2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Semi-supervised Learning for Segmentation of Gigapixel Histopathology Images from Brain Tissues

Zhengfeng Lai[1], Chao Wang[1], Zin Hu[2], Brittany N. Dugger[2],
Sen-Ching Cheung[3], *Fellow, IEEE*, and Chen-Nee Chuah[1], *Fellow, IEEE*

***Abstract*— Automated segmentation of grey matter (GM) and white matter (WM) in gigapixel histopathology images is advantageous to analyzing distributions of disease pathologies, further aiding in neuropathologic deep phenotyping. Although supervised deep learning methods have shown good performance, its requirement of a large amount of labeled data may not be cost-effective for large scale projects. In the case of GM/WM segmentation, trained experts need to carefully trace the delineation in gigapixel images. To minimize manual labeling, we consider semi-surprised learning (SSL) and deploy one state-of-the-art SSL method (FixMatch) on WSIs. Then we propose a two-stage scheme to further improve the performance of SSL: the first stage is a self-supervised module to train an encoder to learn the visual representations of unlabeled data, subsequently, this well-trained encoder will be an initialization of consistency loss-based SSL in the second stage. We test our method on Amyloid-β stained histopathology images and the results outperform FixMatch with the mean IoU score at around 2% by using 6,000 labeled tiles while over 10% by using only 600 labeled tiles from 2 WSIs.**

***Clinical relevance*— this work minimizes the required labeling efforts by trained personnel. An improved GM/WM segmentation method could further aid in the study of brain diseases, such as Alzheimer's disease.**

I. INTRODUCTION

Alzheimer's disease, the sixth leading cause of death, resulted in nearly 122,019 deaths in 2018 and the number of patients is expected to rise to 13.8 million in U.S. by mid-century [1]. To comprehensively study this disease, neuropathologists assess histopathology images to identify extracellular Amyloid-β plaques [2], which have different distributions in grey matter (GM) and white matter (WM) [3]. To determine the density and distribution of these plaques in the two regions, it is imperative to segment GM and WM in histopathology images. Many image processing-based methods have been proposed for histopathology image segmentation, such as [4], [5]. Although these methods are computationally efficient, the inter and intra-variations in staining and color contrast could significantly impair the performances of these methods on a hold-out test set [6].

Recently convolutional neural networks (CNN) have also gained wide popularity in medical segmentation problems. Among these methods, FCN [7] and U-Net [6] based architectures are the predominant choices [8], [9]. In [10], [11], they developed an automated GM and WM segmentation pipeline with promising results and compared different deep learning methods: FCN [7], U-Net [6], ResNet-Patch and ResNet-NCRF. However, these CNNs show their performance through supervised learning, which heavily relies on a large labeled dataset. For example, a recent study [12] claimed that it requires more than 30,000 labeled tiles from gigapixel WSIs to achieve the well-defined performance of CNNs, which requires labor-intensive labeling [13]. Furthermore, the labeling cost could be much higher when annotations must be done by experts (for example, doctors required for medical problems) [14]. Therefore, these challenges in procuring a sufficiently large dataset with annotations limit the wide-adoption of deep learning-based methods in real-world medical problems [15].

As such, it is vital to design an algorithm that not only automates histopathology segmentation but minimizes manual labeling. Semi-supervised learning (SSL) is one that train models without requiring heavy annotations combining a small set of labeled samples with a large amount of unlabeled samples [16]. Consistency loss-based SSL methods involve both pseudo labels and data augmentation, showing their powerful performance on CIFAR-10 [14]. One drawback of these consistency loss-based SSL methods is that the imperfect class conditional distribution is used to generate pseudo labels and the over-reliance on pseudo labels make it difficult to correctly update the class conditional distribution [17]. For example, a recent study [18] applied FixMatch [14] and Mix-Match [16] to a histology dataset and showed that the performance of these state-of-the-art SSL methods are limited due to the above drawback.

To deal with the above issue, inspired by [15] who claims that pre-training a classifier and then transferring it has the potential to outperform SSL in some settings (using 4000 labeled labeled points from CIFAR-10), we design a novel two-stage SSL, SIM-FixMatch, to further reduce the labeling cost when labeled data is too rare for transfer learning. Our first stage is to employ self-supervised learning [19] for learning visual representations, which plays a similar role to pre-training in transfer learning but requires no labels. After the first stage, a pretrained encoder will be fed into

*This work was supported by the NSF HDR:TRIPODS grant CCF-1934568 and the National Institute On Aging of the National Institutes of Health under Award Numbers P30AG010129, and AG062517, a research grant from the University of California office of the president (MRI-19-599956) and a research grant from the California Department Of Public Health (19-10611).

[1]Z. Lai, C. Wang, C.-N. Chuah are with the Department of Electrical and Computer Engineering, University of California Davis, Davis, CA 95616 USA. {lzhengfeng, ecewang, chuah}@ucdavis.edu

[2]Z. Hu, B. N. Dugger are with the Department of Pathology and Laboratory Medicine, University of California Davis, Sacramento, CA 95817 USA. {zinhu, bndugger}@ucdavis.edu

[3]S.-C. Cheung, is with the Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40506 USA. sccheung@ieee.org

the standard consistency loss-based SSL models. We employ our proposed scheme to segment GM and WM in Amyloid-β stained histopathology images. To our best knowledge, our work is the first to tackle this task with minimal labeling cost and our proposed method outperforms FixMatch [14] when the amount of labeled data is much reduced (e.g., to 0.1% of total tiles from WSIs).

II. METHODS

A. SIM-FixMatch Pipeline

In this section, we will introduce SIM-FixMatch, a two-stage SSL approach. In the first stage, we utilize the self-supervised module to pretrain an encoder that learns the visual representations from unlabeled set. Then, we use this encoder as the input into a standard consistency loss-based SSL to leverage the information from both labeled and unlabeled set. Fig. 1 shows the overall architecture.

1) **First Stage - Self-supervised Pre-training:** SimCLR [19] is a simple self-supervised framework for contrastive learning of visual representations on unlabeled images. As shown in Fig. 1, an unlabeled image undergoes two random data augmentation operations t and t' and produces outputs h_i and h_j after going through the encoder network $f(\cdot)$. $g(\cdot)$ is a projection head (multilayer perceptron with one hidden layer) to get $z_i = g(h_i)$. $f(\cdot)$ and $g(\cdot)$ are trained to maximize the agreement using the contrastive loss function

$$l_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where N is the size of a batch (two separate augmentation operators result in $2N$ data points), $\mathbb{I}_{[k \neq i]}$ is an indicator function, which is 1 if $k \neq i$, $\text{sim}(\cdot, \cdot)$ is cosine similarity, and τ is a temperature parameter.

The eventual goal for this stage is to train the encoder $f(\cdot)$ for learning visual representations from unlabeled dataset. In our experiment, we use ResNet [20] for the encoder.

2) **Second Stage - Standard FixMatch:** In this study, we mainly adopt FixMatch [14], which generates artificial labels using both pseudo-labeling and consistency regularization. Specifically, the pseudo label is generated based on a weakly-augmented unlabeled image ($weak$), which will be the target to compare with the output of the model on a strong-augmented version of the same unlabeled image ($strong$) as shown in Fig. 1. As pseudo-labels generated here could be hurt by the imperfect class conditional distribution, we use the encoder $f(\cdot)$ pretrained on unlabeled data from the first stage to provide an initialization for FixMatch. For the optimizer, instead of using standard SGD reported to have the best performance in [14], our experiments show Adam [21] performs better for our WSI dataset. The "strong" augmentation operations include RandAugment and CTAugment [14] while the "weak" includes standard flip-and-shift augmentation.

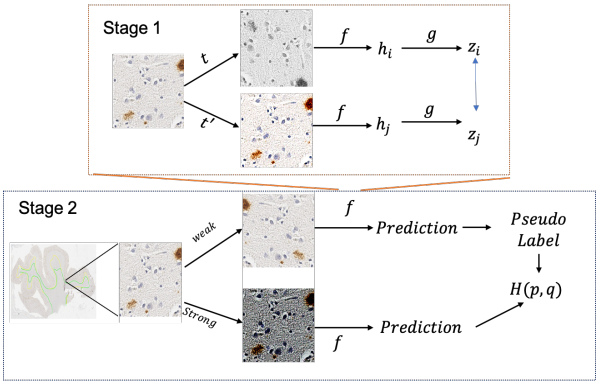


Fig. 1. Two-stage Sim-FixMatch pipeline where encoder f in the 2nd stage is self-trained using the 1st stage.

B. Datasets

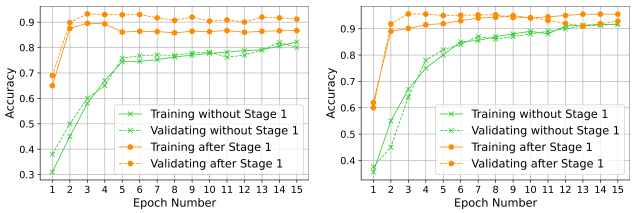
1) Overview: In this study, we utilize 30 Whole Slide Images (WSIs) of 5um formalin fixed paraffin embedded section of human temporal cortex stained with an Amyloid-β antibody (4G8, recognizing residues 17–24, dilution 1:1600, BioLegend (formally Covance) catalog number SIG-39200). These slides were scanned and digitized with an Aperio AT2 at up to $20\times$ magnification, resulting in the average resolution at nearly $60,000 \times 50,000$ pixels each. Among these 30 WSIs, 18 slides (from 10 males and 8 females with an average age at death of 84 ± 7 years) from deceased patients pathologically diagnosed as Alzheimer's disease, and will be referred to as AD cases; the remaining 12 slides lacked a pathological diagnosis of Alzheimer's disease, referred as NAD cases. Among these 12 NAD cases, one had a diagnosis of metastatic carcinoma, and five with cerebrovascular disease. The Ethnoracial make up of the cohort was 22 non-Hispanic White (73%) descendants, 5 African Americans (17%), and 3 Hispanics (10%). To further protect data confidentiality, we refer to the AD cases as WSI-1 to WSI-18 and NAD cases as WSI-19 to WSI-30.

2) Training Data Preparation: As downsampling [13] may lose medical features, we follow a patch-based method in [11] to divide WSIs into 256×256 patches to cope with the ultra-high resolution. In this paper, 20 WSIs (12 AD cases and 8 NAD cases) were randomly selected for training and validation while the remaining 10 WSIs (6 AD cases and 4 NAD cases) were used for hold-out testing and inference. From the 20 WSIs, we selected one AD case and NAD case that have highest inter-rater agreement as the source of labeled patches while we kept the remaining 18 WSIs for generating unlabeled patches (around 600,000 patches). In our setting, we first generated 6,000 labeled patches (nearly 1% proportion of all patches) from 2 labeled WSIs, then we generated 6,00 labeled patches (nearly 0.1%) to further verify the effectiveness of our proposed method.

III. RESULTS

A. Ablation Study

To verify the effectiveness of our first stage, which is to learn visual representations on the unlabeled set and provide an encoder for the second stage, we visualized the training process of the second stage by using our method and baseline FixMatch. We trained both of them over 40 epochs and found



(a) 600 Labeled Patches (b) 6000 Labeled Patches

Fig. 2. Trends of training and validation during the training process of FixMatch with or without the 1st stage.

that they converged quickly within 15 epochs. As shown in Fig. 2, with the first stage, the model will be well-trained after only 3 epochs in the second stage, while it takes almost 20 epochs without the first stage. Besides, our proposed method starts from nearly 50% higher accuracy after the first epoch compared with original FixMatch, which shows the effectiveness of our proposed first stage on learning the representations via contrastive learning.

B. Quantitative Results

TABLE I

PIXEL-WISE IOU SCORES FOR AD, NAD, AND OVERALL TEST SET

	2 Labeled WSIs		0.1% Labeled		1% Labeled	
	FCN	U-Net	FixMatch	**Proposed**	FixMatch	**Proposed**
AD Back	61.04	59.74	93.15	**94.10**	**96.59**	96.33
± STD	± 5.44	± 13.9	± 2.41	± 2.21	± 1.04	± 1.05
AD GM	46.98	37.16	78.57	**84.59**	87.12	**88.21**
± STD	± 2.78	± 9.93	± 3.87	± 3.27	± 3.46	± 3.84
AD WM	27.75	7.57	56.66	**74.31**	73.94	**76.33**
± STD	± 5.50	± 6.02	± 16.4	± 3.36	± 6.89	± 4.77
AD Mean	45.26	35.40	76.13	**84.34**	85.88	**86.95**
± STD	± 3.55	± 7.12	± 5.89	± 1.88	± 3.09	± 2.72
NAD Back	66.66	78.46	**97.07**	96.70	**97.71**	97.63
± STD	± 5.17	± 18.5	± 0.31	± 0.73	± 0.86	± 0.88
NAD GM	50.15	59.59	83.97	**86.58**	90.01	**90.93**
± STD	± 0.49	± 13.6	± 7.76	± 5.44	± 4.02	± 4.12
NAD WM	19.72	3.02	22.72	**62.17**	59.71	**68.79**
± STD	± 13.6	± 3.09	± 19.0	± 7.04	± 9.99	± 7.22
NAD Mean	45.51	47.02	67.92	**81.82**	82.47	**85.78**
± STD	± 3.29	± 10.9	± 6.53	± 2.15	± 2.59	± 2.00
Test Back	63.29	68.28	94.72	**95.14**	**97.04**	96.85
± STD	± 5.81	± 17.2	± 2.71	± 2.17	± 1.08	± 1.15
Test GM	48.25	46.13	80.73	**85.39**	88.27	**89.30**
± STD	± 2.66	± 15.8	± 6.01	± 4.10	± 3.78	± 3.98
Test WM	24.54	5.75	43.08	**69.45**	68.25	**73.31**
± STD	± 9.80	± 5.37	± 24.0	± 7.88	± 10.7	± 6.72
Test Mean	45.36	40.05	72.84	**83.33**	84.52	**86.48**
± STD	± 3.26	± 10.2	± 7.18	± 2.28	± 3.26	± 2.41

AD is the average results on the 6 Alzheimer's disease cases in hold-out test set. NAD is the average results on the 4 non-Alzheimer's disease cases in test set. Test is the average results on all 10 WSIs.

IoU and STD. We first use a standard segmentation metric — Intersection over Union (IoU) to compare the masks from our proposed method and original FixMatch. IoU score is designed for measuring the overlapping degree between two masks. And we also use standard deviation (STD) to evaluate how consistent and robust of our methods across different hold-out test slides. The results of both IoU scores and STD are summarized in Table. I.

We selected the most updated version of FCN [7] and U-Net [6] as the supervised learning (SL) baselines for our comparison. Both of them are trained on only 2 labeled slides (1 AD case + 1 NAD case). Compared to the results reported in [10] that are trained on 20 labeled slides, their performance drastically deteriorates with reduced labeled WSIs from 20 to 2. The mean IoU scores for these two methods are only around 40%. FixMatch and our proposed method are trained on labeled patches (600 and 6000) from the same 2 labeled WSIs while the unlabeled patches are from other 18 unlabeled WSIs. For 6000 labeled patches setting, the labeled ratio is only 1%. FixMatch could achieve 84.52% of mean IoU while our proposed SIM-FixMatch is around 2% higher and has better performance in almost all classes, especially for the WM region in NAD cases (9.08% of improvement). Besides, our proposed method achieves 2.28% lower in terms of STD compared to original FixMatch. To further stress-test of our proposed method, we consider an extreme situation by using only 600 labeled patches (the labeled ratio is down to 0.1%). The improvement of SIM-FixMatch is significant, almost 40% of increase in terms of the WM region in NAD cases and 10.49% of increase in the mean IoU while the STD is still close to original FixMatch.

DICE coefficient. Besides IoU, we also use DICE coefficient [22] to further evaluate the proposed methods. When only 600 patches are labeled, the DICE coefficient of WM increases from 61.32% (FixMatch) to 82.67% (proposed). And when 6000 patches are labeled, it also gains 3% of improvement in WM.

C. Segmentation Visualization

Fig. 3 shows the segmentation visualization of SL methods (FCN [7], U-Net [6]) trained on the same 2 labeled WSIs and SSL methods (original FixMatch [14] and our proposed method) trained using 600 labeled patches from the same two labeled slides and unlabeled patches from the other 18 WSIs (the labeled ratio is only 0.1%). The masks of U-Net (Fig. 3 the 2nd column) indicates that U-Net is unable to distinguish the WM from the GM; the masks of FCN (Fig. 3 the 3rd column) have better visualization than U-Net but there are still many incorrectly labeled regions. FixMatch (Fig. 3 the 4th column) is able to find the rough boundary between GM and WM but there are noisy pixels within WM, indicating it wrongly predicts some WM pixels as GM in the WM region. Our proposed method (Fig. 3 the 5th column) could provide more distinguishable boundary for each region and the masks are the closest to the ground truth masks.

IV. CONCLUSIONS

In this paper, we investigate the applicability of state-of-the-art semi-supervised learning in histology images and propose a two-stage approach to further improve the performance of SSL methods on Amyloid-β stained WSIs at gigapixel level with the minimal labeling efforts. In our two-stage method, we verify the effectiveness of the first stage (self-supervised pretraining) by providing an encoder that has learned adequate visual representations among unlabeled data. Our proposed method outperformed the original FixMatch, especially in the case where labeled tiles are

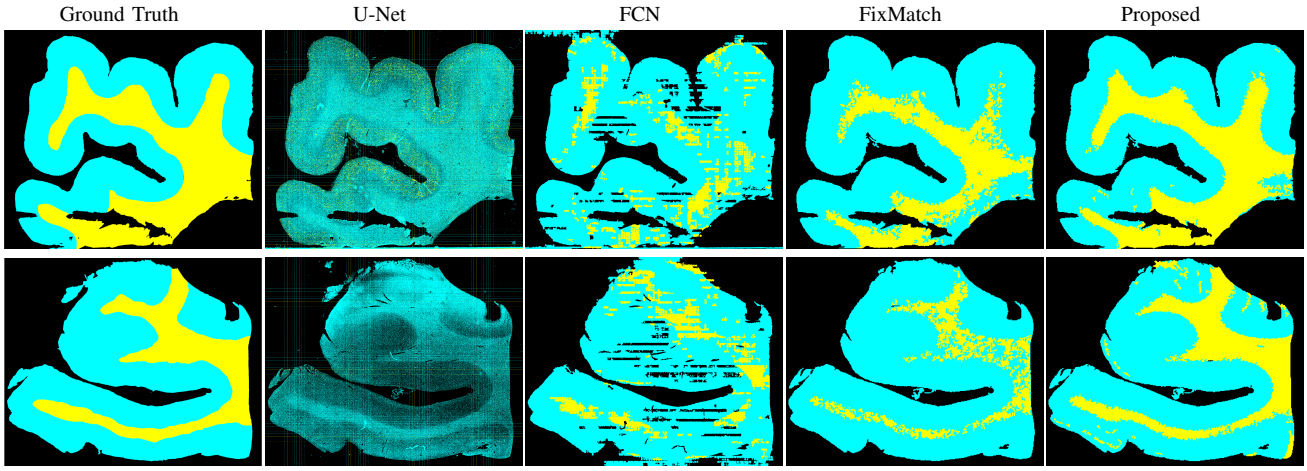


Fig. 3. Segmentation masks visualization by only using 0.1% labeled patches from one AD case (top) and one NAD case (bottom). Both are from hold-out test set. Here GM, WM, and background are indicated by cyan, yellow, and black, respectively.

extremely rare (0.1%). While we showed promising results by running our experiments using randomly selected two WSIs, we will evaluate the selection criteria of WSIs more systematically in our future work.

These techniques have the potential to be applied to other classification and segmentation problems in medical images to minimize the expensive labeling cost. In addition, it takes nearly 3 days for SimCLR to train a good representation. Consequently, our future direction involves developing a task-based architecture to accelerate this process.

ACKNOWLEDGMENT

The authors would like to thank the families and participants of the University of California, Davis Alzheimer's Disease Research Center (UCD-ADRC) for their generous donations as well as the commitments of faculty and staff of the UCD-ADRC.

REFERENCES

[1] A. Association *et al.*, "2020 Alzheimer's Disease facts and figures," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 391–460, 2020.

[2] B. N. Dugger and D. W. Dickson, "Pathology of neurodegenerative diseases," *Cold Spring Harbor perspectives in biology*, vol. 9, no. 7, p. a028035, 2017.

[3] N. Iwamoto, E. Nishiyama, J. Ohwada, and H. Arai, "Distribution of amyloid deposits in the cerebral white matter of the alzheimer's disease brain: relationship to blood vessels," *Acta Neuropathol.*, vol. 93, no. 4, pp. 334–340, 1997.

[4] P. Kłeczek, G. Dyduch, J. Jaworek-Korjakowska, and R. Tadeusiewicz, "Automated epidermis segmentation in histopathological images of human skin stained with hematoxylin and eosin," in *Medical Imaging 2017: Digital Pathology*, vol. 10140. ISOP, 2017, p. 101400M.

[5] D. Bug, F. Feuerhake, and D. Merhof, "Foreground extraction for histopathological whole slide imaging," in *Bildverarbeitung für die Medizin 2015*. Springer, 2015, pp. 419–424.

[6] K. Oskal, M. Risdal, E. Janssen, E. Undersrud, and T. Gulsrud, "A U-net based approach to epidermal tissue segmentation in whole slide histopathological images," *SN Appl. Sci.*, vol. 1, p. 672, 06 2019.

[7] P. Bándi, R. van de Loo, M. Intezar, D. Geijs, F. Ciompi, B. van Ginneken, J. van der Laak, and G. Litjens, "Comparison of different methods for tissue segmentation in histopathological whole-slide images," in *IEEE ISBI 2017*, pp. 591–595.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[10] Z. Lai, R. Guo, W. Xu, Z. Hu, K. Mifflin, C. DeCarli, B. N. Dugger, S.-c. Cheung, and C.-N. Chuah, "Automated segmentation of amyloid-β stained whole slide images of brain tissue," *bioRxiv 2020.11.13.381871*, 2020.

[11] Z. Lai, R. Guo, W. Xu, Z. Hu, K. Mifflin, B. N. Dugger, C.-N. Chuah, and S.-c. Cheung, "Automated grey and white matter segmentation in digitized aβ human brain tissue slide images," in *2020 ICMEW*. IEEE, 2020, pp. 1–6.

[12] Z. Tang, K. V. Chuang, C. DeCarli, L.-W. Jin, L. Beckett, M. J. Keiser, and B. N. Dugger, "Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline," *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[13] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *IEEE CVPR*, 2019, pp. 8924–8933.

[14] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv:2001.07685*, 2020.

[15] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in Neural Information Processing Systems*, vol. 31, pp. 3235–3246, 2018.

[16] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *arXiv:1905.02249*, 2019.

[17] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv:2101.06329*, 2021.

[18] J. V. Pulido, S. Guleria, L. Ehsan, M. Fasullo, R. Lippman, P. Mutha, T. Shah, S. Syed, and D. E. Brown, "Semi-supervised classification of noisy, gigapixel histology images," in *2020 BIBE*. IEEE, 2020, pp. 563–568.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[22] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells III, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports," *Academic radiology*, vol. 11, no. 2, pp. 178–189, 2004.