Automatic annotation of cervical vertebrae in videofluoroscopy images via deep learning

Zhenwei Zhang[a], Shitong Mao[a], James Coyle[b], Ervin Sejdić[a,*]

[a] *Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA.*
[b] *Department of Communication Science and Disorders, School of Health and Rehabilitation Science, University of Pittsburgh, Pittsburgh, PA, 15261, USA.*

Abstract

Judging swallowing kinematic impairments via videofluoroscopy represents the gold standard for the detection and evaluation of swallowing disorders. However, the efficiency and accuracy of such a biomechanical kinematic analysis vary significantly among human judges affected mainly by their training and experience. Here, we showed that a novel machine learning algorithm can with high accuracy automatically detect key anatomical points needed for a routine swallowing assessment in real-time. We trained a novel two-stage convolutional neural network to localize and measure the vertebral bodies using 1518 swallowing videofluoroscopies from 265 patients. Our network model yielded high accuracy that mean distance between predicted points and annotations achieved 4.20 ± 5.54 pixels against human inter-rater errors (4.35 ± 3.12 pixels) and 93% of predicted points are less than five pixels of distance when tested on an independent dataset from 70 subjects. Our model offers more choices for speech language pathologists in their routine clinical swallowing assessments as it provides an efficient and accurate method for anatomic landmark localization in real-time, a task previously accomplished using an off-line time-sinking procedure.

Keywords: videofluoroscopy images, deep learning, dysphagia, vertebrae

*Corresponding author
Email address: esejdic@ieee.org (Ervin Sejdić)

detection

Oropharyngeal dysphagia poses serious health risks to people who suffer from stroke, head and neck cancer, older adults with multiple medical conditions, prematurely born infants and children with neurological, airway and developmental disorders Jones et al. (2018); Leslie et al. (2007). Among the consequences of dysphagia are pneumonia, airway obstruction, inadequate nutrition and hydration, and compromised quality of life He et al. (2019); Arnold et al. (2016); Steele et al. (2019). The evaluation of swallowing function is most often carried out with imaging based instrumental assessment that quantify swallowing function by recording and measuring important physiological events and connecting impairments to swallow outcomes such as inhalation of food/liquids (aspiration) and misdirection of swallowed material. These evaluation tests may include videofluoroscopy swallow studies (VFSS) Zhang et al. (2020), fiberoptic endoscopic evaluation Leder et al. (1998), ultrasound Lopes et al. (2019), CT scans and MRI scans Kumar et al. (2013); Carucci and Turner (2015). The VFSS is the most commonly used radiographic imaging method in clinical practice to confirm the presence and characteristics of dysphagia and assist with intervention planning to mitigate the negative impact of dysphagia. Additionally, manual kinematic measurements, completed by clinicians trained in kinematic analysis, provides the key details in VFSS about the biomechanical nature of swallowing impairments providing clinicians with ideas of potential interventions to improve swallow function. Manual analyses of swallow kinematics and physiology require frame-by-frame denotation of anatomic landmarks, quantification of the duration of key physiologic events and their timing in relation to one another, scoring of airway protection severity, and several other important measurements in order to accurately diagnose the swallowing disorder and derive logical treatment to alleviate it.

In addition to physiologic measurements, scaling of images to compensate for size differences among different patients is a crucial component of the analysis that enables each patient's swallowing function to be compared to norms

that would be expected of a healthy person of the same size. For example, Seo & Molfenter developed a method of scaling images by using the distance between antero-inferior margin of the second and fourth cervical vertebral bodies, in order to correct influence from patient head movement and participant size Molfenter and Steele (2014); Seo et al. (2016). Without scaling, the distance of structural displacements can be over- or under-estimated, leading to inaccurate diagnosis. In practice, each of these landmarks is manually marked on VFSS images, not in real-time but following the examination to serve as the anatomic scalar. The value of kinematic analysis can be exploited only for patients for whom VFSS is available, and unfortunately, kinematic analysis and physiologic measurement are also largely inaccessible in a timely manner in undeserved regions due to the lack of adequate clinical experts. This clinical limitation also leads to the reduction of VFSS analysis into a gestalt judgment of overall function and oversimplified analyses that unnecessarily focus on the completeness of flow of swallowed material instead of on the physiological impairments that lead to misdirection of swallowed food and liquids. Automation of kinematic analyses would provide a needed and important adjunct to swallowing physiologic assessments that could speed up treatment and lower the rates of adverse health consequences associated with dysphagia. The first step in automating kinematic analysis of VFSS images is to develop methods that autonomously produce the image scaling measurements and corrections currently possible only with manual annotation of images.

Automation of vertebrae detection and labeling has been widely investigated in static single-frame imaging such as computed tomography and magnetic resonance imaging Koompairojn et al. (2006); Lessmann et al. (2019); Forsberg et al. (2017); Galbusera et al. (2019); Chen et al. (2019). However, efforts by researchers to quantify and measure kinematic parameters during the swallow process in fluoroscopy, which generates thirty images per second, using computer vision are quite limited. Recent computer vision contributions are mainly involved in a semi-automatic frame-to-frame tracking of the hyoid bone, one important component in swallow kinematic analysis. Typical methods for hyoid

bone tracking include Sobel edge detection Kellen et al. (2010), Haar classifier matching Hossain et al. (2014), and local binary patterns Lee et al. (2017). However, these algorithms are still labor-intensive and time-consuming as they require a selection of a specific time interval from videos and manually defining region of interests in several frames in order to select the correct features in the images and then calculate the coordinate system based on anatomical landmark needed to adjust the subjects movement in each frame. Thus, it is impossible to deploy these methods in real-time during VFSS examinations. However, given recent hardware advances, deep learning techniques have enabled researchers to make a significant progress for various imaging tasks in an efficient and inexpensive way Zhang and Sejdić (2019). Among these techniques, convolutional neural network and or with ResNet block He et al. (2016) is one of state of art architectures that showed prominent achievement in facial, pose detection Angulu et al. (2017); Dong et al. (2018); Zadeh et al. (2017), medical imaging analysis Litjens et al. (2017), segmentation of brain lesions in multi-channel MRI image data Nie et al. (2016), left ventricle segmentation Ngo et al. (2017), micro- and macro-metastases of breast cancer Litjens et al. (2016), blood vessel identification Fu et al. (2016), tissue identification Chang et al. (2017), glomeruli localization, and nucleus detection Xie et al. (2018). Unfortunately, the dysphagia community has yet to utilize the full potential of deep neural network techniques in VFSS analysis. Only few studies achieved preliminary results via deep learning techniques. To investigate the four stages during swallows, Jone *et al.* proposed inated 3D neural network, a state of art architecture, to detect the pharyngeal phase in VFSS videos, yielding on an accuracy of 95% Lee and Park (2018). Zhang *et al.* applied deep learning techniques to detect the hyoid bone location automatically on each VFSS frame of the dysphagia examination and achieved an accuracy of 89% Zhang et al. (2018). Therefore, algorithms to automatically evaluate and assess VFSS dysphagia studies are highly sought after in the dysphagia clinical and scientific communities.

The purpose of this study is to demonstrate how deep learning neural networks can achieve high accuracies which are comparable to human annotators

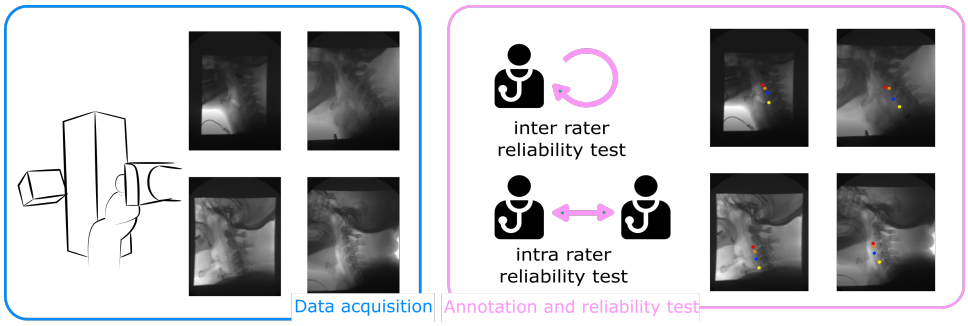

Figure 1: **Data acquisition and annotation procedure** Our dataset included annotated swallows collected from 335 subjects for the model training and evaluation. Video clips were recorded directly during VFSS examination. C2, C3, C4 vertebra locations were manually labeled by one main experienced expert during analysis. Inter-rater reliability test was implemented one month later and intra-rater reliability was tested with two other raters to ensure the accuracy of the judgment.

in anatomical landmark localization that can change the clinical assessment of dysphagia. Most importantly, our models maintain excellent performance even when validated on an independent test dataset, demonstrating its robustness and the generalizability needed for clinical settings. Specifically we present an investigation of deep learning in identifying the necessary anatomic scalar, the distance between the 2nd and 4th cervical vertebral bodies used to correct for size differences among patients, on all frames of a VFSS examination. We further sought to investigate how closely individual vertebral lengths (e.g,, C3 alone) corresponded to the longer C2-C4 segment currently used in kinematic analysis but whose most inferior landmark may not always be visible in VFSS images due to patient posture.

Methods

Videofluoroscopic swallow study dataset

Our dataset was collected from 265 patients with swallowing difficulty and 70 healthy volunteers who underwent videofluoroscopic examination at the Presbyterian University Hospital of the University of Pittsburgh Medical Center

(Pittsburgh, Pennsylvania, USA). The Institutional Review Board at the University of Pittsburgh approved the protocol of this study and all participants provided informed the consent. The first part of the dataset was collected from 265 patients. We didn't use statistical methods to predetermine sample size or subject age range. In this preliminary feasibility study, a convenience sample was used because there are no data upon which to base power calculations for sample size. The age range of these subjects was from 19 to 94 (148/265 males), and the average age was 64.83 $\pm$ 13.56 years old. Forty-four subjects (32/44 males, age: 66.6 $\pm$ 13.7) were diagnosed with stroke. All experiments in this data collection were performed in accordance with relevant guidelines and regulations. Participants in this study include During the VFSS examination, patients were required to swallow liquid boluses of various consistencies and volumes as well as pureed food and cookies, all containing barium. A standard data collection protocol was not followed for the patient data set. Instead, clinicians who conducted the VF modified the protocol for the administration of boluses (e.g. number of swallows, bolus consistencies, bolus volume and patient's head position) based on clinical appropriateness. The following consistencies were used in our studies: Varibar (Bracco Diagnostics, Inc.) thin liquid (<5cPs viscosity), Varibar nectar (300 cPs viscosity), Varibar pudding (5000 cPs viscosity), and Keebler Sandies Mini Simply Shortbread Cookies (Kellogg Sales Company). Patients were seated in the lateral plane with the 9 inch (22cm) image centered on the area of hyoid bone at rest to enable imaging capture of all necessary landmarks for the analyses. Patients were administered barium sulfate boluses either by the examining clinicians in the thin liquid texture from a spoon containing 3-5mL volumes for all consistencies, or self-administered liquid boluses from a cup while instructed to remain as still as comfortable. The second part of the dataset was collected from healthy subjects. The subjects signed up through a research registry. We recruited them based on age deciles and selected them via exclusion criteria questions to determine their eligibility. The questions include whether they have history of dysphagia, and history of head and neck, current pregnancy, or neurological disorders. The age range of

these subjects was from 21 to 87 (31/70 males), and the average age was 62.6 ± 14.70 years old. The subjects were asked to swallow varibar thin via spoon and self-selected volume from a cup.

In our investigation, we analyzed images from two data sets. The first dataset consisted of 265 patients referred for VFSs studies was collected from 2012 to 2015 using the Ultimax system (Toshiba, Tustin, CA) and the second dataset of 70 volunteers was acquired through Precision 500D system (GE Healthcare, LLC, Waukesha, WI) from 2018 to 2019. The videofluoroscopy system was set at a pulse rate of 30 pulses per second (full motion), and data were accrued at a sampling rate of 60 frames per second by a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA) and saved into a hard drive with a LabVIEW program. We down-sampled the video clips to 30 frames/second to eliminate duplicated frames. The first (patient) data set was recorded with 720 $\times$ 1080 resolution in real time while the second (healthy) dataset was captured with 1280 $\times$ 1024 resolution. Given the change in resolution between fluoroscopy units, we performed a laboratory comparison of multiple judgments of both spatial and temporal measurements of sets of videos resampled at both resolutions and found no significant differences in judgments between the two viewing conditions. In some videos, patient motion or anatomic opacification of the landmarks of interest due to patient posture (e.g., patient's shoulder region obscured visualization of the inferior landmark of C4) and underexposure/overexposure of x-ray dose (e.g., lack or saturation of vertebrae information) rendered some data unusable for this study, and those videos (more than half of them) were excluded from analyses, leaving a final data set of 1518 swallow video clips.

Human experts who were trained as previously described in swallow kinematic analysis identified anatomical points of interest (second vertebra and fourth vertebra) in 1518 swallow videos and annotated the landmark frame by frame in MATLAB (R2015b, The MathWorks, Inc., Natick, MA, USA). In addition, the head and tail of third vertebra were labeled on only first three frames of each subjects. Each swallow was segmented to include ll activity beginning with the frame in which the head of the bolus reached the lower

mandibular margin to when the tail end of the bolus passed through the upper esophageal sphincter (UES). Our annotation procedures include ongoing assessment of intra- and inter- reliability by requiring primary judges to repeat blinded annotations on 10% of randomly selected videos on an ongoing basis, and having a second trained ongoing and continuous stability of rater judgments and identifies judges in need of retraining. In our whole procedure, we maintained intraclass correlation coefficient over 0.9 to avoid judgment drift over time.

Image preprocessing and data augmentation

The total number of frames extracted from videos with annotations is 59810 images for our dataset. As we only collected the data from 335 subjects, the head position and image condition of VFSS images from the same patient were quite similar. The problem with the data set from the limited patients is that the trained model may suffer from overfitting and whould not generalize to test dataset. The data augmentation is well accepted practice to directly augment the input data to the model to increase the variety of perturbations in training data information, which more stringently trains the algorithms in detecting events during various common clinical testing conditions. In our dataset, we preprocessed the images from each patient. The augmentation methods included: random flipping half of images horizontally, rotating the images from -45 degree to 45 degree, shearing all images by -10 to 10 degrees, random cropping or padding 75% to 125% to original images, and changing the brightness of the images by multiplying 0.8 to 1.2. To be noted here, shearing technique may introduce incorrect anatomy structures; thus, it's rarely used in majority of medical applicationsNalepa et al. (2019). However, in our study, we find that many subjects have abnormal vertebrae shape which is visually similar to the results of shearing on normal vertebrae. After data augmentation, all of augmented images still contain the C2 - C4 landmarks and the total number of the training images remains unchanged. The deep learning networks highly require computation resources, we resized the input images into 448 $\times$ 448 considering the model training time. The original landmark point is shifted with respect to

the image center, and normalized by (w, h) as given by:

$$(x_i', y_i') = (\frac{x_i - 0.5w}{w}, \frac{y_i - 0.5h}{h}) \quad (1)$$

where (x_i, y_i) are given ground truth coordinate of landmark points and (x_i', y_i') are normalized and centered coordinates, treated as labels for networks training.

Overview of model development

Convolutional neural networks are commonly applied in medical imaging field, which can be used to discover the subtle patterns in a dataset. The main architecture tested in this study was a convolutional neural netwroks which used ResNet blocks followed by two convolutional layers. We implemented a two-stage networks architecture for vertebrae landmark detection. The basic idea of our two-stage network was inspired by Lv et al. (2017). In our design, the netwroks consist of two stages, the global detection network and the local detection network. The global stage provides the rough detection results of vertebrae locations and crops the vertebrae regions. We employed a CNN structure, which contains ResNet block, as our localization model to predict the coarse locations. ResNet block is popular architecture that makes use of the idea of 'short connection', skipping one or several layers and carrying input to the output, which allows to prevent vanishing gradient problem and fasten the training of the networks. We adopt the structure of ResNet-50 in global stage, which performs identity mapping for shortcut connections. We adjusted the last fully connected layer, which was originally designed for classification, to predict the vertebrae region.

Due to the various shape of vertebrae across the population, the global network may not capture all the variations of these difference, especially for the edge and the order of the vertebrae. To overcome the errors of local parts, we introduce the local network for the finer landmark localization, which is essential for accuracy improvement. Images are cropped via the prediction results from the global stage network, then scaled and fed into the local stage network. Similar with the global stage network, we adopt ResNet-34 structure, with the

last fully connected layer adjusted to directly regress the landmark locations on the input images. The inverse transformation function is applied to map the predicted points to the original input image.

Normalization is widely adopted techniques that enables more stable and faster training of deep learning models. In our study, we found that the switchable normalization showed better performance than batch normalization layers in ResNet blocks in the training phase. Switchable normalization combines batch normalization, layer normalization and instance normalization using weight average, which allows the custom choice of normalization depending on the depth of the layer and training batch size (Fig. 3). Batch normalization was proposed and widely implemented in ResNet and similar convolutional network architecture. It reduces internal covariate shift by using mini-batch mean and variance to normalize each mini-batch of data. The normalized version of a mini-batch of inputs $\{x_1, ..., x_m\}$ is computed as follows:

$$\hat{x_i} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad \text{with} \quad \mu = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(xi - \mu)^2$$

The layer normalization normalizes features within each sample, instead of normalizing across samples. The layer normalization is computed over all hidden units (H) in the same layer:

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \quad \sigma^2 = \frac{1}{H}\sum_{i=1}^{H}(a_i^l - \mu^l)^2$$

Similar to layer normalization, instance normalization normalizes features within channels.

The loss function was defined as Euclidean loss for landmark location prediction, which is computed from

$$loss = \frac{1}{2}\sum_{N}^{i=1}((\hat{x_i} - x_i')^2 + (\hat{y_i} - y_i')^2) \tag{2}$$

where $(\hat{x_i}, \hat{y_i})$ are landmark location predicted by the network. We computed the loss function on the training and validation data, and we selected the model with best loss function score on validation dataset as our final model. We fine-tune

the ResNet via transfer-learning and also trained networks from the scratch. The advantages of normalization layers is to regularize the model, reduce the overfitting and improve the model performance. Normalization layers change the distribution in network weights during training.

Training the two-stage network model

In this investigation, two dataset were utilized in the model training and evaluation. The first data was collected from 265 patients using Ultimax system, with 70% of subjects for training, 30% for validation. An extra independent data collected from 70 volunteers was applied for the final testing. We ensure that no person in the training group is in the validation and test group to make it a truly independent group. We also implemented the 5-fold cross-validation on data set collected from patient group and evaluated model performance on patient dataset and independent healthy subjects. In the original paper, the ResNet block utilized the batch-normalization layer. In our model, we implement and tested the residual block using switchable normalization instead of batch-normalization layer. The training curve of batch-normalization and switchable normalization is listed in supplemental files. We trained our two-level neural network models via fine-tune of original ResNet and fully trained switchable ResNet block. The model that performs best on the validation dataset is selected for testing. The switchable normalization showed slight better accuracy compared to the transfer fine-tuning using original ResNet structure. In this study, training and Testing procedures were implemented using Pytorch on the NVIDIA Tesla M40 GPU. We utilized Xavier initialization to initialize weights in the networks, and we used exponential decay learning rate starting from 0.01 and the learning rate was scale by 0.95 after each epoch. The whole-images were resize into 448 $\times$448 and the models were trained over 80 epochs on the first patient dataset with 80 % for training and 20 % for validation. Due to the limitation of C3 annotations, we trained first stage network only with C2 and C4 labels, then the second network were trained with all annotations.

Testing and Analysis

Once the model finished the training, the evaluation of model was implemented on the testing dataset, which independent and not included in the training dataset. All parameters in the models were frozen and we predict landmark points by a forward-pass through the networks. As we rescaled and shifted the landmark points during training phase, these points should be scaled and shifted back to the original image coordinates:

$$(x_i, y_i) = (\hat{x_i'}w + 0.5w, \hat{y_i'}w + 0.5w) \tag{3}$$

The purpose of this study is to locate the key points of vertebrae in the videofluoroscopic images, whose information can be used as an important reference in clinical kinematic analysis. First, we evaluate the mean and standard deviation of location pixel difference between ground truth and points predicted by the models. We also evaluate the percentage of pixel difference compared to whole image size. In addition, we checked how the results were affected when various normalization layers and different input size were applied during model training. We asked three well-trained pathologists (intra- and inter- reliability score greater than 0.9) to manually label the C2-C4 landmarks points, comparing the results tolerance within human and the error distance between humans and machine predictions. Vertebrae information are used to build a coordinate for kinematic analysis in dysphagia field. To evaluate the performance of model, we calculated and compared the ratio of C2, C4 unit, and angle of C2-C4 coordinate. The ratio of C2-C4 is calculated by predicted C2-C4 length over annotated C2-C4 length. The angle of C2-C4 indicates angle between vector of predicted C2-C4 and vector of annotated C2-C4.

Results

We demonstrated an automated pipeline to measure the location, length and orientation of several cervical vertebrae in videofluoroscopic images. First, experienced raters conducted manual anatomic annotation of frame-by-frame

videofluoroscopic data, which was collected from 265 subjects with suspected dysphagia and 70 healthy participants (Method). Raters annotated the location of antero-inferior corner of C2, and the anterior-superior and anterior inferior corners of the C3 and C4 vertebral bodies, as shown in Fig. 1. These measurements served as the ground truth for determining the length of this vertebral axis. Given an input image, the first step is to crop the image by removing the patient information and baffle region (black regions shown in Fig. 1) around the patient's neck region which was used to reduce the radiation during examination. Then the cropped region is scaled to a fixed size and fed into a two-stage network (Fig. 2). The convolutional networks were trained to learn features and patterns from images and mathematically describe the relationship between human annotations and the input images. After training the networks, these parameters were frozen in order to make the prediction on the validation dataset and the test dataset. The first stage network predicted the coarse location of vertebrae landmark regions and the second network finely improved the landmark regions. The model performance is evaluated by measuring mean localization distance, length ratio and angle error. Localization distance measures the actual distance in pixels between predicted landmark coordinates and the labeled landmark coordinates. Length ratio measures the ratio between predicted C3/C2-C4 length and the labeled length while the angle error measures the angles between predicted C3/C2-C4 vector and manually labeled vector. These two metrics are important parameters in the dysphagia analysis, which are widely used to reduce the bias among population in decision making. Thus, we mainly focus on the accuracy of length and orientation measurement.

In the experiment, the model was trained using swallows from 265 consenting patient subjects, and then tested on the second dataset from 70 additional healthy volunteers, which were treated as unseen samples for the deep learning model to evaluate generalization. Notably, our second data was collected three years later and used a different videofluoroscopy machine, which can present the challenge of the invariant performance of our method on vertebrae location given different imaging resources.

In this study, the performance of our model referred to how closely the predicted vertebral locations corresponded to human judgment. An example of a continuous swallowing video captured at 30 images per second, is shown in Fig. 4(a). At each time point, the two-stage model localizes the location of C2, C3 and C4 vertebra. The images on the left show the ground truth and the frame with the largest distance error in vertical direction and the right images right images are those with largest localization error in horizontal direction. Overall, the location results from our model for one subject are reliable. Fig. 4(b) presents several location detection results on the test dataset, with orange for the ground truth, blue for the first stage results and red for our model's final results. The model was applied to the testing set, an independent dataset involving 70 subjects, and mean localization distance (MLD) achieved 4.20 ± 5.54 pixels. In order to verify the advantages of using two-stage networks, we compared the results with the model which uses ResNet50 for training. ResNet50 architecture led to a MLD at 7.44 ± 5.38 pixels. The summaries of localization distance distribution in testing the dataset compared to the human raters' annotations is shown in Fig. 5(b). As there were no established gold standard or previous experiences that could inform our methods. Regarding the acceptable localization distance tolerance, we chose 1 % of the whole image size as our criteria (i.e error less than 5 pixels range). The percentage of acceptable predicted locations via ResNet50 is 49.66% while the two-stage networks gave 87.36 %. The variability across multiple raters is unavoidable due to the limited quality of VFSS images, which is why the reliability test is deployed in routinely in research and routine clinical practice. In this study, the overall kappa ICC between two human raters and between the rater and the model both achieved over 0.9, showing that our model is comparable to human raters. Fig. 5(a) compared the model's predictions errors and one human rater judgment bias on the test data. Ninety percent of the predicted data shows comparable predictions to the second rater judgment while the model still has about 5% of results which demonstrated larger locations errors than the likely errors produced by the human rater during the manual annotation process.

Table 1: **Model performance with 5-fold cross validation** The performance of the model was evaluated with 5-fold cross-validation and each trained model was also tested on the healthy data set.

	Patient Data			Healthy Data		
	MLD	Angle Error	Length Ratio	MLD	Angle Error	Length Ratio
fold1	4.19 ± 4.77	0.04 ± 0.05	1.02 ± 0.04	4.14 ± 5.65	0.03 ± 0.04	1.00 ± 0.05
fold2	4.00 ± 4.26	0.03 ± 0.04	1.01 ± 0.04	4.54 ± 5.66	0.04 ± 0.04	1.00 ± 0.03
fold3	4.13 ± 4.51	0.03 ± 0.05	1.03 ± 0.04	5.37 ± 7.76	0.04 ± 0.04	1.00 ± 0.05
fold4	4.17 ± 6.64	0.02 ± 0.02	0.99 ± 0.03	4.85 ± 5.68	0.05 ± 0.04	0.99 ± 0.05
fold5	3.82 ± 4.90	0.03 ± 0.07	1.00 ± 0.03	4.49 ± 5.44	0.04 ± 0.03	1.01 ± 0.06

Compared to the exact location of vertebrae, estimating the cervical vertebrae length and orientation is highly desired in the clinical settings as these information are usually served as patient-specific criterion referenced correction factor. In our study, we measured the length between C2-C4 and the length of C3 unit. Fig. 5(c) and (d) present the length ratio distribution and angle error distribution between estimated cervical vector and label vector respectively. The mean estimated length ratio from ResNet50 is 1.04 ± 0.09 and 45.95% of them are located in the length ratio range 0.95 to 1.05 while 93.76 % of predictions from two-stage model are located in the same range with mean estimated length ratio 0.99 ± 0.04. The mean absolute angle errors from ResNet 50 is 0.06 ± 0.05 rads and 0.03 ± 0.03 rads for our two-stage model.

To evaluate the performance of the model, we implemented 5-fold cross-validation on patient data and tested each model on healthy data as well. Table 1 presents the MLD, angle error and C2-C4 length ratio for each fold. The average of MLD is 4.07 pixels on patient group and 4.67 pixels on healthy group. The results indicate that the model generalized well on both data set while they were collected from two different video fluoroscopic machines.

Discussion

This study is the first step toward a fully automatic diagnostic image analysis system based upon computational methods, rapidly offering the vertebral scaling information that facilitates objective and accurate measurement in real time. The finding that our two-stage model could accurately and autonomously determine the anatomic scalar necessary for accurate measurements kinematic sets the stage for advancing automated analysis methods from VFSS images. The potential for speeding VFSS interpretations with automated data reduction methods while maintaining precise measurement is broad can improve the consistency of interpretations of VFSS images by providing standard measurements of swallow physiology that lower subjectivity in judgment leading to interventions for dysphagia. In current clinical setting, the importance of an anatomic scalar in VFSS measurement cannot be understated. Given the differences in the sizes of different patients and the direct association between a person's height and the dimensions of the upper aerodigestive tract Steele et al. (2011), the ability to equalize measurements for differences in patient size provides the ability to compare results across patients of different dimensions. Moreover, real-time scaling of images provides immediate raw data for clinical interpretations which accelerates decision-making and increases efficiency of clinical workflow. In dysphagia diagnosis, the use of the vertebral scalar serves as the reference scale for linear measurements commonly used to infer about the nature of a patient's swallowing disorder (e.g., hyoid bone displacement, upper esophageal sphincter opening) that are the basis for determining appropriate treatments and judging the effects of those treatments objectively Molfenter and Steele (2014). In turn, researchers investigating differences in swallow physiology in different disease states, and generation of population-based against which to compare patient function in disease states, provides for accurate determination of the magnitude of various kinematic impairments and a roadmap for determining the success or failure of treatments that restore that function.

Our two-level framework demonstrates the efficacy of using a large dataset

and deep learning architectures for vertebrae landmark localization in videofluoroscopy images. Unlike previous semi-automation attempts for dysphagia keypoints Lee et al. (2017), we conducted our model on a relatively large dataset, including over 300 subjects. Compared to other studies, we included the subjects across the adult age span varying from 19 to 94 years old and included both people with dysphagia and healthy subjects, showing the robustness of the algorithms. Additionally, our dataset not only collected single swallows, but also multiple sequential swallows and swallows in neutral and chin down head positions, all factors that are known to alter judgment of kinematic events when there is large scale motion of the patient during testing. Such diverse dataset prompted us to utilize deep learning approaches, avoiding the attempts of unstable, less powerful traditional image processing methods and classifiers.

Traditional image processing methods focuses on matching local edge and corner features. However, specific frames are rendered unmeasurable with these methods due to noisy edge and corner information in cases of patient motion during the exam, and the effect of the flowing bolus through the video field, influencing the performance of feature matching. In addition, these corner and edge features are influenced by image quality and various vertebral shape across different subjects. To overcome these limits and accurately detect the vertebrae shape with various location and edge shape, we adopted the two level framework in our study, which leverages deep learning technology and learns coarse representation from the VFSS dataset, followed by fine learning from the subregions to localize the keypoints on vertebrae. The coarse detection provides the approximate region of interest which contains C2, C3, C4 vertebrae information, removing the irrelevant information and also reducing the burden of computation for the second stage network. As shown in Fig. 4(b), the second stage network well improved the detection performance from the first network, which shows the importance of the usage of local network structure.

In this study, we have also demonstrated that the current framework can cope with the vertebral locations from videofluoroscopic images via two different videofluoroscopy systems and perform better than transfer learning techniques.

Our framework was built based on ResNet-like structures with switchable-normalization, which is beneficial to the model generalization and stability. To compare the performance, we also trained our model using transfer learning techniques via the pre-trained network on Image-Net, a huge image database which contains various natural images. Transfer learning is a popular method that allows deep learning transferred the pre-knowledges to the new dataset, usually lower training burden and achieve better results. However, our results suggested that the usage of ResNet with switchable normalization instead of batch normalization and training the network from the scratch shows better performance to transfer learning techniques. Shown in the supplement figure, switchable normalization trained from scratch converged better than transfer learning with batch normalization . Furthermore, deeper ResNet structure proved a better accuracy.

Our study has some limitations, notably for the size of individual subjects and imaging resources. While our dataset is relatively large in the dysphagia community, it is still small compared to the popular medical imaging research on organs such as brain and lungs. ~~The~~ First, our data may suffer from the bias of sex with 32/44 male stroke patients. Studies showed that females had higher upper cervical lordosis than males, whereas male had higher lower cervical lordosis than female Been et al. (2017). The intention of a diagnostic application should be gender-neutral Mehrabi et al. (2021),it remains unknown whether the models' bias is from the unbalance of the demography. Future investigations could consider the fairness of the model performance across demography difference. In addition, the sample in this study may not be inclusive of the entire range of variety of anatomic information, which resulted in mis-localization in several cases. As shown in the figure 6, blue dots are the predictions from first stage network, and red/green dots are from second stage network. While second network improved the predictions from first network, its prediction were shifted in both case (a) and (b). In case (a), the C2 and C3 vertebrae contacted in the image due patient's head direction. The model correctly predicted the C2 tail but not other points. In case (b), the model failed

to predict C4 tail due to abnormal C4 and C5 structure. The deep networks not only learned the features from the input image itself and the connection between input and output, they are able to learn the potential relationship between outputs, which might be the reason for this shifted wrong predictions. These abnormal cases such as abnormal bone shape (e.g., cervical osteophytes), postoperative anatomic disruption (e.g., anterior cervical fusion with graft or hardware), altered spinal configuration (e.g., kyphosis, excessive lordosis), or presence of feeding tubes or tracheostomies, provide direction for future research in model training that leads to better generalization of our model across more patient populations. We expect that the model performance will increase as more subjects are included and images are collected from multiple videofluoroscopic machines. On the other hand, other techniques such as multi-stage networks and cascade network have been proposed in facial detection and pose estimation Li et al. (2019); Fan and Zhou (2016). These methods are not constrained by the global and local networks and use several networks to improve landmark locations step by step and may provide advantages that improve detection. However, whether these architectures can improve the performance for the VFSS detection with a larger dataset can improve the performance for the VFSS detection remains an opening question.

In the future, we would ideally extend the localization to other landmarks commonly considered in dysphagia studies (e.g., hyoid bone, arytenoid cartilages, valleculae, and epiglottis) as well as other parameters for swallow measurements. By extending our framework to study a wider range of features and providing a quantitative assessment in swallow videos, we hope that this deep learning approach is able to aid language pathologists' routine evaluation by automating some aspects of daily data analysis. This will enable clinicians to allocate their limited clinical resources on higher-level interpretations of the measurements to provide top-of-license services rather than spending valuable time performing the rote measurement necessary for these interpretations. We also hope that our framework could play an important role in research in order to develop more precise benchmarks for separating disordered from typical

function that aids clinical interpretations, and in characterizing the properties of dysphagia in various disease states.

Conclusion

In this research, we introduced a deep learning neural network-based method for anatomic landmarks localization in videofluoroscopic images. We showed that our two-stage framework are able to accurately estimate the length and angle of cervical vertebrae with mean localization distance comparable to human annotators. We believe that deep learning approach will lead to automation of kinematic analyses that could speed up time to diagnosis and treatment.

Acknowledgments

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institute of Health under Award Number R01HD092239, while the data was collected under Award Number R01HD074819 and 2R01HD074819-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health.

Competing interests

The authors declare no competing interests.

Author contributions

ZZ and SM designed and conducted the experiment, JC provided clinical insights and supports. All authors discussed the results and commented on the manuscript.

References

Angulu, R., Adewumi, A.O., Tapamo, J.R., 2017. Landmark localization approach for facial computing, in: Proceedings of Conference on Information Communication Technology and Society, pp. 1–6.

Arnold, M., Liesirova, K., Broeg-Morvay, A., Meisterernst, J., Schlager, M., Mono, M.L., El-Koussy, M., Kägi, G., Jung, S., Sarikaya, H., 2016. Dysphagia in acute stroke: incidence, burden and impact on clinical outcome. PloS one 11, e0148424.

Been, E., Shefi, S., Soudack, M., 2017. Cervical lordosis: the effect of age and gender. The Spine Journal 17, 880–888.

Carucci, L.R., Turner, M.A., 2015. Dysphagia revisited: common and unusual causes. Radiographics 35, 105–122.

Chang, H., Han, J., Zhong, C., Snijders, A., Mao, J.H., 2017. Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Chen, Y., Gao, Y., Li, K., Zhao, L., Zhao, J., 2019. Vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model. IEEE Transactions on Medical Imaging .

Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018. Style aggregated network for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–388.

Fan, H., Zhou, E., 2016. Approaching human level facial landmark localization by deep learning. Image and Vision Computing 47, 27–35.

Forsberg, D., Sjöblom, E., Sunshine, J.L., 2017. Detection and labeling of vertebrae in MR images using deep learning with clinical annotations as training data. Journal of Digital Imaging 30, 406–412.

Fu, H., Xu, Y., Wong, D.W.K., Liu, J., 2016. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields, in: the 13th IEEE International Symposium on Biomedical Imaging, IEEE. pp. 698–701.

Galbusera, F., Casaroli, G., Bassani, T., 2019. Artificial intelligence and machine learning in spine research. Jor Spine 2, e1044.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, in: Proceedings of European Conference on Computer Vision, pp. 630–645.

He, Q., Perera, S., Khalifa, Y., Zhang, Z., Mahoney, A.S., Sabry, A., Donohue, C., Coyle, J.L., Sejdić, E., 2019. The association of high resolution cervical auscultation signal features with hyoid bone displacement during swallowing. IEEE Transactions on Neural Systems and Rehabilitation Engineering 27, 1810–1816.

Hossain, I., Roberts-South, A., Jog, M., El-Sakka, M.R., 2014. Semi-automatic assessment of hyoid bone motion in digital videofluoroscopic images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2, 25–37.

Jones, E., Speyer, R., Kertscher, B., Denman, D., Swan, K., Cordier, R., 2018. Health-related quality of life and oropharyngeal dysphagia: A systematic review. Dysphagia 33, 141–172.

Kellen, P.M., Becker, D.L., Reinhardt, J.M., Van Daele, D.J., 2010. Computer-assisted assessment of hyoid bone motion from videofluoroscopic swallow studies. Dysphagia 25, 298–306.

Koompairojn, S., Hua, K.A., Bhadrakom, C., 2006. Automatic classification system for lumbar spine x-ray images, in: Proceedings of 19th IEEE Symposium on Computer-Based Medical Systems, pp. 213–218.

Kumar, K.V., Shankar, V., Santosham, R., 2013. Assessment of swallowing and its disorders—a dynamic mri study. European journal of radiology 82, 215–219.

Leder, S.B., Sasaki, C.T., Burrell, M.I., 1998. Fiberoptic endoscopic evaluation of dysphagia to identify silent aspiration. Dysphagia 13, 19–21.

Lee, J.C., Nam, K.W., Jang, D.P., Paik, N.J., Ryu, J.S., Kim, I.Y., 2017. A supporting platform for semi-automatic hyoid bone tracking and parameter extraction from videofluoroscopic images for the diagnosis of dysphagia patients. Dysphagia 32, 315–326.

Lee, J.T., Park, E., 2018. Detection of the pharyngeal phase in the videofluoroscopic swallowing study using inflated 3d convolutional networks, in: Proceedings of International Workshop on Machine Learning in Medical Imaging, pp. 328–336.

Leslie, P., Drinnan, M.J., Zammit-Maempel, I., Coyle, J.L., Ford, G.A., Wilson, J.A., 2007. Cervical auscultation synchronized with images from endoscopy swallow evaluations. Dysphagia 22, 290–298.

Lessmann, N., van Ginneken, B., de Jong, P.A., Išgum, I., 2019. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. Medical Image Analysis 53, 142–155.

Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y., Sun, J., 2019. Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 .

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak,

J., 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports 6, 26286.

Lopes, M., Silva, C., Lima, L., Lima, D., Costa, B., Magalhães, D., Rodrigues, D., Rêgo, T., Pernambucano, L., Santos, A., 2019. A deep learning approach to detect hyoid bone in ultrasound exam, in: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), IEEE. pp. 551–555.

Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X., 2017. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3317–3326.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54, 1–35.

Molfenter, S.M., Steele, C.M., 2014. Use of an anatomical scalar to control for sex-based size differences in measures of hyoid excursion during swallowing. Journal of Speech, Language, and Hearing Research 57, 768–778.

Nalepa, J., Marcinkiewicz, M., Kawulok, M., 2019. Data augmentation for brain-tumor segmentation: a review. Frontiers in computational neuroscience 13, 83.

Ngo, T.A., Lu, Z., Carneiro, G., 2017. Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Medical Image Analysis 35, 159–171.

Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D., 2016. Estimating CT image from MRI data using 3D fully convolutional networks, in: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 170–178.

Seo, H.G., Oh, B.M., Han, T.R., 2016. Swallowing kinematics and factors associated with laryngeal penetration and aspiration in stroke survivors with dysphagia. Dysphagia 31, 160–168.

Steele, C.M., Bailey, G.L., Chau, T., Molfenter, S.M., Oshalla, M., Waito, A.A., Zoratto, D.C., 2011. The relationship between hyoid and laryngeal displacement and swallowing impairment. Clinical Otolaryngology 36, 30–36.

Steele, C.M., Mukherjee, R., Kortelainen, J.M., Pölönen, H., Jedwab, M., Brady, S.L., Theimer, K.B., Langmore, S., Riquelme, L.F., Swigert, N.B., Bath, P.M., Goldstein, L.B., Hughes, R.L., Leifer, D., Lees, K.R., Meretoja, A., Muehlemann, N., 2019. Development of a non-invasive device for swallow screening in patients at risk of oropharyngeal dysphagia: Results from a prospective exploratory study. Dysphagia , 1–10.

Xie, W., Noble, J.A., Zisserman, A., 2018. Microscopy cell counting and detection with fully convolutional regression networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 283–292.

Zadeh, A., Chong Lim, Y., Baltrusaitis, T., Morency, L.P., 2017. Convolutional experts constrained local model for 3d facial landmark detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2519–2528.

Zhang, Z., Coyle, J.L., Sejdić, E., 2018. Automatic hyoid bone detection in fluoroscopic images using deep learning. Scientific Reports 8, 12310.

Zhang, Z., Perera, S., Donohue, C., Kurosu, A., Mahoney, A.S., Coyle, J.L., Sejdić, E., 2020. The prediction of risk of penetration–aspiration via hyoid bone displacement features. Dysphagia 35, 66–72.

Zhang, Z., Sejdić, E., 2019. Radiological images and machine learning: trends, perspectives, and prospects. Computers in Biology and Medicine 108, 354–370.

Supplemental materials

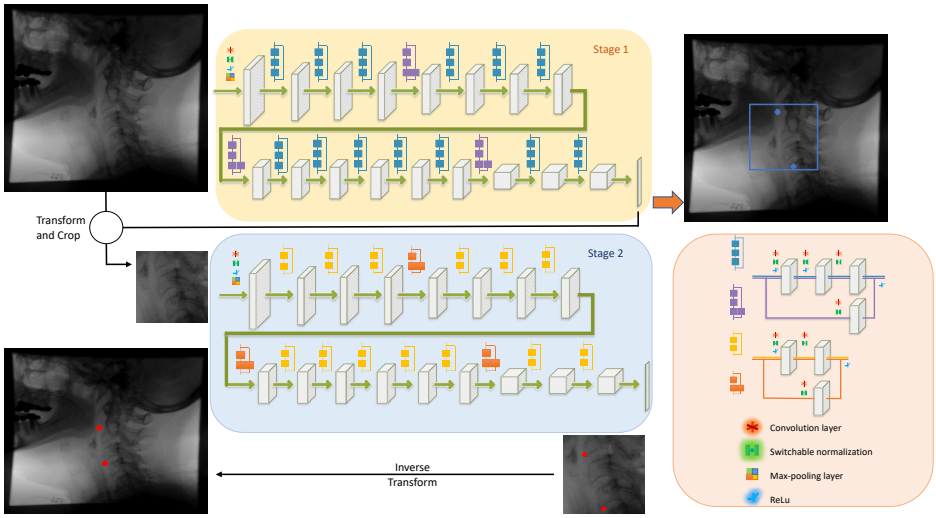

Figure 2: **The pipeline of the proposed two-stage network architecture for vertebrae landmark localization** First, a new input image is preprocessed to remove the patient information and dark regions in the videofluoroscopy image. After preprocessing process, the input image is fed into the first stage of the network to achieve the coarse detection, which allows to crop the image for finer detection. Then, the cropped image, which covers the vertebrae region, is fed into the local stage network for a better landmark localization. The output vectors from the network, which indicates the location of the vertebrae in the cropped image, are projected back to the initial image. The two-stage network consists of several ResNet blocks in each stage network. The first stage network follows the idea of ResNet50 while ResNet34 structure is implement in the second stage network. The ResNet block include several Convolutional layers, followed by normalization layers and a rectified linear unit(ReLU), then an extra identity map create a shortcut between input layer and output layer of the block. Different from the traditional ResNet block, we implemented switchable normalization layers instead of batch normalization layers, which allows to adaptively switch among various normalization techniques.

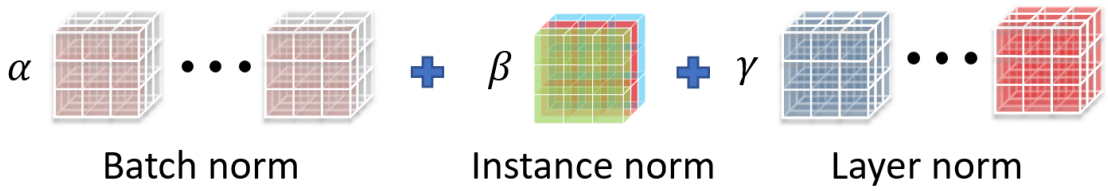

Figure 3: **Switchable normalization** Switchable normalization combines batch normalization, layer normalization and instance normalization using weighted average the means and variances. It allows networks to find the suitable ratios among three normalizations for each layer during training.

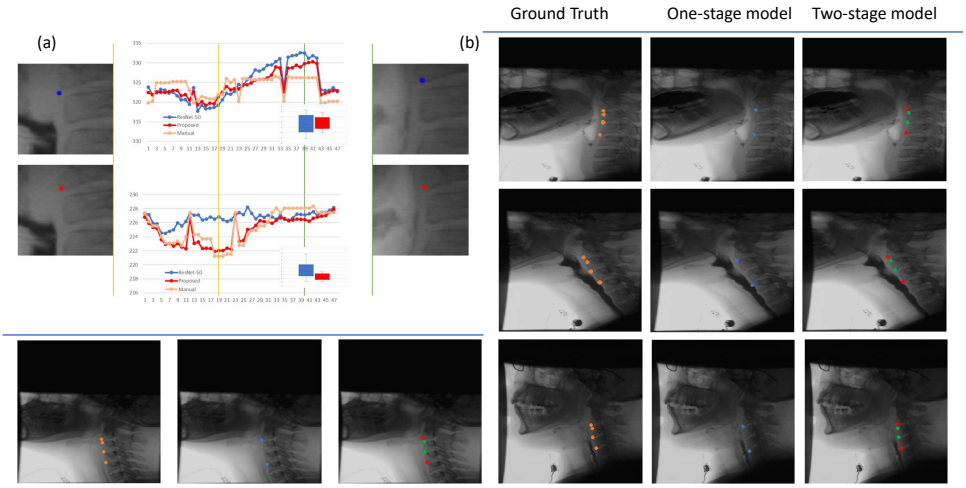

Figure 4: **Landmark localization results demonstrating the two-level model's robustness to variations among patients** (a) localization results predicted on a continuous swallowing video. Blue lines indicate the prediction from predictions, which show larger error variance comparing to red lines (the two-stage model), demonstrating the benefit of our model. Left images illustrates the largest error in y direction and the right images corresponds to the x direction. (b) Examples of the selected videofluoroscopic images with manually annotations, predictions from ResNet50 (first stage) and final prediction results. Note how the second stage achieved invariance to the scale and is able to perform localization despite head pose, vertebrae shape and lighting for different individuals.

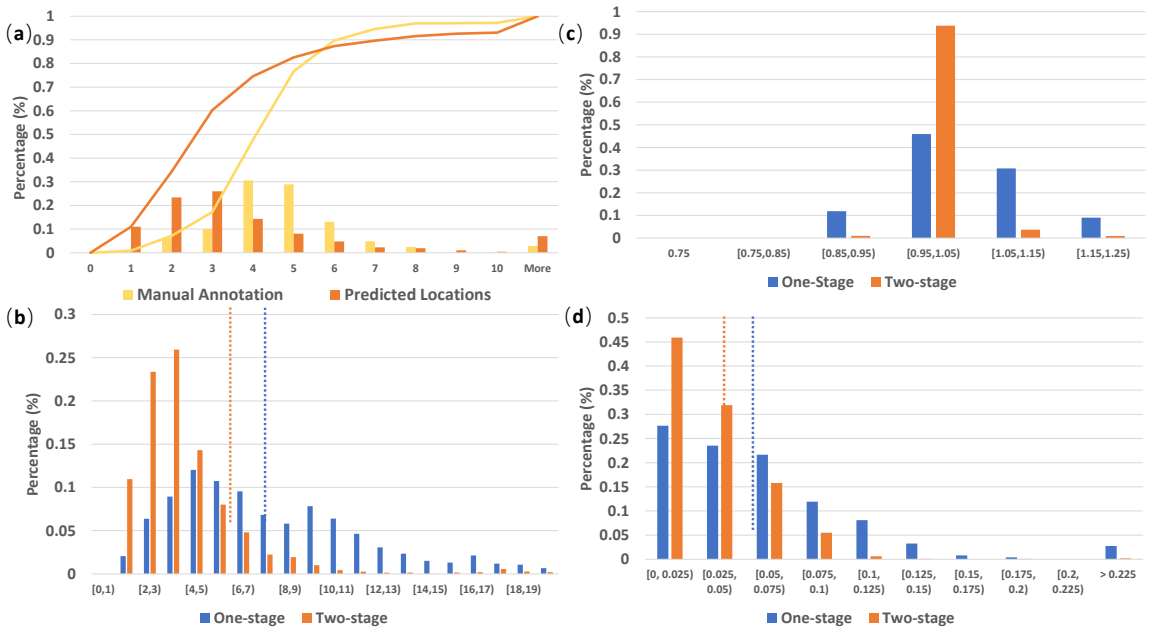

Figure 5: **Human judgment and landmark localization results**(a) The curve indicates the accumulative sum of locations distance errors. Yellow line indicates pixel distances between two human rater judgment and orange line indicates pixel distances between model prediction and one of human raters. (b) Distribution of localization distance errors between predicted and labeled annotation from first stage network and second stage network (c) Length ratio between predicted C2-C4 vector length and the manual annotation (d) Angle errors between predicted vector and manual annotation

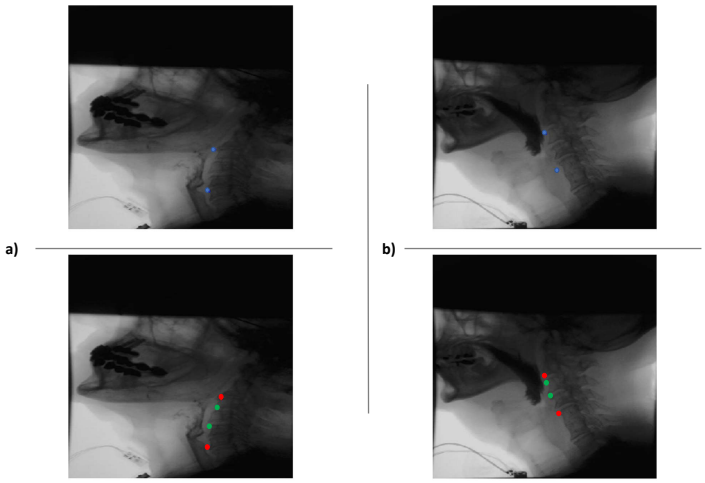

Figure 6: **Failure cases on testing dataset** Blue dots: predictions from one stage network. Green dots: C3 prediction from two stage networks. Red dots: C2, C4 tail edge detection from two statge networks. While two stage networks shows better results in numerical errors, we still can find that the landmark predictions are shifted when subjects are in a extreme posture or with an abnormal vertebra shape.

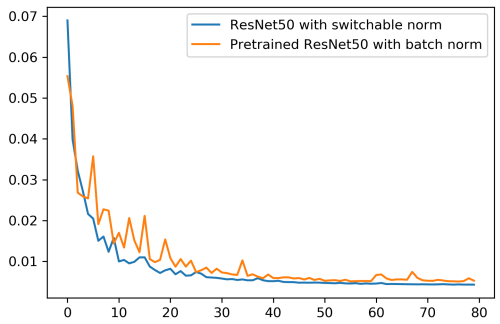

Figure 7: **Validation loss during training phase** Orange: model was trained based on pretrained ResNet50 network. Blue: model was trained from scratch using switchable normalization

We first implemented a preliminary study for our model selection, we compared the model performance with batch normalization trained with inputs without any augmentation and the inputs with augmentation. We used the combination of focal loss and dice coefficient function as our loss function. The input without augmentation led to 0.0074 (training) and 0.0102 (validation), and the input with augmentation led to 0.0060 (training) and 0.0065 (validation). The model with augmented input shows lower loss value in both training and validation dataset. Next, we trained our model with augmented input and switchable normalization on the same data, the model with switchable normalization, it achieved 0.0050 (training) and 0.0054 (validation). Figure 7 shows the training curve for batch normalization and switchable normalization.