Few-shot weakly supervised detection and retrieval in histopathology whole-slide images

Mart van Rijthoven[a], Maschenka Balkenhol[a], Manfredo Atzori[c], Peter Bult[a], Jeroen van der Laak[a, b], and Francesco Ciompi[a]

[a]Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
[b]Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden
[c]Institute of Information Systems, HES-SO (University of Applied Sciences of Western Switzerland), Sierre, Switzerland

ABSTRACT

In this work, we propose a deep learning system for weakly supervised object detection in digital pathology whole slide images. We designed the system to be organ- and object-agnostic, and to be adapted *on-the-fly* to detect novel objects based on a few examples provided by the user. We tested our method on detection of healthy glands in colon biopsies and ductal carcinoma in situ (DCIS) of the breast, showing that (1) the same system is capable of adapting to detect requested objects with high accuracy, namely 87% accuracy assessed on 582 detections in colon tissue, and 93% accuracy assessed on 163 DCIS detections in breast tissue; (2) in some settings, the system is capable of retrieving similar cases with little to none false positives (i.e., precision equal to 1.00); (3) the performance of the system can benefit from previously detected objects with high confidence that can be reused in new searches in an iterative fashion.

Keywords: Detection, Few-Shot, Prototypes, Proposals, Retrieval, Computational Pathology

1. INTRODUCTION

Digital pathology has enabled access to a constantly growing number of histopathology slides in the form of digital whole-slide images (WSI). At the same time, advances in machine learning and deep learning approaches such as convolutional neural networks (CNN), have allowed the automated analysis of WSIs to address tasks such as image classification and segmentation, as well as detection of objects of interest. Training CNNs has so far mostly been done via supervised learning using annotations made by human experts, such as pathologists. For example, to build a system for the detection of colon glands, many visual examples have to be annotated manually, either by drawing the contour of the gland (for segmentation purposes) or by delineating a bounding box around it (for detection purposes).

Creating manual annotations is a time-consuming process that often requires multiple experts to obtain a reliable ground truth, thereby making it very expensive and often unfeasible to obtain a large number of manually annotated WSIs. This limits the amount of data that is available for training. For this reason, alternatives to full supervision have gained popularity in the field of medical imaging, including digital pathology. Examples are CNN model training without manual annotations, i.e., in an unsupervised or self-supervised fashion,[1] or by only using sparse annotations,[2] or only few labels[3] (e.g., weakly supervised learning) during model training. A very recent method used only a few examples for training, and predictions for novel examples were used as pseudo labels whenever the network's confidence for this novel example was high enough.[4] In this way, the model automatically increases the size of the training set, without requiring additional manual annotations. Alternatively, forms of weak supervision can be provided after a CNN is trained, for example, by exposing the model to a single or a few novel target examples, not seen during training, provided by the user at test time. In this scenario, ranking mechanisms have been used in content-based retrieval systems to retrieve regions of interest from an image database that are similar to a user-provided example patch at test time.[5] Although

Further author information: Mart van Rijthoven; E-mail: mart.vanrijthoven@radboudumc.nl

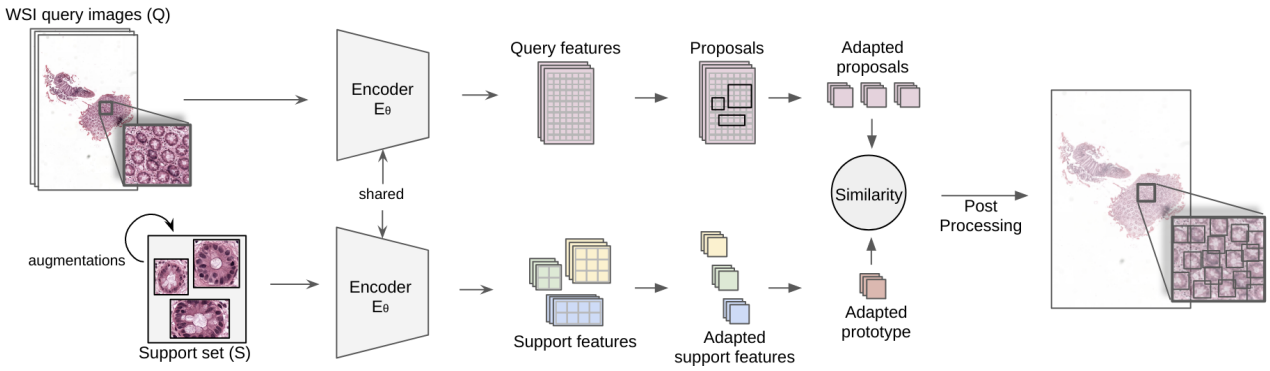

Figure 1: An overview of the proposed system which shows a query pathway (top) and a support pathway (bottom). A convolutional neural network E_θ, encoded both the query and the support examples, resulting in query and support features. Rectangular regions over the query features considered detection proposals, which were adapted to a set of same-sized proposals. The support features were adapted into same sized support features and were combined to create a single adapted prototype. Each adapted proposal was compared to the adapted prototype via a similarity function, and a threshold regulated the detection and retrieval results. Additional post-processing techniques determined the final detection output.

effective, retrieval mechanisms do not guarantee the detection of a specific object, but rather focus on identifying the local field of view, most similar to the requested patch.

Another branch of machine learning that aims to learn from a few examples at test time is *few-shot learning*. In this paper, we focus on a specific approach to few-shot learning based on prototypical networks,[3] which uses *support sets* consisting of a few novel examples (i.e., "few shots") to classify or detect unseen classes that were not in the training set. However, copious annotations of 'base' classes are needed to train those networks, and typically achieve high effectiveness in a multi-class classification setting, which prohibits the use of such a model to detect a single object of interest, as desired in weakly supervised object detection instead.

In this paper, we propose a novel framework for weakly supervised object detection to detect objects and tissue components within a WSI given only a few examples at test time. The proposed method builds upon and combines different technologies such as a) few-shot learning, b) image retrieval, and c) modern object detection approaches such as YOLO.[6] We show that the proposed system can process an entire WSI in about five seconds, and detect objects and tissue components based on only a few novel examples without the need for retraining. We report results on both geometrically simple objects such as colon glands and complex objects such as ductal carcinoma in situ of the breast. Furthermore, we show that this system can be used for image retrieval with high precision.

2. METHOD

The proposed system is schematized in Figure 1. Following the terminology used in few-shot learning, we refer to the collection of WSIs available to the system to detect objects of interest as the *query set* Q, and to the few visual examples provided by the user as the *support set* S. Based on these two concepts, we described the method referring to the query pathway and the support pathway.

2.1 The query pathway

A query, $q \in Q$, was encoded via an encoder E_θ, namely a CNN that outputs an embedding representation with C features. To encode an entire WSI with dimension $\mathbb{R}^{W \times H \times 3}$, we first divided the WSI into a set of non-overlapping patches of size $\mathbb{R}^{R \times R \times 3}$ and then used E_θ to encode each patch into an embedding C, as proposed in Tellez et al.[7] Based on R, we defined a grid G with dimensions $\mathbb{R}^{M \times N \times C}$ over the encoded WSI, where $M = \frac{W}{R}$ and $N = \frac{H}{R}$. On each grid cell, $(m, n) \in G$, rectangles with shapes based on the sizes of the support examples, covering one or more grid cells, were used to extract proposals corresponding to an embedded region.

Subsequently, each proposal was adapted by taking the mean of the embedded region, resulting in a proposal vector of size C:

$$adapted\,proposal = \frac{\sum_{m_a}^{m_b}\sum_{n_a}^{n_b} G(m_a, n_a)}{(m_b - m_a)(n_b - n_a)},$$

where m_a, m_b, n_a, n_b took on values based on the sizes of the support examples in S.

2.2 The support pathway

The same encoder E_θ used in the query pathway, encoded a support example, $s \in S$, into a feature map embedding, similar as was done for a WSI (i.e., with stride R). Therefore, we defined a constraint on the shape of the support images: $\mathbb{R}^{RX \times RY \times 3}$, where $X, Y > 0$, defacto allowing for a grid interpretation of the embedding of s with dimensions $\mathbb{R}^{X \times Y \times C}$. A support image was cropped to the largest possible size whenever it failed the constraint. The embedding of s was adapted by taking the mean with respect to the grid size, resulting in a support vector of size C. Subsequently a single prototype was obtained by calculating the mean vector of all support vectors in a similar fashion as was done in Snell et al.,[3] except we only considered a single support set:

$$adapted\,prototype = \frac{1}{|S|}\sum_{\forall s \in S} \frac{\sum_{y=0}^{Y-1}\sum_{x=0}^{X-1} E_\theta(s_{(Rx,Ry)})}{XY},$$

where $s_{(Rx,Ry)}$ was a region in s, with size R and top-left coordinates (Rx, Ry).

2.3 Detection

Each adapted proposal was compared to the adapted prototype via a distance function $d : \mathbb{R}^C \times \mathbb{R}^C \rightarrow [0, \infty]$. A threshold, based on the distances between the prototype and the individual support vectors, determined if a proposal was sufficiently similar to the prototype. The relative size of the proposal with respect to the whole slide image, and the corresponding grid cell coordinates relative to the WSI, were used to retrieve the final detected bounding box. As a post-processing step, overlapping predictions were resolved by non-maximum suppression.

2.4 Iterations

To gain a potential performance boost, we implemented an iterative approach where the top 1 and top 5 predictions in a WSI were included in a new support set, which was used to run the system in an iterative approach. We refer to the first application with the provided examples as iteration 0 and the second application with the top scored system detections as iteration 1.

3. EXPERIMENTAL RESULTS

To verify if the system was capable of detecting objects and tissue components within WSIs, we applied it on two datasets obtained from the Radboud University Medical Center. The first dataset consisted of colon tissue (polyps) and included five WSIs from five different patients, stained with hematoxylin and eosin (H&E). In these WSIs, a total of seventeen regions with healthy colon glands were exhaustively annotated with bounding boxes delineating the colon glands. The second dataset consisted of breast tissue (resections) and included eleven WSIs from eleven different patients, stained with H&E, where each WSI was exhaustively annotated with coarse regions indicating ductal carcinoma in situ (DCIS) of the breast. For both the colon and the breast dataset, we extracted a single WSI for the creation of the support sets. The remaining WSIs acted as queries. As a preprocessing step, we first apply a tissue versus background segmentation algorithm on the queries to remove all background pixels from downstream processing steps.[8] Furthermore, S was extended by a factor 10X via color, spatial, noise and stain augmentations. Subsequently, the segmented tissue in the queries and the examples in the support set are encoded using a pre-trained encoder from Tellez et al.,[7] which was trained on multiple tasks for histopathology. We tested the system with a single example in the support set (n=1) and a support set, including five examples (n=5). Furthermore, we introduced three different *thresholds* which were applied to the similarity values based on the *minimal*, *mean*, and *maximum* values of the distances between examples in the support set and the prototype. This allows the control sensitivity and specificity of the system.

The similarity between the adapted proposals and adapted prototype was computed via the L2 distance, and overlapping predictions with a 0.1 Intersection over Union score were suppressed via non-maximum suppression. The system ran on a GeForce GTX 2080 Ti and 16 CPUs. We evaluated the performance of the system from two different perspectives. Firstly, from an object-based viewpoint, where multiple detections in one ground truth annotation counted as a single true positive, for which we reported the precision, recall, and the F_1-score of the system. Secondly, from a patch-based viewpoint, where each detection was a true positive if it had $\geq 50\%$ overlap with a ground truth annotation, for which we reported the overall accuracy of the system. Table 1 shows the quantitative results of the system on the colon and breast datasets.

3.1 Detection results for colon glands

The support set for colon glands is shown in Figure 2 (left). Experimental results on the colon query slides show that in iteration 0, with a support set of size of 1, and a minimum threshold value, three examples of colon glands were detected with 100% accuracy. For the mean and maximum thresholds, the recall improved slightly at the cost of precision, providing for six detections with 83% accuracy and 230 detections with 56% accuracy, respectively. Using the best-scored detection in iteration 1, the model did not produce any detections using a minimum threshold. Although recall scores increased for the mean and maximum thresholds, accuracy scores decreased compared to iteration 0, by 23% and 11%, respectively. By increasing the support set size to five examples, in both iteration 0 and 1, recall scores improved for all settings while precision and accuracy scores decreased in iteration 0, and precision and accuracy scores improved in iteration 1 when compared to same settings for the n=1 support set. The best trade-off between precision and recall is observed using n=5, in iteration 0 and a maximum threshold, resulting in an F_1-score of 0.6. The best accuracy score with a score of 100% was obtained via the minimal threshold settings in iteration 0, and produced 3 and 4 detections for n=1 and n=5, respectively. Noteworthy, the next best accuracy with a score of 87% was obtained in iteration 1 with n=5, the mean threshold, and produced 582 detections. Figure 3 shows a qualitative result for a colon query. The system detected colon glands by mainly focusing on the borders and elongated glands were detected partially.

3.2 Dectection results for ductal carcinoma in situ

The support set for DCIS is shown in Figure 2 (right). Results on the breast query slides show that with n=1 in iteration 0, for both the minimum and the mean thresholds, no detections were found. When the maximum value was used as the threshold, 97 detections were obtained with 79% accuracy. Using the best-scored detection in iteration 1, the minimal and mean thresholds produced three detections with 100% accuracy and 76 detections with 61% accuracy, respectively. When the maximum threshold was used accuracy drops 18% when compared to iteration 0. By increasing the size support sets to five examples in both iteration 0 and 1, recall scores improved for all settings while precision and accuracy improved for the minimal and mean thresholds in iteration 0 and the mean threshold in iteration 1. The best trade-off between precision and recall was observed using n=5 in iteration 0, and the mean threshold, resulting in an F_1-score of 0.46. For the same setting, 163 detections were produced with an accuracy of 93%. Figure 4 shows a qualitative result for a breast query where each DCIS region is detected by the system. However, a benign region with rapid cell growth (floral ductal hyperplasia) was falsely detected as DCIS in another breast query (see Figure 5).

3.3 Time performance

The encoding time for a single WSI at 20X magnification was between 2 and 3 minutes. Note that only the tissue was encoded and the background was ignored. Furthermore, the encoding of a WSI only needs to be done once and can be saved to disk and reused with any support set of interest. However the time is dependent on the amount of tissue and the grid defined over the query features. The detection in a WSI takes approximately 10 to 20 seconds when using 16 CPUs. However the time is dependent on the grid size, number of proposals and the shapes of the proposals. We empirically tested that a human can select colon glands at a pace of 1.5 seconds per gland. This results in an optimistic 14 minutes of annotation time for 582 colon glands. Our system is able to detect 582 colon glands with reasonable performance within a minute. However, this can be optimized even more when multiple slides are processed in parallel. Note that in the detection part, no GPU is needed because the WSI query is already encoded and encoding the small support examples can be done on a CPU.

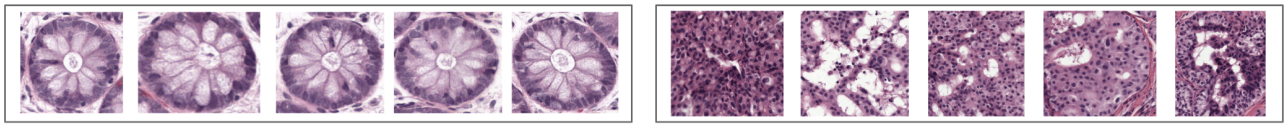

Figure 2: Support set for colon glands (left) and support set for DCIS (right).

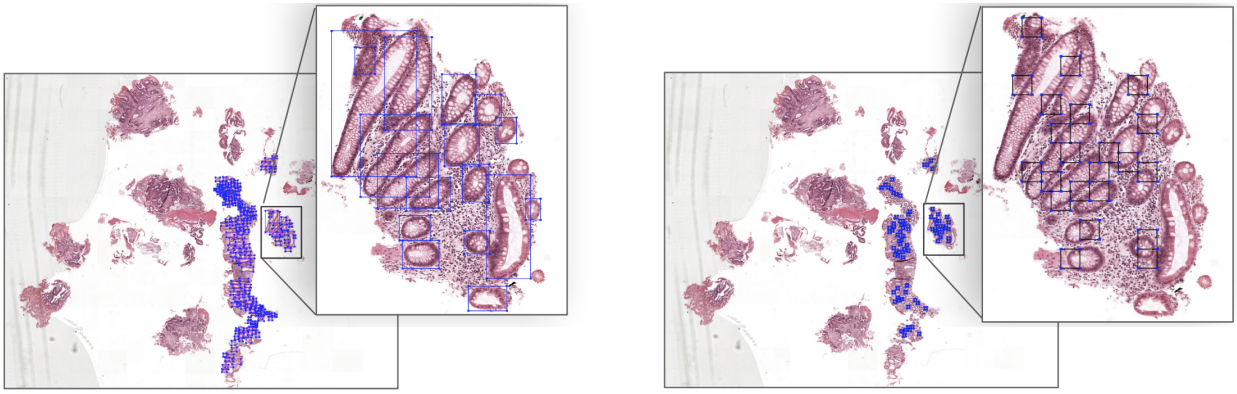

Figure 3: Qualitative results for colon glands. Ground truth (left) and detections made by the system (right) in iteration 1, with 5 support examples and the mean threshold.

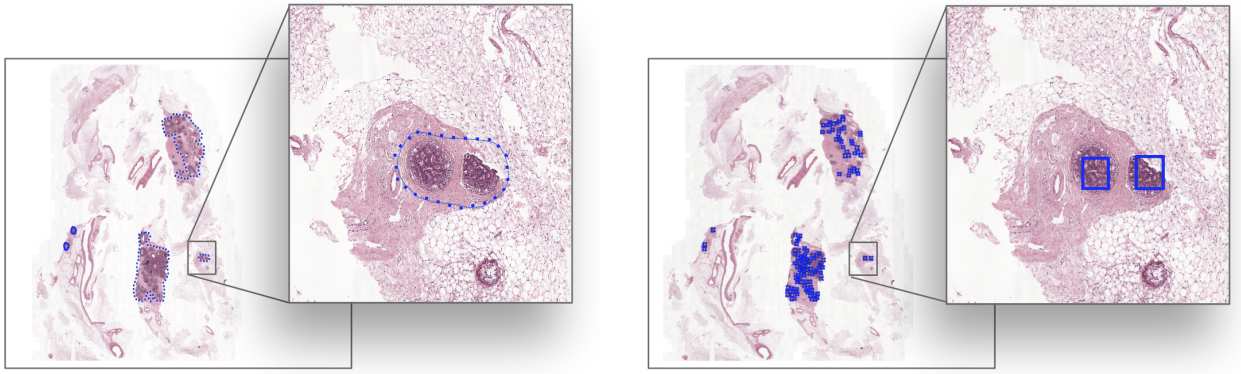

Figure 4: Qualitative results for DCIS regions. Ground truth (left) and detections made by the system (right) in iteration 0 with 5 support examples and the mean threshold.

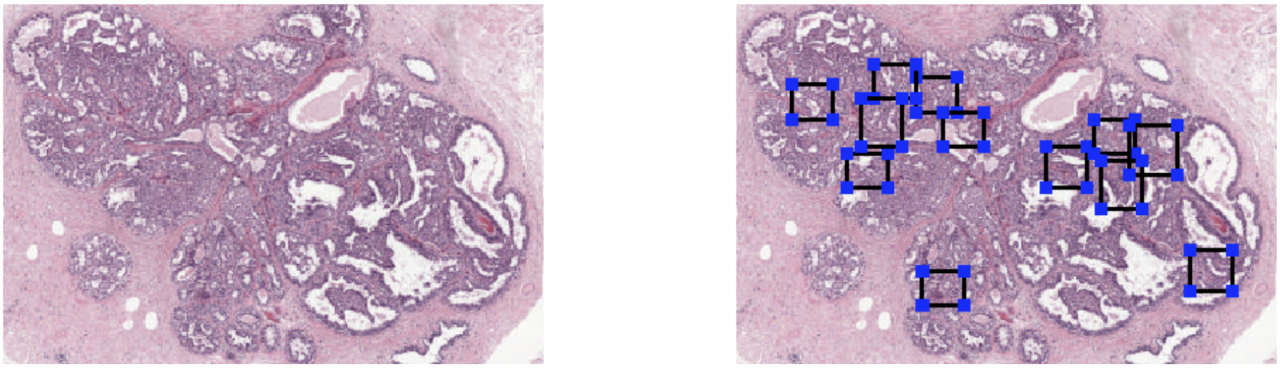

Figure 5: Example of a failure case where a benign region with rapid cell growth (floral ductal hyperplasia) (left) is detected as DCIS by the system (right).

Table 1: Quantitative results for healthy colon glands (top) and ductal carcinoma in situ of the breast (bottom). Precision, recall, and the F1-score were calculated with an object-based viewpoint (i.e., multiple detections in a single ground truth annotation counted as one true positive). The accuracy was computed with a patch-based viewpoint (i.e., a detection was counted as a true positive when it had a $\geq 0.50\%$ overlap with a ground truth annotation). The number between the parentheses after the accuracy depicts the total number of detections.

Colon glands

	n=1						n=5					
	iteration 0			iteration 1			iteration 0			iteration 1		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max
precision	**1.00**	0.83	0.58	0.00	0.60	0.43	**1.00**	0.78	0.36	0.87	0.84	0.49
recall	0.00	0.00	0.10	0.00	0.02	0.35	0.00	0.25	0.66	0.01	0.28	**0.78**
F_1	0.00	0.01	0.16	0.00	0.04	0.39	0.00	0.38	0.47	0.02	0.42	**0.60**
Accuracy	**100%** (3)	83% (6)	59% (230)	0 (0)	60% (47)	48% (1256)	**100%** (4)	78% (486)	50% (3250)	87% (15)	87% (582)	66% (3294)

Ductal carcinoma in situ

	n=1						n=5					
	iteration 0			iteration 1			iteration 0			iteration 1		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max
precision	0.00	0.00	0.26	**1.00**	0.14	0.05	**1.00**	0.40	0.02	0.33	0.15	0.01
recall	0.00	0.00	0.47	0.07	0.34	0.93	0.13	0.53	**1.00**	0.20	0.80	**1.00**
F_1	0.00	0.00	0.34	0.13	0.20	0.09	0.24	**0.46**	0.04	0.25	0.25	0.03
Accuracy	0% (0)	0% (0)	79% (97)	**100%** (3)	61% (76)	61% (707)	**100%** (5)	93% (163)	44% (1220)	50% (12)	75% (280)	36% (1641)

4. DISCUSSION AND CONCLUSIONS

Most image analysis research in computation pathology is based on supervised methods addressing classification, segmentation, or detection. Another research area that is getting more attention is weakly supervised learning, for which only a few labels are available. Examples of weak supervision are the use of a single slide-level label for whole-slide image classification using end-to-end training methods. For the task of image segmentation, weak supervision can be provided via *sparse* annotations (i.e., scribbles or loosely defined regions). However, weak supervision for detection purposes in WSIs has not received much attention yet. In this paper, we attempted to address this problem by proposing a model that builds upon and combines recent few-shot learning methods with well-known detection strategies. Few-shot learning approaches presented in the computer vision community often include a base training phase in which a model is trained on 'base' classes. After this phase, it is expected that the model can produce a distinguished feature vector not only for the training examples but also for novel examples at test time. However, to the best of our knowledge, the field of digital pathology lacks such a dataset. The creation of such a dataset requires labor-intensive work, expertise and is consequently not easily done. To overcome this limitation, in this paper, we relied on the model developed by Tellez et al.,[7] which is pre-trained using digital pathology data from multiple organs and targets multiple tasks to encode digital pathology images into feature map representations. The system included an automatic filtering mechanism via a detection threshold that took on values based on the minimum, mean, and maximum distances between the prototype and the individual support examples. This setup allows the detection of a single class and controls the system's sensitivity and specificity. We showed that when using the minimum threshold, the system can act as a retrieval system with 100% accuracy and could provide the data for large clinical studies by retrieving a collection of WSIs, all containing the same object of interest. However the detection results were not perfect when using the mean or maximum thresholds. The qualitative results showed that the detections could appear partitioned, which is expected in the DCIS regions but was also noticeable for elongated colon glands. Future work could involve an active refinement tool to improve and filter detections, where after the detection could be used in a downstream task.

ACKNOWLEDGMENTS

The authors would like to thank Luca Meesters for her support in the process of annotating the healthy colon glands. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292 (ExaMode).

REFERENCES

[1] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., "A Simple Framework for Contrastive Learning of Visual Representations," *arXiv:2002.05709 [cs, stat]* (2020).

[2] Bokhorst, J.-M., Pinckaers, H., Zwam, P. v., Nagtegaal, I., Laak, J. v. d., and Ciompi, F., "Learning from sparsely annotated data for semantic segmentation in histopathology images," in [*International Conference on Medical Imaging with Deep Learning*], 84–91 (2019).

[3] Snell, J., Swersky, K., and Zemel, R., "Prototypical Networks for Few-shot Learning," in [*Advances in Neural Information Processing Systems 30*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., 4077–4087, Curran Associates, Inc. (2017).

[4] Liu, H. E. and Smith, L. N., "FROST: Faster and more Robust One-shot Semi-supervised Training," *arXiv:2011.09471 [cs, eess, stat]* (Dec. 2020). arXiv: 2011.09471.

[5] Hegde, N., Hipp, J. D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C. J., Amin, M. B., Mermel, C. H., Nelson, P. Q., Peng, L. H., Corrado, G. S., and Stumpe, M. C., "Similar image search for histopathology: SMILY," *npj Digital Medicine* **2**(1), 1–9 (2019).

[6] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You Only Look Once: Unified, Real-Time Object Detection," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 779–788 (2016).

[7] Tellez, D., Hoppener, D., Verhoef, C., Grunhagen, D., Nierop, P., Drozdzal, M., van der Laak, J., and Ciompi, F., "Extending Unsupervised Neural Image Compression With Supervised Multitask Learning," *arXiv:2004.07041 [cs, eess]* (2020).

[8] Bándi, P., Balkenhol, M., Ginneken, B. v., Laak, J. v. d., and Litjens, G., "Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks," *PeerJ* **7**, e8242 (Dec. 2019). Publisher: PeerJ Inc.