Iterative learning to make the most of unlabeled and quickly obtained labeled data in histology

Laxmi Gupta[1] LAXMI.GUPTA@LFB.RWTH-AACHEN.DE
[1] *Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany*

Barbara Mara Klinkhammer[2] BKLINKHAMMER@UKAACHEN.DE
Peter Boor[2] PBOOR@UKAACHEN.DE
[2] *Institute of Pathology, University Hospital Aachen, RWTH Aachen University, Aachen, Germany*

Dorit Merhof[*1] DORIT.MERHOF@LFB.RWTH-AACHEN.DE
Michael Gadermayr[*1,3] MICHAEL.GADERMAYR@FH-SALZBURG.AC.AT
[3] *Salzburg University of Applied Sciences, Salzburg, Austria*

Abstract

Due to the increasing availability of digital whole slide scanners, the importance of image analysis in the field of digital pathology increased significantly. A major challenge and an equally big opportunity for analyses in this field is given by the wide range of tasks and different histological stains. Although sufficient image data is often available for training, the requirement for corresponding expert annotations inhibits clinical deployment. Thus, there is an urgent need for methods which can be effectively trained with or adapted to a small amount of labeled training data. Here, we propose a method to find an optimum trade-off between (low) annotation effort and (high) segmentation accuracy. For this purpose, we propose an approach based on a weakly supervised and an unsupervised learning stage relying on few roughly labeled samples and many unlabeled samples. Although the idea of weakly annotated data is not new, we firstly investigate the applicability to digital pathology in a state-of-the-art machine learning setting.

Keywords: Digital pathology, convolutional neural networks, kidney, segmentation, weakly supervised.

1. Introduction

One of the most significant limiting factors in histological image analysis via machine learning is the need for data annotated by experts (Komura and Ishikawa, 2018). What limits the easy availability of such data is the fact that each whole slide image (WSI) typically has a resolution of about one giga pixels and annotations are required at pixel level. Furthermore, histological images exhibit several dimensions for variability, such as differences in stainings (Gadermayr et al., 2018), etc., which limit generalization of methods developed for a certain data set. Because of these limitations, it is difficult and expensive to obtain annotations.

Researchers have proposed several solutions to deal with these challenges. One option (a) is to optimize the labeling procedure to obtain maximum label data with a fixed manual effort (Komura and Ishikawa, 2018). Another option (b) is to perform effective augmentation of the available labeled data set (Ronneberger et al., 2015). If training data for a similar task is available, transfer

[*] Contributed equally

learning (c) can be applied in order to adjust a pre-trained model (Gadermayr et al., 2018). The pre-trained model is trained on sufficient data on the related task with few labeled samples, unlabeled samples or a mix of both.

These approaches either require sufficient training data for a similar task (c), or they do not focus on making use of unlabeled data (a,b). In the field of digital pathology, however, typically huge amounts of data are routinely captured despite the fact that exhaustive labeling of all data is mostly unfeasible or at least highly uneconomical. In order to make use of the great potential of unlabeled data, here we focus on semi-supervised learning (Khoreva et al., 2017).

Khoreva et al. (2017) evaluate a scenario involving semantic and instance segmentation of 'easy-to-segment' objects using only coarse annotations for training. They generate training labels from bounding boxes (BBs) of the ROIs and train a convolutional neural network iteratively, employing modification cues at each iteration. The cues are based on the labeled BBs and prior information about the objects to be segmented. Their experiments establish that the model benefits from a recursive training returning object shapes significantly better than the input BBs. With their setting, the authors show that segmentation accuracies similar to fully-supervised approaches can be reached using only BBs as annotations.

The first part of our work is inspired by this idea. Their work is based on Pascal VOC12 and COCO datasets, which in comparison to histological images, are easy to segment. In the former, the objects, typically cats and dogs, can be separated from their background class quite distinctly based on color or gradient information. However, histological data consists of rather textured information, with cells, the typical regions of interest (ROIs), resembling their background class quite closely (see Fig. 1).

Contribution:

In this paper, we propose and analyze a pipeline to find an optimum trade-off between segmentation accuracy and manual annotation effort for effectively training segmentation models. For that purpose, we develop a two-stage semi-supervised approach that incorporates a weakly supervised and an unsupervised training method. Inspired by Khoreva et al. (2017), in the first stage we train a fully convolutional network in a weakly supervised way utilizing only a limited amount of quickly obtained rough annotations. In the second stage, we exploit the fact that in digital pathology often large amounts of unlabeled data are available. This data is further used for unsupervised optimization of the model based on specifically developed constraints by incorporating statistical prior knowledge. Here, we investigate the applicability of the method in kidney pathology. Specifically, we segment glomeruli on WSIs of mouse kidney (Kato et al., 2015; Herve et al., 2011) (see Fig. 1). We pose the question for the most effective stage as well as for the best combination of the two stages.

2. Methods

Segmentation Model:

For segmentation, we adapt the method proposed in Gadermayr et al. (2019) (details in Section 3) which is a state-of-the-art approach based on the U-Net architecture (Ronneberger et al., 2015) yielding high accuracies for the same task. As suggested in Gadermayr et al. (2019), we extract training patches randomly from all over the kidney to incorporate data uniformly from the tissue section. We also maintain the ratio of 2:1 for patches with true positives (TPs) and random patches,

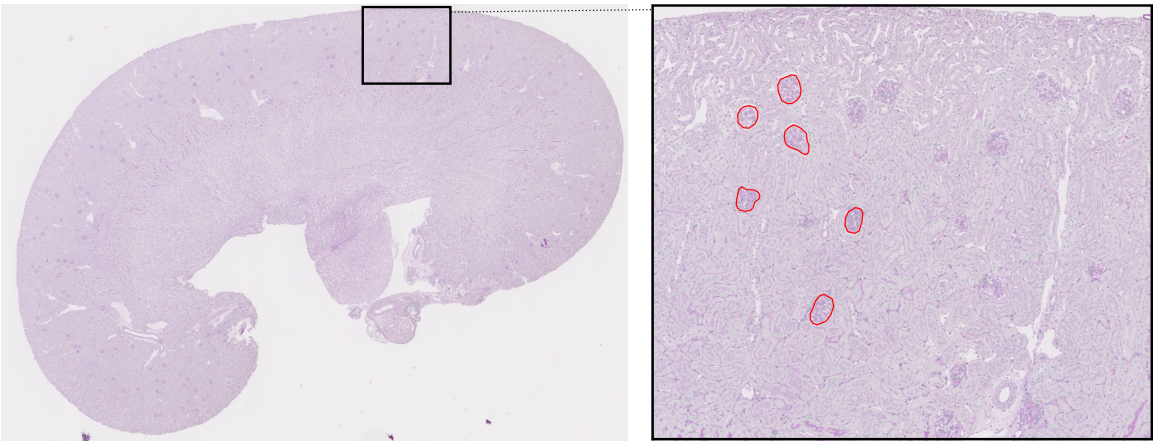


Figure 1: Example WSI of a mouse kidney (left) as well as a magnification of a small patch showing the glomeruli (some are marked in red color, right).

as suggested in the paper for all our experiments. We set the batch size to five, image patch size to 492 and trained with the Adam optimizer on a tensorflow framework. The number of epochs was optimized for each experiment individually (details in Section 2.1).

Proposed Pipeline:

As motivated earlier in this paper, we develop a two stage approach to make the best use of the available weakly labeled (in Stage 1) and unlabeled (in Stage 2) data. The proposed pipeline is outlines in Fig. 2.

As shown in the figure, the training of the network in Stage 1 begins with BBs. The BB of a single object is defined by the minimum rectangle enclosing the object (see Fig. 3). After the first training round, the training images (IS-train1) are segmented using the model thus obtained. These segmentations are then post processed with the help of 'Cues 1' (details in Section 2.1) before using them as the new training labels for the successive iteration. This process is continued for N-iterations.

For Stage 2, the basic iterative training approach remains the same, except for two important differences. Firstly, the initial training labels are generated using the final model trained in Stage 1. This model is applied to the unseen IS-train2 image data set. Secondly, the set of constraints used to modify intermediate results is different (Cues 2) due to the unsupervised setting (we cannot make use of any ground truth (GT) information in Stage 2, for details see Section 2.2).

Ultimately the resulting models from both stages are tested on the test data set IS-test. Further details about the stages are given in the following subsections.

2.1. Stage 1 – Weakly supervised learning

To inspect the applicability of the method to histological images, we adapt the method proposed by Khoreva et al. (2017) to histological data. We apply a very similar approach as proposed in their paper, that involves iterative training and integration of prior information (Cues 1) about the data in the procedure. We initiate the training with BBs as labels. The obtained segmentations, which are

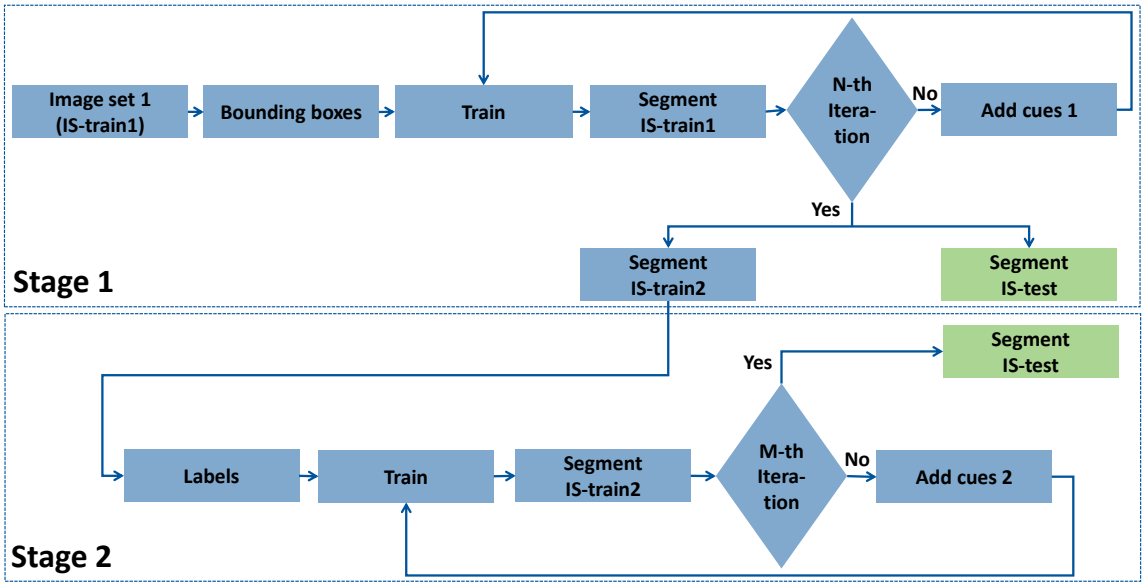


Figure 2: Schematic of the proposed pipeline: Stage 1 shows the weakly supervised training scheme, and Stage 2 shows the unsupervised training scheme.

assumed to be closer to the real object shapes than BBs, are then post processed with Cues 1, which are adapted from Khoreva et al. (2017) to better suit histologial data. The resulting segmentations are in turn are used to generate training labels for the successive iteration.
Cues 1 are namely the following:

- Any object missed completely during segmentation is reset to its corresponding BB. This avoids an increase of the instance of false negatives (FNs) with successive iterations.
- Similar to this, if the segmented object covers less than 50% of the original BB, it is also reset to the latter to avoid training with FNs.
- If any pixel outside the BB of an object is marked as foreground, it is reset as background label because the BBs are assumed to be exhaustive. Thereby, we avoid training with a significant amount of false positives (FPs).

2.2. Stage 2 – Unsupervised learning

We extend the application of the approach further by utilizing the network trained in Stage 1 to facilitate completely unsupervised adaptation in the second step of our pipeline. Effectively, we obtain the training labels for Stage 2 by segmenting a new set of images, IS-train2 (see Fig. 2) using the network output by Stage 1. To keep the training scenario fair, IS-train2 is completely unseen in Stage 1 and we do not consider any GT annotations of IS-train2 data set during training.

For training we again evaluate a similar iterative approach as in Stage 1. However, here we cannot incorporate Cues 1 for the intermediate post processing because of the lack of GT. So, we rely on some basic statistical parameters (area and eccentricity) calculated from the training images

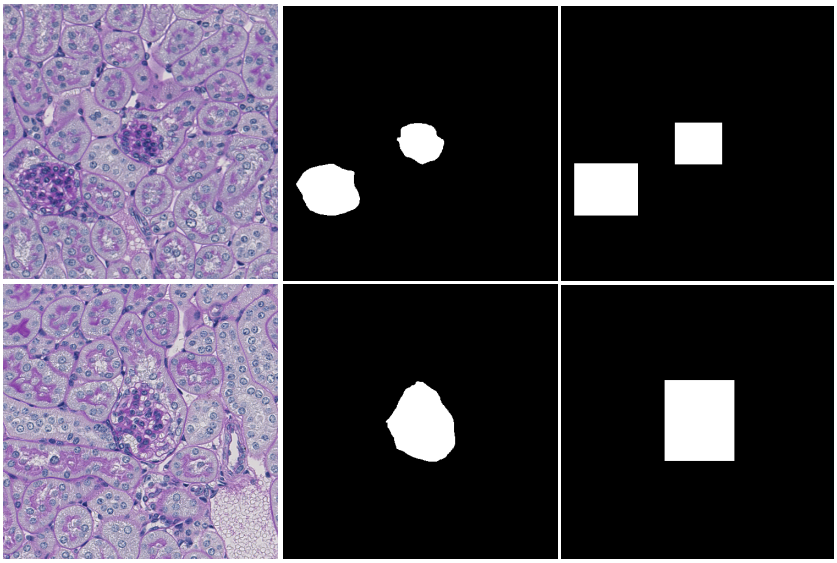


Figure 3: Example showing two 500×500 size patches (left column), the corresponding GT masks (middle column), and the bounding boxes (BBs) (right column).

available for weakly supervised training, IS-train1 (see Fig. 2). For this purpose, we segment IS-train1 with the model resulting in Stage 1 because so far we only had BBs as the GT labels for these images. After segmentation, we obtain 'better than BB segmentations' for the glomeruli on IS-train1, which are then used to calculate the distribution of their area and eccentricity. This information serves as Cues 2.

As explained in Stage 1, here we also train a model followed by segmentation and modification with the constraints (here, Cues 2) as one complete training round. As the initial labels in this stage were not GT, the statistics in Cues 2 was not expected to be a perfect representative of the typical size and shape of glomeruli. In other words, the statistical distribution obtained from these samples also included objects which can be considered as outliers regarding size or shape. Hence, it could be assumed to fit the distribution of the GT only roughly. Taking this into consideration, and to ascertain that unsupervised training incorporates as much correct information about the ROIs as possible, we ignored the objects lying in the marginal distributions (95% confidence interval). Hence, in this stage, after applying Cues 2, we retain only those objects in the segmented images which fall within the 2.5^{th} and 97.5^{th} percentile of the respective distributions. This ensures that we retain only objects which are roughly the size and shape of glomeruli, and use these objects with a high confidence for further training.

3. Image dataset & experimental settings

The dataset used in this work consists of 22 WSIs of resected healthy mouse kidneys which are highly similar to human kidneys. The dataset was divided randomly into two parts, with five WSIs serving as test images (IS-test), and 17 as training images (eight for weakly supervised (IS-train1) and nine for unsupervised (IS-train2)). Image acquisition was performed with a Hamamatsu whole

slide scanner (NanoZommer 2.0-HT (C9600-13)) with a 40 $\times$ magnification. Each WSI was dyed with Periodic Acid-Schiff (PAS) stain and was roughly 37,000 $\times$ 37,000 pixels in size. All the images are processed in the RGB color space.

To determine the optimal parameters for training in Stage 1, we performed some experiments with correctly labeled data for variable training set sizes. We considered 2^3 to 2^8 (8, 16, ..., 256) number of patches for training, extracted equally from a range of 2^0 to 2^3 (0,2, ...,8) number of WSIs. In all, we experimented with 24 different settings, beginning with the extraction of 8 to 256 patches from one WSI to extracting the similar sets of patches from eight WSIs. Here, we optimize the number of epochs based on the number of patches, with higher epochs for lesser training data. We increase the epochs from 2^2 to 2^7 with decreasing number of patches used for training. For each setting, we train the network for 10 iterations. We will refer to this as **Stage 0** in the remainder of the paper.

For **Stage 1**, we extract 16 patches each from eight WSIs (total 128 patches) for training with 32 epochs, similar to the previous experiment with the same number of patches. We train the network for 10 iterations. We then use the networks so obtained at their worst and best scores (networks N1 and N2, respectively) for performing experiments for **Stage 2**. That means we train two unsupervised networks as explained in Section 2.2 to evaluate the effect of Stage 1 on the performance of this step. We take five iterations into consideration for this stage.

4. Results

Fig. 5 shows the results obtained in all the stages based on five test images IS-test.

Stage 0:

Fig. 5(*a*) shows the F-scores with their respective standard deviations for the experiments described in Stage 0. The individual curves for each WSI compare the F-scores reached with different number of patches. At the same time, the overlay of these curves allows the comparison of the effect of training with similar number of patches extracted from different number of WSIs.

Here we see that when training with eight WSIs, the F-score reaches a plateau (=0.89) at 64 patches. Also noteworthy is the low standard deviations in all cases of #WSI=8. On the other hand, when using only one WSI, low F-scores are exhibited with high standard deviations.

When increasing the training data set size in terms of the number of patches extracted from each WSI, the F-scores do not always show an increment. Nevertheless, in except three of the 24 tested settings, we get a mean F-score of above 0.75 even when training the network with as low as eight annotations. However, the standard deviations in all cases are much higher compared to the setting of eight WSIs.

Stage 1:

Fig. 4 shows the results of segmentation after every iterative training of the network in this stage. The figure shows an overlay of the segmentation masks and the GT in one WSI. A perfect overlay of the two is shown in green, falsely segmented objects are red, while (partly) missed objects are shown in blue. A quick glance shows that the FPs reduce greatly with more iterations, as the number of instances of FNs increases.

Fig. 5(*b*) reports the mean F-scores, precision and recall along with their standard deviations for Stage 1. The graph illustrates the values reached on segmenting the test images with the networks obtained after each iteration (total 10) of the training procedure.

Here, we notice that although the recall drops consistently until the sixth iteration, the precision increases, leading to a rise in the F-score. The F-score across all the iterations reaches a peak at this point in the plot.

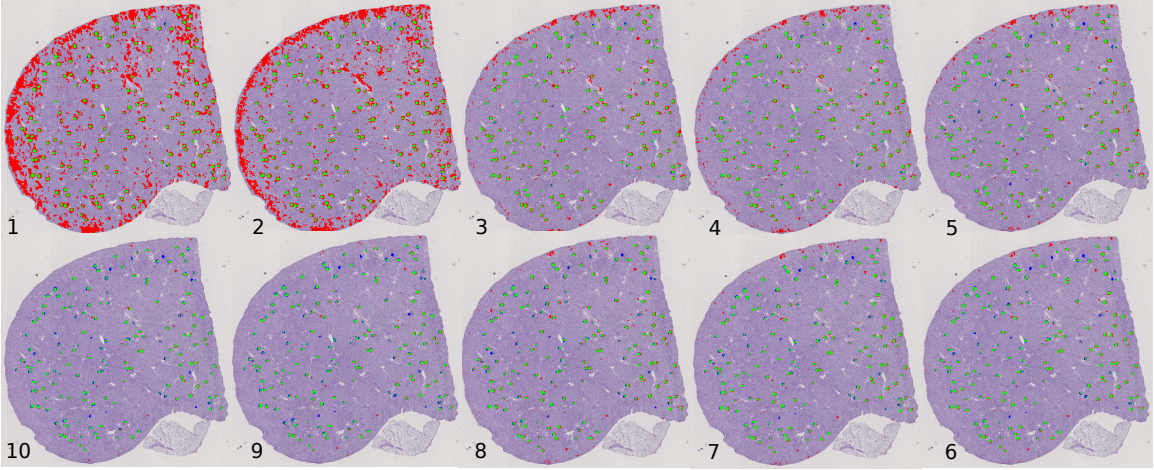


Figure 4: Example showing how the segmentation results evolve on one WSI from 1 to 10 iterations in Stage 1 (clockwise from the top left). Annotation legend: red=FP, blue=FN, green=TP.

Stage 2:

In Fig. 5(*c*) and Fig. 5(*d*), we see the F-score, precision and recall values for the experiments of Stage 2. Here, we see the results for the Stage 2 network (iteration 1 to 5) trained based on the output from the network from the first (N1) and sixth (N2) iteration of Stage 1 (iteration 0), respectively.

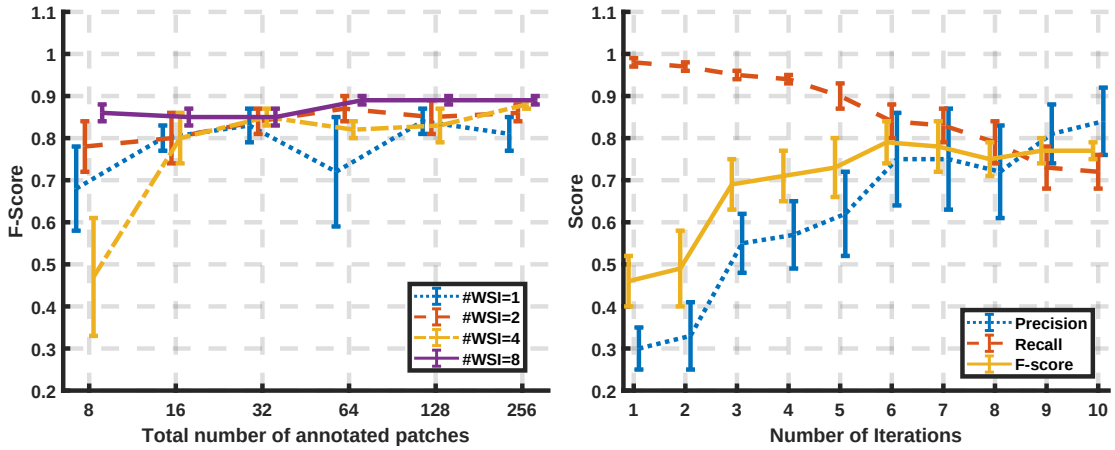
If training Stage 2 after running the first and sixth iterations of Stage 1, respectively, we notice a strong improvement in the F-score between the 0^{th} and 1^{st} iteration. Iteration one here is effectively the the first iteration of Stage 2. Although not much improvement in the values is achieved with more iterations, it is important to note that the standard deviations decrease slightly.

5. Discussion

Referring to Fig. 5(*a*), we may comment that the highest F-scores are obtained with eight WSIs because in this setting we incorporate inter-slide variability much more effectively than in the other settings with fewer number of WSIs. This is reflected also by the comparatively lower standard deviations in the former case.

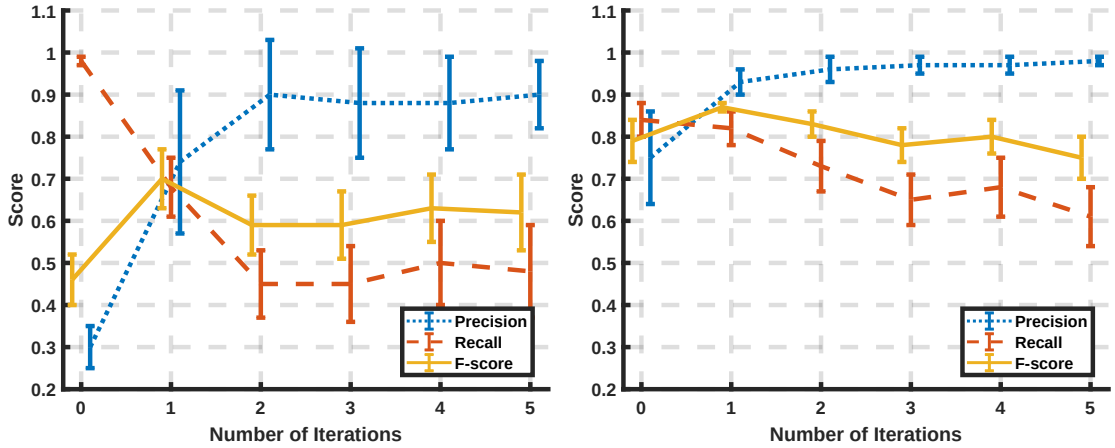
With these observations, we safely chose WSI=8, patches=16 (total patches=128) as the most optimal configuration for Stage 1 (F-score=0.89).

The scores of Stage 1 (Fig. 5(*b*)) may be accounted for as follows. As we initiate training in this stage with BBs, the first iteration has comparatively high number of FPs. This explains the



(*a*) Stage 0: Comparative evaluation of the effect of using varying training set sizes

(*b*) Stage 1: Iterations 1 to 10



(*c*) Stage 2: Based on network (N1) obtained after the first iteration of Stage 1

(*d*) Stage 2: Based on network (N2) obtained after the sixth iteration of Stage 1

Figure 5: Segmentation results on five test images (IS-test) for Stages 0, 1 and 2.

low precision, yet high recall because the network was essentially trained with all the TPs. With each iteration, as the annotations get closer to the GT, precision keeps improving while recall is decreasing. We observe that the overall F-scores increase steadily and the highest score (0.79) is reached at the sixth iteration, when the precision and recall both fall gradually. This is because the iterative training is then gradually incorporating more and more FNs.

The results of Stage 2 (Fig. 5(*c*) and 5(*d*)) show that the network benefits from training on the fully-automatically generated training data because the F-scores improve strongly in the first iteration. However, with the succeeding iterations, the performance decreases. A reason for this could be that Cues 2 used as constraints in this stage were not suitable for the task.

As a qualitative comparison between the annotation efforts with the new approach and the supervised training method, we compare the time taken for labeling in both the cases. For this purpose, we evaluate the experiment setting requiring 16 annotated patches (considering one glomerulus per patch) each, on eight WSIs. The average time taken to label a glomerulus precisely was noted to be about 60 seconds, and to draw a BB around it, approximately 10 seconds. For the proposed method (BB annotations) we needed only 21 minutes (approximately), while for the supervised method (precise annotations), we needed 128 minutes ($>$ 2 hours). Our method saves annotation time significantly (6.5 times), which is an important consideration especially for medical data.

6. Conclusion

Automation of histopathological image analysis procedures typically demands a lot of manually annotated data, which is difficult and time-consuming to obtain. In this work we seek to minimize the limitations caused by the unavailability of fully labeled data by adopting a weakly supervised approach, whereby we require only limited and coarse annotations. This effectively means that we work with imprecise easy-to-collect labels (BBs), refining them with prior knowledge about the dataset, which is easily available. In this endeavor, we achieve mean F-scores of 0.79 when training with as low as eight sparsely and imprecisely annotated WSIs. We address another major limitation in histological image analysis. Although substantial amount of image data is often available, it not routinely utilized. We exploit such unlabeled data to further improve the performance of our weakly supervised models and achieve accuracy values (F-score=0.87) comparable to fully supervised models (F-score=0.89). It is also noteworthy that these results are achieved with significantly reduced annotation effort and time.

Acknowledgement

This work was supported by the German Research Foundation (DFG) under grant no. ME3737/3-1.

References

Michael Gadermayr, Vitus Appel, Barbara M. Klinkhammer, Peter Boor, and Dorit Merhof. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18)*, 2018.

Michael Gadermayr, Ann-Kathrin Dombrowski, Barbara Mara Klinkhammer, Peter Boor, and Dorit Merhof. Cnn cascades for segmenting sparse objects in gigapixel whole slide images. *Computerized Medical Imaging and Graphics*, 2019.

N. Herve, A. Servais, E. Thervet, J.-C. Olivo-Marin, and V. Meas-Yedid. Statistical color texture descriptors for histological images analysis. In *Proceedings of ISBI'11*, pages 724–727, 2011.

Tsuyoshi Kato, Raissa Relator, Hayliang Ngouv, Yoshihiro Hirohashi, Osamu Takaki, Tetsuhiro Kakimoto, and Kinya Okada. Segmental HOG: new descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinformatics*, 16(1), 2015.

Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Aided Interventions (MICCAI'15)*, pages 234–241. Springer International Publishing, 2015.