Vertebrae Identification and Localization Utilizing Fully Convolutional Networks and a Hidden Markov Model

Yizhi Chen, Yunhe Gao, Kang Li, Liang Zhao, and Jun Zhao, *Member, IEEE*

***Abstract*— Automated identification and localization of vertebrae in spinal computed tomography (CT) imaging is a complicated hybrid task. This task requires detecting and indexing a long sequence in a 3-D image, and both image feature extraction and sequence modeling are needed to address the problem. In this paper, the powerful fully convolutional neural network (FCN) technique performs both of these tasks simultaneously because FCNs directly encode and decode the spatial interdependence of different components in images. The key module of our proposed framework is a 3-D FCN trained in an end-to-end manner at the spine level to capture the long-range contextual information in CT volumes. The large increase in the calculation due to the full-size image inputs is alleviated by the scale-down of the inputs and the use of an auxiliary FCN to compensate for the loss of details. The composite network pipeline design enables the integration of local image details and global image patterns. Furthermore, explicit spatial and sequential constraints are imposed by the hidden Markov model (HMM) for a higher robustness and a clearer interpretation of network outputs. The proposed framework is quantitatively evaluated on the public dataset from the MICCAI 2014 Computational Challenge on Vertebrae Localization and Identification and demonstrates an identification rate (within 20 mm) of 94.67%, a mean identification rate of 87.97%, and a mean error distance of 2.56 mm on the test set, thus achieving the highest performance reported on this dataset.**

***Index Terms*— Automatic vertebrae identification and localization, CT image, deep learning, convolutional neural network, fully convolutional network.**

I. Introduction

THE spine serves as a crucial anatomical landmark in three-dimensional (3-D) medical images because it runs along the length of the neck and trunk of the human upper body. Therefore, accurate automated vertebrae detection and localization acts as a substantial component of 3-D image-guided diagnosis, pre-operative planning and post-operative evaluation. This task is also a prerequisite for spine-related structural segmentation (vertebral body or intervertebral disc), idiopathic scoliosis diagnosis and other applications [1]–[3].

The purpose of vertebrae localization is to locate and index the 3-D coordinates of all vertebrae centroids in a computed tomography (CT) volume. Similar to many other image-guided diagnoses in medical images, the vertebrae localization procedure can be accelerated and enhanced by computer-aided diagnosis (CAD) systems because the manual labeling of vertebrae in 3-D medical images can be laborious and tedious, with unpredicted personal subjectivity [4].

Unfortunately, the design of an automatic system for vertebrae identification and localization is not as easy as it seems. The vertebrae localization scenario is actually a sophisticated sequence modeling problem in 3-D space. The presence of pathological diseases, arbitrary field-of-view of CT scans, and metal implant artifacts introduce additional challenges. Some hard cases with severe abnormality are difficult to deal with even for the experienced specialists.

Local details are required to detect the image patterns, separate the vertebrae from each other, and accurately locate the vertebrae centroid coordinates. The similar morphological structure of adjacent vertebrae increases the difficulty in distinguishing the individual vertebrae type from the scope of continuous vertebrae categories. Therefore, global image patterns and vertebrae sequential information are essential to reduce the ambiguity. Moreover, the inter-subject variability introduced by various pathological circumstances, such as scoliosis, kyphosis, vertebra fractures, lumbarization and lumbar sacralization, is a difficult issue, especially if limited training data are available. The probable existence of surgical metal implants, which might cause peculiar image artifacts and blur the border region, adds to the difficulty of vertebrae identification. Fig. 1 displays a difficult case with severe abnormalities and artifacts.

A. Related Work

In the most recent decade, many approaches have been proposed to address the challenging problem of vertebrae identification and localization. Certain early studies pioneered the first work in this area [5]–[7]. In 2012, Glocker *et al.* [8] utilized

Manuscript received May 4, 2019; revised June 30, 2019; accepted July 1, 2019. Date of publication July 8, 2019; date of current version January 31, 2020. This work was supported in part by the National Key Research and Development Program under Grant 2016YFC0104608 and in part by Shanghai Jiao Tong University Medical Engineering Cross Research Funds under Grant YG2017ZD10. *(Liang Zhao and Jun Zhao contributed equally to this work.) (Corresponding authors: Liang Zhao; Jun Zhao.)*

Y. Chen is with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with SenseTime Research, Shanghai 200232, China (e-mail: chenyzsjtu@outlook.com).

Y. Gao and L. Zhao are with SenseTime Research, Shanghai 200232, China (e-mail: gaoyunhe@sensetime.com; zhaoliang@sensetime.com).

K. Li is with the Orthopaedics Department, New Jersey Medical School, Rutgers, The State University of New Jersey, Newark, NJ 07103 USA, and also with the Departments of Computer Science and Biomedical Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854 USA (e-mail: kl419@rutgers.edu).

J. Zhao is with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: junzhao@sjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2927289

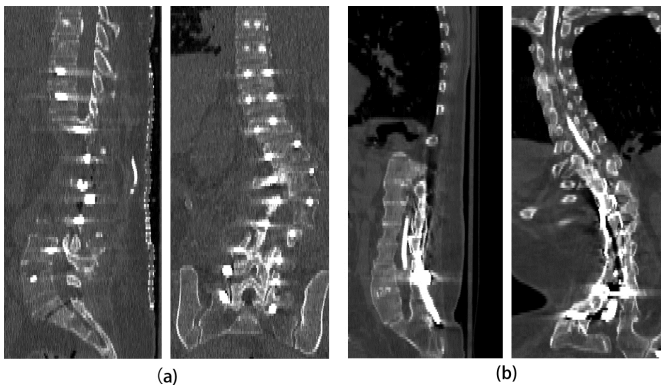

Fig. 1. Sagittal and coronal plane of two difficult cases. (a) A patient suffering from severe scoliosis and lumbarization who was implanted with many metal rods. (b) A long field-of-view CT scan of a patient with scoliosis and kyphosis.

regression forests and the hidden Markov model (HMM) to detect and localize vertebrae in arbitrary field-of-view CT scans. Glocker *et al.* [9] further proposed a localization algorithm that was built on classification forests and trained on generated dense labels. Kim and Lee [10] incorporated local features with global symmetry to improve the localization accuracy. For vertebrae detection in magnetic resonance (MR) images, Zhang *et al.* [11] proposed a robust framework consisting of hierarchical learning and a local articulated model. These methods achieved remarkable identification performance but still suffered from the limitation of hand-crafted image features and also occasionally suffered from the over-fitting problem due to the over-designed framework.

As deep learning gains increasing popularity in the field of CAD for medical images [12]–[16], several methods that use convolutional neural networks (CNNs) have been proposed. Suzani *et al.* [17] applied a fully connected neural network to predict the relative distance from the voxel to each vertebral centroid. Chen *et al.* [18] applied a deep convolutional neural network to identify the vertebrae type based on the centroid proposals generated from the random forest classifier. These methods leveraged the powerful supervised learning tool of CNN techniques and obtained better performance than the previous classic systems. However, the use of classification networks might have overlooked the global latent coherence of the vertebra sequence. More specifically, the differentiation of selected adjacent vertebrae of similar morphology, especially the thoracic vertebrae, relies on global vertebrae sequence information and cannot be classified in one-by-one manner as in the strategies above.

Recently, the application of the fully convolutional network (FCN) technique has contributed to certain powerful vertebrae identification and localization systems. Yang *et al.* [19] proposed a U-Net-like architecture to directly model the vertebrae centroids as a 27-class (26 vertebrae types and background) segmentation-like problem. The network worked well, but simple Gaussian modeling of the vertebrae centroids could not fully exploit the potential of FCNs. More recently, Liao *et al.* [20] used the combination of an FCN and a recurrent neural network (RNN) to leverage both the short- and long-range context information, resulting in a performance boost. It is known that RNN-enabled architectures are capable of addressing the sequential problem [21]–[23], but in Liao's paper, the RNN solely relied on the FCN outputs and could not directly benefit from the original image details. Therefore, the overall system unavoidably suffered from a certain loss of efficiency.

B. Contributions

The FCN is famous for its record-breaking performance in the field of semantic segmentation, and this success lies in its ability to preserve spatial correspondence between the input and output [24], which makes it eligible for the sequence prediction problem in images. A straightforward solution of exploiting the FCN technique is to generate a dense label map around vertebrae centroids as in [9] and model the problem as a 27-class segmentation task in the large 3-D space. However, due to the prohibitively large amount of graphic processing unit (GPU) memory consumption required by the training procedure for large 3-D volumes, the network can only be trained with cropped volume pieces in a sliding-window manner, which results in loss of long-range contextual information in the original CT volume.

Briefly scaling down the CT volume can alleviate the calculation amount issue. However, lower localization accuracy is expected due to loss of the image details required to differentiate the adjacent vertebrae.

To resolve the contradiction mentioned above, namely, capturing the vertebrae spatial pattern in the large search space, an FCN pipeline is designed to decompose the vertebrae localization problem into two separate subproblems, one local problem and one global problem. In this way, the key FCN module of the pipeline can be trained in an end-to-end manner with full-size volume inputs, which benefit the detection capability of the FCN pipeline. In inference time, only one single forward inference is needed to produce results.

Furthermore, a sophisticated post-processing scheme is applied to reduce the dimension of the localization problem to one dimension and constrain the search space for the vertebrae centroids with a probabilistic graphical model. As a consequence, we establish a powerful and robust vertebrae localization system.

Our main contributions are:

1) End-to-end training at the spine level is proposed to allow the FCN to directly learn the long-range image patterns from full-size CT volumes.
2) A composite FCN pipeline is applied to integrate the local and global information of notably large inputs (*i.e.*, $512 \times 512 \times 700$), thus compensating for the loss of details due to the scaling-down of inputs.
3) Better robustness and interpretation are achieved by applying the HMM optimization on the dimension-reduced 1-D CNN outputs.

II. Methodology

The outline of the proposed vertebrae localization framework is presented in Fig. 2. The framework is composed of two main components, the FCN pipeline and the post-processing

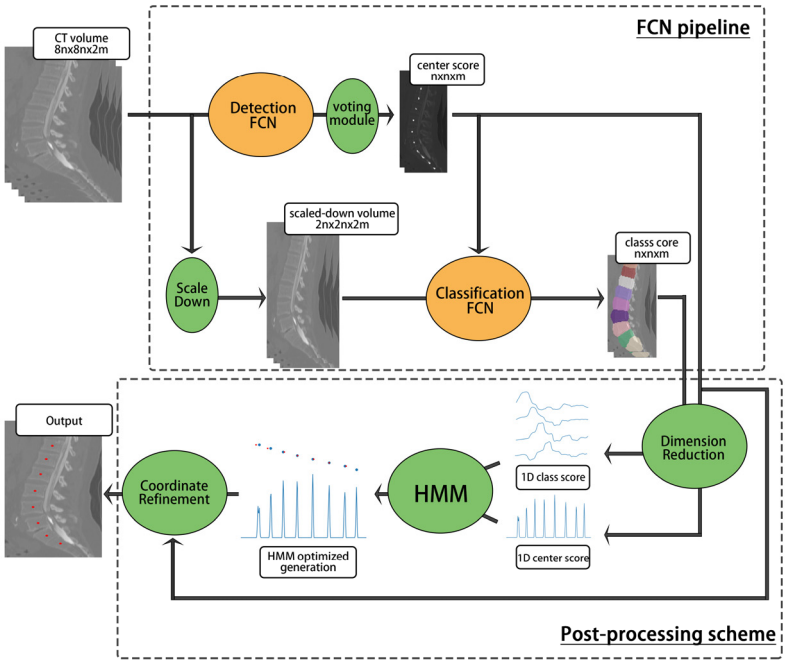

Fig. 2. Overview of the proposed framework for vertebrae identification in the test time. Note that the visual change in the image size of the different components qualitatively implies the volume contraction that occurs during the forward network inference. The scaling down applies only to the first two dimensions of the volume, resulting in a reduction ratio of 1/16. The n and m denote the factor of image size such that 8n and 2m are the length of the original volume along the sagittal (also frontal) and longitudinal axes, respectively.

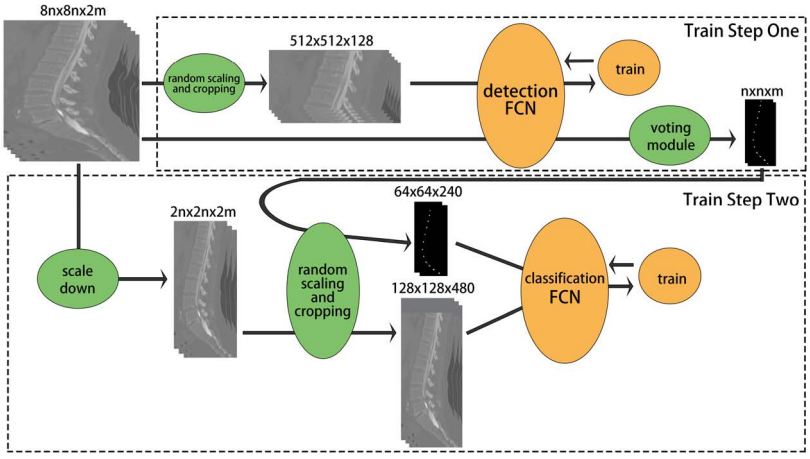

Fig. 3. Illustration of the training procedure for the proposed FCN pipeline. The two networks are trained in order.

scheme, and the FCN pipeline consists of two subnetworks for vertebrae centroid detection and vertebrae classification. First, the raw CT volumes are pre-processed to produce normalized volumes with consistent physical spacing. The resulting volumes are subsequently fed into the FCN pipeline to generate two different network score maps, which are later treated by the post-processing scheme to obtain the final localization result.

Let $V = \{v_c\}$ be the vertebrae centroid set with $v_c \in \mathbb{R}^3$ and $c \in C$, where $C = \{C_1, \ldots, C_7, T_1, \ldots, T_{12}, L_1, \ldots, L_5, S_1, S_2\}$ denote 7 cervical, 12 thoracic, 5 lumbar and the first 2 vertebrae of the sacrum, respectively. The background is denoted as N. Given an image I, the objective is to identify the most likely set of V. It should be noted that the length of V might be less than 26 because we are working with arbitrary field-of-view CT scans.

A. FCN Pipeline

As shown in Fig. 3, the proposed FCN pipeline consists of two independent FCNs. Our objective is to model the vertebrae sequence in the original image space in an end-to-end manner. However, considering the nature of the FCN mechanism, the current goal is to obtain the probability of each voxel in the CT volume belonging to a particular vertebra type, formulated as $P(c|i)$, and $i \in I$ denotes every voxel in the image I.

The main insight of our proposed FCN pipeline is to allow the network to be trained at the spine level. To this

end, two different FCNs are designed for different scales of information. The main FCN, known as the classification FCN, is trained and performs inference at the spine level, focusing on the global-level classification of vertebrae types using scaled-down CT volumes. The other FCN, known as the detection FCN, focuses on vertebrae centroid detection regardless of their types, which relies on fine-grained local image details rather than the global pattern. In this manner, the global sequence information and the local instance information can be combined to train the neural network in a memory-efficient manner, thus making spine-level training available. Fig. 3 illustrates the training procedure for the proposed networks. By decomposing the problem into two subproblems, namely, the vertebrae centroid detection problem and vertebra category classification problem, the proposed pipeline makes better use of the potential of FCN techniques.

During training, the original CT volumes are randomly scaled and cropped to generate fixed-size batches for training the FCNs. However, in the test time, the entire CT volumes are directly fed to the FCN pipeline to avoid troublesome boundary problems because FCNs are naturally capable of handling flexible input sizes.

1) Detection FCN: As mentioned above, the objective of the first FCN is to detect the vertebrae centroids while neglecting their categories, formulated as $P_{detection}(k|i)$, where $k \in \{0, 1\}$ indicates whether the voxel belongs to any vertebra centroids. By directly learning the local-scale image patterns from the original image space, the network is expected to be capable of separating the adjacent vertebrae. An FCN with an encoding-decoding structure, similar to the 3-D U-Net [25]–[27], is used to tackle the problem. The detection network architecture is illustrated in Fig. 4.

The inputs to the network are volume patches ($512 \times 512 \times 128$ during training) randomly scaled and cropped from the training CT volume. For the training labels, it should be emphasized that the annotations of the training dataset are vertebrae centroids coordinates rather than voxel-level segmentation annotations required in the FCNs training. To make use of the supplied coordinate annotations, a network output with 4 channels is carefully designed. An example is shown in Fig. 7.

The network output can be separated into two components, and the generation procedure is described extensively in *Algorithm 1*.

The first component, *i.e.,* the first channel of network outputs, is a label map representing the foreground regions of the vertebrae centroid neighborhoods. This dense label map is used to indicate the neighboring areas of the vertebrae centroids, similar to the dense label map in [9]. Those voxels within a certain range from the vertebrae centroids are tagged as the foreground. However, voxels in the border area between the adjacent vertebrae dense-label regions are excluded because their affiliations are usually ambiguous.

The second component of network outputs, *i.e.,* the last 3 channels, is a coordinate displacement map with 3 channels. To better distinguish the adjacent vertebrae centroids, annotation of the coordinate displacements from each voxel position to the nearest vertebra centroid is generated to form a displacement map. Each of the 3 channels represents one dimension in the 3-D image space. As Fig. 7(c) shows, those voxels around the border area between two adjacent vertebrae on the longitudinal map have scalar values of different signs because the longitudinal axis is the most important among all three axes for adjacent vertebrae differentiation, and the design of displacement map can effectively enhance the network's ability to differentiate adjacent vertebrae centroids.

Algorithm 1 *Pseudo-Code of Coordinate Displacement Generation*

Input: Volume size (X, Y, Z), coordinates of to-be-labelled vertebrae centroids $\{\tilde{x}_j, \tilde{y}_j, \tilde{z}_j\}$.

1. Build a 3D mesh grid of size (X, Y, Z), denoted as $G_i = \{x_i, y_i, z_i\}$
2. For each vertebra centroid j, create a displacement map as $G_{ij} = \{x_i - \tilde{x}_j, y_i - \tilde{y}_j, z_i - \tilde{z}_j\}$, and a distance map as $D_{ij} = \{|x_i - \tilde{x}_j, y_i - \tilde{y}_j, z_i - \tilde{z}_j|\}$
3. For each position i, find out which distance map has the smallest distance $N_i = \{argmin_j (D_{ij})\}$
4. Allocate the position i to the corresponding displacement map, producing the final coordinate displacement map as G_{iN_i}.
5. For each vertebra j, find out the voxels within a provided distance threshold from the vertebrae centroid, denoting as $M_j = \{D_{ij} < d_j\}$.
6. Erode each M_j by a voxel, and then merge all the M_j, resulting in the foreground label map M.

Output: Foreground label map M, coordinate displacement map G_{iN_i}.

The loss for the network training is calculated as the combination of the cross-entropy loss for dense label maps and the weighted mean square loss for coordinate displacement maps and is formulated as follows:

$$\mathcal{L} = \beta \sum_{i \in I} \| \alpha \left(x_i - x_i^{'} \right) \|^2 - \sum_{i \in I} f_i \log \left(f_i^{'} \right) \tag{1}$$

where $x_i, x_i^{'}$ is the ground-truth, predicted coordinate displacement for the voxel I; $f_{ij}, f_{ij}^{'}$ is the ground-truth, predicted dense label map; α is a weighting vector used to enhance the differentiation ability along the longitudinal axis; and β is a scalar applied to balance the two loss functions.

As described in *Algorithm 2*, an explicit vertebrae centroid representation is generated by a voting module from the implicit detection FCN output. Note that the generated score map is now only one channel rather than four channels.

2) Classification FCN: The second FCN performs vertebrae indexing under the global contiguity constraint. The scenario is modeled as a 27-class semantic segmentation problem, denoted as $P_{classification}(c|i)$. The classification network is the core module for our proposed framework because it handles the most critical task, namely, the vertebrae indexing problem in the large 3-D image space. With the help of the preceding detection FCN, the proposed classification FCN can focus on seeking the global patterns rather than the short-range image details.

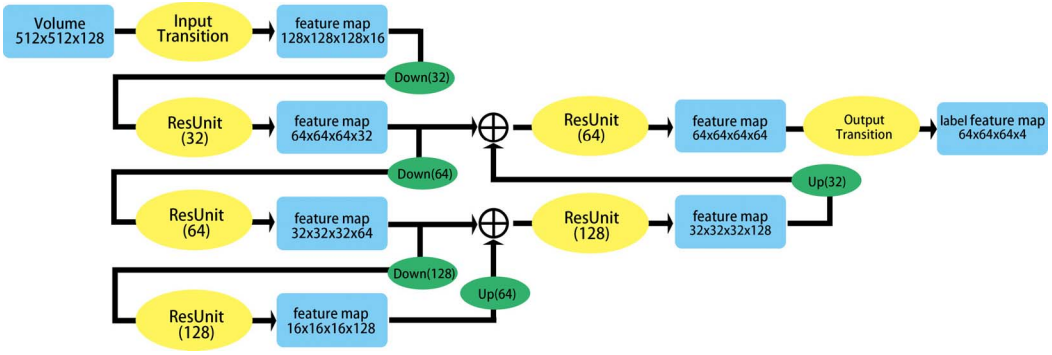

Fig. 4. Structure of the detection FCN. The blue components represent the data units, and the yellow components denote the processing functions. The input transition function includes two $5 \times 5 \times 5$ convolutional layers with a stride of $2 \times 2 \times 1$, each of which are followed by a batch norm (BN) layer and a rectified linear unit (ReLu) activation layer. The output transition function includes two $3 \times 3 \times 3$ convolutional layers each followed by BN and ReLu layers (the last layer does not have any activation layers). The down function is a $3 \times 3 \times 3$ convolutional layer with a stride of $2 \times 2 \times 2$, and the up function is a $2 \times 2 \times 2$ transposed convolutional layer with a stride of $2 \times 2 \times 2$. The ResUnit function is elucidated in detail in Fig. 6. The plus sign indicates concatenation.

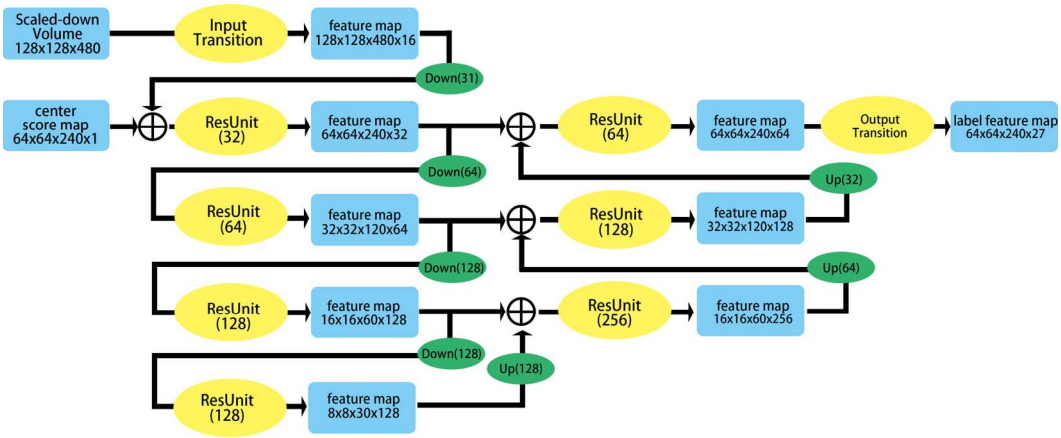

Fig. 5. Structure of the classification FCN. The definitions of the components are largely the same as those in Figure 4, except that the input transition is only a $5 \times 5 \times 5$ convolutional layer with a stride of 1 followed by an instance normalization (IN) layer and a ReLu layer, and the last convolutional layer of the output transition function is appended with a softmax activation.

Algorithm 2 *Pseudo-Code of Voting Module*

Input: Foreground label score map $M_i^{'}$, coordinate displacement score map $G_i^{'} = \{(\Delta x_i, \Delta y_i, \Delta z_i)\}$

1. Foreground label map is generated by thresholding $M_i^{'}$.
2. For each voxel i of coordinate (x_i, y_i, z_i) in the generated foreground label map, calculating its voting position as $v_i = (x_i - \Delta x_i, x_i - \Delta x_i, x_i - \Delta x_i,)$.
3. Centroid score map C is initialized as an empty volume of the same size with $M_i^{'}$.
4. For each v_i, add 1 to the corresponding position in C.
5. C is normalized.

Output: Centroid score map C.

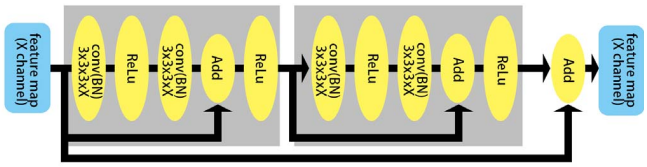

Fig. 6. Structure of the ResUnit function. The X in the ResUnit(X) indicates the number of channels of the convolutional filters. Note that BN layers are replaced by IN layers in the classification network.

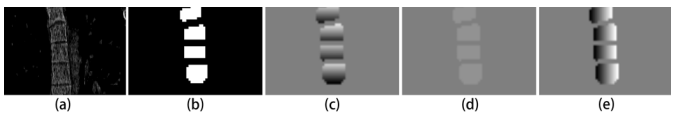

Fig. 7. Example of training data for the detection FCN on the coronal plane. (a) Original CT scan, and (b) foreground dense label map. (c), (d), and (e) are the coordinate displacement maps on the longitudinal, frontal and sagittal axes, respectively.

As in Fig. 5, another FCN with the same encoding-decoding structure is applied. This network is deeper and wider to enlarge the receptive field of the neurons and to strengthen the fitting capability of the network. It should be emphasized that the longitudinal dimension of the input data is set to 480 during training, so that even a volume as long as 640 in the longitudinal axis can be entirely imported into the network after a scaling up of 1.3 in the training time data augmentation.

In the training time, the inputs to the classification FCN are large volume patches ($128 \times 128 \times 480$) cropped from the scaled-down CT volume, and the correspondingly cropped

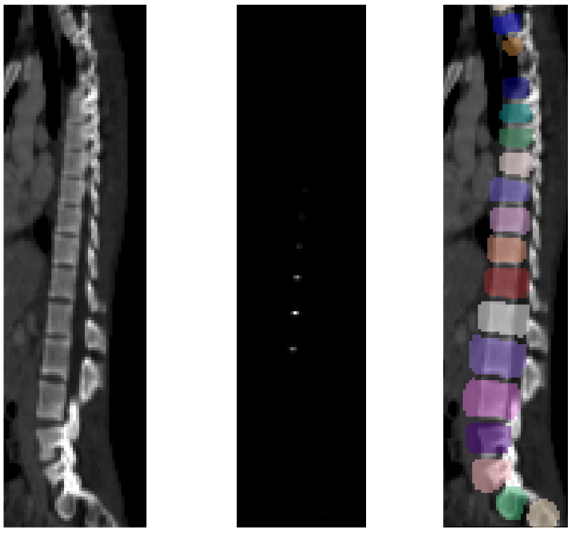

Fig. 8. A slice of a training data example for the classification FCN on the sagittal plane. (a) Inputting scaled-down CT volume. (b) Inputting vertebra centroid score map (only a subset of the vertebrae centroids could be viewed in one particular coronal slice). (c) Output classification labels for different vertebrae (visualized as the combination of all 27 channels with different colors) [28].

vertebra centroid score maps from the detection FCN. A long cropping length in the longitudinal axis allows the FCN to learn to leverage the global sequence information. After scaling in the train-time data augmentation, even a CT volume with a longitudinal length of 600 can be entirely input to the network, which enables network training at the spine level rather than the vertebrae level.

The network output has 27 channels. A dense label map similar to that in the detection FCN, but with an output of 27 different channels for the 26 vertebrae types and the background, is generated as the training label. First, the distributed eroded regions M_j for each vertebra are generated as in *Algorithm 1*, and M_j is considered as the mask for channel j. The channel 0, also known as the background channel, is filled with the complementary mask. All of the remaining channels representing irrelevant vertebrae are filled with empty masks.

It should be emphasized that each vertebra label representation is now stored in one channel of the network output. As noted above, the vertebrae classification in this work is considered as a 27-class semantic segmentation problem to maximize the FCN potential to encode spatial correspondences. An example of the training data is shown in Fig. 8.

The batch normalization layers in the network are replaced with instance normalization layers (IN) [29]. As the price of the network design with large inputs and heavy structures, the batch size of the network training is much smaller than the regular settings. In contrast with the larger batch size of the training of detection FCM, the batch size utilized to calculate normalization statistics is set to 1 in the case of classification FCN, making the application of the mini-batch mean and variation in batch normalization redundant. In inference, the BN layers use the mini-batch mean and variation calculated by averaging the statistics during training, which deteriorates the network performance because the batch in this problem consists of only one instance. Thus, the introduction of IN allows a more reasonable normalization because the normalization statistics are calculated per instance during both training and inference, which is better suited to training of the classification FCN.

The FCN is trained with a simple weighted multiclass cross entropy loss with the addition of the *l2* regularization loss.

Algorithm 3 *Pseudo-Code of Dimension Reduction*

Input: Centroid score map C, classification score map S.

1. Pre-masking.
   a) Vertebra type for each voxel is determined by adopting the one with highest score in S.
   b) Voxels belonging to 26 vertebrae are merged to form a foreground map X.
   c) C is masked by the generated X to get $C^{'}$
2. $C^{'}$ is under threshold to get centroid points, denoted as $\{c_i\}$. SVR is applied to $\{c_i\}$ on the x-z, y-z coordinates, respectively, to get the fully-interpolated spinal curve. The spinal curve in 1-D space is denoted as $\bar{I} = \{\bar{i}\}$.
3. For each point $\bar{i}$, tangent direction in the original 3-D space along the spinal curve is calculated.
4. For each point $\bar{i}$, centroid score is calculated by averaging the centroid scores of points on the vertical plane of i along the tangential direction.
5. For each point $\bar{i}$, the 27-channel classification score is calculated by averaging the 27-channel classification scores of points on the vertical plane of i along the tangential direction.

Output: 1-D spinal curve $\bar{I} = \{\bar{i}\}$, 1-D centroid score map, 1-D classification score map.

B. Post-Processing Scheme

One drawback of applying FCNs to vertebrae identification is that the network outputs are volumetric score maps rather than structured information of the vertebrae coordinates. Therefore, a post-processing procedure is required to extract the desired quantities from the score maps. The framework we propose includes a post-processing scheme composed of dimension reduction, HMM optimization and coordinate refinement. The network outputs are first transformed into 1-D arrays and modeled by an HMM to generate the optimized 1-D vertebrae coordinates. The generated 1-D coordinates are transformed back to 3-D coordinates and refined by the original FCN output.

1) Dimension Reduction: To narrow the search space for the problem, dimension reduction from 3-D to 1-D is applied. Specifically, all of the score maps are re-interpolated along the 1-D spinal curve. Instead of the easier modeling along the longitudinal axis, modeling along the spinal curve is more reasonable and practical.

The dimension reduction procedure is described in *Algorithm 3*, which is the start module for the whole

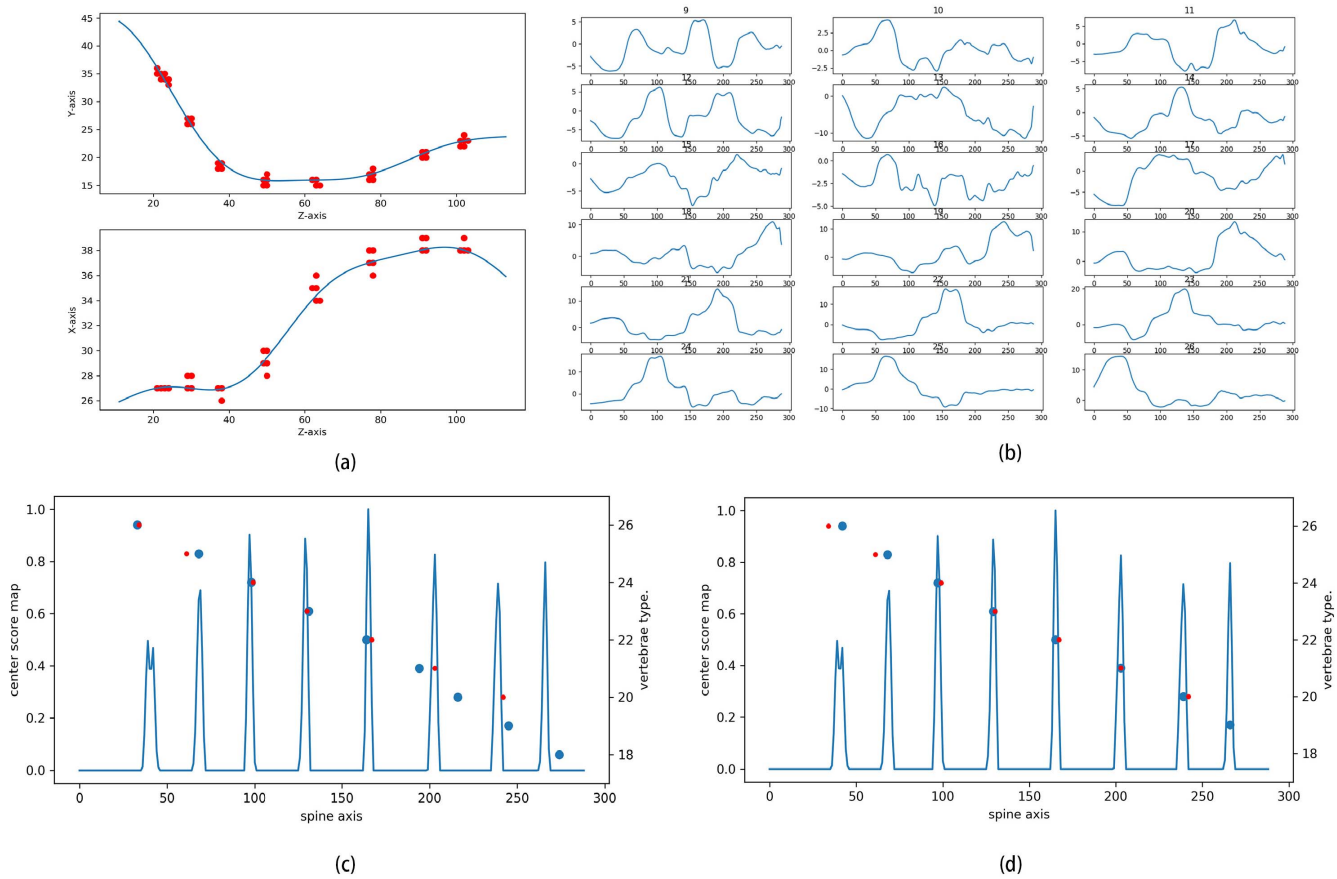

Fig. 9. Example of the post-processing procedure. (a) SVR of a spinal curve. The red dots indicate the vertebrae centroids obtained from thresholding of the detection FCN score map. The blue lines in the upper and lower image represent the fitting spinal curve on the Y-Z plane and the X-Z plane, respectively. (b) Interpolated classification score maps (before softmax) for 12 thoracic, five lumbar and first two sacral vertebrae. The number above each small figure is the affiliated vertebrae index for the figure. The horizontal axis gives the position on the fitted spinal curve, and the vertical axis is the interpolated logit score generated from the classification FCN score map. (c) and (d) The curves are the graphs of the interpolated centroid score maps (after softmax) derived from the detection FCN outputs. The red dots represent the ground-truth vertebrae positions, and the blue dots denote the predictions. It should be noted that the ordinate of the dots is used only to distinguish the different vertebrae types. (c) and (d) were recorded before and after the HMM optimization. It can be observed that the HMM optimization corrected the failure of the classification FCN.

post-processing pipeline and is of great importance. The robustness is guaranteed by two safety belts. Pre-masking of the centroid score map using the classification score map ensures that the centroid points are calculated using all of the available information. However, the application of support vector regression (SVR) [30] makes the spinal curve modeling robust to the centroid point outliers.

Fig. 9(a) shows an example of dimension reduction. Let $\bar{I} = \{\bar{i}\}$ be the voxels along the calculated spinal curve, and thus we obtain $P_{classification}(c|\bar{i})$ and $P_{detection}(k|\bar{i})$.

2) HMM Optimization: Sequential and spatial constraints on the vertebrae categories are implicitly exerted during the classification FCN training. However, failures occur when encountering certain difficult cases. In this work, the restraints of physical distance and the sequential pattern are imposed by an HMM [31].

To simplify the problem, only the vertebrae that probably exist in the volume are taken into consideration. Let $\bar{C} = \{\bar{c}_0, \ldots, \bar{c}_n, \bar{c}_{-1}\}$ be the vertebrae considered, where $\bar{c}_{-1}$ is the background, and $\bar{c}_0$ is the vertebra before the headmost vertebra present in the classification FCN output if it is not already the first vertebra of cervical vertebrae. The probabilistic model can be formulated as follows:

$$P_{HMM} = \prod_{\bar{i}\in\bar{I}} P_{emission}\left(A_{\bar{i}}|v_{\bar{i}} = \bar{c}_j\right) * P_{transition}(v_{\bar{i}} = \bar{c}_j) \tag{2}$$

where $A = \{A_{\bar{i}}\}$ denotes the vertebrae classification prior for each point in $\bar{I}$. The $\mathrm{V} = \{v_{\bar{i}}\}$ are the point states along the calculated spinal curve. The best configuration of hidden states for each point is found by maximizing the HMM function P_{HMM}. The transition function and emission function are calculated as shown:

$$P_{emission}\left(A_{\bar{i}}|v_{\bar{i}} = \bar{c}_j\right) = \frac{P_{classification}\left(v_{\bar{i}} = \bar{c}_j|\bar{i}\right)}{\max\limits_{\bar{c}_j\in C} P_{classification}\left(v_{\bar{i}} = \bar{c}_j|\bar{i}\right)} \tag{3}$$

$$P_{transition}\left(v_{\bar{i}} = \bar{c}_j\right) = P_{detection}\left(k = 1|\bar{i}\right) + \lambda P_{distance}\left(v_{\bar{i}} = \bar{c}_j\right) \tag{4}$$

where

$$P_{distance}(v_{\bar{i}} = \bar{c}_j) = \begin{cases} \left(1 - \dfrac{|L(\bar{i}, \bar{i}_{\bar{c}_{j-1}}) - P_{stan}(\bar{c}_j, \bar{c}_{j-1})|}{P_{stan}(\bar{c}_j, \bar{c}_{j-1})}\right)^3, & \bar{i} > \bar{i}_{\bar{c}_{j-1}} \\ 0, & \bar{i} \leq \bar{i}_{\bar{c}_{j-1}} \end{cases} \quad (5)$$

$$P_{stan}(\bar{c}_j, \bar{c}_{j-1}) = \mathbb{E}_{\Omega}\left(L\left(\bar{i}_{\bar{c}_j}, \bar{i}_{\bar{c}_{j-1}}\right)\right) * \frac{L\left(v_{\bar{c}_1}, v_{\bar{c}_n}\right)}{\mathbb{E}_{\Omega}\left(L\left(\bar{i}_{\bar{c}_1}, \bar{i}_{\bar{c}_n}\right)\right)} \quad (6)$$

where $L(\alpha, \beta)$ denotes the length of the spinal curve between the two voxels α and β, $\bar{i}_{\bar{c}_j}$ denotes the position of the hidden state $\bar{c}_j$, and $\mathbb{E}_{\Omega}(L(\alpha, \beta))$ denotes the average of $L(\alpha, \beta)$ in the training dataset Ω. It should be noted that $L(v_{\bar{c}_1}, v_{\bar{c}_n})$ is calculated based on the initial classification output.

Given a calculated spinal curve $\bar{I}$, the hidden states $\{v_{\bar{c}_j}\}$ can be inferred by maximizing the P_{HMM} likelihood function via dynamic programming.

3) Coordinate Refinement: After the vertebrae locations are identified on the spinal curve, their 3-D coordinates in physical space can be easily obtained by mapping the spine axis back to the original 3-D space. However, considering the fact that the spinal curve is only an approximate model of the real spine, the vertebrae coordinates can be further refined using the original detection FCN output.

First, we locate a local region for each vertebra according to the previously obtained coordinate from the preceding HMM optimization results. Second, we calculate the new centroid position by averaging the detection votes score within the local region. The coordinate of the newly calculated centroid position is the final result for a particular vertebra.

III. Experiments

The framework that we proposed is built, tested and evaluated on a public dataset from the MICCAI Challenge.[1] The dataset contains 242 spine-focused CT training scans containing various types of high-grade pathologies and metal implants, together with 60 other scans for hold-out evaluation. All of the CT images are resampled under spacings of 1.0 mm, 1.0 mm and 1.25 mm along the sagittal, frontal and longitudinal axes, respectively.

For direct comparison with previously published works, the performance of our proposed framework is evaluated using both cross-validation on the training data and independent testing on the test data.

A. Evaluation Metrics

Several metrics are used to quantitatively evaluate the localization performance.

Identification rate (*ID Rate*). A vertebra is considered to be correctly identified if its estimated centroid is within 20 mm of the ground truth, and the closest inferred centroid is the correct centroid, as in [9]. The *ID Rate* has been widely applied as the metric for vertebrae identification in the published literature. However, the distance between two neighboring vertebrae is approximately 20 mm on average and is smaller than 40 mm in most cases. Therefore, the distance between an annotation and its closest correct detection is mostly approximately 10 mm, which makes the constraint of 20 mm too generous. In order to provide a comprehensive understanding of localization performance, a new evaluation metric is introduced below.

Mean identification rate (*Mean ID Rate*). The measure is the average of the *ID Rates* under distance thresholds from 1 to 20 mm with a step size of 1 mm and removal of the constraint that the closest one must be the correct one. This metric is introduced because the shortened distance threshold for the identification rate can be more reasonable for quantitative evaluation of a vertebrae localization system.

Localization error (in mm). Localization error is defined as the average distance (in mm) of the predicted vertebrae centroids from the corresponding ground truth.

TABLE I
COMPARISON WITH THE PUBLISHED METHODS

	CROSS VALIDATION ON THE TRAINING SET			INDEPENDENT TEST ON THE TEST SET		
	Mean Error	Std of Error	ID Rate	Mean Error	Std of Error	ID Rate
SUZANI *ET AL.* [16]	18.2	11.4	-	-	-	-
GLOCKER *ET AL.* [9]	12.4	11.2	70%	13.20	17.80	74.00%
CHEN *ET AL.* [17]	-	-	-	8.82	13.04	84.16%
YANG *ET AL.* [18]	9.10	7.20	80.0%	8.60	7.80	85.00%
LIAO *ET AL.* [19]	-	-	-	6.47	8.56	88.30%
PROPOSED	**4.33**	**6.31**	**88.33**	**2.56**	**3.15**	**94.67%**

B. Implementation Details

The FCN model is implemented on the deep learning framework of PyTorch [32]. Each of the FCNs in the pipeline uses four NVIDIA GTX 1080Ti GPUs for training.

The α, β values for the detection FCN loss function are set to (1, 1, 6), 50. The λ in the HMM module is set to 0.5. The learning rates of two FCNs are both initialized as 0.1 and are changed to 0.01 after 200 epochs.

Data augmentation is an important module of the network training. For the detection network, data augmentation includes translating, scaling, flipping, Gaussian noise polluting, speckle noise polluting and synthesized metal implanting. In particular, synthesized metal implanting places bright cubes as large as a vertebra in random vertebrae positions to simulate metal occlusions similar to those in real CT images. For the classification network, data augmentation is much more straightforward, including translating, scaling and flipping. The inputs to the network are scaled in a ratio from 0.8 to 1.2, so that during

[1] http://csi-workshop.weebly.com/challenges.html

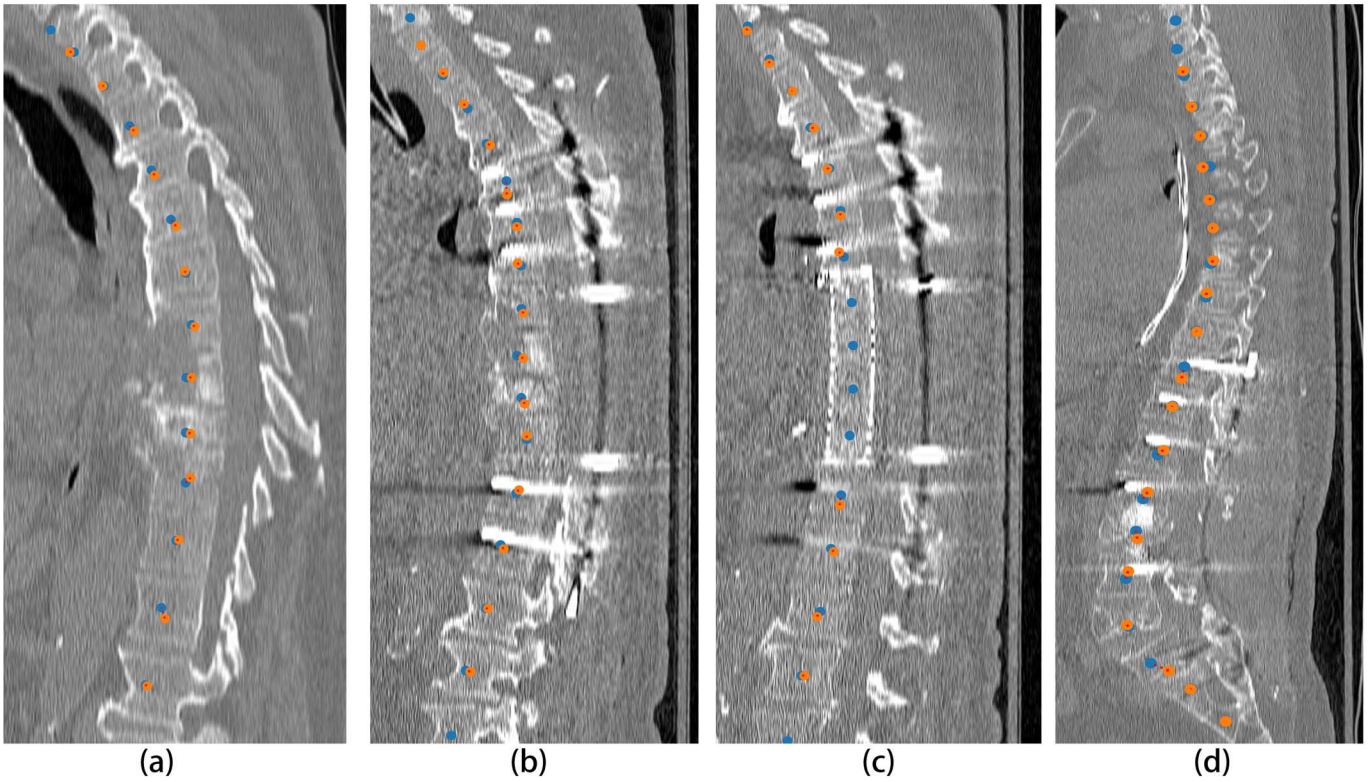

Fig. 10. Selected examples of the final results on the transverse plane. The orange dots are the ground truth of the vertebrae centroids, and the blue dots are the predictions of the proposed framework. Each of the two points corresponding to the same vertebra are linked with a red line. The pictures are of the sagittal plane, and thus sagittal coordinates of vertebrae identifications are deliberately ignored such that every identification can be painted on the figure. (a), (b) and (c) are acquired from the same patient at different times. (d) is acquired from another patient. The blue dots are the results calculated with our proposed model.

TABLE II
DETAILED COMPARISON OF THE SPECIFIC VERTEBRAE CATEGORIES WITH INDEPENDENT TESTS

	Glocker *et al.* [9]			**Chen *et al.*** [17]			**Yang *et al.*** [18]			**Liao *et al.*** [19]			**Proposed**		
	Mean Error	Std of Error	Id Rate	Mean Error	Std of Error	Id Rate	Mean Error	Std of Error	Id. Rate	Mean Error	Std of Error	Id Rate	Mean Error	Std of Error	Id Rate
All	13.20	17.83	74.0%	8.82	13.04	84.2%	8.60	7.80	85.0%	6.47	8.56	88.3%	**2.56**	**3.15**	**94.7%**
Cervical	6.81	10.02	88.8%	5.12	8.22	91.8%	5.60	4.0	92.0%	4.48	4.56	**95.1%**	**2.50**	**3.66**	89.5%
Thoracic	17.35	22.3	61.8%	11.39	16.48	76.4%	9.20	7.90	81.0%	7.78	10.17	84.0%	**2.63**	**3.25**	**95.3%**
Lumbar	13.05	12.45	79.9%	8.42	8.62	88.1%	11.0	10.8	83.0%	5.61	7.68	92.2%	**2.19**	**1.82**	**100.0%**

training, even a CT volume with a longitude length as large as 600 can be inputted as a whole after scaling.

C. Performance

Selected successful results from the trained system are demonstrated in Fig. 11. The trained model is capable of locating vertebrae under complex surroundings and overcoming various types of vertebra fractures and metal implants.

To further quantitatively assess the performance, a comparison with other competing methods evaluated on the same public dataset of the MICCAI challenge is shown in Table I. Moreover, the performance for all vertebrae types, together with the performance for different vertebrae categories (cervical, thoracic and lumbar), is presented in Table II. As shown in the table, the proposed method outperforms all of the other methods in almost all of the measurements, reducing the localization error from 9.10 ± 7.20 to $\mathbf{4.33 \pm 6.31}$ with cross-validation on the training dataset and from 6.47 ± 8.56 to $\mathbf{2.56 \pm 3.15}$ with independent testing on the test dataset. The *ID Rate* is boosted from 80.0% to **88.33%** with cross-validation and from 84.16% to **94.67%** with independent testing. In the later section shown in Table III, we also note that even with minimal post-processing efforts (no HMM module), the single FCN pipeline could still achieve better performance than the other methods. The significant performance improvements prove the feasibility of the proposed methodology.

D. Ablation Studies

In this section, ablation studies are conducted to assess the influence of different modules on the system performance. The model difference between the proposed system and previous ones [17]–[19] mainly lies in the scaled inputs coupled with

TABLE III
COMPARISON OF THE MODEL PERFORMANCE IN THE ABLATION STUDIES

	Proposed			Single FCN			Without HMM			Without Refinement			Without IN		
	Mean Error	Std of Error	Mean Id Rate	Mean Error	Std of Error	Mean Id Rate	Mean Error	Std of Error	Mean Id Rate	Mean Error	Std of Error	Mean Id Rate	Mean Error	Std of Error	Mean Id Rate
All	**2.56**	**3.15**	**88.0%**	5.26	5.11	75.5%	3.56	5.39	83.9%	2.91	3.17	86.2%	2.79	4.08	84.4%
Cervical	**2.50**	**3.66**	**85.2%**	4.77	4.38	75.1%	3.07	4.63	83.3%	3.05	3.80	82.4%	2.55	4.03	81.8%
Thoracic	**2.63**	3.25	**88.6%**	5.36	4.90	75.4%	3.55	5.04	84.1%	2.84	**3.22**	87.6%	3.09	3.75	83.5%
Lumbar	2.19	**1.82**	**91.5%**	5.21	6.65	79.4%	3.80	7.19	86.9%	2.61	1.96	89.6%	**2.12**	**1.82**	90.8%
Sacrum	**3.42**	2.31	**85.9%**	7.33	4.25	65.5%	5.67	4.79	74.1%	3.78	**1.87**	83.4%	3.44	2.33	85.8%

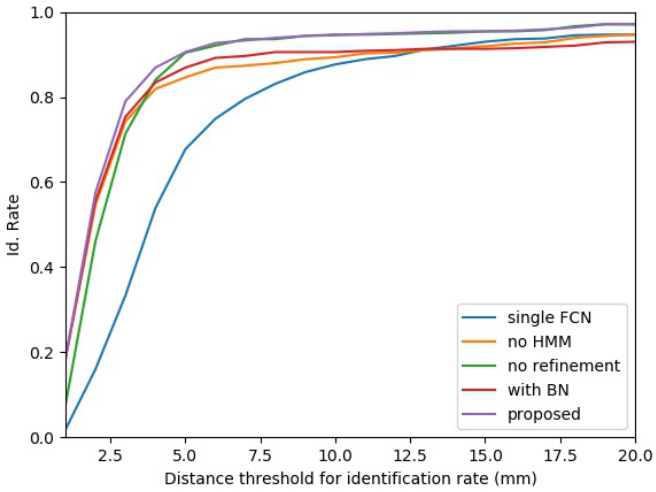

Fig. 11. Plots of *Id. Rates* under different distance thresholds. *Id. Rate* here is calculated without the constraint that the nearest vertebra must be the right vertebra, which makes it different from *Id. Rate* of Table II.

the composite network design, the post-processing scheme, and the improved training with IN replacement. Fig. 11 shows the comparison plots of *Id Rate* (without the constraint that the closest one must be the correct one) calculated with different distance thresholds. Table III records the detailed statistics.

1) Role of the Detection FCN: To assess the role of the detection FCN, the FCN pipeline is replaced with one single FCN similar to the classification FCN, except that the single FCN does not include the second input of the vertebrae centroid score map. The simple FCN model accepts only the scaled inputs and outputs the same classification score map as the one produced by classification FCN. In this case, no detection score map is generated, thereby disabling the original post-processing scheme. Thus, final vertebrae localization results are instead obtained by directly calculating the centroids of the predicted masks for each vertebra type.

As Table III and Fig. 11 indicate, without the local vertebrae centroids information from the detection FCN, the single FCN model performance deteriorates, especially under the smaller distance thresholds. It demonstrates the necessity of incorporating the local image details and proves that the network design of FCN composition does overcome the shortcomings brought by the application of image scaling to network inputs.

The *ID Rate* (under 20mm threshold and with constraint that the nearest vertebra must be the right vertebra) of the single FCN model is 90.87%, slightly higher than the previous state-of-the-art model [20]. However, the single FCN model is much simpler, with little post-processing techniques. The achieved comparable performance by a much simpler model is believed to stem from the difference that the single FCN is trained in an end-to-end manner at the spine level rather than at the vertebrae level as in the previous models [9], [17]–[19].

In addition, the performance gap between the single FCN model and the proposed model is the largest of all the gaps between models as shown in the Fig. 11, which implies that the composite FCN structures contribute mostly to the performance gains of the proposed model with respect to all the listed modules.

Together, the combination of detection FCN and classification FCN allows better network inference at the spine level incorporating both the long-range and short-range image information.

2) Role of the HMM Module: A post-processing scheme without the HMM module is tested, and the other settings remain the same. It can be observed from Fig. 11 and Table III that the network performance also deteriorates, but in this case, it worsens with the larger distance threshold.

The evaluation result confirms our assumption that application of HMM module can globally benefit the system performance by imposing external constraints with the proposed HMM. However, it should also be noted that the improvement in the HMM module on the average performance of the mean error is not significant. If the preceding network component can be further strengthened, a solution without the HMM module is surely feasible, considering that a simpler system is always more advantageous.

3) Role of the Coordinate Refinement Module: A post-processing scheme without the coordinate refinement module is evaluated. It can be easily observed that the module comprehensively reduces the mean localization error and that it improves the system performance over *Id Rate* with small distance thresholds, which substantiates our hypothesis that the true vertebrae centroids line is not a technically smooth curve.

4) Role of Instance Normalization: A simple experiment is carried out to assess the contribution of the IN layers. Normalization layers are set to BNs instead of INs, and it is observed that *mean ID Rate* falls to 84.4% from 88.0%, which is a considerable decline. It implies that the small batch size during the network training does degrade the network performance.

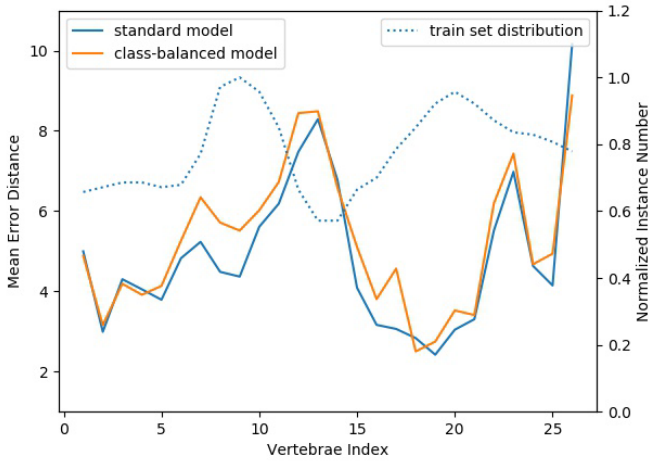

Fig. 12. Comparison of CNN performance between the standard model and the class-balanced model. The left y-axis indicates the mean error distance of each vertebra. The right y-axis indicates the normalized vertebrae instance number.

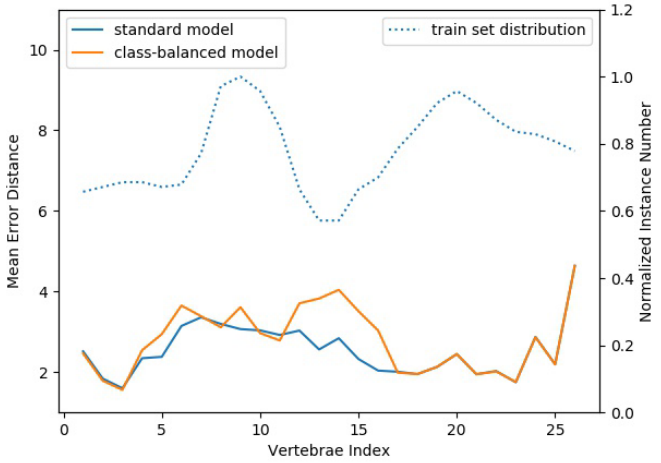

Fig. 13. Comparison of the overall system performance between the standard model and class-balanced model. The axes are defined in the same way as Fig. 12.

5) Role of Data Augmentation: As for data augmentation, on-the-fly data transform during classification FCN training are set to minimal, only with cropping along the longitudinal axis. It is found that the system performance remains of the same level. Overall, the fact that network trained with a very small training dataset and minimal data augmentation can achieve almost the same performance is impressive.

E. Label-Balance Experiment

The training dataset is unbalanced for different vertebrae types, as shown in the train set distribution curve of Fig. 13 and Fig. 14. For the current model, no strategy is applied to address the class imbalance problem under the assumption that the data are not sufficiently unbalanced to be a concern.

An experiment is carried out to examine this assumption. We utilize the common class-balancing strategy of applying the class-balanced loss function. In detail, the weighted cross-entropy loss function is applied to strengthen the loss of less usual vertebrae. The loss of each vertebra class is multiplied with the inverse of the normalized vertebrae instance number in Fig. 12.

Fig. 12 and Fig. 13 illustrate the performance comparison between the proposed model and the class-balanced model. The performance in Fig. 12 is calculated on the CNN outputs with only the simplest post-processing (as in the first ablation study in the manuscript), and the performance in Fig. 13 is calculated on the final system outputs with all the proposed post-processing. After the application of class-balancing module, it can be observed that the performance of the system with simplest post-processing deteriorates slightly, and that the performance of the system with proposed post-processing is of the same level.

With respect to the overall performance of the proposed system, the system performance after adding the class-balancing module remains at the same level. Thus, it is believed that for the proposed scheme the training data are not too unbalanced to be a problem.

F. Evaluation on the External Dataset

The proposed algorithm is further tested on the dataset for the segmentation challenge of the CSI 2014 Workshop [33]. It must be noted that the dataset is a segmentation dataset rather than an identification dataset. The provided labels in the dataset are not vertebrae centroids coordinates as in normal vertebrae identification dataset, but are voxel annotations of vertebrae segmentation. However, the provided segmentation label is of instance type, so that the dataset can be used to evaluate the performance of vertebrae identification system. With higher image quality and lower abnormal complexity, the difficulty in identification of this dataset is considered lower. The main obstacle for vertebrae detection lies in the long longitudinal volumetric length and the vertebrae fracture abnormality.

The dataset contains 20 long-range upper-body CT volumes centered on the spine, all of which are evaluated using the previously trained proposed model (without any further parameter tuning). It is found that all the vertebrae centroids are correctly detected by the proposed algorithm based on manual inspection on every output. Fig. 14 shows two examples with multiple vertebrae fractures. Although the data is relatively simpler, we believe that the successful application of the proposed system to an unseen external dataset shows the robustness of our algorithm to a certain extent.

IV. Discussion

The largest difference between the proposed work and previous work is that the key module of the proposed model, the classification FCN, is trained in an end-to-end manner at the spine level rather than at the vertebrae level. Further applications of novel CNN techniques, such as the attention module [34]–[36], deep supervision [37] or adversarial network [38], [39], may be able to strengthen the system. However, considering the current limited size of available data, the improvement is assumed to be minor if no more training data are supplied.

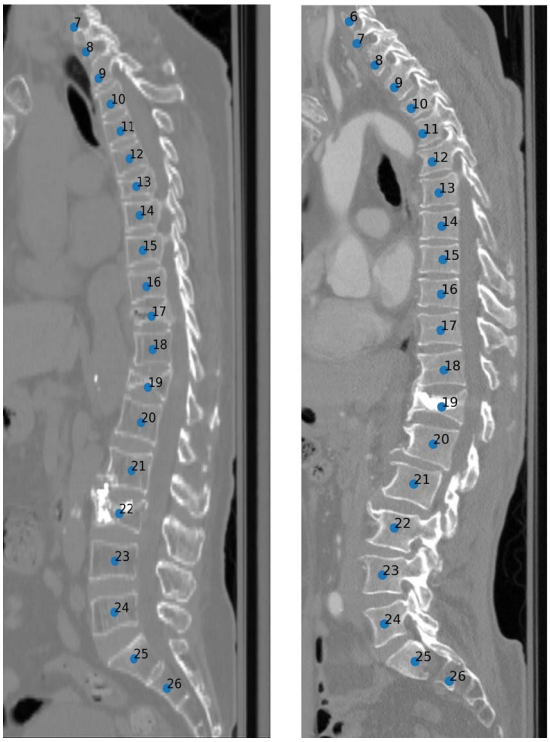

Fig. 14. Evaluation examples of the external dataset from the CSI 2014 Workshop with vertebrae fractures. The blue dots are the results calculated with our proposed model. The picture is of the sagittal plane, and the sagittal coordinates of vertebrae identifications are deliberately ignored such that every identification can be painted on the figure. It can be clearly observed that every vertebra is correctly identified.

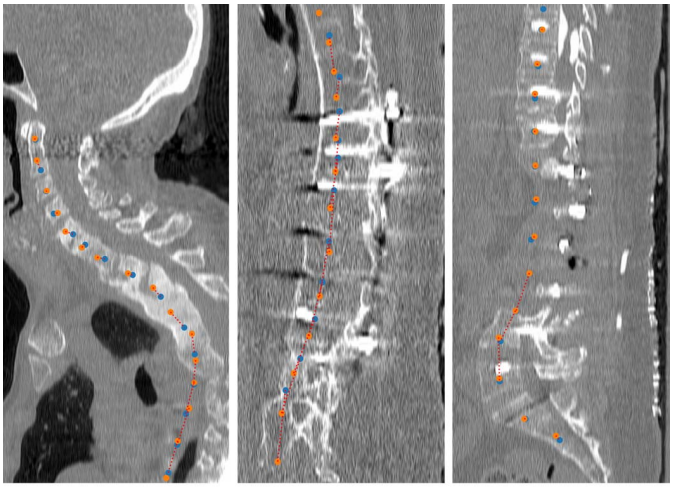

Fig. 15. Examples of failed identifications. (a) Failures could be caused by the lack of data variety in the training dataset, similar to the one in the figure with severe kyphosis and fuzzy bone densities, which is the only case in the dataset. (b) It is also arduous to determine the real vertebrae type when the field of view is restricted inside the thoracic spine which, unlike the other segments, lacks significant geographical landmarks. (c) The design of the classification FCN with fixed output channels makes it difficult to address the lumbarization cases that display one extra lumbar vertebra.

Although the identification rate score of the proposed method appears high, the system is still quite far from perfect. Selected typical failures of the proposed method are shown in Fig. 15. As we know, simulating anomalies is always more difficult than modeling those normal cases because the variation of the former can be much greater. Significant work remains for the construction of a strictly reliable vertebrae localization system.

There exists a potential risk of severe failure from the very beginning of the post-processing module. It is because the dimension reduction from 3-D to 1-D heavily relies on the accurate extraction of the spinal curve, which may corrupt the whole post-processing pipeline in the start. In the previous experiment on the test set of MICCAI dataset, no obvious error occurs in the spinal curve extraction, which confirms our assumption that spinal curve extraction is a relatively easier task. However, algorithm failures are always unavoidable, especially in the sophisticated real-world clinic cases. For now, we assume that results with less than or equal to 3 detected vertebrae classes are not reliable for obtaining an accurate spinal curve. Moreover, the spinal curve could also be presented to the users, and options should be offered to output the CNN results with the simplest post-processing strategies.

In future work, the dilemmas of certain failed cases may be mitigated by introducing additional labeled data of the similar cases. For situations in which extra vertebrae are present, the framework structure may have to be modified to support a flexible output.

Additionally, the generalization capability of the FCN method is somehow surprisingly outstanding. As shown in the left three examples of Fig. 11, the neural network successfully predicts not only the existing vertebrae but also the missing vertebrae that were previously either broken by external forces or were removed in surgery. Note that there is only one data instance with distinctly missing vertebrae in the training dataset. When this example is chosen as the test set in the cross-validation, no other analogous training data exist. Therefore, the network does not have the opportunity to learn to address the cases with missing vertebrae. It can be inferred that the trained network has leant to consider the spatial correspondence of the vertebrae sequence, which again demonstrates the power of CNN techniques.

V. Conclusion

In this paper, we present a novel and high-performance vertebrae identification and localization framework for CT images. The approach combines the FCN technique with an explicit probabilistic graph model to leverage the powerful fitting capability of CNNs and the convenient maneuverability of HMMs. The novelty of this work lies mostly in the design of the FCN pipeline, which allows the system to train the classification FCN at the vertebrae level and to decompose the localization problem into two different tasks that could be respectively resolved at different image scales. Thorough comparisons and extensive experiments on a large public dataset indicate that our design of the FCN pipeline and post-processing scheme substantially improves vertebrae localization accuracy.

ACKNOWLEDGMENT

The authors would like to thank the research team at SenseTime Research, Chang Liu and Xiaofen Li at Shanghai Jiao Tong University for their support.

REFERENCES

[1] I. B. Ayed, K. Punithakumar, R. Minhas, R. Joshi, and G. J. Garvin, "Vertebral body segmentation in MRI via convex relaxation and distribution matching," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.,* 2012, pp. 520–527.

[2] J. Yao, J. E. Burns, H. Munoz, and R. M. Summers, "Detection of vertebral body fractures based on cortical shell unwrapping," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.,* 2012, pp. 509–516.

[3] F. Lecron, J. Boisvert, S. Mahmoudi, H. Labelle, and M. Benjelloun, "Fast 3D spine reconstruction of postoperative patients using a multilevel statistical model," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., 2012, pp. 446–453.

[4] J. E. Burns, "Imaging of the spine: A medical and physical perspective," in *Spinal Imaging and Image Analysis*. Cham, Switzerland: Springer, 2015, pp. 3–29.

[5] J. L. Herring and B. M. Dawant, "Automatic lumbar vertebral identification using surface-based registration," *J. Biomed. Informat.*, vol. 34, pp. 74–84, Apr. 2001.

[6] S. Schmidt *et al.*, "Spine detection and labeling using a parts-based graphical model," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.*, 2007, pp. 122–133.

[7] J. Ma, L. Lu, Y. Zhan, X. Zhou, M. Salganicoff, and A. Krishnan, "Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., 2010, pp. 19–27.

[8] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., 2012, pp. 590–598.

[9] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine CT via dense classification from sparse annotations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., 2013, pp. 262–270.

[10] K. Kim and S. Lee, "Vertebrae localization in CT using both local and global symmetry features," *Comput. Med. Imag. Graph.*, vol. 58, pp. 45–55, Jun. 2017.

[11] Y. Zhan, D. Maneesh, M. Harder, and X. S. Zhou, "Robust MR spine detection using hierarchical learning and local articulated model," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., 2012, pp. 141–148.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[13] Q. Dou, H. Chen, L. Yu, J. Qin, and P.-A. Heng, "Multilevel contextual 3-D CNNs for false positive reduction in pulmonary nodule detection," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1558–1567, Jul. 2017.

[14] L. Xiang *et al.*, "Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image," *Med. Image Anal.*, vol. 47, pp. 31–44, Jul. 2018.

[15] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[16] H. R. Roth *et al.*, "An application of cascaded 3D fully convolutional networks for medical image segmentation," *Comput. Med. Imag. Graph.*, vol. 66, pp. 90–99, Jun. 2018.

[17] A. Suzani, A. Seitel, Y. Liu, S. Fels, R. N. Rohling, and P. Abolmaesumi, "Fast automatic vertebrae detection and localization in pathological CT scans—A deep learning approach," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., Nov. 2015, pp. 678–686.

[18] H. Chen *et al.*, "Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., Nov. 2015, pp. 515–522.

[19] D. Yang *et al.*, "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, May 2017, pp. 633–644.

[20] H. Liao, A. Mesfin, and J. Luo, "Joint vertebrae identification and localization in spinal CT images by combining short- and long-range contextual information," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1266–1275, May 2018.

[21] H. Chen *et al.*, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., Nov. 2015, pp. 507–514.

[22] B. Kong, Y. Zhan, M. Shin, T. Denny, and S. Zhang, "Recognizing end-diastole and end-systole frames via deep temporal regression network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., Oct. 2016, pp. 264–272.

[23] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function," 2017, *arXiv:1707.04912*. [Online]. Available: https://arxiv.org/abs/1707.04912

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., Nov. 2015, pp. 234–241.

[26] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[27] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*., Oct. 2016, pp. 424–432.

[28] P. A. Yushkevich *et al.*, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, pp. 1116–1128, Jul. 2006.

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available:https://arxiv.org/abs/1607.08022

[30] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 155–161.

[31] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1992, pp. 379–385.

[32] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017.

[33] J. Yao *et al.*, "A multi-center milestone study of clinical vertebral CT segmentation," *Comput. Med. Imag. Graph.*, vol. 49, pp. 16–28, Apr. 2016.

[34] O. Oktay *et al.*, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: https://arxiv.org/abs/1804.03999

[35] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 2048–2057.

[36] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*. [Online]. Available: https://arxiv.org/abs/1508.04025

[37] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.,* Oct. 2016, pp. 149–157.

[38] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss," 2018, *arXiv:1804.10916*. [Online]. Available: https://arxiv.org/abs/1804.10916

[39] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.