Weakly-supervised Structured Output Learning with Flexible and Latent Graphs using High-order Loss Functions

Gustavo Carneiro[1] Tingying Peng[2] Christine Bayer[3] Nassir Navab[2,4]
[1]Australian Centre for Visual Technologies, University of Adelaide
[2]Computer Aided Medical Procedures, Technische Universität München
[3]Department of Radiation Oncology, Technische Universität München
[4]Johns Hopkins University

Abstract

We introduce two new structured output models that use a latent graph, which is flexible in terms of the number of nodes and structure, where the training process minimises a high-order loss function using a weakly annotated training set. These models are developed in the context of microscopy imaging of malignant tumours, where the estimation of the number and proportion of classes of microcirculatory supply units (MCSU) is important in the assessment of the efficacy of common cancer treatments (an MCSU is a region of the tumour tissue supplied by a microvessel). The proposed methodologies take as input multimodal microscopy images of a tumour, and estimate the number and proportion of MCSU classes. This estimation is facilitated by the use of an underlying latent graph (not present in the manual annotations), where each MCSU is represented by a node in this graph, labelled with the MCSU class and image location. The training process uses the manual weak annotations available, consisting of the number of MCSU classes per training image, where the training objective is the minimisation of a high-order loss function based on the norm of the error between the manual and estimated annotations. One of the models proposed is based on a new flexible latent structure support vector machine (FLSSVM) and the other is based on a deep convolutional neural network (DCNN) model. Using a dataset of 89 weakly annotated pairs of multimodal images from eight tumours, we show that the quantitative results from DCNN are superior, but the qualitative results from FLSSVM are better and both display high correlation values regarding the number and proportion of MCSU classes compared to the manual annotations.

1. Introduction

Structured output models have become one of the most studied topics in computer vision given their wide applicability in semantic segmentation [18, 32], instance segmentation [25], human pose estimation [28, 3], depth and normal estimation [7], multiple organ detection and segmentation from medical images [19, 30, 20, 36], among other problems. The use of latent variables in structured output learning models [34] is also important in several problems, where an underlying graph helps the design of a more effective approach. Examples of such models are present in 3D human pose estimation [11, 33] and weakly supervised semantic segmentation [16, 23, 33, 9]. High-order loss functions have also become relevant in structured output problems, with, for instance, the use of overlap loss in segmentation problems [27, 24]. Flexible underlying graphs have also been applied in structured output learning problems [26], where the number of nodes and graph structure can vary with the input data. Finally, there are a few problems formulated as weakly supervised latent structured output learning [9, 16, 23], but these approaches rely on low-order loss functions. In summary, current structured output learning

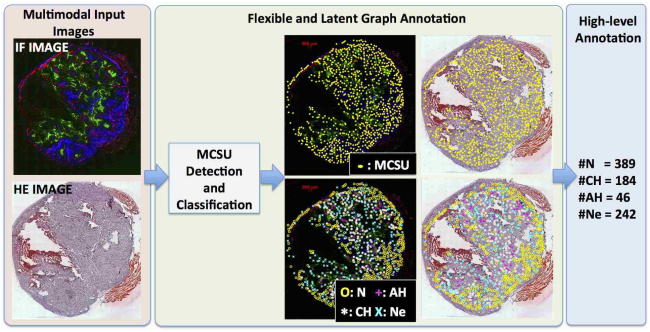


Figure 1. From a pair of multimodal microscopy images (pink left box) acquired from a tumour tissue, the methodology must produce a high-level annotation (blue right box) consisting of the number of MCSU classes found, where the classes are normoxia (N), chronic hypoxia (CH), acute hypoxia (AH), and necrosis (Ne). This annotation is facilitated by a latent graph (green centre box) with nodes representing the MCSUs, which is flexible because the number of nodes and the structure of the graph are not fixed. This figure is better visualised with a pdf reader - please zoom in the IF/HE images to notice the MCSU annotations.

methods that combine latent variables consisting of flexible underlying graphs, weakly supervised training and high-order loss functions, like the approaches being proposed in this paper, have not been proposed for computer vision applications.

In this paper, we address the problem depicted in Fig. 1, which shows the microscopy imaging of a tumour tissue using (immuno-)fluorescence (IF) and hematoxylin and eosin (HE) stainings of the same specimen. It has been observed that tumours containing relatively large number of chronic hypoxic (limitations in oxygen diffusion) and acute hypoxic (local disturbances in perfusion) regions can present resistance to common cancer treatments [2]. This observation led to the development of a manual annotation of such image pairs that produces the number of normoxic (N - normal oxygen diffusion), chronic hypoxic (CH) and acute hypoxic (AH) microcirculatory supply units (MCSU - regions in the tumour tissue supplied by microvessels) [15]. This annotation can then be used in the assessment of the effectiveness of common cancer treatments [2]. Notice in Fig. 1 the presence of the class necrotic (Ne), which is not part of the original manual annotation above, but is nevertheless important to be represented given that an MCSU can be falsely detected in necrotic regions, as explained below in Sec. 2. The main issue with the proposed manual annotation [15] is that it contains only the final number of MCSU classes (N, CH, AH), without indication of MCSU locations, sizes and labels, where an MCSU is loosely defined to have a size of around $200 \times 200\mu m$ with class appearances defined in Fig. 2 [15]. This size is defined by assuming that one pixel in a positron emission tomography (PET) image represents a $200 \times 200\mu m$ region in a microscopy image, which allows a direct comparison between the annotations in these two modalities [15]. In fact, the detection of MCSUs is complicated by the fact that it is based on the detection of a microvessel, which is a non-trivial task given that the tumour tissue section can cut a microvessel in several directions (parallel, oblique, or transversal - see Fig. 3) in addition to the issue that microvessels can vary in diameter. As a result, the task of defining the boundaries of a microvessel in order to form an MCSU is ill-defined, and we propose the use of a flexible and latent graph to facilitate this task. Finally, it is important to note that in spite of its relevance, this manual annotation requires expertise that is generally not available in clinical settings, which makes it a good candidate for automation, particularly considering its potential benefits.

We formulate this problem as a weakly supervised structured output learning using a latent graph that can vary in terms of the number of nodes and structure, where the objective function being optimised consists of a high-order loss function based on the norm of the difference between the number of MCSU classes present in the manual and automated annotations. The need for the flexible and latent graph is based on the idea that it facilitates the complex detection of MCSUs (explained above), from a first trivial detection and classification (into N, CH, AH, and Ne) of microvessel pixels that can be clustered together to form a node in this latent graph, where the clusters are formed based on spatial proximity and classification similarity. We explore two different methodologies to solve this problem. The first approach is based on a latent structured support vector machine model (LSSVM) [34] that uses a flexible underlying graph as the latent variable and minimises a high-order objective function [22] (this first approach is labelled FLSSVM), and the second approach consists of a deep convolutional neural network (DCNN) [13] that minimises a high-order loss function using an implicit flexible and latent underlying graph. This paper claims the following contributions: 1) a new problem to be addressed by computer vision researchers that has the potential to be significant for cancer research, 2) a new LSSVM involving a flexible latent underlying graph and a high-order loss function, 3) a new DCNN model that is able to use a flexible latent underlying graph and minimise a high-order loss function, 4) a weakly annotated dataset of microscopy imaging of cancer tissue [1], and 5) the first methodology that is capable of automatically classifying oxygenation levels of

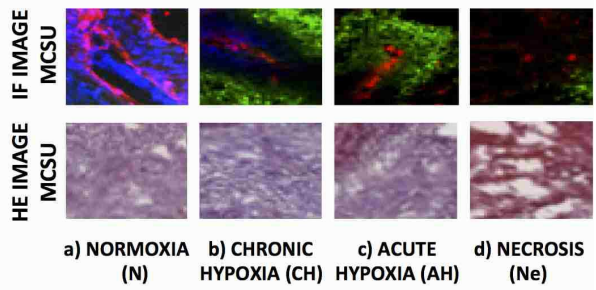


Figure 2. Sketch of the appearance of MCSU classes [15]. Normoxic MCSUs (a) have a red region at the centre of the IF image, representing the microvessel, and a blue region around it denoting normal oxygen supply; chronic hypoxia is denoted again by a red region at the centre (microvessel) with a blue region immediately around it, followed by a green region towards the border, indicating inadequate oxygen supply; acute hypoxia also has a red centre, but immediately followed by green regions; and necrotic regions also have a red centre, but followed by black regions around it. Moreover, normoxic, chronic and acute hypoxic MCSUs have a smooth appearance in the HE image (indicating vital tumour tissue), while necrotic regions have a broken appearance.

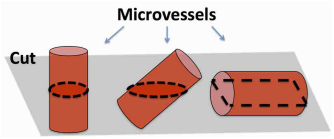


Figure 3. Sketch showing different ways a microvessel can be cut in the preparation for the tumour imaging.

[1] This dataset can be downloaded from the page http://cs.adelaide.edu.au/~carneiro/humboldt/.

MCSUs in multimodal microscopy images of cancer tissue. For the experiments, we use a dataset of 89 pairs of IF and HE images (from eight tumours), where 16 pairs of images from two tumours are used for training a microvessel pixel detector and classifier, and 73 pairs of images from six tumours are used for training and testing the latent structured output learning methodologies. Using a leave-one-tumour-out cross validation experiment, we obtain a high correlation between the manual and automated annotations in terms of the number and proportion of MCSU types for both methodologies, but observe that while the DCNN produces more accurate quantitative results, FLSSVM produces better qualitative results.

2. Methodology

We assume the availability of a dataset represented by $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{v}_n, \mathbf{y}_n)\}_{n=1}^{N}$, where $\mathbf{x} = \{\mathbf{x}^{(\text{IF})}, \mathbf{x}^{(\text{HE})}\}$ is the input IF and HE images, with $\mathbf{x}^{(\text{IF})}, \mathbf{x}^{(\text{HE})} : \Omega \rightarrow \mathbb{R}^3$ ($\Omega \in \mathbb{R}^2$ denotes the image lattice), $\mathbf{v} : \Omega \rightarrow \{0, 1\}$ is a mask that selects regions of the images that contain vital tumour tissue, and $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{N}^3$ denotes the annotation of the number of normoxic (N), chronic hypoxic (CH) and acute hypoxic (AH) MCSUs. Note that the vital tumour mask $\mathbf{v}$ does not delineate precisely the tumour, which means that necrotic (Ne) tissue can still be analysed, so it is important to have the class Ne included in the methodology as a possible class for a detected MCSU, but note that this class is not part of the manual annotation $\mathbf{y}$.

The starting point for our proposed methodologies is the detection and multi-class classification of microvessel pixels from multimodal images (see leftmost image of green box in Fig. 4-(b)), which is detailed in Sec. 2.1. This is followed by the explanation of FLSSVM in Sec. 2.2 and DCNN in Sec. 2.3.

2.1. Microvessel Pixel Detection and Classification

An MCSU is defined as a vital tumour tissue area supplied by a microvessel [15], which has a size of roughly $200 \times 200\mu m$. Microvessel pixels are trivially detected using a threshold on the red channel of the IF image, given that microvessels have a red color in this image modality. In particular, we define a variable $\mathbf{t} : \Omega \rightarrow \{0, 1\}$, where $\mathbf{t}(i) = 1$ if the red channel of the IF image at $i \in \Omega$ is larger than $\tau = 0.1$ (from the range $[0, 1]$), otherwise $\mathbf{t}(i) = 0$ (the yellow dots in the first image of Fig. 4-(b) denote the microvessel pixels). It is also possible to build classifiers to classify a region of size $200 \times 200\mu m$ centred at a microvessel pixel into four classes (N, CH, AH, Ne), using the sketches of Fig. 2. Thus, we annotate a relatively large number of $200 \times 200\mu m$ patches, represented by $\mathbf{x}(i)$, centred at microvessel pixel locations (i.e., image locations $i \in \Omega$, where $\mathbf{t}(i) = 1$) for training the following multi-class classifiers: 1) Adaboost [37], 2) linear SVM [29], 3) random forest [5], and 4) convolutional neural networks [13] (we choose these four classifiers given their superior performances in a recent study [8]). The features used by classifiers 1-3 above are represented by a set of three histograms from the RGB channels of the IF and HE images (one histogram extracted from the centre of the region, another from the border and the other histogram from the region in between the previous two - this accounts for the spatial distribution of the red/blue/green areas in IF, as shown in Fig. 3) and for classifier 4, the features are the RGB values from the vectorized patch (again, from IF and HE images). This process produces four classifiers

$$\{P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)})\}_{k=1}^{K}, \tag{1}$$

with $K = 4$, which are trained and tested with 16 pairs of IF and HE images from 2 tumours (see Sec. 3), and produce the errors in Tab. 1 in a 4-fold cross validation experiment, with each run comprising 8 images for training and the remaining 8 images for testing (we have 1000 annotated patches per image), where error is defined as the proportion of patches $\mathbf{x}(i)$ that are misclassified [6]. We show the results from a majority voting process of the four classifiers in the middle image of Fig. 4-(b).

Table 1. Mean and standard deviation of the errors produced by the microvessel pixel classifiers in the 4-fold cross validation test [6].

Method	Training	Testing
Adaboost	0.132 ± 0.042	0.151 ± 0.053
RandForests	0.080 ± 0.024	0.130 ± 0.041
linear SVM	0.183 ± 0.058	0.210 ± 0.071
CNN	0.047 ± 0.016	0.195 ± 0.055

2.2. Flexible Latent Structure Support Vector Machine (FLSSVM)

The FLSSVM formulation takes as input the microvessel pixel detection and classification from above, where the goal is to build the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ representing the spatial distribution and classification of MCSUs in the image, and use this graph as a hidden variable in a latent structured SVM model that is learned using a high order loss function. More specifically, the estimation of $\mathcal{G}$ starts with the map $\mathbf{t}$ from Sec. 2.1, which represents the locations of microvessel pixels (see the yellow dots in the leftmost image of Fig. 4-(b)). These microvessel pixels are used to form an initial graph, represented by $\mathcal{G}^{ini} = (\mathcal{V}^{ini}, \mathcal{E}^{ini})$, with nodes $v \in \mathcal{V}^{ini}$ labelled with position $i_v \in \mathbb{R}^2$ (where $\mathbf{t}(i_v) = 1$), and classification result $\mathbf{r}_v = [P^{(k)}(c_v|\mathbf{x}, \theta^{(1,k)})]_{c_v \in \{1,...,4\}, k \in \{1,...,K\}} \in \mathbb{R}^{4K}$, (i.e., a vector with the responses from the microvessel pixel classifiers in (1)), and the edges $\mathcal{E}^{ini}$ defined by Delaunay triangulation (leftmost image in FLSSVM box from Fig. 4-(b)). The estimation of $\mathcal{G}$ is based on a minimum spanning tree (MST) clustering [10] that is run on $\mathcal{G}^{ini}$, where the edge weight between nodes v and t (where $v, t \in \mathcal{V}^{ini}$) is defined as $\|i_v - i_t\| \times \|\mathbf{r}_v - \mathbf{r}_t\|$. This clustering algorithm groups nearby microvessel pixels that have similar classification results into the same cluster $\mathcal{C} \subset \mathcal{V}^{ini}$, forming clusters $\{\mathcal{C}_1, ..., \mathcal{C}_{|\mathcal{V}|}\}$, with each cluster denoting

an MCSU. The MST clustering is run using a constraint that guarantees that each cluster $\mathcal{C}$ has a size smaller than $h \times 200\mu m$, with $h \in [0.5, 2]$, where this size is measured by $\max_{v,t \in \mathcal{C}} \|i_v - i_t\|$ (note that h around 1 is related to the definition that an MCSU has a diameter of around $200\mu m$). Therefore the graph $\mathcal{G}$ has nodes $v \in \mathcal{V}$ formed by the clusters $\{\mathcal{C}_v\}_{v=1}^{|\mathcal{V}|}$, where the location of each node v is the centroid of the nodes $t \in \mathcal{C}_v$, and the edge set $\mathcal{E}$ is obtained with Delaunay triangulation (middle of the FLSSVM box of Fig. 4-(b)).

The feature vector $\Psi(\mathbf{x}, \mathbf{y}, h)$ to be used by FLSSVM (right of the FLSSVM box of Fig. 4-(b)) is formed from the labelling of the graph $\mathcal{G}$ that depends on the annotation $\mathbf{y}$ and the nodes $v \in \mathcal{V}$, as follows:

$$\begin{aligned} \underset{\mathbf{M}}{\text{minimise}} \quad & -\|\mathbf{M} \odot \mathbf{P}\|_F^2 + \sum_{c=1}^{3} \left(\mathbf{y}(c) - \|\mathbf{M} \odot \mathbf{E}_c\|_F^2\right)^2 \\ \text{subject to} \quad & \mathbf{1}_4^\top \mathbf{M} = \mathbf{1}_{|\mathcal{V}|}^\top, \ \mathbf{M} \in \{0,1\}^{4 \times |\mathcal{V}|}, \end{aligned} \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{4 \times |\mathcal{V}|}$, with

$$\mathbf{P}(c, v) = \prod_{k=1}^{K} \prod_{t \in \mathcal{C}_v} P^{(k)}(c|\mathbf{x}(i_t), \theta^{(1,k)}) \tag{3}$$

for $c \in \{1,2,3,4\}$ and $v \in \mathcal{V}$, $\mathbf{E}_1 = [\mathbf{1}_{|\mathcal{V}|}, \mathbf{0}_{|\mathcal{V}|}, \mathbf{0}_{|\mathcal{V}|}, \mathbf{0}_{|\mathcal{V}|}]^\top \in \{0,1\}^{4 \times |\mathcal{V}|}$ denotes a matrix with ones in first row and zeros elsewhere (similarly for $c = 2, 3$ with ones in rows 2 and 3), $\mathbf{1}_N$ and $\mathbf{0}_N$ represent a size N column vector of ones or zeros, $\|.\|_F$ denotes the Frobenius norm, $\odot$ represents the Hadamard product, and the summation varies from 1 to 3 because $\mathbf{y}$ has the annotation for three classes only. The optimisation in (2) maximises the label assignment probability and minimises the difference between the number of MCSU classes in $\mathbf{M}$ and in the variable $\mathbf{y}$. We relax the second constraint to $\mathbf{M} \in [0,1]$ to make the original integer programming problem feasible. The resulting matrix $\mathbf{M}$ in (2) allows the labelling of each node $v \in \mathcal{V}$ with $m_v(\mathbf{y}) = \arg\max_{c \in \{1,...,4\}} \mathbf{M}(c, v)$ (Fig. 4-(c)). Notice that the number of microvessel pixels detected is significantly larger than the final number of MCSUs, as shown in Fig. 4(b)-(c). This is because a microvessel is depicted by a large set of red pixels in the IF image, and because of the issues involved in the cutting and imaging of the tumour tissue, as discussed in Sec. 1 and shown in Fig. 3.

The inference in the FLSSVM model is defined by:

$$(\mathbf{y}^*, h^*) = \arg \max_{\mathbf{y} \in \mathcal{Y}, h \in \mathcal{H}} \mathbf{w}^\top \Psi(\mathbf{x}, \mathbf{y}, h), \tag{4}$$

where

$$\Psi(\mathbf{x}, \mathbf{y}, h) = [f_1^{(1,1)}, ..., f_4^{(1,1)}, ..., f_1^{(1,K)}, ..., f_4^{(1,K)}, f^{(2,1)}, ..., f^{(2,L)}]. \tag{5}$$

In (5), the unary features are defined as

$$f_c^{(1,k)} = \sum_{v \in \mathcal{V}} \delta(m_v(\mathbf{y}) - c) \phi^{(1,k)}(c, \mathbf{x}; \theta^{(1,k)}), \tag{6}$$

where $m_v(\mathbf{y}) \in \{1,2,3,4\}$ denotes the label of node $v \in \mathcal{V}$ from (2), $\delta(.)$ is the Dirac delta function and $k \in \{1, ..., K\}$ with $\phi^{(1,k)}(c, \mathbf{x}; \theta^{(1,k)}) = -\log P^{(k)}(c|\mathbf{x}_v, \theta^{(1,k)})$ representing the k^{th} unary potential function in (1) that computes the negative log probability of assigning class c to node v. Also in (5), the binary features are defined as

$$f^{(2,l)} = \sum_{(v,t) \in \mathcal{E}} \phi^{(2,l)}(c_v, c_t, \mathbf{x}; \theta^{(2,l)}), \tag{7}$$

where $l \in \{1, ..., L\}$, $\phi^{(2,1)}(c_v, c_t, \mathbf{x}; \theta^{(2,l)}) = (1 - \delta(c_v - c_t)) g(c_v, c_t, \mathbf{x}; \theta^{(2,l)})$ represents the binary potential function that computes the compatibility (indicated by $g(.)$) between nodes v and t when their labels are different. For instance, we use the following binary potential functions: 1) $g(c_v, c_t, \mathbf{x}; \theta^{(2,1)}) = 1/\|i_v - i_t\|$ (where $i_v \in \Omega$ denotes the position of node v in the image), 2) $g(c_v, c_t, \mathbf{x}; \theta^{(2,2)}) = 1/\|\mathbf{r}_v - \mathbf{r}_t\|$ (where $\mathbf{r}_v = [P^{(k)}(c_v|\mathbf{x}, \theta^{(1,k)})]_{c_v \in \{1,...,4\}, k \in \{1,...,K\}} \in \mathbb{R}^{4K}$ is a vector of the classifier responses for each class in node v); and 3) $g(c_v, c_t, \mathbf{x}; \theta^{(2,3)}) = 1/(\|i_v - i_t\| \times \|\mathbf{r}_v - \mathbf{r}_t\|)$.

The learning process for FLSSVM is formulated by [14]:

$$\begin{aligned} \underset{\mathbf{w}, \{\xi_n\}_{n=1}^N}{\text{minimise}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{n=1}^{N} \xi_n \\ \text{subject to} \quad & \left(\max_{h_n \in \mathcal{H}} \mathbf{w}^\top \Psi(\mathbf{x}_n, \mathbf{y}_n, h_n)\right) - \\ & \left(\mathbf{w}^\top \Psi(\mathbf{x}_n, \widehat{\mathbf{y}}_n, \widehat{h}_n)\right) \geq \Delta(\mathbf{y}_n, \widehat{\mathbf{y}}_n) - \xi_n \\ & \xi_n \geq 0, \forall \widehat{\mathbf{y}}_n \in \mathcal{Y}, \forall \widehat{h}_n \in \mathcal{H}, n = 1, ..., N, \end{aligned} \tag{8}$$

where $\{\xi_n\}_{n=1}^N$ denotes the slack variables and $\Delta(\mathbf{y}_n, \widehat{\mathbf{y}}_n) = \sum_{c=1}^{3} |\mathbf{y}_n(c) - \widehat{\mathbf{y}}_n(c)|$ computes the high-order loss between $\mathbf{y}_n$ and $\widehat{\mathbf{y}}_n$. The learning algorithm to solve (8) is the concave-convex procedure [35], consisting of the following stages: 1) update the latent variable h_n for n^{th} training sample using the latest estimate for $\mathbf{w}$, with $\max_{h_n \in \mathcal{H}} \mathbf{w}^\top \Psi(\mathbf{x}_n, \mathbf{y}_n, h_n)$; and 2) update $\mathbf{w}$ with (8) with $\{h_n\}_{n=1}^N$ from step 1. We use the cutting plane algorithm [12] to estimate $\mathbf{w}$, which iteratively solves a loss augmented inference problem that inserts a new constraint in the set of most violated constraints with $(\widehat{\mathbf{y}}_n, \widehat{h}_n) = \arg\max_{\mathbf{y} \in \mathcal{Y}, h \in \mathcal{H}} \Delta(\mathbf{y}_n, \mathbf{y}) + \mathbf{w}^\top \Psi(\mathbf{x}, \mathbf{y}, h)$. Both the loss augmented inference and the inference in (4) are based on graph cuts (alpha expansion) [4], where the high order loss function $\Delta(\mathbf{y}_n, \widehat{\mathbf{y}}_n)$ is integrated into graph cuts based on the decomposition in [22].

2.3. Deep Convolutional Neural Network (DCNN)

The DCNN model uses as input the maps produced by the four classifiers from (1), which means that the input has 20 channels (four classifiers, each with the output results for five classes), defined by $\mathbf{p}_c^{(k)} : \Omega \to [0,1]$,

$$\mathbf{p}_c^{(k)}(i) = \begin{cases} P^{(k)}(c|\mathbf{x}(i), \theta^{(1,k)}) & \text{if } \mathbf{t}(i) = 1 \\ 0 & \text{if } \mathbf{t}(i) = 0 \end{cases} \tag{9}$$

where $k \in \{1,2,3,4\}$ represents the classifier index, $c \in \{1,2,3,4\}$, and $\mathbf{t}(i) = 1$ indicates a microvessel pixel at

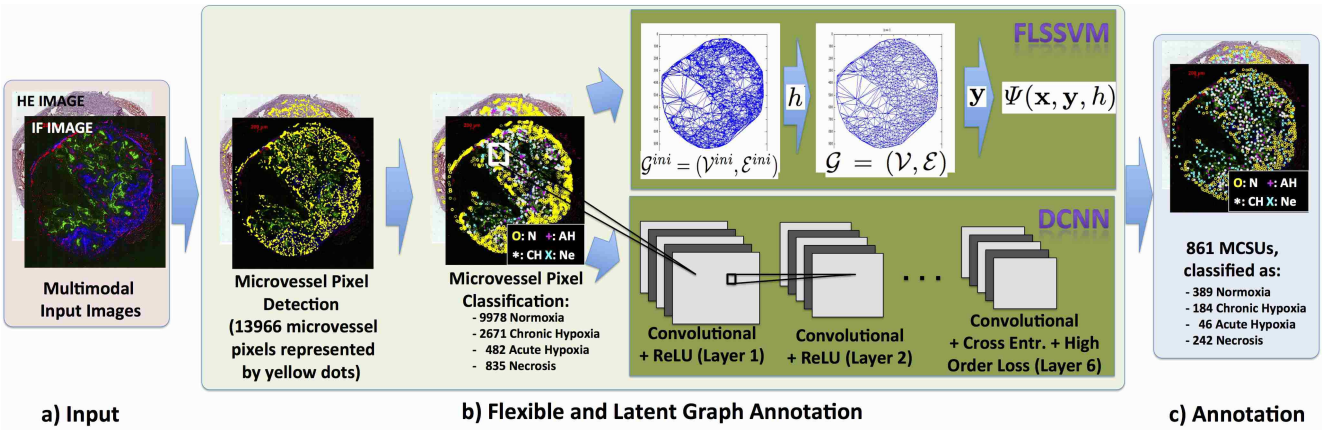


Figure 4. The methodologies proposed in this paper receive as input the IF and HE images (a), then microvessel pixels are detected and classified (first two frames in (b)). Then for the FLSSVM (top) the graph $\mathcal{G}$ is built and labelled using the initial graph $\mathcal{G}^{ini}$ in order to represent the MCSUs and form $\Psi(.)$ for (4) and (8). For DCNN, a series of convolutional layers applied to the microvessel pixel classification images produce a final map containing the MCSUs and their classes. From the outputs of FLSSVM and DCNN, it is trivial to obtain the final annotation in (c). This figure is better visualised with a pdf reader - please zoom in the IF/HE images to notice the MCSU annotations.

location $i \in \Omega$. We also define a new class labelled as Background and indexed by $c = 0$ in (9) with an input defined by $\mathbf{p}_0^{(k)}(i) = 1 - \mathbf{t}(i)$. This fifth class is needed because we minimise a softmax (cross-entropy) loss function with a regularisation term at the last stage of the DCNN, as explained below. The output consists of the number of MCSUs classified as N, CH and AH, and a set of five binary maps $\mathbf{o}_c : \Omega \rightarrow \{0, 1\}$, where $c \in \{1, ..., 4\}$ denotes locations $i \in \Omega$ containing an MCSU classified as N, CH, AH or Ne, and $c = 0$ represents locations without an MCSU (i.e., background). Recall that the location and classification of MCSUs are not available from the training set, so we use the optimisation in (2) to produce a proxy annotation $\mathbf{M}$ that can be used in the DCNN training, where the annotation at location $i \in \Omega$ is defined by

$$m(i) = \begin{cases} \arg\max_{c \in \{1,...,4\}} \mathbf{M}(c, v) & \text{, if } \exists v \in \mathcal{V} \text{ s.t. } i_v = i \\ 0 & \text{, otherwise,} \end{cases} \tag{10}$$

where $v \in \mathcal{V}$, which is the set of nodes of graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, formed as explained in Sec. 2.2. Fig. 5 shows an example of the inputs $\mathbf{p}_c^{(k)}$ (using only one of the classifiers $k \in \{1, ..., 4\}$) and outputs for the DCNN, represented by five binary maps $\mathbf{m}_c : \Omega \rightarrow \{0, 1\}$, where $\mathbf{m}_c(i) = 1$ if $m(i) = c$, and zero otherwise. The error function being minimised in the training of the last layer of the DCNN is the following cross-entropy loss regularised by a high-order loss:

$$\ell = \left(-\sum_{i \in \Omega} \left(\sum_{c=0}^{C} \delta(m(i) - c) \log \frac{\exp(\mathbf{W}_c^\top \mathbf{x}(i))}{\sum_{l=0}^{C} \exp(\mathbf{W}_l^\top \mathbf{x}(i))} \right) \right) + \left(\sum_{c=1}^{3} \left(\sum_{i \in \Omega} \delta(m(i) - c) - \sum_{i \in \Omega} \delta(\widehat{m}(i) - c) \right)^2 \right), \tag{11}$$

where the first term is the usual cross-entropy loss, and the second term is a high-order error that computes the squared difference between the number of MCSUs annotated and classified as N, CH and AH, where the DCNN classification result at image location $i \in \Omega$ is represented by $\widehat{m}(i) = \arg\max_{c \in \{0,...,4\}} \frac{\exp(\mathbf{W}_c^\top \mathbf{x}(i))}{\sum_{l=0}^{C} \exp(\mathbf{W}_l \mathbf{x}(i))}$ (assume here that $\mathbf{x}(i)$ represents the input from the previous layer). The main issue with the loss function (11) is the computation of the derivative of $\delta(\widehat{m}(i) - c)$, so we propose an approximation, consisting of a softmax with a temperature parameter τ, as in $\widetilde{\delta}(\widehat{m}(i) - c) = \frac{\exp\left(\frac{\mathbf{W}_c^\top \mathbf{x}(i)}{\tau}\right)}{\sum_{l=0}^{C} \exp\left(\frac{\mathbf{W}_l^\top \mathbf{x}(i)}{\tau}\right)}$, with $0 < \tau << 1$. This approximation allows for the computation of the following derivative used in the DCNN training:

$$\frac{\partial \ell}{\partial \mathbf{W}_j} = -\sum_{i \in \Omega} \mathbf{x}(i) \left(\delta(m(i) - j) - \frac{\exp(\mathbf{W}_j^\top \mathbf{x}(i))}{\sum_l \exp(\mathbf{W}_l^\top \mathbf{x}(i))} \right) + \sum_{i \in \Omega} 2\mathbf{x}(i) \left(\sum_{c=1}^{3} \left(\delta(m(i) - c) - \widetilde{\delta}(\widehat{m}(i) - c) \right) \times \left(\frac{\exp\left(\frac{\mathbf{W}_c^\top \mathbf{x}(i)}{\tau}\right)}{\sum_l \exp\left(\frac{\mathbf{W}_l^\top \mathbf{x}(i)}{\tau}\right)} - \delta(c - j) \right) \times \frac{\exp\left(\frac{\mathbf{W}_j^\top \mathbf{x}(i)}{\tau}\right)}{\sum_l \exp\left(\frac{\mathbf{W}_l^\top \mathbf{x}(i)}{\tau}\right)} \right). \tag{12}$$

The DCNN model considered in this work consists of 6 convolutional layers with activation functions based on the rectified linear unit (ReLU) [17], except for the last layer, which uses the loss in (11), as shown in Fig. 4. The input image comprises 4×5 channels with the five classes estimated by four classifiers, and is resized to 100×100 and normalized by subtracting the mean. Stages 1-6 use: 1) 10 (5×5) filters, 2) 10 (5×5) filters, 3) 50 (5×5) filters, 4) 100 (5×5) filters, 5) 100 (5×5) filters, and 6) 5 (5×5) filters. This produces an output with five channels (representing classes $\{0, ..., 4\}$) of size 80×80. Training is based

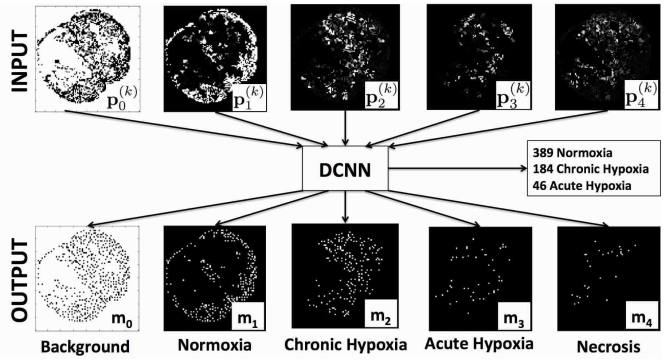


Figure 5. Inputs and outputs for the DCNN.

on backpropagation and inference consists of a feedforward procedure [13].

3. Experimental Setup

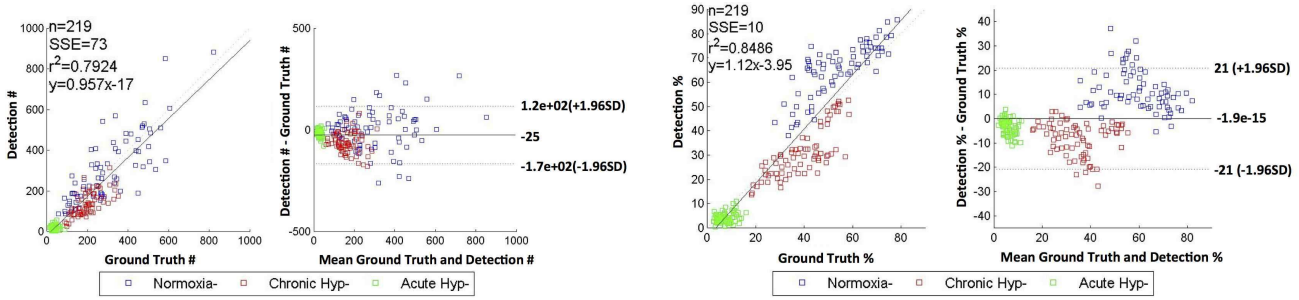
The images used in the experiment is based on the material prepared by Maftei et al. [15], comprising five xenografted human squamous cell carcinoma lines of the head and neck (FaDu), transplanted subcutaneously into the right hind leg of mice. Each tumour cryosection was scanned and photographed with AxioVision 4.7 and the multidimensional and mosaix modules, where the IF images were acquired using three stainings and then the cover slip was removed to stain the same slice with HE to prepare the HE image. For IF image, the green regions were obtained with Pimonidazole to visualise hypoxia, red regions were obtained with CD31 to visualise microvessels, and blue regions were acquired with Hoechst 33342 to display perfusion. This process generates a total of 89 pairs of IF and HE images from eight tumours. The training of the microvessel pixel classifiers $\{P^{(k)}(c|\mathbf{x}(i),\theta^{(1,k)})\}_{k=1}^{4}$ in (1) uses 16 pairs of IF/HE images from two tumours, from which we annotate 1000 microvessel pixels per image, according to the approach described in Sec. 2.1, and the training of the FLSSVM and DCNN models uses the remaining 73 pairs of IF/HE images from six tumours, from which we have manual annotations in terms of the final number of normoxic, chronic hypoxic and acute hypoxic MCSUs. It is worth noting that the location and individual classification of MCSUs are not available in the manual annotation for any of the images above. Finally, the IF and HE images are registered [21] and downsampled to have a size close to 1000×1000 pixels, such that the resolution is approximately 10μm per pixel, and the vital tumour tissue segmentation mask $\mathbf{v}$ is used to mask out the majority of the necrotic regions of the images.

The experiment is based on a six-fold cross validation, where we use the image pairs of five tumours to train and the images from remaining left-out tumour to test (for each of the six tumours). For the FLSSVM, the inference to estimate $\mathbf{y}^*$ and h^* in (4) and the loss augmented inference in (8) to estimate $\widehat{\mathbf{y}}_n$ and $\widehat{h}_n$ are based on graph cuts (alpha-expansion) [4] using a set of possible values for h in $\mathcal{H} = \{0.5, 1, 1.5, 2\}$. Note that during inference, graph cuts produces a labelling for the graph $\mathcal{G}$, but we only take the number of normoxic, chronic hypoxic and acute hypoxic MCSUs to form a vector $\widehat{\mathbf{y}} \in \mathbb{N}^3$, which is subsequently used to build $\Psi(\mathbf{x}, \widehat{\mathbf{y}}, h)$ from the optimisation in (2). For the DCNN training [31], we set temperature parameter $\tau = 0.01$ in (12) and run the training for 100 epochs using mini-batches of size 10, learning rate 0.001, and momentum 0.9.

The quantitative experiment assesses the correlation of the number and proportion of MCSU classes (only for N, CH and AH) between manual and estimated annotations from the proposed FLSSVM and DCNN models in the six test sets (for the six fold cross validation) with the Bland Altman plots [1], which display the number of samples, sum of squared error (SSE), Pearson r-value squared (r^2), linear regression, and p-value. Finally, we also report the inference running time using an un-optimised Matlab code running on a 2.3 GHz Intel Core i7 with 8GB of RAM and Nvidia GeForce 650M.

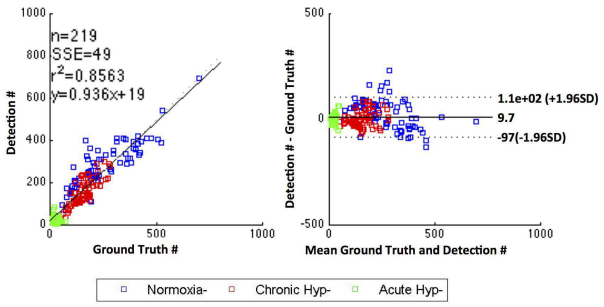
4. Results

The Bland Altman plots for the proposed FLSSVM and DCNN considering the number and proportion of MCSU classes are shown in Figure 6. Note that with respect to the number of MCSU classes, DCNN produces a correlation coefficient $r^2 = 0.85$ and error $SSE = 49$, while FLSSVM has $r^2 = 0.79$ and $SSE = 73$, but both methodologies produce comparable results when considering the proportion of MCSU classes (measured by the percentage of each of the classes N, CH and AH), with $r^2 \approx 0.85$ and $SSE \approx 9$. For the four graphs in Figure 6, the p-values obtained is significantly smaller than 0.01, showing strong correlation results. An additional experiment has been conducted using the loss function in (11) without the high-order loss regularisation, which means that the loss is the usual un-regularised cross-entropy loss. This experiment serves the purpose of testing the validity of the proposed high-order loss for training the DCNN, and the results show that all MCSUs are classified as background (i.e., $c = 0$ in Sec. 2.3) with this un-regularised loss function, which makes sense since this is the most dominant label in the DCNN training. Fig. 7 shows the manual and estimated annotations of 10 different (test) images produced by the proposed methodologies, allowing a qualitative comparison between them not only in terms of the final annotation numbers, but also with respect to the distribution of MCSU classes in the image. Finally, the inference running time of each stage of both methods are as follows (mean average from all test images): microvessel detection (0.03s), microvessel classification (157s), FLSSVM - from microvessels to $\Psi(\mathbf{x}, \mathbf{y}, h)$ (26s), FLSSVM inference in (4) (3.5s), and DCNN inference (0.35s). Thus, the running time for

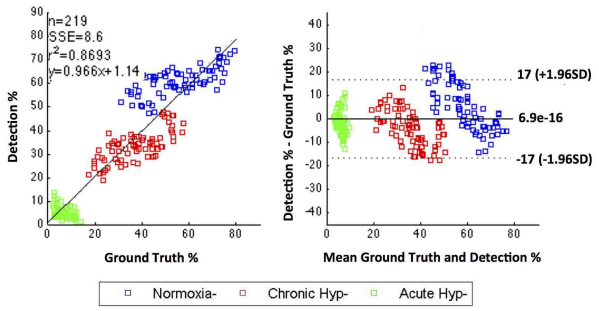


a) FLSSVM (number of MCSU classes)

b) FLSSVM (percentage of MCSU classes)



c) DCNN (number of MCSU classes)



d) DCNN (percentage of MCSU classes)

Figure 6. Bland Altman graphs of the class numbers (left) and proportion (right) results for the FLSSVM (top) and DCNN (bottom).

FLSSVM is 186.53s and for DCNN is 157.38s.

5. Discussion and Conclusion

Both FLSSVM and DCNN show quantitative results with relatively large correlation coefficients and small errors and p-values $<< 0.01$, indicating strong correlation results with the manual annotations. Nevertheless, comparing the results produced by the two proposed methodologies, we can conclude that quantitatively, DCNN produces more accurate results than FLSSVM. However, when looking at the MCSU classification in Fig. 7, we notice that the classification produced by FLSSVM is more coherent with the visual appearance of the MCSU classes shown in Fig. 2. For example, in all IF images of Fig. 7, it is expected that large regions stained in red/blue are annotated with normoxic MCSUs , which is clearly the case for FLSSVM, but not for DCNN. Similarly, regions in IF images showing a transition between blue to green should show chronic hypoxic MCSUs, also clearly seen in the results by FLSSVM, but not by DCNN. Furthermore, green regions in IF images, must display a relatively large number of acute hypoxic MCSUs, which is the case for FLSVMM, but not for DCNN. Finally, necrotic regions appear mostly in the boundaries of the necrotic mask (seen in the image as regions within the tumour tissue without any MCSUs), which is the case for FLSSVM, but not for DCNN. Also, the distribution of MCSUs produced by FLSSVM seems to be more adequate, since the MCSUs are more equally spaced instead of being clustered in some regions of the image. For instance, notice the top region of case 4, where FLSSVM detects a string of MCSUs, while DCNN misses that and instead clusters the MCSUs away from this top region. We believe that the DCNN does not produce adequate qualitative results because of the lack of a spatial prior for the MCSUs, such as the one used in the FLSSVM model. Nevertheless, in terms of the number and proportion of MCSU classes, it is indeed possible to notice the superiority of DCNN, particularly in cases 1-4 of Fig. 7. It is important to reiterate the validity of the cross-entropy loss function regularised by the proposed high-order loss in (11) given the experiment discussed in Sec. 4 that shows that all MCSUs are classified as background when the DCNN model is trained with an unregularised loss function. Finally, the DCNN shows a faster inference, where the main bottleneck is the classification of microvessel pixels, given the large number of microvessels detected from the original IF image.

Acknowledgements: G. Carneiro thanks the Alexander von Humboldt Foundation for the Fellowship for Experienced Researchers and the Australian Research Council's Discovery Projects funding scheme (project DP140102794). T. Peng thanks the Alexander von Humboldt Foundation for the Fellowship for Postdoctoral Researchers.

References

[1] D. G. Altman and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *The statistician*, pages 307–317, 1983. 6

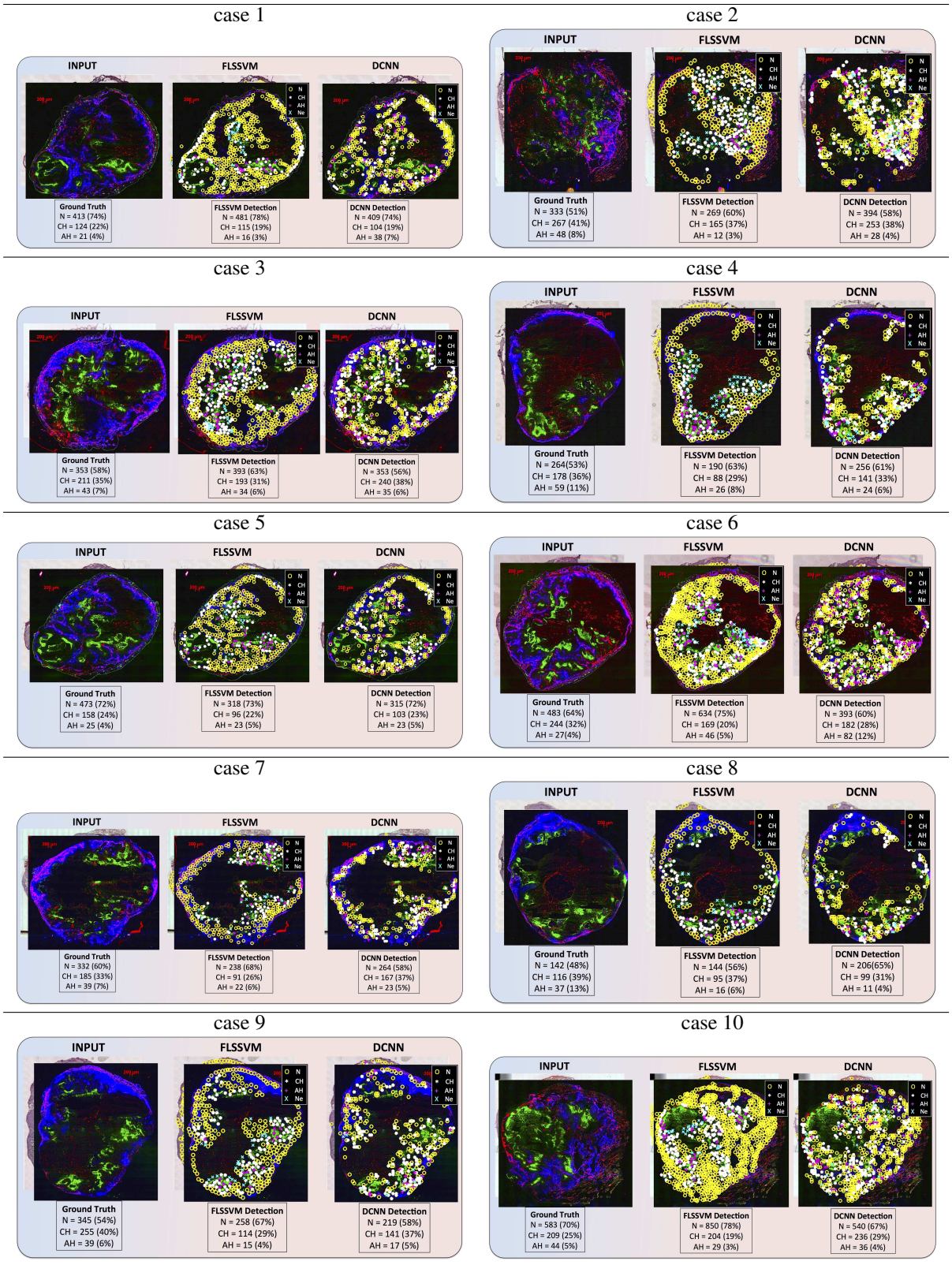


Figure 7. Results of 10 different test images showing a qualitative comparison between FLSSVM (middle image in each case) and DCNN (right image per case) in terms of the distribution of MCSU classes found by each methodology and also with respect to the final number and proportion of MCSU classes compared to the manual annotation (left image per case). This figure is better visualised with a pdf reader - please zoom in the IF/HE images to notice the MCSU annotations.

[2] C. Bayer and P. Vaupel. Acute versus chronic hypoxia in tumors. *Strahlentherapie und Onkologie*, 188(7):616–627, 2012. 2

[3] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In CVPR, pages 1669–1676. IEEE, 2014. 1

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE TPAMI*, 23(11):1222–1239, 2001. 4, 6

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 3

[6] G. Carneiro, T. Peng, C. Bayer, and N. Navab. Automatic detection of necrosis, normoxia and hypoxia in tumors from multimodal cytological images. In *ICIP*, 2015. 3

[7] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *arXiv:1411.4734*, 2014. 1

[8] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *JMLR*, 15(1):3133–3181, 2014. 3

[9] L. Fiaschi, F. Diego, K. Gregor, M. Schiegg, U. Koethe, M. Zlatic, and F. A. Hamprecht. Tracking indistinguishable translucent objects over time using weakly supervised structured learning. In CVPR, pages 2736–2743. IEEE, 2014. 1

[10] O. Grygorash, Y. Zhou, and Z. Jorgensen. Minimum spanning tree based clustering algorithms. In *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*, pages 73–81. IEEE, 2006. 3

[11] C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In ICCV, pages 2220–2227. IEEE, 2011. 1

[12] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009. 4

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2, 3, 6

[14] M. P. Kumar. *Weakly Supervised Learning for Structured Output Prediction*. PhD thesis, Ecole Normale Supérieure de Cachan, 2014. 4

[15] C.-A. Maftei, C. Bayer, K. Shi, S. T. Astner, and P. Vaupel. Changes in the fraction of total hypoxia and hypoxia subtypes in human squamous cell carcinomas upon fractionated irradiation: evaluation using pattern recognition in microcirculatory supply units. *Radiotherapy and Oncology*, 101(1):209–216, 2011. 2, 3, 6

[16] D. Mahapatra, A. Vezhnevets, P. J. Schuffler, J. A. Tielbeek, F. M. Vos, and J. M. Buhmann. Weakly supervised semantic segmentation of crohn's disease tissues from abdominal mri. In *ISBI*, pages 844–847. IEEE, 2013. 1

[17] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010. 5

[18] S. Nowozin, P. V. Gehler, and C. H. Lampert. On parameter learning in crf-based approaches to object class image segmentation. In *ECCV*, pages 98–111. Springer, 2010. 1

[19] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson. A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011. 1

[20] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. M. Möller, S. Nekolla, and N. Navab. Fast multiple organ detection and localization in whole-body mr dixon sequences. In *MICCAI*, pages 239–247. Springer, 2011. 1

[21] T. Peng, M. Yigitsoy, A. Eslami, C. Bayer, and N. Navab. Deformable registration of multi-modal microscopic images using a pyramidal interactive registration-learning methodology. In *Biomedical Image Registration*, pages 144–153. Springer, 2014. 6

[22] P. Pletscher and P. Kohli. Learning low-order models for enforcing high-order statistics. In *International Conference on Artificial Intelligence and Statistics*, pages 886–894, 2012. 2, 4

[23] G. Quellec, M. Laniard, G. Cazuguel, M. D. Abràmoff, B. Cochener, and C. Roux. Weakly supervised classification of medical images. In *ISBI*, pages 110–113. IEEE, 2012. 1

[24] M. Ranjbar, A. Vahdat, and G. Mori. Complex loss optimization via dual decomposition. In *CVPR*, pages 2304–2311. IEEE, 2012. 1

[25] N. Silberman, D. Sontag, and R. Fergus. Instance segmentation of indoor scenes using a coverage loss. In *ECCV*, pages 616–631. Springer, 2014. 1

[26] M. Szummer, P. Kohli, and D. Hoiem. Learning crfs using graph cuts. In *ECCV*, pages 582–595. Springer, 2008. 1

[27] D. Tarlow and R. S. Zemel. Structured output learning with high order loss functions. In *International Conference on Artificial Intelligence and Statistics*, pages 1212–1220, 2012. 1

[28] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014. 1

[29] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, page 104. ACM, 2004. 3

[30] Z. Tu, K. L. Narr, P. Dollár, I. Dinov, P. M. Thompson, and A. W. Toga. Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE TMI*, 27(4):495–508, 2008. 1

[31] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, 2014. 6

[32] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, pages 845–852. IEEE, 2012. 1

[33] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE, 2011. 1

[34] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176. ACM, 2009. 1, 2

[35] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003. 4

[36] S. K. Zhou. Discriminative anatomy detection: Classification vs regression. *Pattern Recognition Letters*, 43:25–38, 2014. 1

[37] J. Zhu, H. Zou, S. Rosset, and T. Hastie. Multi-class adaboost. *Statistics and Its*, 2009. 3