

Optimization of Transfer Learning of Foundation Models for Hyperspectral Histologic Imaging

Michael D. Hellman^{a,b}, Ling Ma^{a,b}, James Yu^{a,b,c}, 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 (HSI) is a promising modality for digital pathology, but it is not yet widely adopted compared to traditional red-green-blue (RGB) histologic imaging. This study aims to develop techniques for transferring knowledge from histopathological foundation models trained on conventional RGB image datasets to models that can process data acquired by hyperspectral imaging. We used a dataset of 89 whole-slide hyperspectral histologic images from 54 patients to fine-tune three different foundation models. We also performed a hyperparameter search for each model and technique to identify general hyperparameter combinations well-suited for this task. Our results show that performing end-to-end fine-tuning of models generally outperforms other knowledge transfer paradigms, and that low learning rates and high weight decays tend to perform best for the transfer learning process. These findings partially contradict the common wisdom of first performing training only in the embedding layer, where gradients are concentrated. This study demonstrates a set of effective techniques for applying foundation models trained on RGB images to hyperspectral images for computational histopathology.

Keywords: hyperspectral imaging (HSI), histopathology, foundation model, transfer learning.

1. INTRODUCTION

Computational pathology has drawn growing interest in the past several years, particularly with the development of deep learning algorithms, which promise to substantially reduce human error in diagnosis and accelerate the process¹⁻³. Foundation models are a subclass of these deep learning models, which are trained on especially broad datasets and designed to be applicable to a range of different downstream tasks⁴. Several such foundation models have been recently introduced within the field of histopathology and have achieved state-of-the-art performance on numerous downstream tasks⁵⁻⁷. These models also exhibit promising few-shot performance metrics which make them strong candidates for application in domains where data is not widely available⁵. In cases where data is not abundant for pre-training, it may be possible to instead fine-tune a foundation model to achieve strong performance metrics on a particular task.

Hyperspectral imaging (HSI) is an emerging modality that is seeing growing use in biomedical imaging⁸. For example, in *ex vivo* tissues, Aboughaleb et al⁹ successfully applied hyperspectral imaging in the diagnosis and segmentation of breast cancer tissue from normal tissue, Halicek et al¹⁰ utilized hyperspectral data to classify squamous-cell carcinoma and thyroid cancer, and Tran et al¹¹ applied HSI for the segmentation of nervous tissue. Within histopathology, hyperspectral imaging has shown promise in the diagnosis of head and neck cancers^{12, 13}, colorectal cancer¹⁴, and membranous nephropathy¹⁵, as well as the analysis of hematological samples towards a variety of different ends¹⁶. Compared to conventional imagery, HSI features a substantially larger channel space, typically consisting of dozens or more channels, which represent different wavelengths of light. This increased spectral resolution allows for the identification of tissue features which may not be apparent using conventional imagery, improving diagnostic accuracy. Acquisition of hyperspectral imagery is more difficult and features larger overheads than conventional image acquisition methods. Hyperspectral images take up more space and require substantially longer scanning time to take high resolution images, and hyperspectral cameras are not as widely available as RGB cameras⁸. For these reasons, the availability of hyperspectral data is limited in comparison to RGB data.

The lower availability of hyperspectral data combined with the limited feasibility of mass acquisition makes it difficult for a foundation model to be trained on hyperspectral data from the ground up. However, performing fine-tuning on a model pre-trained on a large corpus of RGB data using a relatively small HSI dataset may potentially enable the model to use hyperspectral information to identify the same spatial features which the pre-trained RGB network was able to identify, in addition to new spectral features enabled by the increased spectral resolution of the system. There is a dearth of research on performing fine tuning from the conventional domain to the spectral, and more broadly on fine-tuning visual transformers on new, channel-dense information. While there is literature within the medical field covering transfer learning from RGB-model to RGB-model^{17,18}, there is far less covering the inverse, when images are of similar subjects but in a different format than what pre-training was performed on. To these ends it may therefore be valuable to investigate practices which yield the best results for performing transfer learning between different spectral domains.

In this work, we performed a study of the performance of various hyperparameter combinations and fine-tuning techniques across three different training paradigms and three different foundation models. We then used the results of this experiment to establish fine-tuning paradigms which we expect to translate well across all Vision Transformer¹⁹ (ViT)-based models.

2. METHODS

2.1. Dataset Preparation

For this work, we used 89 histologic slides sampled from 54 different patients with head and neck cancer squamous cell carcinoma. For 35 of these patients, both a cancer-positive (T) slide and a cancer-negative (N) slide was scanned, for 13 of the patients only a cancer-negative sample was scanned, and for 6 patients only a cancer-positive sample was scanned. Hyperspectral whole-slide histologic images of these slides were obtained using our developed automated hyperspectral whole-slide scanning microscope comprising of an inverted brightfield microscope and a customized snapscan hyperspectral camera covering 460-720 nm wavelength range²⁰. Each whole slide was scanned in tiles with a size of 1000 pixels \times 1000 pixels \times 87 bands, totaling 400,127 tiles across all patients. Hyperspectral images were taken at 4 \times objective magnification to reduce image size and scanning time, and for each slide, a digitized whole-slide image was obtained from our previous study²¹. The low-resolution hyperspectral image tiles and high-resolution RGB images were automatically registered²⁰. To overcome the low spatial resolution of the hyperspectral data, we use a technique proposed by Ma et al²², utilizing a pan-sharpening model which takes in high-resolution RGB images and low-resolution HSI images to reconstruct a high-resolution hyperspectral image. These reconstructions can be used in lieu of native high-resolution hyperspectral data when performing classification. These image pairs were then used to generate a dataset of the same number of reconstructed high-resolution hyperspectral images^{20,22}, each labeled as cancerous or normal. Finally, all registered RGB and hyperspectral image tiles were cropped into patches. The dimensions of each patch were 200 pixels by 200 pixels, and each contained 87 channels of spectral information within the 460-720 nm wavelength range. The full details of the image acquisition pipeline for this dataset are outlined in Ma et al²⁰.

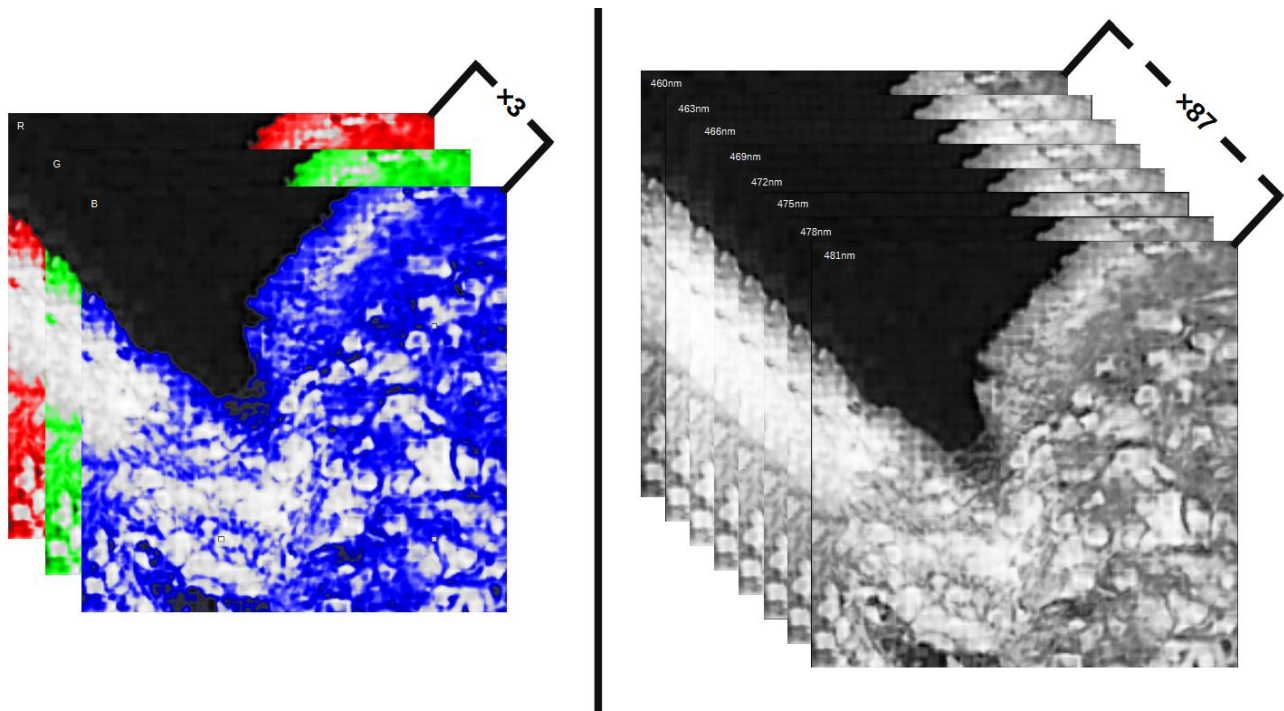


Figure 1. An illustration of the hyperspectral (right) and RGB (left) data domains.

2.2. Histopathology Foundation Models

To evaluate the performance of various transfer learning techniques, we selected three pathological foundation models which were pre-trained on RGB data, namely UNI⁵, Prov-Gigapath⁶, and CONCH⁷. To accommodate the greater channel space of the HSI data, we replaced the encoding layer for each model with a new layer whose convolution accepts eighty-seven channels instead of three. To attempt to preserve the knowledge of the pre-trained embedding layer in our new embedding layer, each spectral channel was weighted based on a product of their expected contribution to each of the R, G, and B channels and the embedded weights for each channel in the original foundation model's embedding layer. This information regarding each channel's contribution is drawn from the algorithm described in Ma et al.¹², itself derived from the work of Vos²³. Additionally, the classification head of each model was entirely replaced with a new binary classifier, as our dataset only has two classes.

2.3. Transfer Learning Methodology

To find the best approach to the transfer of knowledge from RGB-trained foundation models to HSI, we evaluated three different approaches to transfer learning. These were (1) end-to-end training, in which all learnable model parameters were trained for five epochs; (2) embedding-only training, where exclusively the embeddings and classification head were trained for five epochs; and (3) embedding-first training, where embedding-only training occurred for an epoch, followed by four epochs of end-to-end training.

Hyperparameter search was performed using the Bayesian optimization algorithm²⁴ across a parameter space consisting of learning rate, weight decay, and first and second betas. These were then passed to the AdamW optimizer²⁵, which was used to train the model. To prevent data leakage and ensure that our results were robust across the entire dataset, we employed a 3-fold cross-validation strategy for our training. First, we selected a stratified holdout set of approximately 10% of the dataset to be used for later testing, followed by the generation of three stratified folds. All stratification was performed along patient lines, ensuring that when performing inference on the test and validation sets, no samples from those patients had been trained on at any point. Cross-validated accuracy values for a hyperparameter set were taken as the average of the accuracies of each of the three folds. Bounds for each parameter space in

hyperparameter selection are outlined in Table 1. The best performing hyperparameter set was then selected, and a final test accuracy was generated by performing inference on the holdout set.

Table 1. Bounds of hyperparameter search. Distribution indicates how values were selected from within the sample. Uniform randomly selects a value from within the range, whereas Log Uniform randomly selects a value between the common logarithm of the lower and upper limits.

HYPERPARAMETER BOUNDS				
	Learning Rate	Weight Decay	Beta1	Beta2
Lower Limit	1e-7	1e-4	0.4	0.9
Upper Limit	0.1	0.5	0.9	0.999
Distribution	Log Uniform	Log Uniform	Uniform	Uniform

All training and inferences were run in parallel across 4 Nvidia A6000 GPUs, and data preprocessing was performed on two AMD EPYC 7302 64-core processors. Total training time for a fold averaged two hours, and a single parameter evaluation averaged six hours. Each of the three models were trained using each of the three different techniques, and ten hyperparameter sets were evaluated for each technique, summing to ninety hyperparameter evaluations overall. Our data augmentation and regularization pipeline consisted of vertical and horizontal flips, 90-degree rotations, and resizing of each image to 224×224 pixels. Data was then standardized, which was stratified along patient lines to avoid any data leakage. This was accomplished in a performant manner by pre-computing the per-channel mean and variance for each patient individually and combining these statistics at runtime for each patient in our training set. In this way, we achieved single-pass computation of all statistics required for standardization for any subset of our dataset.

3. RESULTS

In this work, we tested different transfer learning approaches on three RGB-trained histopathology foundation models, namely UNI, CONCH, and Prov-Gigapath. Since they are all based upon the ViT model architecture, similar behavior was exhibited with regards to hyperparameter combinations and training paradigms. Thus, we present the results for UNI as a representative sample of hyperparameter impact on all three. Prov-Gigapath and CONCH are not represented separately here. The results of hyperparameter selection for UNI for each fine-tuning method we described are outlined in Figure 2 to Figure 4.

Figure 2 shows the results of performing embedding-first tuning on UNI with 10 hyperparameter sets using our search scheme. We see that the three strongest-performing hyperparameter combinations featured learning rates lower than 10^{-3} , with the best performance falling between 10^{-5} and 10^{-6} . These optimal three also feature weight decays greater than 10^{-2} , with the best performance exhibited just above 10^{-1} . Interestingly, many of the best performing runs have beta 1 values lower than 0.6, however the best performance is exhibited closer to 0.8. No clear pattern emerges as performance relates to beta 2.

Figure 3 shows the results of performing embedding-only tuning on UNI ten times using our search scheme. It is immediately apparent that the performance of embedding-only tuning is closely bounded, the worst and best performing runs are only separated by an accuracy of about two points. This performance is substantially worse than the best case in the other two figures, two to four points worse than exhibited in end-to-end and embedding-first tuning. Also notable is the lack of any apparent groups of high-performance in the set. The best performance is exhibited with a learning rate of close to 10^{-1} , and a weight decay of roughly the same. This run's performance is unique among the set, it achieves validation accuracy improvements of nearly a full point over the next-best model.

Figure 4 shows the results of performing end-to-end tuning on UNI using our search scheme. This set has several high performing runs which settle at or above a validation accuracy of more than 0.915. The best performance is exhibited

with a learning rate of close to 10^{-6} and is very closely matched in performance by a run which originates with a learning rate of about the same. Both of these runs use a weight decay near 10^{-1} and diverge substantially in their beta values. Beta 1 values of runs which perform well tend to be greater than 0.6, and the best performers have values greater than 0.7. Both low and high Beta 2 values see successful runs.

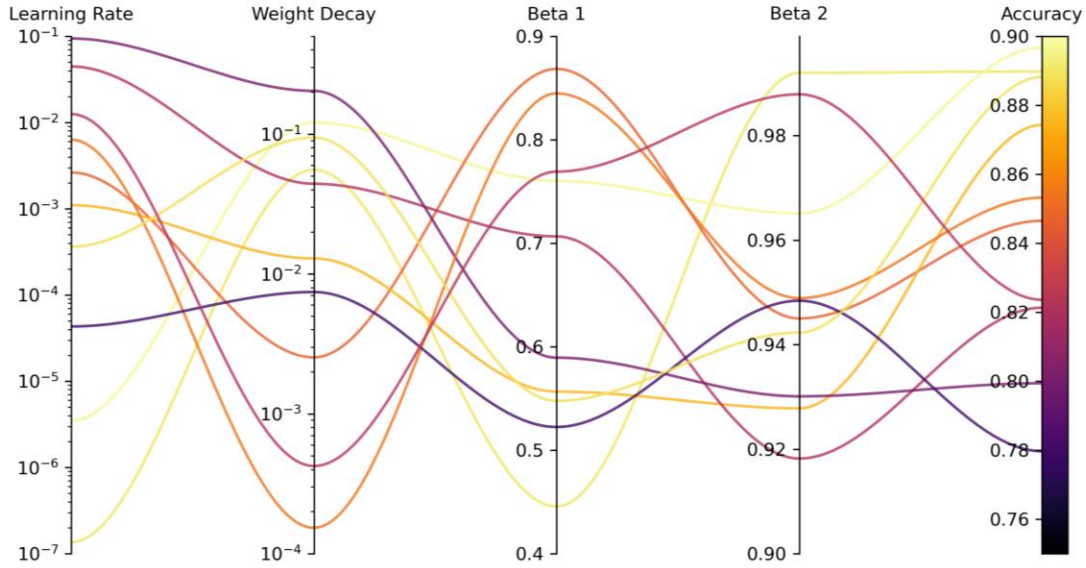


Figure 2. Results of Embedding-First Tuning for the UNI model. Each line running left to right represents a parameter set being evaluated, and each vertical line represents a particular parameter. The final line on the right represents the best accuracy achieved by a particular parameter combination. Parameter combinations represented by lines that are brighter yellow performed better than those represented by purple lines.

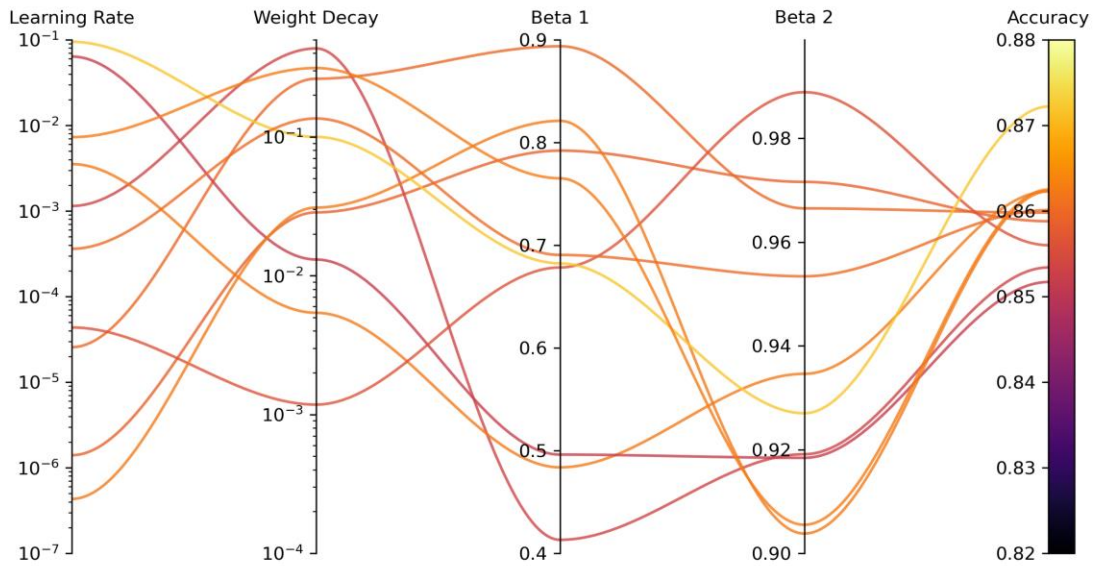


Figure 3. Results of Embedding-Only tuning for the UNI model. Each line running left to right represents a parameter set being evaluated, and each vertical line represents a particular parameter. The final line on the right represents the best accuracy achieved by a particular parameter combination.

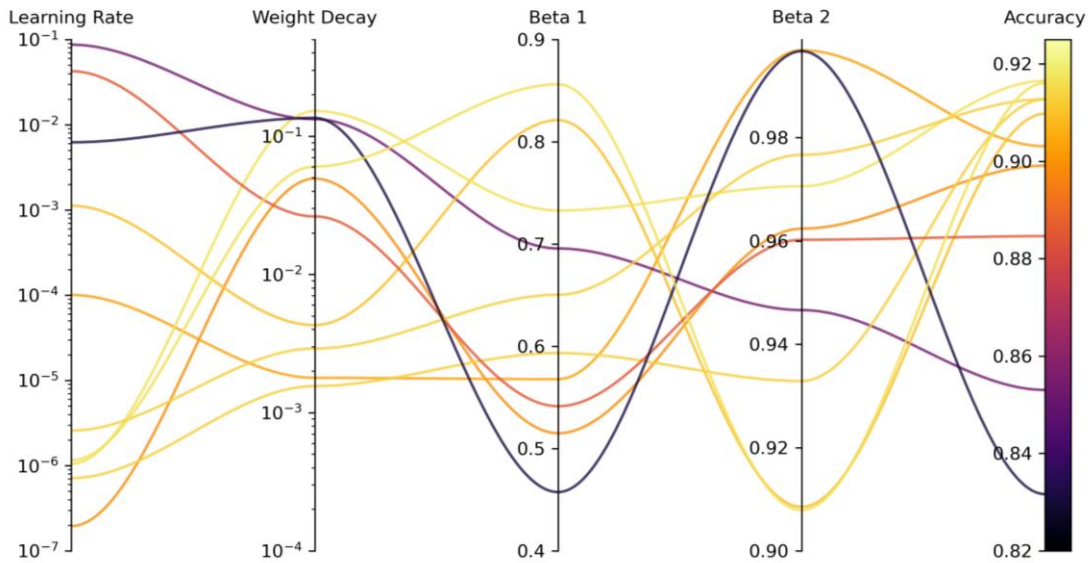


Figure 4. Results of end-to-end tuning for the UNI model. Each line running left to right represents a parameter set being evaluated, and each vertical line represents a particular parameter. The final line on the right represents the best accuracy achieved by a particular parameter combination.

The best performing hyperparameter combinations for each technique and model can be found in Table 2. For end-to-end and embedding-first fine tuning, learning rates between $1e-4$ and $1e-6$ tend to perform best. High weight decays also may have some positive correlation with accuracy, and the best performance across all models, i.e., UNI in end-to-end training, featured a weight decay of 0.152. This is a tenuous relationship, however, as strong performance was also achieved in end-to-end tuning with a weight decay of 0.003, and several passes performed poorly with high learning rates and weight decays. It would be difficult to point to any clear relationship between beta values and performance, possibly indicating that momentum does not substantially impact training performance, as beta values primarily govern the degree to which momentum is preserved across training iterations. Embedding-only training lacks any clear correlation between hyperparameter combination and performance outcomes, though its variance of accuracy is far less than that of the other two approaches.

Table 2. Best-performing hyperparameter combinations based on validation accuracy.

	Optimal Hyperparameters								
	Embedding-First			Embedding-Only			End-to-End		
	UNI	Prov-G	CONCH	UNI	Prov-G	CONCH	UNI	Prov-G	CONCH
Learning Rate	3.558e-6	0.008515	5.223e-6	0.09473	0.004351	0.02551	1.052e-6	6.208e-5	0.8967
Weight Decay	0.1208	0.2273	0.1883	0.09987	0.008268	0.2267	0.1522	0.04261	0.4789
Beta 1	0.7602	0.7938	0.8156	0.6823	0.8251	0.8682	0.7328	0.8532	0.7984
Beta 2	0.9651	0.9385	0.9861	0.9270	0.9329	0.9877	0.9706	0.9952	0.9472
Validation Accuracy	0.8967	0.8632	0.8602	0.8722	0.7943	0.8563	0.9165	0.8774	0.8978
Test Accuracy	0.8789	0.8453	0.8476	0.8845	0.8585	0.8532	0.9205	0.8904	0.8866

Note: Prov-Gigapath is abbreviated as Prov-G.

In Table 2, we see that across the board end-to-end training performed the best, followed by embedding-first and then embedding-only. Performance was largely consistent between the test and validation statistics, with the exception of the anomalously poor embedding-only performance exhibited by Prov-Gigapath on our validation set. Though comparisons between models are not the focus of this study, we note that UNI tends to perform the best out of the three in all cases. We further note that, in some cases, embedding-first and embedding-only reach near-parity with each other, reinforcing that, on our dataset, training the embedding layer in isolation seems to find a local minimum which becomes difficult to train the model out of.

4. DISCUSSION & CONCLUSION

In this work, we developed a robust approach for evaluating the efficacy of fine-tuning RGB-trained foundation models on HSI data. We tested three different common fine-tuning approaches across three different histopathology foundation models and isolated effective hyperparameter combinations for each. We find that end-to-end training was the most effective approach for performing transfer learning from RGB-trained to HSI-trained models within all models we tested. This approach is generally most effective when used in combination with very low values for learning rate and large weight decays. This combination of hyperparameters is generally associated with the prevention of overfitting, which may indicate that a relatively small amount of new learning has to be done by the models when adapting from RGB to HSI or may indicate that our dataset is too small to train for long periods of time without overfitting occurring.

Of the three learning techniques, end-to-end tends to perform the best, while embedding-only training tends to perform the worst. Embedding-only tuning fails to find a competitive fit and reliably achieves validation accuracy metrics between about 0.85 and 0.87. One possible explanation for this behavior is the possibility that catastrophic forgetting is occurring within the model due to the need to adjust our heavily modified embedding layer. Despite the measures described earlier, which attempt to align our new embedding weights with the original 3-channel weights, the process is not perfect, and it's likely that the embedded representations of data in our new model are substantially different from the embedded representations that the rest of the model is tuned to expect. This sudden adjustment of the embedding layer may be causing the model to enter a space with an uneven loss landscape, which is correlated with the occurrence of catastrophic forgetting in machine learning models²⁶. The model then finds a local minimum for loss and fails to exit it or ever properly develop embedded representations which closely correspond to those expected by the transformer layers. As an extension of this, embedding-first training may partially recover from this initial instance of catastrophic forgetting through adjustment of the transformer blocks, but this may also result in adverse effects on the transformer block parameters learned in the original training steps. Together, we see that the approach of adjusting only the embedding layer, initially or otherwise, tends to have an adverse effect on model performance.

End-to-end tuning of model parameters outperformed both other paradigms, exhibiting both a substantially higher best and average-case performance. Though leaving all parameters unfrozen throughout training should not affect gradient accumulation in the embedding layer, it is possible that catastrophic forgetting in the embedding layer is still being mitigated as the transformer blocks can here immediately adapt to the new embedding weights, reducing the magnitude of any destructive changes to embedding layer parameters on successive iterations of the learning process. This represents a substantial distinction between the accepted best practice for fine-tuning in more typical intra-domain applications, and our less typical inter-domain (RGB to HSI) fine-tuning application, and is a novel finding of this report. The effective conclusion is that while in intra-domain fine-tuning freezing the transformer blocks may help preserve model knowledge, this principle is predicated on the embedding layer remaining applicable for the new data. In our inter-domain application, this principle no longer holds as substantial modifications are required in the embedding layer, which cannot perfectly preserve the knowledge gained in the pre-training process. In instances such as these, it may therefore be more effective to perform damage control in early model training by leaving all parameters unfrozen to prevent a concentration of knowledge loss occurring in the embedding layer.

Future work within this space should focus on assessing the relative performance of these models which have been fine-tuned on HSI images against their performance on the corresponding RGB images which represent the same spatial information. Though not as easily done, it may also be interesting to assess the impact of dataset size and scope on the efficacy of the performed transfer learning, as it's possible that better performance could be extracted from a dataset which more fully represents the global space for normal and cancerous tissue. Within our dataset certain patients are overrepresented compared to others. It may help performance to utilize a balanced sampling scheme to mitigate those

effects. Further, it may be valuable to look at the impact of introducing a spectral attention mechanism to these models to see if that improves our results.

In this study, we conducted a comprehensive investigation into techniques and hyperparameters for transferring knowledge from RGB-trained histopathology foundation models to models that can process hyperspectral histological images. Specifically, we explored the efficacy of different knowledge transfer paradigms as applied to inter-domain transfer learning. These paradigms covered the common cases of training all model parameters simultaneously, as well as training only the embedding layer either for a subset or for the entirety of the training time. We perform a systematic hyperparameter search and identify effective hyperparameter combinations for each of the knowledge transfer paradigms employed and can identify the relative positive and negative impacts of learning rate, weight decay, and beta values on knowledge transfer. These collectively provide new insight into the optimal approaches to the less commonly-studied process of performing transfer learning between data domains such as preserving model performance even as the structure of the underlying data is changed.

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] Wu, Y., Cheng, M., Huang, S., Pei, Z., Zuo, Y., Liu, J., Yang, K., Zhu, Q., Zhang, J., Hong, H., Zhang, D., Huang, K., Cheng, L., and Shao, W., "Recent Advances of Deep Learning for Computational Histopathology: Principles and Applications," *Cancers*, 14, 1199 (2022).
- [2] Vergheze, G., Lennerz, J., Ruta, D., Ng, W., Thavaraj, S., Siziopikou, K., Naidoo, T., Rane, S., Salgado, R., Pinder, S., and Grigoriadis, A., "Computational pathology in cancer diagnosis, prognosis, and prediction - present day and prospects," *The Journal of pathology*, 260, (2023).
- [3] Berbís, M. Á., McClintock, D., Bychkov, A., van der Laak, J., Pantanowitz, L., Lennerz, J., Cheng, J., Delahunt, B., Egevad, L., Eloy, C., Farris, A., Fraggetta, F., Moral, R., Hartman, D., Herrmann, M., Hollemans, E., Iczkowski, K., Karsan, A., Kriegsmann, M., and Shen, J., "Computational pathology in 2030: a Delphi study forecasting the role of AI in pathology within the next decade," *eBioMedicine*, 88, 104427 (2023).
- [4] Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D., "Foundational models in medical imaging: A comprehensive survey and future vision," *arXiv preprint arXiv:2310.18689*, (2023).
- [5] Chen, R., Ding, T., Lu, M., Williamson, D., Jaume, G., Song, A., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L., Wang, J., Vaidya, A., Le, L., Gerber, G., Sahai, S., Williams, W., and Mahmood, F., "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, 30, 850-862 (2024).
- [6] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., and Poon, H., "A whole-slide foundation model for digital pathology from real-world data," *Nature*, 630, 1-8 (2024).
- [7] Lu, M., Chen, B., Williamson, D., Chen, R., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L., Gerber, G., Parwani, A., Zhang, A., and Mahmood, F., "A visual-language foundation model for computational pathology," *Nature Medicine*, 30, 863-874 (2024).
- [8] Lu, G., and Fei, B., "Medical hyperspectral imaging: a review," *J Biomed Opt*, 19(1), 10901 (2014).
- [9] Aboughaleb, I. H., Aref, M. H., and El-Sharkawy, Y. H., "Hyperspectral imaging for diagnosis and detection of ex-vivo breast cancer," *Photodiagnosis and Photodynamic Therapy*, 31, 101922 (2020).
- [10] Halicek, M., Lu, G., Little, J., Wang, X., Patel, M., Griffith, C., El-Deiry, M., Chen, A., and Fei, B., "Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging," *Journal of Biomedical Optics*, 22, (2017).

- [11] Tran, M., Bryarly, M., Ma, L., Yousuf, M. S., Price, T., and Fei, B., "Nerve Detection and Visualization Using Hyperspectral Imaging for Surgical Guidance," *Proc. SPIE 12930, Medical Imaging 2024: Clinical and Biomedical Imaging*, 129302A (2024).
- [12] Ma, L., Little, J., Chen, A., Myers, L., Sumer, B., and Fei, B., "Automatic detection of head and neck squamous cell carcinoma on histologic slides using hyperspectral microscopic imaging," *Journal of biomedical optics*, 27, (2022).
- [13] Tran, M., Ma, L., Litter, J., Chen, A., and Fei, B., "Thyroid Carcinoma Detection on Whole Histologic Slides Using Hyperspectral Imaging and Deep Learning," *Proc. SPIE 12039, Medical Imaging 2022: Digital and Computational Pathology*, 120390H (2022).
- [14] Leavesley, S. J., Walters, M., Lopez, C., Baker, T., Favreau, P. F., Rich, T. C., Rider, P. F., and Boudreaux, C. W., "Hyperspectral imaging fluorescence excitation scanning for colon cancer detection," *Journal of Biomedical Optics*, 21(10), 10 (2016).
- [15] Zhang, X., Li, W., Gao, C., Yang, Y., and Chang, K., "Hyperspectral pathology image classification using dimension-driven multi-path attention residual network," *Expert Systems with Applications*, 230, 120615 (2023).
- [16] Ortega, S., Halicek, M., Fabelo, H., Marrero Callico, G., and Fei, B., "Hyperspectral and multispectral imaging in digital and computational pathology: a systematic review [Invited]," *Biomedical Optics Express*, 11, (2020).
- [17] Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., and Mellit, A., "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustainability*, 15(7), 5930 (2023).
- [18] Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T., "Transfer learning for medical image classification: a literature review," *BMC Medical Imaging*, 22(1), 69 (2022).
- [19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint at arXiv:2010.11929* (2020).
- [20] Ma, L., Ha, A., Zainab, I., Rathgeb, A., Mubarak, H., and Fei, B., "An automatic processing framework for hyperspectral histologic images and benchmark dataset." *Proc. SPIE 13413, SPIE Medical Imaging 2025: Digital and Computational Pathology* (2025).
- [21] Halicek, M., Dormer, J. D., Little, J. V., Chen, A. Y., Myers, L., Sumer, B. D., and Fei, B., "Hyperspectral Imaging of Head and Neck Squamous Cell Carcinoma for Cancer Margin Detection in Surgical Specimens from 102 Patients Using Deep Learning," *Cancers (Basel)*, 11(9), (2019).
- [22] Ma, L., Rathgeb, A., Mubarak, H., Tran, M., and Fei, B., "Unsupervised super-resolution reconstruction of hyperspectral histology images for whole-slide imaging," *Journal of Biomedical Optics*, 27(5), 056502 (2022).
- [23] Vos, J. J., "Colorimetric and photometric properties of a 2° fundamental observer," *Color Research & Application*, 3(3), 125-128 (1978).
- [24] Frazier, P. I., "A tutorial on Bayesian optimization," *arXiv preprint arXiv:1807.02811*, (2018).
- [25] Loshchilov, I., "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, (2017).
- [26] Li, H., Ding, L., Fang, M., and Tao, D., "Revisiting Catastrophic Forgetting in Large Language Model Tuning," *arXiv preprint arXiv:2406.04836*, (2024).