

Masked image modeling in medical hyperspectral imaging: reconstruction evaluation and downstream tasks

Kelden Pruitt ^{a,b}, Hemanth Pasupuleti ^{a,b}, James Yu ^{a,b,c}, Weston DeAtley ^{a,b}, Baowei Fei ^{a,b,c} *

^a University of Texas at Dallas, Department of Bioengineering, Richardson, TX;

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

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

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

ABSTRACT

Self-supervised pre-training has been shown to improve deep learning networks in various tasks including natural language processing and computer vision. While this approach has shown promise in various fields, more development and translation need to be dedicated to medical imaging applications. Current literature scarcely focuses on thorough assessment of implemented pre-training approaches as well, potentially hindering performance in downstream tasks. In this work, we leverage a state-of-the-art pre-training architecture with hyperspectral imaging (HSI) to effectively encode spatial and spectral features of various *ex vivo* tissues. We utilize a masked image modeling scheme to perform pre-training on an internal dataset captured with a high-speed hyperspectral laparoscopic imaging system. Our network implements sequential spectral and spatial attention, factorizing the model for efficiency. Evaluation of both pre-training and finetuned classification was performed on a validation dataset unseen in either set to prevent data leakage. Pre-training results are qualitatively assessed through reconstruction visualization and quantitatively assessed with mean absolute error (MAE), achieving a value of 0.0294 on the validation dataset. To test the capabilities of the pre-trained model, we finetuned the network as an abdominal tissue classifier, achieving 87.9% accuracy on 17 classes with frozen model weights. Overall, we present a masked autoencoding framework for the pre-training of hyperspectral images with an emphasis on the evaluation of the network for potential improvements in downstream tasks such as tissue classification and segmentation.

Keywords: Hyperspectral imaging, masked autoencoding, deep learning, abdominal tissue

1. INTRODUCTION

Attention-based deep learning models, transformers, have been extensively explored since their inception [1]. Variations of the architecture have been constructed for a variety of tasks, including a divergence from natural language processing tasks to computer vision (CV) [2], with many of the leading models using transformer components. A recent application of this architecture was in the domain of self-supervised training for improved performance called masked autoencoding. The use of masked autoencoders (MAEs) has been shown to increase accuracy and improve training times in various CV tasks when used for pre-training. The basic premise includes introducing "masking" to random patches of pixels in the input. The model is then exposed to the limited input (unmasked patches) and tasked with predicting the masked patches. Slightly different approaches exist with variations introduced throughout the pipeline, including the method of reconstruction, which can be a lightweight transformer [3, 4] or a simple linear layer [5], and how or when to introduce masking. Applications have been primarily in RGB image analysis, which have seen medical translation in works such as Prov-GigaPath [6], UNI [7], and CONCH [8] which focus on whole-slide histology tasks.

Hyperspectral imaging (HSI) captures both spectral and spatial features of imaged scenes [9] and has shown the ability to outperform RGB imaging in tasks with comparable dataset sizes [10, 11]. Previously, networks for medical HSI have largely been convolutional-based [12, 13] or simply artificial neural networks [14], with applications of modern transformer architectures scarce and largely attributed to the remote sensing community [15]. While convolutional approaches lend themselves well to CV tasks generally, due to the inductive bias towards spatial reasoning and translational equivariance present via shared weights, transformers have been shown to outperform them given sufficient model and dataset size. The self-attention mechanism that underpins the transformer allows for long-range dependencies

to be made and global features to be extracted from datasets. As such, our work seeks to leverage the capabilities of transformers, particularly in a masked autoencoder, and assess its downstream performance in medical HSI through pre-training on an *ex vivo* tissue dataset we have curated.

2. METHODS

2.1 Dataset acquisition

Hyperspectral images were acquired using a setup similar to our previous work [16]. Optical modifications were made to incorporate three high-speed hyperspectral cameras that utilize spectrally resolved detector arrays. Hypercubes were registered and combined to produce a final image with a raster size of 510 x 270 x 55, including band images from 460 nm – 960 nm. A C++ application developed for the system was used for imaging and further processing was performed in MATLAB. The dataset utilized in this work consisted of *ex vivo* abdominal tissues from porcine, bovine, and galline animal models. Standard HSI processing including white reference calibration and dark current correction were performed.

The dataset used in this work consisted of over 1100 hyperspectral images from up to eight organs across four animal models, resulting in 17 unique classes. The overall class distribution can be seen in Figure 1. A train/test split of 75/25 was utilized resulting in 864 images for training and 289 for testing. The final 10 bands were dropped due to noise and center cropping was performed as a preprocessing step, resulting in a final image size of 224 x 224 x 45. With a patch size of 16 x 16 x 5, this results in 1764 patches per image, resulting in over 1.5 million spatial-spectral patches for training.

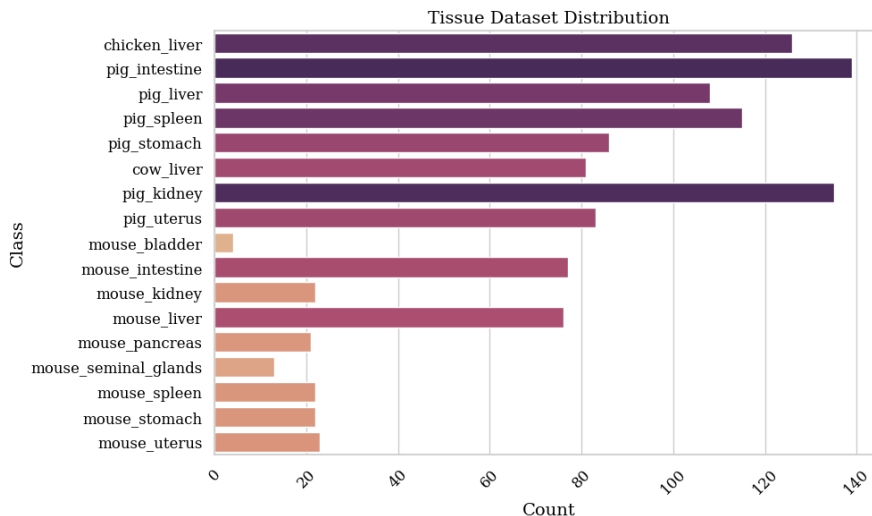


Figure 1. Hyperspectral imaging dataset of various types of tissue for training and evaluation.

2.2 Network architecture for pre-training

We leverage an architecture based on MaskedSST [15] with modifications made to suit our application. The architecture was originally designed for use in remote sensing which generally has lower spatial resolution and higher spectral dimensions when compared to medical HSI. As such, input parameters were significantly altered to reflect this difference, including larger spatial and decreased spectral patch sizes. The architecture can be visualized in Figure 2. Random masking is utilized at a ratio of 0.6 and block-wise patch embeddings, a feature leveraged in the original work, was implemented. This allows each spectral block to have a unique embedding as opposed to a single learned embedding for all patches. We used a transformer with an embedding dimension of 1024, multi-layer perceptron layer with 256 dimensions, depth of 12 layers, and 12 heads. Pre-training was performed for 1000 epochs using AdamW optimizer [17] with a learning rate of $1e^{-4}$ on a Nvidia RTX A6000 GPU for 18 hours.

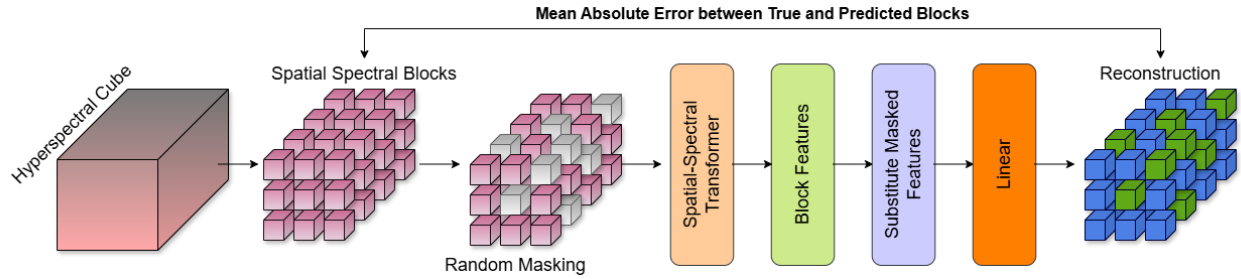


Figure 2. Block diagram outlining masked autoencoder workflow for hyperspectral image reconstruction. Masked and unmasked features are fed into a linear layer to predict missing intensities. Mean absolute error of masked pixels then used for training.

2.3 Evaluation and Finetuning

Qualitative evaluation of self-supervised pre-training consisted of visual assessment of the reconstructed hypercube. Scripting in Python was performed to visualize masked patches and the corresponding reconstruction in the validation set. As random masking was performed, the reconstruction visualization varies in each band image. To assess performance throughout the captured spectral range, we visualize six spectral images sampled throughout the captured hypercube. Quantitative assessment of the pre-training task is gauged by the MAE of reconstructed patches compared to their ground truth intensities, which are calculated as a per pixel average over all masked patches in the validation set.

In order to assess the ability of the model to learn high-level features of the data, we further tested the performance of the pre-training task by training a classifier with the learned model weights. To implement this, the spatial-spectral transformer from pre-training was used without altering the weights. The validation set from earlier was reserved for testing while the model was finetuned on a new split of the training data. The output from the model was averaged across all patches, producing a single feature vector which was used as input to a fully-connected perceptron layer. Therefore, only model parameters associated with this layer were tuned during backpropagation. Top-1 and top-3 accuracies were evaluated as metrics to assess the performance of the finetuned model, with top-1 accuracy gauging the frequency with which the model correctly predicted the class as its top choice and top-3 accuracy including predictions if the ground truth label was in the model's top three choices.

3. RESULTS

3.1 Qualitative reconstruction results

Upon completion of pre-training, we evaluated the performance of the autoencoder by visualizing reconstructed bands from hypercubes outside of the training set. Figure 3 demonstrates the performance of the network to reconstruct imaged pork intestine at band images throughout the input. The image shows a variety of high- and low-level features, as well as varying reflectance intensities as assessed from the visible to near-infrared range. General contours of the tissue are recovered despite the limited unmasked input. Some finer details, such as the small notch present in the bottom right corner of the sample, are accurately reconstructed as seen in the 29th band image. Other features, such as the remaining dark area characteristic of the endoscopic view in the top left corner of later bands, aren't as accurately reconstructed when masked out. Overall, the stitched images indicate the successful learning of spectral and spatial features by the network.

3.2 Quantitative performance and finetuning

Upon completion of the pre-training task, our model achieved an average MAE of 0.0244 and 0.0294 on the training and validation sets, respectively. While the model was trained for 1000 epochs, MAE values plateaued after approximately 600 epochs. Training error was relatively lower than the validation set, suggesting slight overfitting by the model potentially. Figure 4 highlights the performance of the network when finetuned for classification, showing the predictions against the ground truth labels. Overall, the diagonal is the most heavily populated region of the plot, suggesting the features learned during pre-training are significant and can be used to distinguish tissues accurately. Top-1 and top-3 accuracies of 87.9% and 98.3% were achieved. Cells outside the diagonal indicate misclassifications. Identical organs

from different animal models or spectrally similar organs contribute to a large portion of these errors, particularly kidney, liver, and spleen tissue across the imaged animals.

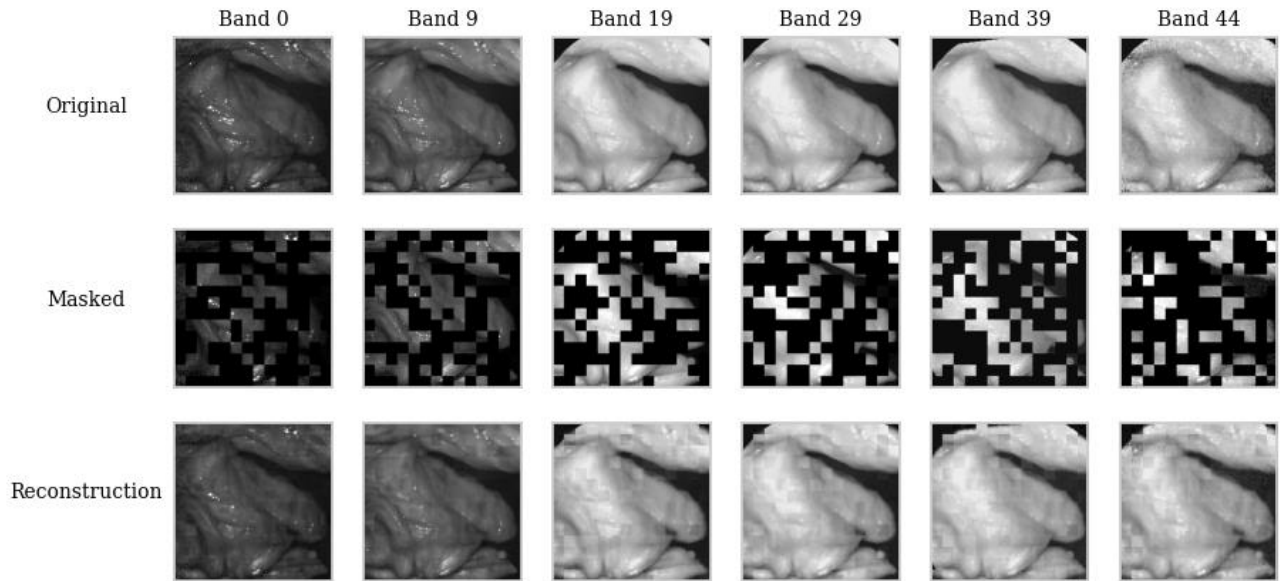


Figure 3. Masked reconstruction performance compared to original band images. Each column corresponds to a different band image from the hypercube. The top row shows the original band image. The middle row displays the remaining visual information for the network while the bottom row displays the stitched reconstruction.

Tissue Classification Confusion Matrix

True Label \ Predicted Label	pig_stomach	pig_liver	pig_spleen	pig_kidney	pig_intestine	pig_uterus	cow_liver	chicken_liver	mouse_stomach	mouse_liver	mouse_spleen	mouse_kidney	mouse_intestine	mouse_uterus	mouse_bladder	mouse_seminal_glands	mouse_pancreas
pig_stomach	20	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
pig_liver	0	23	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0
pig_spleen	0	2	24	0	0	1	1	0	0	1	0	0	0	0	0	0	0
pig_kidney	0	3	0	29	1	0	0	1	0	0	0	0	0	0	0	0	0
pig_intestine	0	0	0	0	34	0	1	0	0	0	0	0	0	0	0	0	0
pig_uterus	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0
cow_liver	0	2	0	0	0	0	15	2	0	0	1	0	0	0	0	0	0
chicken_liver	2	0	1	0	0	0	0	29	0	0	0	0	0	0	0	0	0
mouse_stomach	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
mouse_liver	0	0	0	0	0	0	0	0	0	18	0	0	1	0	0	0	0
mouse_spleen	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	0	0
mouse_kidney	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
mouse_intestine	0	0	0	0	0	0	0	0	1	0	0	0	18	0	0	0	0
mouse_uterus	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0
mouse_bladder	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
mouse_seminal_glands	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
mouse_pancreas	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	4

Figure 4. Confusion matrix of the finetuned model on tissue classification task.

4. DISCUSSION AND CONCLUSION

We present a deep learning architecture for pre-training for medical HSI that has the potential to improve downstream tasks such as cancer classification and tissue segmentation in the future. We emphasize the thorough evaluation of the pre-training performance of the model, as optimal pre-training allows networks to learn relevant features of the target dataset more effectively. Further, we assess the application of the model on a downstream classification task on a dataset with 17 classes. While these architectures have seen use in other fields, medical HSI has yet to be an area of focus and improvement with these networks.

Deep learning techniques pose substantial promise as they can parse and interpret the large amounts of data and features offered by HSI. Transformer architectures, with their ability to scale with large amounts of data, could be the key to unlocking the full potential of HSI. We demonstrate the ability of a masked image modeling scheme to perform unsupervised pre-training on a set of *ex vivo* tissues imaged with a high-speed HSI setup and later perform classification with raw features extracted from input hypercubes. While the performance of our model showed competence in reconstruction with regards to spectral features, there is room for improvement and expansion. The hyperparameters could be further tuned (masking ratio, patch size) and larger, more diverse datasets can augment our current corpus. Alternative masking strategies, such as tube and spectral masking, can also be investigated for the ability to improve the model's performance and ability to learn robust features. Future work will emphasize not only the improvement in pre-training, but also investigate performance boosts in downstream tasks such as patch-wise classification and pixelwise segmentation. In the future, this model and architecture may allow for state-of-the-art classification abilities when coupled with HSI.

5. ACKNOWLEDGMENTS

Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA288379 and R01CA204254 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under Award Number RP240289 and RP240542. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

- [1] A. Vaswani *et al.*, "Attention Is All You Need," presented at the Neural Information Processing Systems, Long Beach, CA, USA, 2017.
- [2] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," presented at the ICLR, June 3, 2021.
- [3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16000-16009.
- [4] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10078-10093, 2022.
- [5] Z. Xie *et al.*, "SimMIM: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653-9663.
- [6] H. Xu *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, pp. 1-8, 2024.
- [7] R. J. Chen *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850-862, 2024/03/01 2024, doi: 10.1038/s41591-024-02857-3.
- [8] M. Y. Lu *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 863-874, 2024/03/01 2024, doi: 10.1038/s41591-024-02856-4.
- [9] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *J Biomed Opt*, vol. 19, no. 1, p. 10901, Jan 2014, doi: 10.1117/1.JBO.19.1.010901.
- [10] 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.

- [11] S. Ortega *et al.*, "Hyperspectral Imaging for the Detection of Glioblastoma Tumor Cells in H&E Slides Using Convolutional Neural Networks," *Sensors (Basel)*, vol. 20, no. 7, Mar 30 2020, doi: 10.3390/s20071911.
- [12] A. Studier-Fischer *et al.*, "Spectral organ fingerprints for machine learning-based intraoperative tissue classification with hyperspectral imaging in a porcine model," *Sci Rep*, vol. 12, no. 1, p. 11028, Jun 30 2022, doi: 10.1038/s41598-022-15040-w.
- [13] S. Seidlitz *et al.*, "Robust deep learning-based semantic organ segmentation in hyperspectral images," *Med Image Anal*, vol. 80, p. 102488, Aug 2022, doi: 10.1016/j.media.2022.102488.
- [14] B. Jansen-Winkel *et al.*, "Feedforward Artificial Neural Network-Based Colorectal Cancer Detection Using Hyperspectral Imaging: A Step towards Automatic Optical Biopsy," *Cancers (Basel)*, vol. 13, no. 5, Feb 25 2021, doi: 10.3390/cancers13050967.
- [15] L. Scheibenreif, M. Mommert, and D. Borth, "Masked Vision Transformers for Hyperspectral Image Classification," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 17-24 June 2023 2023, pp. 2166-2176, doi: 10.1109/CVPRW59228.2023.00210.
- [16] K. Pruitt *et al.*, "Design and validation of a high-speed hyperspectral laparoscopic imaging system," *Journal of Biomedical Optics*, vol. 29, no. 9, p. 093506, 2024. [Online]. Available: <https://doi.org/10.1117/1.JBO.29.9.093506>.
- [17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.