Tongue tumor detection in hyperspectral images using deep learning semantic segmentation

Stojan Trajanovski*, Caifeng Shan*†, Pim J.C. Weijtmans, Susan G. Brouwer de Koning, and Theo J.M. Ruers

***Abstract*—*Objective*: The utilization of hyperspectral imaging (HSI) in real-time tumor segmentation during a surgery have recently received much attention, but it remains a very challenging task. *Methods*: In this work, we propose semantic segmentation methods and compare them with other relevant deep learning algorithms for tongue tumor segmentation. To the best of our knowledge, this is the first work using deep learning semantic segmentation for tumor detection in HSI data using channel selection and accounting for more spatial tissue context and global comparison between the prediction map and the annotation per sample. *Results and Conclusion*: On a clinical data set with tongue squamous cell carcinoma, our best method obtains very strong results of average dice coefficient and area under the ROC-curve of 0.891 ± 0.053 and 0.924 ± 0.036, respectively on the original spatial image size. The results show that a very good performance can be achieved even with a limited amount of data. We demonstrate that important information regarding tumor decision is encoded in various channels, but some channel selection and filtering is beneficial over the full spectra. Moreover, we use both visual (VIS) and near-infrared (NIR) spectrum, rather than commonly used only VIS spectrum; although VIS spectrum is generally of higher significance, we demonstrate NIR spectrum is crucial for tumor capturing in some cases. *Significance*: The HSI technology augmented with accurate deep learning algorithms has a huge potential to be a promising alternative to digital pathology or a doctors' supportive tool in real-time surgeries.**

***Index Terms*—Tumor Segmentation, Hyperspectral imaging, Deep Learning.**

I. Introduction

Worldwide, head and neck cancer accounts for more than 650,000 cases annually [1]. Tongue cancer is a specific case of head and neck cancer, and patients suffering from tumors in tongue tissue are generally treated by removing the cancer surgically. Accurate segmentation of the tumor tissue from the healthy part is challenging; besides tumor tissue, the surgeon removes a margin of non-tumorous tissue around the tumor, which is called a resection margin in surgical oncology [2]. Moreover, literature stresses the importance of adequate resection margins, as it is a powerful predictor of the 5-year survival rate [3], [4], [5].

Smits *et al.* [6] conducted a clinical review regarding resection margins in oral cancer surgery and found that 30% to 65% of the resection margins in surgical results are inadequate. The removal of an oral tumor and determining the resection margins are challenging procedures, since they rely on palpation. This entails that the surgeon classifies tumor tissue based experience and on the sense of touch and sight. The accuracy of resection margins is assessed by pathologists and they analyze the removed tissue after a surgery. This is time-consuming process, and it is less effective since it provides surgeons with information after the surgery instead of during the procedure. From this perspective, there is an opportunity for technological solutions to assist surgeons in the assessment of tumor tissue and to determine adequate resection margins, by providing real-time feedback during oral cancer surgery. Moreover, there are no commonly accepted clinical standards or adopted techniques for this purpose.

Hyperspectral imaging (HSI), originally developed for remote sensing [7], has been successfully used in many fields such as food quality and safety, resource control, archaeology and biomedicine [8]. With advancements in hardware and computational power, it has become an emerging imaging modality for medical applications [9]. HSI has the potential advantages of low cost, relatively simple hardware and ease of use. This makes HSI a candidate for intra-operative support of a surgeon. Compared to regular RGB images it is challenging to process the HSI data due to the size of the data: hundreds of color bands for each pixel in a patient image with a large spatial size results in large files with varying amounts of redundant information.

Previous studies have mostly focused on tumor *classification tasks* [10], [11], [12], [13] in HSI images. Fei *et al.* [11] have evaluated the use of HSI (450-900nm) on specimen from patients with head and neck cancer. They achieved an area under the ROC-curve (AUC) of 0.94 for tumor classification with a linear discriminant analysis on a data set of 16 patients in which 10 were verified to have squamous cell carcinoma (SCCa). However, their testing was done on specimens from the same patient as the classifier was trained on. Lu *et al.* [14] and Halicek *et al.* [12] acquired multiple specimen from 50 head and neck cancer patients with 26 having squamous cell carcinoma (SCCa) in the same visual spectral range 450-900nm. They [12] did a classification task with deep convolutional neural networks on 25x25 patches with leaving-

*S.T. and C.S. have contributed equally. This research was done while S.T. and C.S. were with Philips Research, Eindhoven, The Netherlands.

† Corresponding author.

The authors declare no competing financial interests. A preliminary version of this research has appeared as (non-proceeding) extended abstract at MIDL (International Conference on Medical Imaging with Deep Learning), July 8-10, 2019, London, UK.

Stojan Trajanovski is with Microsoft Inc., Bellevue, WA, USA & London, UK (e-mail: sttrajan@microsoft.com).

Caifeng Shan is now with Shandong University of Science and Technology, Qingdao, China (e-mail: caifeng.shan@gmail.com).

Pim J.C. Weijtmans was with Eindhoven University of Technology (TU/e), Eindhoven, The Netherlands.

Susan G. Brouwer de Koning and Theo J.M. Ruers are with Netherlands Cancer Institute, Amsterdam, The Netherlands and fMIRA Institute, University of Twente, Enschede, The Netherlands (e-mails: {s.brouwerdekoning,t.ruers}@nki.nl).

one-patient-out cross-validation and reported accuracy of 77% for the SCCa group. Animal study on mice [15] with induced tumors was conducted by Ma *et al.* [10], achieving an accuracy of 91.36% with convolutional neural networks in a leave-one-out cross-validation also for a classification task. A similar animal study on prostate cancer and on head and neck cancer in mice were conducted in [16] and [17]. Ravì *et al.* [18] use random forest-based approaches on hyperspectral images for brain tumor segmentation. Other machine learning techniques, minimum spanning trees [19], support vector machines [20], [13], k-nearest neighbors algorithm [21], [22], naïve [22] Bayes, gaussian mixture models [23] and well-performing deep learning architectures [24] (e.g., inception [25]), have also be used for hyperspectral images, mostly for classification tasks. In order to simplify and reduce the large number of channels, standard techniques such as tensor decomposition [26] or principal component analysis (PCA) [27] are applied. In all mentioned studies the focus lies entirely on spectral information in the visible range around 450-900nm.

All of the above research works: (i) have utilize only the visible part of the spectra; (ii) have focused mainly on animal cases; (iii) have mostly utilize PCA or other well-established techniques for channels/dimensionality reduction; or (iv) have mostly focused on classification task as a global malignancy assessment per patient. However, the need of real-time and precise intraoperative feedback requires accurate segmentation between the tumor and non-tumor in human tissues. In this work, we have examined several structural, spectral and semantic segmentation deep learning models (such as U-Net [28] variants) taking patches of images with all or predefined channels selection, but assessing the global performance on a per patient/specimen base. Compared to the previous work, we have used much broader spectra utilizing both the visible (VIS) and near-infrared (NIR) spectral ranges [29] that has been rarely employed [30], [31], especially in the context of deep learning methods. The contributions of the paper are the following:

- With the best semantic segmentation method, we have achieved competitive performance of average dice coefficient and area under the ROC-curve of 0.891 ± 0.053 and 0.924 ± 0.036, respectively, in the leave-patients-out cross-validation with on a clinical data set of 14 patients.
- We have demonstrated that the (often omitted) near-infrared spectra are crucial: first for spotting/classifying; and second for correctly segmenting the tumor tissue in some cases, although VIS channels remain the most significant ones for the performance on average.
- We have proposed a novel channel selection/reduction technique, rather than standard reduction techniques that eliminate pure channel information (e.g., PCA), and we have demonstrated that there is a selected set of number of channels that leads to optimal performance; meaning that with smaller number of channels part of the signal is lost and higher number of channels brings extra noise, both contributing to reduced performance.

The remainder of the paper is organized as follows. Details of the clinical dataset are given in Section II. The proposed methods and the obtained results are provided in Section III and Section IV, respectively. We conclude in Section V.

II. Clinical Dataset

The clinical data was collected at the Netherlands Cancer Institute (NKI) - Antoni van Leeuwenhoek Hospital in Amsterdam, the Netherlands. Hyperspectral images of specimens from 14 patients undergoing surgery for the removal of squamous cell carcinoma of the tongue were acquired. All ethical guidelines required for *ex vivo* human studies were followed.

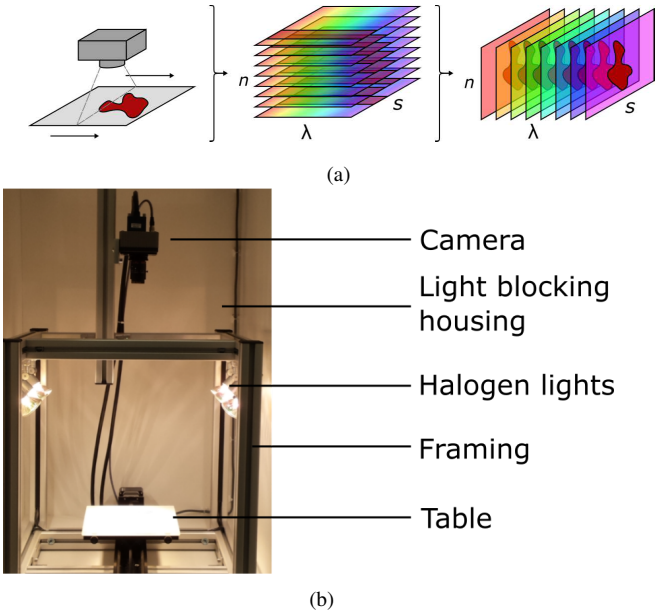
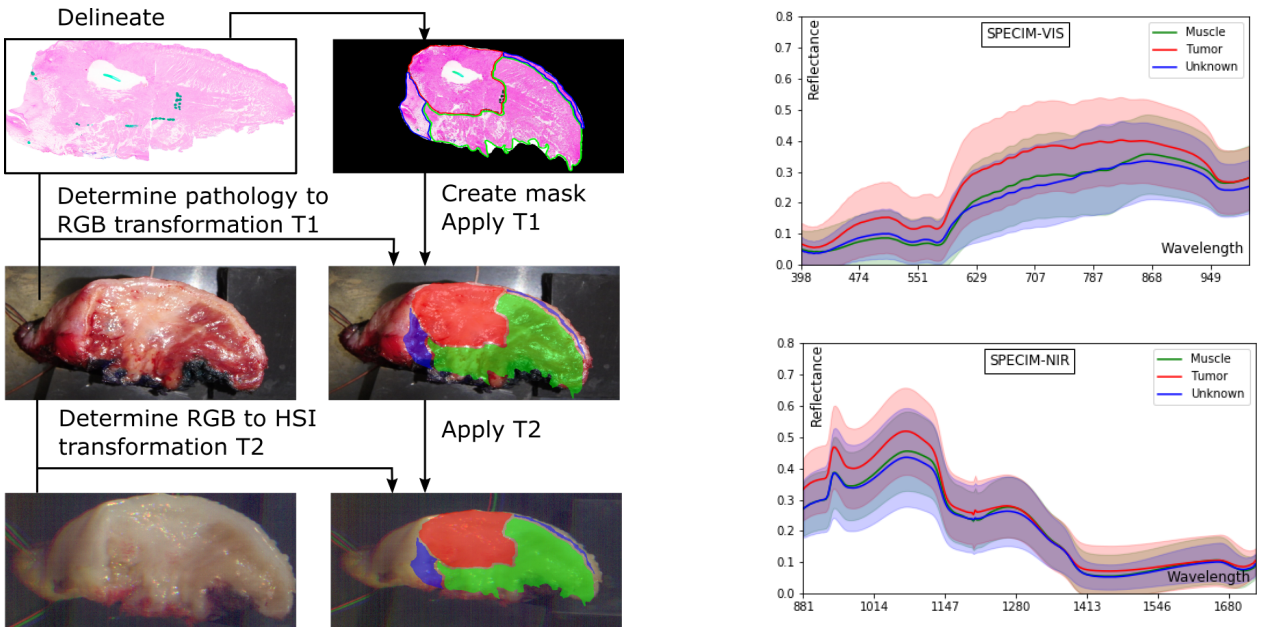


Fig. 1: (a) A table shifts tissue under the camera. Lines are recorded and assembled into a data volume. (b) The imaging system used to collect the HSI data, where a tissue is placed on the table in the bottom.

Directly after resection, the specimen was brought to the pathology department, where the resection margins were inked according to the routine clinical workflow. The pathologist localized the tumor by palpation and subsequently cut the specimen through the middle of the tumor. First, a RGB image was taken from the cut surface with a regular photo camera. Immediately after and without touching the specimen, the cut surface of the specimen was imaged with two hyperspectral cameras (Spectral Imaging Ltd., Oulu, Finland), one operating in the visible wavelength range (VIS) and the other in the near-infrared wavelength range (NIR). The instrumentation with the hyperspectral cameras and the supporting equipment for data acquisition are shown in Figure 1. After HSI imaging, the specimen was subjected to further routine pathological processing, and the pathologist annotated different tissue types on the histopathology slide. Additional details on the data acquisition can be found in [32].

Both HSI cameras are push broom line scan system. The VIS camera captures 384 wavelengths in the 400nm-1000nm range with 1312 samples per line, for 612 lines. The NIR



(a) (b)

Fig. 2: (a) Annotation and registration of the hyperspectral data: tumor (red), healthy tongue muscle (green), and healthy epithelium or other non-tumor tissue (blue). (b) Reflectance of the acquired data for VIS and NIR. (Blue curves, labeled as "unknown", represent epithelium or other non-tumor tissue and it has been used with term "non-tumor", interchangeably.)

camera captures 256 wavelengths in the 900nm-1700nm range with 320 samples per line, for 191 lines.

In order to label the HSI data, a histopathological slide is taken from the surface that has been scanned. The slide is digitized and delineated to mark the tumor (red), healthy muscle (green) and epithelium & non-tumor tissue (blue). This is the first step shown in Figure 2a. From the delineation a mask is created. During histopathological processing the specimen was deformed and to correct this, a non-rigid registration algorithm is used. Obvious matching points in the histopathological and RGB images were visually selected. Using these points, the mask is transformed to match the RGB picture. This is depicted in Figure 2a in middle row as transformation T1. The point-selection [33] is done again on the RGB and HSI data to derive transformation T2, which is used to transform the mask again to match the HSI data. The VIS and NIR datasets contain the same patients, therefore the data cubes could be combined into a broad spectrum ranging from 400nm to 1700nm. During the registration process, the points that were used to determine the transformations were stored, which could then be used to align the VIS and NIR data cubes. Transformation T2 (NIR data) was inversed to transform the HSI NIR to the RGB shape, and subsequently transform T2 (VIS data) was applied to the VIS data. The NIR data had a lower resolution, therefore it was up sampled during the transformation. At this point, the data volumes had the same shape and could be concatenated along the spectral axis. Due to the overlap in spectral range, which was approximately 120nm, half of 120nm from VIS cube and the other half from NIR were removed. Because the first and last bands of the data cube were noisy, bands from both sets were removed instead of removing from only the NIR or VIS cube. The reflectance as a function of the wavelength depicted in Figure 2b demonstrates that a merger/stack of VIS & NIR (top and bottom subfigures if summed up) well captures the full spectra. See [32] for more details on the data registration and preprocessing.

The number of pixels for the two datasets are shown in Tables I and II. There is a small difference between the VIS and NIR sets, despite they are made with the same tissue and annotations. This could be explained in several ways. For instance, the tissue could have moved between measurements, causing a deformation of the tissue, or that the tissue was imaged from a different angle. Another explanation might be that the difference is a result of an error while transforming the histopathological image to match the HSI data.

III. Methods

Having the annotation masks along with the hyper-cube allows for supervised machine learning. With the annotated data at hand, we consider two types of methods: 1) pixel-wise classification: spectral, structural, hybrid, 2) semantic segmentation; and compare their performance. It is also important to stress that we have a prediction for two classes: healthy[1] and tumor tissues.

A. Channel selection, spectral, structural and hybrid approaches

Using small patches spanning all bands the spectral information can be captured. By selecting bigger patches, structural information becomes available. By combining both inputs, the full spectral and some structural information are available for the network.

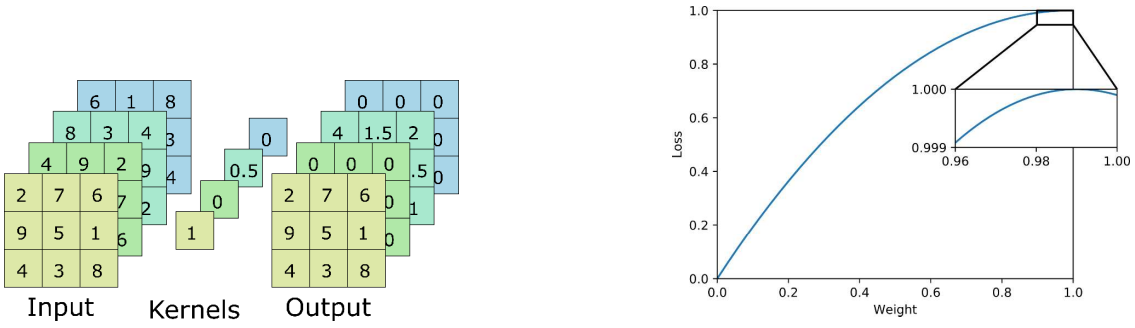
[1]As mentioned earlier healthy tissue accounts for epithelium, muscle and other non-tumor tissue

TABLE I: The number of pixels per patient in the SPECIM-VIS dataset. Patient 30 was the only patient without 'non-tumor' tissue in the annotation. The term 'non-tumor' refers to epithelium or other ordinary or non-tumor tissue.

Patient ID	Total	Tumor		Muscle		Non-tumor		Background	
25	94K	9K	9.6%	7K	7.2%	2K	1.9%	77K	81.3%
26	30K	1K	3.5%	11K	34.9%	1K	3.1%	18K	58.5%
28	181K	25K	14.0%	16K	8.8%	16K	8.6%	124K	68.5%
29	52K	6K	12.2%	10K	19.1%	2K	3.2%	34K	65.5%
30	43K	3K	6.8%	13K	30.8%	-	-	27K	62.4%
31	45K	1K	2.1%	11K	23.6%	135	0.3%	34K	74.0%
33	33K	2K	7.7%	9K	27.8%	1K	3.7%	20K	60.8%
34	42K	1K	2.9%	15K	35.5%	1K	1.3%	26K	60.4%
35	33K	3K	8.0%	9K	26.4%	149	0.5%	22K	65.2%
36	36K	2K	5.1%	9K	25.0%	442	1.2%	25K	68.7%
37	39K	4K	10.8%	11K	28.8%	1K	3.2%	22K	57.2%
39	36K	3K	9.4%	10K	26.7%	2K	4.2%	21K	59.7%
40	34K	1K	3.5%	5K	15.0%	462	1.4%	27K	80.1%
41	40K	2K	5.1%	8K	20.9%	299	0.8%	29K	73.3%
Total	738K	65K	8.7%	143K	19.4%	26K	3.5%	504K	68.3%

TABLE II: The number of pixels per patient for the SPECIM-NIR dataset.

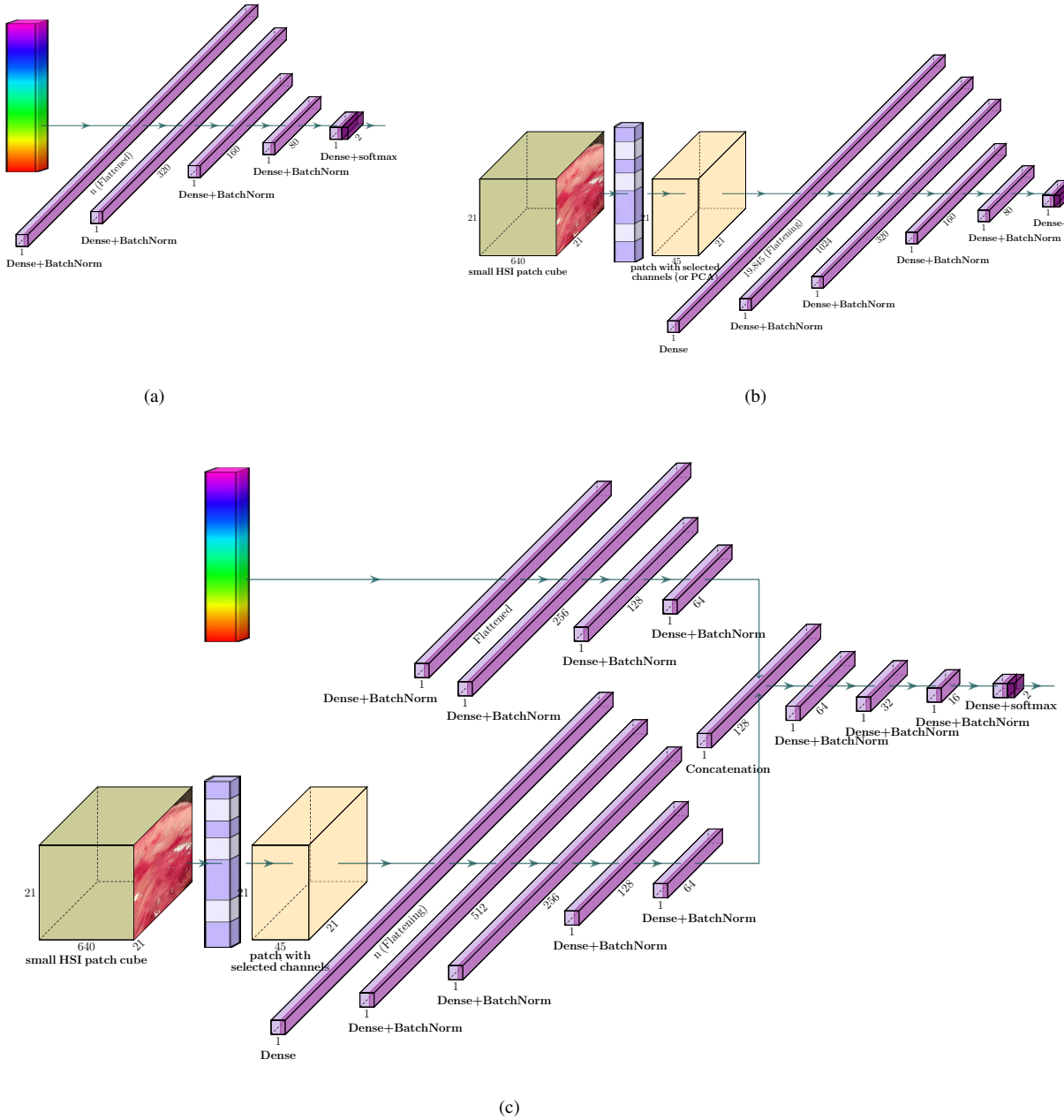
Patient ID	Total	Tumor		Muscle		Non-tumor		Background	
25	12,880	1,128	8.8%	919	7.1%	210	1.6%	10,623	82.5%
26	5,696	170	3.0%	1,185	20.8%	99	1.7%	4,242	74.5%
28	22,240	2,943	13.2%	1,900	8.5%	1,762	7.9%	15,635	70.3%
29	8,268	785	9.5%	1,012	12.2%	171	2.1%	6,300	76.2%
30	6,723	287	4.3%	1,470	21.9%	-	0.0%	4,966	73.9%
31	6,237	80	1.3%	1,157	18.6%	13	0.2%	4,987	80.0%
33	4,615	253	5.5%	997	21.6%	113	2.4%	3,252	70.5%
34	6,059	127	2.1%	1,596	26.3%	59	1.0%	4,277	70.6%
35	5,780	290	5.0%	1,004	17.4%	15	0.3%	4,471	77.4%
36	5,159	220	4.3%	951	18.4%	48	0.9%	3,940	76.4%
37	6,624	460	6.9%	1,273	19.2%	157	2.4%	4,734	71.5%
39	5,330	380	7.1%	1,067	20.0%	161	3.0%	3,722	69.8%
40	5,168	117	2.3%	556	10.8%	51	1.0%	4,444	86.0%
41	6,160	205	3.3%	984	16.0%	26	0.4%	4,945	80.3%
Total	106,939	7,445	7.0%	16,071	15.0%	2,885	2.7%	80,538	75.3%



(a) (b)

Fig. 3: (a) Depth-wise convolutions outcome. (b) Behavior for the channel selection loss function.

1) Channel selection: The channel selection gives the most important channels of a dataset. This step is also important as it ranks/separates more important channels containing signal from the less important ones containing noise, which improves the performance. It is used in both pixel-wise and semantic segmentation. The selection method is based on l_1 regularization concept. By adding an additional layer between the input and the first layer of the model, it is possible to examine what weights are given to the channels. The patches we use are three-dimensional and we want to find a subset of channels, so a two-dimensional depth-wise convolution [35] is used, with a 1x1 kernel and 1x1 stride (see Figure 3a). The weights of this selection layer are constrained to range from 0 to 1, to disable and enable channels respectively. To encourage channels selection, an incentive is added. Without this, all weights would simply be 1 and there would be no effective change. l_1 weight regularization is considered for this incentive. With l_1 the regularization parameter, w_i is the layer weight of matching channel i and n the number of channels. The result J is added to the loss during training, so there is a penalty on having non-zero weights. However, now all weights will go to zero, some faster than others. Ideally, our aim is that the important channels go to one and all others go to zero. The formula is adapted by applying a function to the weights before summing them and this is added to the



(a) (b) (c)

Fig. 4: Benchmark neural networks, visualized using [34]. (a) Spectral neural network. (b) Structural neural network with channels filtering. (c) Dual-stream spectral and structural neural network.

loss $l_1 \sum_{i=1}^{n} \left(1 - (\frac{|w_i|}{p} - 1)^2\right)$. This function is a second degree polynomial and has maximum 1 at p. The value of p is set at .99 creating a small pocket at $(0.99, 1)$ where the loss slightly increases while the weight is reduced from 1 to .99. After that small pocket, the loss will decrease until the weight is zero as demonstrated in Figure 3b. More details of the selected channels are given in Section IV.

2) Spectral neural network architectures: To exploit the full extent of the spectral information, a network is designed for using 1x1 patches including the full spectrum of the HSI data. Five fully connected (FC) layers were used. The first layer has as many units as wavelengths in the input data (640 for stacked, where the contributions of VIS and NIR are 384 and 256, respectively). The last layer is the output layer, and therefore has only two units. The hidden layers in between had 320, 160 and 80 units, and those layers were followed by batch normalization to stabilize the training process. The neural network is visualized in Figure 4a. Alternatively, instead of fully connected layers, convolutional neural networks [36] (i.e. shareable weights and translation-invariant filters) have been also explored, but these resulted in slightly worse or comparable performance. The next thing worth considering is expanding to the spatial context of the data cube. However, taking more pixels in a combination to having all channels is becoming computationally challenging

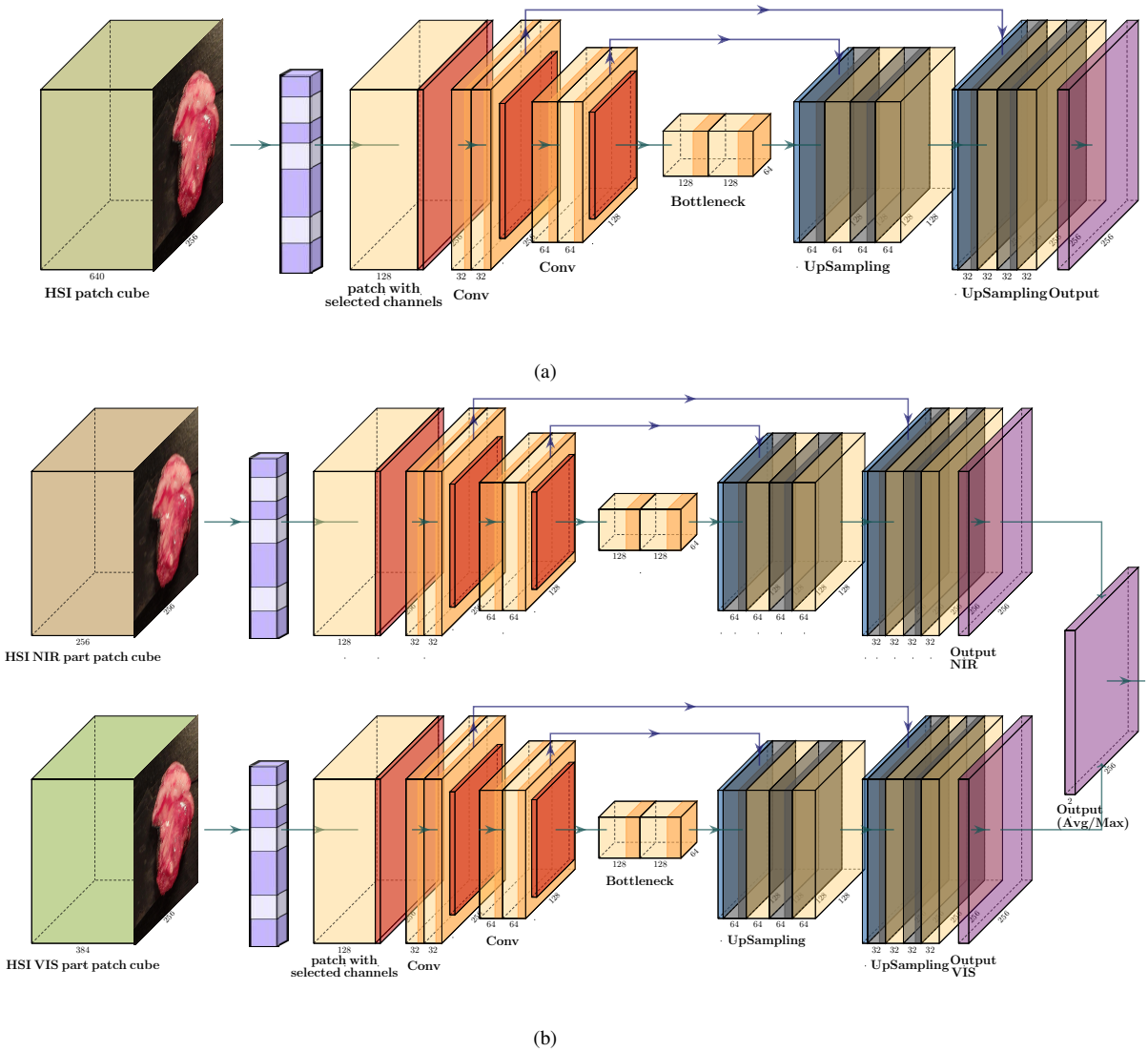


Fig. 5: U-Net based neural network architectures for HSI data. The visualization is drawn using PlotNeuralNet software [34] (https://github.com/HarisIqbal88/PlotNeuralNet). (a) Singe-stream U-Net for stacked (VIS and NIR) patches. (b) Dual-stream U-Net based with separate VIS and NIR streams.

(e.g., significantly slower or memory demanding) and it also brings extra redundancy and possible conflicting context in some of the channels with the annotation. Some channel importance prioritization and filtering are in order as elaborated in the following sections.

3) Structural neural network architectures: By increasing the patch size, morphological features in the HSI data can be used to classify the tissue. With the most important channels known either reduced by the channels selection (as described earlier) or alternatively a principle component analysis (PCA), a model that focuses on structure can be trained. Data with a high resolution will benefit the most from this approach, as it will contain more structural information compared to low resolution data. With 45 VIS channels selected and spatial patch of 21×21, input data cubes are defined. The number of VIS channels used have been selected empirically. We conducted experiments with different number of channels, and found that (i) initially the performance improved when more channels were included; (ii) the performance did not get any better when using more than 45 VIS channels [37, Page 19, Figure 16]. This is followed by flattening which results in 19,845 input units. The network is similar to the spectral model, with fully connected hidden layers and it is shown in Figure 4b. The difference lies in the input shape and the first fully connected layer. Instead of matching the number of units of the flattened channel-filtered input, 1024 units are used. This is done to reduce the number of parameters in the model. As in the spectral neural networks, convolutional neural blocks are also tried, but these result in worse or similar performance over the fully connected layers.

4) Hybrid spectral & structural neural network architectures: To utilize both spectral and structural contexts, we derive a dual-stream spectral and structural model. The model is designed, using a 1x1 patch that covers all wavelengths to incorporate spectral features and a bigger patch with selected channels for the structural information. The network is visu-

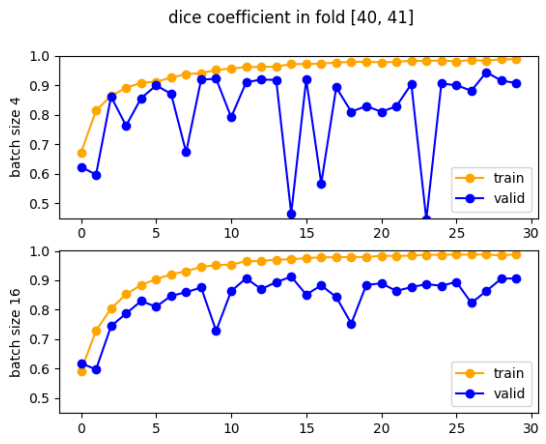


Fig. 6: The influence of the batch size on the training.

alized in Figure 4c. A spectral patch is selected from the data volume, and fed into four fully connected layers. The first of those layers has the same number of units as the amount of wavelengths in the spectral sample. The following hidden layers have 256, 128 and 64 units respectively. Additionally, a structural patch is selected from the data using the channel ranking as discussed in the channel selection section. The patch is flattened and followed by four hidden dense layers with 512, 256, 126 and 64 units. After concatenation of the spectral and structural branches, four additional dense layers follow with 64, 32, 16 and 2 output units. After every layer, batch normalization is applied to stabilize the training process.

B. Semantic segmentation based on U-Net neural network variants

Based on the hyper-cubes and corresponding annotations, we create HSI input and annotation patches with wider spatial context. The reason of using patches, with a fixed size per experiment, is two-fold: (i) we have a limited data from 14 patients and in this way, we create a train and validation cohort of patches that can lead to reasonable results and (ii) in the semantic segmentation, the spatial dimensions (length and width) of the HSI data cubes are different, thus some patching is needed in order to have a unique input data shape for any neural network. Moreover, having convolutional layers at the initial layer still allows to examine the test performance on the full spatial size for different patients with one predict forward pass in the neural network without changing it at all. We use leave-patients-out cross validation, thus there are never patches from the same patient in both train and validation sets.

In our approach, we use 100 random patches of size 256×256 for each patient with the central pixels being 50/50 tumor and healthy classes and appropriate channel selection of the most significant channels as explained in Subsection III-A1. This means that we do 7-fold cross-validation with 1200 patches (12 patients) of size $256 \times 256 \times$ #channels for training and 200 patches (2 patients) for validation. (We have conducted additional experiments having standard train/validation/hold out set data partition or having nested cross-validation – thus leaving less patients for reporting results, but the results are similar to those with this partition scheme.) Each patch contains both tumor and healthy tissue and in fact, covers most of the spatial part of each HSI cube. It depends on the tissue, but in each patch for training both healthy and tumor tissues are decently represented. To achieve better generalization, we apply standard data-augmentation techniques such as rotation and flipping of the patches. As a loss function dice coefficient is used for for training and validation that compares the overlap of the full-size prediction map with the annotation map. Additionally, we have experimented with alternative losses like the focal loss, but comparable (or worse) and less stable results were obtained. Although, this loss could be promising, it also brings two additional parameters that have to be tuned. With these input patches and annotations (size 256x256x2 tumor/no tumor), we train a U-Net neural network [28] variant with batch normalization. The architecture is visualized in Figure 5a.

In order to separately examine the contribution of the VIS and NIR parts of the data cube, we derive dual-stream U-Net based neural network. As visualized in Fig. 5b, there is a separate stream/branch of the VIS and NIR accepting the corresponding parts of the data cube, followed by appropriate averaging or maximization for the final prediction.

1) Details on the training process: We have experimented with different batch size and, as expected, reasonable batch size of 16 or 32 leads to more stable validation curves compared to smaller batch size (e.g., batch size of 4). This is demonstrated in the Fig. 6 where the train and validation learning curves per epochs are show for two experiments that only differ in the batch size. We have experimented with different loss optimizers like Adam [38], standard SGD (Stochastic gradient descent), RMSprop [39] etc., but it seems the performance is not affected much by this choice, thus we decided for the former in most of the experiments. We have experimented with different values of the learning rate and values in the range 10^{-4} or smaller leads to slow validation loss improvement per epoch and it might require many epochs to achieve a decent validation loss. On the other hand, learning rates in the range of 10^{-3} or bigger leads to having a perfect train loss in the earlier epoch that often leads to overfitting in some of the folds. Therefore, a moderate learning rate of 0.5×10^{-4} is often the best choice although this can vary per fold.

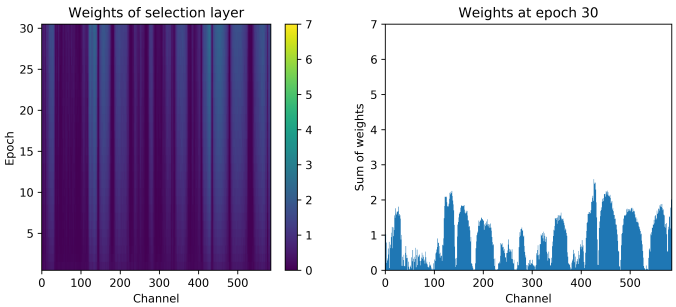


Fig. 7: Channel selection results.

TABLE III: Results of the experiments with different deep neural networks.

Architecture	Dice	AUC	Acc.	Sens	Spec
Structural 9x9 x 10ch (from VIS) (Fig. 4b)	0.793	0.855	0.769	0.738	0.839
Structural 21x21 x 45ch (from VIS)	0.744	0.844	0.811	0.755	0.838
Spectral NIR (Fig. 4a)	0.810	0.866	0.787	0.783	0.794
Spectral VIS	0.840	0.881	0.818	0.847	0.755
Spectral Stacked	0.839	0.917	0.844	0.795	0.894
Dual stream Stacked (Fig. 4c)	0.853	0.895	0.834	0.843	0.813
HSI U-Net NIR all 256ch (Fig. 5a)	0.821	0.879	0.932	0.801	0.957
HSI U-Net VIS all 384ch	0.860	0.915	0.948	0.866	0.964
HSI U-Net Stacked 32ch	0.848	0.893	0.938	0.828	0.959
HSI U-Net Stacked 64ch	0.870	0.912	0.950	0.857	0.968
HSI U-Net Stacked 128ch most important	**0.891**	0.924	**0.958**	0.873	**0.975**
HSI U-Net Stacked 256ch	0.870	0.914	0.951	0.860	0.968
HSI U-Net Stacked 128+64ch central	0.877	0.919	0.955	0.866	0.973
HSI U-Net Stacked 256+128ch central	0.864	0.923	0.953	0.879	0.967
HSI U-Net Stacked 128+64ch most important	0.851	0.910	0.948	0.855	0.966
HSI U-Net Stacked 88+64ch most important	0.851	0.909	0.948	0.851	0.967
HSI U-Net Stacked 64+64ch most important	0.878	0.923	0.956	0.874	0.972
HSI Dual U-Net 128+64ch central (avg. output) (Fig. 5b)	0.873	**0.938**	0.940	**0.935**	0.941
HSI Dual U-Net 88+40ch most important (avg. output)	0.870	0.926	0.947	0.896	0.956
HSI Dual U-Net 256+128ch most important (max output)	0.838	0.927	0.941	0.906	0.948

IV. Results and Discussion

A. Channel selection

The proposed method starts with a selection of important channels. Figure 7 (left) shows the weights of the selection layer as the training progresses. The choice of channels does not change over the epochs, but does become more distinct as shown by the increasing weights of selected channels. Figure 7 (right) shows the weights at epoch 30 and illustrates how selected channels are clustered together. It also shows that no channels are given the maximum weight in all folds indicating that selected channels are given low weights in some folds. From Fig. 7 (right), we can see that important channels lie both in the VIS and NIR part of the spectra, and it will be demonstrated later (see Fig. 10) that both are crucial for spotting the tumor. There is a difference in the selection band for a single patient and the conclusion on which channels are more important is made based on all patients thus those having recognized as the most important in the majority of the patients being ranked higher as a cumulative contribution across all patients. This is the only (i) fair and general approach, (ii) it is important if we bring additional data from a new patient to have this selection done a-priory (and not repeating it over again for data that might not be present anymore); and it is crucial for deployment in practice as band selection is time consuming so inference phase is more feasible in this way. A general conclusion is that the very initial and very end bands are the least significant; while various bands are quite important in the middle of the range as well as reasonably at the beginning and close to the end in the range. After the important channels are selected, we can proceed with smaller size cube patch input with noise channels filtered for the proposed neural networks.

B. Performance results and comparison

We evaluate the results of the proposed semantic segmentation method and compare it with the spectral, structural, dual-stream pixel-wise approaches proposed. We also want to stress that we have evaluated several other alternatives for: U-Net depth, the number of input channels, regularization techniques, hyper-parameter optimization, but due to space limitation, we show the best and most representative results in Table III as the others are worse or with comparable performance and properties. We use several performance metrics like the dice coefficient, the area under the ROC-curve (AUC), accuracy, sensitivity, and specificity for evaluating the validation results. Using all these performance metrics allows to evaluate the performance reasonably, even when the tumor/healthy presence in the data is imbalanced without the need to choose a classification threshold. The validation results of these metrics are given in Table III[2], showing that semantic segmentation by HSI U-Net variant with 128 input channels that can capture both spectral and spatial aspects has the best performance (in bold). It is important to stress that by both using less and more than 128 channels the performance is degraded (Table III) as either some tumor information is ignored or some noisy channels dominate in the decision, respectively. It is also interesting to mention that the proposed HSI U-Net variant still works well, although starting with less initial filters compared to the input channels, opposite to standard U-Net for RGB images (where 3 channels are significantly less than the initial number of filters), thus realizing an immediate pooling/selection effect. With the best performing HSI U-Net, we achieve a mean dice coefficient and AUC validation scores of 0.891 ± 0.053 and 0.924 ± 0.036, respectively. In Fig. 8, the performance on a per-patient base is shown, demonstrating the algorithm works well (with some reasonable variance) in all cases. In addition, it is interesting to mention U-Net experiments with RGB patches show significantly worse performance (around 0.8 for both dice and AUC) than those with multiple HSI channels,

[2](i) "HSI U-Net Stacked Xch" refers to all 640 channels (384 VIS and 256 NIR) are stacked in a cube, channel importance selection algorithm is applied and the X most significant are selected; (ii) "HSI U-Net Stacked $X + Y$ch central" means the central X VIS channels are selected (from 384 in total), the central Y NIR channels are selected (from 256) and then they are stacked together in a HSI cube; (iii) "HSI U-Net Stacked $X + Y$ch most important" means channel importance is separately applied for VIS and NIR, such that the most important X VIS channels are selected, the most important Y NIR channels are selected and then they are stacked together in a HSI cube.

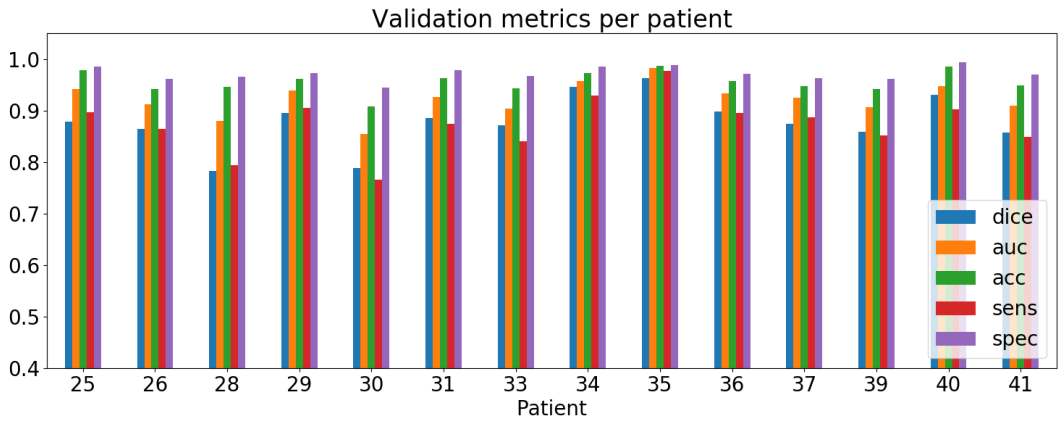


Fig. 8: Validation performance represented per patient by the dice coefficient, AUC, accuracy, sensitivity and specificity. On the x-axis are the patient ID. (The ID values are not of particular meaning or importance.)

which suggests that the HSI channels richness is important for improved precision. On the other hand, Halicek *et al.* [41] conducted research focusing on this aspect and found HSI did not provide significant advantages over the identification of tumor margins in ex-vivo tissue compared to RGB imagery. Therefore, we admit more in-depth research is needed in this direction. To better illustrate the accuracy of the prediction compared to the ground truth labels, we depicted the hard prediction map and the ground truth maps for 4 patients in Figure 9. It is also important to mention that although the network is trained on fixed patches, the reported test results are based on the original HSI dimension (that is different for each HSI image) directly using the obtained model, because the convolutional kernels of the initial layers can take arbitrary input size with a single pass. We have done experiments with U-Net networks of different number of layers, different units per layer, different dropout values or other regularizers, but the performance is worse or comparable to those reported in Table III.

C. The importance of VIS and NIR channels

From the results in Table III, we can see VIS channels are generally more significant that NIR for tumor segmentation. However, there are rare cases where the NIR channels are the most crucial and decisive for tumor segmentation, while the tumor is not spotted by VIS channels. The importance of the NIR spectral alone or even better in a combination with VIR in a dual stream U-Nets architecture (Fig. 5b) is visualized in Figure 10. In this figure, for this particular patient, we can see that by using NIR channels, the algorithm captures the tumor tissue, while this is not a case by using VIS channels only. Although there are such cases as in Figure 10, which highlights the importance of the NIR channels, in the majority of the cases VIS channels are those contributing the most in spotting the tumor tissue.

V. Conclusion

Real-time tumor segmentation during surgery is an important and challenging task. On the other hand, recent hyperspectral camera developments offer additional possibility for better quality and more insights that can lead to more accurate segmentation. Several techniques have been proposed in the past, mostly based on standard machine learning and pixel-wise approaches. To the best of our knowledge, this is the first work using deep learning U-Net [28] semantic segmentation for tumor detection and trainable channels selection for both NIR and VIS HSI spectra. The proposed semantic segmentation shows superior performance over the other alternatives (average dice coefficient and area under the ROC-curve of 0.891 ± 0.053 and 0.924 ± 0.036, respectively). We also demonstrate that channel selection and filtering is beneficial over the full spectra for achieving better performance, that both VIS and NIR channels are important and very good performance can be achieved even with a limited amount of data. Moreover, we have shown that the often omitted near-infrared (NIR) spectra is crucial for detecting the tumor in some cases. The hyperspectral cameras have been demonstrated to be powerful tools in many fields; and the technology, augmented with accurate algorithms, has a huge potential in biomedical engineering and medicine and with promising results available could be a doctors' supportive tool for real-time surgeries and an alternative to digital pathology.

Acknowledgments

The work of Caifeng Shan is partially supported by the National Natural Science Foundation of China under Grant 61972188.

References

[1] F. Bray *et al.*, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] E. Healthcare, "Tumor resection," https://www.emoryhealthcare.org/orthopedic-oncology/tumor-resection.html.

[3] T. Yee Chen *et al.*, "The clinical significance of pathological findings in surgically resected margins of the primary tumor in head and neck carcinoma," *International Journal of Radiation Oncology Biology Physics*, vol. 13, no. 6, pp. 833 – 837, 1987.

[4] T. R. Loree and E. W. Strong, "Significance of positive margins in oral cavity squamous carcinoma," *The American Journal of Surgery*, vol. 160, no. 4, pp. 410 – 414, 1990, papers of the Society of Head and Neck Surgeons presented at the 36th Annual Meeting.

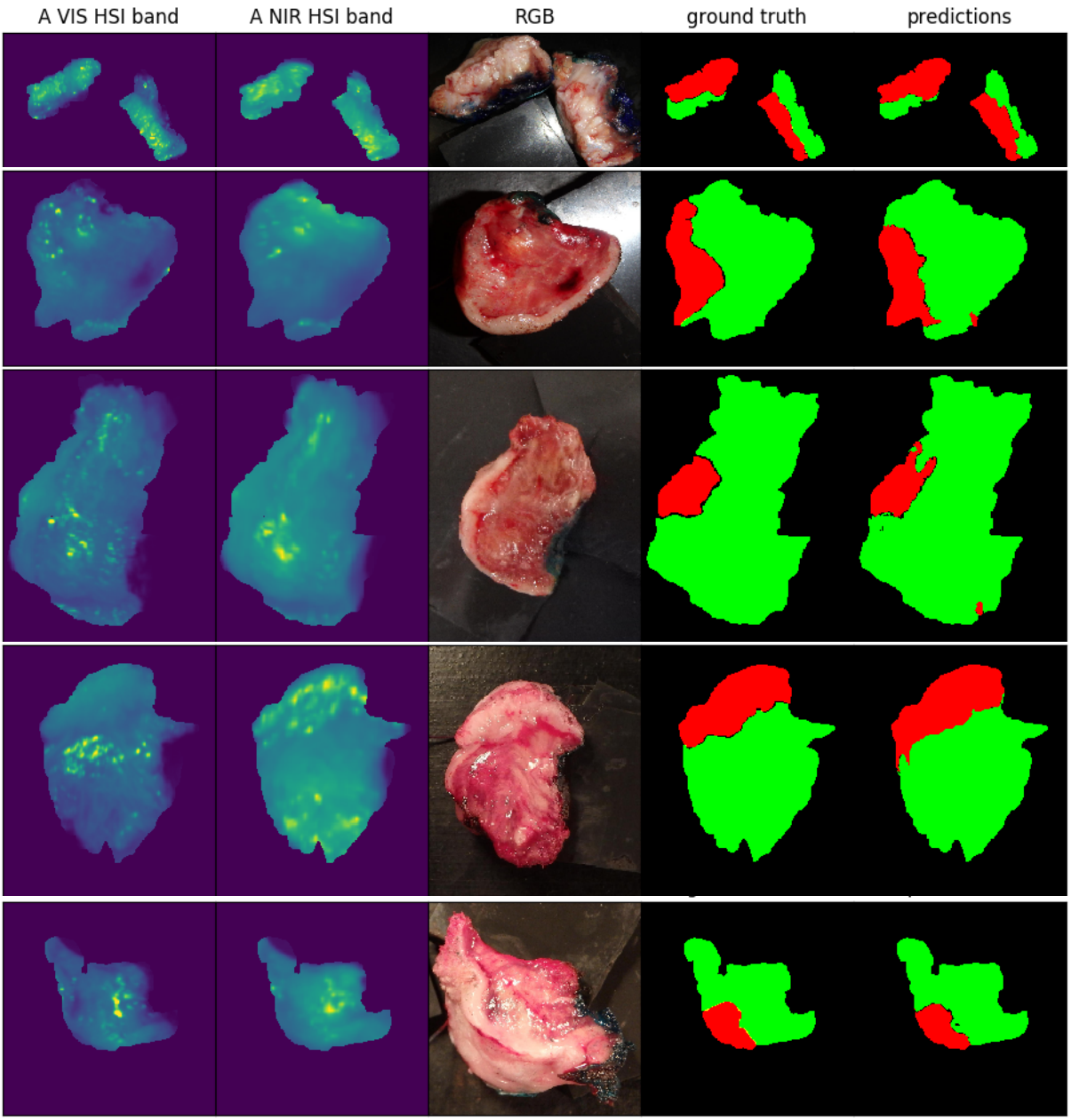


Fig. 9: A VIS (1st column) and NIR (2nd column) HSI bands, the RGB representations (3rd column), ground truth (4th column; green is for healthy, red is for tumor tissue) and hard predictions of our method (5th column; predicted pixels are green for values smaller than 0.5, or red for values at least 0.5) for 4 patients. (For the two HSI slices (first two columns), standard *viridis* color map has been used, see e.g., [40].)

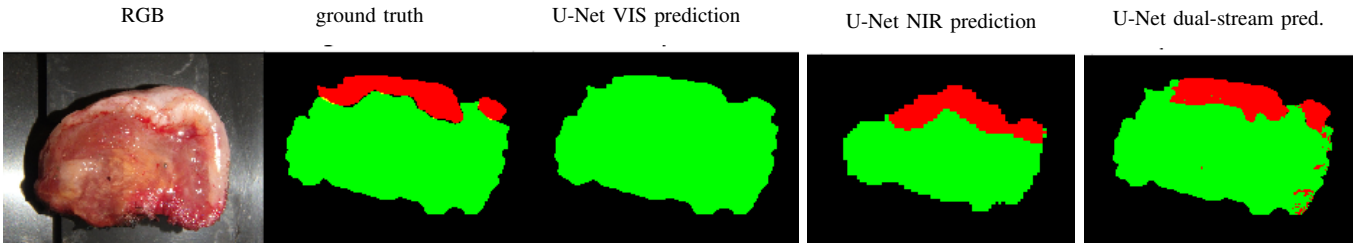


Fig. 10: An exceptional case. Ground truth (green for healthy, red for tumor) and hard predictions (threshold 0.5) by VIS and NIR U-Net.

[5] A. Binahmed *et al.*, "The clinical significance of the positive surgical margin in oral cancer," *Oral Oncology*, vol. 43, no. 8, pp. 780 – 784,

2007.
[6] R. W. Smits *et al.*, "Resection margins in oral cancer surgery: Room for improvement," *Head & Neck*, vol. 38, no. S1, pp. E2197–E2203, 2016.
[7] A. F. Goetz, "Three decades of hyperspectral remote sensing of the earth: A personal view," *Remote Sensing of Environment*, vol. 113, 2009, imaging Spectroscopy Spec. Iss.
[8] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of Biomedical Optics*, vol. 19, no. 1, p. 010901, 2014.
[9] M. A. Calin *et al.*, "Hyperspectral imaging in the medical field: Present and future," *Applied Spectroscopy Reviews*, vol. 49, no. 6, pp. 435–447, 2014.
[10] L. Ma *et al.*, "Deep learning based classification for head and neck cancer detection with hyperspectral imaging in an animal model," *Proc.SPIE*, vol. 10137, 2017.
[11] B. Fei *et al.*, "Label-free reflectance hyperspectral imaging for tumor margin assessment: a pilot study on surgical specimens of cancer patients," *Journal of Biomedical Optics*, vol. 22, no. 8, pp. 1–7, 2017.
[12] M. Halicek *et al.*, "Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging," *Journal of Biomedical Optics*, vol. 22, no. 6, 2017.
[13] E. Kho *et al.*, "Broadband hyperspectral imaging for breast tumor detection using spectral and spatial information," *Biomed. Opt. Express*, vol. 10, no. 9, pp. 4496–4515, Sep 2019.
[14] G. Lu *et al.*, "Detection of head and neck cancer in surgical specimens using quantitative hyperspectral imaging." *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 23 18, pp. 5426–5436, 2017.
[15] ——, "Detection and delineation of squamous neoplasia with hyperspectral imaging in a mouse model of tongue carcinogenesis," *Journal of Biophotonics*, vol. 11, no. 3, p. e201700078, 2018.
[16] H. Akbari *et al.*, "Hyperspectral imaging and quantitative analysis for prostate cancer detection," *Journal of Biomedical Optics*, vol. 17, no. 7, 2012.
[17] L. Ma *et al.*, "Adaptive deep learning for head and neck cancer detection using hyperspectral imaging," *Visual Computing for Industry, Biomedicine, and Art*, vol. 2, no. 1, p. 18, Nov 2019.
[18] D. Ravì *et al.*, "Manifold embedding and semantic segmentation for intraoperative guidance with hyperspectral brain imaging," *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1845–1857, Sep. 2017.
[19] R. Pike *et al.*, "A minimum spanning forest-based method for noninvasive cancer detection with hyperspectral imaging," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 653–663, March 2016.
[20] H. Fabelo *et al.*, "Spatio-spectral classification of hyperspectral images for brain cancer detection during surgical operations," *PLOS ONE*, vol. 13, no. 3, pp. 1–27, 03 2018.
[21] G. Florimbi *et al.*, "Accelerating the k-nearest neighbors filtering algorithm to optimize the real-time classification of human brain tumor in hyperspectral images," *Sensors*, vol. 18, no. 7, 2018.
[22] G. Lu *et al.*, "Hyperspectral imaging of neoplastic progression in a mouse model of oral carcinogenesis," in *Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging*, B. Gimi and A. Krol, Eds., vol. 9788, International Society for Optics and Photonics. SPIE, 2016, pp. 252 – 259.
[23] B. Regeling *et al.*, "Hyperspectral imaging using flexible endoscopy for laryngeal cancer detection," *Sensors*, vol. 16, no. 8, 2016.
[24] S. Ortega *et al.*, "Detecting brain tumor in pathological slides using hyperspectral imaging," *Biomed. Opt. Express*, vol. 9, no. 2, pp. 818–831, Feb 2018.
[25] M. Halicek *et al.*, "Hyperspectral imaging of head and neck squamous cell carcinoma for cancer margin detection in surgical specimens from 102 patients using deep learning," *Cancers*, vol. 11, no. 9, 2019.
[26] G. Lu *et al.*, "Spectral-spatial classification for noninvasive cancer detection using hyperspectral imaging," *Journal of Biomedical Optics*, vol. 19, no. 10, pp. 1 – 18, 2014.
[27] H. Chung *et al.*, "Superpixel-based spectral classification for the detection of head and neck cancer with hyperspectral imaging," in *Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging*, B. Gimi and A. Krol, Eds., vol. 9788, International Society for Optics and Photonics. SPIE, 2016, pp. 260 – 267.
[28] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.
[29] S. G. Brouwer de Koning *et al.*, "Near infrared hyperspectral imaging to evaluate tongue tumor resection margins intraoperatively (Conference Presentation)," in *Optical Imaging, Therapeutics, and Advanced Technology in Head and Neck Surgery and Otolaryngology 2018*, B. J. F. W. M.D. *et al.*, Eds., vol. 10469, International Society for Optics and Photonics. SPIE, 2018.
[30] E. Kho *et al.*, "Hyperspectral imaging for resection margin assessment during cancer surgery," *Clinical Cancer Research*, 2019.
[31] H. Akbari *et al.*, "Cancer detection using infrared hyperspectral imaging," *Cancer Science*, vol. 102, no. 4, pp. 852–857, 2011.
[32] S. G. Brouwer de Koning *et al.*, "Toward assessment of resection margins using hyperspectral diffuse reflection imaging (400-1,700nm) during tongue cancer surgery," *Lasers in Surgery and Medicine*, vol. n/a, no. n/a, 2019.
[33] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, June 1989.
[34] H. Iqbal, "Harisiqbal88/plotneuralnet v1.0.0," Dec. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.2526396
[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.
[36] I. Goodfellow *et al.*, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[37] P. Weijtmans, "From hyperspectral data cube to prediction with deep learning," Master's thesis, Eindhiven University of Technology (TU/e), Eindhoven, The Netherlands, March-April 2019.
[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (International Conference for Learning Representations)*, ser. ICLR '15, 2015, pp. –.
[39] T. Tieleman and G. Hinton, "Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.
[40] Matplotlib, "Perceptually uniform sequential colormaps." [Online]. Available: https://matplotlib.org/3.2.1/tutorials/colors/colormaps.html#miscellaneous
[41] M. Halicek *et al.*, "Tumor detection of the thyroid and salivary glands using hyperspectral imaging and deep learning," *Biomed. Opt. Express*, vol. 11, no. 3, pp. 1383–1400, March 2020.