Spine-GAN: Semantic Segmentation of Multiple Spinal Structures

Zhongyi Han[a,b,c,d], Benzheng Wei[a,b,*], Ashley Mercado[c,d], Stephanie Leung[c,d], Shuo Li[c,d,*]

[a] *College of Science and Technology, Shandong University of Traditional Chinese Medicine, Jinan, SD, China*
[b] *Computational Medicine Lab, Shandong University of Traditional Chinese Medicine, Jinan, SD, China*
[c] *Digital Image Group (DIG), London, ON, Canada*
[d] *Dept. of Medical Imaging, Western University, London, ON, Canada*

Abstract

Spinal clinicians still rely on laborious workloads to conduct comprehensive assessments of multiple spinal structures in MRIs, in order to detect abnormalities and discover possible pathological factors. The objective of this work is to perform automated segmentation and classification (i.e., normal and abnormal) of intervertebral discs, vertebrae, and neural foramen in MRIs in one shot, which is called semantic segmentation that is extremely urgent to assist spinal clinicians in diagnosing neural foraminal stenosis, disc degeneration, and vertebral deformity as well as discovering possible pathological factors. However, no work has simultaneously achieved the semantic segmentation of intervertebral discs, vertebrae, and neural foramen due to three-fold unusual challenges: *1)* Multiple tasks, i.e., simultaneous semantic segmentation of multiple spinal structures, are more difficult than individual tasks; *2)* Multiple targets: average 21 spinal structures per MRI require automated analysis yet have high variety and variability; *3)* Weak spatial correlations and subtle differences between normal and abnormal structures generate dynamic complexity and indeterminacy. In this paper, we propose a Recurrent Generative Adversarial Network called Spine-GAN for resolving above-aforementioned challenges. Firstly, Spine-GAN

*Corresponding author
Email addresses: wbz99@sina.com (Benzheng Wei), slishuo@gmail.com (Shuo Li)

explicitly solves the high variety and variability of complex spinal structures through an atrous convolution (i.e., convolution with holes) autoencoder module that is capable of obtaining semantic task-aware representation and preserving fine-grained structural information. Secondly, Spine-GAN dynamically models the spatial pathological correlations between both normal and abnormal structures thanks to a specially designed long short-term memory module. Thirdly, Spine-GAN obtains reliable performance and efficient generalization by leveraging a discriminative network that is capable of correcting predicted errors and global-level contiguity. Extensive experiments on MRIs of 253 patients have demonstrated that Spine-GAN achieves high pixel accuracy of 96.2%, Dice coefficient of 87.1%, Sensitivity of 89.1% and Specificity of 86.0%, which reveals its effectiveness and potential as a clinical tool.

Keywords: spine, magnetic resonance imaging, segmentation, classification, generative adversarial network, LSTM, autoencoder, computer-aided detection and diagnosis

2010 MSC: 00-01, 99-00

1. Introduction

Spinal diseases greatly limit body mobility, block nervous system, and deteriorate quality of life worldwide. For instance, neural foraminal stenosis (NFS) [1] often causes muscle weakness and body disability, and NFS has attacked about 80% of the elderly population (Kaneko et al., 2012; Rajaee et al., 2012). In another instance, intervertebral disc degeneration (IDD) easily induces chronic back pain and body functional incapacity, while IDD is responsible for over 90% of spine surgical procedures (He et al., 2017a). Also, lumbar vertebral deformities (LVD) always goes along with scoliosis and spondylolisthesis and severely

[1] Neural foraminal stenosis refers to a spinal nerve root compressed by a narrowing neural foramen of the peripheral nervous system, which consists of the nerves and ganglia outside of the brain and spinal cord, serving as a relay between the brain and spinal cord and the rest of the body.

deteriorates the quality of life of children, young athletes and older people (Sun et al., 2017; Cai et al., 2017). However, clinical diagnosis of these spinal diseases still relies on laborious workloads and suffers from high inter-observer variations, because experienced radiologists make decisions based on their level of expertise and always get different radiological grading results even according to the same grading criterion (Lee et al., 2010). Mis-grading and inaccuracy clearly affect the decision of optimal therapeutic planning. Therefore, this work primarily focuses on an automated analysis of NFS, IDD, and LVD at the same time, which is extremely useful in providing reliable and objective assessment and relieving clinicians from laborious workloads.

Automated semantic segmentation of neural foramen, intervertebral discs, and vertebrae simultaneously have more clinical significance than conventional spinal segmentation or spinal classification of one spine component solely. In practice, since NFS, IDD, and LVD have strong pathological correlations (Lee et al., 2010; Krämer, 1983; McClure, 2000), automated semantic segmentation of the three structures can significantly promote clinical pathogenesis-based diagnosis [2] of NFS, IDD, and LVD. The specific reason is that each of NFS, IDD, and LVD is a crucial pathogenic factor of the others as well as a vital predictor of the others. Specifically, *1)* not only could IDD with space collapse lead to NFS but also indicate that nerve roots in the surrounding neural foramen will be compressed even when NFS has not occurred (Lee et al., 2010); *2)* not only could IDD with space collapse cause the surrounding vertebrae to be deformed but also indicate scoliosis or spondylolisthesis in the future; *3)* not only could LVD compress surrounding intervertebral discs but also restrict the room of neural foramen to pressure nerve roots. Therefore, this work finally devotes to semantic segmentation of neural foramen, intervertebral discs, and vertebrae

[2] Pathogenesis-based diagnosis is a key step to prevent and control spinal diseases in clinic, where spinal clinicians conduct both early diagnosis and comprehensive assessment by drawing crucial pathological links between spinal diseases and their pathogenic factors, which are the biological mechanisms that lead to the diseased state.

simultaneously, which can present their pathogenic factors for assist clinicians in discovering pathogenesis of them. For instance, if our system accurately predicts that both IDD and NFS occur in the same area of a spine, the system can present that a pathogenesis of the NFS is the IDD. Our system thereby has strong sensitivity and objectivity of pathological changes. This work thus significantly promote early diagnosis when a pathogenic factor is solely occurring. This work also can be helpful for eradicating a spinal disease when its pathogenic factor is presented, which contributes to building comprehensive pathological analysis and benefits to clinical surgery planning. In addition, this work leverages magnetic resonance imaging (MRI) since MRI is widely used in clinical diagnosis of spinal diseases as is better to demonstrate spinal anatomy and is the most appropriate test for imaging neural foramen and intervertebral discs (Kim et al., 2015).

However, no work has simultaneously achieved the semantic segmentation of neural foramen, intervertebral discs, and vertebrae due to three unusual challenges. Firstly, multiple tasks must be satisfied to ensure a successful system: *1)* simultaneous and pinpoint segmentation of neural foramen, intervertebral discs, and vertebrae to recognize pathogenic sites (i.e., locations of target diseases); *2)* simultaneous and accurate classification of neural foramen, intervertebral discs, and vertebrae to analyze corresponding diseases and present the pathological correlations between them. Secondly, multiple structures are required to be analyzed since each MRI has on average 21 target structures, which include seven neural foramen, seven intervertebral discs, and seven vertebrae [3]. The normal structures or abnormal structures of each disease have high variety and variability since they do differ with race, sex, and age. For instance, the heights of normal neural foramen on average range from 0.61 cm to 2.27 cm

[3]Clinical MRIs of lumbar spine follow the same imaging standard that each scan comprises of eight vertebrae from the 11-th thoracic vertebra to the final caudal vertebra. This work excludes the final caudal vertebra and analyzes the seven vertebrae from 11-th thoracic vertebra to the 5-th lumbar vertebra (L5), which are easiest to get sick following clinical statistics.

while the widths range from 0.41 cm to 1.90 cm according to our quantitative statistics. Thirdly, the normal and abnormal structures of each disease have subtle differences since they share a high degree of visual similarity. The weak spatial correlations on MRIs between neural foramen, intervertebral discs, and vertebrae lead to dynamic complexity and indeterminacy, which impedes the presentation of pathogenic factors such that cannot comprehensively analyze target diseases to prevent the further recurrences of morbidity. The locations of pathogenic factors on MRIs are always hard to be presented in spine surgery when patients do not feel obvious pain. For example, it is hard to decide that the main pathogenic factor of NFS is from which side when a neural foramen is compressed (Lee et al., 2010).

To overcome the aforementioned challenges, we propose a deep Recurrent Generative Adversarial Network called Spine-GAN and achieve the automated semantic segmentation of neural foramen, intervertebral discs, and vertebrae simultaneously. The proposed Spine-GAN comprises a segmentation network and a discriminative network within an integrated end-to-end framework. The segmentation network contains two modules: an atrous convolution (i.e., convolution with holes or dilated convolution) autoencoder module that is capable of obtaining deep task-aware representation and preserving fine-grained information as well; and a local long short-term memory module (LSTM) that is capable of modeling the spatial correlations between neural foramen, intervertebral discs, and vertebrae. The discriminative network effectively smooths and strengthens the higher-order spatial consistency of semantic segmentation by playing an adversarial role to distinguish its input whether from the segmentation network or ground truth. The end-to-end training of above three integrated modules ensures Spine-GAN acquire global information in conforming to the presence of pathological correlations and pathologies of spinal structures.

1.1. Related work on spinal image analysis

To the best of our knowledge, no work has achieved the segmentation and classification of multiple spinal structures simultaneously. Existing works in-

clude but are limited to the manual assessments, automated detection or radiological classification of one or two types of spinal structure.

1.1.1. Manual assessments

Existing manual assessment works of spine demonstrate that *(1)* manual assessment has inevitable subjectivity, and *(2)* pathological correlations do exist among intervertebral discs, neural foramen, and vertebrae. For instance, Lee et al. (2010); Park et al. (2012); Kim et al. (2015) aimed at evaluating the reproducibility of different manual grading criteria of neural foramen. Results of these studies show that experienced radiologists have different assessment results even according to the same grading criterion. Cinotti et al. (2002); Panjabi et al. (2006); Kaneko et al. (2012); Peck et al. (2016) studied the pathogenesis and treatment of spine diseases about how it is affected by segmental deformities and anatomical variations. Their results prove that the pathogenesis of spine diseases are from multiple pathogenic factors and can be attributed to the combination of accelerated degeneration of the intervertebral discs, dysplasia of the vertebral bones, and narrowing of the neural foramen. Although manual works have gained on grading criterion and pathogenic factors, they are time-consuming and inefficient.

1.1.2. Automated detection

Existing automated detection of the spine can be attributed to three types. The first type is the automated localization of one or two types of spinal structure (Alomari et al., 2011; Corso et al., 2008; Štern et al., 2009; Zhan et al., 2012; Cai et al., 2015), which are only capable of presenting specific anatomic structure. The second type is automatic segmentation of one or two types of spinal structure (He et al., 2017b,c; Wang et al., 2015; Yao et al., 2016). The third type is simultaneous localization and segmentation (Ghosha et al., 2011; Huang et al., 2009; Klinder et al., 2008; Peng et al., 2006; Shi et al., 2007; Kelm et al., 2013). Existing detection methods although achieved accurate recognition of one or two types of spinal structure, they cannot achieve the radiological

classification of spinal structures. The proposed method, instead, achieves simultaneously pinpoint segmentation and accurate radiological classification of three types of spinal structure.

1.1.3. Radiological classification

A few radiological classification works of spine include but are limited to one type of spinal structure, such as lumbar neural foramen grading (He et al., 2016), lumbar disc generation grading (He et al., 2017a; RajaS et al., 2011), spondylolisthesis grading (Cai et al., 2017). Besides, SpineNet proposed by Jamaludin et al. (2017) has achieved multiple disc diseases grading; however, it only takes the preprocessed discs' volume as input so that only analyzes one type of spinal structure. In this study, not only do we analyze three types of spinal diseases at three types of spinal structure but also contribute to presenting spatial pathological links between them.

1.2. Related work on deep learning

Atrous convolution was originally developed for the efficient computation of the undecimated wavelet transform (Holschneider et al., 1990) and used before in semantic segmentation (Chen et al., 2016; Yu and Koltun, 2015; Chen et al., 2017), object recognition (Sermanet et al., 2013), and image scanning (Giusti et al., 2013).

LSTM was proposed by Hochreiter and Schmidhuber (1997), and next was extended to language modeling (Sundermeyer et al., 2012), image captioning (Krause et al., 2017; Johnson et al., 2016; Vedantam et al., 2017; Karpathy and Fei-Fei, 2015; Xu et al., 2015), scene labeling (Byeon et al., 2015), and semantic object parsing (Liang et al., 2016b,a). In medical image analysis community, LSTM was used for cardiac assessment (Xue et al., 2017b; Poudel et al., 2016) by capturing the temporal dependencies between MR sequences.

The novel generative adversarial theory, corresponding to minimax two-player game, was proposed by Goodfellow et al. (Goodfellow et al., 2014) to generate virtual samples. The adversarial theory was first extended to super-

vised semantic segmentation task by Luc et al. (2016) and then used for weakly unsupervised semantic segmentation (Souly et al., 2017), unsupervised video summarization (Mahasseni et al., 2017), prostate cancer detection (Kohl et al., 2017), brain MRI segmentation (Moeskops et al., 2017; Kamnitsas et al., 2017), and unsupervised anomaly detection (Schlegl et al., 2017). These works have gained different levels of improvements, which demonstrates the effectiveness of adversarial learning.

1.3. Contributions

The primary contributions of this study include:

- For the first time, simultaneous semantic segmentation of multiple spinal structures are achieved, which is an advance compared to conventional segmentation or classification of spinal structures.
- GAN, local LSTM, and atrous autoencoder are combined in an integrated end-to-end framework. The novel framework is powerful in representation learning, spatial context learning, and adversarial learning. The integration can be efficiently extended to semantic segmentation, object parsing, and scene labeling.

The rest of this paper is organized as follows: In Section 2 we present the advantages of Spine-GAN in terms of advanced modules, strategy, and algorithm. In Section 3 we introduce the details about dataset and experiment settings. In Section 4 we conduct comprehensive performance analysis of Spine-GAN on this challenging study. In Section 5 we conclude this paper and then discuss the important influence of this study for other tasks.

2. Spine-GAN

2.1. An overview of Spine-GAN

The newly-proposed Spine-GAN has advanced modules (see Section 2.2) with hybrid learning strategy (see Section 2.3) and dynamic optimization algorithm (see Section 2.4). As illustrated in Fig. 1, Spine-GAN comprises of

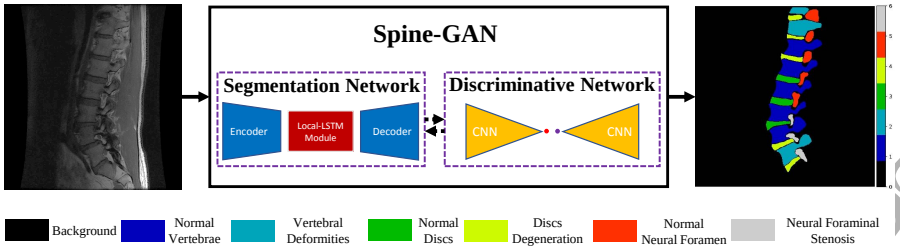


Figure 1: Spine-GAN directly localizes all target structures, while classifies all target structures with crucial radiological diagnoses (i.e., normal and abnormal) with pixel-level precision (i.e., each pixel has the possibility of seven classes including the background.). The goal of the segmentation network is to trick the discriminative network while the discriminative network is to urge the segmentation network to produce higher correct predictions, which efficiently prompts both accurate and convergence performance.

two newly-designed networks that are constituted of three tightly integrated modules. Specifically, a segmentation network is designed for segmentation and radiological classification of neural foramen, intervertebral discs, and vertebrae. The segmentation network includes a deep atrous convolution autoencoder module (ACAE) for spinal image representation and pixel-level classification (i.e., each pixel has the possibility of seven classes including the background.), and a local LSTM (Local-LSTM) based recurrent neural network (RNN) module for spatial dynamic modeling of spatial pathological correlations between spinal structures. Inspired by the GAN theory (Goodfellow et al., 2014), a discriminative network is designed to supervise and encourage the segmentation network to produce correct predictions. The discriminative network includes a convolutional neural network (CNN) module composed of auxiliary convolutional layers. The proposed hybrid learning strategy and dynamic optimization algorithm prevent Spine-Gan from over-fitting and also contribute to a robust generative adversarial learning.

The primary advantages of Spine-GAN include:

- The ACAE module enables Spine-GAN to not only address the high vari-

ability and complexity of spinal appearance in MRIs explicitly, but also effectively preserve fine-grained differences between normal and abnormal structures. ACAE module can produce deep semantic representation with only a few parameters and large receptive fields, avoiding too many stacked downsampling and up-sampling operations which severely reduce the feature resolution among low-dimensional manifold.

- The Local-LSTM module enables Spine-GAN to model the latent yet crucial spatial correlations between neighboring structures dynamically. Local-LSTM module is capable of selectively memorizing and forgetting semantic information of previous spinal structures when transforming the high-level semantic features into sequential inputs of LSTM units. Therefore, Spine-GAN achieves improved performance through leveraging spatial context from spinal structures' prior information.

- The discriminative network enables Spine-GAN to correct predicted errors and break through small dataset limitation, so as to achieve continued gains on global-level contiguity and avoid over-fitting. In addition, the robust learning strategy and flexible optimization algorithm allow Spine-GAN to handle the adversarial engagement between the segmentation network and the discriminative network smoothly during training. Spine-GAN thus obtains efficient convergence and generalization.

2.2. Three advanced modules for two networks

As illustrated in Fig. 1, an ACAE module, and a Local-LSTM module constitute a segmentation network, while a CNN module constitutes a discriminative network. The segmentation network aims at tricking the discriminative network by generating true-to-nature pixel-level segmentation maps. On the contrary, the discriminative network makes great efforts to discriminate its inputs whether they are fake maps from segmentation network or real maps from ground truth. This adversarial process between the two networks practically corrects higher-order inconsistencies between predicted segmentation maps and

their ground truth. The details of three advanced modules for two networks are as follows.

2.2.1. ACAE module of segmentation network

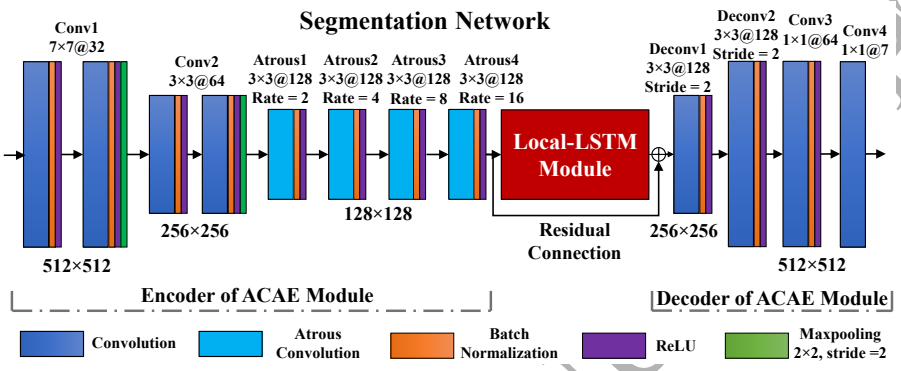


Figure 2: The segmentation network aims at generating pixel-level segmentation and radiological classification of multiple spinal structures. The ACAE module does not only preserve fine-grained information but also recovers input spatial dimension and object details. The Local-LSTM learns spatial pathological correlations between surrounding abnormal spinal structures.

As illustrated in Fig. 2, the ACAE module comprises atrous convolution layers as an encoder and deconvolution layers as a decoder. ACAE, therefore, has both advantages of atrous convolution network and traditional convolution autoencoder (CAE). ACAE module does not only preserve fine-grained information with few parameters and suitable receptive fields but also recovers input spatial dimension and object details. The difference between our ACAE module and existing CAE method (Masci et al., 2011; Chen et al., 2017; Zeiler et al., 2010; Ronneberger et al., 2015; Xue et al., 2017a) is that ACAE module extends atrous convolutions as an encoder for efficient expressive presentation and uses deconvolution as a decoder instead of the common interpolation method (Chen et al., 2017) with a learning manner to recover input spatial dimension. Moreover, the reason of using two max-pooling layers and two deconvolution layers rather than completely using atrous convolutions is to reduce the size of

feature maps and enlarge receptive fields more efficiently, such that Spine-GAN has faster computational speed and wider receptive fields.

To better understand atrous convolution of ACAE, we consider standard two-dimensional convolution layers first. For each location i on the output y and each kernel k on the weight w and bias b, standard convolutions are applied over the input feature map x with the stride of 1:

$$y[i] = f(\sum_k x[i+k] \times w[k] + b[k]). \tag{1}$$

f denotes the kernel-wise nonlinear transformation of ReLU and batch normalization (Ioffe and Szegedy, 2015). The atrous convolution adds an atrous rate r to each kernel k:

$$y[i] = f(\sum_k x[i+r\cdot k] \times w[k] + b[k]), \tag{2}$$

where the $r \cdot k$ is equivalent to convolving the input x with upsampled kernels, which is produced by inserting $r-1$ zeros between two consecutive values of each kernel along each spatial dimension (Chen et al., 2017). Standard convolution is a special case of atrous convolution with $r = 1$. Our ACAE module adopts progressive rates r of $\{2, 4, 8, 16\}$ after cross-validation, which allows ACAE module to modify kernel's receptive fields adaptively without stacked downsampling operators. Adaptive receptive fields of ACAE module ensures Spine-Gan acquire global information in conforming the presence of pathology of spines, such as vertebral morphology and posture.

Deconvolution layers (or transposed convolution) have been widely employed to recover the spatial resolution (Liu et al., 2015; Mostajabi et al., 2015; Badrinarayanan et al., 2015a; Pinheiro and Collobert, 2014; Papandreou et al., 2015). Unlike existing methods using stacked deconvolution layers with huge trainable parameters, our ACAE module only uses two deconvolution layers, which are actually the transpose (gradient) of standard convolution layer with stride of 2.

2.2.2. Local-LSTM module of segmentation network

Local-LSTM module is explicitly an LSTM units based RNN and deployed to model the spatial pathological correlations between neighboring structures,

which efficiently improves the radiological classification accuracy of spinal structures. For instance, in terms of spatial context, current neural foramina has a high probability of being abnormal when neighboring discs or vertebra are abnormal. The Local-LSTM module, therefore, is able to memorize long-period spatial pathological correlations from local neighboring structures and avoid the gradient vanishing/exploding problem in traditional RNNs. The difference between our Local-LSTM module and existing LSTM works is that Local-LSTM integrates LSTM units into autoencoder module and is for the first time used for spinal image analysis.

As illustrated in Fig. 3, the newly proposed Local-LSTM module has a top-down structure. Specifically, assuming $M \in \mathbb{R}^{n\times n\times c}$ represents a set of deep convolutional feature maps generated by the encoder of ACAE module with widths of n, heights of n , and channels of c. Firstly, the Local-LSTM module downsamples its input feature maps to $M^{'} \in \mathbb{R}^{\frac{n}{i}\times\frac{n}{i}\times c}$, where i is the size of 4 according to the receptive fields of spinal structures. Secondly, the Local-LSTM module patch-wisely unstacks these downsmapled feature maps $M^{'}$ to a set of spatial sequences $M^{''} \in \mathbb{R}^{(\frac{n}{i})^2\times c}$. Thirdly, the Local-LSTM recurrently memorizes long-period context information between spatial sequences and generates outputs $S \in \mathbb{R}^{(\frac{n}{i})^2\times c}$. Finally, the Local-LSTM module adaptively upsamples the outputs S into $S^{'} \in \mathbb{R}^{n\times n\times c}$ using two deconvolution layers. Accordingly, Local-LSTM module has $(\frac{n}{i})^2$ LSTM units and c-dimensions cell state.

As illustrated in Fig. 4, each LSTM unit of the Local-LSTM module contains peephole connections that are capable of learning the fine distinction between sequences and generating very stable sequences of high nonlinearity, which is one popular LSTM variant introduced by Gers and Schmidhuber (2000). Each LSTM unit also contains an input gate i_t, an output gate o_t, an forget gate f_t, and an memory cell m_t, which allows the Local-LSTM module to learn when to forget previous hidden state c_{t-1} and when to update current hidden states c_t. The connective mechanism of each LSTM unit is shown below:

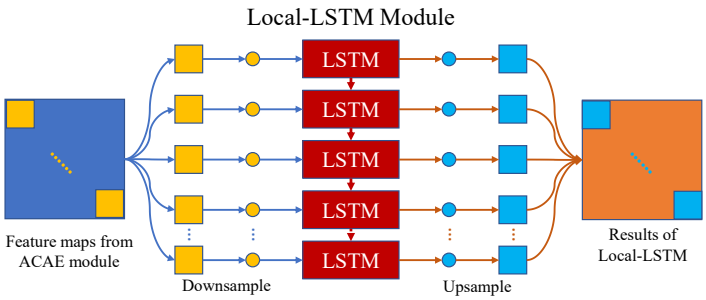


Figure 3: Local-LSTM module has a top-down structure that is able to capture the spatial dynamics of neighboring spinal structures and memorize their long-period pathological correlations after taking the encoder results of ACAE module as input.

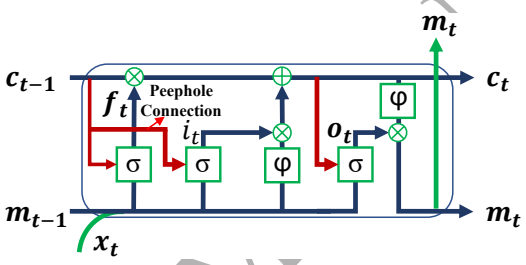


Figure 4: An illustration of the various gates and peephole connections (in red color) of an LSTM unit.

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i), \tag{3}$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f), \tag{4}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_{t-1} + b_o), \tag{5}$$

$$g_t = \varphi(W_{cx}x_t + W_{cm}m_{t-1} + b_c), \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{7}$$

$$m_t = o_t \odot \varphi(c_t), \tag{8}$$

where $\sigma(\cdot)$ and $\varphi(\cdot)$ are element-wise logistic sigmoid and hyper-bolic tangent non-linearity functions respectively. $\odot$ terms are element-wise products. W terms denote weight matrices (e.g. W_{fx} is the matrix of weights from the forget

gate to the input), W_{ic}, W_{fc}, W_{oc} are diagonal weight matrices for peephole connections. The b terms denote bias vectors (e.g. b_f is the forget gate bias vector). g_t is the cell input activation functions using $\varphi(\cdot)$. The equations of 3, 4 and 5 efficiently do map the current input and previous hidden state to the input gate i_t, the forget gate f_t and the output gate o_t with peephole connections. This strategy enables Local-LSTM module to adaptively control the information flow, so that Local-LSTM module adaptively memorizes and accesses spatial context of spinal structures in long term.

2.2.3. The CNN module of discriminative network

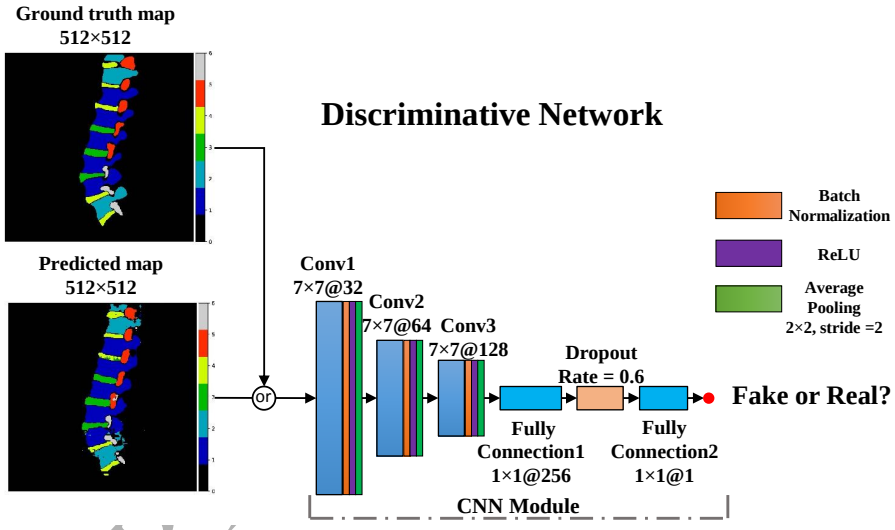


Figure 5: The CNN module of the discriminative network receives either the predicted maps or ground truth maps and outputs a single scalar to discriminate it input whether fake or real. Its adversarial role eagerly boosts the segmentation network to look out mismatches in a wide range of higher-order statistics, which prompt the global field-of-view and long-range spatial contiguity.

Our CNN module extensively plays an adversarial role to prompt the global-level field-of-view, convergence, and generalization of the segmentation network. The CNN module enforces long-range spatial label contiguity, without adding

complex post-processing (e.g. conditional Markov random fields) used at test time (Luc et al., 2016). Our CNN module differs from existing ones because our module integrates LSTM and atrous convolution into an end-to-end framework.

As illustrated in Fig. 5, the CNN module of the discriminative network comprises of three convolutional layers with large kernels, three batch normalizations, three average pooling layers, and two fully connected layers with dropout. In regard to training, the CNN module first receives either the predicted maps from the segmentation network or the manual maps from ground truth, then outputs a single scalar representing whether the input is from segmentation network or ground truth. When strong confrontation occurs, the discriminative network eagerly prompts the segmentation network to look out mismatches in a wide range of higher-order statistics between predicted segmentation maps and ground truth.

2.3. Hybrid learning strategy

The learning strategy of Spine-GAN empirically has a stable performance by combining the robust losses of the segmentation network and the discriminative network. To better understand the learning strategy of Spine-GAN, we consider primary GAN (Goodfellow et al., 2014) first. According to the theory of two-player minimax game, the objective of primary GAN is to minimize the probability of the samples (generated by the generative network G) to be recognized while maximizing the probability of the discriminative network D making a mistake (Mahasseni et al., 2017), which is formulated as the following minimax optimization:

$$\min_{G}\max_{D}[E_{y\sim p_{data}(y)}[\log D(y)] + E_{x\sim p_x(x)}[\log(1 - D(G(x))]], \tag{9}$$

where $y \sim p_{data}(y)$ denotes the true data samples, while $D(y)$ represents the probability that y came from the true data rather than generated (fake) data. $x \sim p_x$ denotes the prior input noise of generative network, while $G(x)$ represents the newly generated data. However, how to train GANs more stably is an ongoing question.

In regard to our task, $x \sim p_x$ represents the clinical spinal MRIs, which are the $512 \times 512 \times 1$ matrices. $y \sim p_y$ represents the ground truth maps, which have the same size as MRIs. The value of each pixel in ground truth denotes its class (seven class including background). When training, the discriminative network $D(\cdot;\theta_d)$ simultaneously receives the segmentation maps generated by the segmentation network $S(x;\theta_s)$ and ground truth y. Specifically, $D(\cdot)$ denotes the probability of its inputs where are from, while $S(x)$ represents the generated segmentation maps. θ_s and θ_d denote the trainable parameters of the segmentation network and the discriminative network respectively.

Since our aim is to generate best segmentation maps and each point of maps represents a radiological classification, Spine-GAN uses a hybrid loss function that is a weighted sum of two terms comprising a segmentation loss and GAN loss (Equation 9). To encourage the segmentation network generating best segmentation maps and compel the discriminative network making mistakes, the hybrid loss function $\mathcal{L}(\theta_s, \theta_d)$ is defined as:

$$\mathcal{L}(\theta_s,\theta_d) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\mathcal{L}_{mcl}(S(x_n), y_n)}_{\text{Segmentation network}} - \lambda[\underbrace{\mathcal{L}_{bcl}(D(y_n),1) + \mathcal{L}_{bcl}(D(S(x_n)),0)}_{\text{Discriminative network}}]. \tag{10}$$

λ is set to one in order to maintain the balance of adversarial learning. The segmentation loss ($\mathcal{L}_{mcl}$) is a balanced multi-class cross-entropy loss which encourages the segmentation network to predict right pixel-wised class labels. Given a dataset of N training MRIs x_n with dimension of $H \times W \times 1$, and corresponding ground truth maps y_n with dimension of $H \times W \times 1$ and C classes, the weighted multi-class cross-entropy loss is

$$\mathcal{L}_{mcl} = -\sum_{i=1}^{H\times W} w_c y_i \log\left(\frac{e^{\hat{y}_{ic}}}{\sum_{c=1}^{C} e^{\hat{y}_{ic}}}\right), \tag{11}$$

where $\hat{y}$ denotes predicted pixel-level results, $w_c = \sum_{n=1}^{N}(H \times W)/M_c$ denotes the c-th class's weight, where M_c is the pixel amounts of the c-th class in training dataset.

The second term $\mathcal{L}_{bcl}$ is a binary cross-entropy loss which makes it hard for the discriminative network to recognize whether its input has generated fake maps or ground truth maps. The binary cross-entropy loss is

$$\mathcal{L}_{bcl} = -[z \log \hat{z} + (1 - z) \log(1 - \hat{z})], \tag{12}$$

where z is the input labels (fake maps are zeros and ground maps are ones), while $\hat{z}$ are the single scalar of the output of the discriminative network.

2.4. Dynamic optimization algorithm

Spine-GAN is efficiently optimized by a dynamic minimizing algorithm using a suboptimal thinking, which stably boosts the convergence and generalization. Specifically, minimizing Equation (10) can be decomposed into two subproblems with respect to optimizing the segmentation network and optimizing the discriminative network respectively and simultaneously.

2.4.1. Optimizing the segmentation network

To trick the discriminative network, the segmentation network should generate best segmentation maps that can not be distinguished. Therefore, optimizing the segmentation network is equivalent to update θ_s by

$$\min_{\theta_s} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{mcl}(S(x_n), y_n) - \mathcal{L}_{bcl}(D(S(x_n)), 0). \tag{13}$$

Following Goodfellow et al. (2014); Radford et al. (2015); Luc et al. (2016), $-\mathcal{L}_{bcl}(D(S(x_n)), 0)$ is replaced by $\mathcal{L}_{bcl}(D(S(x_n)), 1)$, which leads to a stronger gradient signal when the discriminative network accurately predicts x_n such that speeds up training. Therefore, Eq.(13) is changed to

$$\min_{\theta_s} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{mcl}(S(x_n), y_n) + \mathcal{L}_{bcl}(D(S(x_n)), 1), \tag{14}$$

where y_n is after one-hot coding with the dimension of $H \times W \times C$ consisting of zeros and ones (i.e. true class places are 1, others are 0).

2.4.2. Optimizing the discriminative network

To guarantee the adversarial strength, optimizing the discriminative network is equivalent to update θ_d by

$$\min_{\theta_d} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{bcl}(D(y_n), 1) + \mathcal{L}_{bcl}(D(S(x_n)), 0), \tag{15}$$

where y_n is after weighted one-hot coding consisting of τ and ones (i.e. true class places are $1 - \tau$, the others are τ). In our task, τ is set to 0.01 after cross-validation, which can prevent the discriminative network to distinguish the ground truth maps easily when it detects its input maps consisting of zeros.

2.4.3. Global optimization of Spine-GAN

To ensure that the discriminative network is functioning, the weights of both θ_s and θ_d are initialized with Xavier initialization (Glorot and Bengio, 2010) without pre-training. Two types of optimizers are implemented on each network respectively. The first optimizer employed in the segmentation network is RMSProp algorithm (Tieleman and Hinton, 2012) with exponential decay based learning rate η_1. The second optimizer employed in the discriminative network is Adam algorithm (Kingma and Ba, 2014) with fixed learning rate η_2. Considering the task of the segmentation network is harder than the discriminative network, the initial learning rate η_1 of RMSProp is set to 0.01, while the fixed learning rate η_2 of Adam is 0.001. Based on these settings, the dynamic minimizing suboptimal method implementation is shown in Algorithm 1.

3. Data and Experiment

3.1. Data

The proposed Spine-GAN has been intensively evaluated on a challenging dataset which includes 253 multicenter clinical patients. Average years of patient age is 53±38 with 147 females and 106 males. Since these multicenter patients are examined by different models of vendors, their MRI scans have different parameters. The range of repetition time is from 380 ms to 4,000 ms with

Algorithm 1 Spine-GAN optimization

Input: A dataset of N training MRI x; Ground truth maps y; Class balanced weights W; The loss balanced weight λ; Learning rates η_1, η_2; Mini-batch size n; Maximum epochs M;

Output: Learned parameters $\{\theta_s, \theta_d\}$ within checkpoints;

Initialize all parameters $\{\theta_s, \theta_d\}$ randomly and construct model graph;
for step in $\frac{MN}{n}$ **do**
  x_n, $y_n \leftarrow$ fed mini-batch n from shuffled training dataset;
  /* The **forward** propagation of $S(x_n)$: */
  $e_n = encoder(x_n)$;
  $lstm_n = Local\text{-}LSTM(e_n)$;
  $S(x_n) = decoder(lstm_n)$;
  /* The **forward** propagation of $D(\cdot)$: */
  $D(S(x_n)) = CNN(S(x_n))$;
  $D(y_n) = CNN(y_n)$;
  /* The **backward** propagation for $S(x_n)$: */
  $\theta_s \xleftarrow{+} -\eta_1 \nabla(\mathcal{L}_{mcl}(S(x_n), y_n) + \mathcal{L}_{bcl}(D(S(x_n)), 1))$
  /* The **backward** propagation for $D(\cdot)$: */
  $\theta_d \xleftarrow{+} -\eta_2 \nabla(\mathcal{L}_{bcl}(D(y_n), 1) + \mathcal{L}_{bcl}(D(S(x_n)), 0))$
end for

mean of 1,529 ms; echo time from 8.144 ms to 151 ms with mean of 38.35 ms under a magnetic field length of 1.5 T; slice thickness from 0.879 mm to 4 mm with mean of 3.14 mm; and in-plane pixel spacing from 0.38 mm × 0.38 mm to 0.86 mm × 0.86 mm with mean of 0.56 mm × 0.56 mm. Among sequential T1/T2-weighted MRI scans of each patient, one lumbar middle scan was selected to better present neural foramen, discs, and vertebra simultaneously in the sagittal direction. So the dataset comprises 253 lumbar scans from 253 patients and no patient is placed in both sets of training and testing. The dataset has 1779 neural foramen (738 normals, 1041 abnormals), 1818 discs (623 normals, 1195 abnormals), and 1746 lumbar vertebrae (1058 normals, 678 abnormals). The segmentation ground truth is labeled by our lab tool according to the clinical criterion. The ground truth of classification was annotated by extracting from clinical reports of spinal surgery, which are double-checked by two experienced spinal radiologists. In regard to neural foramen classification, normal neural foramen refer to the absence of foraminal stenosis, while abnormal neural foramen refer to mild or severe foraminal stenosis showing perineural fat obliteration in vertical and/or transverse directions without or with morphologic changes. In regard to intervertebral disc classification, normal discs refer the absence of deformation and degeneration, while abnormal discs refer to the appearance of either deformation or degeneration; In regard to vertebrae classification, normal vertebrae refer the absence of osteoarthritic changes in the facet joints and cephalad subluxation of the superior articular process of the inferior vertebra, while abnormal vertebrae refer the appearance of either osteoarthritic changes in the facet joints or cephalad subluxation of the superior articular process.

3.2. Configuration

Standard five-fold cross-validation is employed for performance evaluation and comparison. Since the number of the dataset is limited, we did not divide it into training, validation and testing sets directly. Following the standard five-fold cross-validation, the original dataset is randomly partitioned into five equal size sub-datasets. Of the five sub-datasets, a single sub-dataset is retained as

the testing data for testing the model, and the remaining four sub-datasets are used as training data. When training five same models, we choose the same training epochs to save the trained models and which is used to evaluate the performance. The five results from the folds can then be averaged to produce a single result. The advantage of this method is that all observations are used for both training and testing, and each observation is used for testing exactly once. Spine-GAN directly handles clinical MRIs without any pre/post-processing and data augmentation. In terms of RMSProp optimizer, decay is 0.9, momentum is 0.9, and ϵ is 1e-10. In terms of Adam optimizer, β_1 is 0.9, β_2 is 0.999, and ϵ is 1e-08. The Spine-GAN is implemented on Python 2.7 and Tensorflow 1.2 library (Abadi et al., 2016). Mini-batch size is 4 and training epochs is 300 using one Nvidia GPU Titan X with cuDNN v5.1 and Intel CPU Xeon(R) E5-2620@2.5GHz.

3.3. Evaluation criteria

Pixel-level accuracy, Dice coefficient, Specificity, and Sensitivity are adopted for demonstrating the advantages of Spine-GAN. The segmentation and radiological classification of one spinal structure is correct if this structure is pixel-wisely segmented and classified correctly. Average pixel-level accuracy is defined as

$$Pa = \frac{1}{NC}\sum_{n=1}^{N}\sum_{n=1}^{C}\frac{p_c}{P_c}, \tag{16}$$

where N is the number of testing samples, C is the number of class, p_c denotes the amount of right classified pixels of class c while P_c denotes all pixels of class c. Dice coefficient of one class c is defined as

$$Dsc = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \tag{17}$$

where TP_c is the true positives of class c, while FP_c and FN_c are false positives and false negatives of class c respectively. Dice coefficient, Specificity, and Sensitivity are standard metrics computed on pixel-level confusion matrix.

3.4. Experimental design

Extensive experiments are conducted to validate the effectiveness of our Spine-GAN from the following aspects.

Firstly, comprehensive performance evaluation of Spine-GAN is conducted on the semantic segmentation of neural foramen, intervertebral discs, and vertebrae with our dataset following the standard five-fold cross-validation protocol.

Secondly, the advantages of the segmentation network and the discriminative network of Spine-GAN are extensively analyzed with ablation experiments following the same five-fold cross-validation protocol. First, with and without the CNN modules (as ACAE+LSTM) are investigated for demonstrating their important adversarial ability. Second, with and without the Local-LSTM modules (as ACAE+CNN) are investigated for proving their strengths in spatial pathological correlation modeling ability. Third, both CNN module and Local-LSTM module are removed (as ACAE) for proving the abilities of ACAE module also the necessity of the end-to-end integration of proposed three modules. When the CNN module is removed, the weighed multi-class cross-entropy loss described in Eq. 11 is used for the experimental comparisons.

Finally, inter-comparisons are conducted to demonstrate the strengths of Spine-GAN. Four other popular segmentation networks are examined with the same five-fold cross-validation protocol. The first one is U-Net proposed by Ronneberger et al. (2015) for biomedical image segmentation, which has 19 convolution layers, 4 up-convolution layers, and 4 max-pooling layers [4]. The second one is the Fully convolutional network (FCN) proposed by Shelhamer et al. (2017) for natural image segmentation. Its FCN-VGG16 version is used for comparison with Spine-GAN. The deconvolution layers of FCN-VGG16 are initialized by bilinear upsampling and its weights initialization uses VGG weights as original paper [5]. The third one is DeepLabv3+ model (Chen et al., 2018), which extends DeepLabv3 (Chen et al., 2017) by adding a simple yet effective decoder module

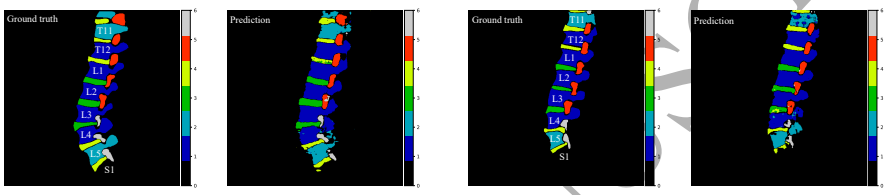
[4] U-Net: `https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/`

[5] FCN: `https://github.com/shelhamer/fcn.berkeleyvision.org`

to refine the segmentation results. The forth one is SegNet (Badrinarayanan et al., 2015b). The four networks are implemented by ourselves on Tensorflow strictly following the original papers using the same number of training batch and training epochs as Spine-GAN.

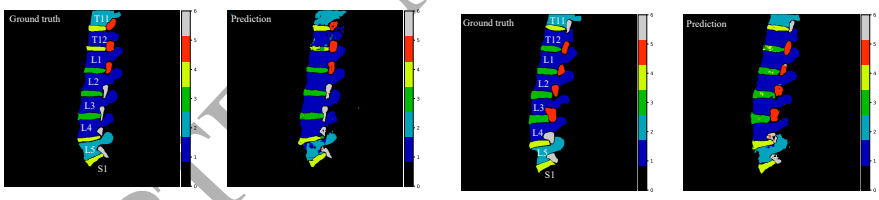
4. Result and Analysis

4.1. Comprehensive analysis



(a) Spine-GAN automatically presents that the intervertebral discs at L4-L5 and L5-S1 not only caused deformed vertebra and narrowed foramen but also indicate that scoliosis or spondylolisthesis may occur in the future.

(b) Spine-GAN presents that the pathogenic sites of L4-L5 NFS are at L5 vertebra and L4-L5 intervertebral disc, which contributes to clinical surgery planning to prevent and reduce the further recurrences of morbidity.



(c) Spine-GAN simultaneously predicts the four NFSs, four LVDs, and two IDDs, which contributes to building comprehensive treatment plans.

(d) Spine-GAN presents that the NFSs at L4-L5 and L5-S1 is caused by the L5 vertebral deformation, L4-S1 intervertebral discs' degeneration.

Figure 6: Spine-GAN has achieved reliable performance in the semantic segmentation of neural foramen, intervertebral discs, and vertebrae, which demonstrate it is a great tool for clinical settings. Color bars represent: **0:background; 1:normal vertebrae; 2:LVD; 3:normal disc; 4:IDD; 5:Normal foramen; 6:NFS** (Best in color).

As shown in Fig. 6 and Table 1, the effectiveness, and advantages of Spine-GAN have been demonstrated. Spine-GAN has simultaneously achieved accu-

Table 1: Spine-GAN (ACAE+LSTM+CNN) has superior segmentation and radiological classification effectiveness demonstrated by inter-comparisons and intra-comparisons.

Method	Pixel accuracy	Dice coefficient	Specificity	Sensitivity
FCN	0.917±0.004	0.754±0.033	0.754±0.035	0.712±0.032
SegNet	0.945±0.002	0.760±0.032	0.795±0.043	0.719±0.024
DeepLab v3+	0.953±0.001	0.812±0.021	0.799±0.035	0.827±0.017
U-Net	0.920±0.004	0.797±0.013	0.816±0.027	0.770±0.026
ACAE	0.958±0.002	0.841±0.013	0.862±0.018	0.823±0.024
ACAE+LSTM	0.959±0.002	0.848±0.009	0.865±0.021	0.837±0.025
ACAE+CNN	0.960±0.004	0.863±0.006	0.873±0.015	0.855±0.027
Spine-GAN	**0.962±0.003**	**0.871±0.004**	**0.891±0.017**	**0.860±0.025**

rate segmentation, precise radiological classification of neural foramen, intervertebral discs, and vertebrae. This is beneficial to clinical treatment processes such as therapeutic schedules, surgery plans and etc. For instance in Fig. 6(b), Spine-GAN automatically presents that the pathogenic factors of the abnormal neural foramina (NFS) between L4-L5 are its surrounding L5 vertebra (LVD), L4-L5 intervertebral disc (IDD). Also in Fig. 6(c), Spine-GAN rightly presents that the abnormal L5 vertebra is the pathogenic factor of the L4-L5 IDD and the L4-L5 NFS. In terms of the prediction and prevention of spinal diseases, as shown in Fig. 6(a), Spine-GAN precisely presents that the intervertebral discs at T12-L1, T11-L2, and T10-T11 not only caused deformed vertebra but also indicate that scoliosis or spondylolisthesis may occur in the future. Even both the weak spatial correlations between three diseases and various spine structures lead to unusual difficulties, Spine-GAN obtains accurate performance which demonstrates its strengths in addressing the spatial pathological correlations and high structures variability.

4.2. Modules analysis by intra-comparison

As shown in Fig. 7 and Table 1, 2, 3 (from 4th row to 7th row), the three modules endow Spine-GAN a superior performance for the segmentation and radiological classification of intervertebral discs, neural foramen, and verte-

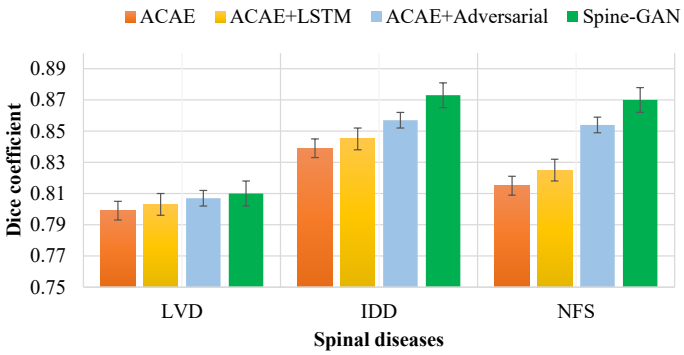


Figure 7: Spine-GAN shows superior performance in the semantic segmentation when compared with its ablated versions after removing Local-LSTM modules or/and CNN modules.

brae. As a baseline, Spine-GAN on average achieves 96.02%±0.3 pixel accuracy, 87.01%±1.00 Dice coefficient, 89.10%±1.70 Specificity, and 86.00%±1.70 Sensitivity. After only preserving the ACAE module, pixel accuracy and Dice coefficient are 95.8%±0.2 and 84.1%±1.3 decreased by 0.4% and 3.0% respectively. This not only demonstrates the effectiveness of Local-LSTM module, CNN module but also proves that the ACAE module is capable of obtaining deep semantic representation and preserving fine-grained detailed differences between normal and abnormal structures. Then, after removing the CNN module without generative adversarial training, pixel accuracy and Dice coefficient are 95.9%±0.2 and 84.8%±0.9, a decrease of 0.3% and 2.3% respectively, which demonstrates the CNN module can effectively correct the errors of semantic segmentation. Finally, after removing Local-LSTM module, Specificity and Sensitivity are 87.3%±0.02 and 85.5%±0.03, a decrease of 1.8% and 0.5% respectively, which demonstrates the capability of Local-LSTM module in modeling spatial pathological correlations between surrounding abnormal structures. Representative feature maps in Fig. 8 is better to demonstrate the ability of the Local-LSTM module intuitively. Moreover, in regard to radiological classification, Spine-GAN also achieves higher Specificity and Sensitivity than its ablated versions. Therefore, the combination of ACAE module, Local-LSTM

module, and CNN module makes Spine-Gan an efficient and reliable resolution for the semantic segmentation of multiple spinal structures.

Regard to the Local-LSTM module, since it is true that the output of encoder network is divided into smaller parts and fed into LSTM sequentially, the order and the size of parts will affect the performance on the semantic segmentation of multiple spinal structures. We, therefore, conducted several compared experiments in terms of different orders and sizes. Firstly, the order of the part of our final network is from top left to bottom right (top-down) as illustrated in Fig.3. The reason for determining this order is that spinal structures inherently have the top-down spatial correlation, which can be used for LSTM to learn the prior knowledge to improve the semantic segmentation performance. We have conducted three compared experiments in terms of top-down, down-top, and bidirectional orders. The result has demonstrated that the top-down achieved the best performance as shown in Table 4. Secondly, the size of each part is 4×4. The reason of determining the size is that We find that 4×4 is most suitable to the size of receptive fields of each filter of the convolutional layer at the former of Local-LSTM module, after we made a statistics of the real size of spinal structures on MR image, and computed the receptive field of each filter of the former convolutional layer. We have conducted three compared experiments in the size of 2×2, 4×4, and 8×8. The result has demonstrated that the 4×4 size achieved the best performance as shown in Table 4.

Table 2: Spine-GAN shows superior segmentation effectiveness demonstrated by the pixel-level Dice coefficient of normal and abnormal spinal structures from inter-comparisons and intra-comparisons.

Method	Dice coefficient					
	Normal vertebrae	**LVD**	Normal disc	**IDD**	Normal foramen	**NFS**
FCN	0.870±0.017	0.701±0.046	0.730±0.055	0.725±0.070	0.785±0.039	0.711±0.018
SegNet	0.889±0.009	0.760±0.023	0.695±0.019	0.776±0.014	0.756±0.037	0.684±0.021
DeepLabv3+	0.895±0.012	0.765±0.036	0.746±0.035	0.824±0.060	0.833±0.042	0.808±0.013
U-Net	0.878±0.007	0.726±0.036	0.772±0.025	0.803±0.020	0.821±0.017	0.782±0.010
ACAE	0.917±0.009	0.799±0.026	0.809±0.028	0.839±0.011	0.863±0.029	0.815±0.019
ACAE+LSTM	0.918±0.010	0.803±0.015	0.817±0.012	0.845±0.013	0.868±0.022	0.825±0.020
ACAE+CNN	0.929±0.010	0.807±0.015	0.835±0.021	0.857±0.015	0.887±0.009	0.854±0.014
Spine-GAN	**0.930±0.011**	**0.810±0.016**	**0.840±0.026**	**0.873±0.011**	**0.900±0.011**	**0.870±0.018**

Table 3: Spine-GAN shows superior radiological classification effectiveness demonstrated by the pixel-level Specificity and Sensitivity of three spinal diseases from inter-comparisons and intra-comparisons.

Method	Specificity			Sensitivity		
	LVD	IDD	NFS	LVD	IDD	NFS
FCN	0.875±0.025	0.638±0.072	0.745±0.041	0.737±0.085	0.726±0.039	0.672±0.029
SegNet	0.906±0.002	0.731±0.012	0.746±0.015	0.738±0.017	0.755±0.032	0.662±0.022
DeepLabv3+	0.894±0.010	0.717±0.027	0.786±0.035	0.761±0.020	0.852±0.025	0.865±0.012
U-Net	0.889±0.042	0.746±0.031	0.814±0.057	0.729±0.079	0.811±0.049	0.769±0.060
ACAE	0.907±0.027	0.810±0.039	0.867±0.032	0.817±0.070	0.844±0.027	0.821±0.027
ACAE+LSTM	0.909±0.027	0.818±0.055	0.870±0.032	0.822±0.079	0.854±0.025	0.829±0.026
ACAE+CNN	0.918±0.019	0.804±0.028	0.893±0.020	0.830±0.080	0.869±0.029	0.856±0.039
Spine-GAN	**0.921±0.020**	**0.844±0.063**	**0.907±0.047**	**0.831±0.084**	**0.871±0.029**	**0.876±0.029**

4.3. Superior analysis by inter-comparison

The inter-comparison results in Table 1, 2, 3 (from 2nd row to 3rd row) reveal the great advantages of Spine-GAN as well as its hybrid learning strategy and dynamic optimization algorithm. When compared with existing segmentation network as shown in Fig. 9, Spine-GAN significantly outperforms the FCN network by 4.5% pixel accuracy and 11.7% average Dice coefficient. FCN gets 91.7%±0.4 pixel accuracy and 75.4%±3.3 average Dice coefficient. Also, Spine-GAN outperforms the U-Net network by 4.2% pixel accuracy and 7.4% average Dice coefficient. U-Net obtains 92.0%±0.4 pixel accuracy and 79.7%±1.3 average Dice coefficient. In regard to radiological classification, Spine-GAN achieves more higher Specificity and Sensitivity than U-Net and FCN. Spine-GAN also has fewer parameters than FCN and U-Net, which effectively decreases the test time in practice. Specifically, Spine-GAN only has 104M parameters, while FCN has 134M parameters and U-Net has 268M parameters. Therefore, Spine-GAN enjoys strong superiority of prediction performance and application ability in the computer-aided diagnosis of spinal diseases.

5. Conclusion and Discussion

Automated semantic segmentation of multiple spinal diseases in one shot has been achieved by the innovative Recurrent Generative Adversarial Network

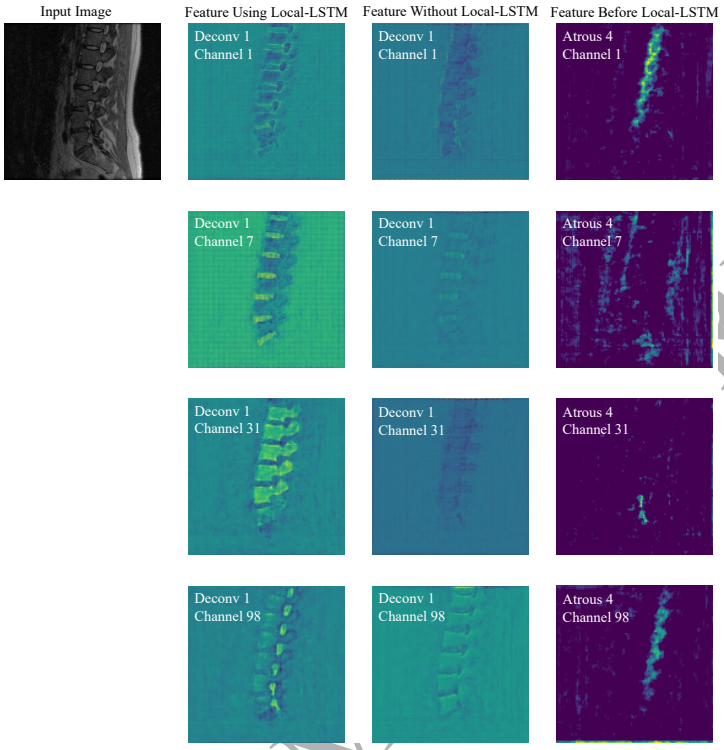


Figure 8: The feature maps generated after the Local-LSTM module shows the superior representative ability of the Local-LSTM for modeling spatial correlations between vertebrae, intervertebral disc, and neural foramen when compared with feature maps generated without Local-LSTM module and before Local-LSTM module, respectively. These feature maps are randomly selected from the same channels and comparable layers.

Spine-GAN, which is helpful for clinical pathogenesis-based diagnosis of spinal diseases. Spine-GAN combines the advantages of the powerful atrous convolution and autoencoder for semantic fine-grained representation, leverages the specialty of LSTM for modeling spatial pathological relationship among abnormal spinal structures, and also depends on an auxiliary CNN module for correcting predicted errors. When validated on spinal MR images of 253 patients, Spine-GAN achieved accurate segmentation, radiological classification, and pathological correlations representation of three types of spinal diseases. Specifically, Spine-GAN has achieved the pixel accuracy of 96.2 %, which demonstrates that

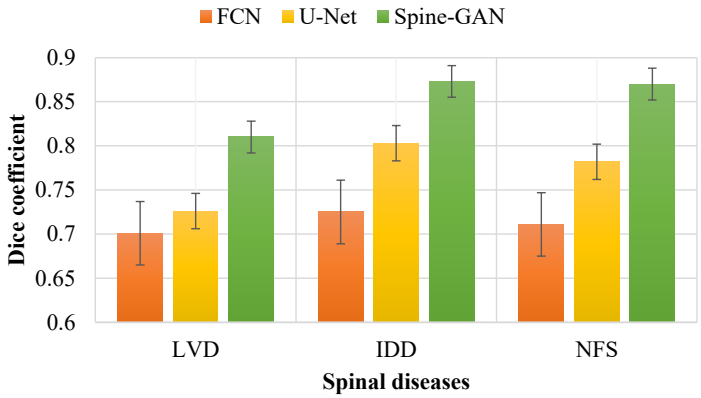


Figure 9: Spine-GAN shows the best performance for the segmentation and radiological classification of intervertebral discs, neural foramen, and vertebrae when compared with popular segmentation networks FCN and U-Net.

Table 4: The inter-comparison at different parameters of the Local-LSTM module.

parameter	Method	Pixel accuracy	Dice coefficient	Specificity	Sensitivity
Order	down-top	0.959±0.008	0.861±0.004	0.888±0.015	0.846±0.024
	bidirectional	0.961±0.002	0.855±0.004	0.888±0.015	0.848±0.024
	top-down	**0.962±0.003**	**0.871±0.004**	**0.891±0.017**	**0.860±0.025**
Size	2×2	0.959±0.003	0.867±0.002	0.876±0.018	0.853±0.023
	8×8	0.960±0.003	0.867±0.003	0.886±0.015	0.852±0.024
	4×4	**0.962±0.003**	**0.871±0.004**	**0.891±0.017**	**0.860±0.025**

our system can provide a very visual presentation for clinical application. Spine-GAN has also achieved 89.1 % specificity and 86 % sensitivity on both three types of spinal structures simultaneously, which demonstrates that our system is capable of assisting clinicians to improve the diagnosis efficiency. This newly-proposed automated framework has paved a great way for automated spinal disease's computer-aided diagnosis and can be adapted to other organs' semantic segmentation.

The limitations of our work have mainly two aspects: 1) this task only achieves segmentation and classification of spinal structures, which cannot real-

ize the human understanding of MR images. One future work accordingly is to realize clinical radiological report generation to directly help clinicians to improve their diagnosis efficiency. 2) The method of this paper has certain space to improve, one can embed clinical prior knowledge of spine diagnosis into this framework, which should be another future work.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was partly funded by Natural Science Foundation of China (No. 61872225); the Natural Science Foundation of Shandong Province (No. ZR2015FM010); the Project of Shandong Province Higher Educational Science and Technology Program (No. J15LN20); the Project of Shandong Province Medical and Health Technology Development Program (No.2016WS0577).

References

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 .

Alomari, R.S., Corso, J.J., Chaudhary, V., 2011. Labeling of lumbar discs using both pixel- and object-level features with a two-level probabilistic model. IEEE Transactions on Medical Imaging 30, 1–10. doi:10.1109/TMI.2010.2047403.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015a. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 .

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015b. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. CoRR abs/1511.00561. URL: http://arxiv.org/abs/1511.00561, arXiv:1511.00561.

Byeon, W., Breuel, T.M., Raue, F., Liwicki, M., 2015. Scene labeling with lstm recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3547–3555.

Cai, Y., Leungb, S., Warringtonb, J., Pandeyb, S., Shmuilovichb, O., Lib, S., 2017. Direct spondylolisthesis identification and measurement in mr/ct using detectors trained by articulated parameterized spine model, in: Proc. of SPIE Vol, pp. 1013319–1.

Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S., 2015. Multi-modality vertebra recognition in arbitrary views using 3d deformable hierarchical model. IEEE Transactions on Medical Imaging 34, 1676–1693. doi:10.1109/TMI.2015.2392054.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915 .

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611 .

Cinotti, G., De Santis, P., Nofroni, I., Postacchini, F., 2002. Stenosis of lumbar intervertebral foramen: anatomic study on predisposing factors. Spine 27, 223–229.

Corso, J.J., RajaS, A., Chaudhary, V., 2008. Lumbar disc localization and labeling with a probabilistic model on both pixel and object features, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 202–210.

Gers, F.A., Schmidhuber, J., 2000. Recurrent nets that time and count, in: Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, IEEE. pp. 189–194.

Ghosha, S., RajaS, A., Chaudharya, V., Dhillonb, G., 2011. Automatic lumbar vertebra segmentation from clinical ct for wedge compression fracture diagnosis. work 9, 11.

Giusti, A., Ciresan, D.C., Masci, J., Gambardella, L.M., Schmidhuber, J., 2013. Fast image scanning with deep max-pooling convolutional neural networks, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE. pp. 4034–4038.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

He, X., Landisa, M., Leunga, S., Warringtona, J., Shmuilovicha, O., Lia, S., 2017a. Automated grading of lumbar disc degeneration via supervised distance metric learning, in: Proc. of SPIE Vol, pp. 1013443–1.

He, X., Lum, A., Sharma, M., Brahm, G., Mercado, A., Li, S., 2017b. Automated segmentation and area estimation of neural foramina with boundary regression model. Pattern Recognition 63, 625–641.

He, X., Yin, Y., Sharma, M., Brahm, G., Mercado, A., Li, S., 2016. Automated diagnosis of neural foraminal stenosis using synchronized superpixels

representation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 335–343.

He, X., Zhang, H., Landis, M., Sharma, M., Warrington, J., Li, S., 2017c. Unsupervised boundary delineation of spinal neural foramina using a multi-feature and adaptive spectral segmentation. Medical image analysis 36, 22–40.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Holschneider, M., Kronland-Martinet, R., Morlet, J., Tchamitchian, P., 1990. A real-time algorithm for signal analysis with the help of the wavelet transform, in: Wavelets. Springer, pp. 286–297.

Huang, S.H., Chu, Y.H., Lai, S.H., Novak, C.L., 2009. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal mri. IEEE transactions on medical imaging 28, 1595–1605.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, pp. 448–456.

Jamaludin, A., Kadir, T., Zisserman, A., 2017. Spinenet: Automated classification and evidence visualization in spinal mris. Medical Image Analysis 41, 63 – 73. URL: http://www.sciencedirect.com/science/article/pii/S136184151730110X, doi:http://dx.doi.org/10.1016/j.media.2017.07.002. special Issue on the 2016 Conference on Medical Image Computing and Computer Assisted Intervention (Analog to MICCAI 2015).

Johnson, J., Karpathy, A., Fei-Fei, L., 2016. Densecap: Fully convolutional localization networks for dense captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565–4574.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial

networks, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 597–609.

Kaneko, Y., Matsumoto, M., Takaishi, H., Nishiwaki, Y., Momoshima, S., Toyama, Y., 2012. Morphometric analysis of the lumbar intervertebral foramen in patients with degenerative lumbar scoliosis by multidetector-row computed tomography. European Spine Journal 21, 2594–2602.

Karpathy, A., Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137.

Kelm, B.M., Wels, M., Zhou, S.K., Seifert, S., Suehling, M., Zheng, Y., Comaniciu, D., 2013. Spine detection in ct and mr using iterated marginal space learning. Medical image analysis 17, 1283–1292.

Kim, S., Lee, J.W., Chai, J.W., Yoo, H.J., Kang, Y., Seo, J., Ahn, J.M., Kang, H.S., 2015. A new mri grading system for cervical foraminal stenosis based on axial t2-weighted images. Korean journal of radiology 16, 1294–1302.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Klinder, T., Wolz, R., Lorenz, C., Franz, A., Ostermann, J., 2008. Spine segmentation using articulated shape models. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008 , 227–234.

Kohl, S., Bonekamp, D., Schlemmer, H.P., Yaqubi, K., Hohenfellner, M., Hadaschik, B., Radtke, J.P., Maier-Hein, K., 2017. Adversarial networks for the detection of aggressive prostate cancer. arXiv preprint arXiv:1702.08014 .

Krämer, G., 1983. Whiplash injuries of the cervical spine. pathogenesis of cerebral involvement and persistent post-traumatic disorders. Die Medizinische Welt 34, 1134–1140.

Krause, J., Johnson, J., Krishna, R., Fei-Fei, L., 2017. A hierarchical approach for generating descriptive image paragraphs, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lee, S., Lee, J.W., Yeom, J.S., Kim, K.J., Kim, H.J., Chung, S.K., Kang, H.S., 2010. A practical mri grading system for lumbar foraminal stenosis. American Journal of Roentgenology 194, 1095–1098.

Liang, X., Shen, X., Feng, J., Lin, L., Yan, S., 2016a. Semantic object parsing with graph lstm, in: European Conference on Computer Vision, Springer. pp. 125–143.

Liang, X., Shen, X., Xiang, D., Feng, J., Lin, L., Yan, S., 2016b. Semantic object parsing with local-global long short-term memory, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3185–3193.

Liu, W., Rabinovich, A., Berg, A.C., 2015. Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 .

Luc, P., Couprie, C., Chintala, S., Verbeek, J., 2016. Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 .

Mahasseni, B., Lam, M., Todorovic, S., 2017. Unsupervised video summarization with adversarial lstm networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Masci, J., Meier, U., Cireşan, D., Schmidhuber, J., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. Artificial Neural Networks and Machine Learning–ICANN 2011 , 52–59.

McClure, P., 2000. The degenerative cervical spine: pathogenesis and rehabilitation concepts. Journal of Hand Therapy 13, 163–174.

Moeskops, P., Veta, M., Lafarge, M.W., Eppenhof, K.A., Pluim, J.P., 2017. Adversarial training and dilated convolutions for brain mri segmentation. arXiv preprint arXiv:1707.03195 .

Mostajabi, M., Yadollahpour, P., Shakhnarovich, G., 2015. Feedforward semantic segmentation with zoom-out features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3376–3385.

Panjabi, M.M., Maak, T.G., Ivancic, P.C., Ito, S., 2006. Dynamic intervertebral foramen narrowing during simulated rear impact. Spine 31, E128–E134.

Papandreou, G., Kokkinos, I., Savalle, P.A., 2015. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 390–399.

Park, H.J., Kim, S., Lee, S.Y., Park, N.H., Rho, M.H., Hong, H.P., Kwag, H.J., Kook, S.H., Choi, S.H., 2012. Clinical correlation of a new mr imaging method for assessing lumbar foraminal stenosis. American Journal of Neuroradiology 33, 818–822.

Peck, S.H., Casal, M.L., Malhotra, N.R., Ficicioglu, C., Smith, L.J., 2016. Pathogenesis and treatment of spine disease in the mucopolysaccharidoses. Molecular genetics and metabolism 118, 232–243.

Peng, Z., Zhong, J., Wee, W., Lee, J.h., 2006. Automated vertebra detection and segmentation from the whole spine mr images, in: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, IEEE. pp. 2527–2530.

Pinheiro, P., Collobert, R., 2014. Recurrent convolutional neural networks for scene labeling, in: International Conference on Machine Learning, pp. 82–90.

Poudel, R.P., Lamata, P., Montana, G., 2016. Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation, in: International Workshop on Reconstruction and Analysis of Moving Body Organs, Springer. pp. 83–94.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 .

Rajaee, S.S., Bae, H.W., Kanim, L.E., Delamarter, R.B., 2012. Spinal fusion in the united states: analysis of trends from 1998 to 2008. Spine 37, 67–76.

RajaS, A., Corso, J.J., Chaudhary, V., Dhillon, G., 2011. Toward a clinical lumbar cad: herniation diagnosis. International journal of computer assisted radiology and surgery 6, 119–126.

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

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 146–157.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 .

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 39, 640–651.

Shi, R., Sun, D., Qiu, Z., Weiss, K.L., 2007. An efficient method for segmentation of mri spine images, in: Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on, IEEE. pp. 713–717.

Souly, N., Spampinato, C., Shah, M., 2017. Semi and weakly supervised semantic segmentation using generative adversarial network. arXiv preprint arXiv:1703.09695 .

Štern, D., Likar, B., Pernuš, F., Vrtovec, T., 2009. Automated detection of spinal centrelines, vertebral bodies and intervertebral discs in ct and mr images of lumbar spine. Physics in medicine and biology 55, 247.

Sun, H., Zhen, X., Bailey, C., Rasoulinejad, P., Yin, Y., Li, S., 2017. Direct estimation of spinal cobb angles by structured multi-output regression. in: International Conference on Information Processing in Medical Imaging, Springer. pp. 529–540.

Sundermeyer, M., Schlüter, R., Ney, H., 2012. Lstm neural networks for language modeling, in: Thirteenth Annual Conference of the International Speech Communication Association.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4, 26–31.

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., Chechik, G., 2017. Context-aware captions from context-agnostic supervision. arXiv preprint arXiv:1701.02870 .

Wang, Z., Zhen, X., Tay, K., Osman, S., Romano, W., Li, S., 2015. Regression segmentation for m3 spinal images. IEEE transactions on medical imaging 34, 1640–1648.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, pp. 2048–2057.

Xue, W., Islam, A., Bhaduri, M., Li, S., 2017a. Direct multitype cardiac indices estimation via joint representation and regression learning. IEEE Transactions on Medical Imaging PP, 1–1. doi:10.1109/TMI.2017.2709251.

Xue, W., Nachum, I.B., Pandey, S., Warrington, J., Leung, S., Li, S., 2017b. Direct estimation of regional wall thicknesses via residual recurrent neural

network, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 505–516.

Yao, J., Burns, J.E., Forsberg, D., Seitel, A., Rasoulian, A., Abolmaesumi, P., Hammernik, K., Urschler, M., Ibragimov, B., Korez, R., et al., 2016. A multi-center milestone study of clinical vertebral ct segmentation. Computerized Medical Imaging and Graphics 49, 16–28.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 .

Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R., 2010. Deconvolutional networks, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 2528–2535.

Zhan, Y., Maneesh, D., Harder, M., Zhou, X.S., 2012. Robust mr spine detection using hierarchical learning and local articulated model, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 141–148.