

CrowdDeep: Deep Learning from the Crowd for Nuclei Segmentation

Parmida Ghahremani* and Arie E Kaufman*

*Computer Science Department, Stony Brook University, Stony Brook, NY, US

ABSTRACT

In recent years, deep convolutional neural networks (CNNs) have shown tremendous success in solving many biomedical tasks. However, the development of deep convolutional networks requires access to large quantities of high-quality annotated images for training and evaluation. As image annotation is a tedious task for biomedical experts, recruiting non-expert crowd workers is an economical and efficient way to provide a rich dataset of annotated images. We present an approach to improve the accuracy of segmenting nuclei in Hematoxylin and Eosin (H&E) slides by hiring crowd workers. We first present a crowdsourcing framework that enables fast and efficient acquisition of nuclei-segmented masks from the crowd by providing manual and semi-automatic annotation methods. Then, we present *CrowdDeep*, a novel technique to improve the accuracy of deep learning models trained on expert annotation by efficiently hiring crowd-annotated data. CrowdDeep consists of two sub-networks: Crowd-Subnet, and Expert-Subnet. The Crowd-Subnet is trained on the crowd-annotated images to extract crowd-related features from the crowd-annotated masks, while the Expert-Subnet is trained on the expert-driven annotations to extract expert-related features from the expert-annotated masks. Then, it calculates the final segmentation mask from the generated segmentation masks by two sub-networks. The results show that CrowdDeep outperforms a CNN model trained on solely expert-derived annotations in terms of F1-Score, IOU, and Pixel Accuracy. This approach is multi-organ and generalizes across different organs, staining, and disease states and is easily expandable by crowdsourcing images with an assortment of nuclei shapes and sizes from any desirable body tissue.

Keywords: Nuclei Segmentation, CNN, Crowdsourcing, Pathology, Histopathological Images, H&E Slides

1. INTRODUCTION

Pathologists diagnose cancer by interpreting the nuclear morphometric and appearance features, such as nuclei density, nucleus-to-cytoplasm ratio, average size, and pleomorphism in a microscopic image of a histopathological section of tissue removed from a patient. With accurate segmentation, these morphological features can be extracted and assessed for grading cancer and predicting treatment effectiveness.

The number of clinical and scientific medical images produced grows quickly. Therefore, computers seem obligatory to assist pathologists in diagnosis, treatment planning, and clinical studies. There are several goals for computer-aided diagnosis, including automating the process to handle a large number of cases with high accuracy, achieving fast and accurate results, and avoiding personal errors resulting from fatigue, data overload, or missing manual steps. There is no universal algorithm for the segmentation of every medical image. On the other hand, machine learning algorithms need large sets of labeled data as their training sets. Obtaining this data is a time-consuming and expensive process acquiring a considerable amount of experts' effort. Thus, crowdsourcing seems an efficient and economical way to generate ground-truth data for these algorithms or even be used directly for medical assessments.

There are several challenges in the way of spreading crowdsourcing for medical diagnosis and treatments. One of the major challenges in crowdsourcing is maintaining the security of medical data. By crowdsourcing the resources, medical data can be accessed by unauthorized parties and random people who may be malicious

Further author information: (Send correspondence to Parmida Ghahremani)

Parmida Ghahremani: E-mail: pghahremani@cs.stonybrook.edu

Arie Kaufman: E-mail: ari@cs.stonybrook.edu

and play with sensitive data leading to losing confidentiality.¹ Therefore, it is crucial to check workers every once in a while.² The other challenge is the possibility of noise generation and complicated care due to using not well-founded information.³ In detecting cysts and tumors, even specialists often cannot reach a consensus, so integrating results from workers is another challenge. As crowdsourcing is an inexpensive way for medical evaluations, people prefer this inexpensive solution over the heavily expensive doctors, resulting in doctors taking exception to it.

We introduce CrowdDeep to segment nuclei in histopathological images accurately. CrowdDeep improves on the state-of-the-art in three ways: First, it introduces an interactive crowdsourcing framework that allows the workers to annotate a large number of images with hundreds of nuclei in them in a short time, enabling the creation of a large dataset of annotations using crowd workers. Second, CrowdDeep takes advantage of this large dataset to train a CNN model to segment nuclei in histopathological images. The CrowdDeep is multi-organ. We train the model with the histology images of seven body organs with variable staining, resulting in a universal model for nuclei segmentation independent of the tissue, organ, and staining method.

2. RELATED WORKS

Accurate nuclear detection and segmentation are necessary for extracting features to be used in computational methods. Over the past decade, automatic algorithms have been developed to analyze histopathological slides and assist pathologists in cancer diagnosis. Accurate nuclei detection and segmentation are necessary for extracting features to be used in these computational methods. In recent years, deep convolutional networks have become the best performing state-of-the-art techniques for various segmentation tasks.⁴⁻⁶ Despite the limited number of annotated histology images, some recent works achieved promising results.

Naylor et al.⁷ present a fully automated workflow for nuclei segmentation in histopathology images by using deep neural networks trained with a set of manually annotated images and post-processing probability maps in order to split jointly segmented nuclei. In another work, Naylor et al.⁸ formulate the segmentation problem as a regression task of the distance map to overcome the challenge of separating touching nuclei in H&E stained images. Alom et al.⁹ use Recurrent Residual Convolutional Neural Networks based U-Net (R2U-Net) to segment nuclei in pathology slides. Cui et al.¹⁰ overcome the segmentation challenge by presenting a nucleus-boundary model that predicts nuclei and their boundaries simultaneously using a fully convolutional neural network. Mask R-CNN and U-Net have shown remarkable performance in medical image segmentation tasks. Vuola et al.¹¹ compare these two segmentation frameworks in the nuclei segmentation task, show the strength and failures of each model, and finally present an ensemble model to combine their predictions that outperforms both models. Koohbanani et al.¹² focus on separating overlapping nuclei by proposing a spatially-aware network (SpaNet) to capture spatial information in a multi-scale manner. They perform pixel-wise segmentation and centroid detection maps of nuclei, followed by applying spectral clustering on the output of the last SpaNet using the nuclear mask and the Gaussian-like detection map for determining the connected components and associated cluster identifier, respectively. Conventional CNN-based approaches lack structured prediction capabilities required for distinguishing the overlapping nuclei. Hence, Mahmood et al.¹³ present a conditional generative adversarial network (cGAN) trained with synthetic and real data to overcome this issue.

The major challenge in developing tools in computational biomedical imaging is access to high-quality annotated images for training and evaluating these algorithms. The large size of whole slide images (WSIs) is an issue for examining these images, and WSIs typically contain thousands of regions-of-interest (ROIs) and tens of thousands of nuclei. Thus, it would be impossible to provide a vast database of expert-derived labeled images. Since image annotation is a tedious task for biomedical experts, recruiting non-expert crowd workers is an economical and efficient way to provide a rich dataset of annotated images. Crowdsourcing is an online distributed problem-solving model where an individual, an institution, or a non-profit organization offers a voluntary task via a flexible open call to a group of individuals of varying knowledge, heterogeneity, and number, known as the crowd.¹⁴ There is no limit on the size of a crowd. However, a crowd is defined as a large network of people which is poorly defined and diverse. Crowds can exist before the start of a crowdsourcing process. Though, most of the time, crowds become active after engaging through an open call for performing crowdsourced tasks.¹⁵

Crowdsourcing can be used for rapidly obtaining annotations for detection, segmentation, and classification of anomalies in biomedical imaging data. In recent works, crowdsourcing was employed to detect polyps in

virtual colonoscopy¹⁶ and nodules in lung computed tomography (CT).¹⁷ Irshad et al.¹⁸ use the crowd as an alternative method for obtaining large-scale image annotations for nucleus detection and segmentation. They ask the non-expert individuals to annotate each nucleus by drawing a boundary around it, which takes a tremendous amount of time to annotate an image with a large number of nuclei. Their results show that non-expert crowd workers perform at a similar level to research fellows and automatic methods for nucleus detection and segmentation. However, since their technique requires the user to draw a line around each cell, a significant amount of time and effort is needed to segment an image with many cells. As their results show, increasing the image size has a negative impact on users' performance, leading to the failure of their method in annotating large images. Albarqouni et al.¹⁹ detect mitosis in histopathological slides by passing crowd votes on detected mitosis candidates to an existing CNN to refine the trained models on different image scales. The primary issue is the need for a sizeable multi-scale dataset of annotated images to train the initial multi-scale model to provide mitosis candidates to the crowd, requiring experts to spend a massive amount of time annotating images.

Crowdsourcing platform is the stage where crowdsourcing is carried out. There are several available platforms for assigning tasks to workers. Amazon Mechanical Turk (AMT or MTurk)*, Innocentive[†], and Openideo[‡] are several example platforms that provide service of creating jobs including several tasks and assigning them to users for a small amount of payment for each task. AMT is an online crowdsourcing marketplace that allows organizations, known as requesters to publish small tasks, known as human intelligence tasks (HITs) to be performed by members of the crowd, known as workers, for a small financial reward.²⁰ For example, Information Retrieval (IR) community hired AMT for annotation^{21,22} or relevance evaluation.²³ Also, it has been used in medical domain for Computer Aided Diagnosis.²⁴ Although, AMT was the first crowdsourcing platform and is still the leading market, a wide range of platforms are emerging. Aside from these services, a small number of open source projects such as Turkit²⁵ and CrowdForge²⁶ offer crowdsourcing services.

3. OVERVIEW

We present CrowdDeep, a novel approach for multi-organ nuclei segmentation in histology images. As shown in Figure 1, this approach consists of three consecutive steps. In the first step, we collect a large dataset of crowd-annotated images using our interactive crowdsourcing framework in a short time, as described in Section 4.1. In the second step, we train a CNN model consisting of two subnetworks. One subnetwork is trained with collected annotations from the crowd to extract crowd-related features, and the second subnetwork is trained with expert annotations to extract expert-related features. In the last step, we combine the generated masks from two subnetworks to create the final segmentation mask. We compare the CrowdDeep with a CNN model trained on solely expert annotations and a CNN model trained on solely crowd annotations, showing that CrowdDeep outperforms the other models by extracting features from both annotation types.

4. METHODS

4.1 Crowdsourcing Framework

In this framework, we ask the crowd workers to annotate 512×512 histology images containing approximately 20 to 300 nuclei. The jobs are published in AMT to be completed by the crowd who are certified workers by AMT. Each job is a collection of five tasks, and each task consists of two images, as shown in Figure 2; the original image is displayed on the left, while the segmented image is shown on the right. Displaying two images side by side enables workers to identify unsegmented nuclei more precisely. Each worker, before entering a job, is trained by reading the instructions, including correct and incorrect examples of annotations and how to interact with the framework. Each job consists of two phases, starting with the question phase and followed by the judgment phase. In the question phase, the worker must reach at least 75% accuracy to enter the judgment phase. In the judgment phase, the worker is asked to continue annotating the images.

We provide two annotation methods in the framework, allowing faster and more accurate segmentation: (1) The worker draws a straight line inside an unsegmented nucleus, and the system automatically annotates the

*<http://www.mturk.com/>

†<https://www.innocentive.com/>

‡<https://www.openideo.com/>

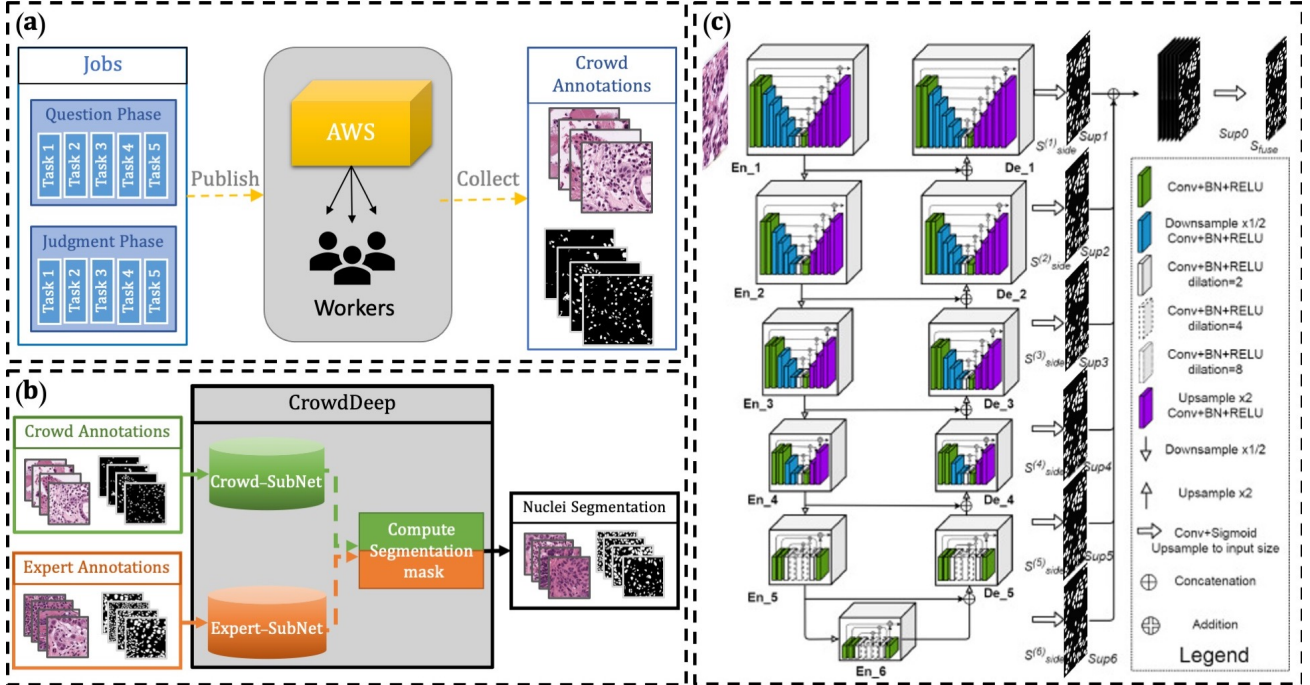


Figure 1. Overview of CrowdDeep. (a) Overview of crowdsourcing framework. First, jobs containing two phases of question and judgment consisting of five tasks are created. These jobs are published through AWS to be completed by the workers. After completion of all tasks, the annotations generated by the users are collected and processed for training the model. (b) Overview of segmentation mask generation. In this step, two U2Net based models are trained using the crowd annotations and the expert annotations, respectively, called Crowd-SubNet and Expert-SubNet. The generated segmentation mask are merged to create the final segmentation mask. (c) The architecture of the Crowd-SubNet and Expert-SubNet models. We use a two-level nested U-structure, called U2Net²⁷ to capture more contextual information from different scales and increase the architecture depth without significantly increasing the computation cost.

remaining nuclei with similar pixel intensity. The result is shown to the worker in the right image. The worker can continue specifying unsegmented nuclei. (2) The worker specifies the boundary of a nucleus precisely by drawing a line around it. The worker is also able to remove a region that has been mistakenly annotated by specifying the region and removing it. The second annotation method is designed to enhance the results of the first annotation technique and make the boundaries more precise, if needed. This interactive framework allows users, including experts and non-experts, to accurately annotate nuclei in large histology images in a short time.

4.2 Segmentation

4.2.1 Ground-truth Dataset

Our expert-derived training set is taken from MonuSeg Training Dataset, a publicly available dataset containing 30 images and approximately 22,000 nuclear boundary annotations, released to the public as a dataset article in IEEE Transactions on Medical Imaging in 2017.²⁸ Each image in this set is tiled into nine parts with overlaps between images to avoid information loss, resulting in 270 images of size 512×512 in $40\times$ magnification zoom. Our full dataset covers 7 body organs (6 breasts, 6 livers, 6 kidneys, 6 prostates, 2 bladders, 2 colons and 2 stomachs), which leads to a complete evaluation of our techniques for segmenting nuclei in histopathological slides.

We first randomly selected 150 images of size 512×512 in $40\times$ from the WSIs in The Cancer Genome Atlas Program (TCGA)[§]. We published each image in three different jobs to be segmented by three different users.

[§]<https://portal.gdc.cancer.gov/>

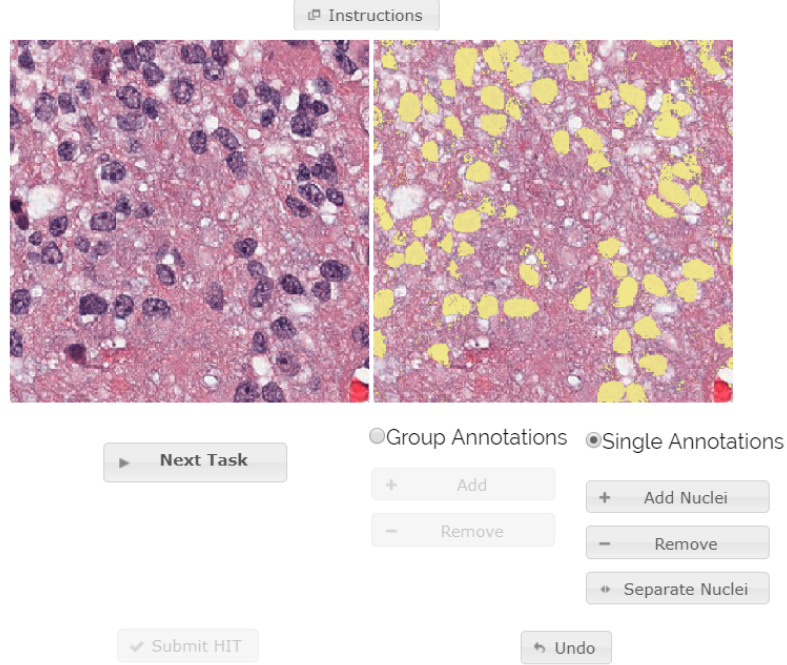


Figure 2. Our crowdsourcing workers’ interface

Due to the failure of some tasks (such as incomplete jobs caused by network issues or lack of interest of the user), we ended up with 354 images in the crowd training set.

We use the MonuSeg Testing Data as our testing set.²⁸ This dataset contains 14 images and approximately 7000 nuclei. Each image in this set is tiled into nine parts with overlaps between images to avoid information loss, resulting in 126 images of size 512×512 in 40× magnification zoom.

4.2.2 Stain Normalization

We use images with inconsistent staining from various body parts. To overcome inconsistencies in staining, we normalized the histopathological images using several stain normalizing techniques^{29–31} and figured out that the stain normalizing method of Macenko et al.²⁹ provides more consistent images leading to more accurate segmentation. We evaluated all techniques using both raw and normalized images. The results show that normalizing images has a significant effect on improving precision.

4.2.3 Model Architecture

Recently, deep convolutional networks have shown promising results in segmentation and classification tasks. However, their success is limited by the size of the training datasets. In the last years, U-structured networks have been employed to solve different biomedical segmentation problems and outperformed the state-of-the-art in many tasks such as the segmentation of neuronal structures in EM stacks, cell segmentation in light microscopy images, and retina blood vessel segmentation.^{32,33} U2-Net²⁷ is a two-level nested U-structure, designed for salient object detection. This design is able to capture more contextual information from different scales due to the mixture of the receptive fields of different sizes in the Residual U-blocks (RSU). It also increases the depth of the whole architecture without significantly increasing the computation cost because of the pooling operations used in RSU blocks. The architecture enables us to train a network from scratch without using backbones from image classification tasks. As shown in Figure 1, CrowdDeep consists of two U2-Net-based sub-networks: (1) Crowd-SubNet, which is trained on crowd-annotated images to extract crowd-related features; and (2) Expert-SubNet which is trained on expert-driven annotations to extract expert-related features. The final segmentation mask is computed by weighted average of the segmentation masks generated by each sub-network.

4.2.4 Experiments

We designed three different experiments to precisely evaluate the use of crowd in training models for nuclei segmentation. In all experiments, we train a U2-Net based model for 500 epochs with a batch size of 4. We used binary cross entropy loss and Adam optimizer with learning rate of 0.001.

Experiment 1 (Expert-Net): In the first experiment, we train our model using expert-derived training set described in Section 4.2.1.

Experiment 2 (Crowd-Net): We train our model using annotations collected from the crowd described in Section 4.2.1.

Experiment 3 (CrowdDeep): We train CrowdDeep with the combination of the crowd and expert-driven annotations. CrowdDeep consists of two sub-networks: Crowd-Subnet, and Expert-Subnet. The Crowd-Subnet is trained with the crowd-annotated images, while the Expert-Subnet is trained on the expert-driven annotations. Then, we calculate the final segmentation mask from the generated segmentation masks by two sub-networks using the following equation:

$$S_i = E_i \times \lambda + C_i \times (1 - \lambda) \quad (1)$$

where E_i is the segmentation mask generated by Expert-Subnet and C_i is the segmentation mask generated by Crowd-Subnet. We empirically found that λ set to 0.6 results in the highest accuracy.

5. RESULTS AND EVALUATION

5.1 Performance Metrics

For comparing the results of the three methods, we calculate different validation measures including $Precision = \frac{TP}{TP+FP}$ (positive predictive value), $Recall = \frac{TP}{TP+FN}$ (true positive rate), $F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision}$ (the harmonic mean of recall and precision), IOU as the intersection over the union, and pixel accuracy (PixAcc) as $\frac{TP}{TP+FP+FN}$, where TP, FP, and FN represent the number of true positive, false positive, and false negative pixels, respectively.

5.2 Evaluation

We evaluate our approaches in terms of the defined performance metrics for seven body organs and report the quantitative analysis of the three experiments in Table 1. Each row of the table represents the three defined metrics for each experiment separately, which are calculated on the set of normalized histology images. In the CrowdDeep row, the results of training the CNN model with a balanced combination of crowd and expert-derived annotations are reported. The results show that a model trained with crowd annotation can work almost as well as a model trained with expert-derived annotations. The crowd detects a different set of features while annotating the images compared to the experts, as merging the annotations generated by Crowd-Net and Expert-Net results in higher accuracy of the segmentation task.

Precision-Recall (PR) curve is plotted using a set of precision-recall pairs. The pairs are computed from a predicted saliency probability map by changing the thresholding value from 0 to 1 and averaging the precision and recall scores of saliency maps of samples. Figure 3 depicts the PR curve of three experiments and the box plots of F1, IOU, and PixAcc for the three experiments. We also show three samples along with the generated segmentation masks for all experiments in Figure 3.

The results of the three experiments show that the Crowd-Net model performs better in finding higher proportions of the nuclei, where the Expert-Net model can segment the nuclei more precisely. It seems that while annotating nuclei, the crowd and experts look for different characteristics in the images resulting in different segmentation masks. As shown in Table 1, the Crowd-Net achieves the highest recall showing that the crowd attempts to find a greater number of nuclei in an image, and the highest precision of Expert-Net shows that experts focus on annotating the most obvious nuclei precisely. In CrowdDeep, we take advantage of both models to find more nuclei with higher precision. As shown in Table 1 and Figure 3, CrowdDeep improves the segmentation accuracy in terms of F1-score, IOU, and Pixel Accuracy, while attempting to get close to the highest precision and recall.

Table 1. Quantitative analysis of the three experiments.

Model	Precision	Recall	F1	IOU	PixAcc
Crowd-Net	71.78	82.87	76.58	62.16	88.72
Expert-Net	83.50	75.73	79.21	65.73	91.19
CrowdDeep	82.05	80.69	81.15	68.41	91.66

6. CONCLUSION

In this work, we designed a crowdsourcing framework for segmenting nuclei. The results show that the crowd performs well in segmenting the nuclei in the H&E slides. The higher recall of the crowd-trained model implies that the crowd pays more attention to segmenting as many nuclei as possible with lower precision, while the higher precision of the expert-trained model depicts that the experts are interested in segmenting nuclei more precisely. We took advantage of the differences in characteristics and features of the annotated masks by crowd and experts to improve the segmentation accuracy of models trained on solely expert annotations. In the future, we will study the use of the CrowdDeep approach in other biomedical tasks such as tumor segmentation and classification, cell and membrane segmentation in fluorescence images.

ACKNOWLEDGMENTS

This project was supported in part by NSF grants CNS1650499, OAC1919752, ICER1940302, and IIS2107224.

REFERENCES

- [1] Krishna, R., Kelleher, K., , and Stahlberg, E., “Patient confidentiality in the research use of clinical medical databases,” *American Journal of Public Health* **97**(4), 654–658 (2007).
- [2] Thawrani, V., Londhe, N. D., and Singh, R., “Crowdsourcing of medical data,” *IETE Technical Review* **31**(3), 249–253 (2014).
- [3] Smith, J. R., “A virtual opinion,” *IEEE MultiMedia* **19**, 2–3 (Feb 2012).
- [4] He, K., Gkioxari, G., Dollár, P., and Girshick, R., “Mask R-CNN,” *Proceedings of the IEEE International Conference on Computer Vision* , 2961–2969 (2017).
- [5] Long, J., Shelhamer, E., and Darrell, T., “Fully convolutional networks for semantic segmentation,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 3431–3440 (2015).
- [6] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *Proceedings of the IEEE International Conference on Computer Vision* , 2223–2232 (2017).
- [7] Naylor, P., Laé, M., Reyat, F., and Walter, T., “Nuclei segmentation in histopathology images using deep neural networks,” *IEEE 14th International Symposium on Biomedical Imaging (ISBI)* , 933–936 (2017).
- [8] Naylor, P., Laé, M., Reyat, F., and Walter, T., “Segmentation of nuclei in histopathology images by deep regression of the distance map,” *IEEE Transactions on Medical Imaging* **38**(2), 448–459 (2019).
- [9] Alom, M. Z., Yakopcic, C., Taha, T. M., and Asari, V. K., “Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net),” *IEEE National Aerospace and Electronics Conference (NAECON)* , 228–233 (2018).
- [10] Cui, Y., Zhang, G., Liu, Z., Xiong, Z., and Hu, J., “A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images,” *Medical & Biological Engineering & Computing* **57**(9), 2027–2043 (2019).
- [11] Vuola, A. O., Akram, S. U., and Kannala, J., “Mask-RCNN and U-Net ensembled for nuclei segmentation,” *IEEE 16th International Symposium on Biomedical Imaging (ISBI)* , 208–212 (2019).
- [12] Koohbanani, N. A., Jahanifar, M., Gooya, A., and Rajpoot, N., “Nuclear instance segmentation using a proposal-free spatially aware deep learning framework,” *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2019).
- [13] Mahmood, F., Borders, D., Chen, R. J., McKay, G. N., Salimian, K. J., Baras, A., and Durr, N. J., “Deep adversarial training for multi-organ nuclei segmentation in histopathology images,” *IEEE Transactions on Medical Imaging* **39**(11), 3257–3267 (2020).

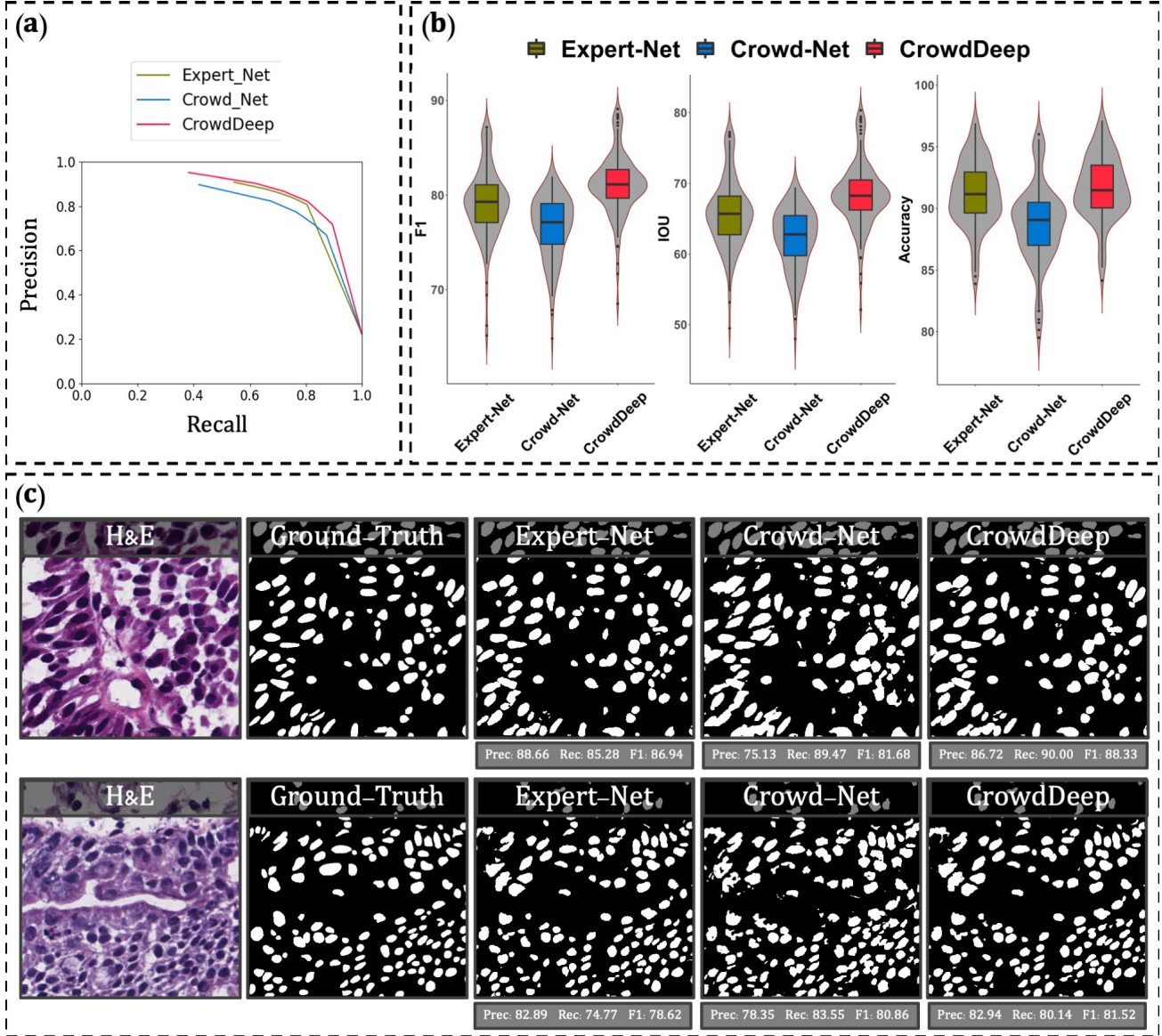


Figure 3. Quantitative and qualitative analysis of three experiments. (a) PR curve of three experiments. (b) F1-Score, IOU, and Pixel Accuracy box plots comparing three experiments. (c) Three sample H&E images along with the ground-truth segmentation mask and the generated masks by three models. As the results show, the CrowdDeep outperforms models trained solely on the crowd- or expert-derived annotations.

- [14] Estellés-Arolas, E. and González-Ladrón-De-Guevara, F., “Towards an integrated crowdsourcing definition,” *Journal of Information Science* **38**, 189–200 (Apr. 2012).
- [15] Cullina, E., Conboy, K., and Morgan, L., “Measuring the crowd: A preliminary taxonomy of crowdsourcing metrics,” *Proceedings of the 11th International Symposium on Open Collaboration (OpenSys)* , 7:1–7:10 (2015).
- [16] Park, J. H., Mirhosseini, S., Nadeem, S., Marino, J., Kaufman, A., Baker, K., and Barish, M., “Crowdsourcing for identification of polyp-free segments in virtual colonoscopy videos,” *Proc. of SPIE Medical Imaging* (2017).
- [17] Boorboor, S., Nadeem, S., Park, J. H., Baker, K., and Kaufman, A., “Crowdsourcing lung nodules detection and annotation,” *Proc. of SPIE Medical Imaging* (2018).

- [18] Irshad, H., Montaser-Kouhsari, L., Waltz, G., Bucur, O., Nowak, J., Dong, F., Knoblauch, N. W., and Beck, A. H., "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd," *Pacific Symposium on Biocomputing* (2014).
- [19] Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., and Navab, N., "Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Transactions on Medical Imaging* (2016).
- [20] Kaufmann, N., Schulze, T., and Veit, D., "More than fun and money. Worker motivation in crowdsourcing - a study on mechanical turk," *AMCIS* (2011).
- [21] Alonso, O. and Mizzaro, S., "Can we get rid of TREC assessors? Using mechanical turk for relevance assessment," *SIGIR Workshop on The Future of IR Evaluation* (2009).
- [22] Nowak, S. and Rüger, S., "How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation," *Proceedings of the International Conference on Multimedia Information Retrieval*, 557–566 (2010).
- [23] Alonso, O., Rose, D. E., and Stewart, B., "Crowdsourcing for relevance evaluation," *SIGIR Forum* **42**, 9–15 (Nov. 2008).
- [24] Nguyen, T. B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., Burns, J. E., and Summers, R. M., "Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography," *Radiology* **262**, 824–33 (2012).
- [25] Little, G., Chilton, L. B., Goldman, M., and Miller, R. C., "Turkit: Human computation algorithms on mechanical turk," *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST)*, 57–66 (2010).
- [26] Kittur, A., Smus, B., Khamkar, S., and Kraut, R. E., "Crowdforge: Crowdsourcing complex work," *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 43–52 (2011).
- [27] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M., "U2-Net: going deeper with nested U-structure for salient object detection," *Pattern Recognition* **106** (2020).
- [28] Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., and Sethi, A., "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging* **36**(7), 1550–1560 (2017).
- [29] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E., "A method for normalizing histology slides for quantitative analysis," *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (2009).
- [30] Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P., "Color transfer between images," *IEEE Computer Graphics and Applications* (2001).
- [31] Zanjani, F. G., Zinger, S., et al., "Deep convolutional gaussian mixture model for stain-color normalization of histopathological images," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2018).
- [32] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241 (2015).
- [33] Xiancheng, W., Wei, L., Bingyi, M., He, J., Jiang, Z., Xu, W., Ji, Z., Hong, G., and Zhaomeng, S., "Retina blood vessel segmentation using a u-net based convolutional neural network," *Proc. of International Conference on Data Science (ICDS)* (2018).