Structure Prediction for Gland Segmentation with Hand-Crafted and Deep Convolutional Features

Siyamalan Manivannan, *Member, IEEE,* Wenqi Li, *Member, IEEE,* Jianguo Zhang, Emanuele Trucco, and Stephen McKenna, *Senior Member, IEEE*

***Abstract*—We present a novel method to segment instances of glandular structures from colon histopathology images. We use a structure learning approach which represents local spatial configurations of class labels, capturing structural information normally ignored by sliding-window methods. This allows us to reveal different spatial structures of pixel labels (e.g., locations between adjacent glands, or far from glands), and to identify correctly neighbouring glandular structures as separate instances.**

Exemplars of label structures are obtained via clustering and used to train support vector machine classifiers. The label structures predicted are then combined and post-processed to obtain segmentation maps. We combine hand-crafted, multi-scale image features with features computed by a deep convolutional network trained to map images to segmentation maps.

We evaluate the proposed method on the public domain GlaS dataset, which allows extensive comparisons with recent, alternative methods. Using the GlaS contest protocol, our method achieves the overall best performance.

***Index Terms*—Molecular and cellular imaging; Gastrointestinal tract; Segmentation.**

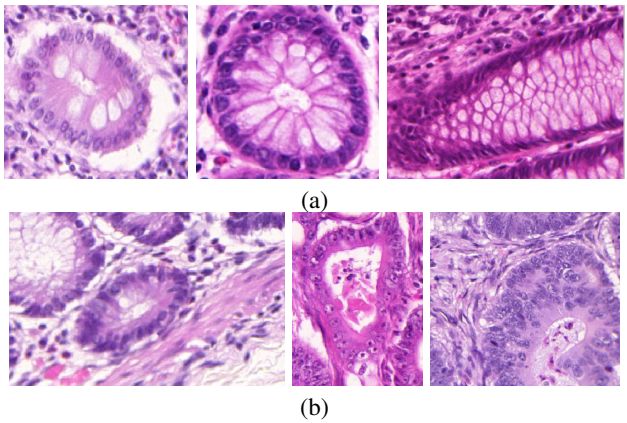


Fig. 1: Glandular structures in the Warwick-QU dataset [2]. (a) Glands in healthy tissue. (b) Left to right: adenoma, moderately differentiated, and poorly differentiated adenocarcinoma.

I. Introduction

Histological assessment of gland formation and morphology informs diagnosis, prognosis and treatment planning of patients [1]. It is useful for grading of adenocarcinomas in colon, breast, and prostate. Such assessment is labour intensive, performed by highly trained pathologists, and often has limited reproducibility. The emergence of whole-slide imaging is increasing the volume of digital histology image data to be analysed, exacerbating the problem. Algorithms capable of reliably segmenting glandular structures automatically would accelerate analysis and provide reproducible, quantitative measures of gland morphology. The development of such algorithms is challenging because malignancy results in irregular morphology and poorly differentiated gland boundaries, and because glandular structures can be closely packed together but need to be segmented as separate instances. Glands in healthy epithelial tissue have a clear structure with interior lumen surrounded by columnar epilthelial cells (Fig. 1 (a)). This structure degenerates in moderately or poorly differentiated adenocarcinomas (Fig. 1 (b)).

Gland segmentation, and more generally semantic pixel labelling, often incorporates a sliding window classification procedure based on features extracted from a local window centred at each image location (e.g. [3], [4]). Such a procedure ignores the class labels' spatial structure. Instead, we propose to learn discriminative models for segmentation in which local spatial structures are encoded in the label (output) space as well as in the feature (input) space. By directly employing label structure we can more reliably separate objects and thus improve instance segmentation. The number of possible label structures grows exponentially as the size of the local region considered increases, posing a challenge. We show how this large output space can be handled by combining small numbers of local structure exemplars obtained via clustering.

We combine hand-crafted features with learned deep convolutional features to capture image context information. We conduct experiments with the publicly available GlaS dataset [2] showing that the proposed use of local structure prediction improves gland segmentation compared with using binary classifiers. Direct comparison with other published results indicates that our method is the top ranked.

This paper follows on from short conference papers that evaluated choices of image features [4] and a preliminary version of local structure prediction [5]. Contributions of this paper with respect to those earlier ones include the following.

1) We summarise the literature on gland segmentation, reviewing progress over the last decade.
2) We incorporate features learned using fully convolutional networks (FCN) into the local structure prediction framework, whereas previously [5] we tested the

S. Manivannan is with the University of Jaffna, Sri Lanka. W. Li is with University College London, UK. J. Zhang, E. Trucco, and S. J. McKenna are with the University of Dundee (Computer Vision and Image Processing (CVIP), School of Science and Engineering), UK. All authors were with University of Dundee during the initial stages of the research reported here.

Manuscript received Month day, year; revised Month day, year.

feasibility of structure learning using only hand-crafted features.
3) We evaluate combining feature types in the context of local structure prediction, whereas previously [4] we combined features only in a binary (gland vs non-gland) setting.
4) We investigate the effect of the number of structure exemplars (clusters) at both training and test time.
5) We employ the full GlaS dataset [2] for all evaluations presented in this paper whereas our previously published experiments used only a subset of it.
6) The GlaS dataset allows extensive comparisons with recent, alternative methods; using the GlaS contest protocol, our method achieves the best overall performance.

II. Related Work

Here we review how gland segmentation has progressed over the past decade. Early methods attempted to segment glandular structures by first explicitly identifying substructures such as nuclei and lumen. Farjam et al. [6] performed k-means clustering of local texture features to distinguish stroma and lumen from regions more densely populated with nuclei. However, robust gland segmentation requires more domain knowledge to be incorporated whether modeled explicitly or acquired via machine learning with a supervision mechanism. A popular approach has been to classify pixels based on colour, identify candidate lumen regions, and run either region growing or contour-based segmentation initialised at each of these candidates. Wu et al. [7] described a region growing algorithm initialised in lumen regions obtained by thresholding. Naik et al. [8] used supervised pixel colour classification to label pixels as nuclear, cytoplasmic, or lumen. Candidate gland lumen regions were identified based on size and bordering epithelial cytoplasm. These were used to initialise level-sets contour segmentation; contours evolved outward with stopping gradient based on nuclei likelihood. In a similar spirit, Nguyen et al. [9] grouped nuclear and cytoplasmic pixels to obtain gland boundary segments and then grew lumen regions in a controlled way until they met with surrounding gland boundary segments. Gunduz-Demir et al. [10] used k-means colour clustering to identify pixel clusters corresponding approximately to nuclei and lumen. They ran an iterative algorithm to fit discs inside nuclear regions and lumen regions. They then clustered lumen discs into two clusters based on features including size and displacement of neighbouring discs. Finally, lumen discs in the cluster more likely to represent glands were used to seed region growing constrained by line segments joining proximal pairs of nuclear discs. More recently, Cohen et al. [11] classified pixels as nuclei, immune system, lumen, cytoplasm, stroma, and goblet cells based on local colour statistics using two stages of random forest classification. Candidate lumen boundaries were then used to initialise active contours with external forces designed to attract the contour to nuclear pixels and repel it from stroma and immune system pixels, encouraging it to stop at the boundary of the nuclear layer at the gland periphery. A final classification step reduced false positives based on shape features and pixel labels. The methods described above, based on iterative segmentation initialised at lumina and terminated based on constraints provided by pixels classified as nuclear, can work well for well-formed glandular structures. However, they will fail when the spatial assumptions on which they are based are badly violated. This will often be the case for malignant glands with deformation of gland morphology.

Ben Cheikh et al. [12] used colour classification to locate cell nuclei. They applied advanced morphological operators to nuclear objects to obtain candidate epithelial layers and gland central regions. These were combined to obtain glandular structures. Nguyen et al. [13] formulated gland segmentation as a graph cuts problem, constructing graphs with nuclei and lumen as nodes. Nuclei were detected based on radial symmetry and classified as epithelial or stromal using a support vector machine (SVM) based on local texture features. This method was able to detect glands without lumen and glands with multiple lumina. Sirinukunwattana et al. [14], [15] found candidate glands by classifying superpixels as gland or non-gland based on colour and texture features extracted from superpixel neighbourhoods. In [15] they initialised a polygonal contour model for each such candidate and inferred both the number of vertices in the polygon and their location based on reversible-jump Markov chain Monte Carlo. After post-processing to remove some false positives, the MAP contours obtained compared favorably with several previous algorithms. Scattering coefficients have been used as texture features in the computation of glandular structure maps [14]. Their use as input to a convolutional neural network (CNN) in addition to raw image values to detect tumour cells in histology images, showed that CNN can perform better when the input consists of a combination of handcrafted features and raw data [16].

Research on gland segmentation was invigorated by the GlaS contest [2]. Its focus on shared data and comparative evaluation of methods in a controlled setting yielded an informative snapshot of the state-of-the-art. Several of the most highly ranked GlaS entries were based on sliding window classifiers incorporating CNNs. The 5th-placed entry, CVML [2], used a CNN trained to classify small 19×19 pixel windows into three classes representing gland lumen, epithelial cells forming the gland boundary, and inter-gland tissue; class probability maps thus obtained were used to drive level set segmentation. In general, sliding window classifiers trained to classify gland versus non-gland pixels can result in neighbouring glandular structures being erroneously merged. In an effort to prevent this, the 2nd-placed entry from ExB [2] trained a CNN with two paths, one to classify gland versus non-gland pixels, the other to classify pixels close to gland boundaries versus all other pixels. For the same reason, Kainz et al. [3] annotated pixels close to at least two gland objects in the training data and trained a CNN to classify windows as centred on such pixels or not. A second CNN was trained to classify windows as centred on gland or non-gland from either benign or malignant tissue (i.e. four classes). An additive combination of outputs from these two classifiers gave better segmentation results than the latter CNN alone. Finally, gland segmentation was refined using convex geodesic active contours. The 4th-placed entry, from our group, used

SVM classifiers with features from a CNN combined with features extracted from multi-scale patches [4]. Fu et al. [17] developed a sliding window detector in which windows were circular. A conditional random field (CRF) model was trained, after transforming windows to polar coordinates, to find the closed contour in each window most likely to be a gland boundary. Support vector regression on pyramid HOG features was used to select the strongest gland candidates from the CRF. This method assumes glands are star shaped (in the sense of Veksler [18]) which is not always the case, especially with malignant glands.

Fully convolutional networks (FCNs) can be trained end-to-end to map images directly to their segmentation maps [19]. Their typical contracting network architecture, in which consecutive convolution layers are interspersed with spatial pooling operations, can result in FCN outputs having low resolution. However, subsequent upsampling operators and convolutions can be used to learn more precise output. Four methods for gland segmentation based on FCN variants with image-to-segmentation-map training have been proposed in the literature [20–23]. Ben Taieb and Hamarneh [20] used a loss function for a deep FCN with penalty terms that encouraged gland boundary smoothness and correct label hierarchy. An indicator function was used in the loss function to indicate whether or not an assignment was valid. Their experiments suggest that this can help improve gland segmentation. Three other methods based on FCN incorporate some mechanism to help avoid nearby neighbouring gland structures from merging erroneously. The U-net of Ronneberger et al. [22] (the third ranked team in GlaS) is an FCN modified to yield more precise segmentation. It learns to map a raw RGB image to a binary gland segmentation. As well as a contracting analysis path and an up-sampling synthesis path with many feature channels, this network combines high-resolution features from the contracting path with the upsampling layers so that a successive convolution layer can learn to assemble a more precise output. A high pixel-wise loss was used for pixels in gaps between glandular structures in the training set. The winning GlaS entry from Chen et al. [21] used an FCN combining upsampling from layers at different depths to enhance multi-scale analysis due to the varying effective receptive field size. Their network was trained simultaneously to output both a gland foreground map and a gland boundary map. These maps were then logically combined to obtain a gland foreground map in which nearby glands were kept separate. More recently, Xu et al. [23] incorporated boundary maps into an FCN with a complex structure. A deep convolutional channel predicted a gland foreground map; outputs from N of its convolutional layers were fed as inputs to a side channel which predicted a gland boundary map through a linear combination of maps computed from each of the N stages it was fed. A final CNN stage combined these maps to predict a gland instance map. They reported state-of-the-art results on the GlaS dataset.

In summary, today's most successful methods are based on supervised machine learning, typically incorporating CNNs enhanced by some mechanism to prevent neighbouring gland structures from merging. This contrasts with earlier methods which tended to use pipelines reliant on detection of components such as lumen and nuclei to seed and constrain multiple instances of region growing or contour search.

III. Method

Gland segmentation takes a histology image as input and outputs a label image in which pixel values denote gland or non-gland. We formulate this in terms of local label structure prediction. In summary, for each location on a rectangular grid we extract image features and apply support vector machine classifiers to predict a local label patch centred at that location. These label patches are obtained as combinations of label structure exemplars. Neighbouring label patch predictions overlap so they are averaged. A post-processing step is applied to identify the regions corresponding to individual glandular structures. We first describe label structure classification and post-processing before giving details of the features used to capture image context using both handcrafted features and deep convolutional networks.

A. Structure prediction

We denote a labelled data set as $\{(I_i, G_i)\}, i = 1, \ldots, n$, where I_i is an image and G_i its ground truth annotation. In G_i, each gland region is represented by pixels assigned a unique positive integer while non-gland (background) regions are zero. Figs. 10 and 11 show examples with different gland regions mapped to different colours. Let $\mathbf{x}_{ij} \in \mathbb{R}^d$ denote the feature representation at point $\mathbf{s}_{ij}$ of image I_i. Let $\mathbf{u}_{ij}$ denote a label patch extracted at the same location from the corresponding binary ground truth segmentation map. Fig. 3 shows some examples of label patches. Let $u_{ijk} \in [0, 1]$ denote the k^{th} element of $\mathbf{u}_{ij}$, i.e., the k^{th} location in the label patch. If this location is definitely foreground (gland) then $u_{ijk} = 1$. If it is definitely background (non-gland)then $u_{ijk} = 0$.

A common approach (e.g. [24]) is to train a binary classifier on a set of labeled windows, $\{\mathbf{x}_{ij}, y_{ij}\}$, where $y_{ij} \in \{0, 1\}$. A binary label, y_{ij}, can be computed from $\mathbf{u}_{ij}$ by thresholding:

$$y_{ij} = \begin{cases} 1 & \frac{1}{d'} \sum_{k=1}^{d'} u_{ijk} > t \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where t is a user-specified threshold. Instead, our proposed method directly finds a mapping from the input feature space to a set of *label exemplars* $\{\mathbf{v}_k\}, k = 1, \ldots, K$. At test time, the method directly predicts the local structure of the labels for any given image location (see Fig. 2) using these exemplars. The exemplars can be thought of as visual words for binary images. A labelling can be reconstructed as a weighted combination of those exemplars. The exemplars can be obtained, for example, by clustering training label patches $\{\mathbf{u}_{ij}\}$ and treating each cluster center, $\mathbf{v}_k$, as an exemplar. Fig. 4 shows exemplars obtained using K-means.

Once we have K exemplars, structure classifiers are defined, each of which separates a label configuration $\mathbf{v}_k$ from other configurations $\mathbf{v}_m, \forall m, m \neq k$. We use linear classifiers,

$$f_k(\mathbf{x}_{ij}) = \mathbf{w}_k^T \mathbf{x}_{ij} + b_k \quad (2)$$

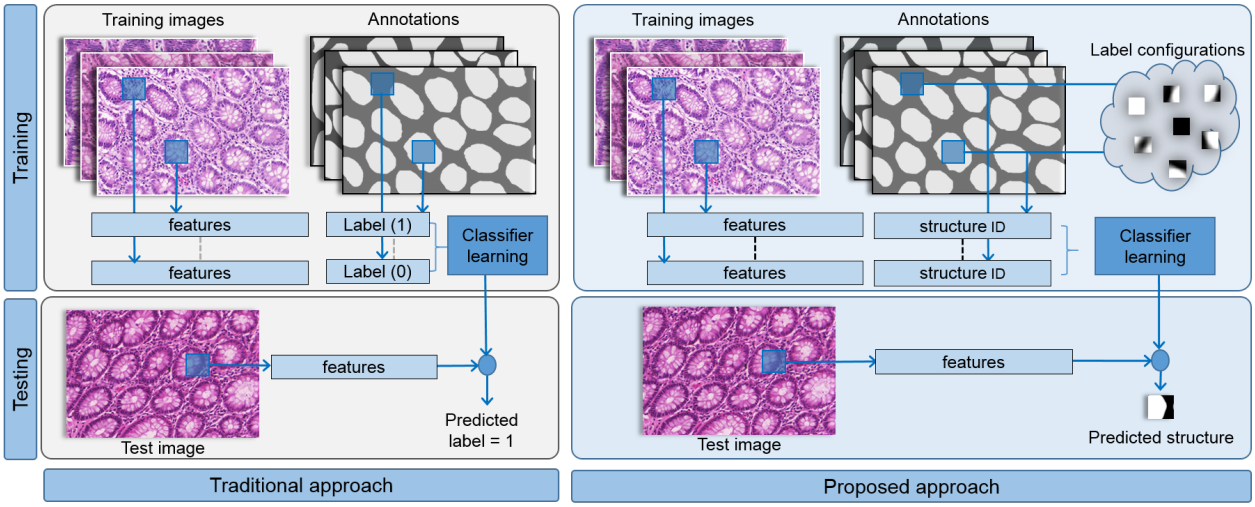
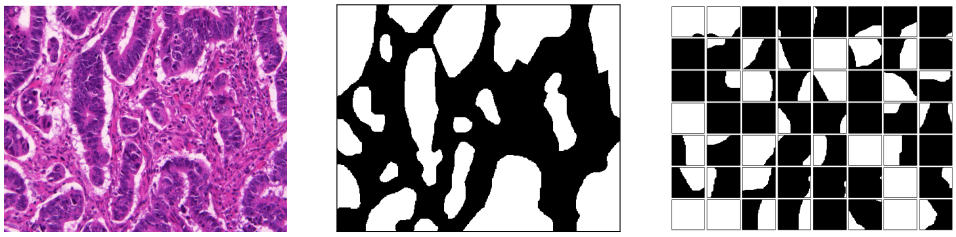


Fig. 2: Left: a traditional approach to patch-based segmentation. Right: using local structure learning and prediction.



(a) (b) (c)

Fig. 3: (a) An image, (b) its ground truth, and (c) example label patches extracted from (b).

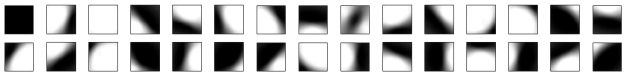


Fig. 4: Label exemplars obtained using K-means ($K = 30$).

and learn $(\mathbf{w}_k, b_k)$ using an SVM optimization,

$$\begin{aligned}\arg\min_{\mathbf{w}_k, b_k} \frac{1}{2}\|\mathbf{w}_k\|_2^2 + \frac{\lambda}{|\mathcal{U}|}\sum_{i,j\in\mathcal{U}} \max(0, 1 - \mathbf{w}_k^T\mathbf{x}_{ij} - b_k) \\ + \frac{\lambda}{|\bar{\mathcal{U}}|}\sum_{i,j\in\bar{\mathcal{U}}} \max(0, 1 + \mathbf{w}_k^T\mathbf{x}_{ij} + b_k)\end{aligned} \quad (3)$$

where λ is a regularization parameter and $\mathcal{U}$ is defined as

$$\mathcal{U} = \{i, j \mid \|\mathbf{u}_{ij} - \mathbf{v}_k\|_2^2 \leq \|\mathbf{u}_{ij} - \mathbf{v}_m\|_2^2,\ \forall m, m \neq k\}. \quad (4)$$

$\bar{\mathcal{U}}$ is the complement of $\mathcal{U}$. We used the LibLinear library [25] to implement (3). The output of each classifier, f_k, is calibrated using Platt scaling [26] to obtain probabilities $p_k(\mathbf{x}_{ij})$ using the logistic function,

$$p_k(\mathbf{x}_{ij}) = \frac{1}{1 + \exp^{-A_k f_k(\mathbf{x}_{ij}) - B_k}} \quad (5)$$

where A_k and B_k are two free parameters to be learned. For a given test image location $\mathbf{s}_{ij}$, the learned classifiers $\{(\mathbf{w}_k, b_k)\}, k = 1, \ldots, K$ output the probabilities:

$$\mathbf{p}_{ij} = [p_1(\mathbf{x}_{ij}), \ldots, p_K(\mathbf{x}_{ij})]\,. \quad (6)$$

Let $\mathcal{P}$ represent a set of $r(\leq K)$ indices which correspond to the largest r values in $\mathbf{p}_{ij}$. Renormalising,

$$q_k(\mathbf{x}_{ij}) = \frac{p_k(\mathbf{x}_{ij})}{\sum_{m\in\mathcal{P}} p_m(\mathbf{x}_{ij})},\ \forall k \in \mathcal{P}, \quad (7)$$

we obtain a distribution $\mathbf{q}_{ij} = [q_1(\mathbf{x}_{ij}), \ldots, q_r(\mathbf{x}_{ij})]$ that indicates the extent to which the label structure at $\mathbf{s}_{ij}$ is exemplified by each of the r most relevant exemplars. The label patch $\mathbf{u}_{ij}$ at a given test image location s_{ij} can be reconstructed by weighting the label exemplars accordingly:

$$\mathbf{u}_{ij} \approx \sum_{r\in\mathcal{P}} q_r(\mathbf{x}_{ij})\mathbf{v}_r \quad (8)$$

In this way the local label structure can be reconstructed from a few exemplars. When $r = 1$ this amounts to selecting the exemplar with the highest Platt-scaled classification score.

B. Post-processing

Structure prediction (Section III-A) outputs a label window centred on each location in a test image (Eq. (8)). Nearby

label windows overlap and so are averaged to obtain a map in which higher values correspond to probable gland locations. Example label maps are shown in Figs. 10 and 11. To segment individual glands from this map, we apply a fixed threshold of T (estimated from training data, see IV-C) followed by morphological erosion with circular structuring element of radius 5 pixels to help reduce any connectivity between adjacent glands. Small connected components (area < 900 pixels) are discarded. (To give an idea of scale, images in Fig. 10 are 775×522 pixels). Dilation with structuring element of radius 10 pixels restores objects to their original size given that the system was trained using the ground truth images which had been eroded using a structuring element of radius 5. Finally, a hole filling algorithm is run to remove holes in foreground regions.

C. Feature representation

Let $\mathbf{s}_{ij}$ be the j^{th} sampling point from image I_i. We use two sets of features to represent image context around $\mathbf{s}_{ij}$: deep features extracted by applying a trained FCN, and HC features with locality-constrained linear coding (LLC) [27]. These two representations are normalized independently using the square-root and L2 normalizations as in [28] and concatenated.

1) Fully convolutional neural networks: To capture multi-level contextual information we make use of a fully convolutional neural network [19]. This network can be trained in an end-to-end (image-to-image) manner, which takes an image as input and produces a correspondingly-sized probability map in a single forward propagation. The network contains a down-sampling path and an up-sampling path (Fig. 5). The down-sampling path contains convolutional and max pooling layers, and aims at extracting the high level (coarse) abstraction information. The up-sampling path contains convolution and up-sampling layers which try to extract the fine (pixel-level) detail. We use a transfer learning approach to mitigate the challenge of insufficient training data. Our starting point is the pretrained FCN model from [19], an FCN-8s architecture with 21 layers, trained using ImageNet and fine-tuned on the Pascal VOC dataset. For gland segmentation, we append two convolutional layers. The first, containing 512 output channels, acts as a feature extractor; the second, containing 2 channels, provides foreground (gland) and background (non-gland) scores. By leveraging an existing pre-trained network, this design keeps relatively low the number of new parameters to be learned for gland segmentation thus reducing the cost of training and the risk of over-fitting on a dataset of relatively small scale. In structure prediction experiments we use as features the penultimate layer's output (512 channels).

The network was implemented using Caffe [29]. Parameters in the new layers were initialized using the "Xavier" initialization [30] and the other 21 layers were initialized as the pre-trained FCN from [19]. The whole network was then trained in an end-to-end fashion by stochastic gradient descent with maximum number of iterations set to $75,000$. Since we need to fine-tune the FCN-8s architecture (shaded blue in Fig. 5) and to learn from scratch the parameters of the new layers (shaded red in Fig. 5), we set a higher learning rate for the latter (10^{-4}) than for the former (10^{-5} for FCN-8s).

The network was trained on randomly cropped sub-images of size 384×384 pixels. Data augmentation (rotations and flipping) was used to increase effectiveness of the network while reducing risk of over-fitting. At test time, we extracted overlapping sub-images of size 384×384 with overlap of 200 pixels. We averaged adjacent probability maps to obtain the final prediction for an image. This sliding window approach reduces memory requirements compared to an FCN applied to entire images.

2) Hand-crafted features: Our second representation is inspired by *zoom-out features* [31] and was built by concatenating window descriptors computed from concentric windows of sizes 48×48, 80×80, 128×128, and 200×200 centered at $\mathbf{s}_{ij}$, as well as from the entire image. In addition, to capture local fine structure, the 48×48 window was divided into nine 16×16-pixel windows and the feature representations from these windows also concatenated. This is illustrated in Fig. 6.

Within each window, root-SIFT [32], vectorized raw-pixel values, and multi-resolution local patterns [33] were extracted from patches of size 16×16 with a step size of 2 pixels. For each feature type, features extracted from the three color channels (R, G, and B) were concatenated. Average pooling was used to get window representations from the dictionary-encoded features with a dictionary size of 200. (See [4] for experiments with different dictionary sizes). We used square-root and L2 normalizations [28] to normalize the pooled encoded features from each individual window before concatenation. The final dimensionality of HC features was 8400 (14 windows $\times$ size of the dictionary $\times 3$ features).

IV. Experiments

A. Dataset

The Warwick-QU dataset was used for evaluation [15]. It formed the basis of the Gland Segmentation (GlaS) Challenge Contest hosted by MICCAI [2] and is now publicly available[1]. It consists of 165 images each annotated with an associated 'ground truth' segmentation and histological grade (*benign* or *malignant*). These 165 images were extracted from a set of 52 visual fields which had been selected from 16 H&E-stained whole-slide images (from 16 patients) of stage T3 or T4 colorectal adenocarcinoma. They had been scaled to have pixel resolution of $0.620\mu m$ ($20\times$ magnification) and were typically of size 775×522 pixels. An expert pathologist had provided ground truth which involved assessing the grade of each visual field and delineating the glandular structures. In line with GlaS Challenge protocol, we used the provided split into a training part and two test parts: test part A and test part B. The training part has 85 images: 37 from visual fields graded as benign and 48 from fields graded as malignant. Test A has 60 images (33 benign, 27 malignant) and test B has 20 images (4 benign and 16 malignant). Figs. 10 and 11 show example images. Table I summarises dataset composition.

[1] http://www2.warwick.ac.uk/fac/sci/dcs/research/combi/research/bic/glascontest/

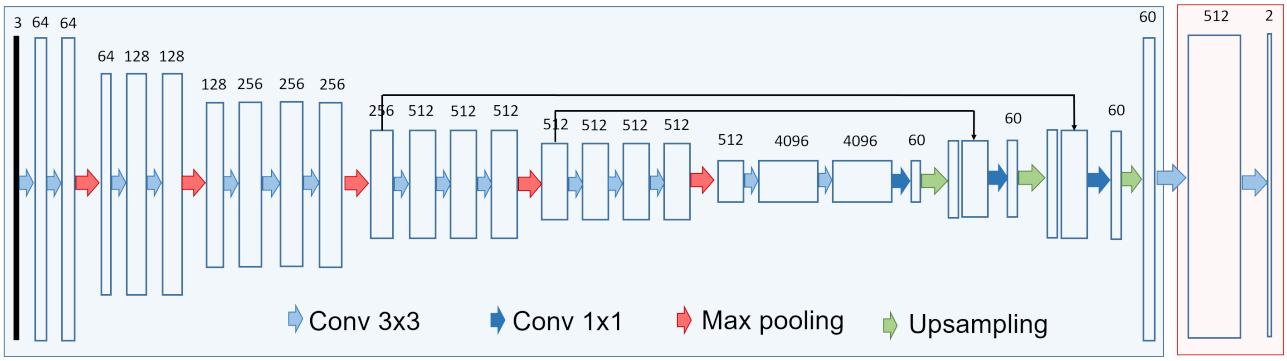


Fig. 5: Proposed FCN architecture. Blue: FCN-8s [19]. Red: new layers appended in this work.

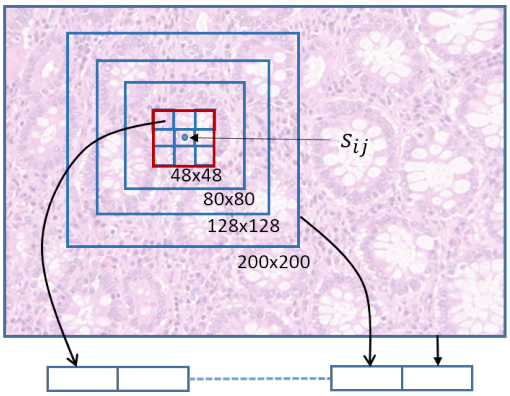


Fig. 6: Feature representation at s_{ij} by HC features.

Histological grade	Number of images		
	Training	Test A	Test B
Benign	37	33	4
Malignant	48	27	16

TABLE I: Dataset composition.

B. Evaluation measures

Evaluation used three criteria, following [2]: detection accuracy, segmentation score, and shape dissimilarity. Mean values over all the test images under each criteria were used to rank different methods.

1) Detection accuracy: The F1 score is employed to measure the detection accuracy of individual glandular objects:

$$F1 = \frac{2PR}{P+R},\ P = \frac{N_{TP}}{N_{TP}+N_{FP}},\ R = \frac{N_{TP}}{N_{TP}+N_{FN}}$$

where N_{TP}, N_{FP}, and N_{FN} denote the number of true positives, false positives, and false negatives, respectively. Correspondence is established between each segmented instance and the ground truth object that has maximum overlap with it. A segmented instance that intersects with at least 50% of its corresponding ground truth object is considered a true positive, otherwise it is considered a false positive. A ground truth object that has no corresponding segmented instance, or that has less than 50% of its area overlapped by its corresponding segmented instance, is considered a false negative.

2) Segmentation score: Pixel-level Dice score of a segmentation $\mathcal{O}$ with ground truth $\mathcal{G}$ is defined as

$$D_p(\mathcal{G},\mathcal{O}) = \frac{2|\mathcal{G}\cap\mathcal{O}|}{|\mathcal{G}|\cup|\mathcal{O}|}$$

where $|\cdot|$ denotes set cardinality. Object-level Dice score is then defined as

$$D_o(G_i,O_i) = \frac{1}{2}\left[\sum_{j=1}^{n_O}\omega_i D_p(G_{ij},O_{ij}) + \sum_{j=1}^{n_G}\tilde{\omega}_i D_p(\tilde{G}_{ij},\tilde{O}_{ij})\right]$$

where $w_j = |O_{ij}|/\sum_{k=0}^{n_O}|O_{ik}|$ and $\tilde{w}_j = |\tilde{G}_{ij}|/\sum_{k=0}^{n_G}|\tilde{G}_{ik}|$. G_{ij} is the j^{th} ground-truth object that maximally overlaps with the segmentation O_{ij}, and $\tilde{O}_{ij}$ is the j^{th} segmentation which maximally overlaps with the ground-truth object $\tilde{G}_{ij}$. n_G and n_O are the total number of ground-truth objects, and segmented objects in the images G_i and O_i respectively.

3) Shape dissimilarity: Shape dissimilarity of segmented object and ground truth is measured as Hausdorff distance:

$$H(\mathcal{G},\mathcal{O}) = \max\{\sup_{x\in\mathcal{G}}\inf_{y\in\mathcal{O}}\|x-y\|, \sup_{x\in\mathcal{O}}\inf_{y\in\mathcal{G}}\|x-y\|\}$$

An object-level measure is then defined as

$$H_o(G_i,O_i) = \frac{1}{2}\left[\sum_{j=1}^{n_O}\omega_i H_p(G_{ij},O_{ij}) + \sum_{j=1}^{n_G}\tilde{\omega}_i H_p(\tilde{G}_{ij},\tilde{O}_{ij})\right]$$

C. Experimental settings

Structure predictors were applied at locations on a grid with spacing of 8 pixels. We set the size of each patch, $\mathbf{u}_{ij}$, to 48×48 pixels, approximately $30\times30\mu m^2$. We used k-means to learn 48×48-pixel exemplars (Fig. 4). Parameter λ in Eq. (3) was set to $\lambda = 1$; solvers in liblinear are known not to be very sensitive to this parameter [25]. For all the reported binary segmentation methods we used a relatively high value of $t = 0.8$ (Eq. (1)) to encourage separation of adjacent glands. We used data augmentation at training and test time. Four instances of each classifier were trained, each on a rotated version of the training data ($\{0°, 90°, 180°, 270°\}$). At test time, 16 prediction maps were obtained for each image (4

Methods	Mean values						Median values					
	F1 score		Obj dice		Obj Hausdorff		F1 score		Obj dice		Obj Hausdorff	
	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B
Binary (HC)	0.718	0.675	0.692	0.695	111.0	160.0	0.724	0.697	0.690	0.729	93.2	140.3
Binary (CNN)	0.758	0.690	0.688	0.671	94.9	166.5	0.772	0.769	0.724	0.746	66.9	133.1
Binary (HC+CNN)	0.784	0.722	0.799	0.776	75.0	139.0	0.809	0.817	0.820	0.879	59.2	74.3
Ours (HC)	0.838	0.723	0.834	0.747	67.0	122.4	0.857	0.714	0.859	0.809	55.9	116.1
Ours (CNN)	0.870	0.749	0.868	0.832	55.8	105.4	0.882	0.845	0.874	0.899	39.7	**44.8**
Ours (HC+CNN)	**0.892**	**0.801**	**0.887**	**0.853**	**51.2**	**87.0**	**0.930**	**0.857**	**0.941**	**0.914**	**23.4**	45.7

TABLE II: The proposed method (K=100) vs. binary prediction with different features. Mean and median values of the evaluation measures over the images are reported.

Post-processing	Binary (HC+CNN)						Ours (HC+CNN)					
	F1 score		Obj dice		Obj Hausdorff		F1 score		Obj dice		Obj Hausdorff	
	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B	Test A	Test B
No post-processing	0.747	0.629	0.743	0.746	92.7	141.4	0.657	0.546	0.849	0.817	63.6	102.0
remove small isolated regions	0.781	0.724	0.743	0.747	93.2	141.5	0.849	0.763	0.852	0.818	63.0	101.5
Erosion + remove small isolated regions + dilation	**0.795**	**0.729**	0.795	0.769	**73.9**	140.5	**0.892**	**0.801**	0.881	0.845	52.1	88.5
Erosion + remove small isolated regions + dilation + fill holes	0.784	0.722	**0.799**	**0.776**	75.0	**139.0**	**0.892**	**0.801**	**0.887**	**0.853**	**51.2**	**87.0**

TABLE III: Effect of post-processing on mean values of performance measures. (Post-processing steps are explained in Section III-B).

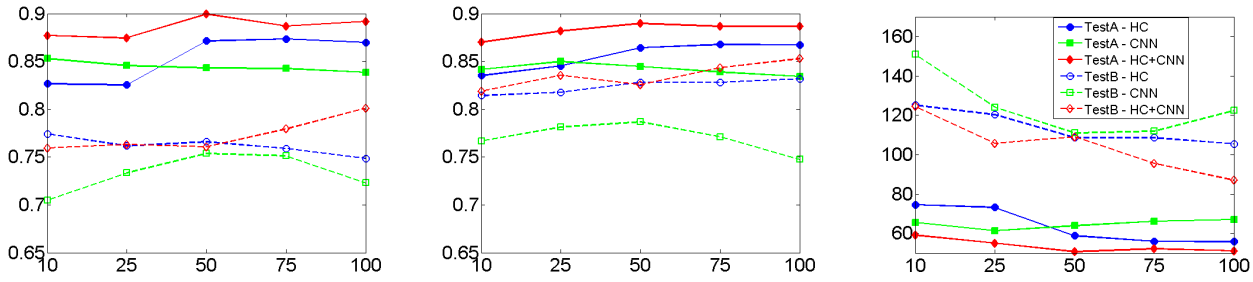
rotations of the test image $\times$ 4 classifiers). These prediction maps were averaged to get the final prediction map for that image (as in [33]). The threshold T in III-B was automatically selected such that the selected T maximizes the average of object level F1 and Dice scores on training data.

D. Evaluation

Results in Table II compare the proposed method with the binary prediction method of Eq. (1). Structure prediction gave better mean and median values than binary prediction in terms of all measures. (Higher values are better for F1 and Dice scores; lower values are better for Hausdorff distance.) Fig. 7 explores the effect of the number of structure exemplars, K. When CNN features were used, performance peaked at around $K = 50$ for test set B and at around $K = 25$ for test set A. This is consistent with the fact that test set B has a higher proportion of malignant cases; these contain irregularly shaped glandular structures and so we would expect a more complex representation of label structure to be of benefit. When HC features are used either alone or in combination with CNN features, it appears that still larger values of K can be beneficial. Boxplots in Fig. 8 provide a comparison of structure prediction (with different values of K) and binary prediction, in the case of combined HC and CNN features. Considering the median values (red lines) these are consistent with Fig. 7. These boxplots highlight that performance varies considerably between images with a few particularly challenging images being very poorly segmented. Fig. 10 shows examples for which stucture prediction gave better segmentation than binary prediction; adjacent gland boundaries are more reliably separated by our method. Fig. 11 shows two cases for which binary prediction gave better agreement with the annotated ground-truth, although qualitative subjective comparison may lead the reader to question whether it did in fact give better segmentations. Fig. 12 shows two challenging examples that resulted in structure predictions that were outliers.

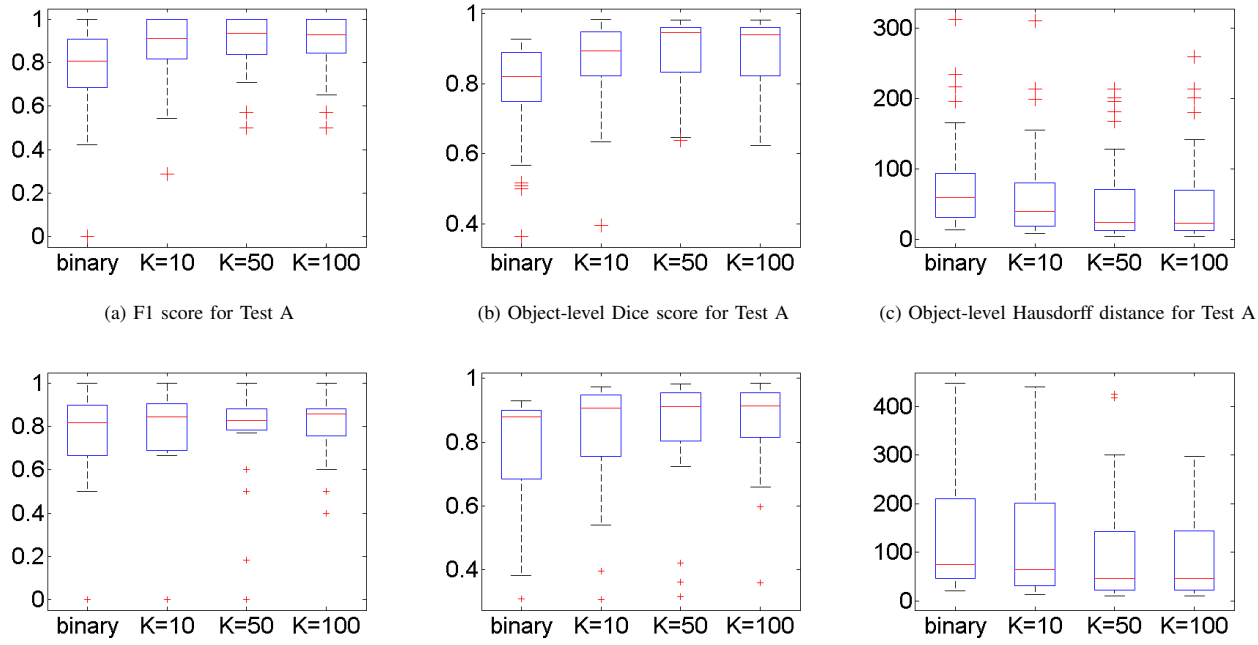
Statistical tests were used to compare the eight methods featured in Table II and Fig. 8 based on their performance on the complete test set (Test A and Test B combined). For each of the three measures (F1, Dice and Hausdorff) a non-parametric Friedman test of differences was conducted and in all three cases the Chi-square value was significant ($p < .01$). Post-hoc Nemenyi tests were then conducted. A critical difference (CD) value of 1.1739 was obtained (at significance level 0.05); the difference in performance of two methods can be considered statistically significant if the rank difference is more than the CD. Figure 9 reports the results as CD diagrams [34]. Structure prediction was significantly better than all of the binary predictors provided that CNN features were used. Structure prediction with combined HC and CNN features (HC+CNN) and $K \in \{50, 100\}$ gave the best results; the Dice scores obtained were significantly better than all other methods, and the F1 and Haussdorff measures were significantly better than structure prediction with HC features alone and significantly better than binary prediction. The gains over binary prediction were $\sim$0.11 (Test A) and $\sim$0.08 (Test B) in terms of mean object-level F1 score, and $\sim$0.09 (Test A) and $\sim$0.08 (Test B) in terms of mean object-level Dice score.

Effect of post-processing. The post-processing steps are



(a) F1 score (b) Object-level Dice score (c) Object-level Hausdorff distance

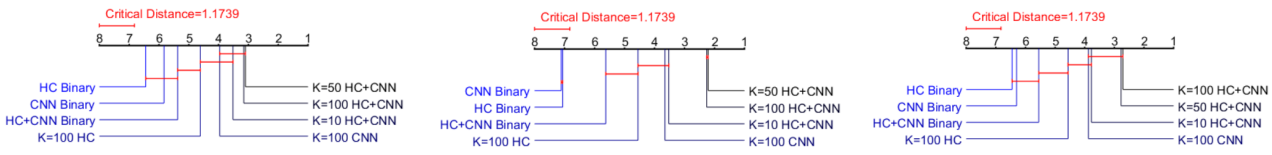
Fig. 7: Results (mean values) using structure prediction with different features and different values of K.



(a) F1 score for Test A (b) Object-level Dice score for Test A (c) Object-level Hausdorff distance for Test A

(d) F1 score for Test B (e) Object-level Dice score for Test B (f) Object-level Hausdorff distance for Test B

Fig. 8: Box plots comparing structure prediction with binary prediction (with HC + CNN features).



(a) Object-level F1 score (b) Object-level Dice score (c) Object-level Hausdorff distance

Fig. 9: Nemenyi post-hoc test results. The horizontal scale numbered 1 to 6 shows the average rank of each method. Smaller ranks are better. Red horizontal lines indicate no significant difference between the methods they connect.

explained in Section III-B. Table III explores the effect of post-processing on binary and structure prediction methods. In both cases post-processing improved the overall scores.

Effect of r. Thus far, the value of r in Eq. (7) was set to $r = 1$. Table IV reports performance for different values of r on test set A and suggests that increasing r does not help. Similar results were obtained on test set B. Larger values of r sometimes resulted in overly smoothed gland boundaries.

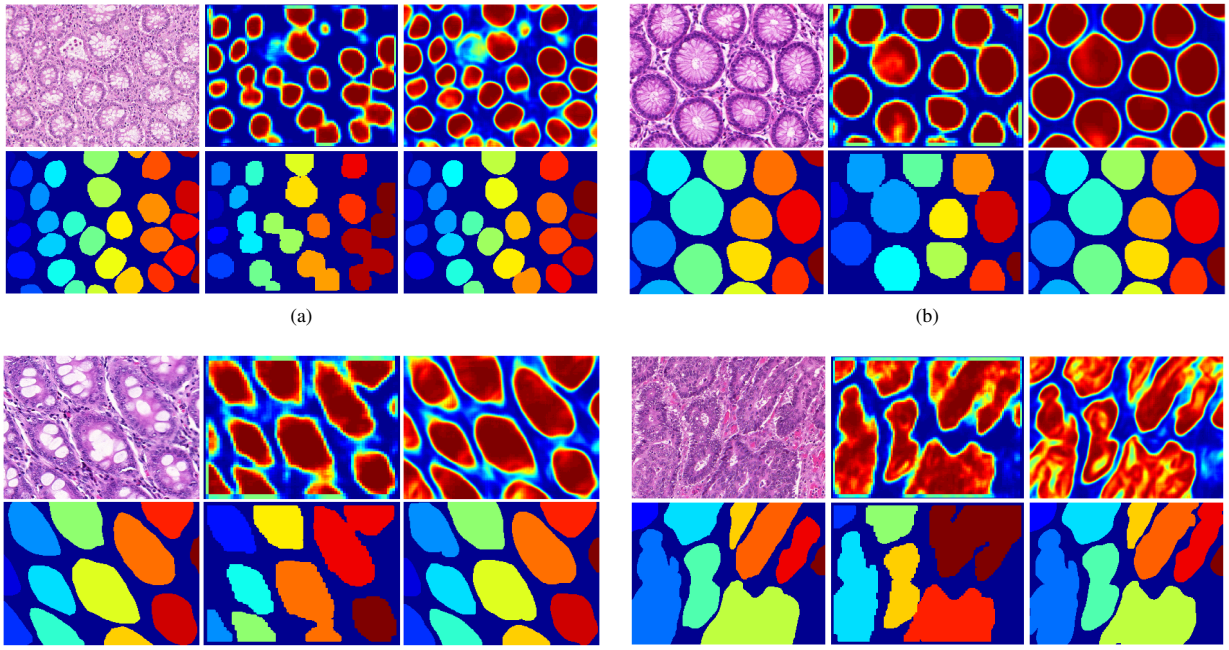


Fig. 10: Four examples for which structure prediction (with HC+CNN features) gave better object-level detection scores than binary prediction (with HC+CNN features). In each sub-figure, the first column shows the original image (top) and its ground truth (bottom), the second column shows the probability map and segmentation obtained using binary prediction, and the last column shows the probability map and segmentation obtained using the proposed method.

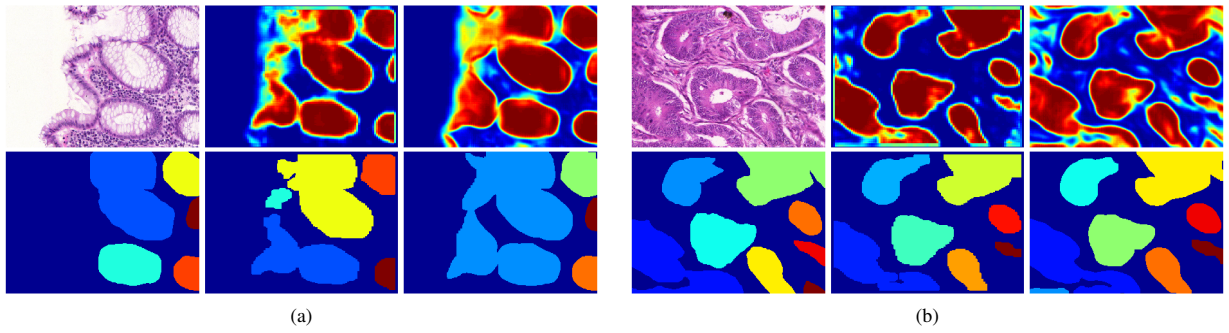


Fig. 11: Two examples for which binary prediction (with HC+CNN features) gave better object-level detection scores than local structure prediction (with HC+CNN features). Sub-figure layout is similar to Fig. 10. (a) A mismatch with ground-truth was caused by torn glands at the edge of the tissue sample which the annotator had chosen not to annotate. (b) A mismatch with ground-truth occurs at the lower-left of the image, probably due to boundary effects.

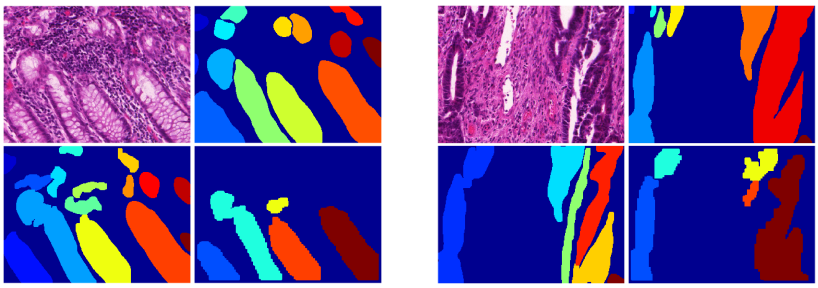
r	K=50			K=100		
	F1 score	Obj Dice	Obj Hausd.	F1 score	Obj Dice	Obj Hausd.
1	0.90	0.89	50.6	0.89	0.89	51.2
3	0.89	0.88	50.6	0.88	0.88	53.3
5	0.88	0.87	51.1	0.87	0.87	55.2
10	0.86	0.85	57.0	0.86	0.86	53.7

TABLE IV: Effect of varying r (HC+CNN features).

E. Computational cost

The network (Fig. 5) took 12 hours to converge on an NVidia Tesla K40 GPU with 12GB memory[2]. Training a structured output classifier with 100,000 randomly sampled features (HC+CNN) took around 2 hours on a core i7 machine with 32 GB of RAM using Matlab 2015a. The average time to extract the HC features for a test image of size 755×522 was about $80s$. The total time required to obtain the segmentation from a test image by our unoptimized Matlab code was

[2]NVidia Corporation donated the Tesla K40 GPU used for this research.



(a) (b)

Fig. 12: Example predictions that correspond to outliers in Fig. 8 (using HC+CNN features). Upper-left: original image. Upper-right: ground truth annotation. Lower-left: structure prediction. Lower-right: binary prediction.

Method	F1 score				Obj Dice				Obj Hausdorff				Rank sum
	Test A		Test B		Test A		Test B		Test A		Test B		
	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank	
Ours	0.892	3	0.801	1	0.887	3	0.853	1	51.175	2	86.987	1	11
Xu et al. [23]	0.858	8	0.771	2	0.888	2	0.815	2	54.202	3	129.930	2	19
CUMedVision2	0.912	1	0.716	5	0.897	1	0.781	6	45.418	1	160.347	8	22
ExB1	0.891	4	0.703	6	0.882	6	0.786	4	57.413	8	145.575	3	31
ExB3	0.896	2	0.719	4	0.886	4	0.765	7	57.350	7	159.873	7	31
Freiburg2	0.870	5	0.695	7	0.876	7	0.786	4	57.093	5	148.463	5	33
CUMedVision1	0.868	6	0.769	3	0.867	10	0.800	3	74.596	10	153.646	6	38
ExB2	0.892	3	0.686	8	0.884	5	0.754	8	54.785	4	187.442	10	38
Freiburg1	0.834	9	0.605	10	0.875	8	0.783	5	57.194	6	146.607	4	42
CVIP Dundee	0.863	7	0.633	9	0.870	9	0.715	9	58.339	9	209.048	12	55
CVML	0.652	11	0.541	11	0.644	13	0.654	10	155.433	13	176.244	9	67
LIB	0.777	10	0.306	13	0.781	11	0.617	11	112.706	12	190.447	11	68
vision4GlaS	0.635	12	0.527	12	0.737	12	0.610	12	107.491	11	210.105	13	72

TABLE V: Comparison with the state of the art.

$300s$. This time includes extracting all features (HC+CNN) from 4 rotated versions of the image and obtaining the final segmentation. However note that this could be improved by processing the 4 rotated versions of an image in parallel.

F. Comparison with the State of the Art

We compare our method (structure classifier with HC+CNN features and $K = 100$) with published methods for gland segmentation. We use the performance measures and ranking criteria from the GlaS challenge for consistency [2]. Specifically, the ranking method is as follows. Methods are first ranked based on each of the three criteria (IV-B) on each of the two test sets, giving 6 rank scores for each method. The sum of these 6 scores, termed the *rank sum*, is used as an indicator of overall performance; lower rank sums are better.

Table V compares our method with other methods using results reported in [2] and [23]. It also reports the ranks and rank sums we computed from these measures[3]. Our method's rank sum of 11 compares favorably with the rank sum of 22 obtained by the winning method in the GlaS contest (CUMedVision2 [21]) and with the rank sum of 19 obtained by the recently proposed method of Xu et al. [23].

[3]Our submission to the GlaS contest (denoted *CVIP Dundee* [2]) used a seven-layer CNN based on AlexNet [35] trained on fixed patch size of 96×96.

V. Discussion and Conclusion

We have proposed a method for learning to segment object instances that takes into account the local spatial structure of labels by training classifiers using a set of structure exemplars obtained via clustering. The encouraging results and the fact that this approach is relatively straightforward to implement by modifying existing classification pipelines lead us to believe that our approach will be of interest to researchers working not only on gland segmentation but on other similar problems in biomedical image analysis.

In other experiments not reported here we tried alternatives to k-means for learning the exemplars. Specifically, we tried to jointly learn the exemplars in a discriminative way together with the structure output classifiers rather than learning the exemplars by unsupervised clustering. We did this in a joint

optimization framework that minimizes the overall reconstruction error between the binary ground truth maps and the predictions reconstructed using the discriminatively learned exemplars learned. However our initial results did not show any advantage for this formulation.

The FCN approach typically employs heavy downsampling, reducing the spatial resolution of intermediate feature maps. The proposed pipeline utilises the multi-scale representations of FCN and retains local spatial information. This scheme may partially explain its relatively good performance.

The implementation used in this paper is too slow for deployment to a clinical application processing whole-slide images. However, processing time could certainly be reduced. HC feature extraction and structure prediction were implemented in Matlab without optimising the code for speed and without use of GPUs. The method lends itself straightforwardly to parallelisation across multiple cores or GPU cards: the rotated versions of each image can be processed in parallel and images themselves (extracted from whole slide images) can be processed in parallel. Combined with rapid improvements being made in hardware infrastructure, scaling this kind of method to whole slide imaging should become feasible without prohibitive cost in the not so distant future.

The GlaS challenge [2] provided an important dataset and protocol for comparison of algorithms for glandular structure segmentation, adding impetus to this aspect of digital pathology image analysis. We obtained state of the art results using the proposed method on the GlaS dataset. While useful as an indicator of performance relative to other methods, rank-based 'league table' comparisons are not always robust and the statistical (or clinical) significance of differences in rank is not always obvious. Methods such as bootstrapping could be usefully employed in future for systematic comparison of methods.

Results had higher variance on test set B than on test set A. This is as expected given that test set B consists of fewer images which are mostly from malignant tissue with moderately or poorly differentiated adenocarcinoma. The proposed method ranked first by all performance measures on test set B indicating that it copes well with malignancy.

We followed the GlaS contest protocol and data set splits in order to facilitate direct comparison with other published methods. As acknowledged in [2], different visual fields from the same slide can appear in different parts of the dataset (i.e. training and test parts) because the data were not stratified by patient. They were however stratified according to the histologic grade and the visual field before splitting. Design of any future gland segmentation dataset should ensure that no data from the same patient can be present in both training and test splits. Despite this limitation, GlaS results may in fact be pessimistic because images have been subdivided into small sub-images, introducing artificial image borders that cut many gland structures into incomplete parts. This makes learning to segment the structures artificially difficult. When applied to whole-slide imaging, the difficulties arising from viewing arbitrary 2D slices through 3D structures will remain but need not be exacerbated by cropping the 2D slice.

VI. Acknowledgement

We are grateful to Korsuk Sirinukunwattana for answering our queries about the GlaS dataset and results, and to the other GlaS organisers (J. Pluim, D. Snead, N. Rajpoot) for making the dataset available.

References

[1] M. Fleming, S. Ravula, S. F. Tatishchev, and H. L. Wang, "Colorectal carcinoma: Pathologic aspects," *Gastrointestinal Oncology*, vol. 3, pp. 153—-173, 2012.

[2] K. Sirinukunwattana, J. P. W. Pluim, H. Chen, X. Qi, P. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, and N. M. Rajpoot, "Gland segmentation in colon histology images: The GlaS challenge contest," *Medical Image Analysis*, vol. 35, pp. 489 – 502, 2016.

[3] P. Kainz, M. Pfeiffer, and M. Urschler, "Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation," *arXiv preprint arXiv:1511.06919*, 2015.

[4] W. Li, S. Manivannan, S. Akbar, J. Zhang, E. Trucco, and S. J. McKenna, "Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks," in *ISBI*, 2016.

[5] S. Manivannan, W. Li, S. Akbar, J. Zhang, E. Trucco, and S. J. McKenna, "Local structure prediction for gland segmentation," in *ISBI*, 2016.

[6] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, and R. A. Zoroofi, "An image analysis approach for automatic malignancy determination of prostate pathological images," *Cytometry*, vol. 72B, p. 227–240, 2007.

[7] H.-S. Wu, R. Xu, N. Harpaz, D. Burstein, and J. Gil, "Segmentation of intestinal gland images with iterative region growing," *Journal of Microscopy*, vol. 220, no. 3, pp. 190–204, 2005.

[8] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "Gland segmentation and computerized Gleason grading of prostate histology by integrating low-, high-level and domain specific information," in *MIAAB Workshop*, 2007.

[9] A. J. K. Nguyen and R. Allen, "Automated gland segmentation and classification for Gleason grading of prostate tissue images," in *International Conference on Pattern Recognition*, Istanbul, August 2010.

[10] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Medical Image Analysis*, vol. 14, no. 1, pp. 1 – 12, 2010.

[11] A. Cohen, E. Rivlin, I. Shimshoni, and E. Sabo, "Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation," *Comput Med Imaging Graph.*, vol. 43, pp. 150–64, July 2015.

[12] B. Ben Cheikh, P. Bertheau, and D. Racoceanu, "A structure-based approach for colon gland segmentation in digital pathology," in *Medical Imaging: Digital Pathology*, ser. SPIE Proceedings, vol. 9791, 2016.

[13] K. Nguyen, A. Sarkar, and A. K. Jain, "Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei," *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2254–2270, December 2014.

[14] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot, "A novel texture descriptor for detection of glandular structures in colon histology images," pp. 94 200S–94 200S–9, 2015.

[15] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images." *IEEE Transactions on Medical Imaging*, vol. 34, no. 11, pp. 2366–2378, 2015.

[16] M. N. Kashif, S. E. A. Raza, K. Sirinukunwattana, M. Arif, and N. Rajpoot, "Handcrafted features with convolutional neural networks for detection of tumor cells in histology images," in *IEEE International Symposium on Biomedical Imaging*, 2016, pp. 1029–1032.

[17] H. Fu, G. Qiu, J. Shu, and M. Ilyas, "A novel polar space random field model for the detection of glandular structures," *IEEE TMI*, vol. 33, no. 3, pp. 764–776, 2014.

[18] O. Veksler, "Star shape prior for graph-cut image segmentation," in *ECCV '08 Proceedings of the 10th European Conference on Computer Vision*, vol. III. Springer-Verlag, 2008, pp. 454–467.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[20] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *MICCAI*, 2016.

[21] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "DCAN: Deep contour-aware networks for accurate gland segmentation," in *CVPR*, 2016.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, ser. LNCS. Springer, 2015, vol. 9351, pp. 234–241.
[23] Y. Xu, Y. Li, M. Liu, Y. Wang, M. Lai, E. I. Chang *et al.*, "Gland instance segmentation by deep multichannel side supervision," *arXiv preprint arXiv:1607.03222*, 2016.
[24] S. Manivannan, H. Shen, W. Li, R. Annunziata, H. Hamad, R. Wang, and J. Zhang, "Brain tumor region segmentation using local co-occurrence features and conditional random fields," in *MICCAI – Brain Tumour Digital Pathology Segmentation Challenge*, 2014.
[25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.
[26] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, no. 3, 2007.
[27] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE CVPR*, 2010.
[28] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.
[29] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.
[31] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," *CoRR*, vol. abs/1412.0774, 2014.
[32] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE CVPR*, 2012.
[33] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, and S. J. McKenna, "An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells and specimens," *Pattern Recognition*, vol. 51, pp. 12–26, 2016.
[34] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, p. 1–30, 2006.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.