicted pixel classifications. Our ELU nets had 14,297,174 nodes, and 11,739,880 free parameters (weights), and our MaxOut nets had 23,872,834 nodes, and 23,479,760 free parameters (weights).

MHz, Aloka) was taped (Transpore medical tape) to the posterior neck at level C4-6, to allow free movement of the head and an image of 5 bilateral layers of muscles. Images were acquired at 25 Hz using a frame grabber card (Data Translation DT 3120) synchronised to the Vicon recording.

B. Ultrasound Image Ground Truth

Pitch and yaw rotations of the head relative to the trunk were computed from Vicon motion data using Visual 3D (C-Motion). Forty US images were selected per participant to sample uniformly their range of pitch and yaw rotations giving 1,100 images. Two annotators, trained to recognise the cervical muscle boundaries (Fig. 1), used a custom Matlab graphical user interface to annotate approximately 550 images each. The annotation procedure was to select a segment (trapezius, splenius, …, etc.) using the keyboard and mark the muscle side of the boundary of each segment, medial to lateral, in a clockwise manner, closing each segment. To aid visualisation of texture, annotators were able to adjust the local contrast by normalising on patches between (10 and 50 pixels). Local contrast normalization (LCN) was applied via the keyboard. Where segment boundaries were ambiguous or invisible, annotators used the general pattern informed by anatomical atlas [50] to complete each segment to the best estimate. For a given participant, if available the annotated MRI and partially registered MRI was presented for reference. Following the first annotation, each image was presented with the previous annotation for reference. Annotators usually cleared the existing annotation and annotated each segment boundary from scratch but had the option to update existing boundary points.

C. Equipment and Software

All software was developed entirely within the group at Manchester Metropolitan University, and all neural network code/software was written from scratch solely by Ryan Cunningham using C/C++ and CUDA-C (NVidia Corporation, Santa Clara, California). No libraries other than the standard C++ 11 library and standard CUDA libraries (runtime version 8.0 cuda.h, cuda_runtime.h, curand.h, curand_kernel.h, cuda_occupancy.h, and device_functions.h) were used. All 84 neural networks were trained on an Intel Xeon CPU E5-2697 v3 (2.60GHz), 64GB (2133 MHz), with two NVidia GTX 1080 Ti GPUs.

D. Network Architecture

Multiple network architectures were trained to investigate the effect of transfer functions on the deconvolutional (up-sampling) part of the network. First, we applied the ELU) throughout the network (encoder and decoder parts). Second, we applied the ELU only in the encoder part or the network, where a linear transfer function was used in the decoder part. Finally, we applied MaxOut units throughout the network, where the transpose of the MaxOut function is used in the decoder part (argmax of corresponding units in the encoder part), therefore implements naturally a linear deconvolutional network with respect to the derivatives of the units.

In the final deconvolutional (output) layer, we use the same spatial properties (stride, input/output dimensions) and number of units (feature channels) as the first convolutional (input) layer, but we modify the local field of each unit to have an extra channel of length n (where n = number of pixel categories). The SoftMax activation function is applied over this channel, separately, for every pixel in the output map, after all features have been deconvolved. This yields, for every pixel, n real (single precision) values between 0 and 1, which sum to 1. The maximum of these n values over a pixel reveals the most likely classification of that pixel with respect to the neural network. The second largest value yields the second most likely classification, and so on.

All networks were initialized according to the following scheme, based on the literature and our experience of training neural networks. Linear, SoftMax and MaxOut unit weights were drawn from a real (single precision) uniform distribution in the range $\left[-\sqrt{\frac{3}{fan\ in}} \quad +\sqrt{\frac{3}{fan\ in}}\right]$. ELU weights were drawn from a real (single precision) uniform distribution in the range $\left[-\sqrt{\frac{6}{fan\ in}} \quad +\sqrt{\frac{6}{fan\ in}}\right]$. The *fan in* is the total number of the local (spatial) and feature inputs to any given unit in a layer. For deconvolutional layers, the *fan in* of a unit represents the total number of local (spatial) outputs (transposed inputs) to that unit.

E. Network Optimization

Adaptive moment estimation [54] (ADAM) was used with default $\beta_1 = 0.9$ and $\beta_2 = 0.999$ parameters, but with a smaller ($\alpha = 0.00002$) than suggested $\alpha = 0.002$ parameter (learning rate) to account for non-batch (batch size of 1) learning. A small L^2-norm weight penalty was used (5e-4) as recommended to aid convergence (rather than to regularize). All parameters were empirically selected using a subset of the data, to check for quick (with respect to weight updates) and stable (no 'exploding gradients') convergence.

F. Data Augmentation

Each US image and corresponding label was reflected about the vertical line of symmetry, artificially doubling the size of the data set. The order of images was randomised before each pass of the entire training set. Each learning iteration includes a forward pass, an error calculation, a backward pass, and a weight/parameter update. Before each learning iteration the input image was normalized to unit variance and zero mean and subjected to a random transformation (horizontal and vertical translation, followed by a rotation). Translation and rotation parameters were sampled from a random real (single precision) uniform distribution in the ranges of [-25 +25] pixels and [-20 +20] degrees, respectively. Image and label pixels were resampled during transformation using bilinear, and nearest neighbour interpolation, respectively.

G. Network Training and Validation

Cross validation was executed with 14 folds. For each set of network parameters and properties (architecture/units) 14 identical networks were trained separately using 26 of the 28 participants' images and labels; the remaining 2 participants' images and labels were used to test each network. For each of the 14 networks, neither of the test participants' images/labels were used to test any of the other 13 networks. This process yielded genuine held-out test results for all 28 participants.

The cross-entropy cost function was used to minimize the training error between labels and network prediction. Training for each fold consisted of online learning, interrupted every quarter pass (550 learning iterations) through the training set, to record cross entropy test results from the 2 test sets. If the cross-entropy loss for either test set was lower than any previous recorded loss for that test set, the

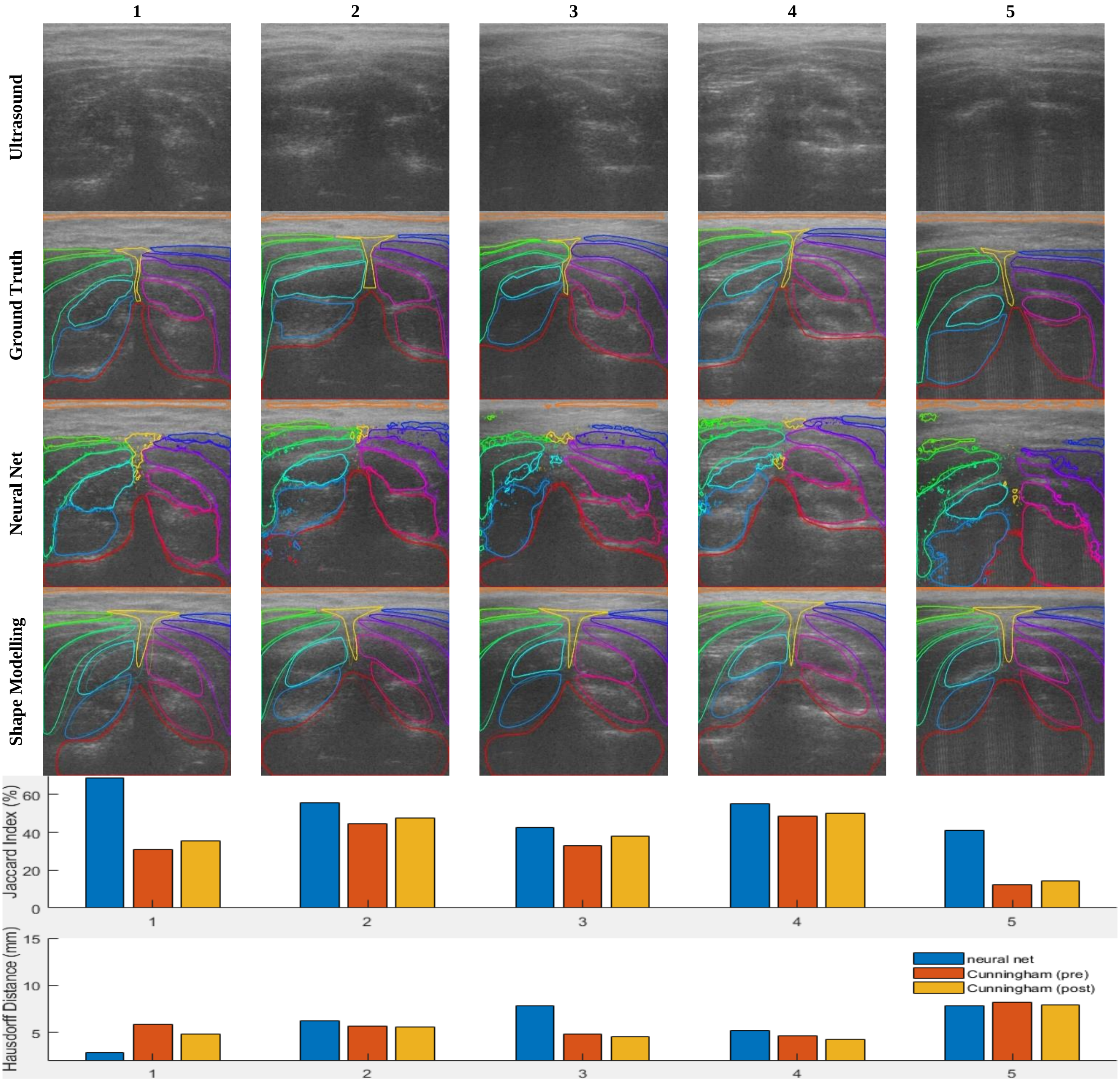

Figure 3. Representative segmentation maps.

Col 1: Good neural net segmentation, typical shape model segmentation.

Col 2: Typical neural net segmentation, good shape model segmentation.

Col 3: Challenging US image; both the human labeller (ground truth) and the neural net have moderate difficulty in correctly classifying regions (left side of the neck/image).

Col 4: Challenging image; the neural net is mostly able to identify the correct regions, but the shape model [6] is not robust to the assymetry.

Col 5: A very challenging US image, where most of the deep features are invisible, and the human labeller (ground truth) is able to identify the segments, but none of the automatic methods do very well, although the neural net does a better job than the shape modelling method.

Representative results: Neural net method (Jaccard/Hausdorff in the range: 41-69(%)/2.8-7.8(mm); Shape Model method 14-50(%)/4.2-7.9(mm).

network was saved to long term storage. When neither test set recorded a lower loss for 25 consecutive test iterations, training was terminated. Following training, the network associated with the lowest loss for test set 1 was loaded to acquire results for test set 2, and *vice versa*, yielding true held out optimal results for both test sets. The same configuration of test/train sets was used for all variations of network parameters and architecture.

III. Results

A. *Comparison of DCNN Architectures*

Three neural network architectures were trained to compare performance with different transfer functions. Performance was measured by Jaccard index (JI) on the complete set of labels (A), and on a reduced set of labels (B, no skin-ligament) to enable comparison with previous literature (Table 1). All networks were trained both with and without dropout using full 14-fold cross-validation, and results were computed over all 28 held out participants. Results were also calculated for top 2 and top 3 SoftMax outputs. Observation 1: results for the reduced class set (B) were notably more accurate than for the full class set (A) in all cases (set B: 49.6%-53.2%, set A: 45.1%-48.7%). Observation 2: adding dropout improved results for both sets (A, B), and for all networks (45.1%-52.2% without dropout; 46.47%-53.18% with dropout). Observation 3: performance

Table 1. Comparison of proposed neural network methods.
Shows mean Jaccard Index (x 10^2), JI, over all segments, for the 3 neural network methods.
Class set A includes skin, muscles, vertebra and Ligamentum Nuchae.
Class set B includes only the muscles and vertebra.
Top 1-3 results are presented: top 1 is the result if we take the max class over the SoftMax units; top 2 is the result if we can take either the max or the second max (whichever is the correct class) over the SoftMax units; and so on.

Method	Class Set	JI ± σ (%)		
		Top 1	Top 2	Top 3
MaxOut + dropout 0.5	B	**53.2 ± 14.2**	75.8 ± 13.3	88.0 ± 11.7
MaxOut		52.2 ± 14.4	75.0 ± 13.5	87.5 ± 12.0
ELU conv, linear deconv + dropout 0.5		51.2 ± 12.4	**76.2 ± 12.6**	**88.7 ± 10.8**
ELU conv, linear deconv		50.5 ± 14.1	73.7 ± 14.6	86.6 ± 14.1
ELU conv/deconv + dropout 0.5		51.2 ± 12.4	76.2 ± 12.6	88.7 ± 10.8
ELU conv/deconv		49.6 ± 13.4	74.5 ± 13.1	87.3 ± 10.7
MaxOut + dropout 0.5	A	**48.7 ± 12.6**	74.7 ± 11.8	86.7 ± 10.8
MaxOut		47.8 ± 12.6	74.1 ± 11.8	86.3 ± 10.8
ELU conv, linear deconv + dropout 0.5		46.5 ± 10.9	75.3 ± 11.2	87.3 ± 10.0
ELU conv, linear deconv		45.9 ± 12.6	72.5 ± 13.5	85.1 ± 13.2
ELU conv/deconv + dropout 0.5		46.5 ± 10.9	**75.3 ± 11.2**	**87.3 ± 9.9**
ELU conv/deconv		45.1 ± 12.3	73.4 ± 12.2	85.6 ± 10.5

improved with more linearity used in the network: MaxOut networks were best at 53.18%, followed by ELU convolutions with linear deconvolution at 51.21%, followed finally by ELU convolutions with ELU deconvolutions at 51.21%. Including top 2 and top 3 SoftMax outputs revealed the same general patterns, and performance increased from approximately 50% JI to 75% JI for top 2, and approximately 88% JI for top 3.

B. Comparison with the Benchmark

These DCNNs are compared to the benchmark shape model method [6] and to agreement between annotators (Ground-Truth) from their training set of US images in the static (MRI) head position.

Observation 4: For all segments grouped (Table 2), the DCNN methods perform best (JI 53 ± 14%, Hausdorff Distance (HD) 5.7 ± 2 mm) and lie close to the reference defined by ground truth (JI 58 ± 13%, HD 4.4 ± 2.1 mm). The benchmark initial segmentation method (pre) performs worst at 37.5% JI and 6.6 mm HD, then the optimised shape fitting (post) provides an improvement at 40.5% JI and 6.2 mm HD. Performance is better on the reduced class set B v A.

Observation 5: Using JI, the DCNN method performs better on the vertebra and muscles segments than the skin and ligamentum nuchae, and performs better than the benchmark method excepting only the ligament (Table 3). Using HD, the DCNN method out performs the benchmark shape model on the vertebra and all muscles except trapezius (Table 3).

Table 2. Comparison of proposed neural network with the benchmark. Shows Jaccard Index (x 10^2), JI, and Hausdorff Distance, HD, mean ± S.D. for all segments.
Neural Net: DCNN method.
Benchmark: Shape model method [6]: (pre) shows results of an initial segmentation (zero fitting iterations); (post) represents the results of a refined segmentation (25 fitting iterations).
Static labels: Agreement between labelers for ground truth training images (class set A). N.B. subset of 25 participants collected using same US machine (Aloka).

Method	Class Set	JI ± σ (%)	HD ± σ (mm)
Neural Net	B	**53.1 ± 14**	**5.74 ± 1.9**
Benchmark (post)		40.5 ± 13	6.29 ± 2.06
Benchmark. (pre)		37.5 ± 14	6.66 ± 2.24
Neural Net	A	**48.7 ± 12.6**	**5.84 ± 1.8**
Benchmark (post)		36.4 ± 11.3	5.96 ± 2.0
Benchmark (pre)		34.2 ± 12.2	6.21 ± 2.1
Static labels	-	58 ± 13	4.4 ± 2.1

Visual inspection of the resulting semantic maps (Fig. 3, Figs. C, D Supp. Material) reveals the inflexibility of the benchmark method, and the robustness of the DCNN method, which can segment very challenging held-out test images.

IV. Discussion

A. Purpose of study

The purpose of this study was to test whether the quality of US image obtained during head movement, the direct manual annotation of US images for ground truth, and deep machine learning is sufficient to meet our primary objective: to provide robust, accurate segmentation of cervical muscles from transverse US images of the human neck during functional head movement.

The cervical muscles traverse a fundamental mechanical and sensory node within the human neuromuscular system [5]. Pain and dysfunction within this muscle group has profound consequences on quality of life [55], [56]. Personalised analysis and diagnosis of cervical muscles is under-developed, and so there is benefit to developing automatic segmentation, particularly for use during functional movement. Ultrasound is readily available in clinics and laboratories. Online segmentation of the cervical muscles allows visualisation and analysis of their dynamic structure and function and allows attribution of measurement and targeting of treatment to individual muscles. Muscle-specific measurement and targeted injection is valuable, specifically for diagnosis and treatment of involuntary muscles contractions in cervical dystonia [1], [6].

Segmentation of cervical muscles within US images is challenging. Even, using a defined static head position, there is a limit to the accuracy with which muscle boundaries can be defined from US. In this study, using a high-quality 3D atlas of the head and neck muscles, and using participant specific MRI defined boundaries, trained human labellers agreed muscle boundaries to a JI of 58 ± 13% and a HD of 4.4 ± 2.1 mm. Head movement beyond the MRI-defined head position increases variability in the appearance of muscles and their boundaries. These results demonstrate that DCNNs provide segmentation of muscles in moving heads to an accuracy comparable with the agreement achieved by trained humans in less challenging static images (Tables 2, 3).

For neural networks to generalise successfully, the data must contain features which exist consistently across training and testing cases and those features must associate consistently with ground-truth labels. The accuracy of muscle-vertebra segmentation (JI 53 ± 14%, HD 5.7 ± 1.9 mm) validates the combination of US image quality, inherent information content, ground-truth labelling and deep learning to achieve our objective: to provide robust, accurate segmentation of cervical muscles during functional head movement.

B. Relationship of contribution to current literature

This study introduces deep learning to the problem of automatic labelling of US images of skeletal muscle, and provides a public data set to stimulate continued development of this field. This application to a challenging modality [15], requires accurate, online segmentation of 10 muscles, the vertebra, skin and ligamentum nuchae in full resolution

Table 3. Group results for individual segments. Shows Jaccard Index (JI), and Hausdorff Distance (HD), mean ± S.D. for individual segments.
Neural Net: DCNN method.
Benchmark: Shape model method [6]: (post as above).
Static labels: Ground-truth training images as above: agreement between labelers.

Deep → Superficial

Method	Metric	Vertebra	Multifidus	Spinalis Cervicis	Spinalis Capitis	Splenius	Trapezius	Ligamentum Nuchae	Skin
Neural Net	JI	**78.1 ± 14**	**55.4 ± 14**	**49.5 ± 20**	**53.3 ± 17**	**50.6 ± 21**	**44.7 ± 25**	14.2 ± 10	**33.5 ± 15**
Benchmark	(%)	69.9 ± 15	34.9 ± 17	33.15 ± 17	44.84 ± 15	40.99 ± 19	33.86 ± 23	**19.75 ± 11**	7.37 ± 17
Static labels		81 ± 6	62 ± 9	59 ± 8	60 ± 13	58 ± 14	54 ± 20	37 ± 17	52 ± 18
Neural Net	HD	**6.29 ± 4.57**	6.53 ± 3.52	**6.02 ± 3.89**	**5.98 ± 4.97**	**5.65 ± 3.32**	4.23 ± 3.33	7.36 ± 4.27	5.51 ± 5.38
Benchmark	(mm)	7.44 ± 3.31	7.59 ± 2.95	6.61 ± 3.42	6.59 ± 3.19	6.02 ± 4.44	**4.06 ± 2.40**	**5.74 ± 2.58**	**2.62 ± 3.16**
Static labels		5.0 ± 1.5	4.8 ± 1.4	5.2 ± 3.0	6.6 ± 3.8	5.5 ± 3.6	2.3 ± 1.4	4.5 ± 1.9	1.0 ± 0.4

(525x491) b-mode US images of the posterior neck. Given the lack of public data, there is a lack of literature in this domain. Our previous method [6] requires MRI for the generation of an accurate US texture-shape dictionary covering the intended range of head positions and participant conditions. This requirement for MRI proved fatal to our current need for a method (i) applicable to participants with cervical dystonia who cannot remain still within an MRI machine, and (ii) to cover a range of head positions not available from MRI scanning (iii) requiring large volumes of labelled data. The application of deep learning, combined with direct annotation of US images within the required range or head positions, overcomes these limitations and provides additional benefits. Deep neural networks are robust to dropout of large regions of the image, can predict missing segments and can communicate high-resolution regional confidence.

Deconvolutional neural network (DCNN) architectures are appropriate for our application. Within our investigation of different DCNN architectures, we tested different activation transfer functions. ELUs have smooth derivatives near the mean activation and are linear for positive input. ELU units address the 'dying ReLU' problem more efficiently [57] than other popular methods like batch (re)normalisation [53], [58]. We explored MaxOut units with 2 linear components because, without introducing further nonlinearity, they can up-sample more than 1 local texture map per spatial location. Results confirmed MaxOut units gave superior performance, JI > 53%, compared with the best ELU architecture, 51.2% (Table 1). Use of ELUs in the deconvolutional part of the network reduced performance and we observed consistent performance, between ELU/ELU and ELU/Linear, both with and without dropout (Table 1). For DCNNs, MaxOut units increase model complexity, and in a way that is highly regularized. The convolutional (encoder) part of the network, is a piecewise linear function which behaves like any other nonlinear network transfer function. The hidden complexity lies in the storage of multiple reconstruction pathways

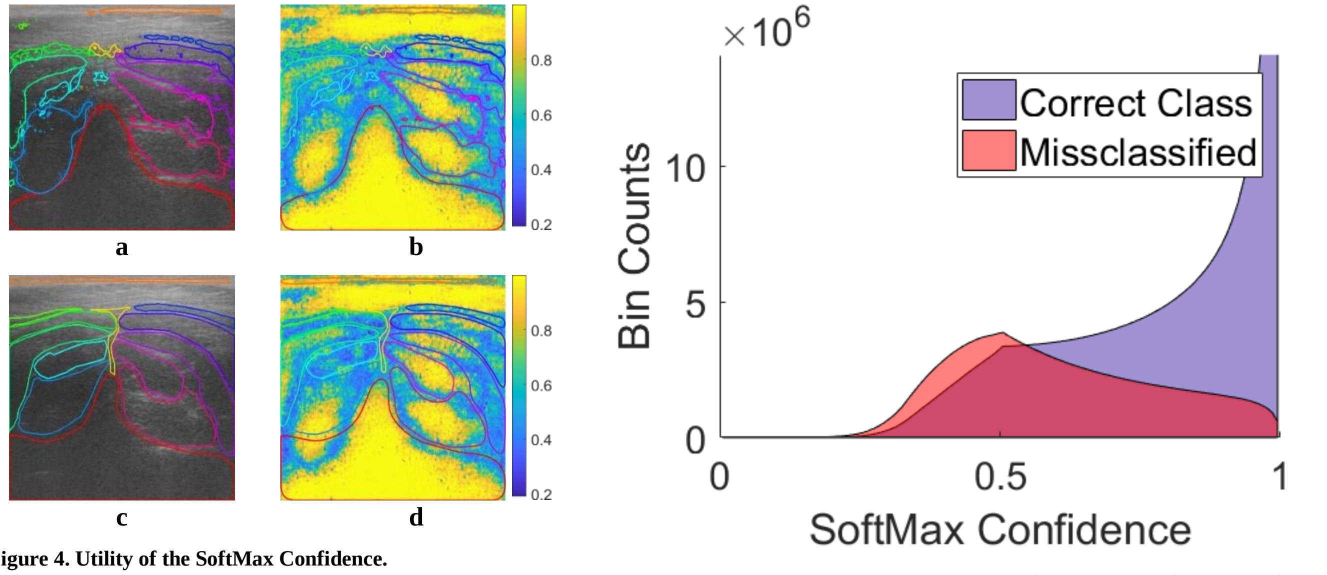

Figure 4. Utility of the SoftMax Confidence.
Left: **(a)-(d)** refer to the same US image and segmentation seen in Fig. 3 column 3. **(a)** neural net segmentation. **(b)** max for each pixel, of the raw SoftMax output layer of the neural net, as a heat map with the neural net segmentation superimposed. **(c)** ground truth segmentation. **(d)** same heat map as in (b), with the ground truth segmentation superimposed. On the left hand side of the US image, the neural net has some difficulty in identifying the regions belonging to the muscles, and in fact completely misses out the left trapezius muscle. Observation of the SoftMax output (b & c) (which we can think of as the probability/confidence of the correct classification of each pixel) reveals that where the network is incorrect on the left (Spinalis Cervicis, splenius, and trapezius), it has relatively low confidence, and where it is correct, the network has relatively high confidence. On the right hand side of the US image, the network is mostly correct: where it is not correct (e.g. far right side of multifidus, or lower-right side of Spinalis Capitis) the confidence is relatively low. Confidence is a useful diagnostic value to assess whether to accept the segmentation given by the neural net, or to optimize the image by positioning the probe, or changing image settings on the US machine.
Right: shows, for the entire held-out dataset (2,200 images), for all pixels (2,200x525x491 = 567,105,000 pixels) the max of the raw output of the SoftMax layer of the neural net, binned into correct or incorrect class histograms. The two distributions are clear; for misclassified the majority of the pixels are associated with relatively lower SoftMax confidence, whereas for correct class the majority of the pixels are associated with relatively higher SoftMax confidence.

exploited by the deconvolutional part, yet learning is fully controlled by the convolutional part.

The benchmark shape modelling method [6] enforced patterns of segments. The shape model ensures smooth segment boundaries and no segment holes. The neural network can predict no visible class (class = *background*) and does not enforce explicit shapes. The higher dimensionality of the neural network prediction is useful for representing muscle specific changes from active contraction and from head rotation. However, segment holes and non-smooth boundaries are a weakness for thin segments (e.g. skin, trapezius). For example, the neural network could predict holes in skin across the entire segment (Fig. 3, row 3, col 4). In contrast, the shape model produces a smooth shape but any small vertical position error causes minimal overlap with ground truth. For skin and trapezius segments, the shape model gave better HD and worse JI values (Table 3).

The neural network produces a 'segmentation confidence map' (Fig. 4) for any given segmentation. This intrinsic feature of the SoftMax layer gives a measure of confidence between 0 and 1 for every pixel. A value close to 1 means the network is confident in its prediction for that pixel. A value close to 0 means the neural network is confident that pixel could be one of multiple classes. Figure 4 shows that higher confidence is associated with correct classification. This confidence provides online, operator feedback of the quality of an image segmentation. For example, if a clinician needs to target a muscle for injection (standard treatment for cervical dystonia), the clinician would optimise the placement of the probe, and the US machine settings to increase the confidence of the neural network in the target muscle. The clinician would then target an injection point to the centre of a contiguous cluster of high confidence, and appropriately classified pixels (equal to target muscle) within the target muscle.

As described in the methods, the 1,100 images and their labels were sampled from a larger data set of full-articulated head motion, containing hundreds of thousands of images without associated labels. Our segmentation neural network will be useful for automatically generating labels for these images for analysis of muscle behaviour and also for pre-training neural-networks.

V. Conclusions

Online segmentation of US images of cervical muscles during functional head movement is required for visualisation, analysis of deep muscle structure and function, and monitoring of patient specific treatment protocols for neck/back pain/injury, work related disorders, myopathies, neuropathies and movement disorders. The lack of publicly available labelled data, benchmark methods and results, hinders development of this domain. This study contributes a labelled dataset of 1,100 US images of the human posterior cervical muscles, with a description of the process we followed to create the labels. For the first time, we have applied deep learning methods, DCNNs, to this application and we demonstrate superior performance and flexibility over the current published technique (shape modelling and heuristic contour fitting). We have shown that MaxOut networks outperform ELU networks in the deconvolutional architecture, which will influence future research on this data. This proof of principal shows that a robust, accurate, bespoke muscle US segmentation system can be constructed with deep learning, and with a little as 1,100 annotated images. This data set stimulates further development in this domain. This methodology for labelling ground-truth and training automated labelling networks is applicable for dynamic segmentation of moving muscles and for participants with involuntary movement disorders who cannot be still. This contribution is relevant in healthcare for conditions such as cervical dystonia, and with future development, will likely lead to clinical software systems for guiding and monitoring of treatment, and training of doctors, nurses and radiographers.

References

[1] A. Siddique, R. Cunningham, M. Silverdale, P. Harding, I. Loram, and C. Kobylecki, "Segmentation of neck muscles using ultrasound in cervical dystonia," *Mov. Disord.*, vol. 33, no. suppl 2, p. S322:S322, 2018.

[2] G. Rankin, M. Stokes, and D. J. Newham, "Size and shape of the posterior neck muscles measured by ultrasound imaging: Normal values in males and females of different ages," *Man. Ther.*, vol. 10, no. 2, pp. 108–115, 2005.

[3] A. V. Dieterich *et al.*, "Shear wave elastography reveals different degrees of passive and active stiffness of the neck extensor muscles," *Eur. J. Appl. Physiol.*, vol. 117, no. 1, pp. 171–178, 2016.

[4] M. Stokes, J. Hides, J. Elliott, K. Kiesel, and P. Hodges, "Rehabilitative ultrasound imaging of the posterior paraspinal muscles," *J. Orthop. Sports Phys. Ther.*, vol. 37, no. 10, pp. 581–595, 2007.

[5] I. D. Loram, B. Bate, P. Harding, R. Cunningham, and A. Loram, "Proactive Selective Inhibition Targeted at the Neck Muscles: This Proximal Constraint Facilitates Learning and Regulates Global Control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 4320, no. 4, pp. 357–369, 2017.

[6] R. J. Cunningham, P. J. Harding, and I. D. Loram, "Real-Time Ultrasound Segmentation, Analysis and Visualisation of Deep Cervical Muscle Structure," *IEEE Trans. Med. Imaging*, vol. 36, no. 2, 2017.

[7] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, 2013.

[8] D. Grangier, L. Bottou, and R. Collobert, "Deep Convolutional Networks for Scene Parsing," *ICML 2009 Deep Learn. Work.*, vol. 3, no. i, 2009.

[9] S. Kamijo and M. Sakauchi, "Segmentation of vehicles and pedestrians in traffic scene by spatio-temporal Markov random field model," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2, pp. II285-II288, 2003.

[10] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, 2017.

[11] B. Peng, L. Zhang, X. Mou, and M. H. Yang, "Evaluation of Segmentation Quality via Adaptive Composition of Reference Segmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 1929–1941, 2017.

[12] Y. Zhang, X. Chen, J. Li, C. Wang, C. Xia, and J. Li, "Semantic Object Segmentation in Tagged Videos via Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.

40, no. 7, pp. 1741–1754, 2018.

[13] L. Najman, Y. Xu, E. Carlinet, and G. Thierry, "Tree-Based Shape Spaces," vol. 39, no. 3, pp. 457–469, 2017.

[14] J. Pont-Tuset and F. Marques, "Supervised Evaluation of Image Segmentation and Object Proposal Techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1465–1478, 2016.

[15] J. a. Noble and D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Trans. Med. Imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[16] W. Qiu *et al.*, "Automatic segmentation approach to extracting neonatal cerebral ventricles from 3D ultrasound images," *Med. Image Anal.*, vol. 35, pp. 181–191, 2017.

[17] P. Yan, S. Xu, B. Turkbey, and J. Kruecker, "Adaptively learning local shape statistics for prostate segmentation in ultrasound," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3 PART 1, pp. 633–641, 2011.

[18] F. Destrempes, J. Meunier, M. F. Giroux, G. Soulez, and G. Cloutier, "Segmentation in ultrasonic B-mode images of healthy carotid arteries using mixtures of Nakagami distributions and stochastic optimization," *IEEE Trans. Med. Imaging*, vol. 28, no. 2, pp. 215–229, 2009.

[19] E. Kozegar, M. Soryani, H. Behnam, M. Salamati, and T. Tan, "Mass Segmentation in Automated 3-D Breast Ultrasound Using Adaptive Region Growing and Supervised Edge-Based Deformable Model," *IEEE Trans. Med. Imaging*, vol. 37, no. 4, pp. 918–928, 2018.

[20] C. Garnier *et al.*, "Prostate segmentation in HIFU therapy," *IEEE Trans. Med. Imaging*, vol. 30, no. 3, pp. 792–803, 2011.

[21] Y. Li, C. P. Ho, M. Toulemonde, N. Chahal, R. Senior, and M. X. Tang, "Fully automatic myocardial segmentation of contrast echocardiography sequence using random forests guided by shape model," *IEEE Trans. Med. Imaging*, vol. 37, no. 5, pp. 1081–1091, 2018.

[22] G. Unal, S. Bucher, S. Carlier, G. Slabaugh, T. Fang, and K. Tanaka, "Shape-driven segmentation of the arterial wall in intravascular ultrasound images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 3, pp. 335–347, 2008.

[23] G. Q. Zhou and Y. P. Zheng, "Automatic Fascicle Length Estimation on Muscle Ultrasound Images With an Orientation-Sensitive Segmentation," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 12, pp. 2828–2836, 2015.

[24] X. Zang, R. Bascom, C. Gilbert, J. Toth, and W. Higgins, "Methods for 2D and 3D Endobronchial Ultrasound Image Segmentation.," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1426–1439, 2015.

[25] D. D. B. Carvalho *et al.*, "Lumen Segmentation and Motion Estimation in B-Mode and Contrast-Enhanced Ultrasound Images of the Carotid Artery in Patients With Atherosclerotic Plaque," vol. 34, no. 4, pp. 983–993, 2015.

[26] S. Sun, M. Sonka, and R. R. Beichel, "Graph-based IVUS segmentation with efficient computer-aided refinement," *IEEE Trans. Med. Imaging*, vol. 32, no. 8, pp. 1536–1549, 2013.

[27] A. Faisal, S. C. Ng, S. L. Goh, J. George, E. Supriyanto, and K. W. Lai, "Multiple LREK Active Contours for Knee Meniscus Ultrasound Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 2162–2171, 2015.

[28] N. Almeida *et al.*, "Left-Atrial Segmentation from 3-D Ultrasound Using B-Spline Explicit Active Surfaces with Scale Uncoupling," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 63, no. 2, pp. 212–221, 2016.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv Prepr. arXiv1511.00561*, vol. 39, no. 12, pp. 2481–2495, 2015.

[30] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, vol. 11–18–Dece, pp. 1520–1528.

[31] A. Romero, M. Drozdzal, A. Erraqabi, S. Jégou, and Y. Bengio, "Image Segmentation by Iterative Inference from Conditional Score Estimation," pp. 1–11, 2017.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," vol. 40, no. 4, pp. 834–848, 2014.

[33] G. Chen, S. Member, X. Zhang, Q. Wang, and F. Dai, "Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images," vol. 11, no. 5, pp. 1–12, 2018.

[34] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[35] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, no. December 2012, pp. 60–88, 2017.

[36] A. A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Buhler, "Fully Convolutional Architectures for Multi-Class Segmentation in Chest Radiographs," *IEEE Trans. Med. Imaging*, vol. 37, no. 8, pp. 1865–1876, 2018.

[37] Y. Li, L. Shen, and S. Yu, "HEp-2 Specimen Image Segmentation and Classification Using Very Deep Fully Convolutional Network," *IEEE Trans. Med. Imaging*, vol. 36, no. 7, pp. 1561–1572, 2017.

[38] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic Image Segmentation via Multistage Fully Convolutional Networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, 2017.

[39] Y. Yuan, M. Chao, and Y. Lo, "Fully Convolutional Networks with Jaccard Distance," vol. 0062, no. c, pp. 1–11, 2017.

[40] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Auto-context Convolutional Neural Network for Geometry-Independent Brain Extraction in Magnetic Resonance Imaging," *IEEE Trans. Med. Imaging*, vol. 36, no. 11, pp. 2319–2330, 2017.

[41] Y. Han and J. C. Ye, "Framing U-Net via Deep Convolutional Framelets: Application to Sparse-View CT," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1418–1429, 2018.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Miccai*, pp. 234–241, 2015.

[43] L. Wu, X. Yang, S. Li, T. Wang, P. Heng, and D. Ni, "Cascaded Fully Convolutional Networks for Automatic Prenatal Ultrasound Image Segmentation," *Biomed. Imaging*, pp. 663–666, 2017.

[44] N. Wang, Y. Wang, H. Wang, B. Lei, T. Wang, and D. Ni, "Auto-context fully convolutional network for levator

hiatus segmentation in ultrasoudn images," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2018–April, no. Isbi, pp. 1479–1482, 2018.

[45] F. C. Ghesu *et al.*, "Marginal Space Deep Learning: Efficient Architecture for Volumetric Image Parsing," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1217–1228, 2016.

[46] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 968–982, 2012.

[47] R. Cunningham, M. Sánchez, G. May, and I. Loram, "Estimating Full Regional Skeletal Muscle Fibre Orientation from B-Mode Ultrasound Images Using Convolutional, Residual, and Deconvolutional Neural Networks," *J. Imaging*, vol. 4, no. 2, p. 29, Jan. 2018.

[48] R. J. Cunningham, P. J. Harding, I. D. Loram, and I. D. Cunningham, Ryan J, Harding, Peter J, Loram, "Deep residual networks for quantification of muscle fiber orientation and curvature from ultrasound," in *Medical Image Understanding and Analysis.*, Springer, Cham, 2017, pp. 63–73.

[49] R. J. Cunningham, P. J. Harding, and I. D. Loram, "The application of deep convolutional neural networks to ultrasound for modelling of dynamic states within human skeletal muscle," *arXiv*, 2017.

[50] "Anatomy TV: 3D Atlas of the Head and Neck." .

[51] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," pp. 1–8, 2013.

[52] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013," *Comput. Vision–ECCV 2014*, vol. 8689, pp. 818–833, 2014.

[53] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv*, 2014.

[54] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," pp. 1–15, 2014.

[55] V. C. W. Hoe, D. M. Urquhart, H. L. Kelsall, and M. R. Sim, "Ergonomic design and training for preventing work-related musculoskeletal disorders of the upper limb and neck in adults," *Cochrane Collab. Cochrane Database Syst. Rev.*, no. 8, 2012.

[56] D. Hoy *et al.*, "The global burden of neck pain: estimates from the Global Burden of Disease 2010 study," *Ann. Rheum. Dis.*, 2014.

[57] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," pp. 1–14, 2015.

[58] S. Ioffe, "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models," pp. 1–6, 2017.