A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology

Neeraj Kumar,* Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi

***Abstract*— Nuclear segmentation in digital microscopic tissue images can enable extraction of high-quality features for nuclear morphometrics and other analysis in computational pathology. Conventional image processing techniques, such as Otsu thresholding and watershed segmentation, do not work effectively on challenging cases, such as chromatin-sparse and crowded nuclei. In contrast, machine learning-based segmentation can generalize across various nuclear appearances. However, training machine learning algorithms requires data sets of images, in which a vast number of nuclei have been annotated. Publicly accessible and annotated data sets, along with widely agreed upon metrics to compare techniques, have catalyzed tremendous innovation and progress on other image classification problems, particularly in object recognition. Inspired by their success, we introduce a large publicly accessible data set of hematoxylin and eosin (H&E)-stained tissue images with more than 21 000 painstakingly annotated nuclear boundaries, whose quality was validated by a medical doctor. Because our data set is taken from multiple hospitals and includes a diversity of nuclear appearances from several patients, disease states, and organs, techniques trained on it are likely to generalize well and work right out-of-the-box on other H&E-stained images. We also propose a new metric to evaluate nuclear segmentation results that penalizes object- and pixel-level errors in a unified manner, unlike previous metrics that penalize only one type of error. We also propose a segmentation technique based on deep learning that lays a special emphasis on identifying the nuclear boundaries, including those between the touching or overlapping nuclei, and works well on a diverse set of test images.**

***Index Terms*— Annotation, boundaries, dataset, deep learning, nuclear segmentation, nuclei.**

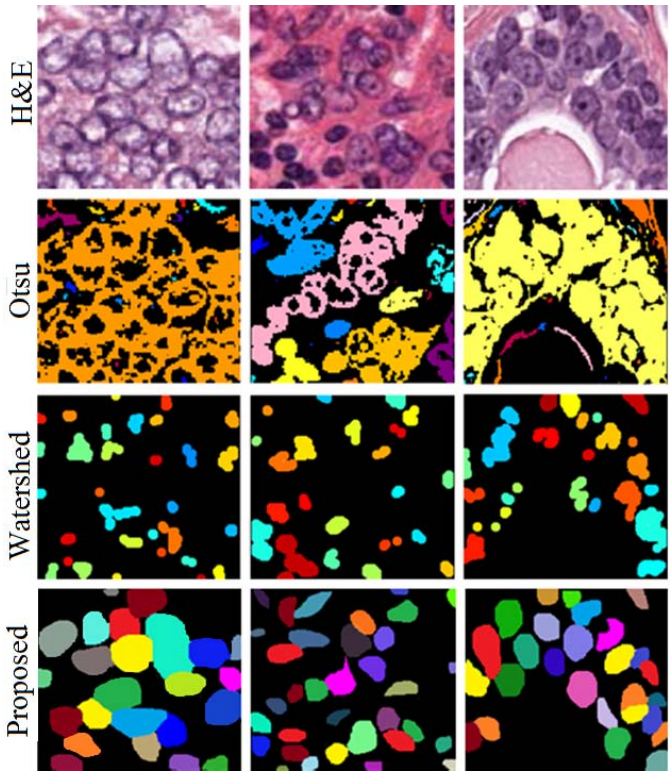


Fig. 1. Challenges in nuclear segmentation: Original H&E stained tissue images show crowded and chromatin-sparse nuclei. Otsu thresholding [9] leads to merged nuclei (under-segmentation). Marker controlled watershed segmentation [10] leads to fragmented nuclei (over-segmentation). Proposed technique detects and segments almost all nuclei well. Each segmented nucleus is shown in a separate color.

I. Introduction

With improvements in computer vision techniques and hardware, some of the problems of manual assessment of histology images, such as inter- and intra-observer variability, inability to assess subtle visual features, and the time taken to examine whole slides [1], [2], are being alleviated by computational pathology [3]. A key module in several computational pathology pipelines is the one that segments nuclei. Nuclear morphometric and appearance features such as density, nucleus-to-cytoplasm ratio, average size, and pleomorphism can be helpful for assessing not only cancer grades but also for predicting treatment effectiveness [4]–[7]. Identifying different types of nuclei based on their segmentation can also yield information about gland shapes, which, for example, is important for cancer grading [8]. Thus, techniques that accurately segment nuclei in diverse images spanning a range of patients, organs, and disease states, can significantly contribute to the development of clinical and medical research software.

The primary goal of this work is to help those working in computational pathology be able to accurately segment nuclei in a diverse set of H&E stained histology images. For this purpose, we are releasing a large dataset of images with annotated nuclear boundaries that are difficult to segment.

Manuscript received January 9, 2017; revised February 10, 2017; accepted February 22, 2017. Date of publication March 6, 2017; date of current version June 29, 2017. *Asterisk indicates corresponding author.*

*N. Kumar is with IIT Guwahati, Guwahati 781039, India (e-mail: neeraj.kumar.iitg@gmail.com).

R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi are with IIT Guwahati, Guwahati 781039, India (e-mail:amitsethi@gmail.com).

Digital Object Identifier 10.1109/TMI.2017.2677499

This will enable training and testing of *readily usable* (or generalized) nuclear segmentation pipelines. This dataset is sampled from the whole slide images obtained from several hospitals and covers multiple organs, patients, and disease states. Additionally, we propose a metric to evaluate nuclear segmentation techniques that treats nucleus detection and segmentation errors in a unified manner. Finally, we propose a nuclear segmentation technique based on deep learning and have released its source code that works well in comparison to other publicly available techniques.

Most of the earlier work on nuclear segmentation did not take challenging cases into account. For example, in pathological conditions (such as hyperplasia or certain cancer subtypes) nuclei enlarge and exhibit margination of chromatin, such that they have a lighter inner body and slightly darker outer ring when stained using hematoxylin (a commonly used bluish-purple dye). Additionally, under such conditions, prominent nucleoli that are darker than rest of the nucleus appear inside the nuclear boundary. Popular image segmentation techniques such as Otsu thresholding [9], or marker controlled watershed segmentation [10]–[12] anticipate relatively uniform and distinguishable colors or textures within a nucleus. This assumption leads to under-segmentation or over-segmentation, as shown in Figure 1, even if we control for the imaging modality by using only one of the most common staining method based on hematoxylin and eosin (H&E) and magnification (40x objective with standard 10x eyepiece leading to approx. $(0.25\mu m)^2$ per pixel). Throughout this paper, we use the term *under-segmentation* for a segment that is unduly large. That is, it almost completely covers a ground truth nucleus and additionally covers significant area outside that nucleus, which may include neighboring nuclei. Conversely, when most of the pixels of an unduly small segment belong to only one nucleus but it leaves out a large proportion of other pixels from the same nucleus uncovered, we call it *over-segmentation*. This also includes cases where a nucleus is split into multiple detected objects. A rare phenomenon that we ignore is where there is only a small overlap between areas of a segment and a ground truth nucleus, which is neither under- nor over-segmentation, but wrong nonetheless.

Using machine learning, new techniques have shown the potential to accurately segment nuclei in the challenging images [13], [14]. A significant barrier in evaluating, using, or improving these techniques is the unavailability of large and diverse annotated training datasets. We address this problem in Section III by introducing a publicly accessible dataset of histology images (including several challenging ones) with more than 21,000 manually annotated nuclear boundaries. Our dataset spans 30 patients and seven organs. We requested an expert pathologist to assess our annotations to ensure that they were of high quality. Further, we suggest how to divide the dataset into a training and two testing sets. One of the testing sets is even more challenging than the other because it covers organs that are not in the training set.

Although several metrics for measuring the nuclear detection and segmentation quality have been reported in the literature, these metrics do not penalize object-level (detection) and pixel-level (segmentation) errors in a unified manner. We show in Section II-D that the previous metrics are inadequate to capture the segmentation quality. In Section IV-B, we suggest a metric that overcomes these challenges.

In Section V, we introduce a readily usable deep learning-based nuclear segmentation technique that seems to work well on the challenging test cases. Our proposed technique is motivated by the need to identify pixels on the nuclear boundaries irrespective of whether they lie on the boundary between a nucleus and surrounding cytoplasm, or between touching or overlapping nuclei. We, therefore, introduce a third class of pixels (nuclear *boundary*), in addition to the two usual classes – background (*outside* all nuclei) and foreground (*inside* any nucleus). We compare this technique with the other open-source software and another technique based on deep learning in Section VI. We are also releasing the source code of our technique to aid its use, evaluation, and improvement. We conclude in Section VII.

II. Background and Related Work

In this section, we review the importance of H&E stained images, nuclear segmentation techniques and metrics, and features of the publicly accessible datasets for computer vision and nuclear segmentation problems.

A. Hematoxylin and Eosin (H&E) Stained Images

Histologic structure of a tissue primarily consists of epithelium (glands), lumen (ducts within glands), adipose (fat), and stroma (connective tissue that holds the glands together). Shape, size, color, and crowding of glands, as well as various nuclei in epithelium and stroma reveal a lot of information about the health of the tissue to a pathologist. The combination of hematoxylin and eosin, or H&E, is a ubiquitous, general, and inexpensive staining (dyeing) scheme. Hematoxylin renders nuclei dark blueish purple and epithelium light purple, while eosin renders stroma pink. Together, H&E enhance the contrast between nuclei, epithelium, and stroma for examination under a microscope.

There seems to be a vast amount of untapped information in H&E stained images that can be used for specific diagnoses such as cancer molecular sub-types determination [15], mortality prediction [4], and treatment effectiveness prediction [7], [16]. Due to its low cost, widespread use for primary diagnosis, and potential for use in highly predictive models, our dataset covers H&E stained images. However, most machine learning techniques can easily be trained on tissue images with other types of stains.

B. Nuclear Segmentation Techniques for H&E Stained Images

Most state-of-the art nucleus segmentation techniques use watershed segmentation, morphological processing, color-based thresholding, active contours, and their variants along with a multitude of pre- and post-processing techniques to achieve the aforementioned goals [10], [11], [17]–[20]. However, such methods fail to generalize across a wide spectrum of tissue morphologies (Figure 1) due to inter- and intra-nuclear color variations in crowded and chromatin sparse nuclei.

Techniques based on machine learning can give better results on the challenging cases of nuclear segmentation because they can be trained to recognize nuclear shape and color variations. One class of learning-based methods use hand-crafted features such as color-texture, blue ratio, color histograms, Laplacian of Gaussian response, geometric features from the gradient or Hessian profiles and other image characteristics in standard learning based models to segment nuclei and non-nuclei regions [5], [21]–[23]. A second class of learning based methods use deep learning – specifically, convolutional neural networks (CNNs) – instead of hand crafted features and have outperformed previous techniques in nuclear detection or segmentation [13], [14], [24]. These techniques estimate a probability map of the nuclear and non-nuclear (two-class) regions based on the learned nuclear appearances, and rely on complex post-processing methods to obtain the final nuclear shapes and separation between touching nuclei. For example, a graph partitioning method was used by [24], while a distance transform of the nuclear map followed by H-minima thresholding and region growing was used by [13]. A more comprehensive review of state-of-the-art nuclei segmentation algorithms can be found in [25] and [26]. These techniques have not yet been demonstrated to work on multiple organs or disease states right out of the box (without re-training) and their source codes are not publicly available.

We propose to explicitly find inter-nuclear boundaries using a third class of pixels to separate crowded nuclei later in this paper to improve upon the binary classes learned by previous deep learning techniques. We also release both an instance of our technique trained to work on multiple organs, as well as its source code.

C. Computer Vision Datasets

Significant progress has been made on certain computer vision problems due to a healthy competition enabled by publicly available datasets and evaluation metrics for benchmarking such as ImageNet [27] and CIFAR [28] for object recognition in images, and UCF for action recognition in videos [29]. Medical imaging community has eventually started to follow this lead with the release of well-annotated datasets and organization of competitions for segmentation, classification, and detection [30]–[33].

For detecting and segmenting nuclei, a few large datasets have recently been released. One of them has around 29,000 marked nuclear centers for detection [14], but does not have annotated nuclear boundaries required for training and testing segmentation techniques. Additionally, it contains images from only one organ. Datasets with annotated boundaries of a few thousand nuclei for single organs have previously been released [34]–[36]. We introduce one of the first large datasets of diverse nuclei from multiple organs with annotated boundaries that contains more than twice the number of annotated nuclei compared to a previous notable effort to introduce a multi-organ, multi-disease state dataset [20]. Additionally, we cover cases of crowded nuclei more extensively.

A useful concept utilized in some human action recognition datasets is that of *groups* [29]. A group is a set of samples within a class that are similar in the way the data was acquired. For example, videos shot in the same background of different instances of the same action could be clubbed as a group. Keeping an entire test group out of the training set assesses the generalization ability of the classification methods. In histology, we propose that the groups can correspond to patients, batches of slides, disease states, or organs.

D. Methods for Evaluation of Segmentation Techniques

A good evaluation criterion for nuclear segmentation techniques should penalize both object-level (nucleus detection) and pixel-level (nucleus shape and size) errors listed below:

1) Missed detection of ground truth (annotated) objects,
2) False detection of ghost objects,
3) Under-segmentation of correctly detected objects, and
4) Over-segmentation of correctly detected objects.

A commonly used object detection metric is the F1-score. For ground truth objects G_i indexed by i and segmented objects S_j indexed by j, the F1-score is based on true positives TP (count of all ground truth objects G_i with an associated segmented (detected) object S_j), false positives FP (count of all segmented objects S_j without a corresponding ground truth object G_i), and false negatives FN (count of all ground truth objects G_i without a corresponding detected object S_j). F1-score is defined as follows:

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

The key detail in evaluating an F1-score is a criterion for deciding whether a ground truth object G_i has an associated segmented object S_j, that is, whether G_i has been detected successfully. This has been done in different ways, where some ways are more generalizable than others for evaluating the detection quality. For example, the association criterion for a ground truth nucleus proposed in [13] was based on finding its closest segmented nucleus and checking if their centroids were also less than a threshold apart. This is not generalizable because the distance threshold is set subjectively and may need to change with magnification, imaging modality, object type, organ, and disease state. The association criterion for ground truth glands proposed in [31] was based on finding segmented glands that cover at least 50% of the ground truth (annotated) gland, which is more generalizable for detection than a distance threshold-based criterion.

A major shortcoming of F1-score using any object association criteria is that it does not take pixel-level (segmentation) errors into account. For example, while the association criterion in [31] does not penalize under-segmentation, the one in [13] does not penalize over-segmentation of crowded nuclei. Thus, most work on nuclear segmentation reports two metrics, one to evaluate detection (object-level errors), and another to report shape concordance (pixel-level errors) between the ground truth objects and their associated segmented objects.

To compute shape concordance between a ground truth object G_i and its associated segmented object S_j one of the following metrics is often used – Jaccard index [37], Dice's coefficient [38], or Housdroff distance [39]. Of these, Hausdorff distance, which penalizes distance of furthest pixels on contours of two shapes, is less popular because of its

TABLE I
COMPOSITION OF THE DATASET AND ITS PROPOSED DIVISION FOR TRAINING, VALIDATION, AND TESTING

Data subset ↓	Nuclei	Images							
	Total	Total	Breast	Liver	Kidney	Prostate	Bladder	Colon	Stomach
Training and validation	13,372	16	4	4	4	4	–	–	–
Same organ testing	4,130	8	2	2	2	2	–	–	–
Different organ testing	4,121	6	–	–	–	–	2	2	2
Total	**21,623**	**30**	**6**	**6**	**6**	**6**	**2**	**2**	**2**

computational complexity being $O(|G_i||S_j|)$ while that of the other two is $O(|G_i| + |S_j|)$, where G_i and S_j are the sets of pixels in ground truth object i and segmented object j respectively. Jaccard index J and Dice's coefficient D are closely related and measure the relative area of overlap between the two sets, and are widely used. Only the true positives among the detected objects affect these shape concordance metrics, and thus object-level errors are not accounted for by such metrics. Later in this paper, we modify Jaccard index because it is easy to understand and covers the same information as Dice's coefficient. Our modification penalizes both detection and segmentation errors.

For two sets, which in our case are pixels of a ground truth nucleus G_i and its associated segmented nucleus S_j, Jaccard index is defined as follows:

$$J = \frac{|G_i \cap S_j|}{|G_i \cup S_j|} \tag{2}$$

It isn't clear how two nuclei segmentation techniques can be compared if one has a better object detection (e.g. F1-score), and the other has a better average shape matching for detected objects (e.g., Jaccard index). A unified metric that combines both object-level and pixel-level performance is desirable. Additionally, we show in Sections IV-B and VI-C that the currently popular metrics do not reflect segmentation quality as expected. Moreover, a detection criterion that is free of hyper-parameters is also needed so that it can be applied to different magnifications and image types.

III. ANNOTATED DATASET

Finding, downloading, and annotating tissue slides is a time-consuming task requiring expertise, which may impede the development of new nuclear segmentation software that can be used in computational pathology. Our publicly hosted dataset with a diverse set of tissue images and painstakingly annotated nuclear boundaries can fill this gap. It can be used by the research community to develop and benchmark generalized nuclear segmentation techniques that work on diverse nuclear types. Our dataset consists of one of most the commonly used images – H&E stained and captured at 40x magnification. Although we used H&E stained tissue slides digitized at 40x magnification for developing the proposed algorithm, but our approach can be easily applied to the commonly available 20x slides by re-training or appropriate upsampling to 40x using super-resolution techniques tailored for such images [40].

We downloaded 30 whole slide images (WSIs) of digitized tissue samples of several organs from *The Cancer Genomic Atlas* (TCGA) [41] and used only one WSI per patient to maximize nuclear appearance variation. These images came from 18 different hospitals, which introduced another source of appearance variation due to the differences in the staining practices across labs. The details of the dataset can also be found in the supplementary materials available in the supplementary files /multimedia tab or our http://nucleisegmentationbenchmark.weebly.com/website. Since computational requirements for processing WSIs are high, we cropped sub-images of size 1000×1000 from regions dense in nuclei, keeping only one such cropped image per WSI and patient. To further ensure richness of nuclear appearances, we covered seven different organs *viz.*, breast, liver, kidney, prostate, bladder, colon and stomach, including both benign and diseased tissue samples. An illustration of the spectrum of tissue appearances and their nuclei is shown in Figure 3, row 1.

After obtaining 1000×1000 sub-images, we annotated more than 21,000 nuclear boundaries in Aperio ImageScope®. Images were enlarged to $200\times$ on a 25" monitor such that each image pixel occupied 5×5 screen pixels for clear visibility, and the nuclear boundaries were annotated with a laser mouse. The annotators were engineering students, and were trained to identify nuclear boundaries by the co-authors. The generated XML files containing pixel coordinates of the annotated nuclear boundaries are available on our website.[1] Our annotations included both epithelial and stromal nuclei. For overlapping nuclei, we assigned each multi-nuclear pixel to the largest nucleus containing that pixel. A representative set of annotations is shown in row 2 of Figure 3, and the composition of the dataset is shown in Table I.

We sent annotated images to an expert pathologist for examination of annotation quality. In a PowerPoint® deck, we used one image per slide. On a slide, we put the unannotated and annotated images side by side to cover a large portion of the slide. The pathologist viewed the slide on a 25" monitor, and was instructed to place an arrow shape on every problematic annotation, whether it was a false positive, a false negative, an over-segmented, or an under-segmented nucleus. We counted all the arrows and divided the count by the number of annotated nuclei in those images to estimate that our annotators made less than 1% errors on any given image. We left these errors uncorrected due to their low count.

IV. TRAINING AND TESTING PROTOCOL

To facilitate the development of generalized nuclear segmentation techniques, we propose how to split this dataset into training and testing sets, as well as an evaluation metric that penalizes both object-level and pixel-level errors.

[1] http://nucleisegmentationbenchmark.weebly.com/

A. Training and Testing Sets

As shown in Table I, we propose to keep images corresponding to 16 patients equally divided among four organs – breast, kidney, liver, and prostate – in the training and validation set. This corresponds to over 13,000 annotated nuclei. For any segmentation technique based on pixel classification, this corresponds to several hundred thousand *pixels* (along with their surrounding and overlapping patches) belonging to nuclei class that can be used to train a machine learning system (e.g. a CNN) to produce binary maps (e.g. in [13]), even without data augmentation (e.g. by rotating and flipping the surrounding patches). We have divided rest of the images into two test sets.

1) Same Organ Test Set: The first test set has *images* from the same organs – breast, kidney, liver, and prostate – that are represented in the training set, although, from different *patients*. Most nuclear segmentation techniques based on machine learning train and test on only one organ. Using this dataset, their generalization to four different organs can be tested.

2) Different Organ Test Set: The second test set is even more challenging because its images are taken from organs not represented in the training set – bladder, colon, and stomach. An algorithm that performs well on the six images from this set can be expected to generalize nuclear segmentation pretty well for H&E stained images imaged at 40x magnification.

B. Proposed Evaluation Criterion

We propose a parameter-free detection criterion that works regardless of nuclear size and magnification. We also propose a unified evaluation metric that penalizes both object-level and pixel-level errors to overcome the shortcomings of other evaluation criteria discussed in Section II-D. We use the spirit of the Jaccard index (Equation 2) for both these goals of detecting a nuclei and evaluating the segmentation results. We also give primacy to the ground truth nuclei.

1) Detection Criterion: With each ground truth nucleus indexed by i represented as a set of pixels G_i, we associate a detected nucleus that maximizes their pixel-wise Jaccard index as per Equation 2. This criterion does not depend on a subjective distance or pixel overlap threshold and can be applied across magnifications or object types. Thus, one detected object can correspond to more than one ground truth objects (e.g., when under-segmented), but not the other way around. When combined with the evaluation metric proposed below, the detection criterion accounts for the four types of errors listed in Section II-D.

2) Evaluation Metric: We propose the following metric to evaluate the performance of a nuclear segmentation method over an image or a dataset of images, which we call aggregated Jaccard index (AJI). It computes an aggregated intersection cardinality numerator, and an aggregated union cardinality denominator for all ground truth and segmented nuclei under consideration. For each ground truth nucleus G_i in an image (or a dataset), after associating a segmented nucleus S_j, we add the contributions to the aggregated Jaccard index by adding the pixel count of $G_i \cap S_j$ to AJIs numerator, and that of $G_i \cup S_j$ to the denominator. This naturally adds pixels of those ground truth nuclei that do not find an intersecting segmented nucleus (detection false negatives) to the denominator. We also add the pixel counts of all unclaimed segmented nuclei (detection false positives) to the denominator. Because this metric adds the pixel counts of false positives and false negatives to the denominator in addition to the pixels of non-overlap among ground truth and detected nuclei (true detection), it penalizes all four types of errors listed in Section II-D. It is worth noting that the previous segmentation metrics are only computed over true positives, and the pixels in false positives and false negatives are completely ignored in evaluating the segmentation quality. When an under-segmented nucleus corresponds to multiple ground truth nuclei, the proposed metric has the potential to count several falsely detected pixels multiple times in the denominator. This is necessary, because otherwise it is possible to obtain low *mean* Jaccard index by biasing a segmentation system towards slight under-segmentation. Thus, aggregated Jaccard index (AJI), in general, has a lower value than F1-score and mean Jaccard index. In Section VI-C we show empirical evidence of how higher AJI better represents segmentation (and detection) quality.

Our detection criterion and evaluation metric are described in detail in Algorithm 1. In case no segmented nucleus intersects with a ground truth nucleus, the intersection pixel cardinality is zero, and union pixel cardinality is the same as $|G_i|$ in step 4 of the algorithm. AJI will range between 0 for the worst case (no intersection between ground truth and segmented objects), and 1 for the best case (perfect detection and segmentation).

We made additional enhancements to step 3 of Algorithm 1. In case of a tie of Jaccard indices between more than one segmented nuclei, the one that maximized the intersection with the ground truth nucleus was selected.

Algorithm 1 Computing Aggregated Jaccard Index (AJI)

Input: A set of images with a combined set of annotated nuclei G_i indexed by i, and a segmented set of nuclei S_k indexed by k.

Output: Aggregated Jaccard Index A.

1: Initialize overall correct and union pixel counts: $C \leftarrow 0$; $U \leftarrow 0$
2: **for** Each ground truth nucleus G_i **do**
3: $\quad j \leftarrow \arg\max_k(|G_i \cap S_k|/|G_i \cup S_k|)$
4: $\quad$ Update pixel counts: $C \leftarrow C + |G_i \cap S_j|$; $U \leftarrow U + |G_i \cup S_j|$
5: $\quad$ Mark S_j *used*
6: **end for**
7: **for** Each segmented nucleus S_j **do**
8: $\quad$ If S_k is not *used* then $U \leftarrow U + |S_k|$
9: **end for**
10: $A \leftarrow C/U$

V. Deep Learning-Based Nuclear Segmentation

When a large number of annotated examples are available, deep learning techniques, especially CNNs, have shown

state-of-the-art performance for image classification [42]. CNNs are also being used on large medical images to produce probability maps for detecting various tissue segments based on processing small fixed size patches sampled from large images. Our main comparative example is a two-class CNN that classifies whether the central pixel of a patch belongs to a nucleus or not [13]. Such a CNN, when applied to all possible overlapping patches of a fixed size, will give a binary or nuclear probability map. To resolve touching nuclei, it has been proposed that distance transform be computed on the binary map, which is likely to yield two separate local minimas near the centers of two touching nuclei due to the concavity in the shape of multiple nuclei. Shape concavity has been used by others to separate touching nuclei as well [11], [22], which works well for simple cases of two touching nuclei (like a figure of eight), but not for more complex cases, as we shall demonstrate in Section VI-B.

We propose a deep learning based technique, in which a CNN is used to produce a ternary map, unlike a binary map that has been used in the past [13]. While two-class CNNs can distinguish between the inside and outside of nuclei even in chromatin-sparse nuclei, the additional advantage of a third class of the nuclear boundary pixels is that it can find inter-nuclear boundaries irrespective of the configuration of the crowded nuclei. That is, it can also help in segmenting crowded nuclei in addition to chromatin sparse nuclei. Our technique is based on this insight, which has also benefited gland segmentation [31], [43]. Below, we also describe its pre-processing and post-processing steps.

A. Pre-Processing: Color Normalization

One of the challenges for tissue segmentation is a large variation in image colors due to H&E reagents, staining process, and sensor response [44], [45]. Color normalization has been shown to improve tissue segmentation by accounting for the variations in the staining and scanning processes, especially one that preserves biological structure by basing color mixture modeling on sparse nonnegative matrix factorization (SNMF) [44], [45]. This technique first assesses a stain density map for a given image as follows. For a pixel vector $\overrightarrow{x}$ with (R, G, B) components in the 8-bit pixel range $[0, 255]$, Beer-Lambert transform converts it into an optical density vector $\overrightarrow{y}$ with components (r, g, b) such that, $r = -\log(R/255)$, $g = -\log(G/255)$, $b = -\log(B/255)$. Then, stain density is obtained using SNMF over the data matrix $\mathbf{Y}$ corresponding to all pixels such that one matrix contains the (r, g, b) optical density components of each stain prototype (two prototypes for H&E), and the other contains the stain mixing components of each pixel, or stain density maps. The tradeoff between sparseness and reconstruction accuracy is controlled by a hyper-parameter λ that is multiplied with the L1 norm of the weight (stain density) matrix, which was recommended to be set to 0.1 [44]. Introduction of sparseness emulates the specificity of target material for each stain in the staining process, and thus preserves the biological structure. A color normalized image is computed by multiplying its stain density map with a color prototype basis matrix of a standard image, and taking the inverse Beer-Lambert transform. The target image was selected by a pathologist based on what he subjectively determined to be a well-stained image and has been made available in our dataset online. Sample results of the color normalization procedure can be assessed by comparing rows 1 and 4 in Figure 3. Specifically, intra-image inter-class (nucleus to stroma) contrast is preserved, while inter-image intra-class (nucleus to nucleus or stroma to stroma) contrast is reduced due to color normalization.

TABLE II
CNN3 ARCHITECTURE FOR SEGMENTING NUCLEI

Layer	Filter size	Activation	Output size	Dropout rate
Input	–	–	$51 \times 51 \times 3$	–
Conv 1	4×4	ReLU	$48 \times 48 \times 25$	0.1
Pool 1	2×2	Max	$24 \times 24 \times 25$	–
Conv 2	5×5	ReLU	$20 \times 20 \times 50$	0.2
Pool 2	2×2	Max	$10 \times 10 \times 50$	–
Conv 3	6×6	ReLU	$5 \times 5 \times 80$	0.25
Pool 3	2×2	Max	$3 \times 3 \times 80$	–
FC 1	–	ReLU	1024	0.5
FC 2	–	ReLU	1024	0.5
Output	–	SoftMax	3	–

B. Three-Class CNN Emphasizing Nuclear Boundaries

Our novel idea for the CNN architecture is to emphasize detection of nuclear *boundary* pixels, including those between two touching nuclei, by explicitly introducing a third class of pixels in addition to the usual binary of foreground (*inside* any nucleus) and background (*outside* every nucleus). As we show later, this simplifies nuclear segmentation of even touching nuclei with a simple post-processing method. For estimating the 3-class probability assignment for each pixel, we trained a 3-output node CNN that took a color normalized patch of size 51×51 centered at that pixel as input. The size of the patch was selected to cover most of the large nuclei, and was arrived at using validation such that it was the smallest patch size (for efficient computation) that displayed saturation in performance. To generate the target output, we obtained a ternary mask for each image using the annotated nuclear boundaries. *Boundary* class pixels were those that were within a distance of ± 1 from the annotated boundary, forming a ring around each annotated nucleus (Figure 3, row 3). The two (usual) remaining classes – *inside* and *outside* (not shown) – were labeled such that they did not overwrite the *boundary* class.

The architecture of the CNN designed to predict the probability of the three classes is shown in Table II. The last layer was a softmax layer so that the three probabilities sum to one. We empirically determined that three convolutional layers and two fully connected layers were well suited for our problem. Adding more layers increased the computation time without an appreciable positive impact on accuracy. Other CNN hyper-parameters such as the size and the number of filters in convolutional layers, number of nodes in hidden layers and input-output size were selected such that the resulting architecture (see Table II) gave the best performance on the held-out validation samples. We used rectified linear

unit (ReLU) nonlinearity in all hidden layers, which speeds up training by avoiding vanishing gradient [46], [47].

For training, we extracted 171,000 patches from 12 training images and validated on 45,000 patches from 4 validation images (Table I). We also used dropout with different dropout rates for each layer [47]. We increased the dropout rate while going deeper in the network because a higher dropout rate in the initial layers results in information loss from the image but it acts as a good regularizer in the deeper layers to avoid over-fitting. We sampled an equal number of pixel locations from the three classes – inside, outside, and on the boundaries of nuclei – with uniform probability of selection within a class using annotations for the training images. To ensure correctness in patch sampling for the outside nuclei class, we used a standard protocol of applying a threshold on the average intensity of the color-deconvolved image as proposed in [48]. The rationale is based on the fact that the outside class pixels should not strongly absorb hematoxylin. Sample estimated probability maps for the boundary class are shown in Figure 3, row 5.

The CNN was trained using Torch [49] on an NVIDIA Tesla K20c® graphics processing unit (GPU). Initial learning rate for the network was 0.01 and was decreased by a factor of 10 when the validation error was approximately constant. We trained the CNN for 115 epochs (16.5 hours) as the pixel level classification performance on the validation set saturated beyond that. We used GPU time from AWS EC2 G2® instances for testing. Using GPUs led to a testing time of around 30 seconds for each 1000×1000 image.

C. Post-Processing: From Three-Class to n-Ary Nuclear Maps

Nuclei were seeded (detected) easily by thresholding the *inside* class probability map at 0.5. The seeded nuclei were grown iteratively by pushing the boundary out by one pixel per iteration in each direction. As a seeded region grows, the average *inside* class probability of its contour pixels decreases, while their average *boundary* class probability increases. When the average *boundary* class probability of the pixels on the contour of a region reaches a local maximum for the iterations (assessed by remembering the contour pixels of the last iteration), then we stop the nuclei from growing further. We also stop the nuclei from growing in directions where it would intrude into other growing regions (nuclei), or where its boundary (outside) class probability will decrease (increase). This leads to an anisotropic region growing mechanism where initial markers are grown at different rates in different directions, allowing them to acquire non-circular shapes. Sample segmented nuclei obtained after pre-processing, CNN application, and post-processing are shown in the Figure 3(h).

We term our technique *CNN3* for its use of three classes, and show its training and inference scheme in Figure 2.

VI. Experimental Results

We tested our single CNN-based generalized nucleus segmentation scheme on the 14 test images of size 1000×1000 and compared it with two other open source software for nuclear segmentation and a 2-class CNN-based technique similar to [13], which we call CNN2. Before we discuss our results of comparison with other techniques, we first visualize how each step in our process worked in Figure 3.

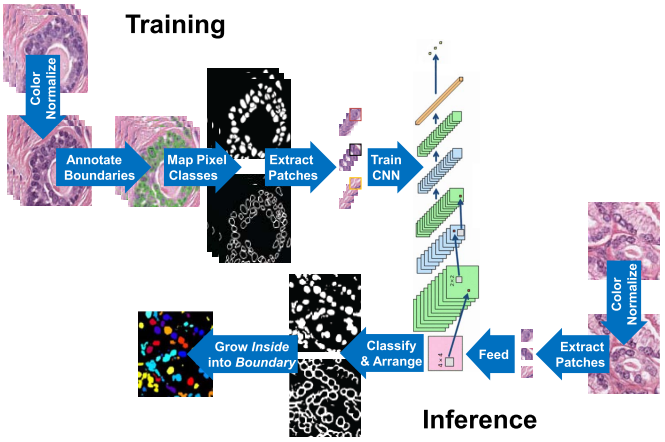


Fig. 2. Proposed training and inference to segment nuclei using CNN3.

A. Visualizing Results Step-by-Step

Cropped versions of the seven test images, one for each organ, are shown in row 1 of Figure 3, including three from the organs not used for training. These images demonstrate the variation in the tissue and nuclear appearances represented in our dataset. The annotations are shown in row 2 (we invite the reader to notice its quality). The ideal (target) class maps for boundary class are shown in row 3 for illustration only, and were not used for producing test segmentation (but the corresponding maps for training images were used to train the CNN). The ideal inside and outside class maps are easy to infer and have been omitted. Row 4 illustrates how color normalization reduces appearance variability of H&E stains [44]. Row 5 shows the estimated boundary class maps. We again omit showing maps for the other two classes. Finally, the segmented nuclei maps are shown in row 6 as n-ary masks to observe which nuclei were merged (under-segmented) or unnecessarily split (over-segmented) and which were not. It is clear that even the crowded and chromatin-sparse nuclei (with bright holes) were correctly segmented using *CNN3*. The generalization power of the proposed method is demonstrated by testing on unseen patients and multiple organs, including some unseen organs.

B. Comparison With Other Open-Source Software and CNN2

For a comparison between ours and another technique based on deep learning, we trained a segmentation system that was very similar to the one proposed by Xing *et al.* [13], which we call CNN2. There was no boundary class of pixels, and thus we sampled equal number of nuclei and non-nuclei pixel centered patches from our annotated training dataset. A CNN with a binary output (inside and outside nucleus) was trained on color normalized patches centered at the labeled pixels, similar to training the CNN with ternary output for CNN3. The CNN

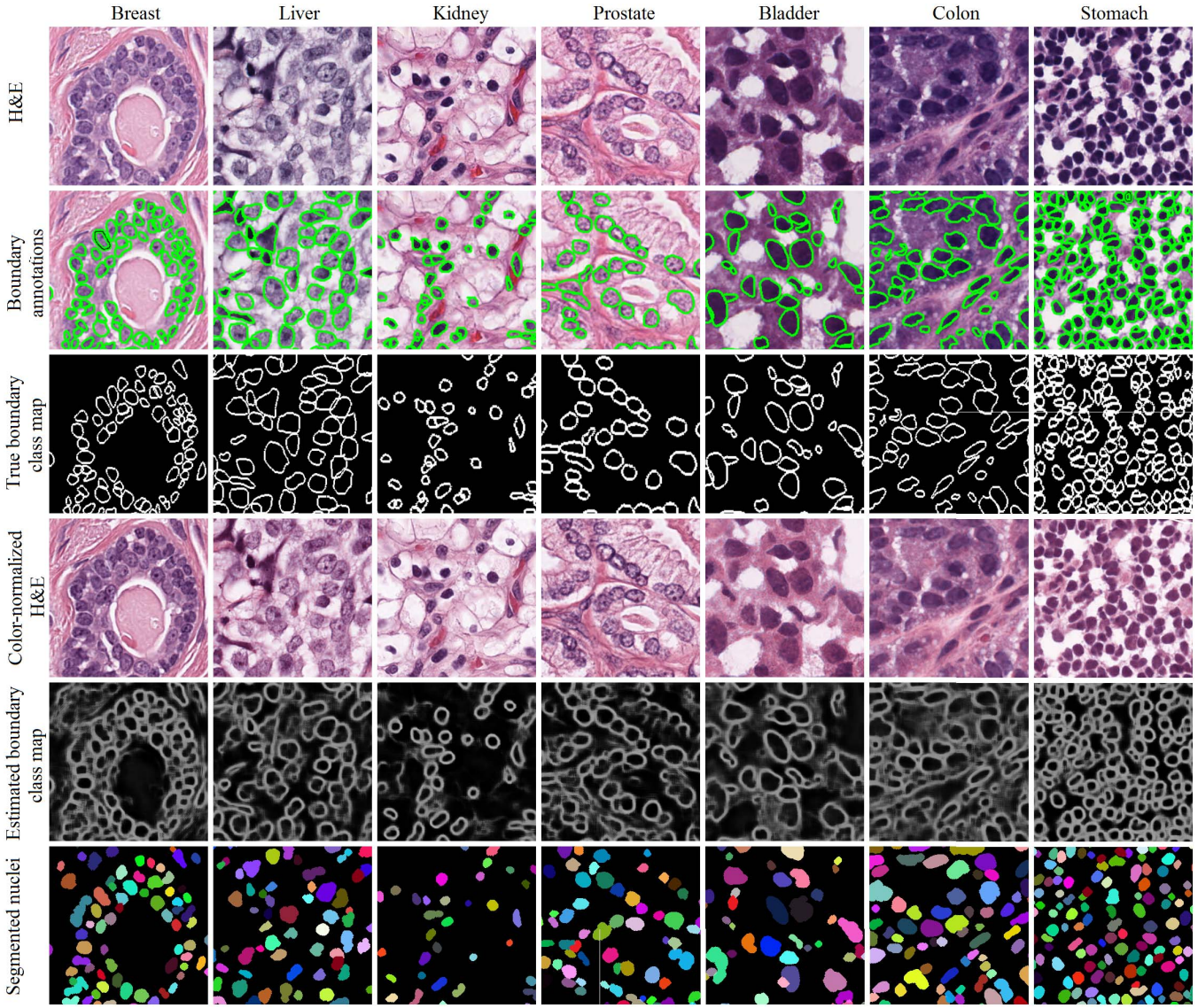


Fig. 3. Examples of sub-images taken from the test images for different organs (columns) showing challenging cases based on variation in nuclear appearance and crowding. Intermediate and final results of our segmentation process are shown in rows. Annotation boundaries of the test images were only used for evaluation.

architecture was very similar for CNN2 and CNN3 (Table II), except that CNN2 had two output nodes instead of three. In our experiments, our architecture for binary pixel classification gave 92% accuracy, while the one proposed in [13] gave only 78% accuracy. Note that, pixel accuracy itself was not the end-goal. It was used to compare (and select) CNN architectures and hyper-parameters.

To facilitate the comparison between CNN2 and CNN3, the nuclear (marker) detection scheme for CNN2 was identical with that of [13], which allowed for seeding of multiple nuclear markers within each segment to avoid under-segmentation. On the inside (nuclear) probability maps obtained using the two class CNN, a binary map was computed by thresholding at 0.5, small (noisy) nuclear regions with less than 30 pixels were removed, and the distance transform followed by H-minima suppression (depth parameter was 2) was computed to obtain the nuclear markers, as was done in [13]. However, [13] used an isotropic region growing mechanism that simply added a layer of pixels to the marker boundaries until the new boundaries touched any neighboring nuclei. This made their intermediate nuclear shapes unnecessarily circular, which they proposed to refine using a complicated shape dictionary-based method whose manual tuning of parameters has not been adequately described and thus requires a separate investigation. Hence, the region growing scheme starting from nuclear markers was kept very similar for CNN2 and CNN3 for fair comparison. We adapted our anisotropic region growing approach (Section V-C) for CNN2 such that the nuclear markers for CNN2 were iteratively grown over the inside class in the binary map to obtain the nuclear segments, without merging two segments that grew to touch each other. Region growing was stopped in the directions where it ran into the outside class, thus allowing for non-circular nuclear shapes. We found empirically that

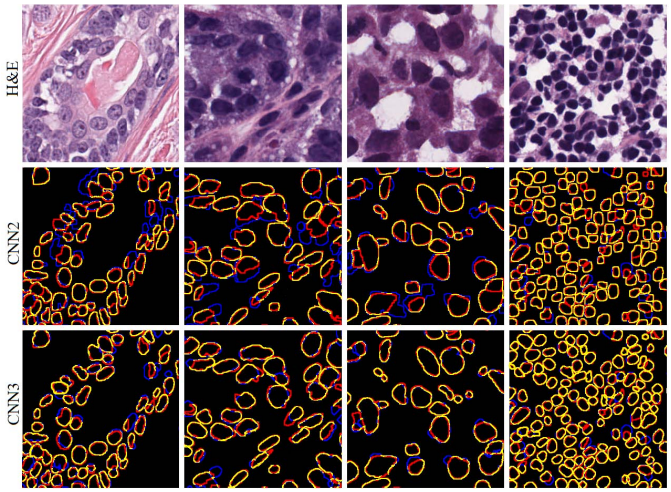


Fig. 4. Comparing segmentation methods based on deep learning on H&E stained images (top row). In the middle and bottom rows, ground truth (annotated) boundaries are red, detected are blue, and overlap between the two are yellow. Predominance of yellow (correct) boundaries can be seen with the use of proposed CNN3 in the bottom row compared to the CNN2 similar to [13] in the middle row. For easy cases where the nuclei are uniformly colored and well-separated, both methods perform equivalently as seen in the rightmost column.

our anisotropic region growing outperformed Xing *et al.*'s isotropic region growing even for the two class case. Hence, we report the results of CNN2 with the nuclei maps obtained using anisotropic region growing from initial markers for a fair comparison with CNN3.

Basically, we are comparing the two methods on their ability to seed the correct nuclear markers using their respective CNN outputs followed by anisotropic region growing with or without boundary class. The results of this comparison are shown in Figure 4, in which it is obvious that while CNN2 works well for chromatin-sparse nuclei, it under-segments crowded nuclei where the third class (boundary pixels) of CNN3 gives a definite advantage in separating them.

We also computed the proposed metric – aggregated Jaccard index AJI (Algorithm 1) – on our segmentation maps for the 14 test images. Additionally, we used two other open source software – Cell Profiler [50] and Fiji (ImageJ) [52] – for comparison. Fiji is a Java-based software that has a watershed transform-based nuclear segmentation plugin available [51]. Cell Profiler is a python-based software with several suggested pipelines for computational pathology [50]. We used a part of one such pipeline which implemented nuclear segmentation by converting an H&E image into gray scale, followed by manual intensity thresholding. We experimented to get the best results by using color normalization, and adjusting the intensity threshold. We also experimented with the advanced routines such as watershed segmentation, available in Cell Profiler but report only the best results obtained. As shown in Table III, both these software resulted in much lower aggregated Jaccard index (AJI). For illustration, we also compared other popular metrics for the four techniques, even though these metrics do not treat object-level and pixel-level errors in a unified framework. AJI for all images was smaller compared to average Dice's coefficient even though both have the same range because the former also penalizes detection errors,

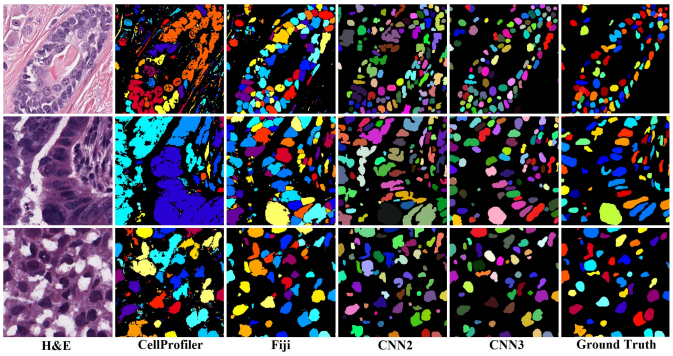
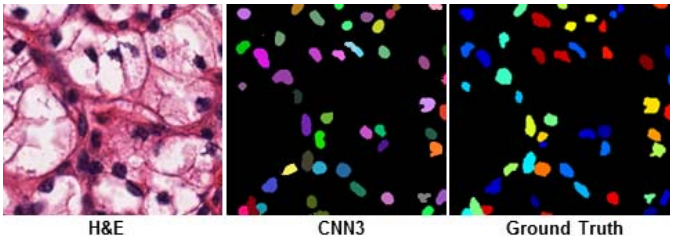
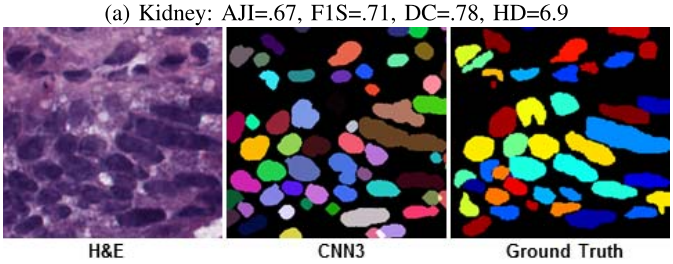


Fig. 5. H&E stained images with comparison of segmentation results using open-source software – cell profiler [50] and Fiji [51], – and deep learning methods CNN2 (similar to [13]), and CNN3 (proposed) – along with ground truth segmentation maps.



(a) Kidney: AJI=.67, F1S=.71, DC=.78, HD=6.9



(b) Colon: AJI=.49, F1S=.71, DC=.77, HD=7.1

Fig. 6. The proposed metric aggregated Jaccard index (AJI) captures the better segmentation quality of (a) Kidney with respect to (b) Colon, while F1-score (F1S), mean DiceâĽ™s coefficient (DC) and mean Hausdorff distances (DC) are very similar for the two images.

while the latter only penalizes segmentation errors computed on the true positives. Similarly, F1-score does not penalize segmentation errors, and is thus higher than AJI. In Figure 5, we compare the four techniques on a subset of the test data. Our results demonstrate the advantages of deep learning based techniques in general, and the proposed *CNN3* in particular, especially for diffused-chromatin and crowded nuclei of breast, prostate, and colon.

C. Utility of the Proposed Metric

We also examined different images with roughly equal F1-scores, mean Dice coefficients, and mean Hausdorff distances but different segmentation qualities. As can be seen in Figure 6, AJI is able to distinguish between such segmentation results. Cropped portions of Kidney image 2 and Colon image 1 (see Table III) are shown in Figure 6 to ensure illustrative clarity. The main reason behind the AJI being better able to capture segmentation quality is its unified treatment of detection and segmentation. For example, in Figure 6, while

TABLE III
PERFORMANCE COMPARISON OF TWO OTHER OPEN SOURCE SOFTWARES – CELLPROFILER (CP) [50] AND FIJI [51] – AND A TWO-CLASS CNN (CNN2) SIMILAR TO [13] WITH THE PROPOSED THREE-CLASS CNN (CNN3)

Organ	Image	AJI (proposed)				Average Hausdorff distance				Average Dice's Coefficient				F1-Score			
		CP	Fiji	CNN2	CNN3	CP	Fiji	CNN2	CNN3	CP	Fiji	CNN2	CNN3	CP	Fiji	CNN2	CNN3
Breast	1	0.0641	0.2772	0.3852	0.4974	11.6527	9.4535	9.4896	8.7632	0.4990	0.5514	0.6101	0.6885	0.1674	0.5739	0.6532	0.7478
	2	0.2353	0.2252	0.4663	0.5796	8.3015	8.2584	7.1423	6.0411	0.5156	0.5957	0.6825	0.7476	0.5758	0.7804	0.8407	0.9149
Liver	1	0.1576	0.2539	0.4086	0.5175	9.2489	9.3594	8.9887	7.6578	0.4778	0.4989	0.5998	0.6726	0.4047	0.6322	0.7835	0.8568
	2	0.2063	0.2826	0.3325	0.5148	7.2725	7.2924	6.4830	6.0651	0.5177	0.5711	0.6188	0.7036	0.7567	0.8161	0.8863	0.9409
Kidney	1	0.0631	0.2429	0.3129	0.4792	7.7820	7.8715	7.2364	6.5476	0.5544	0.5676	0.5928	0.6606	0.2798	0.4522	0.7385	0.7869
	2	0.2838	0.3290	0.5010	0.6672	7.8746	7.7466	7.5797	6.9249	0.6903	0.7089	0.7195	0.7837	0.3449	0.5161	0.6752	0.7132
Prostate	1	0.1576	0.2356	0.2707	0.4914	8.9397	8.9397	8.4306	6.9583	0.6021	0.6915	0.7643	0.8306	0.4872	0.7854	0.8075	0.8717
	2	0.0707	0.1592	0.1848	0.3761	12.1742	12.1742	11.3288	10.1323	0.5135	0.5888	0.6438	0.7537	0.2980	0.5651	0.6819	0.7452
Bladder	1	0.0784	0.3730	0.3498	0.5465	8.7602	8.3941	8.1843	7.3168	0.7615	0.7949	0.8809	0.9312	0.1692	0.7309	0.7693	0.8184
	2	0.0529	0.2149	0.2876	0.4968	12.7769	12.7769	12.5471	11.7198	0.5119	0.5128	0.5483	0.6304	0.3131	0.4234	0.5721	0.7506
Colon	1	0.0102	0.2295	0.3043	0.4891	10.5964	9.7941	9.1577	7.1296	0.5929	0.6381	0.6516	0.7679	0.1045	0.5654	0.5875	0.7136
	2	0.0061	0.2685	0.3125	0.5692	9.3570	9.0281	9.1741	8.4603	0.6556	0.6620	0.6761	0.7118	0.1891	0.5887	0.6581	0.7746
Stomach	1	0.1776	0.3757	0.3961	0.4538	7.3267	7.1828	8.2824	6.7192	0.7313	0.8663	0.8610	0.8913	0.8295	0.9503	0.9601	0.9781
	2	0.1604	0.3586	0.3618	0.4378	7.8166	7.0381	7.6697	6.8247	0.7398	0.8428	0.8491	0.8982	0.7446	0.9280	0.9395	0.9609
Overall		**0.1232**	**0.2733**	**0.3482**	**0.5083**	**9.2771**	**8.9507**	**8.6924**	**7.6615**	**0.5974**	**0.6493**	**0.6928**	**0.7623**	**0.4046**	**0.6649**	**0.7538**	**0.8267**

the proportion of the missed or false detections are roughly the same for both examples, but the area of pixels in those erroneous cases are vastly different. This difference is not captured by the shape quality metrics (mean Dice coefficient and Hausdorff distance) as these metrics are only computed on true detections. Thus, even when the detection and shape metrics are roughly equal, the difference between the overall segmentation quality is obvious, and is captured by AJI.

VII. CONCLUSIONS, FUTURE WORK, AND DISCUSSION

We introduced a large dataset of human tissue images with annotated nuclear boundaries from a diverse set of patients and organs, captured in one of the most widely used setting in digital pathology – H&E stained tissue captured at 40x magnification. We hope that this dataset along with the proposed metric will aid the development and benchmarking of generalized nuclear segmentation techniques. We plan to add more annotations to this dataset over the next few years and invite others to contribute as well.

We also proposed a technique and allowed public access to the software for nuclear segmentation using a CNN. We showed that it gives reasonable results even on organs on which it was not trained, thus demonstrating generalization. We have also released executable and source files for the software that will make it usable right out-of-the-box, and allow its integration as a plugin with more general purpose tools such as Cell Profiler [50] and Fiji [51] that do more than just segmentation and have graphical user interfaces.

While our trained CNN will work for H&E stained 40x digital pathology images, our training code should allow re-estimation of the CNN parameters for other stains and imaging modalities. In case of increase in the magnification, the network architecture will need further changes to scan larger window sizes to accommodate nuclei and their spatial context. A few modifications may be required to accommodate a different number of pixel classes if nuclear classification is also desired.

We hope that this work will contribute to the development of nuclear morphometry and computational pathology software for research and clinical use. Such tools can be used in computational pathology for more effective treatment planning by allowing basic diagnosis, disease subtype identification, triaging, and predicting treatment outcome.

ACKNOWLEDGMENT

The authors thank Yachee Gupta, Trina Naskar, and Niladri Bhatacharya for annotating several thousand nuclei each, Prof. Peter H. Gann for assessing annotation quality, Amazon Web Services® CS Educate program for donating cloud computing credits, and the anonymous IEEE TMI reviewers for suggesting new experiments to strengthen the paper.

REFERENCES

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, Oct. 2009.

[2] H. Llewellyn, "Observer variation, dysplasia grading, and HPV typing: A review," *Amer. J. Clin. Pathol.*, vol. 114, pp. S21–S35, Nov. 2000.

[3] D. N. Louis *et al.*, "Computational pathology: A path ahead," *Arch. Pathol. Lab. Med.*, vol. 140, no. 1, pp. 41–50, 2015.

[4] A. H. Beck *et al.*, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Sci. Transl. Med.*, vol. 3, no. 108, p. 108ra113, 2011.

[5] H. Chang *et al.*, "Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association," *IEEE Trans. Med. Imag.*, vol. 32, no. 4, pp. 670–682, Apr. 2013.

[6] P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Trans. Med. Imag.*, vol. 32, no. 12, pp. 2169–2178, Dec. 2013.

[7] A. Sethi, L. Sha, R. J. Deaton, V. Macias, A. H. Beck, and P. H. Gann, "Computational pathology for predicting prostate cancer recurrence," in *Proc. AACR 106th Annu. Meeting*, Aug. 2015, p. LB-285.

[8] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Proc. 5th IEEE Int. Symp. Biomed. Imag., Nano Macro*, May 2008, pp. 284–287.

[9] J.-H. Xue and D. M. Titterington, "*t*-tests, *F*-tests and Otsu's methods for image thresholding," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2392–2396, Aug. 2011.

[10] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 53, no. 11, pp. 2405–2414, Nov. 2006.

[11] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. W. Pluim, "Automatic nuclei segmentation in H&E stained breast cancer histopathology images," *PLoS ONE*, vol. 8, no. 7, p. e70221, 2013.

[12] A. Vahadane and A. Sethi, "Towards generalized nuclear segmentation in histological images," in *Proc. IEEE 13th Int. Conf. Bioinform. Bioeng. (BIBE)*, Nov. 2013, pp. 1–4.

[13] F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 550–566, Feb. 2016.

[14] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[15] R. Verma, N. Kumar, A. Sethi, and P. H. Gann, "Detecting multiple sub-types of breast cancer in a single patient," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2648–2652.

[16] G. Lee *et al.*, "Co-occurring gland angularity in localized subgraphs: Predicting biochemical recurrence in intermediate-risk prostate cancer patients," *PLoS ONE*, vol. 9, p. e97954, May 2014.

[17] J. Cheng and J. C. Rajapakse, "Segmentation of clustered nuclei with shape markers and marking function," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 3, pp. 741–748, Mar. 2009.

[18] S. Ali and A. Madabhushi, "An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery," *IEEE Trans. Med. Imag.*, vol. 31, no. 7, pp. 1448–1460, Jul. 2012.

[19] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, Apr. 2010.

[20] S. Wienert *et al.*, "Detection and segmentation of cell nuclei in virtual microscopy images: A minimum-model approach," *Sci. Rep.*, vol. 2, Nov. 2012, Art. no. 503.

[21] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and cell splitting," *IEEE Trans. Med. Imag.*, vol. 30, no. 9, pp. 1661–1677, Sep. 2011.

[22] M. E. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4568–4580, Nov. 2012.

[23] M. Zhang, T. Wu, and K. M. Bennett, "Small blob identification in medical images using regional features from optimum scale," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1051–1062, Apr. 2015.

[24] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2421–2433, Oct. 2015.

[25] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review—Current status and future potential," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 97–114, Dec. 2013.

[26] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 234–263, Jan. 2016.

[27] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[28] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. 2, 2009.

[29] K. Soomro, A. R. Zamir, and M. Shah. (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild." [Online]. Available: https://arxiv.org/abs/1212.0402

[30] *HER2 Contest, Her2 Scoring in Breast Cancer Histology Images*, Univ. Warwick, Coventry, U.K., 2016.

[31] K. Sirinukunwattana *et al.*, "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, Jan. 2017.

[32] *DREAM Challenge, The Digital Mammography Dream Challenge*, Univ. Warwick, Coventry, U.K., 2016.

[33] J. Odstrcilik *et al.*, "Retinal vessel segmentation by improved matched filtering: Evaluation on a new high-resolution fundus image database," *IET Image Process.*, vol. 7, no. 4, pp. 373–383, Jun. 2013.

[34] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J. Pathol. Inform.*, vol. 7, p. 29, Jul. 2016.

[35] E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath, "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinform.*, vol. 10, no. 1, p. 368, 2009.

[36] H. Irshad *et al.*, "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd," in *Proc. Pacific Symp. Biocomput.*, 2015, pp. 294–305.

[37] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[38] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[39] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[40] A. Vahadane, N. Kumar, and A. Sethi, "Learning based super-resolution of histological images," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 816–819.

[41] *The Cancer Genome Atlas (TCGA)*, accessed on May 14, 2016. [Online]. Available: http://cancergenome.nih.gov/

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[43] S. Manivannan, W. Li, S. Akbar, J. Zhang, E. Trucco, and S. J. McKenna, "Local structure prediction for gland segmentation," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 799–802.

[44] A. Vahadane *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Trans. Med. Imag.*, vol. 35, no. 8, pp. 1962–1971, Aug. 2016.

[45] A. Sethi *et al.*, "Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images," *J. Pathol. Inform.*, vol. 7, p. 17, Apr. 2016.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, 2010, pp. 807–814.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[48] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, Aug. 2001.

[49] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: A modular machine learning software library," Idiap Res. Inst., Martigny, Switzerland, Tech. Rep. EPFL-REPORT-82802, 2002.

[50] A. E. Carpenter *et al.*, "CellProfiler: Image analysis software for identifying and quantifying cell phenotypes," *Genome Biol.*, vol. 7, no. 10, p. R100, 2006.

[51] F. Dong *et al.*, "Computational pathology to discriminate benign from malignant intraductal proliferations of the breast," *PLoS ONE*, vol. 9, no. 12, p. e114885, Dec. 2014.

[52] J. Schindelin *et al.*, "Fiji: An open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, Jul. 2012.