To pretrain or not to pretrain? A case study of domain-specific pretraining for semantic segmentation in histopathology

Tushar Kataria^{1,2}, Beatrice Knudsen³, and Shireen Elhabian^{1,2}

Kahlert School of Computing, University Of Utah
 Scientific Computing and Imaging Institute, University of Utah

³ Department of Pathology, University of Utah

{tushar.kataria,shireen}@sci.utah.edu, beatrice.knudsen@path.utah.edu

Abstract. Annotating medical imaging datasets is costly, so fine-tuning (or transfer learning) is the most effective method for digital pathology vision applications such as disease classification and semantic segmentation. However, due to texture bias in models trained on real-world images, transfer learning for histopathology applications might result in underperforming models, which necessitates the need for using unlabeled histopathology data and self-supervised methods to discover domainspecific characteristics. Here, we tested the premise that histopathologyspecific pretrained models provide better initializations for pathology vision tasks, i.e., gland and cell segmentation. In this study, we compare the performance of gland and cell segmentation tasks with domain-specific and non-domain-specific pretrained weights. Moreover, we investigate the data size at which domain-specific pretraining produces a statistically significant difference in performance. In addition, we investigated whether domain-specific initialization improves the effectiveness of out-of-domain testing on distinct datasets but the same task. The results indicate that performance gain using domain-specific pretraining depends on both the task and the size of the training dataset. In instances with limited dataset sizes, a significant improvement in gland segmentation performance was also observed, whereas models trained on cell segmentation datasets exhibit no improvement.

Keywords: Domain Specific pretraining \cdot Gland and Cell Segmentation \cdot Transfer Learning.

1 Introduction

Deep learning models typically require a substantial amount of data to effectively learn generalized latent space representations [4]. However, acquiring large medical image datasets is more challenging compared to real-world image datasets for three primary reasons. Firstly, the annotation process for medical images involves domain-specific knowledge from pathologists [16,11,19] and radiologists to manually outline anatomical structures. This is challenging given the global scarcity

2 T. Kataria et al.

of pathology and radiology experts; Secondly, the image annotation interfaces are inefficient generating labor-intensive workflows. Thirdly, inter-observer disagreement among medical professionals necessitates the involvement of multiple experts to repeat each annotation task [5]. Lastly, in addition to the annotation challenges there are biases in medical data. Biases in histopathology images arise from variations in tissue quality, staining protocols leading to difference in color and texture [10], scanning protocols and slide scanners [10,14]. These biases are often site-specific and can cause major domain shifts between different data sets, which in term reduces the generalization of deep learning models. [10,14]. Other forms of domain shifts in cancer cohorts include discrepancies between cancer and normal tissue histology, the proportion of histologic cancer subtypes, grades and stages, and variations in clinical, demographic, and race-related variables. These variables generate data imbalances that can degrade the performance of deep learning models during testing.



Fig. 1. Do different weight initialization matter? The study is designed from the perspective of an AI user who can choose between multiple pretrained model options for a given task. The best pretrained model is the one that is least effected domain shift. This study provides a framework to choose amongst pretrained models and select the most advantageous for the task.

In medical image vision tasks, fine-tuning pretrained models (also known as transfer learning) has become a common approach [6,15]. These tasks are important for automated diagnosis, cancer grading and predictions of patients outcomes across all cancer types. Using supervised or self-supervised methods, deep learning models exhibit strong capabilities to learn effective latent representations [3]. However, they may suffer from domain-specific texture bias [9], which can impede their performance [18]. Previous research indicates that if sufficient data is available for training, a model trained de-novo (i.e., from scratch) may outperform a fine-tuned model [15,18]. This suggests a potential benefit of domain-specific pretraining [24,12] over transfer learning from ImageNet [8].

Because large, annotated data sets are difficult to obtain for pretraining on histopathology images, self-supervised and annotation free methods (SSL) provide an alternative strategy for pretraining models to learn valid representation in the latent space [2,23,1]. Models can then be further fine-tuned with a few annotations to produce acceptable results on test datasets. However, no studies systematically evaluated the impact of domain-specific pretaining for histopathology models that are tasked to learn cell and gland segmentation. The closest matching work to this study is an investigation of pretraining on classification and instance segmentation tasks [12]

Because gland and cell segmentation differ from instance segmentation and classification, the effect of pretraining on the analysis of out-of-distribution (OOD) datasets also remains unknown. The contributions of this paper are as follows:-

- Comparison of de-novo trained models with pretrained models on the ImageNet dataset using class supervision [8] and self-supervision [23] for semantic segmentation tasks in histopathology.
- Finetuning pretrained domain-specific models [12] for gland and cell segmentation. These comparisons will indicate whether domain-specific pretraining aids cell and gland segmentation in out-of-distribution data sets after finetuning of models.
- Determining the effect of compute resources and data quantity on model performance improvements.
- Investigating whether domain-specific training leads to a better generalization of models.

$\mathbf{2}$ **Different pretraining Strategies**

To investigate whether domain-specific pretraining leads to generalization in gland and cell segmentation tasks, the study aims to address the following research questions:

- Is domain pretraining, which involves initializing the weights with domainspecific images, more effective for transfer learning compared to pretrained weights from ImageNet?
- Do self-supervised outperform supervised weight initializations?
- Does domain-specific pretraining enhance the quality of features and improve the model's performance on datasets with domain shifts?

All initializations are compared against random initialization (i.e., training from scratch), which serves as the baseline to identify initializations (mentioned below) that outperform random. The flow diagram of the study is shown in Figure 1.

Models are trained with 3 different types of initializations: (1) pretrained weights using class supervision on ImageNet data: default weights are provided in Pytorch for ImageNetV1 and ImageNetV2. The top-1 accuracies in the initialization amount to 76.13 and 80.85, respectively. These weights are obtained by training a ResNet50 [8] model with class supervision. For two other initialization, weights are obtained using a self-supervised technique called Barlow Twins [23]. (2) Pretrained weights with ImageNet data using SSL 4 T. Kataria et al.

(SSLImage): Self-supervised weights were obtained after training on data from ImageNet without using labels. (3) **Domain-Specific pretraining using SSL** (SSLPathology): This model is released as part of the study in [12] for domainspecific pretraining on histopathology data. The model was pretrained using more than three million histopathology image patches sampled from various cancers at different magnifications. More details about the pretraining method and the dataset can be found in [12].

2.1 Dataset Details

We have experimented with gland and cell segmentation tasks on these five histopathology datasets:

Gland Segmentation Datasets: Colon cancer datasets, GlaS and CRAG [21,7], possess ground truth gland segmentation annotations for normal and colon cancer glands. The GlaS dataset has 88 training & 80 testing images of size less 700x600 pixels, whereas the CRAG dataset has 160 training & 40 testing images of size 1512x1512 pixels.

Cell Segmentation Datasets: Three cell segmentation datasets are used for experimentation KUMAR [13], CPM17 [22] and TNBC [17] possess ground truth annotations of nuclear outlines.

 Table 1. Cell Segmentation Dataset Details. Sample examples of the dataset are shown in supplementary Figure 9.

Datasets	Train Imgs	Test Imgs	Img Size	No. of Annotated Nuclei
KUMAR [13]	16	14	1000×1000	21623
CPM17 [22]	31	31	500×500	7570
TNBC [17]	34	16	512x512	4022

2.2 Implementation Details

A U-Net [20] model is used with Resnet50 [8] backbone for semantic segmentation application(gland & cell both). The decoder is always the same for all models. Models are trained using PyTorch and a data split of 80-20 for training and validation. The best model possessing a minimum loss on validation data is further evaluated on the test dataset. Testing data is only used for inference.

During training the patch size is 256x256, sampled randomly in the whole image. At inference, predictions are averaged over a window size of 128 pixels. The learning rate is fixed to 0.0001 and the number of epochs for all experiments is set to 4000 for gland segmentation and 2000 for cell segmentation. The models are trained five times and average metrics are reported, this ensures that variations due to stochasticity caused by the dataset loader are factored out. Data augmentation includes horizontal and vertical flips, random rotation, and translation. All models are trained on NVIDIA V100 GPUs.

Evaluation Metrics: Dice and Jaccard scores (also known as the intersection over union) serve as metrics for segmentation tasks [21,20].

3 Results

Gland Segmentation Results: The line plots (variation marked as shading) of performance measures for different initialization are shown in Figure 2,6-A⁴. We trained models with different backbone initializations on an increasing amount of data. The following observations emerged from these experiments:- (a) Increasing the quantity of data improves performance for all initializations and decreases variation. (b) At all levels of target domain training data, models with pretrained weight initializations outperform those with random initializations, but the performance gap between random initialization and pretraining decreases as the quantity of data increases. (c) For small datasets, domain-specific pretraining has a significant performance advantage over other initializations. However, as the size of the dataset grows, the effect of domain-specific pretraining diminishes.



Fig. 2. Gland Segmentation Results for Different Initializations on GlaS[21]. (A) Dice and Jaccard Score for different percentage of training data used. We can clearly observe that increasing data increases model performance, but with more data domain specific pretraining doesn't have a significant effect on performance. (B) Average dice score variations with different amounts of training time, i.e., number of epochs. We clearly see a difference in performance for different initialization for low dataset size and lesser epochs. Results on CRAG dataset are shown in Supplementary Figure 6.

Variation in performance due to different amounts of training epochs for all datasets is shown in Figure 2,6-B. For very small datasets(10% and 30% graph), domain-specific pretraining outperforms all other initializations at all epochs. However, for larger datasets(100% data), ImageNet supervised weights also outperform at lower epochs as well. This show that domain-specific pretraining is

⁴ All images are best viewed on a digital device following magnification.

6 T. Kataria et al.

dataset diversity dependent and not computational power. If a dataset is not diverse or small in size, then domain-specific pretraining is beneficial, but other initialization can be better for higher diversity and higher epochs. Qualitative results are shown in supplementary Figure 8, domain-specific fine-tuned models have more accurate gland outlines and fewer false positive pixels than other models.

Cell Segmentation Results : The performance of various initializations is depicted in Figure 3. Even though some of the observations are similar to those of previous experiments, novel observations emerge from cell segmentation results:- (a) Model performances with KUMAR [13] data are an exception where random initialization is outperforming or competitive with other initializations. (b) Domain-specific pretraining is performing similar to or worse than ImageNet initialization for most cases. Altogether our results demonstrate that domain-specific pretraining does not improve the performance of the U-Net/ResNet model for cell segmentation tasks. Qualitative results are shown in supplementary Figure 9.



Fig. 3. Cell Segmentation Results. Different initialization has similar patterns, i.e., with increasing data variation in performance decreases and mean performance increases. But for cell segmentation domain specific pretraining doesn't seem to be better than image-net pretrained weights for different data sizes.

UMAP Results: We sampled 300 random patches from the test sets of GlaS and CRAG to generate projections for encoders and decoders shown in Figure 4. Feature values were extracted from the first encoder layer in U-Net, the deepest encoder layer, and the last decoder layer.

In the network's first layer, the projections of features from various initializations form clouds that overlap. We interpret this observation to conclude that the initial layers of deep neural networks capture low-level statistics and that all initializations capture comparable attributes. As encoding depth increases, the representations become more distinct and the overlap decreases, indicating that networks pretrained in different ways may be learning different representations of the same data. This is counterintuitive, as we would expect that each of the pretrained models generates similar high-level representations when performing identical tasks and using the same dataset. However, the distribution of features in the UMAP projection of latent layer representations appears to have topological similarity across initializations which indicates that features for different initialization may be related via a rigid transformation in latent space. A similar conclusion is valid for the decoder UMAP. Together, these results suggest that distinct initializations, despite being clustered at different locations in the UMAP, might learn similar relational feature characteristics between samples in the dataset.



Fig. 4. UMAP for different Gland Segmentation models for GlaS and CRAG datasets. We generate UMAP with nearest neighbor=25, distance=0.1, and metric=cosine. Comparing the latent representation of the initial encoder to that of the deepest encoder, there is substantially less overlap between initializations, but the distribution of points is topologically similar.

3.1 Out Of Domain Testing Results

For OOD testing we use the pretrained models for out-of-box (without finetuning) testing on other datasets. This analysis reveals the bias to the domain that learned with various initializations.

Gland Segmentation Results The results for OOD testing for the gland segmentation task are shown in 5. At a low amount of data, the domain-specific, finetuned models perform best and using random initializations results in the greatest relative performance drop compared to all other initializations.

Cell Segmentation Results: The results of OOD testing for different datasets are shown in supplementary Figure 7 and lead to the following observations: (a) pretrained models are better than models with random initialization at the same task on unseen datasets from KUMAR [13] and CPM17 [22]). In contrast, models with random initialization and trained on TNBC [17] outperform or



Fig. 5. Average dice score for OOD testing. Y-axis shows the dataset used for training of the model. X-axis is the performance on the corresponding test sets without any fine-tuning. A model trained on CRAG datasets transfers effectively to GlaS, but not vice versa. Domain-specific pretrained models are generally better at out-of-domain performance.

perform the same as the pretrained initialized model. (b) A drop in performance exists on TNBC data for models trained on KUMAR [13] and CPM17 [22] but not for models trained on TNBC [17] or KUMAR [13] and applied to CPM17. (c) Domain-specific pretrained models when tested on OODdata demonstrate a lesser drop in performance compared to other pretraining approaches.

4 Conclusion and Future Work

In this study, we demonstrate that a domain-specific pretraining backbone can be beneficial for gland and cell segmentation when data are limited or of low diversity data for the task at hand. However, the need for domain-specific pretraining decreased for gland and cell segmentation as the amount of training data increases. The results of cell segmentation indicate that domain-specific pretraining may not be advantageous for all types of tasks. The results of UMAP projections indicate that the initial layers of domain-specific and non-domainspecific models learn similar features, but that the deeper encoders are distinct. Although the topology of latent feature representations is similar for the different initialization, models may be learning similar high-level characteristics within the latent feature spaces. Lastly, during out-of-distribution testing, domain-specific pretraining suffers the same performance degradation as other initializations, i.e. domain-specific pretrained models may not be effective at learning site-independent features. Our final conclusion from this study is that domain-specific pretraining may be beneficial for specific tasks and datasets, but benefits are not universal. Domain-specific pretraining suffers from the same issues as pretraining on image-net. Lastly, we would like to make the reader aware that this study did not cover medical vision tasks such as multi-class semantic segmentation and cell detection. We also did not utilize models pretrained using vision-language models. Both these comparisons are left for future work.

References

- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems 33, 9912–9924 (2020)
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems 33, 22243–22255 (2020)
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A largescale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
- Eaden, J., Abrams, K., McKay, H., Denley, H., Mayberry, J.: Inter-observer variation between general and specialist gastrointestinal pathologists when grading dysplasia in ulcerative colitis. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland 194(2), 152–157 (2001)
- Erhan, D., Courville, A., Bengio, Y., Vincent, P.: Why does unsupervised pretraining help deep learning? In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 201–208. JMLR Workshop and Conference Proceedings (2010)
- Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.A., Snead, D., Tsang, Y.W., Rajpoot, N.: Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Medical image analysis 52, 199–211 (2019)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
- Hermann, K., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. Advances in Neural Information Processing Systems 33, 19000–19015 (2020)
- Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature communications 12(1), 1–13 (2021)
- Jajosky, R.P., Jajosky, A.N., Kleven, D.T., Singh, G.: Fewer seniors from united states allopathic medical schools are filling pathology residency positions in the main residency match, 2008-2017. Human Pathology 73, 26–32 (2018)
- Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
- Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging 36(7), 1550–1560 (2017)
- Liu, Q., Dou, Q., Yu, L., Heng, P.A.: Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE transactions on medical imaging 39(9), 2713–2724 (2020)

- 10 T. Kataria et al.
- Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., Ferrari, V.: Factors of influence for transfer learning across diverse appearance domains and task types. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(12), 9298– 9314 (2021)
- Metter, D.M., Colgan, T.J., Leung, S.T., Timmons, C.F., Park, J.Y.: Trends in the us and canadian pathologist workforces from 2007 to 2017. JAMA network open 2(5), e194337–e194337 (2019)
- Naylor, P., Laé, M., Reyal, F., Walter, T.: Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE transactions on medical imaging 38(2), 448–459 (2018)
- Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems 32 (2019)
- Robboy, S.J., Gross, D., Park, J.Y., Kittrie, E., Crawford, J.M., Johnson, R.L., Cohen, M.B., Karcher, D.S., Hoffman, R.D., Smith, A.T., et al.: Reevaluation of the us pathologist workforce size. JAMA network open 3(7), e2010648–e2010648 (2020)
- Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
- Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis 35, 489–502 (2017)
- Vu, Q.D., Graham, S., Kurc, T., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., et al.: Methods for segmentation and classification of digital microscopy tissue images. Frontiers in bioengineering and biotechnology p. 53 (2019)
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
- Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)



5 Supplementary

Fig. 6. Gland Segmentation Results for Different Initializations on CRAG[7]. (A) Dice and Jaccard Score for different percentage of training data used. (B) Average dice score variations with different amounts of training time, i.e., number of epochs.



Fig. 7. Average dice score for OOD testing. Y-axis shows the dataset used for training the model. X-axis is the performance on the test set of the datasets without any fine-tuning.



Fig. 8. Qualitative results for Gland Segmentation Experiments. We can observe that pretrained models have better qualitative results than Random Initializations. Domain Specific pretraining models perform better for gland segmentation tasks. These models are better at recognizing the outlines of the gland compared to other initializations.



Fig. 9. Qualitative results for Cell Segmentation Experiments. We can observe that pretrained models have better qualitative results than Random Initializations. But all of the pretrained models make similar mistakes in the outlines of cells, touching cells and distinguishing between cell and background.