Improving Spatial Context in CNNs for Semantic Medical Image Segmentation

Russel Mesbah
rassoul@cs.otago.ac.nz

Brendan McCane
mccane@cs.otago.ac.nz

Steven Mills
steven@cs.otago.ac.nz

Anthony Robins
anthony@cs.otago.ac.nz

Department of Computer Science, University of Otago, Dunedin, New Zealand

Abstract

Convolutional Neural Networks (CNNs) have been widely used in the semantic segmentation of medical images. Current CNN-based approaches don't fully exploit information about the local neighbourhood of the pixels being classified. Furthermore, the average pixel-wise accuracy gives the average likelihood of correct classification across all the pixels in the frame where this likelihood is actually not the same for every pixel. We propose a new approach to address these issues by using multiple neighbourhoods around the pixel of interest and aggregating different hypotheses about the pixel's label. The results produced by this method are comparable with the state of the art solutions. In addition, the method is capable of detecting less accurate regions by assessing the consistency of labelling while shifting the sampling frame across these pixels.

1. Introduction

Convolutional neural networks (CNNs) are characterised by their weight sharing and down-sampling in which the former aims to detect local dependencies and the latter applies the resolution decrease to the feature maps. This helps to sustain location and noise invariance in CNN-based architectures [9]. It also provides more abstract features of a lower resolution by moving upward through the pyramid of convolutional layers until single or multiple concepts can be deduced at the apex of the abstraction pyramid [22].

CNNs show a good performance in many semantic segmentation problems by inferring the label for a single pixel from all the pixel values in the sampling frame. Due to the resolution decrease in the abstraction pyramid, the output can be noisy and less accurate at object boundaries [14]. In practice, even increasing the depth of the architecture cannot help to alleviate the deterioration of the accuracy in these regions. We propose a new approach to address this issue by employing multiple neighbourhoods around the pixel of interest and inferring the pixel label based on a set of potential labels extracted from different focus frames.

2. Related Works

Research on the deterioration of segmentation accuracy at object boundaries using feed-forward CNN-based architectures can be roughly divided into three categories: regulating the local dependencies in feature maps, fully convolutional networks, and encoder-decoder architectures.

Chen *et al.* [5] apply conditional random fields (CRF) to a deep CNN's outputs during training. Similarly, Liu *et al.* [15] employed a CRF to regulate CNN feature maps to avoid premature decision-making and add more homogeneity to the label space. Toca *et al.* [26] introduced a new convolutional layer called an "AutoMarkov Layer" which is capable of applying local dependencies between the feature maps to the loss function. CRFs can help segment dominant regions such as homogeneous backgrounds or salient objects rather than small regions or the areas with a higher frequency of variation in labelling. This is an impediment to employing this technique for object boundary detection for images with high frequency of variations in labelling.

Fully convolutional networks (FCNs) [16] map each feature space onto a bigger frame(s) and the corresponding interpolation function is learnt during back-propagation. FCNs provide a simultaneous classification of all pixels in the frame which includes local dependencies to the feature space and output layer. FCNs are sensitive to the scale of objects, so that if the object is very big or too small compared with the size of the receptive field, the segmentation accuracy may decrease. The fact that the reconstruction of the features based on FCN architectures are not capable of approximating functions with a high frequency of variation can be considered as an another shortcoming of this

approach.

Various encoder-decoder methods have been proposed to deal with localisation accuracy [2, 3, 4, 18], but none of them fully exploit context information. Moreover, due to the non-homogeneity of the average accuracy in the sampling frame, the correct classification rate for a particular pixel can be affected by the position of the pixel in the frame [17].

In order to address the lack of sufficient context information in the sampling frame, Roth *et al.* [21], Setio *et al.* [24], and van Grinsven *et al.* [27] sample the input image by shifting, scaling, or rotating these frames randomly several times and feeding them into multiple parallel CNN architectures to produce a set of probabilities. Then a cascaded classifier is employed to infer the right label for the pixel of interest from the set of generated probabilities.

Despite the fact that the authors could successfully include more context information in training and testing phases, the techniques are unable to model the interactions between neighbouring output pixels directly. To remedy this issue, we proposed to shift the sampling frames with a simple deterministic pattern and feed them into a set of parallel convolutional encoder-decoder architectures. A set of cascaded classifiers can be used to label the region of interest, accurately. Unlike [21, 24, 27] in which multiple parallel CNN architectures are engaged in a single expensive training phase, our approach is based on only one trained encoder-decoder architecture.

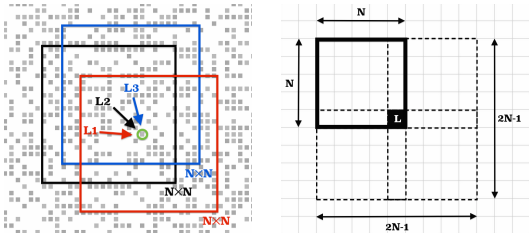
3. Shifting the Sampling Frames

With a convolutional encoder-decoder architecture of enough depth for a particular problem, the error can be derived from three main sources: lack of enough spatial context information in the dataset to generalise the learnt hypothesis, the localisation issue of DL-based approaches and the positional average accuracy across the sampling frame.

The former can be addressed by increasing the size of the sampling frame which can result in a noticeable growth in the computation cost. The second factor can be thought as the difficulty in detection of the object boundaries which are often of regions with a low confidence score (less certainty) in the classification. Finally, often the closer a pixel is to the borders of the sampling frame the lower the associated correct classification rate is [17]. Our approach alleviates all these three sources of error by incorporating more spatial context information at run time.

A convolutional encoder-decoder segments all the pixels in the input frame simultaneously. The network classifies each pixel based on the neighbouring input pixels and their corresponding inferred labels in the sampling frame. The network can segment the whole image at the same time. In practice, since numerous samples are needed for training the network, and due to the high computation cost for training an architecture with a large sampling frame, simultaneous segmentation of large images is not practical. Our solution is to employ non-overlapping sub-frames of the input image for training and testing the network. We propose to use overlapping frames to generate multiple labels for per pixel and then intelligently combine the labels.

By shifting the sampling frame across the pixels their confidence score is usually affected by both the different set of neighbouring pixels and the position of the pixels in the sampling frame. The corresponding labels for erroneous regions are less consistent compared to the ones for other areas of the image. This characteristic is the core idea to detect these less accurate regions at runtime.



(a) (b)

Figure 1. Shifting the sampling frame in the runtime: (a) L1, L2, and L3 are inferred labels for a single pixel based on different frame positions ; (b) Expansion of the receptive field by shifting the frame across a pixel.

Given a convolutional encoder-decoder architecture with an $N \times N$ receptive frame, each pixel in the image can be classified using $N \times N$ different sets of neighbouring pixels by shifting the sampling frame in vertical and horizontal directions, as shown in Fig. 1.b.

Let $I(x, y)$ and $L_i(x, y)$ represent the intensity value and the corresponding label for a pixel at coordinate (x, y) using the ith shifted frame for segmenting the pixel, respectively. Let $\Omega = \{\omega_1, \omega_2, ..., \omega_\alpha\}$ be the set of all possible labels (classes) where α is the number of classes. The probability of assigning $\omega_j \in \Omega$ to the pixel at coordinate (x, y) can be interpreted as

$$P\{\, L_i(x,y) = \omega_j \mid I(x,y), \forall (L_i(a,b), I(a,b)) : [(a,b) \in F_i, (a,b) \neq (x,y)] \,\} \quad (1)$$

where F_i is the set of all pixels in the ith shifted frame and

$$\mathbf{L}(x,y) = \{L_1(x,y), L_2(x,y), ..., L_\beta(x,y)\} \quad (2)$$

and $\beta = \frac{N^2}{J^2}$ where the frames are shifted with the stepsize of J. Thus, the vector is associated with an $(2N-1) \times (2N-1)$ receptive field as is shown in Fig. 1.b.

The vector $\mathbf{L}(x, y)$ can be simply mapped onto the Ω space by counting the number of occurrence of each

ω_j among the elements of $\mathbf{L}(x, y)$. Consider $\mathbf{W} = \{W_1, W_2, ..., W_\alpha\}$ as the corresponding label vector in the Ω space (assume $\alpha < \beta$ which is almost always valid). $\| \mathbf{W} \|_1 = \beta$ where $\| \mathbf{W} \|_1$ is the Manhattan norm. As a result, inclination of $\mathbf{W}$ toward one component affects its scalar projection onto other components. Based on this fact, we define the consistency of labelling for a pixel as

$$\textbf{consistency} := \frac{\textbf{max}\ (\mathbf{W})}{\| \mathbf{W} \|_1} \tag{3}$$

Pixels with high consistency ratio can be thought to be associated with low information entropy. Consistency and entropy are highly correlated especially when the number of classes is moderate. Although the relation between consistency and the corresponding label is not straightforward (especially in pixels with low consistency rate), we believe that the higher the consistency, the more likely the assigned label is to be correct. This conjecture is evaluated in the experiments we have performed in this work. The correct label can be inferred from the vector $\mathbf{L}(x, y)$ by feeding it into a classifier.

4. Datasets

We have evaluated our approach using four medical image segmentation datasets. Each has its own unique properties in terms of size and adjacency of the objects, dominance of the regions, and presence of noise.

Dataset 1 - greyscale microscopy images of the excised mouse spinal cord provided by University of Pennsylvania Medical Center: Semantic segmentation of these images offers the potential to study autoimmune diseases [10] and also brain connectivity and maturation [19] since myelin volume has an intimate relationship with nerve signal transmission performance in mammals [11]. Myelin can be found in lower intensity regions surrounding axons which are often brighter. However, the relationship between intensity values and labels is not always straightforward as there are regions with uniform labels and of a variety of intensity values.

Dataset 2 - retinal vessel segmentation database from Lincoln School of Computer Science: Estimating the width of retinal vessels can help to diagnose a variety of diseases such as arteriosclerosis, diabetic retinopathy, and so on [1, 13]. Although the edges of retinal vessels are of a darker material and the internal vessel regions are brighter, this intensity characteristic is not unique to vessels. In addition, the presence of noise and very low vessel diameter make the segmentation task challenging.

Dataset 3 - sclera and eye recognition benchmarking competition (SSERBC 2017): Sclera segmentation is recognised as an ocular biometric and is of a great practical importance in many security-based applications [6]. Despite the fact that the sclera regions are mostly brighter than other ocular areas, the segmentation cannot always be accurate as there are regions in the sclera which are of a lower intensity and some parts of the peri-ocular or iris are as bright as sclera pixels.

Dataset 4 - human lumbar vertebrae lateral X-ray images provided by Department of Computer Science, University of Otago: Lumbar vertebrae segmentation has an important role to play in computer-aided diagnosis of a variety of pathological conditions such as lumbar wedge compression, vertebral and intervertebral disc abnormalities, etc [8]. Bone regions are of a variety of intensity values not very different to the other areas. Although the vertebrae boundary pixels are roughy brighter than the internal and external vertebrae regions, the segmentation task cannot be performed easily due to the presence of noise.

5. Method

5.1. Encoder-Decoder Design

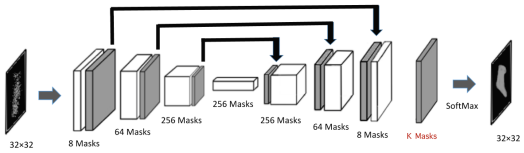


Figure 2. Convolutional encoder-decoder architecture

Our architecture consists of three main structures: encoder, fully convolutional, and decoder units (Fig. 2). The first includes three convolutional layers each followed by a 2×2 MaxPooling layer. A ReLU activation function is applied to each MaxPooling unit. There are 8, 64, and 256 3×3 convolutional masks in the first, second, and third convolutional layers, respectively. Since the receptive field is a 32×32 matrix and due to the resolution decrease through the MaxPooling layers, there are 256 4×4 feature maps at the apex of the encoder's pyramid of abstractions.

The next layer connects encoder and decoder units using fully convolutional operation. The un-pooling functions in the decoder unit remember the sub-sampling indices of the corresponding MaxPooling layers. The decoder includes three convolutional layers with 256, 64, and 8 3×3 masks, respectively. Each convolutional module placed after its associated un-pooling unit.

In the architecture designed for the first dataset, a convolutional layer with three 3×3 convolutional masks followed by a SoftMax function infers three classes at the backend of the decoder unit. These classes are represented in the first dimension of the $3 \times 32 \times 32$ output matrix. Since the three other datasets are associated with two-class segmentation problems, in their architecture, there is one convolutional mask replaced by the three convolutional masks

and the SoftMax unit of the former design. The output is a $1 \times 32 \times 32$ matrix in which the first dimension represents the two classes.

5.2. Shifting the Sampling Frames at Runtime

We have applied our technique to the convolutional encoder-decoder architecture by shifting the sampling frame across the points with the stepsize $J = 4$. As a result, there are eight shifted frames in each direction (horizontal left, horizontal right, vertical left, and vertical right). The consistency check has been performed using $\beta = 64$ inferred labels for each pixel.

The relation between the consistency ratio and the correct label needs to be learnt especially at the pixels with low consistency rate in which the right label cannot be inferred by just voting for the most frequent one among the set of hypothesised labels. To optimise the performance, these labels are fed into a multilayer perceptron (MLP) architecture in which the hidden layer has 256 nodes followed by a hyperbolic tangent activation function. For the first dataset, the output is a three-dimensional vector each dimension corresponds to one of the three classes. The architecture for the two other datasets includes a one-dimensional output vector (one node) which is associated with the two classes to be inferred.

5.3. Training and Experimental Setup

To verify our segmentation approach, we designed the trials by excluding one image and randomly sampling the other images of the dataset using 32×32 sampling frames. The corresponding labels are 3 dimensional matrices in which the first dimension represents the three/two classes. For the first dataset, one of the elements of the first dimension is 1 while others are 0. In other datasets, the element is either 1 or -1 indicating the presence or absence of the sought class. Table 1 shows the distribution of the classes in the datasets. There are 102000, 70000, 78000, and 76000 samples in the four datasets, in order.

–	Class A	Class B	Class C
Dataset 1	24.8%	43.5%	31.7%
Dataset 2	9.5%	90.5%	–
Dataset 3	23.8%	76.2%	–
Dataset 4	55.7%	44.3%	–

Table 1. Distribution of the datasets

We have implemented all the experiments using the Torch7 library (www.torch.ch) on an iMac with a 3.2 GHz Core i5 quad-core processor and an 1024 MB NVIDIA GPU. Detailed information about the training parameters are shown in Table 2.

	LR	Epochs	BS	GPU	Training
Setup 1	0.01	25	1	NO	375 min
Setup 2	0.1	10	16	YES	30 min
Setup 3	0.1	5	16	YES	15 min
Setup 4	0.1	7	16	YES	21 min

Table 2. Training setup: Learning Rate (LR), Batch Size (BS)

6. Results

The approach was compared against the most recent and accurate solutions [7, 17, 20] for the first three datasets. Since the research on the lumbar vertebrae segmentation is mainly based on the volumetric imaging techniques (MRI, CT, etc.) [23, 25] and as the works on X-ray vertebrae images are mostly devoted to object boundary detection [12], we are not able to use their reported accuracies as a validation test against our approach (we employed pixel-wise accuracy metric for the validation test). The correct classification rates and the corresponding standard deviations across the trials are given in Fig. 3.

We used paired sample t-tests to examine the significance of the segmentation accuracy between the results produced by our approach and the ones produced by the base encoder-decoder technique. For each dataset, the two sets of samples are the average pixel-wise per-image accuracies, respective to the techniques. The corresponding P values for the four datasets are 2.4×10^{-4}, 1.1×10^{-4}, 4.1×10^{-4}, and 2.5×10^{-4}, respectively, indicate a significant (< 0.05) improvement for all the four datasets.

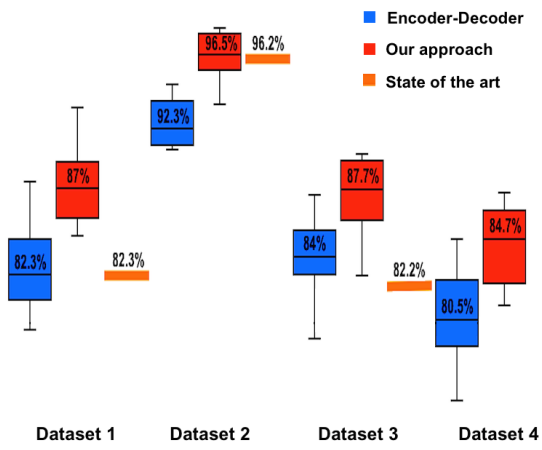


Figure 3. Comparison between pixel-wise accuracies: As the standard deviations associated with the state of the art solutions are not in hand, we only report the corresponding average pixel-wise accuracies.

7. Conclusion & Discussion

In this paper, we propose a self-assessment approach based on the convolutional encoder-decoder architecture. The method gives the likelihood of correct classification for each individual pixel in the image which can be important for a variety of applications. We address the lack of sufficient spatial context-related information in the sampling frame of encoder-decoder architectures by shifting the frames across pixels and inferring the correct label from the set of guessed labels, at runtime.

The method assesses the consistency of the labelling for the pixel of interest using different shifted sampling frames. As it can be seen in Fig. 4, the low consistency regions can be roughly classified into two main groups: the ones with a high frequency of variations in intensity and the pixels with lower classification confidence. While the former can be seen mostly at object boundaries, the latter can also be found in homogeneous regions. The learnt hypothesis differentiates the two sources of low consistency and the classifier acts accordingly. Examples of compensating for the lack of accuracy in both the object boundaries and the homogeneous regions can be found in all the four segmentation samples in Fig. 4. However, the approach shows higher performance in correcting the first type of misclassified pixels than the ones of the second group.

The higher the consistency rate, the better the classification performance. There are many boundaries and edges to be detected in the first two datasets unlike the others with many uniform regions.

Since the approach is based on the trained convolutional encoder-decoder architecture, the computation cost is comparable with traditional encoder-decoders, in the training phase. Assume the simultaneous classification of all pixels in an $N \times N$ image using convolutional encoder-decoder takes T seconds. The runtime for the proposed approach is equal to $N^2 \times J^{-2} \times T$ where J is the stepsize (the amount of time required for the artificial neural network to infer the right label should also be taken into the consideration).

Our Contributions - The contributions of this work can be summarised as follows:

- We proposed a self-assessment approach based on the convolutional encoder-decoder architecture. The technique gives the likelihood of correct classification for each individual pixel.

- The current convolutional encoder-decoder approaches don't fully exploit information about the local neighbourhood of the pixels being classified. We addressed the issue by using multiple neighbourhoods around the pixel of interest and aggregating different hypotheses about the pixel's label. The approach is capable of inferring the correct label based on the collected spatial context information at runtime.

- We applied the proposed technique to four medical image semantic segmentation datasets and the results show a significant improvement against the state of the art in three out of the four datasets.

References

[1] B. Al-Diri, A. Hunter, D. Steel, M. Habib, T. Hudiab, and S. Berry. Review - a reference data set for retinal vessel profiles. *Int Conf of the IEEE Eng in Medicine and Biology.*, pages 2262–2265, 2008. 3

[2] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *CVPR*, 2014. 1

[3] V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561*, 2015. 1

[4] T. Brosch, L. T. Y. Yoo, D. Li, A. Traboulsee, and R. Tam. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. *MICCAI*, 9351:3–11, 2015. 1

[5] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICIR*, pages 1–14, 2014. 1

[6] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein. Multi-angle based lively sclera biometrics at a distance. *Symp on Comp Intelligence in Biometrics and Identity Management (CIBIM)*, 2014. 3

[7] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein. Ssrbc 2016: Sclera segmentation and recognition benchmarking competition. *2016 Int Conf on Biometrics, ICB 2016*, 2016. 4

[8] S. Ghosh, R. S. Alomari, V. Chaudhary, and G. Dhillon. Automatic lumbar vertebra segmentation from clinical ct for wedge compression fracture diagnosis. *SPIE Medical Imaging Conference*, 3:796303–796309, 2011. 3

[9] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. 1

[10] C. E. Hafer-Macko, K. A. S. MBBS, C. Y. Li, T. W. Ho, D. R. Cornblath, G. M. McKhann, A. K. Asbury, and J. W. Griffin. Immune attack on the schwann cell surface in acute inflammatory demyelinating polyneuropathy. *Annals of Neurology*, 39:625–635, 1996. 3

[11] J. E. Hall and A. C. Guyton. *Guyton and Hall Textbook of Medical Physiology*. Saunders, Philadelphia: pa, 2011. 3

[12] B. Ibragimov, R. Korez, B. Likar, F. Pernus, L. Xing, and T. Vrtovec. Segmentation of pathological structures by landmark-assisted deformable models. *IEEE Transactions on Medical Imaging*, 0062:1–1, 2017. 4

[13] D. K. Kumar, B. Aliahmad, and H. Hao. Retinal vessel diameter measurement using unsupervised linear discriminant analysis. *ISRN ophthalmology*, 2012:151369–151376, 2012. 3

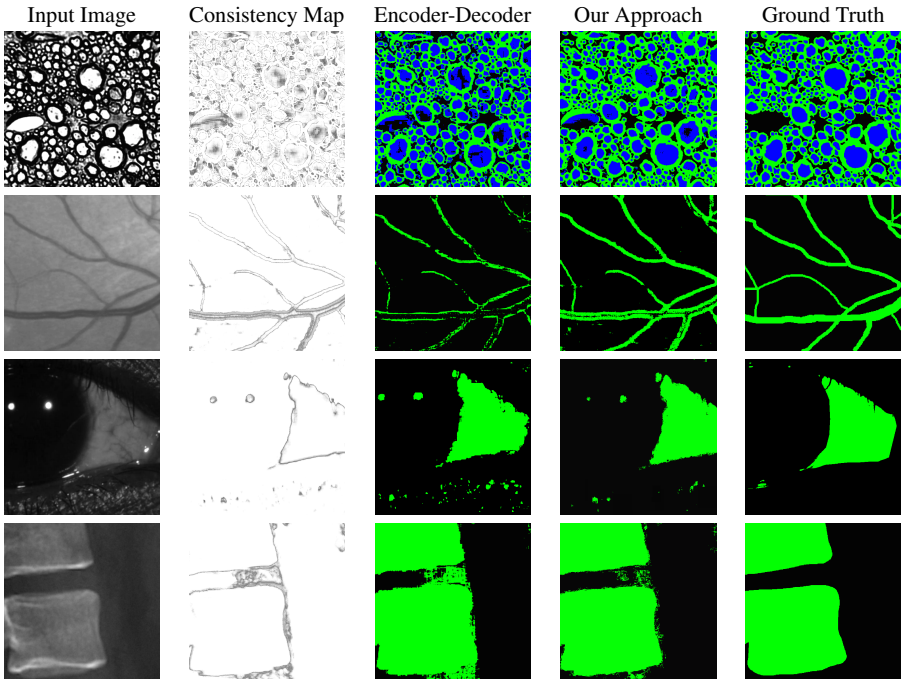


Figure 4. From left to right: input image which was unseen to the network during the training, consistency map, encoder-decoder segmentation output, segmentation based on our approach, and ground truth: in the consistency maps, pixels with lower consistency ratios are shown with lower intensity values.

[14] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 52:436–444, 2015. 1

[15] F. Liu, G. Lin, and C. Shen. Crf learning with cnn features for image segmentation. *Pattern Recognition*, 48:2983–2992, 2015. 1

[16] J. Long, E. Shehamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 39:3431–3440, 2015. 1

[17] R. Mesbah, B. McCane, and S. Mills. Deep convolutional encoder-decoder for myelin and axon segmentation. *IVCNZ*, pages 1–6, 2016. 1, 2, 4

[18] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *ICCV*, 1:1520–1528, 2015. 1

[19] H. H. Ong, A. C. Write, S. L. Wehrli, A. Souza, E. D. Schwartz, S. N. Hwang, and F. W. Wehri. Indirect measurement of regional axon diameter in excised mouse spinal cord with q-space imaging: Simulation and experimental studies. *NeuroImage*, 40:1619–1632, 2008. 3

[20] D. Pandey, X. Yin, H. Wang, and Y. Zhang. Accurate vessel segmentation using maximum entropy incorporating line detection and phase-preserving denoising. *Computer Vision and Image Understanding*, 155:162–172, 2016. 4

[21] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans. on Med. Imaging*, 35(5):1170–1181, 2016. 1, 2

[22] J. Schmidhuber. Deep learning in neural networks: An overview. *Elsevier*, 61:85–117, 2015. 1

[23] A. Sekuboyina, A. Valentinitsch, J. S. Kirschke, and B. H. Menze. A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets. *arXiv:1703.04347*, pages 1–10, 2017. 4

[24] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sanchez, and B. van Ginneken. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Trans. on Med. Imaging*, 35(5):1160–1169, 2016. 1, 2

[25] L. Tai, Q. Ye, and M. Liu. Pca-aided fully convolutional networks for semantic segmentation of multi-channel fmri. *arXiv:1610.01732*, 2016. 4

[26] C. Toca, C. Patrascu, and M. Ciuc. Automarkov dnns for object classification. *ICPR*, pages 3441–3446, 2016. 1

[27] M. J. J. P. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sanchez. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Trans. on Med. Imaging*, 35(5):1273–1284, 2016. 1, 2