Sparse Structured Prediction for Semantic Edge Detection in Medical Images

Lasse Hansen HANSEN@IMI.UNI-LUEBECK.DE
Mattias P. Heinrich HEINRICH@IMI.UNI-LUEBECK.DE
Institute of Medical Informatics, University of Lübeck, DE

Abstract

In medical image analysis most state-of-the-art methods rely on deep neural networks with learned convolutional filters. For pixel-level tasks, e.g. multi-class segmentation, approaches build upon UNet-like encoder-decoder architectures show impressive results. However, at the same time, grid-based models often process images unnecessarily dense introducing large time and memory requirements. Therefore it is still a challenging problem to deploy recent methods in the clinical setting. Evaluating images on only a limited number of locations has the potential to overcome those limitations and may also enable the acquisition of medical images using adaptive sparse sampling, which could substantially reduce scan times and radiation doses.

In this work we investigate the problem of semantic edge detection in CT and X-ray images from sparse sampling locations. We propose a deep learning architecture that comprises of two parts: 1) a lightweight fully convolutional CNN to extract informative sampling points and 2) our novel sparse structured prediction network (SSPNet). The SSPNet processes image patches on a graph generated from the sampled locations and outputs semantic edge activations for each patch which are accumulated in an array via a weighted voting scheme to recover a dense prediction. We conduct several ablation experiments for our network on a dataset consisting of 10 abdominal CT slices from VISCERAL and evaluate its performance against strong baseline UNets on the JSRT database of chest X-rays.

Keywords: sparsity, structured prediction, edge detection, deep learning.

1. Introduction

The vast majority of medical image acquisition and analysis has so far focused on reconstructing and processing dense data. This is mainly motivated by the simplicity of representing data points and their spatial relationships on regular grids and storing or visualizing them using arrays. In particular convolutional operators for feature extraction and pooling have seen increased importance for denoising, segmentation, registration and detection due to the rise of deep learning techniques. Learning spatial filter coefficients through backpropagation is well understood and computationally efficient due to highly optimized matrix multiplication routines for both CPUs and GPUs. However, for many computer vision tasks in medical image analysis such as landmark or edge detection it seems unnecessary and expensive (in terms of time and memory limitations) to process images end-to-end with dense methods, e.g. fully-convolutional networks or encoder-decoder architectures. Therefore, in this work, we aim to show new possibilities in the area of deep learning to process image data on sparse and irregular instead of dense grids. The feasibility of our suggested approach is demonstrated on the problem of semantic edge detection in CT and X-ray images.

Related Work: Of all hierarchical feature learning models, CNNs have shown to be one of the most successful approaches for a wide variety of tasks such as classification, bounding box regression and segmentation (Ronneberger et al., 2015; He et al., 2017). Lately, another class of works (graph convolutional neural networks (GCNNs)) attempts to transfer these well-known concepts from the two dimensional image domain to non-Euclidean and irregular domains. Spectral CNNs, defined on graphs, were first introduced in (Bruna et al., 2013). The main drawback of the proposed method is that it relies on prior knowledge of the graph structure to define a local neighborhood for weight sharing. (Henaff et al., 2015) extended the ideas to graphs where no prior information on the structure is available. While (Bruna et al., 2013; Henaff et al., 2015) relied on splines for the formulation of their graph convolutional operators, (Kipf and Welling, 2017) uses truncated Chebyshev polynomials that allow for clear description of the support size of the learned spectral filters. (Bronstein et al., 2017) provides a comprehensive review of current research on this topic. In the medical domain GCNNs were successfully applied in a number of applications such as population-based disease prediction (Parisot et al., 2017), metric learning for brain connectivity graphs (Ktena et al., 2017) and survival analysis on pathological images (Li et al., 2018).

Edge detection is a key task in computer vision applications and is studied for decades (Canny, 1986). (Dollár and Zitnick, 2013) chose a data-driven approach using random decision forests to predict structured labels from input image patches. This technique was successfully applied in the medical domain for multi-modal registration of ultrasound and CT/MRI images (Oktay et al., 2015). (Xie and Tu, 2015) is the first deep learning method to explicitly learn edges. Features are extracted with a modified VGGNet and all layers are trained with deep supervision. In the end side outputs from different VGG Layers are fused to output a final edge map. State-of-the-art detectors for semantic edge detection mainly resemble encoder-decoder architectures that are trained with specialized loss terms (Yu et al., 2017; Liu et al., 2018).

Contributions: In this work we make a first step towards dense prediction from a few sparse sampling points using deep learning methods. We bring together the robustness of grid based CNNs and the flexibility of GCNNs in a single framework for pixel-level structured prediction. In this, our work differs from (Li et al., 2018), which used GCNNs for global context aggregation for image labeling. Our main contributions is the sparse structured prediction network (SSPNet). Furthermore, we successfully provide a first proof-of-concept for our new approach by evaluating it on the challenging task of semantic edge detection in medical images.

2. Methods

In this section, we present our proposed approach for sparse structured prediction for semantic edge detection. Figure 1 illustrates the general idea of our method. Input to our pipeline is an image x. A light-weight CNN ϕ, called sample CNN, extracts potentially informative locations from the image and outputs a single channel sample map $\phi(x)$. A fixed number N of sample coordinates $((x_1,y_1),\dots,(x_N,y_N))$ are drawn following a multinomial distribution with probabilities proportional to the sample map's values. Depending on the application many alternatives of extracting sampling locations are conceivable, e.g. for landmark detection one could initialize the sample map with the mean locations of the landmarks in the training set. At the given positions, patches $(p_1,\dots,p_N)$ are extracted from the input image x. Furthermore, a simple distance graph G_σ

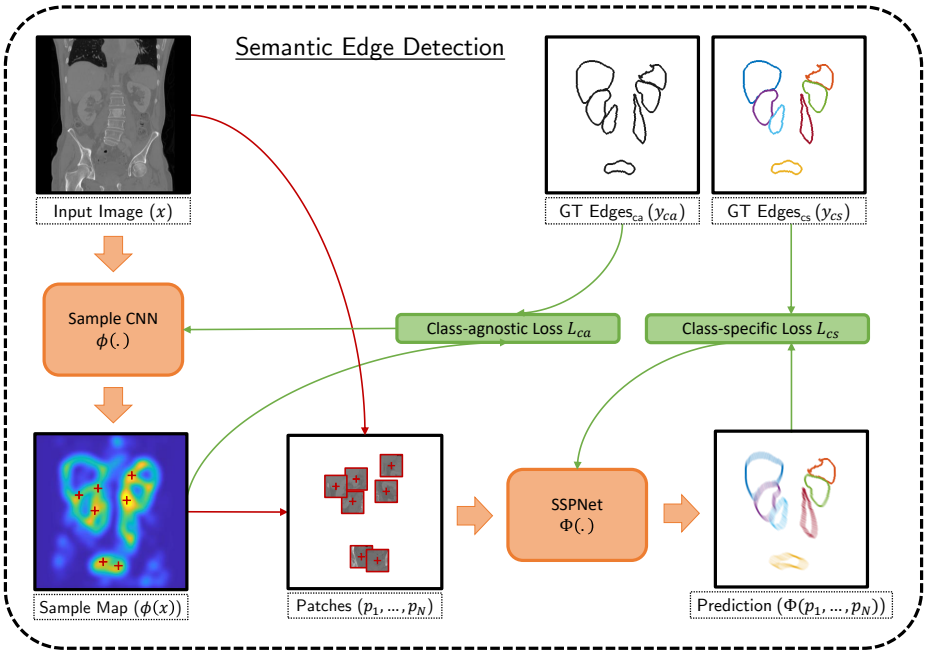


Figure 1: Our general idea for sparse semantic edge detection. We train a lightweight fully-convolutional CNN with a class-agnostic loss to output an informative heatmap from which samples are drawn with probabilities proportional to its values. Image patches are extracted around the chosen locations and our proposed SSPNet processes the generated patch graph to output a semantic edge activation for each sampling point. To recover a dense prediction all edges are accumulated in an array and the class-specific loss is applied to update the SSPNet's parameters.

is generated. The adjacency matrix $\mathbf{A}$ of the graph G_σ is given by entries

$$a_{ij} = \exp\left(\frac{-d_{ij}^2}{2 \cdot \sigma^2}\right),$$

where σ is a scalar diffusion coefficient and d_{ij} denotes the euclidean distance between two sampling locations (x_i, y_i) and (x_j, y_j). Again, depending on the application and given priors the graph may be initialized accordingly. Next, the extracted image patches $(p_1, \ldots, p_N)$ as well as the graph G_σ serve as input to our proposed SSPNet (explained in detail below), which predicts edges for each input image patch and accumulates all predictions on a dense grid weighted by their class-specific confidence. This semantic edge map is our final output. While the focus of this work is clearly on the SSPNet, in the following we also shortly describe the training of the sample CNN.

2.1. Sample CNN

The sample CNN ϕ is based on a lightweight version of the holistically-nested architecture in (Xie and Tu, 2015). We significantly cut the networks capacity by removing deeper layers and use

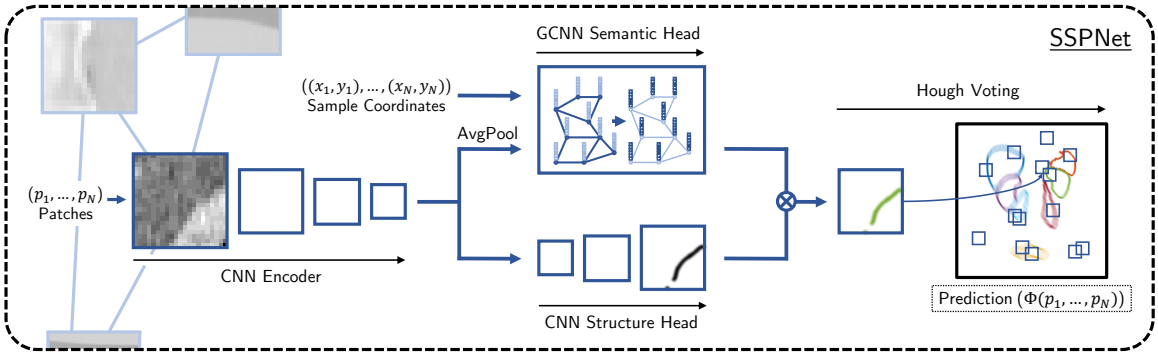


Figure 2: The proposed sparse structured prediction net (SSPNet) expects a graph of patches sampled at informative image locations. For each patch a CNN encoder extracts a set of feature maps, which are further processed by 1) the structure head that predicts local edge activations and 2) the semantic head where global context is aggregated by a GCNN. We perform a weighted Hough voting to accumulate all predictions and recover a dense semantic edge map.

reduced numbers of filters. In total the network consists of only three layers (each with two 3×3 convolutions + relu activation). Layer 2 and 3 start with convolutions with stride 2 resulting in the network's receptive field size of 23. After each layer a side output $\hat{y}_i$ is generated by a further 1×1 convolution and sigmoid activation. Side outputs are concatenated and fused to form a final prediction $\hat{y}_0$ by a 1×1 convolution and sigmoid activation. The sample CNN is trained with deep supervision on all outputs using the loss function from (Deng et al., 2018) which combines the binary cross-entropy (*BCE*) and Dice loss (*DICE*) to

$$L_{ca} = \sum_{i=0}^{3} \alpha BCE(\hat{y}_0, y_{ca}) + \beta DICE(\hat{y}_0, y_{ca}).$$

In the loss term y_{ca} depicts the class-agnostic version of the ground truth edge map and α and β control the weighting of the two losses. Trading robustness for precise localization the final sample map is obtained from prediction y_0 after multiple average pooling steps with stride 1.

2.2. SSPNet

As stated above input for our SSPNet Φ are the extracted image patches $(p_1,\ldots,p_N)$ as well as the graph G_σ. The network itself consists of a CNN encoder part, a structure and semantic head and a final Hough voting step to recover a dense prediction from the single patches. An overview of our proposed SSPNet is given in Figure 2. The CNN encoder applies four convolutions (kernel sizes: 5, 3, 3, 3) with relu activations on each image patch. The resulting feature maps are further processed by two network heads. The structure head consists of three transposed convolutions (kernel sizes 3, 3, 3) and relu activations. The final structured prediction is obtained by a 1×1 convolution with sigmoid activation. The semantic head aggregates global context information and is modeled with a GCNN. Input features on our graph are the average pooled feature maps from the CNN encoder. We

also experiment with explicitly adding the sampling coordinates as additional informative features. For this part of our work we decided to strive for simplicity and use a simple random walk diffusion with a single σ kernel to pool features across our input graph G_σ (Atwood and Towsley, 2016; Hansen et al., 2018). The diffusion process can be described by the diffusion matrix

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A},$$

where $\mathbf{I}$ denotes the identity matrix and the degree matrix $\mathbf{D}$ is solely defined by its diagonal elements $d_{ii} = \sum_j a_{ij}$. By matrix multiplication with the input feature vector a weighted average pooling across edges of the graph is employed. The diffusion pooling is followed by two 1×1 convolutions with relu activations. In total we employ two of the described graph convolutions. Final semantic confidence scores for each node (sampling locations) on the graph are obtained by a 1×1 convolution with sigmoid activation. As Hough voting has been proven to be effective for locating shapes in images (Ballard, 1981; Lindner et al., 2015) we accumulate the structured predictions from all image patches on a dense grid (with C channels, where C corresponds to the number of semantic classes) and weight each prediction with the corresponding semantic confidence score. Furthermore, each grid point is normalized by the number of predictions made for this point. Note that by construction the final semantic edge map holds values between 0 and 1 and we can apply our class-specific similar to our class-agnostic loss as

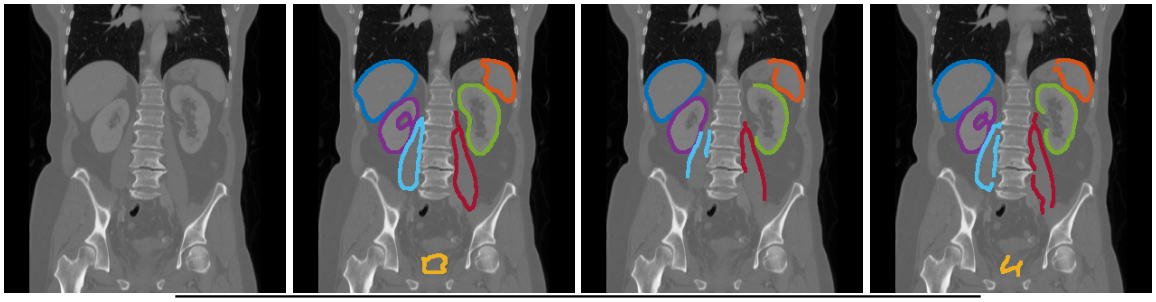
$$L_{cs} = \sum_{i=0}^{C-1} w_i(\alpha BCE(\hat{y}^{(i)}, y_{cs}^{(i)}) + \beta DICE(\hat{y}^{(i)}, y_{cs}^{(i)})),$$

where y_{cs} depicts the one-hot encoded semantic ground truth edges, such that pixels can belong to multiple labels. Classes may be weighted by the parameters w_i.

3. Experiments and Results

We validate the feasibility of our approach on two different datasets for the task of semantic edge detection. The first dataset consists of 10 2D coronal slices of abdominal CT scans from VISCERAL (Jimenez-del Toro et al., 2016) and the second dataset is the JSRT dabase of 247 chest X-ray images (Shiraishi et al., 2000). As validation metric we use the F-score on the fixed contour threshold (ODS), where the threshold is determined from all images in the test dataset. Before evaluation the thresholded predictions are thinned and spurious detections (< 10 pixels) are removed. ODS metrics are computed for each semantic class individually and we report the mean value. We compare our approach against three different 5-Layer UNet implementations (UNet-S, UNet-M, UNet-L). The UNet-S has a comparable capacity in terms of learnable parameters as our SSPNet, whereas the UNet-L has almost 2.5 as many parameters.

Implementation Details: All models were trained for 300 and 100 epochs for VISCERAL and JSRT, respectively. ADAM optimization was used with an initial learning rate of .02. We employ batch normalization with a mini batch size of 4 and an exponential learning rate schedule with a multiplicative factor of 0.99 to stabilize training. The images are augmented with a random affine transformation. The graph for the SSPNet is computed with a σ value of 0.1 and normalized coordinates. We set the α and β parameters of the loss terms to .001 and 1, respectively. Class weights were applied corresponding to the organ label occurrences for all experiments with the UNet variants. During training and at test time we sample patches at 500 and 2000 locations, respectively. All hyperparameters were determined by grid search for our simplest baseline method and kept fixed for all further experiments.



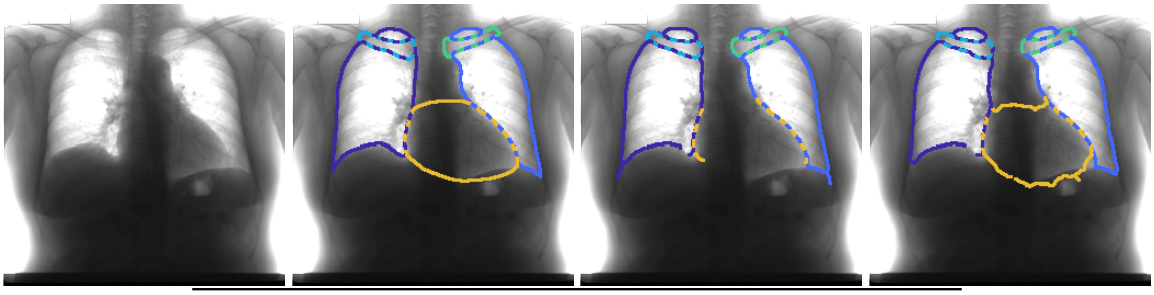
Method	Parameters	Samples	ODS
SSPNet – 1×1 conv.	˜ 500k	2000	.690
SSPNet – GCNN	˜ 500k	2000	.786
SSPNet – 1×1 conv. + coords.	˜ 500k	2000	.801
SSPNet – GCNN + coords.	˜ 500k	2000	**.827**
UNet-S	˜ 500k	dense	.769
UNet-M	˜ 900k	dense	.791
UNet-L	˜ 1300k	dense	**.834**

Figure 3: Qualitative and quantitative results on VISCERAL. The images (from left to right: original CT slice, ground truth, UNet-L, SSPNet) show edge overlays from seven anatomical structures: liver ■, spleen ■, bladder ■, left kidney ■, right kidney ■, left psoas major muscle (pmm) ■ and right pmm ■. Our approach outlines edges of the psoas muscles much clearer and also detects the urinary bladder.

3.1. VISCERAL

We perform initial experiments on the 10 2D coronal slices of abdominal CT scans from VISCERAL in a leave-one-out fashion. The images are resampled to an isotropic pixelsize of 1.5mm^2 and cropped to dimensions of 320×312 without any guidance. We consider ground truth labels for seven anatomical structures: liver ■, spleen ■, bladder ■, left kidney ■, right kidney ■, left psoas major muscle (pmm) ■ and right pmm ■. Besides our described architecture we test three other baselines of the approach: A SSPNet employing only 1×1 convolutions instead of the GCNN, the GCNN without sampling coordinates as additional input features and the network of 1×1 convolutions with sampling coordinates as additional features.

Results: Qualitative and quantitative results are depicted in Figure 3. The GCNN outperforms the network with only 1×1 convolutions in both cases with and without sampling coordinates as additional features, although the result is much clearer in the second case (ODS of .690 against .786). The best SSPNet with an ODS of .827 yields a higher score than the UNet-S and Unet-M and performs only slightly worse than the UNet-L (ODS of .769, .791 and .834 respectively). Without class-weighting the UNet variants perform worse with ODS values of .763, .773 and .817, respectively. In contrast, the SSPNet showed similar results with and without class weighting. The visual comparison shows a clearer outline of the psosas muscels and a better detection of the unary bladder in favor of the SSPNet.



Method	Parameters	Samples	ODS
SSPNet – GCNN + coords.	˜ 500k	2000	**.900**
UNet-S	˜ 500k	dense	.874
UNet-M	˜ 900k	dense	.878
UNet-L	˜ 1300k	dense	**.884**

Figure 4: Qualitative and quantitative results on the JSRT chest X-ray database. The images (from left to right: original X-ray, ground truth, Unet-L, SSPNet) show edge overlays from five anatomical structures: left lung ■, right lung ■, left clavicle ■, right clavicle ■ and heart ■. The UNet misses parts of the edges of the heart whereas our approach successfully follows informative gradients along its outline.

3.2. JSRT

The JSRT database consists of 247 chest X-ray images that were downsampled to dimensions of 256×256. A four-fold cross validation was employed to compute the results. We test the SSP-Net with additional sampling coordinates as input features against the three UNet implementations UNet-S, UNet-M and UNet-L. Ground truth labels are generated from the provided landmarks for five anatomical structures: left lung ■, right lung ■, left clavicle ■, right clavicle ■ and heart ■ .

Results: Qualitative and quantitative results are depicted in Figure 4. The SSPNet yields a slightly higher OSD score of .900 than the UNet-L with .884, though visual results are mostly comparable. However, in some cases the UNet misses parts of the edges of the heart whereas the SSPNet can follow informative gradients along its outline.

4. Discussion and Conclusion

In this work we proposed a new approach for structured prediction for semantic edge detection from a few sparse sampling locations on an image. To the best of our knowledge the SSPNet is the first deep learning network that combines structured prediction with CNNs and global context aggregation with graph convolutions to recover a dense output. In our experiments on VISCERAL and JSRT we showed that the SSPNet performed better or on par with several UNet variants while also having the lowest number of trainable parameters.

For future work, incorporating the SSPNet in an end-to-end learning framework instead of working with an explicitly trained sample CNN is clearly of high interest. This may be achieved by using a more complex GCNN model with attention mechanisms, e.g. (Monti et al., 2018), which could lead the selection of sampling locations. With an extension to 3D volumes, our approach can be evaluated on medical datasets with stronger memory and computational limitations. While in this work the focus was on edge detection, other tasks for structured prediction, such as landmark detection in medical images, may also be suited well for our approach.

In conclusion, we showed that our SSPNet is feasible for semantic edge detection in medical images and we believe that it can be used as a potential alternative to dense encoder-decoder architectures for general pixel-level image tasks in deep learning.

Acknowledgments

We would like to thank the reviewers for their many insightful comments and suggestions helping to improve our paper. We gratefully acknowledge the support of the NVIDIA Corporation with their GPU donations for this research.

References

James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1993–2001, 2016.

Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *Signal Processing Magazine*, 34(4):18–42, 2017.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2013.

John Canny. A computational approach to edge detection. *Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In *European Conference on Computer Vision (ECCV)*, pages 562–578, 2018.

Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *International Conference on Computer Vision (CVPR)*, pages 1841–1848, 2013.

Lasse Hansen, Jasper Diesel, and Mattias P Heinrich. Multi-kernel diffusion cnns for graph-based learning on point clouds. *arXiv preprint arXiv:1809.05370*, 2018.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *Transactions on Medical Imaging*, 35(11):2459–2475, 2016.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Distance metric learning using graph convolutional networks: Application to functional brain networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 469–477, 2017.

Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 174–182, 2018.

C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes. Robust and accurate shape model matching using random forest regression-voting. *Transactions on Pattern Analysis & Machine Intelligence*, 37(9):1862–1874, 2015.

Yun Liu, Ming-Ming Cheng, JiaWang Bian, Le Zhang, Peng-Tao Jiang, and Yang Cao. Semantic edge detection with diverse deep supervision. *arXiv preprint arXiv:1804.02864*, 2018.

Federico Monti, Oleksandr Shchur, Aleksandar Bojchevski, Or Litany, Stephan Günnemann, and Michael M Bronstein. Dual-primal graph convolutional networks. *arXiv preprint arXiv:1806.00770*, 2018.

Ozan Oktay, Andreas Schuh, Martin Rajchl, Kevin Keraudren, Alberto Gomez, Mattias P Heinrich, Graeme Penney, and Daniel Rueckert. Structured decision forests for multi-modal ultrasound image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 363–371, 2015.

Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrerro Moreno, Ben Glocker, and Daniel Rueckert. Spectral graph convolutions for population-based disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 177–185, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development

of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *International Conference on Computer Vision (CVPR)*, pages 1395–1403, 2015.

Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26, 2017.