Multi-Class Cell Detection Using Spatial Context Representation

Shahira Abousamra, David Belinsky, John Van Arnam, Felicia Allard, Eric Yee,
Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, Chao Chen
Stony Brook University
Stony Brook, NY 11794, USA

Abstract

In digital pathology, both detection and classification of cells are important for automatic diagnostic and prognostic tasks. Classifying cells into subtypes, such as tumor cells, lymphocytes or stromal cells is particularly challenging. Existing methods focus on morphological appearance of individual cells, whereas in practice pathologists often infer cell classes through their spatial context. In this paper, we propose a novel method for both detection and classification that explicitly incorporates spatial contextual information. We use the spatial statistical function to describe local density in both a multi-class and a multi-scale manner. Through representation learning and deep clustering techniques, we learn advanced cell representation with both appearance and spatial context. On various benchmarks, our method achieves better performance than state-of-the-arts, especially on the classification task.

1. Introduction

We propose the first joint cell detection and classification method that explicitly learns a spatial-context-aware representation of cells. We demonstrate that incorporating spatial context will significantly improve the performance, especially for the classification task.

Identification of various types of cells such as tumor cells, lymphocytes, and stromal cells from whole-slide histology images is an important step towards automatic diagnosis and prognosis in digital pathology. The spatial arrangement of different cells can comprehensively characterize the interaction between tumor and immune cells and be correlated with clinical outcomes [29, 48, 23]. One good example is the detection and measurement of tumor infiltrating lymphocytes (TILs), i.e., lymphocytes residing within the border of invasive tumors [36]. The prevalence of TILs has been shown to be associated with better clinical outcomes [37, 38, 41]. Aside from lymphocytes, the presence of isolated or small clusters of tumor cells at the invasive tumor front, a phenomenon known as tumor budding, is a prognosis biomarker associated with an increased risk of lymph node metastasis in colorectal carcinoma and other solid malignancies [28]. Other examples include the assessment of lymphovascular invasion and perineural invasion [27] and the identification and measurement of intraepithelial lymphocytes for the diagnosis of celiac disease [34]. All these studies necessitate an effective algorithm to accurately identify cells of different types.

Multi-class cell identification involves both cell detection and cell classification. Cell detection has been studied extensively in the past few decades [40, 21, 45]. Existing approaches either adopt the object detection algorithm from computer vision [20, 47], or treat the problem as an instance segmentation problem and segment nuclei one-by-one [18, 19, 32, 25, 30]. Although segmentation methods provide detailed nuclei morphology, their training requires highly detailed and thus time-consuming nuclei mask annotation. To circumvent this bottleneck, one may use weakly-supervised methods [31, 46, 43, 11] to segment nuclei based only on *point annotations*, i.e., points placed at the centers of nuclei. Point annotation is a much more affordable annotation for large scale training.

Despite the success in the cell detection, our progress on cell classification is not as advanced. Indeed, classification is a challenging task even for human experts. Cells of different kinds can manifest with similar appearance. Meanwhile, cells of the same type may exhibit large variation of morphology and texture in regions of neoplasia and inflammation. To correctly classify cells in such challenging scenarios, pathologists not only use appearance, but also rely on the contextual information of surrounding cells, their spatial relationships and tissue architecture. For example, degenerating or apoptotic cells often cluster together within the luminal spaces of gland forming tumors and can be readily identified in this context despite the spectrum of morphologic features they display; a context which can be identified through cellular architectural patterns in a larger scale. Similarly, architectural patterns can be used to help distinguish reactive stromal cells from tumor cells when their shape and chromatin pattern are indistinguishable.

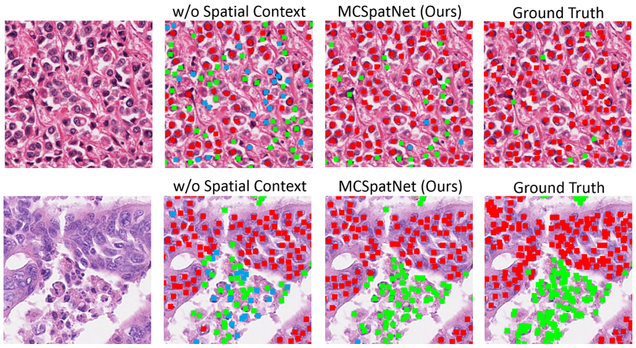

Figure 1. Sample results with and without spatial contexts. Blue, red, and green dots represent inflammatory, epithelial, and stromal cells, respectively. From left to right: original patches, detection and classification results without spatial contexts, our method using spatial contexts, ground truth point annotation. Our method better classify cells using spatial contexts.

To design an ideal classification algorithm, it is essential to imitate a pathologist's thought process and incorporate the spatial context into the decision making. Unlike existing methods which only learns the context implicitly [19, 40], we propose a novel algorithm that explicitly leverages spatial context. To model the spatial context, we introduce the classic Ripley's K-function [15] from spatial statistics [5]. The K-function encodes spatial relationship between cells in a multi-class, multi-scale manner. It has been shown to be a powerful descriptor of cellular architectures [48, 7]. However, existing studies use the K-function mainly for downstream analysis, not for better cell identification as we do.

Our major challenge is that the spatial context is not available at inference time; one cannot infer the K-function without first identifying cells and their classes. To address this challenge, we assume a deep neural network has sufficient learning power and **propose to learn a spatial-context-aware representation through a multi-task learning framework.** We train a deep neural net that jointly performs cell detection, cell classification and spatial context prediction (i.e., predicting K-functions). Through training, the network learns a representation that incorporates both appearance and spatial context. At inference stage, only the detection and classification modules are used. But the spatial-context-aware representation ensures a superior performance. See Figure 1 for an illustration.

To learn with such a multi-task framework is challenging, because the tasks are very different in nature. The cell classification/detection modules predict class labels. Whereas the spatial context prediction module needs to predict K-functions, which are high-dimensional continuous-valued vectors. To better learn with these tasks, we introduce a deep clustering module, inspired by works in unsupervised and weakly-supervised learning [10, 13]. In this module, we introduce pseudo-labels generated by clustering of the deep representation. These pseudo-labels serve as a connection between cell class labels and K-functions, thus facilitate the collaboration of different modules and the fusion of appearance and spatial information. See Figure 2 for an overview of our model.

We apply our method, called MCSpatNet, to a joint cell detection and classification task. We evaluate our method on three benchmark datasets (breast cancer, colorectal cancer and lung cancer), all of which with multi-class point annotations.[1] Our method outperforms various SOTA baselines, demonstrating the power of the spatial-context-aware representation.

To summarize, our contributions are as follows:

- We propose a novel cell detection and classification method which, for the first time, explicitly learns a spatial-context-aware representation of cells via multi-task learning.
- We introduce spatial statistical functions (K-function) as an effective descriptor of cells' spatial context.
- We introduce spatial context prediction module and deep clustering module to facility learning of the representation.
- Our code is available here: https://github.com/TopoXLab/MCSpatNet

1.1. Related Work

Most cell detection from point annotation methods operate by training a regression network to predict a probability density map that has high values at the ground truth points and the probability attenuates as you go further away. The detected cells are then the peaks in the regressed maps [40, 21]. In [45, 24] a patch-wise classification is performed of whether a patch contains a cell nucleus in the center. [20, 47] use faster-RCNN object detection algorithm [33]. One can naturally extend the method by using other similar object detection algorithms [8, 14, 9]. Our proposed method learns to detect cells by pixel-wise binary classification. It is trained with ground truth binary maps where each connected component represents a cell nucleus and the components are restricted to not overlap so as to respect the boundaries between cells.

When nuclei segmentation masks are available, fully supervised instance segmentation methods can be trained to predict nuclei contours [50, 19, 30]. Alternatively, weakly supervised nucleus segmentation methods learn to segment

[1] Our model can naturally be extended to mask annotations. In this paper, we focus on point annotations as they are much more affordable. Indeed, our experiments reveal that our spatial-context-aware model trained with point annotations can be as good as segmentation-based models trained with mask annotations w.r.t. detection and classification.

nuclei using ground truth point annotations [31, 46, 43, 11]. These methods mostly implement some form of detection followed by refinement to get more accurate boundaries.

There are fewer methods that target cell classification, mostly operating in 2-stages. [40] has a 2-stage detection and patch-wise classification, where the patches are small windows centered at the detected cells. [12, 48] perform a 2-stage segmentation and patch-wise classification. [48] further applies spatial kernel smoothing and cancer region detection to improve the classification accuracy. [51] also applies a classification smoothing on top of [40] using conditional random field (CRF) built on cells and superpixels. [49] classifies readily available patches that are coarsely centered around the nuclei. [19] proposes a fully supervised segmentation and pixel-wise cell classification method using 3 network branches for segmentation, boundary detection, and classification tasks.

Deep clustering [10, 3, 26, 22] has been known to enhance feature representation in unsupervised and weakly supervised learning [22, 44, 13]. Mostly, deep clustering methods generate clusters from feature representations and learn to predict the cluster for each input instance [10, 26]. Other methods try to learn clusters that maximize the information across classes [3, 22]. Similar to [10, 13], we iteratively generate pseudo labels for sub-classifying cells by deep clustering. We use features encoding the spatial statistics functions and the class texture features thus combining the spatial and visual contexts for a better representation.

2. Method

We propose a method for joint cell detection and classification on H&E stained images. The prediction and the ground truth are in the form of multi-class point annotations; one point is positioned in the approximate center of each nucleus, with specific cell class labels: inflammatory, epithelial[2], and stromal cells. See Fig. 3 for an illustration.

Our model has several different modules corresponding to different tasks. These modules share the same input and the same feature extractor. But they have their own blocks of convolutional layers for their own different prediction tasks. This way these tasks learn and benefit from a common representation without conflicts.

The architecture is shown in Fig. 2. Aside from the cell detection and classification modules, our method has two additional modules. First, we introduce a spatial distribution prediction module, which learns to predict the cells' associated spatial statistics functions. As a result, it learns to pool features that describe the spatial context. To further improve the feature representation, we introduce a cell-level deep clustering module. The module iteratively cluster cells based on the feature representation and predict their clusters. It integrates both appearance and spatial contexts into a better feature representation.

In Sec. 2.1, we formulate the spatial context information, and illustrate its intuition. In Sec. 2.2, we provide details of all four modules and the common feature extractor.

2.1. Cellular spatial context

We define the spatial context of a cell as the distributions of cells of different classes in its surrounding neighborhood. To describe the spatial context, we introduce **Ripley's K-function**, a spatial statistics function that describes point patterns [15, 5]. For a cell of interest, called the *source*, we can measure the number of neighbors (called targets) within a distance r from the source. Aggregating over all observed points, we have the K-function as the cumulative distribution function that represents the expected number of neighbors within increasing distances r_i from the source. See the illustration in Fig. 4. Formally, given a 2D point set X of size n, the K-function is:

$$K(r) = \frac{1}{\lambda} \sum\nolimits_{s \in X} \sum\nolimits_{t \in X \setminus \{s\}} \frac{1}{n-1} [\![d(s,t) < r]\!] \tag{1}$$

where $d(s,t)$ is the Euclidean distance between the source and target points s and t. $[\![\cdot]\!]$ is the Iverson bracket, which is one if the condition inside it is true, and zero otherwise. λ is the intensity function used to normalize w.r.t. the density of the source points. Depending on the assumption, the intensity function can be a constant, i.e., $\frac{Area}{n}$, (homogeneous setting) or location dependent (inhomogeneous).

We can compare the calculated K-function with the baseline, i.e., the K-function of a random point process (Poisson point process). When the K-function of interest is above the baseline at a certain range of r, we know there are more target points than one would expect from a random point process, and thus the points are clustered. When the K-function is below the baseline, we have less target points than one would expect from a random point process, hence the points are dispersed. For multi-class point sets, we extend the K-function so that the source and target can be from different classes of points. In such case, the K-function is also called the K-cross function.

Cell-specific K-functions and vectorization. In cell detection and classification, we focus on individual cells and inspect the spatial context of each cell rather than at the population level. For a given cell s as the source, we restrict to a fixed-size region surrounding it. We only consider target cells falling within this local region or patch. We consider target cells of different classes one-by-one. For class c, denote by X_s^c all class-c cells within the local patch centered at s. The K-function of class c is

$$K_s^c(r) = \frac{1}{n_{max}} \sum\nolimits_{t \in X_s^c \setminus \{s\}} [\![d(s,t) < r]\!] \tag{2}$$

[2] Our patches are selected from cancerous regions, in which epithelial cells are tumor cells.

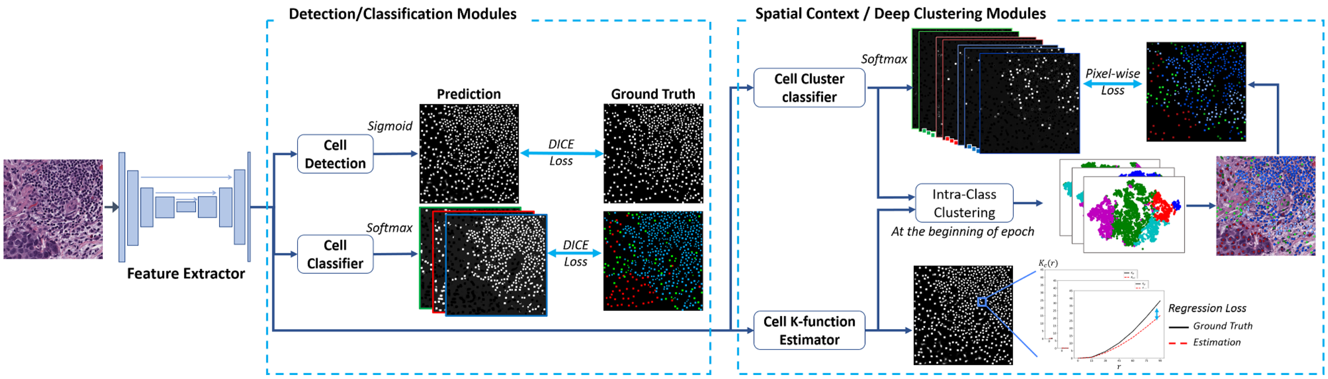

Figure 2. Model architecture. The feature extractor generates shared features for all four following modules: cell detection, classification, spatial context, and deep clustering. Each module has its own layers to generate its own task-specific representation. The spatial context module learns the K-function of each cell. The deep clustering module performs dynamic clustering based on spatial context representations and cell classification representations. The model is trained end-to-end and thus learns a spatial-context-aware feature representation.

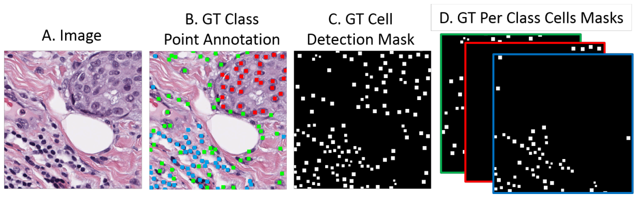
A. Image
B. GT Class Point Annotation
C. GT Cell Detection Mask
D. GT Per Class Cells Masks

Figure 3. Input and ground truth (GT) training maps. (A) Sample training patch. (B) The GT point annotation. (C) The detection GT mask. (D) The classification GT masks. Blue, red, and green indicate inflammatory, epithelial, and stromal cells/channels, respectively.

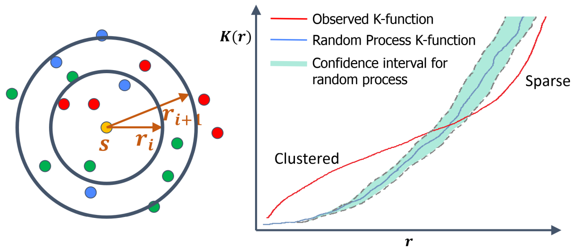

Figure 4. Ripley's K function. Left: The K-function considers the number of neighboring target points (cells) of different classes at increasing radii from a source (cell) s. The sources and targets can each be points from a specific class. Right: Plot of an observed K-function together with the baseline from a Poisson point process. The dashed lines are the confidence interval envelopes for the random process function.

where $n_{max} = \max_s \sum_c |X_s^c|$ denotes the maximum number of target cells for any local patch.

In practice, we set the patch size to 180×180. We also vectorize the K-function by uniform sampling at a finite set of radii: $r = 15, 30, 45, \ldots, 90$ pixels. In total, we have three classes, each with six dimension. We call this 18 dimensional vector the **K-function vector** of s. By learning to predict this K-function vector, our model learns the spatial representation.

K-functions and cells' spatial behavior. We finalize this subsection by providing real examples to illustrate how K-functions can help refine cells' spatial representation. In Fig. 5, we visualize different classes of cells and their K-functions. Since K-function is high dimensional, we cannot directly show their values. Instead, we cluster cells of each class into sub-categories based on their K-function vectors, and then visualize different sub-categories with different colors (from dark blue to light blue for inflammatory cells, from dark green to light green for stromal cells, from dark red to light red / pink for epithelial cells).

We observe that different sub-categories clearly exhibit distinct spatial behavior. For epithelial cells, cells clustered within a tumor nest often belong to the pink sub-category, whereas those that are more dispersed and closer to stromal cells belong to the red sub-category. For stromal cells, the ones from the light green sub-category are often closer to other classes (inflammatory or epithelial cells). Meanwhile, other stromal cells that are far from inflammatory/epithelial cells tend to belong to the dark green sub-category. Similar behavior can be observed for inflammatory cells: light blue cells are clustered while dark blue cells are dispersed. Moreover, Fig. 6, shows the average K-functions for the different pairs of source and target cell classes. It is obvious that different pairs of cells classes exhibit different spatial behavior.

This demonstrates that K-functions do stratify cells into sub-categories with distinct spatial behavior. It motivates us to learn the spatial representation via K-functions.

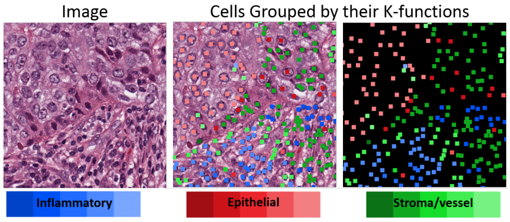

Figure 5. Visualization of cells with distinct K-functions. We group cells into sub-categories based on their K-function vectors. Different sub-categories exhibit different spatial behavior. Please note that the choice of darker to lighter colors does not imply a sequential relationship between sub-categories.

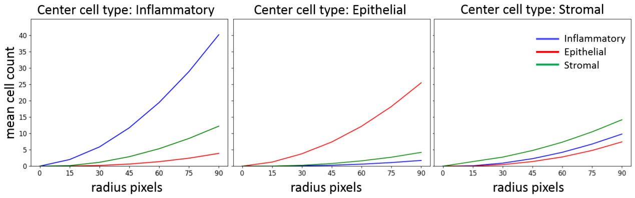

Figure 6. Average K-functions when the centers are inflammatory (left), epithelial (center), and stromal cells (right). In each plot, blue, red, and green curves represent different neighboring cell types.

2.2. Multi-task learning for a spatial-context-aware representation

Our proposed model has four modules for four different tasks: cell detection, cell classification, deep clustering, and spatial context prediction. The four modules share a common feature extractor that learns a shared feature representation. The deep clustering module and the spatial context prediction module are only used in training. They help learn a spatial-context-aware representation which improves the performance of cell detection and classification. Please see Fig. 2 for the model architecture.

Feature extractor. The feature extractor is a variant of U-Net model [35] with a VGG-16 [39] backbone. The output of the feature extractor has 96 channels and the same spatial resolution as the input. The output feature representation is shared by all four task-specific modules.

Each of the task-specific modules also has its own block of convolutional layers. For all the task-specific blocks, the spatial resolution of the input and output are the same as the input image. But the number of output channels are different for different tasks. This allows different task-specific blocks to tune the features to better suit their own tasks without conflict. The tasks are very different; involving both classification and regression. The deep clustering task is very dynamic in nature. Without its own task-specific block, it will destabilize the feature extractor and affects other tasks negatively.

Next, we explain different task-specific modules.

Cell detection and classification modules. For cell detection, the model predicts a single channel likelihood map over all pixels and compare it with a binary ground truth mask. The ground truth mask is generated by dilating the point annotation slightly. For cells close to each other, we use smaller dilation radii to avoid overlapping. Each connected component in the mask corresponds to one cell. See Fig. 3 for an illustration. The output of the cell-detection block has a single channel with sigmoid activation. The training for this module uses DICE loss as it tends to preserve small objects in the image.

For the classification task, we create a similar ground truth mask as the detection ground truth mask, except that pixels in each connected component has a specific cell class. The cell-classification block output has three channels, corresponding to the three cell classes. We use a softmax activation and use DICE loss for training.

During inference, we only use the detection and classification branches. We threshold the cell-detection block output and use the centroids of all connected components as the predicted cell locations. We classify each cell using the cell-classification block output at its predicted location.

Spatial context prediction module. The spatial context prediction module predicts cell-specific K-function vectors (as defined in Eq. 2). The intuition is that by predicting spatial context, the learnt feature representation is spatial-context-aware and can help detection and classification modules. This will be verified in the experimental section via an ablation study.

We predict an 18 dimensional K-function vector for each cell. For ground truth, we use the ground truth cell detection mask generated for the cell detection task. For each positive pixel, i.e., the pixels within the dilated components, we compute the K-function vector. Our spatial-context-prediction block output has the same spatial resolution as the input image, and has 18 channels corresponding to the 18 dimensional K-function vectors. No additional activation is needed. Only the predictions at the positive pixels are compared with the ground truth K-function vectors.

To compare two K-function vectors, we can use the Kolmogorov–Smirnov test [6] for comparing sample estimates of cumulative distribution functions. Formally, it is the supreme norm of the difference between the prediction and ground truth:

$$KM(K_{pred}, K_{gt}) = \sup_r |K_{pred}(r) - K_{gt}(r)|.$$

In practice, we found out that the supreme norm is less efficient and we use L1 norm as a surrogate loss.

Deep clustering module. The spatial context prediction module learns representations of spatial context. However, in practice, we observe that it does not collaborate well with the detection and classification modules. We hypothesize that the appearance feature and the spatial context feature do not fuse well. This could be due to the very different natures of different tasks. The detection and classification tasks are per-pixel classification tasks with a small number of classes. Whereas the spatial context prediction task is a regression problem of high dimensional output.

In order to re-calibrate the appearance features and spatial features learnt from these different tasks, we propose a deep clustering module, which plays the role of self-supervision and lies in the mid-ground between the other modules. The deep clustering module executes a per-pixel classification task, but with a higher number of pseudo-classes derived from both spatial and appearance information. Deep clustering is known to enhance feature representation [10, 13] especially for unsupervised and weakly-supervised tasks. This also fits our setting; we do not have a golden standard of sub-categories of cells. Instead, we derive the sub-classes dynamically based on the feature representation and train the model to predict them.

In particular, we apply k-means clustering on the intermediate feature representation to obtain pseudo-sub-classes that further divide cells of each class. Each class of cells is stratified into 5 pseudo-sub-classes. The deep-clustering block then learns to predict these pseudo labels. The design of this block is similar to the cell classification module. We note that the intermediate feature representation for clustering is a concatenation of the features from the deep-clustering block and the spatial-context-prediction block. This ensures that the clustering, and thus the learnt feature representation, integrates both spatial and appearance information in a well-calibrated way.

The clustering and derived pseudo-sub-class labels are re-generated on the start of every epoch. When running k-means clustering, the clusters' centroids are initialized with the previous epoch's centroids to avoid excess jumping between cluster mapping. The training is very similar to the cell classification task, using DICE loss.

The overall loss of our model is the weighted sum of losses from all four modules:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{Det}} + \lambda_2\mathcal{L}_{\text{Class}} + \lambda_3\mathcal{L}_{\text{Spatial}} + \lambda_4\mathcal{L}_{\text{DeepCluster}}.$$

In practice, we simply set all weights to one.

Technical details. We finalize this subsection with a few more details of our model architecture. The feature extractor has VGG-16 encoder and has 4 decoder blocks. The last decoder block is a single deconvolution layer with 96 output channels. The detection, classification, deep clustering, and spatial context blocks, all have a similar architecture: two 3×3 convolutions with 64 channels output, followed by 1×1 convolution to give the final output. Each convolution except for last is followed by a ReLU activation. More details are in Section 3 and supplementary material.

3. Experiments

We evaluate our method, MCSpatNet, on three datasets of different cancer types: breast cancer, lung cancer, and colorectal cancer. The breast cancer dataset, BRCA-M2C, consists of 120 patches belonging to 113 patients, collected from TCGA [42]. The lung cancer dataset, SEER-Lung, is a collection of 57 patches from the SEER cohort [16]. Each patch is sampled from different whole slide images or tissue samples to maximize the generalizability. The colorectal cancer dataset, Consep [19], is publicly available. It has 41 patches. The patches from all 3 datasets are of size $\approx 500 \times 500$ at 20x magnification; they are large enough to provide spatial context. The lung and breast cancer datasets are annotated by pathologists with ground truth points at approximate centers of cells with an associated class: inflammatory, epithelial, or stromal. It is worth mentioning that all the epithelial cells in these patches are tumor cells. The Consep dataset additionally has nuclei contour masks. Details of these datasets are in the supplemental material.

Implementation details of MCSpatNet. We train the model on patches taken at 20x magnification, which is around 0.5 microns per pixel. We use dilated ground truth dot masks with dilation up to 9 pixels In all datasets, the model is trained on the 3 major classes: inflammatory, epithelial, and stromal cells. For the deep clustering, we use kmeans. Each cell class has k=5 clusters (15 in total).

To train the spatial context module, we use an R-package to generate the ground truth K-functions with border correction. These K-functions only need to be computed once before training starts. It takes around 0.1 min per patch. The radius r's range=$[0, 90]$ with step=15 pixels. To get the formula as in Eq. 1, the returned K-function is multiplied by $\frac{\text{number of cells in region}}{area \times n_{max}}$, where n_{max} is a constant set to 100. During inference, we apply a threshold=0.5 on the detection output and get rid of tiny connected components with area less than 5 pixels. The location of the predicted cells are then the centroids of the resulting connected components.

Evaluation. We train and evaluate our method against state-of-the-arts on all three datasets. We split each training dataset into train and validation sets and evaluate all methods on the same split. The performance is evaluated with the F-score metric. We report the F-scores on the detection and classification tasks. Similar to existing approaches [19, 40], we say a predicted cell is true positive (TP) if it is within 6 pixels (around 3 microns) from one of the ground truth points. Otherwise it is a false positive (FP). A ground truth point is false negative (FN) if it does not have a nearby prediction. F-score is then calculated as $\frac{TP}{TP+0.5(FP+FN)}$.

The detection F-score is computed over *all* detected cells

Table 1. Results on all three datasets. For each dataset, we report five scores: F-scores for individual classes (inflammatory, epithelial, stromal), mean F-score over all three classes (Mean), and detection F-score over all cells (Det.). For each score, the best method is highlighted with bold fonts.

	BRCA-M2C					Consep					SEER-Lung				
Method	Infl.	Epi.	Stro.	Mean	Det.	Infl.	Epi.	Stro.	Mean	Det.	Infl.	Epi.	Stro.	Mean	Det.
U-Net	0.498	0.744	0.476	0.572	0.838	0.681	0.613	0.561	0.618	0.724	0.779	0.809	0.571	0.720	0.856
Faster-RCNN [33]	0.572	0.718	0.490	0.594	0.806	0.259	0.523	0.446	0.410	0.492	0.192	0.769	0.411	0.457	0.616
Cascade RCNN [8]	0.564	0.708	0.505	0.592	0.796	0.644	0.633	0.515	0.597	0.682	0.710	0.793	0.454	0.653	0.759
PointSeg[31]	0.249	0.407	0.300	0.319	0.538	0.104	0.603	0.210	0.306	0.435	0.748	0.773	0.537	0.686	0.848
HoverNet-Weakly	0.582	0.702	0.513	0.599	0.817	0.549	0.377	0.365	0.431	0.518	**0.823**	0.808	0.535	0.722	0.846
HoverNet [19]	-	-	-	-	-	0.633	0.651	0.624	0.636	0.730	-	-	-	-	-
MCSpatNet	**0.635**	**0.785**	**0.553**	**0.658**	**0.849**	**0.724**	**0.695**	**0.628**	**0.682**	**0.762**	0.799	**0.817**	**0.600**	**0.739**	**0.860**

Table 2. Ablation study on the 3 datasets. The metrics and the highlight convention are the same as in Table 1.

	BRCA-M2C					Consep					SEER-Lung				
Method	Infl.	Epi.	Stro.	Mean	Det.	Infl.	Epi.	Stro.	Mean	Det.	Infl.	Epi.	Stro.	Mean	Det.
U-Net	0.498	0.744	0.476	0.572	0.838	0.681	0.613	0.561	0.618	0.724	0.779	0.809	0.571	0.720	0.856
U-Net+Deep Clus.	0.593	0.763	0.505	0.620	0.851	0.613	0.644	0.504	0.587	0.714	0.768	0.804	0.576	0.716	0.849
U-Net+Spat. Pred.	0.597	0.771	0.507	0.625	**0.853**	0.682	0.650	0.522	0.618	0.761	0.787	0.809	0.587	0.728	0.851
Using NN Dist.	**0.641**	0.698	0.447	0.595	0.829	0.656	0.616	0.568	0.613	0.719	0.752	0.808	0.558	0.706	0.848
Using Density	0.563	0.753	0.525	0.614	0.851	0.686	0.687	0.624	0.666	**0.764**	0.797	0.807	0.572	0.725	0.850
MCSpatNet	0.635	**0.785**	**0.553**	**0.658**	0.849	**0.724**	**0.695**	**0.628**	**0.682**	0.762	**0.799**	**0.817**	**0.600**	**0.739**	**0.860**

regardless of their class. The classification F-score is evaluated on cells of each class (inflammatory F-score, epithelial F-score and stromal F-score). We also report the mean F-score over the three classes as an overall metric for the classification performance.

Baselines. We compare with several SOTA methods which can jointly segment/detect and classify cells. **U-Net** is a baseline method for joint detection and classification. It uses a U-Net architecture with VGG-16 backbone. It is essentially our method (MCSpatNet) without the spatial prediction and deep clustering modules. We also compare against SOTA computer vision multi-class detection algorithms, e.g., **Faster-RCNN** [33] and **Cascade RCNN** [8].

Aside from detection based methods, segmentation-based methods can also be applied. Since the ground truth nuclei masks are not available, we apply a SOTA weakly supervised nuclei segmentation method (**PointSeg**) [31], which only requires point annotation for training. To classify the segmentation results, we train a CNN classifier (**SSPP**) [40] on local patches enclosing each predicted cell segment. We also apply **HoverNet** [19], a SOTA joint segmentation and classification method. Training HoverNet requires fully annotated nucleus masks, which is only available for Consep dataset. We only apply the full HoverNet on Consep. Meanwhile, for all three datasets, we apply a weakly-supervised HoverNet (**HoverNet-Weakly**) by training on pseudo-masks acquired from point-annotation-anchored superpixels. More details can be found in the supplemental material.

Results and discussions. Table 1 shows the results of our method and all baselines on all three datasets. Our method consistently outperform all baselines on different metrics. Generally, we observe a bigger advantage in the classification task (inflammatory, epithelial, stromal F-scores and mean F-score) than in the detection task (Det. F-score). This is expected as spatial context is used by pathologists to help determine cell classes in ambiguous situation, whereas appearance is the main cue for detection.

Figure 7 shows the qualitative results of our method and different baselines. In all patches, the appearance of cells are not very discriminating and baseline models are unable to correctly classify them. Our model is able to differentiate epithelial vs. stromal cells and epithelial vs. inflammable cells. More results are in the supplemental material.

Ablation study. We evaluate our proposed method against a few variations to show the efficacy of different components. In all the runs we use the same experimental setting as above. We first evaluate the efficacy of the proposed two novel modules: spatial prediction and deep clustering. To this end, we evaluate 3 baselines: U-Net (the model with both modules removed), U-Net plus deep clustering module, and U-Net plus spatial prediction module. See Table 2 for the results. Comparing U-Net with the second and third baselines, we observe both modules contribute to performance improvement. Furthermore, the full version of our model, MCSpatNet, outperforms all three baselines. This

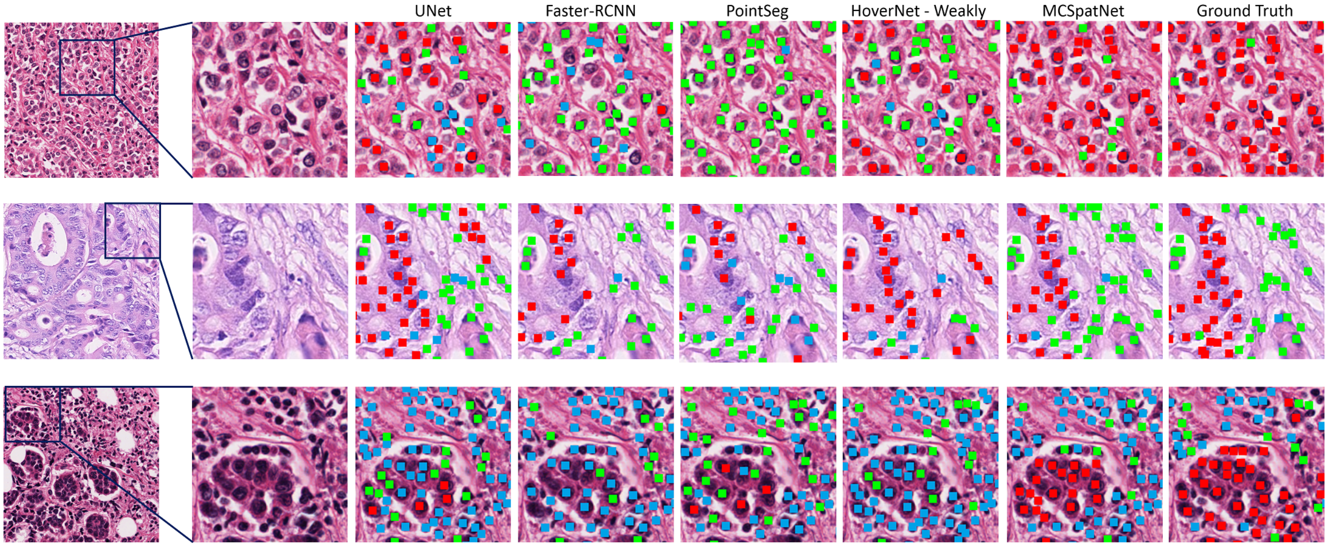

Figure 7. Qualitative Results. Blue, red, and green dots represent inflammatory, epithelial, and stromal cells, respectively.

establishes the necessity of both modules.

To further investigate the efficacy of the proposed K-function, we compare with using different spatial context descriptors: nearest neighbor distance function and density function. For the first baseline, in the spatial prediction module, we replace the K-function vector with the distance to the nearest neighbor cell of each of the 3 classes. For the second baseline, we replace the K-function vector with the density of the cells from each of the 3 classes within the neighborhood area. As shown in Table 2, K-function based spatial prediction outperforms both of these baselines. This demonstrates that the K-function provides richer spatial information and helps the model learn a spatial-context-aware representation that best suits the cell detection and classification tasks.

More ablation studies are in the supplemental material.

Statistical significance. To verify that the benefit of our method is robust, we ran our method and two top baselines: UNet and HoverNet-Weakly, for three more splits on BRCA-M2C (Table 3). Our method is consistently the best over all scores. Furthermore, we ran a paired T-test comparing our method with these baselines. We highlight a baseline result if it is not statistically significantly different from ours (i.e., p-value > 0.05). These "close second" results often have much lower average, but high standard deviation; they are very unstable.

4. Conclusion

In this paper, we propose a novel method for joint cell detection and classification. The novel contribution is to explicitly introduce spatial context information and train the model to learn a spatial-context-aware cell representation of cells. We also use a deep clustering method to better recalibrate the spatial and appearance features. The proposed method outperforms different SOTA methods, demonstrating the significance of spatial context.

Table 3. Average and standard deviation of scores on four random splits of BRCA-M2C.

	MCSpatNet	U-Net	Hovernet-Weakly
Infl.	$\mathbf{0.71 \pm 0.07}$	$\mathbf{0.58 \pm 0.22}$	0.54 ± 0.04
Epi.	$\mathbf{0.76 \pm 0.03}$	0.72 ± 0.02	$\mathbf{0.55 \pm 0.19}$
Stro.	$\mathbf{0.56 \pm 0.08}$	0.51 ± 0.09	0.49 ± 0.08
Mean	$\mathbf{0.67 \pm 0.04}$	$\mathbf{0.60 \pm 0.10}$	0.53 ± 0.06
Det.	$\mathbf{0.85 \pm 0.02}$	0.84 ± 0.01	0.78 ± 0.03

In the future, we will extend the method to other forms of contextual information, e.g., topology [4, 2], and other stainings, e.g., multiplex immunohistochemistry [1, 17].

Acknowledgement. We thank anonymous reviewers for their constructive feedback. This work was partially supported by grants NSF IIS-1909038, CCF-1855760, NCI 1R01CA253368-01, NCI UH3-CA225021 and NCI U24CA215109 as well as generous donor funding from Betsy and Bob Barton.

References

[1] Shahira Abousamra, Danielle Fassler, Le Hou, Yuwei Zhang, Rajarsi Gupta, Tahsin Kurc, Luisa F Escobar-Hoyos, Dimitris Samaras, Beatrice Knudson, Kenneth Shroyer, Joel Saltz, and Chao Chen. Weakly-supervised deep stain decomposition for multiplex ihc images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 481–485. IEEE, 2020. 8

[2] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *AAAI*, 2021. 8

[3] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. 3

[4] Andrew Aukerman, Mathieu Carrière, Chao Chen, Kevin Gardner, Raúl Rabadán, and Rami Vanguri. Persistent homology based characterization of the breast cancer immune microenvironment: a feasibility study. In *36th International Symposium on Computational Geometry (SoCG 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020. 8

[5] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, 2015. 2, 3

[6] Vance W. Berger and YanYan Zhou. *Kolmogorov–Smirnov Test: Overview*. American Cancer Society, 2014. 5

[7] Joshua A Bull, Philip S Macklin, Tom Quaiser, Franziska Braun, Sarah L Waters, Chris W Pugh, and Helen M Byrne. Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. *Scientific reports*, 10(1):1–12, 2020. 2

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 7

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2021. 2

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018. 2, 3, 6

[11] A. Chamanzar and Y. Nie. Weakly supervised multi-task learning for cell detection and segmentation. In *IEEE International Symposium on Biomedical Imaging*, 2020. 1, 3

[12] Y. H. Chang, G. Thibault, O. Madin, V. Azimi, C. Meyers, B. Johnson, J. Link, A. Margolin, and J. W. Gray. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 3

[13] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 6

[14] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting RCNN: On awakening the classification power of faster RCNN. In *European Conference on Computer Vision*, 2018. 2

[15] Philip M Dixon. R ipley's k function. *Wiley StatsRef: Statistics Reference Online*, 2014. 2, 3

[16] Kemi M. Doll, Alfred Rademaker, and Julie A. Sosa. Practical Guide to Surgical Data Sets: Surveillance, Epidemiology, and End Results (SEER) Database. *JAMA Surgery*, 153(6):588–589, 06 2018. 6

[17] Danielle J Fassler, Shahira Abousamra, Rajarsi Gupta, Chao Chen, Maozheng Zhao, David Paredes, Syeda Areeha Batool, Beatrice S Knudsen, Luisa Escobar-Hoyos, Kenneth R Shroyer, Dimitris Samaras, Tahsin Kurc, and Joel Saltz. Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images. *Diagnostic pathology*, 15(1):1–11, 2020. 8

[18] S. Graham, D. Epstein, and N. Rajpoot. Dense steerable filter cnns for exploiting rotational symmetry in histology images. *IEEE Transactions on Medical Imaging*, 39(12):4124–4136, 2020. 1

[19] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 1, 2, 3, 6, 7

[20] Jane Hung, Allen Goodman, Deepali Ravel, Stefanie Lopes, Gabriel Rangel, Odailton Nery, Benoît Malleret, Francois Nosten, Marcus Lacerda, Marcelo Ferreira, Laurent Renia, Manoj Duraisingh, Fabio Costa, Matthias Marti, and Anne Carpenter. Keras r-cnn: library for cell detection in biological images using deep neural networks. *BMC Bioinformatics*, 21:300, 2020. 1, 2

[21] Henning Höfener, André Homeyer, Nick Weiss, Jesper Molin, Claes F. Lundström, and Horst K. Hahn. Deep learning nuclei detection: A simple approach can deliver state-of-the-art results. *Computerized Medical Imaging and Graphics*, 70:43–52, 2018. 1, 2

[22] Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *International Conference on Computer Vision*, 2019. 3

[23] M. Kavianpour, M. Saleh, and J. Verdi. The role of mesenchymal stromal cells in immune modulation of COVID-19: focus on cytokine storm. *Stem Cell Research & Therapy*, 11(1):404, 09 2020. 1

[24] Mina Khoshdeli, Richard Cong, and Bahram Parvin. Detection of nuclei in H&E stained sections using convolutional neural networks. In *IEEE-EMBS International Conference on Biomedical and Health Informatics*, 2017. 2

[25] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017. 1

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. *International Conference on Learning Representations*, 2021. 3

[27] Peng Li, Hao-Qiang He, Chong-Mei Zhu, Yi-Hong Ling, Wan-Ming Hu, Xin-Ke Zhang, Rong-Zhen Luo, Jing-Ping Yun, Dan Xie, Yuan-Fang Li, et al. The prognostic significance of lymphovascular invasion in patients with resectable gastric cancer: a large retrospective study from southern china. *BMC Cancer*, 15:370, 2015. 1

[28] Alessandro Lugli, Inti Zlobec, Martin D Berger, Richard Kirsch, and Iris D Nagtegaal. Tumour budding in solid can-

cers. *Nature Reviews Clinical Oncology*, 18(2):101–115, 2021. 1

[29] Sidra Nawaz and Yinyin Yuan. Computational pathology: Exploring the spatial dimension of tumor ecology. *Cancer letters*, 380(1):296–303, 2016. 1

[30] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38(2):448–459, 2019. 1, 2

[31] Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Gregory M. Riedlinger, Subhajyoti De, and Dimitris N. Metaxas. Weakly supervised deep nuclei segmentation using points annotation in histopathology images. In *International Conference on Medical Imaging with Deep Learning*, 2019. 1, 3, 7

[32] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M. Rajpoot. Micro-Net: A unified model for segmentation of various objects in microscopy images. *Medical Image Analysis*, 52:160–173, 2019. 1

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems*, 2015. 2, 7

[34] Marie E Robert, Sheila E Crowe, Lawrence Burgart, Rhonda K Yantiss, Benjamin Lebwohl, Joel K Greenson, Stefano Guandalini, and Joseph A Murray. Statement on best practices in the use of pathology as a diagnostic tool for celiac disease. *The American journal of surgical pathology*, 42(9):e44–e58, 2018. 1

[35] O. Ronneberger, P.Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 5

[36] Roberto Salgado, Carsten Denkert, S Demaria, N Sirtaine, F Klauschen, Giancarlo Pruneri, S Wienert, Gert Van den Eynden, Frederick L Baehner, Frederique Pénault-Llorca, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international tils working group 2014. *Annals of oncology*, 26(2):259–271, 2015. 1

[37] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports*, 23(1):181–193, 2018. 1

[38] M. Shibutani, K. Maeda, H. Nagahara, T. Fukuoka, Y. Iseki, S. Matsutani, S. Kashiwagi, H. Tanaka, K. Hirakawa, and M. Ohira. Tumor-infiltrating Lymphocytes Predict the Chemotherapeutic Outcomes in Patients with Stage IV Colorectal Cancer. *In Vivo*, 32(1):151–158, 2018. 1

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[40] K. Sirinukunwattana, S. E. A. Raza, Y. 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 Transactions on Medical Imaging*, 35(5):1196–1206, 2016. 1, 2, 3, 6, 7

[41] Sasha E. Stanton and Mary L. Disis. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *Journal for ImmunoTherapy of Cancer*, 4(1), 2016. 1

[42] TCGA development team. The Cancer Genome Atlas. `https://tcga-data.nci.nih.gov/docs/publications/tcga/`. 6

[43] Kuan Tian, Jun Zhang, Haocheng Shen, Kezhou Yan, Pei Dong, Jianhua Yao, Shannon Che, Pifu Luo, and Xiao Han. Weakly-supervised nucleus segmentation based on point annotations: A coarse-to-fine self-stimulated learning strategy. In *Medical Image Computing and Computer-Assisted Intervention*, 2020. 1, 3

[44] Xiang Wang, Huimin Ma, and Shaodi You. Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes. *Neurocomputing*, 381:20–28, 2020. 3

[45] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, 35(1):119–130, 2016. 1, 2

[46] Inwan Yoo, Donggeun Yoo, and Kyunghyun Paeng. Pseudoedgenet: Nuclei segmentation only with point annotations. In *Medical Image Computing and Computer-Assisted Intervention*, 2019. 1, 3

[47] S. Yousefi and Y. Nie. Transfer learning from nucleus detection to classification in histopathology images. In *IEEE International Symposium on Biomedical Imaging*, 2019. 1, 2

[48] Yinyin Yuan, Henrik Failmezger, Oscar M. Rueda, H. Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F. Schwarz, Christina Curtis, Mark J. Dunning, Helen Bardwell, Nicola Johnson, Sarah Doyle, Gulisa Turashvili, Elena Provenzano, Sam Aparicio, Carlos Caldas, and Florian Markowetz. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science Translational Medicine*, 4(157):157ra143–157ra143, 2012. 1, 2, 3

[49] L. Zhang, Le Lu, I. Nogues, R. M. Summers, S. Liu, and J. Yao. Deeppap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1633–1643, 2017. 3

[50] Yanning Zhou, Omer Fahri Onder, Qi Dou, Efstratios Tsougenis, Hao Chen, and Pheng-Ann Heng. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In *International Conference on Information Processing in Medical Imaging*, 2019. 2

[51] Konstantinos Zormpas-Petridis, Henrik Failmezger, Shan E Ahmed Raza, Ioannis Roxanis, Yann Jamin, and Yinyin Yuan. Superpixel-based conditional random fields (SuperCRF): Incorporating global and local context for enhanced deep learning in melanoma histopathology. *Frontiers in Oncology*, 9:1045, 2019. 3