

A Video Transformer Network for Thyroid Cancer Detection on Hyperspectral Histologic Images

Minh Ha Tran ^a, Ofelia Gomez ^a, and Baowei Fei^{a,b,c,*}

^a Center for Imaging and Surgical Innovation, University of Texas at Dallas, Richardson, TX

^b Department of Bioengineering, University of Texas at Dallas, Richardson, TX

^c Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX

*Corresponding author: bfei@utdallas.edu, Website: <https://fei-lab.org>

ABSTRACT

Hyperspectral imaging is a label-free and non-invasive imaging modality that seeks to capture images in different wavelengths. In this study, we used a vision transformer that was pre-trained from video data to detect thyroid cancer on hyperspectral images. We built a dataset of 49 whole slide hyperspectral images (WS-HSI) of thyroid cancer. To improve training, we introduced 5 new data augmentation methods that transform spectra. We achieved an F-1 score of 88.1% and an accuracy of 89.64% on our test dataset. The transformer network and the whole slide hyperspectral imaging technique can have many applications in digital pathology.

Keywords: Hyperspectral imaging, whole-slide imaging, thyroid cancer, deep learning, transformer attention-based neural networks, image classification.

1. INTRODUCTION

Head and neck cancer (HNC) is the seventh most common type of cancer worldwide [1]. With close to 1 million cases annually, it includes cancer of the oropharynx, the tongue, the nasal cavity, the thyroid, and the larynx. Thyroid cancer is one of the most common types of HNC [1]. It is different from other types of HNC in which it is the cancer of the follicular cells, whereas most HNC are cancers of the squamous cells [1, 2]. Surgical resection remains the primary method of treatment. If there are cancer cells present at the edge of the resected tissue, the margin is labeled as positive. An acceptable margin range from 5 to 20 mm [3]. Because the thyroid region is vital for quality of life, it is important to preserve the normal tissue. Precise tumor margin will help improve surgical outcome and quality of life [4, 5].

Computer-aided diagnosis (CAD) uses advanced machine learning techniques, such as support vector machines (SVM) and deep neural networks, to help identify tumors [6-9]. Deep neural networks, such as convolutional neural networks (CNN), have been used to identify classify normal and tumor tissue. Deep learning with pre-training networks are able to learn morphological features, which can be finetuned on specific datasets at a fraction of the cost and time. Halicek *et al.* [7] trained a pretrained Inception network to classify different types of HNC. They achieved an area under the receiving operating characteristics curve (AUC-ROC) value of 0.916 for the detection of squamous cell carcinoma.

Hyperspectral imaging (HSI) is an optical imaging technique that captures both the spatial and spectral information of tissues. It is non-invasive and label-free. HSI has been used to detect tumor on hematoxylin and eosin (H&E)-stained slides [8-12]. Our previous work [12] and the works of Ma *et al.* [13] show that using hyperspectral images of HNC to train neural networks can result in improvements of patch-wise classification over using regular RGB images. However, the lack of publicly available hyperspectral data makes it difficult to train a neural network. Researchers in other fields leverage pretrained networks and augmentation to solve the problem. However, there are no pretrained network publicly available for the task of hyperspectral image classification. It is still not known what wavelengths contribute to differentiating tumor from normal tissue. If the question of wavelengths is addressed, future researchers

can focus on specific wavelengths that is relevant to identification tasks, reducing the complexity of their data and networks.

In recent years, vision transformers are being used alongside CNN for the task of image classification at large scales [14]. Transformers are multi-layer neural networks that use the mechanism of self-attention. In a self-attention layer, the input is linearly transformed using three learnable weights to produce “keys,” “values,” and “queries” (K, V, Q respectively). The terms were inspired by database retrieval, where a query can be used to retrieve the values based on the keys. Here, each token in the input has its own key-value pairs to be searched for by other tokens, and its own queries to match up with other tokens. The measurement of how relevant each token is in relation to each other is a matrix called the attention matrix, and it is calculated using the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK}{\sqrt{d}}\right)V$$

Where d refers to the vector dimensionality of the keys. To increase the performance of self-attention, multiple attention heads run parallel to each other where each head captures a different type of dependency. The results were then concatenated in what is called “multi-head self-attention” and linearly transformed through another learnable weight W_o to produce the output.

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{Output} = W_o \begin{bmatrix} \text{head}_1 \\ \vdots \\ \text{head}_n \end{bmatrix}$$

Where W^Q, W^K, W^V are the weights to linearly transform the queries, keys, and values. Vision transformer is already used in digital histopathology [15-18]. However, training them often requires large amounts of data to overcome the lack of inductive bias. With HSI, it is not often feasible to produce data at such a scale. Steiner *et al.* [19] suggested that proper augmentation is a must for training vision transformer with small amount of input. They suggested that using RandAugment technique, one can achieve the same validation accuracy with only 50% of the original data. RandAugment is an augmentation method that reduces the search space significantly by focusing on only two hyperparameters m and n ; m refers to the strength of the augmentation; and n refers to the number of times the operations are successively applied. Örnberg *et al.* [20] used RandAugment augment spectral data. However, their transformations are not combined with spatial transformations. Several augmentation methods specific toward histology data were developed in the past. Tellez *et al.* [21] and Faryna *et al.* [22] augmented hematoxylin-eosin-DAB contents digitally. They achieved this by using a known matrix of RGB responses of different types of stain to estimate the amount of stain within each picture. However, when the number of spectra is large, it is difficult to estimate the spectral responses correctly.

In this paper, we propose the use of a pretrained video classifier network, TimeSformer [23], for the classification of tissue on HSI data. TimeSformer is a modified vision transformer network that has temporal attention on top of spatial attention. Researchers explored using transformer networks for the task of hyperspectral images classification [24-26]. He *et al.* [27] discussed the use of video classifiers for hyperspectral classification and pointed out the differences in temporal and spectral data. However, since vision transformer has little inductive bias [14], we believe that it can be finetuned on hyperspectral data. Vision transformers require large amounts of input data, which is often not feasible in hyperspectral imaging. We proposed new augmentation methods for RandAugment in the spectral domain, which are based on both real physical phenomena and non-negative factorization (NMF) decomposition. Our contributions include: (1) we showed that a pretrained video classifier can be successfully adapted to classify hyperspectral images, reducing the need for large amount of training; (2) we proposed new transformation methods to augment hyperspectral data; (3) we tested our network on a relatively large dataset of WS-HSI.

2. METHODS

2.1 Histologic Slides from Thyroid Cancer Patients

We described how the histologic slides were obtained in our previous studies [12, 13, 28-30]. In this study, a total of 49 histologic slides were used. This makes our study the largest repository of WS-HSI as of date. The histologic slides were from 32 patients. The slides were fixed and stained with H&E stains. We used an automatic acquisition system to capture hyperspectral whole slide images. Our system consists of an optical microscope, a compact hyperspectral snapsan camera, a high-precision X-Y motorized stage, and a high precision Z motorized stage. With this platform, we capture 84-band images at the wavelength range of 467-721 nm. Whole slide images were divided into non-overlapping patches of size $84 \times 250 \times 250$, which were then resized into $84 \times 224 \times 224$ pixels. Table 1 details the number of slides and image patches used for training, validation, and testing. This dataset contains 39 slides that were either entirely tumor or normal tissue, and 10 slides that include both cancerous and normal tissue. Tumor margins were annotated by an HNC pathologist, using RGB scans of the whole slides under an automatic RGB histology scanner.

Table 1. Summary of the number of patients, whole slide hyperspectral images (WS-HSI), and patches used in training, validation, and testing. All cancerous WS-HSI are thyroid follicular carcinoma.

	Training	Validation	Testing	Total
Number of patients	17	5	10	32
Number of WS-HSI	30	9	10	49
Number of image patches	89,406	21,652	39,384	150,442
Number of positive patches	39,419	11,379	16,403	67,201
Number of negative patches	49,987	10,273	22,981	83,241

2.2 Transformer Network

We used weights from a pretrained TimeSformer as a starting point [23]. The image patches was first divided into 14×14 equal-size sub-patches, and then each sub-patch was converted into a one-dimensional sequence using a convolutional kernel [14]. Positional and class embeddings were used in a similar manner as in vision transformer [14]. In TimeSformer, there were two attention layers: a spatial attention layer and a temporal attention layer [23]. The spatial attention layer compared the sub-patch with other sub-patches of the same time frame, and the temporal attention layer compared the sub-patch with itself at different time frames. We modified the network so that temporal attention became spectral attention (Figure 1a). Figure 1b describes the layout of a divided spectral-spatial attention block. The network contains 12 spectral-spatial blocks, as shown in Figure 1c.

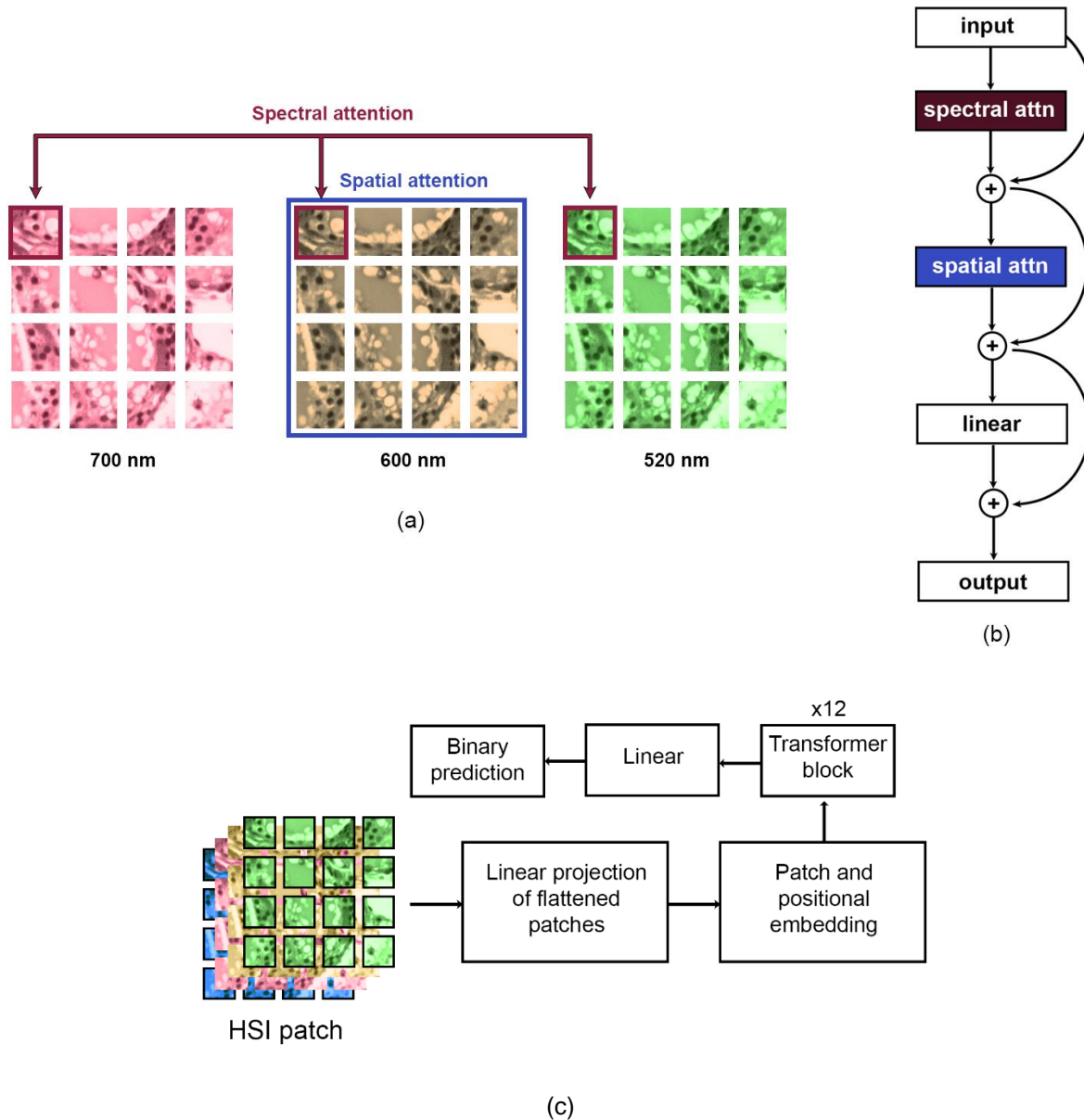


Figure 1. (a) Divided attention mechanism. Spatial attention compared the sub-patch with other sub-patches within the same spectra, whereas spectral attention compared the sub-patch with itself from different spectrum. (b) Sequence of a spectral-spatial attention transformer block. (c) Layout of the transformer network. Figures were adapted from [23].

2.3 Augmentation of Spectral Data

In addition to the list of spatial transformations already used in RandAugment, we append the following tailored transformations: spectral noise, spectral crosstalk, spectral shift, spectral spike, and NMF estimation. Spectral noise, spectral crosstalk, and spectral shift all modified the entire spectrum of the image. Spectral noise randomly shift the signal by a random uniform value. Spectral crosstalk averages the signals in the spectral dimension with five of its closest neighbors. Spectral spike transformation introduced irrelevant data in the form of saturated values at certain spectra. NMF estimation estimated the content of different tissue stains and modified them to simulate different

staining outcomes [31]. Figure 2 shows the algorithms overview for RandAugment algorithm, where we found that the values of $m=5$ and $n=3$ is the most suitable for our study. Figure 3 shows the spectrum before and after different types of transformations using $m=5$ and $n=3$.

```

Algorithm 1. RandAugment


---


Transformations:
{ Identity, Rotate, ShearX, ShearY, TranslateX,
  TranslateY, Brightness, Contrast, SpectralNoise,
  CrossTalk, SpectralShift, SpectralSpike, NMF }
for  $i \leftarrow 1$  to  $n$ :
  Randomly select a transformation
  image = transformation(image,  $m$ )


---



```

Figure 2. RandAugment algorithm, modified for hyperspectral data.

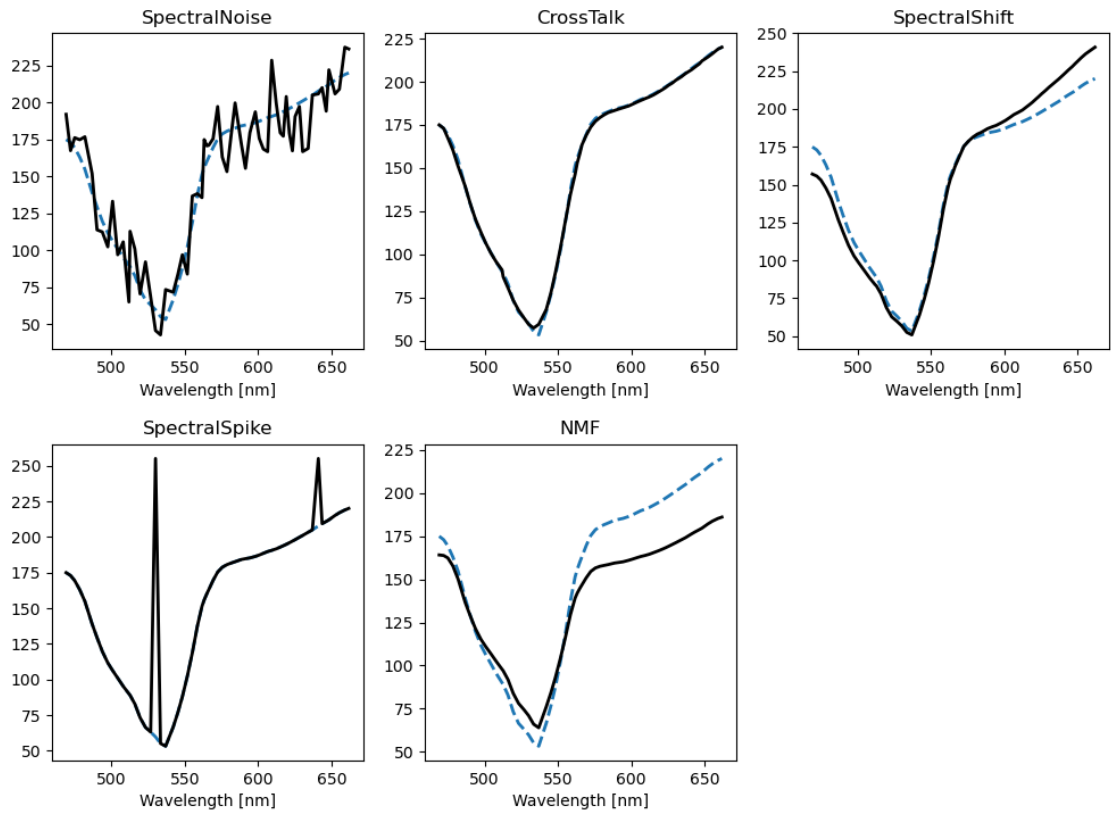


Figure 3. Spectral data augmentation by using specific spectral transformations. Curves show the average spectra of the hyperspectral patch before (dashed blue lines) and after (solid black line) the transformation.

2.4 Experimental Setup

The network was implemented in PyTorch and trained on NVIDIA RTX A6000s with 48Gb GPU. We finetuned all weights of the transformer network for 3 epochs. We used batch gradient descent optimizer with Nesterov momentum

of 0.9. The batch size and initial learning rate for the transformer network were 64 and 1×10^{-3} . We divided the development set into 30 training slides and 9 validation slides. The testing data were separate from the training data and were from a group of different 10 patients. The metric we used for best validation is the F-1 score, which is the average of the precision and recall. Precision and recall are calculated from the true positive (TP), true negative (TN), and false positive (FP) rates. Only the epoch that achieves maximum validation F-1 score was selected for testing.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = \frac{\text{Precision} + \text{Recall}}{2}$$

2.5 Visualization of Attention

We visualize the weights of the finetuned attention layers using the attention unrolling technique. This method aggregates all attention weights across all layers into one. At each layer, a single head produce an attention map $A_{h,L}$ which is the SoftMax multiplication of the keys $K_{h,L}$ and queries $Q_{h,L}$. The attention map is averaged across all heads and added with the identity matrix I to produce the layer attention \hat{A}_L . The rollout attention is calculated as the dot product of all attention maps at all layers.

$$\text{Attention}_{h,L} = \text{softmax}(Q_{h,L}K_{h,L}^T)$$

$$\hat{A}_L = I + \text{average}(\text{Attention}_L)$$

$$\text{rollout} = \hat{A}_1 \odot \hat{A}_2 \odot \dots \odot \hat{A}_{12}$$

Because this network has a visual attention layer and a spectral attention layer, we produced two attention maps. The visual attention maps were rescaled to match the original dimension of the patch image. The spectral attention maps were averaged across the spectral dimension to produce a relative attention map. The results were two attention maps for every input patch. Each map shows relative importance of the spatial and spectral segment to the overall classification output.

3. RESULTS

3.1 Test Result

On the validation dataset, the network achieves a best accuracy score of 95.82% and a best F-1 score of 0.9608. On the testing dataset, the network achieves an accuracy of 89.64 % and an F-1 score of 88.08 %. We measured a weighted recall value of 89.64 %. Table 1 shows the per-slide performance.

Table 2. Per-slide performance of TimeSformer network on the test dataset.

Slide	F-1	Accuracy
T1	0.6593	0.7801
T2	0.7765	0.8686
T3	0.9167	0.8826
T4	0.8422	0.8456
T5	0.9318	0.9530
T6	0.9634	0.9700

T7	0.9169	0.9398
T8	0.9498	0.9305
T9	0.8612	0.8809
T10	0.8970	0.9060
Weighted average	0.8808	0.8964

3.2 Visualization of Visual Attention

We selected four image patches with high prediction confidences. The result is shown in figure 4. In this figure, the warmer colors (orange, yellow) correspond to regions of the image that are more relevant toward the output prediction, and the cooler colors (blue, green) correspond to regions that are less relevant. Note that the attention is only relevant within the pictures themselves, so it is not appropriate to compare the attention maps between two pictures. In the positive predictions samples (true positive and false negative), the most relevant regions are the regions of connective tissues, whereas the regions of the nuclei were given less attention. This might be due to how the network associate morphology of the connective tissue regions with positive predictions. On the other hand, in the negative prediction samples (false positive and true negative), the highest attention was given to the colloid regions, indicating that they are the most relevant morphological features when predicting negative classes.

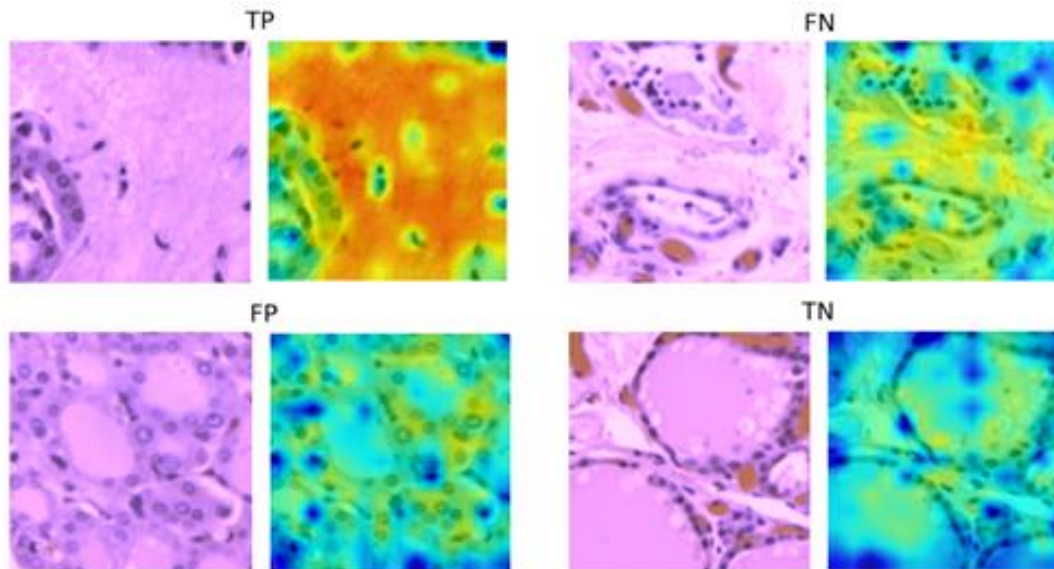


Figure 4. RGB representation of sample image patches that the network predicted. Next to them on the right are the attention maps overlaying the patches, ranging from orange (high attention) to blue (low attention). TP: true positive, FN: false negative, FP: false positive, TN: true negative. Positive prediction means cancer and negative prediction means normal tissue.

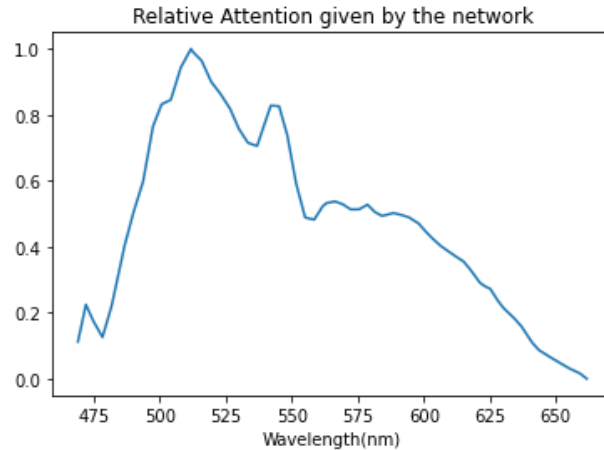


Figure 5. Relative attention in the spectral domain, achieved by taking the average value of the weights of all spectral attention layers and then normalized by the maximum value.

Figure 5 shows the flattened attention map of the average spectral attention when the network produced true positive predictions. This attention curve is already normalized, with the higher values demonstrating higher relevancy in prediction outcome. The curve shows that the network paid the most attention toward the wavelength range 500 – 550 nm, which were the wavelengths with the highest contrast between nuclei and extracellular connective tissues [12]. Even though the attention curve should not be taken literally as a demonstration of clinically important spectra, we believe that the attention curve holds considerable importance in wavelength selection.

4. CONCLUSION AND DISCUSSION

We presented a relatively large whole slide hyperspectral image database of thyroid cancer patients. Using this dataset, we demonstrated the use of a fully attention-based neural network for the task of hyperspectral image classification. Previously, transfer learning using a video classifier has not been seriously considered due to the differences between temporal dimension and spectral dimension. However, our work shows that by using an attention-based network, correlations in the spectral dimension can be learned by the network successfully. By visualizing the attention map, we showed that the network was using relevant spatial and spectral features to classify images. One challenge with an attention-based network is the size of the network. Our network contains 125M trainable parameters, which is significantly larger than that of convolutional networks. A training loop takes a long time even on a large GPU hardware. Finetuning the network should be used instead of re-training the network from scratch, in order to provide a good starting point of convergence [19]. We believed that distillation, a technique of using a large network to train a smaller network, can be the key to achieve both high accuracy and small models [32].

To improve network training, we proposed new data transformations that modify the spectral components. Our data transformation is based on the RandAugment technique and combines both spatial and spectral transformations. The spectral transformations were inspired by real physical phenomena that impacts the quality of hyperspectral images. We showed that by using data augmentation, we overcame the limitation of small dataset and achieved both high train and validation accuracy.

We visualized the attention of our model in the visual and spectral dimensions. The visual attention map shows that the network is relying on real biological morphology to make predictions whether the patch is normal or tumorous. However, it also potentially explains how the network can make incorrect predictions. Even though the morphology of intracellular matrix is important for tumor identification, it is not all that the histologist rely on to properly make decisions. In the future, a large-scale study using a large database of hyperspectral images, along with more than one histologist, is needed in order to properly validate all improvements of the vision transformer on hyperspectral images over regular convolutional networks on regular RGB images.

5. ACKNOWLEDGEMENTS

This research was supported in part by the U.S. National Institutes of Health (NIH) grants (R01CA156775, R01CA204254, R01HL140325, and R21CA231911), the Cancer Prevention and Research Institute of Texas (CPRIT) grant RP190588, and the Eugene McDermott Graduate Fellowship 202009 at the University of Texas at Dallas.

REFERENCES

- [1] L. Q. Chow, "Head and neck cancer," *New England Journal of Medicine*, vol. 382, no. 1, pp. 60-72, 2020.
- [2] S. Marur and A. A. Forastiere, "Head and neck cancer: changing epidemiology, diagnosis, and treatment," in *Mayo Clinic Proceedings*, 2008, vol. 83, no. 4: Elsevier, pp. 489-501.
- [3] M. Amit *et al.*, "Improving the rate of negative margins after surgery for oral cavity squamous cell carcinoma: a prospective randomized controlled study," *Head & neck*, vol. 38, no. S1, pp. E1803-E1809, 2016.
- [4] K. T. Robbins *et al.*, "Surgical margins in head and neck cancer: Intra-and postoperative considerations," *Auris Nasus Larynx*, vol. 46, no. 1, pp. 10-17, 2019.
- [5] M. W. Kubik *et al.*, "Intraoperative margin assessment in head and neck cancer: a case of misuse and abuse?," *Head and neck pathology*, vol. 14, no. 2, pp. 291-302, 2020.
- [6] J. Folmsbee *et al.*, "Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018: IEEE, pp. 770-773.
- [7] M. Halicek *et al.*, "Head and Neck Cancer Detection in Digitized Whole-Slide Histology Using Convolutional Neural Networks," *Sci Rep*, vol. 9, no. 1, p. 14043, Oct 1 2019, doi: 10.1038/s41598-019-50313-x.
- [8] S. Ortega *et al.*, "Hyperspectral imaging and deep learning for the detection of breast cancer cells in digitized histological images," *Proc SPIE Int Soc Opt Eng*, vol. 11320, Feb 2020, doi: 10.1117/12.2548609.
- [9] 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 (Basel)*, vol. 11, no. 9, Sep 14 2019, doi: 10.3390/cancers11091367.
- [10] S. Ortega *et al.*, "Hyperspectral and multispectral imaging in digital and computational pathology: a systematic review," *Biomedical Optics Express*, vol. 11, no. 6, pp. 3195-3233, 2020.
- [11] L. Ma *et al.*, "Pixel-level tumor margin assessment of surgical specimen in hyperspectral imaging and deep learning classification," presented at the Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, 2021.
- [12] M. H. Tran *et al.*, "Thyroid carcinoma detection on whole histologic slides using hyperspectral imaging and deep learning," in *Medical Imaging 2022: Digital and Computational Pathology*, 2022, vol. 12039: SPIE, pp. 101-111.
- [13] L. Ma, A. Rathgeb, H. Mubarak, M. Tran, and B. Fei, "Unsupervised super-resolution reconstruction of hyperspectral histology images for whole-slide imaging," *Journal of Biomedical Optics*, vol. 27, no. 5, p. 056502, 2022.
- [14] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [15] Z. Gao *et al.*, "Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021: Springer, pp. 299-308.
- [16] R. J. Chen and R. G. Krishnan, "Self-supervised vision transformers learn visual concepts in histopathology," *arXiv preprint arXiv:2203.00585*, 2022.
- [17] M. A.-E. Zeid *et al.*, "Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers," in *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2021: IEEE, pp. 224-230.

- [18] H. Chen *et al.*, "Gashis-transformer: A multi-scale visual transformer approach for gastric histopathology image classification," *arXiv e-prints*, p. arXiv: 2104.14528, 2021.
- [19] A. Steiner *et al.*, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.
- [20] O. Örnberg, "Semi-supervised methods for classification of hyperspectral images with deep learning," ed, 2020.
- [21] D. Tellez *et al.*, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical image analysis*, vol. 58, p. 101544, 2019.
- [22] K. Faryna *et al.*, "Tailoring automated data augmentation to H&E-stained histopathology," in *Medical Imaging with Deep Learning*, 2021.
- [23] G. Bertasius *et al.*, "Is space-time attention all you need for video understanding?," in *ICML*, 2021, vol. 2, no. 3, p. 4.
- [24] X. He *et al.*, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sensing*, vol. 13, no. 3, p. 498, 2021.
- [25] Y. Qing *et al.*, "Improved transformer net for hyperspectral image classification," *Remote Sensing*, vol. 13, no. 11, p. 2216, 2021.
- [26] B. Yun *et al.*, "Spectr: Spectral transformer for hyperspectral pathology image segmentation," *arXiv preprint arXiv:2103.03604*, 2021.
- [27] M. He *et al.*, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017: IEEE, pp. 3904-3908.
- [28] L. Ma *et al.*, "Hyperspectral microscopic imaging for the detection of head and neck squamous cell carcinoma on histologic slides," in *Medical Imaging 2021: Digital Pathology*, 2021, vol. 11603: International Society for Optics and Photonics, p. 116030P.
- [29] L. Ma *et al.*, "Pixel-level tumor margin assessment of surgical specimen in hyperspectral imaging and deep learning classification," in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2021, vol. 11598: International Society for Optics and Photonics, p. 1159811.
- [30] 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, p. 1367, 2019.
- [31] G. Lu *et al.*, "Estimation of tissue optical parameters with hyperspectral imaging and spectral unmixing," in *Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 2015, vol. 9417: SPIE, pp. 199-205.
- [32] H. Touvron *et al.*, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021: PMLR, pp. 10347-10357.