Whole-Slide Image Focus Quality: Automatic Assessment and Impact on AI Cancer Detection

Timo Kohlberger[1], Yun Liu[1], Melissa Moran[1], Po-Hsuan Cameron Chen[1], Trissia Brown[2], Jason D. Hipp[1,3], Craig H. Mermel[1], Martin C. Stumpe[1,4]

[1]Google Health, Palo Alto, CA, USA, [2]Work done at Google Health via Advanced Clinical, Deerfield, IL, USA, [3]Current Affiliation: AstraZeneca, Gaithersburg, MD, USA, [4]Current Affiliation: Tempus Labs, Chicago, IL, USA

Abstract

Background: Digital pathology enables remote access or consults and powerful image analysis algorithms. However, the slide digitization process can create artifacts such as out-of-focus (OOF). OOF is often only detected on careful review, potentially causing rescanning, and workflow delays. Although scan time operator screening for whole-slide OOF is feasible, manual screening for OOF affecting only parts of a slide is impractical. **Methods:** We developed a convolutional neural network (ConvFocus) to exhaustively localize and quantify the severity of OOF regions on digitized slides. ConvFocus was developed using our refined semi-synthetic OOF data generation process and evaluated using seven slides spanning three different tissue and three different stain types, each of which were digitized using two different whole-slide scanner models ConvFocus's predictions were compared with pathologist-annotated focus quality grades across 514 distinct regions representing 37,700 35 μm × 35 μm image patches, and 21 digitized "z-stack" WSIs that contain known OOF patterns. **Results:** When compared to pathologist-graded focus quality, ConvFocus achieved Spearman rank coefficients of 0.81 and 0.94 on two scanners and reproduced the expected OOF patterns from z-stack scanning. We also evaluated the impact of OOF on the accuracy of a state-of-the-art metastatic breast cancer detector and saw a consistent decrease in performance with increasing OOF. **Conclusions:** Comprehensive whole-slide OOF categorization could enable rescans before pathologist review, potentially reducing the impact of digitization focus issues on the clinical workflow. We show that the algorithm trained on our semi-synthetic OOF data generalizes well to real OOF regions across tissue types, stains, and scanners. Finally, quantitative OOF maps can flag regions that might otherwise be misclassified by image analysis algorithms, preventing OOF-induced errors.

Keywords: Computer-aided diagnostics, digital pathology, focus quality, out-of-focus, quality control

INTRODUCTION

Digital pathology is advancing into clinical workflows,[1,2] enabled by the recent regulatory approval of the first whole-slide image (WSI) scanner for primary diagnosis in the U.S.,[3] as well as wider availability of cheaper storage and large technical infrastructure to manage gigapixel-sized image files. Digitization has several compelling use cases: archiving, telepathology for remote consults or diagnosis, teaching, and increasingly, facilitation of powerful image analysis algorithms.[4,5] The process of digitization, however, can add artifacts to the imaging process, including color or contrast variations and out-of-focus (OOF) areas. These artifacts, particularly OOF areas, can hinder the rendering of accurate diagnoses by pathologists,[1,6] or impact the accuracy of automated image analysis.[7]

We here distinguish between three general categories of OOF in decreasing severity: *global OOF* that affects the entire slide, *regional OOF* that affects larger tissue patches, and *local OOF* that affects individual cells or subcellular structures [Figure 1]. Global OOF can be caused by the whole-slide scanner erroneously focusing on coverslip debris. Regional OOF can be particularly prominent in specific tissue types (such as fat), or affect an entire section of the slide (e.g., "scan lane"). Finally, local OOF is common because typical tissue sections of approximately 5 μ are thicker than

Address for correspondence: Dr. Timo Kohlberger,
Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA, USA.
E-mail: tkohlber@google.com

How to cite this article: Kohlberger T, Liu Y, Moran M, Chen PC, Brown T, Hipp JD, *et al*. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. J Pathol Inform 2019;10:39.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2019/10/1/39/272775

the depth of field of standard digital pathology scanners at high magnification. While a potential solution is the use of multiple focal depths ("z-stacking"), such techniques currently increase scan time and file sizes to an impractical extent for routine use.

The different OOF categories have different impacts on the clinical workflow. Pathologists or histotechnicians may flag images with global OOF as low quality and order a rescan, which potentially results in reporting and workflow delays. Regional and local OOF can be much more difficult to detect consistently, particularly when reviewing a slide at low magnification. Rescans may only be requested after the pathologist has spent significant time reviewing other areas in the slide, before proceeding to evaluate specific areas of concern at higher magnification. Importantly, a technician can prescreen all digital slides for global OOF, but that would be impractical for manual review for smaller OOF regions. OOF image artifacts can have even more severe consequences in

automated image analysis by directly impacting detection and classification. Some studies found that systematic errors can be attributed to suboptimal focus quality, such as OOF germinal centers being mistaken for tumor metastases by an algorithm.[7]

While every WSI scanner has built-in focus evaluation that can be used for automatic rescans of the affected regions or for quality reporting,[8] there are several shortcomings in existing methods: (1) despite built-in auto-focus control, most currently available WSI scanners still produce scans with focus issues,[9,10] (2) focus evaluation methods across scanners are different, inhibiting comparison across devices, (3) focus metrics are not typically exported to the user with detail regarding the spatial distribution of focus quality, and (4) evaluation does not take the clinical relevance of the focus quality into account. For example, cytology diagnoses that more strongly rely on subcellular details require a higher focus quality than diagnoses that are based primarily on
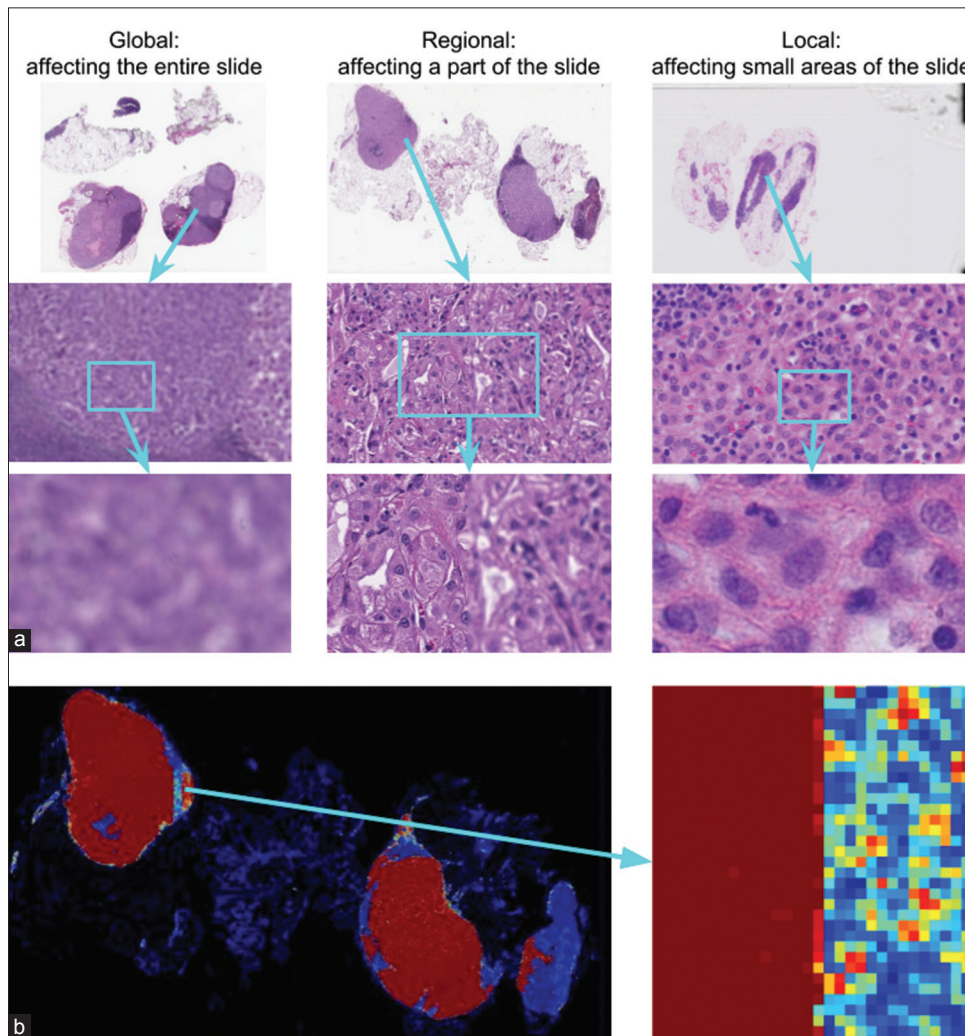

Figure 1: Examples of different types of OOF. (a) Examples of different types of OOF: "global" affecting the whole slide, "regional" affecting a large expanse on a slide, and "local" affecting small tissue or cellular areas. (b) Example of the OOF scan lane in the regional OOF above (middle), that causes a striking artifact in a leading cancer detection algorithm. Left: the heatmap visualization of cancer predictions; black indicates nontissue-containing regions; and other colors range from blue (nontumor) to red (tumor). The OOF scan lane is predicted to be nontumor. Right: a zoom in of the OOF scan lane boundary. OOF: Out-of-focus

tissue architectural patterns, such as prostatic adenocarcinoma Gleason grading.

A sizable body of work exists for automated focus quality assessment in microscopy images using manually engineered image features.[11-15] Deep convolutional neural network approaches, on the other hand, learn the discriminative features and yield higher accuracies,[7,16-20] for disease diagnosis and tissue classification for image quality classification in particular.[21-23] The main challenge, particularly for the neural networks approaches, is the availability of training data to help generalization across a large variety of tissue morphologies and stain properties. To help generalization, previous works have generated simulated OOF examples using real in-focus images through synthetic Gaussian blurring,[23] and additional perturbations such as brightness perturbations and artificial sensor noise.[22]

This work further improves the synthetic data generation approach by more closely mimicking the image acquisition process of real OOF artifacts. These improvements allowed us to build a robust, highly sensitive and fine-grained OOF predictor (ConvFocus) that distinguishes a large spectrum of focus degrees across different tissue and stain types. Our approach provides a generally applicable metric that is highly concordant with manually annotated focus quality and provides information about focus quality across every region in a WSI. Moreover, we quantify the sensitivity of a leading tumor detection algorithm to focus quality, which has not been shown in prior related work.

METHODS

Overview

Given a gigapixel-sized image of a whole pathology slide, our goal was to automatically detect and grade OOF regions at an accuracy level matching that of a pathologist across tissue, biopsy and stain types. To that end, we employed a convolutional neural network approach, which has shown superior performance over more classical machine learning approaches using hand-crafted features.[23] We term our approach ConvFocus [Figures 2 and 3].

Training data

Developing accurate neural networks requires large volumes of training data. Unfortunately, accurate delineation and grading of OOF areas in WSIs is a highly time-intensive task. For example, our pathologist spent 8–16 h per slide to find, delineate, and grade OOF regions (see the "Test Data" section for more details). Consequently, generating a sufficiently large pool of OOF training examples purely by manual annotation is impractical. Instead, we first manually assessed whether patches were completely in-focus or not. For each patch that was labeled "in-focus" by three independent raters, we then synthetically generated multiple additional OOF patches with varying but known degrees of OOF. The real in-focus image patches and the synthetically "de-focused" versions served as training examples. We next describe this process in detail.

Whole-slide scans used for training

26,526 different slides spanning a large variety of tissue, biopsy and stain types were used to train and validate our convolutional neural network. Of these, 8,135 slides had surgical sectioning site information available in a structured form [Table 1]. 26,099 slides were scanned with an Aperio AT2 (Leica Biosystems, Germany) at × 40 (pixel size: 0.251 μm × 0.251 μm), using a semi-automatic scan mode, and 427 slides were scanned using a Hamamatsu NanoZoomer XR (Hamamatsu Photonics, Japan) also at × 40 (pixel size: 0.227 μm × 0.227 μm).

Manual annotation of in-focus image patches

From each of the aforementioned WSIs, 300 × 300 pixel-sized patches at × 40 were sampled from tissue-containing regions, and subsequently evaluated to be either in-focus or OOF by three independent human raters. First, a "tissue mask" was generated by applying the criterion $0.0 < Y \leq 0.8$. Y was defined as the average intensity and computed from the RGB values: $Y = 0.212 \times R + 0.715 \times G + 0.072 \times B$. Next, 300 × 300 pixel-sized patches were randomly sampled from tissue-containing areas. Of those, the top eight (for the AT2) and top 24 (for the NanoZoomer XR) patches with the lowest average luma value were picked for the manual assessment of whether the patch was in-focus. More patches were selected from NanoZoomer XR images because fewer images from that scanner were available in our database, thus this was necessary to achieve a more balanced training set with respect to scanner. This procedure was designed to select the densest tissue parts, which in most cases provided the highest density of sharp image gradients and best exposed OOF issues.

Table 1: Surgical site information for a 8135 slide subset of the training slides (for which this was available)

| Surgical site	# slides	Surgical site	# slides
Prostate, seminal vesicles	3857	Submandibular gland	51
Breast	1622	Fallopian tube, ovary	48
Uterus	664	Ovary	47
Thyroid	524	Kidney	23
Colon	499	Uterus, fallopian tube, ovary	22
Lymph node	239	Products of conception	21
Tonsil	149	Appendix	20
Vas deferens	144	Jejunum	12
Placenta	93	Fetus	10
Duodenum	87	Liver	8
Bladder	87	Stomach, esophagus	8
Stomach	87	Aorta	3
Omentum	83	Testicle	3
Ileum	79	Spermatic cord	2
Esophagus	67	Vulva	2
Breast, lymph node	59	Gallbladder	1
Colorectal	58	Tonsil, uvula	1
Skin	54	Parotid	1
Fallopian tube	53		
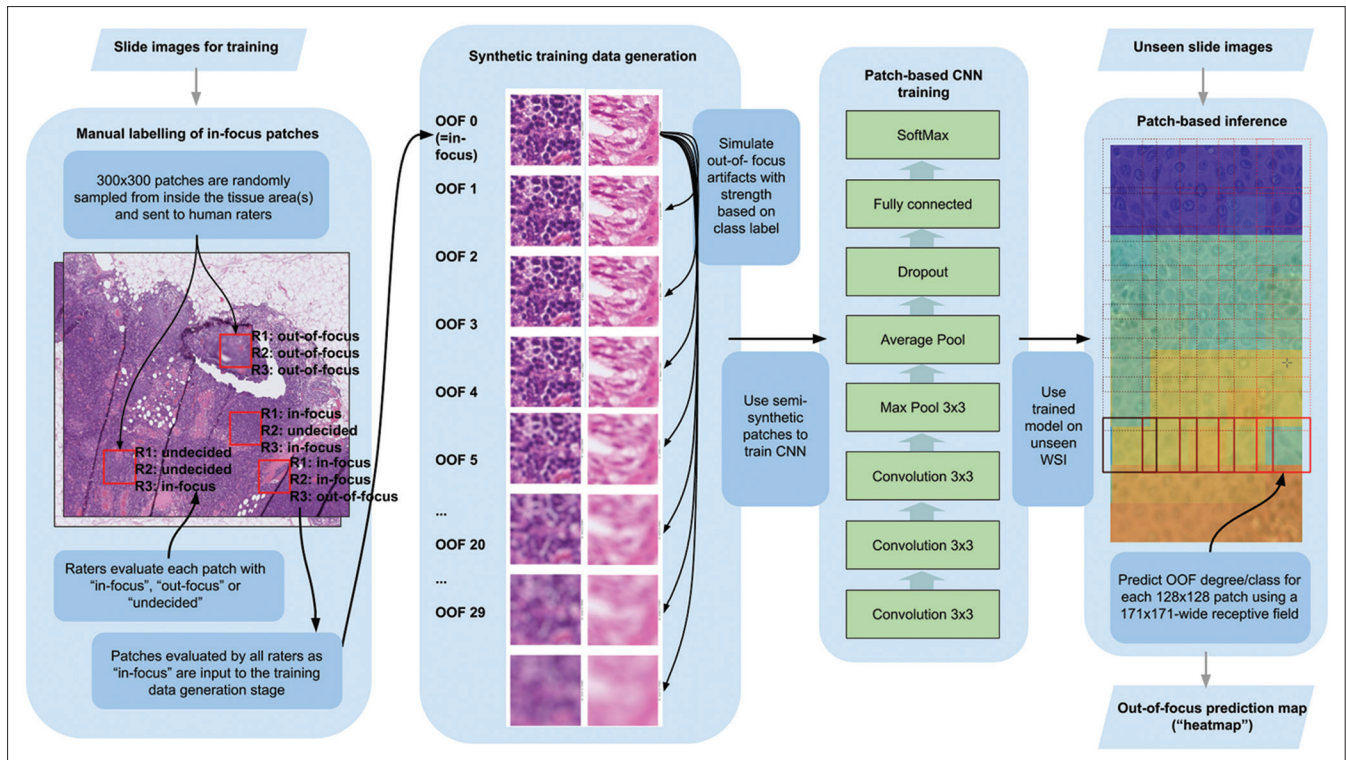
Figure 2: Overview of labeling and convolutional neural network approach to automated out-of-focus grading: ConvFocus

These "densest" patches were then sent to trained nonpathologist raters for independent evaluation. Raters applied one of the following labels to each patch: "in-focus" if at least 75% of tissue area was in-focus, "out-focus" if at least 75% was OOF to any degree, and "undecided" otherwise. We selected approximately 166,000 patches that all three raters labeled as "in-focus" (i.e., consensus rating) to create semi-synthetic training data.

Semi-synthetic training data generation

For each rater-determined in-focus patch (which is labeled with OOF class 0), we synthetically generated 29 additional pseudo-OOF versions with increasing blur magnitude, labeled as OOF classes 1–29. Classes 1–28 represented fine-grained, exponentially increasing OOF magnitudes, and class 29 covered a large range of strong blurring artifacts. Early empirical results indicated that linearly increasing blurring magnitudes for classes 1–28 yielded inferior detection results.

We explored two kinds of blur filters: either a convolution with a 2D Gaussian kernel of size σ - as was done by prior work - or, alternatively, with a two-dimensional (2D) heaviside step function of radius r (both measured in pixels). The latter simulates "bokeh" defocus blur[24,25] which, based on qualitative observations, were found to resemble more the OOF artifacts observed in scanned WSIs. "Bokeh" defocus blur artifacts are well-known in photography, where an OOF point light source generates an approximately homogeneously illuminated image of the lense aperture, rather than a 2D Gaussian distribution.

Specifically, to generate a blurred patch for an intermediate OOF class $c \in [1, 28]$ a σ-value for Gaussian blurring was

chosen randomly from the interval $[0.926 \times \exp((c-1) \times 3/28)$, $0.926 \times \exp(c \times 3/28)]$, while the radius r for simulated bokeh blurring was chosen from $[1.4 \times \exp((c-1) \times 3/28)$, $1.4 \times \exp(c \times 3/28)]$. Strong OOF patches for class 29 were generated by randomly choosing Gaussian σ-values from the interval $[0.926 \times \exp(3), 132]$, and bokeh radii r from $[1.4 \times \exp(3), 200]$, respectively. The rationale for randomly sampling from class-specific intervals over just using each interval's center value was to help the network capture a wider range of blur magnitude during training. In the formulae above, the scaling factors 0.926 and 1.4 and the maximum values 132 and 200 were calibrated to yield similar blur strengths for either method or the same OOF class. Specifically, these scaling factors were determined by minimizing the sum of squared differences between RGB values of the images over all OOF classes [Figure 4].

Postblurring Joint Photographic Experts Group and noise artifacts

In initial experiments, convolutional neural networks trained on Bokeh or Gaussian-blurred synthetic examples yielded poor prediction accuracy on real OOF images; erroneously predicting almost all OOF test patches as in-focus (see Results). We hypothesized that this was caused by the artificial smoothing removing other types of real artifacts. For example, grid-like artifacts at the edges of scan lanes and Joint Photographic Experts Group (JPEG) blocks are smoothed out in artificially blurred images, but would be present in real OOF images. Thus, several categories of other artifact types were re-added after synthetic blurring [Figure 5].
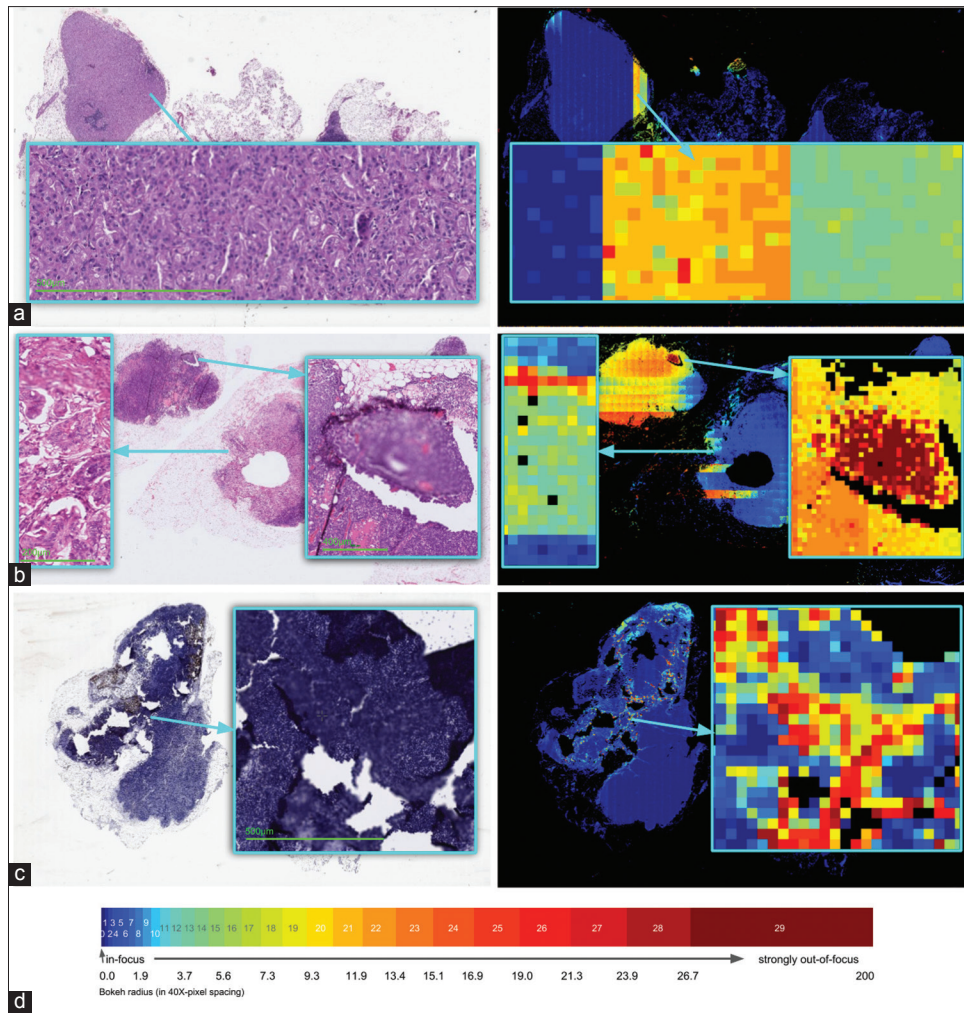
Figure 3: Sample predictions from ConvFocus. Left: WSI of a lymph node biopsy from a colon cancer case, exhibiting both regional (along scan lanes) and local OOF at varying degrees. Right: predicted OOF classes expressed using a "jet" colormap, which ranges from blue (in-focus) to red (strongly OOF), as illustrated in panel (d). (a) Left: WSI of the lymph node from Figure 1b, exhibiting regional OOF artifacts. Right: predicted OOF classes. (b) Lymph node biopsy from colon cancer case (H and E). (c) Lymph node biopsy from breast cancer case (immunohistochemistry stain). (d) Color-coding of predicted OOF degrees and mapping to disk radii used in Bokeh blurring. WSI: Whole-slide image; OOF: Out-of-focus

Visual comparison at high magnification of synthetically blurred real in-focus images to real OOF revealed other artifacts beyond OOF: pixel noise, likely from the image sensor; and JPEG compression artifacts, likely from the lossy JPEG compression applied by scanners postdigitization (scanner quality settings in our data ranged from 70 to 90). In synthetically blurred images however, both of these artifacts ranged from faint to absent, depending on the synthetic blurring magnitude, even if they were present in the original in-focus input images. This is because both pixel noise and JPEG artifacts typically consist of high spatial frequencies, which are diminished by blurring.

Consequently, simulated JPEG compression artifacts were added back into the synthetically blurred images, implemented through JPEG encoding and decoding and an image quality parameter chosen between 70% and 90%. In addition, pixel noise was added because image sensors collect Poisson-distributed noise. Pixel-wise poisson noise was simulated separately for

each color channel value $x_c \in [0,1], c \in \{R, G, B\}$, by applying the mapping $f(x_c) = P(x_c / S) \times S$ to each, where P is the probability mass function of the Poisson distribution, and S is a parameter that inversely controls the signal-to-noise ratio. As the latter was observed to vary significantly between different scanners and objective magnifications, the noise portion during training was varied by randomly sampling S from the interval (0.01, 64.0) for each training patch.

Test data

The OOF algorithms were evaluated using pathologist-graded real OOF artifacts (as opposed to synthetic OOF in the training set). Specifically, three prostate resection and four lymph node biopsy slides were used, including hematoxylin and eosin (H and E) and immunohistochemistry stains. Each WSI was scanned using two scanners at × 40: A Leica Aperio AT2 (pixel size: 0.251 μm × 0.251 μm) using its semi-automatic mode, and a Hamamatsu NanoZoomer S360

scanner (pixel size: 0.230 μm × 0.230 μm) using its fully automatic mode. This resulted in two test sets: seven WSIs per scanner.

A pathologist then manually and nonexhaustively identified, delineated, and graded in-focus and OOF regions, using integral grades ranging from 0 (in-focus) to 6 (very strong OOF), [Figure 6]. Half-grades (e.g., 1.5) were occasionally used when the degree of OOF was interpreted as between integral grades. Annotations were corrected by the pathologist as desired (blinded to algorithm predictions), to help achieve grading consistency across the different tissue and stain types, and scanner models. The delineated regions were rasterized using a 128 × 128 pixel-spaced grid to enable direct comparison with the predictions of the OOF classifier (see below). To ensure label purity, only patches completely contained within a delineated OOF region were used. These annotations produced a total of 37,715 patches.

In addition to pathologist annotations, we collected "z-stack" scans to further evaluate ConvFocus performance on a set of images that have real OOF with a consistent pattern. We digitized a lymph node biopsy slide with z-axis increments of 0.4 μm, and spanning +4 μm to −4 μm (relative to the scanner-determined in-focus depth), [Figure 7a]. This produced a total of 21 WSIs of the same glass slide.

Convolutional neural network-based algorithm
Network architecture
Our convolutional neural network architecture was a series of convolutions and pools, equivalent to the Inception (v3) architecture[7,26] truncated at a lower layer ("MaxPool_3a_3 × 3"[26]) and with a reduced number of filters per layer ("depth_multiplier" = 0.1) to reduce computation [Figure 2]. For this task, we empirically observed that neither modification deteriorated performance (Results). The last layer was attached to a 30-way classification layer to predict the 30 OOF classes. To handle the large image sizes of each WSI, we adopted a patch-based approach. Small crops of 139 × 139 pixels-sized patches (corresponding to approximately 35 μm × 35 μm at × 40) for each WSI were used for training. To help improve performance, we applied data augmentation techniques described in,[7] randomly perturbing the orientation, brightness, contrast, hue, saturation, and adding random translational jitter. The network was trained using the softmax cross-entropy loss function, using the same learning rate schedule and other hyperparameters as previously described.[7]

Inferring out-of-focus heatmaps on whole-slide images
To obtain algorithm predictions of OOF for every region on the slide, we applied ConvFocus in a sliding window fashion across each WSI to produce algorithm predictions for every 128 × 128 patch at × 40. This stride was chosen to match the stride used for cancer detection (see below), and is adjustable. The final predicted OOF class for each patch was the class with maximal probability ("activation") in the final layer of

the network. To visualize these predictions, we used the "jet" colorspace, ranging from blue for class 0 (in-focus) to red for class 29 (strongly OOF) [Figure 3].

ConvFocus evaluation
Each algorithm's performance was evaluated by comparing the predicted OOF class (in the range 0–29) to the corresponding pathologist-annotated OOF grades (0–6 in half-point increments, 13 grades overall) among all patches with annotations within the two test sets (one for each scanner: AT2 and S360). However, the numbers of patches across the 13 different pathologist-assigned grades were unevenly distributed within each test set, and the distributions differed between the two test sets. Therefore, for each test set and grade, 3000 patches were randomly sampled with replacement to obtain evenly distributed classes.

Similar to prior work,[23] the Spearman's rank correlation coefficient was used as the main evaluation metric. To assess the statistical significance of the correlation, we used a two-sided test.[27] Next, we computed a linear regression to evaluate deviation of the intercept from 0.0, since an "in-focus" patch (annotated grade: 0.0) should ideally be predicted as "in-focus" (ConvFocus class: 0). The regression's slope was used to assess deviations from 4.83 (=29/6), which reflects a linear mapping of the ConvFocus OOF class range of (0, 29) to the graded OOF range of (0, 6).

In addition, ConvFocus was further evaluated on the z-stack test set by plotting the predicted OOF class against the z-level for a number of patches. Two aspects were qualitatively assessed. First, we checked for the presence of a "v"-shaped plot indicating that predicted OOF class increases (towards poorer focus) as the z-level goes further from in-focus (in either direction). Second, we checked if the "in-focus" plane (z = 0) was generally predicted as the lowest OOF class.

Measuring cancer detection algorithm performance as a function of focus quality
After validating ConvFocus, we used the algorithm to study the impact of OOF (both real and synthetic) on the performance of a cancer detection algorithm on the publicly available Camelyon 2016 challenge test dataset,[16] consisting of 80 non-tumor and 48 tumor slides with pixel-level annotations of tumor locations. In the first experiment (real OOF), we employed the current best-performing breast cancer metastasis detector "LYNA,"[7,28] which we configured to provide binary predictions at the same granularity as the OOF classifier. That is, for each image patch, predictions from both the breast metastasis and ConvFocus were available. Next, we stratified the patches by predicted OOF class. We then measured the LYNA algorithm's performance for patches in each OOF class, merging the higher OOF classes to ensure sufficient numbers of both tumor and nontumor patches. In a second experiment, we added artificial Bokeh blur at a selected strength to all patches in the test set, ran LYNA on all patches, and evaluated LYNA's performance. We repeated this for a
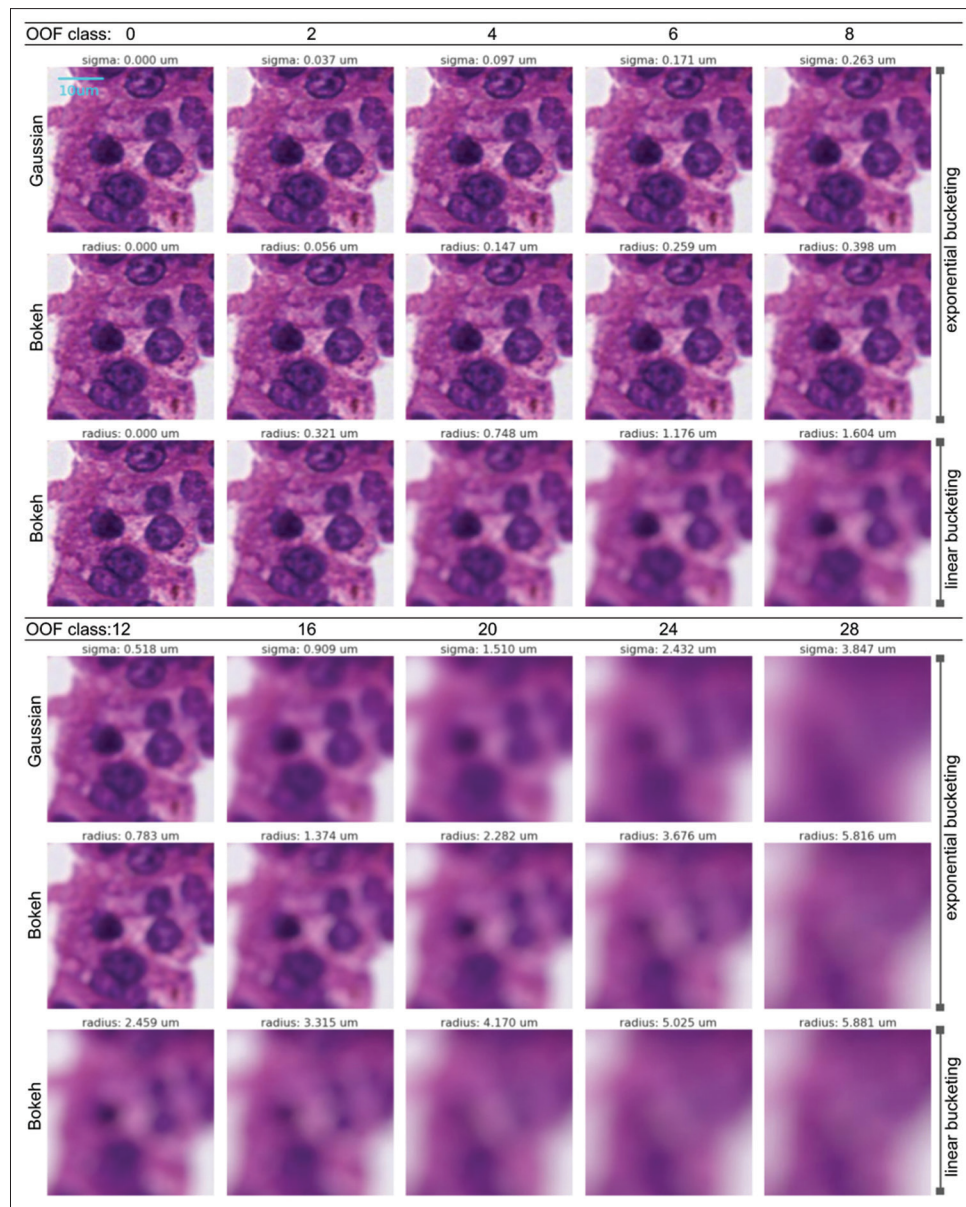
Figure 4: Examples of different degrees of synthetic OOF used for training. Gaussian or Bokeh blurring were applied at increasing levels to an in-focus image (see images at top left with σ =0 and radius = 0). To maintain visually similar blur degrees for the same OOF class label, the σ values and Bokeh disk radii were aligned by minimizing the sum of squared differences between the blurred images. OOF: Out-of-focus

range of Bokeh-blur strengths. In both experiments, we used the patch-level area under receiver operating characteristic curve (AUC[29]). Confidence intervals for the patch-level AUC were computed using the nonparametric bootstrap approach,[30] with $n = 500$ samples at the slide level.

Results

ConvFocus evaluation

Sample visualizations of ConvFocus's predictions are shown in Figures 3 and 8, depicting regional OOF patterns such as OOF scan lanes and OOF tissue edges. Quantitative evaluation using a total of 37,715 pathologist-graded OOF patches is reported in Table 2 and visualized in Figure 9, showing high Spearman

rank correlations >0.8 ($p < 0.001$) between pathologist-graded OOF and ConvFocus predictions for slides scanned on both scanners tested: AT2 and S360.

Each step in our proposed semi-synthetic data generation approach was derived using qualitative assessments of whether regions with strong OOF were appropriately detected [Figure 8]. To quantify the improvements derived from each step, we also performed ablation experiments using several configurations [Table 2]. Using Bokeh synthetic blurring instead of Gaussian yields only a small improvement in correlation metrics, but a large increase in the regression's slope parameter values. Since the synthetic Gaussian and Bokeh blurring magnitudes were visually and
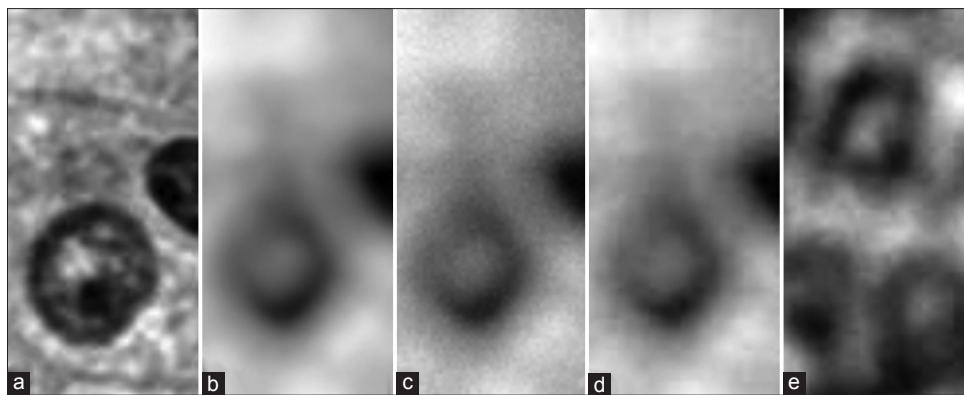
Figure 5: Detailed comparison of real OOF to simulated OOF images. (a) An in-focus image of a lymphocyte, shown in grayscale to improve visibility of artifacts: particularly, grid-like artifacts from JPEG compression and Poisson noise from the sensor. (b) When synthetically generating OOF training examples, just applying Gaussian or Bokeh blurring alone smooths these artifacts. (c) Adding Poisson noise and (d) JPEG artifacts restores these artifacts, resulting in more realistic images. (e) An real OOF image of the same cell type for comparison. OOF: Out-of-focus; JPEG: Joint Photographic Experts Group
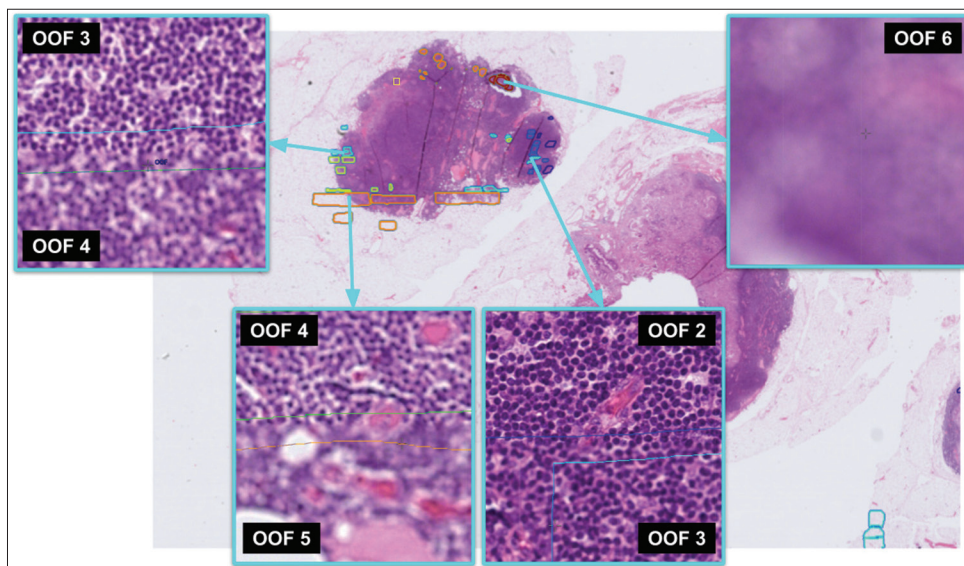

Figure 6: Examples of OOF annotations by pathologist. The pathologist identified, delineated and graded the regions highlighted in this lymph node from a colon cancer case (scanned at × 40 on Leica AT2). Colors from the jet palette indicate the manual OOF annotation grade, ranging from 0 (in-focus, dark blue) to 6 (strong OOF, red). OOF: Out-of-focus

quantitatively aligned (Methods), similar slope values were expected for the Gaussian-trained configuration. Instead, the Gaussian configuration classified patches annotated with OOF grades 5.0–6.0 as mid-range OOF classes 14–20 [Figure S1] compared to 20 and above with the Bokeh configuration [Figure 9]. Moreover, the majority of patches with no or weak OOF (annotated Grades 0.0–1.5) were predicted as OOF classes 0–4 by the Bokeh configuration (ConvFocus), compared to 5 and higher with the Gaussian configuration.

We next assessed ConvFocus's predictions for patches from a set of 21 "z-stack" images [Figure 10]. Almost all patches were predicted to be in-focus at a z-level, though not necessarily at z = 0 [Figure 7]. A strong "v"-shaped trend was also observed: patches were generally predicted to be more OOF at z-levels further from the predicted in-focus z-level (on average at

z = 0). However, using the Gaussian configuration, few patches were predicted to have OOF class <5 [Figures S2a and S3]. Similarly, the configuration with Bokeh blur but linearly increasing convolution mask size (instead of exponential) resulted in less fine-grained sensitivity towards weak OOF degrees [Figures S2b and S4].

Impact of focus quality on cancer detection algorithm performance

Figure 11 shows the effect of OOF on patch-level AUC for the LYNA metastatic breast cancer detector. Results for the first experiment (effects of real OOF) are shown in black dots with gray bars that indicate 95% confidence intervals. Since few image patches in the test set were predicted as higher OOF classes, we bucketed the classes into five groups: 0–4, 5–9, 10–14, 15–19, and 20–29. Results indicated a gradual drop in
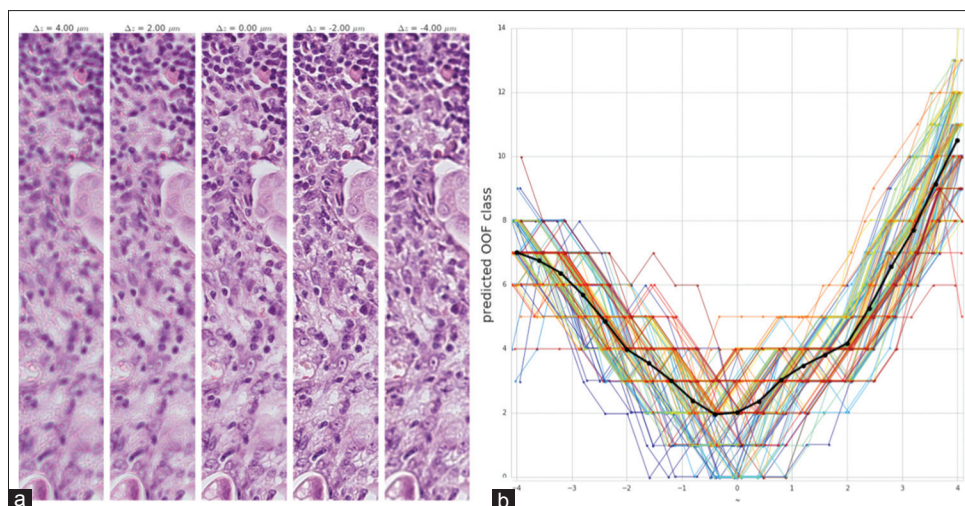
Figure 7: Z-stack examples and OOF predictions along z-axis. (a) Sample high magnification views at different z-levels. (b) Predicted OOF class plotted against z-level for 114 patches from the white rectangle in the z = 0 μm panel of Figure 9. Jitter has been added in the x and y dimensions for better visualization. Different colors are used to clearly indicate different patches. Mean class predictions across these patches are shown in black to highlight the average trend. Spearman's rank correlation rho between the absolute value of z and mean class predictions averages are: >0.999 for $z \leq 0$ and 1.0 for $z \geq 0$ ($P < 0.001$ for both). OOF: Out-of-focus

Table 2: Effects on performance of adding each step in our proposed semi-synthetic data generation process

| Configuration# | Description | Blur method | Spearman rank correlation | | Linear regression | | | |
| | | | | | Slope | | Intercept | |
			AT2	S360	AT2	S360	AT2	S360
1	Default (only blur)	Gaussian	−0.274	−0.134	−0.023	−0.211	2.504	1.949
2	Configuration #1, plus simulated Poisson noise	Gaussian	−0.140	−0.325	−0.057	−0.265	0.502	1.234
3	Configuration #2, plus JPEG artifacts	Gaussian	0.785	0.922	2.851	2.851	2.801	2.801
4	Configuration #3 (ConvFocus), replacing Gaussian with Bokeh blurring	Bokeh	0.808	0.936	3.903	4.408	−2.945	−1.730
5	Configuration #4, replacing ConvFocus's simpler convolutional neural network with a more complex, modified Inception architecture	Bokeh	0.773	0.934	3.889	4.504	−3.024	−1.969

The final configuration illustrates that a more complex architecture is not required for this specific task. JPEG: Joint Photographic Experts Group

cancer detector performance with predicted OOF, as evidenced by decreasing AUC values and the associated upper bounds of the confidence intervals.

To assess the effects of finer-grained OOF, we conducted a second experiment: synthetically adding OOF, as depicted by the blue curve. In contrast to the first experiment, since synthetic blur at each specific degree was applied uniformly across all slides, no class-merging was necessary. Results similarly showed the gradual and consistent drop in cancer detection with increasing OOF.

CONCLUSIONS

We have developed and evaluated an automated focus quality detector (ConvFocus) to accurately locate and classify OOF regions as small as 32 μm × 32 μm in gigapixel-sized WSIs. Our algorithm correlates well with pathologist-annotated focus quality grades on real OOF regions, and across multiple scanners despite different imaging characteristics (e.g., pixel size in the S360 is 9% smaller than in the AT2). Because different scanners do not generally digitize the same slide locations with the same focus quality, the actual image patches used in this evaluation are different. As a result, a direct and fair inter-scanner performance comparison is not possible.

In a second evaluation using z-stacks to produce digitized images of varying OOF degrees, ConvFocus similarly produced sensible trends of increasing predicted OOF classes in both directions away from the scanner-determined in-focus plane. In this evaluation, while the scanner-determined in-focus plane (z = 0) was on average predicted to be close to or the most in-focus, this was not necessarily true of all image patches. Further analysis is needed to determine whether this is caused by inaccuracies in focus detection by the scanner's algorithm, or by ConvFocus. The asymmetry in this "v"-shaped trend may be because the z-levels' span of 8 μm exceeds the thickness of a tissue section (5 μm), and thus, some z-levels will overlap the coverslip or glass slide. This effect may be further exacerbated if the z = 0 plane is not centered within the tissue section.
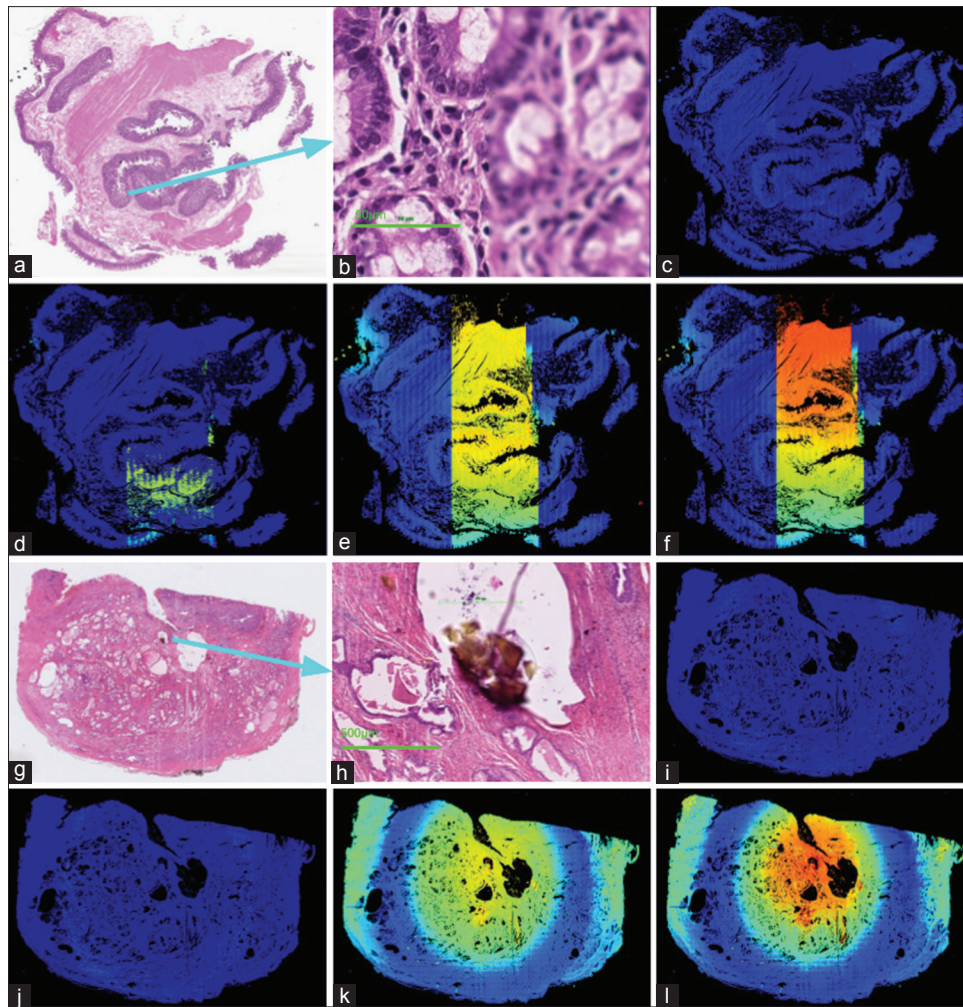
Figure 8: Qualitative assessment of the impact of each step in our semi-synthetic OOF data generation process. (a and g) Slide images of a duodenum and prostate specimen. (b and h) Magnified views of regions with OOF artifacts of panels a and g, respectively. (c-f and i-l) Algorithm-predictedOOF classes; color map is shown in Figure 2d. Four different configurations were applied [Table 2]. (c and i) Algorithm trained with Gaussian blurring and brightness perturbations only. (d and j) Model trained with simulated JPEG artifacts in addition. (e and k) Algorithm trained with simulated Poisson noise in addition. (f and l): Algorithm trained with Bokeh blurring instead of Gaussian. OOF: Out-of-focus

A significant advantage of our semi-synthetic data generation approach is an easy extension to additional scanner models, tissues, and stain types. Specifically, input data can be generated by non-expert raters, improving the speed and cost of data collection. Our results indicate that our method of synthesizing OOF examples of different degrees enables better generalization to real OOF images. In terms of granularity, ConvFocus was developed for fine-grained discrimination between 30 OOF classes, significantly more than both our pathologist's grading scale and prior works: six classes[23] and twelve classes.[22,23] When the granularity was further increased in our experiments (e.g., to 40 classes, data not shown), training accuracy decreased, indicating that 30 classes is close to a meaningful upper bound of granularity. Our qualitative [Figure 4] and quantitative [Figure S1] analyses also found that human perception of OOF followed an exponential relationship with blur magnitude, as opposed to linear. The exponential relationship expands the granularity of fine-grained OOF classes, and is particularly relevant for use cases like lymphoma diagnosis, where even mild OOF may interfere with the ability to resolve crucial nuclear details.

Prior works in synthetic data generation for training focus quality predictors have leveraged Gaussian blurring for histopathology[23] and Gaussian blurring with Poisson noise for cytology specimens from transmission light microscopes.[22,23] By contrast, our work systematically demonstrates the value of adding Poisson noise and JPEG compression artifacts in histopathology, and using a synthetic blur (Bokeh) that more closely mimics real OOF. Our results indicate that generating synthetic OOF data using Gaussian blurring underestimates strong OOF. Based on these improvements, we showed that focus quality can be accurately predicted on a fine-grained scale using fairly shallow neural networks, which improves the speed of applying ConvFocus to large WSIs.
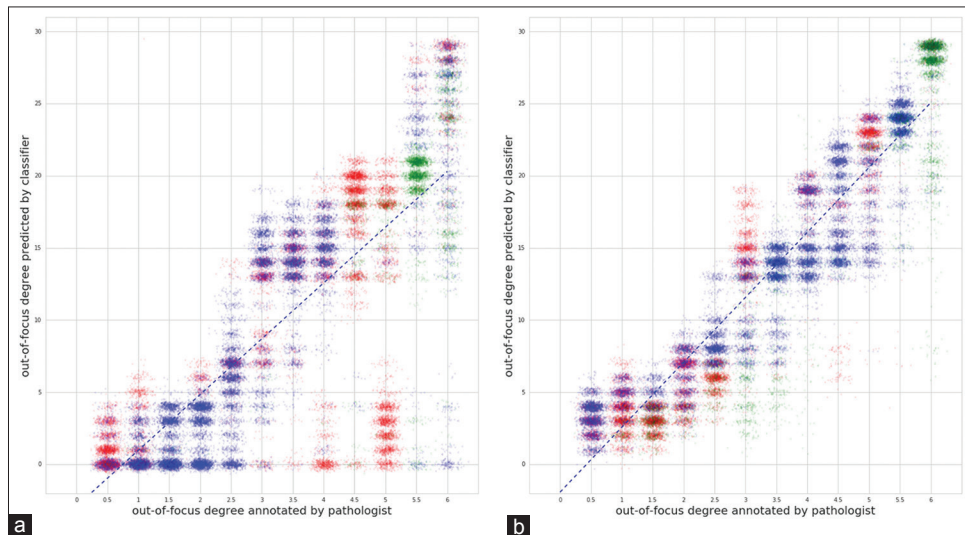
Figure 9: Scatter plots of ConvFocus predictions against pathologist-annotated OOF grades. Plot show 36,000 annotated image patches across seven different slides scans per scanner model. To enable visualization of point density, small amounts of x, y jitter were added. Left: results for AT2 (Spearman's ρ =0.808). Rights: results for S360 (Spearman's ρ =0.936). Colors indicate patches from different specimens or stains; red: lymph node; blue: prostate; green: immunohistochemistry stained slides. OOF: Out-of-focus
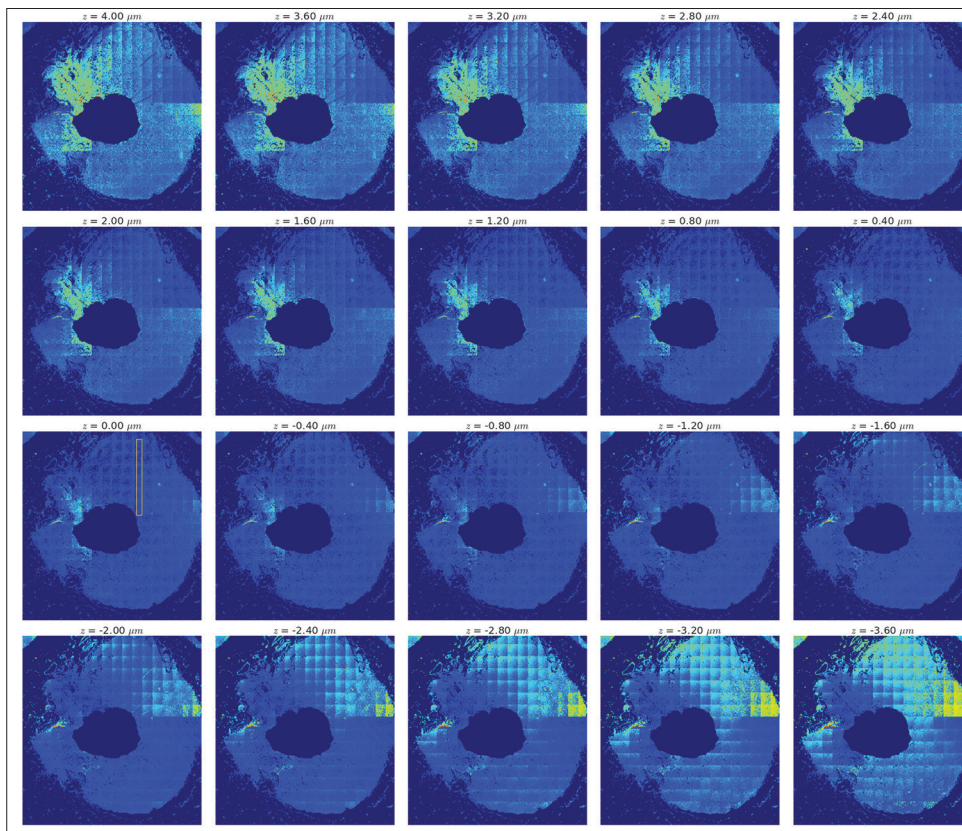
Figure 10: ConvFocus predictions for z-stack scans from a lymph node biopsy of a colon cancer case. This is the middle lymph node in Figure 3b. WSIs were acquired in z-stack mode on a NanoZoomer S360 ranging from $+4$ μm (top left) to -3.6 μm (bottom right) in 0.4 μm increments. z = 0 μm indicates the scanner-determined "in-focus" plane. Plots of the OOF predictions for patches in the white rectangle in the z = 0 μm are shown in Figure 10

We further investigated the impact of focus quality on an otherwise accurate algorithm's interpretation,[7] showing that algorithm performance was lower in image patches identified by ConvFocus as OOF. Furthermore, our controlled experiment of synthetically "de-focusing" the image patches showed similar trends, demonstrating
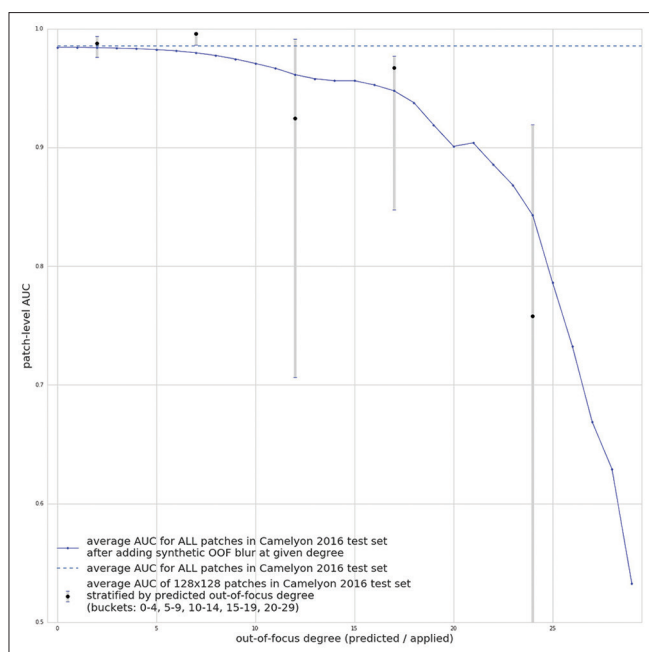
Figure 11: Performance of a breast cancer metastasis detection algorithm as a function of OOF degree. Patch-level area under receiver operating characteristic curve was computed using annotations in the publically-available Camelyon 2016 challenge's test set. Performance degrades whether looking at real images of different OOF degrees (black dots with gray bars that indicate 95% confidence intervals), or synthetically blurred OOF images (blue line). Dotted line represents the patch-level AUC of the cancer detection algorithm across all patches (independent of OOF degree). OOF: Out-of-focus

that synthetic OOF causes degradations in algorithm performance. These findings suggest that focus quality can negatively impact algorithms, and that focus quality should be comprehensively assessed at the region level to avoid both false-positive and false-negative interpretations by an algorithm. This is particularly important for algorithms because most algorithms are developed to assess smaller cropped image patches from large WSIs (e.g., 300 pixels for a cropped patch compared to 100,000 pixels across for the entire slide). By contrast, pathologists assessing an image generally review the tissue at lower magnifications, and also have the option of zooming out to assess a region of interest at a lower magnification.

The present work contains some limitations. First, the cost of manual grading of focus quality has limited our test dataset size and diversity in terms of stains, tissue types, scanner models, as well as number of annotators. Further work will be needed to further validate the generalizability of ConvFocus against multiple rater annotations,[23] and quantifications of z-stack prediction performance. Second, using our algorithm to compare focus quality across different scanners was confounded by the fact that different image patches are identified by our pathologist as OOF. Additional work will be needed to obtain scanner-agnostic or tissue-independent metrics of focus. Finally, we have focused on the high magnification, ×40,

which most clearly shows OOF artifacts. The extension to the next-lower ×20 is the subject of future work.

Conflicts of interest

T.K., Y.L., M.M., P.C., J.H., C.M., and M.S. are/were employees of Google, LLC and own/owned Alphabet stock. T.B. was compensated by Google, LLC for the participation as a pathologist in these studies.

REFERENCES

1. Pantanowitz L. Digital images and the future of digital pathology. J Pathol Inform 2010;1:15.
2. Hashimoto N, Bautista PA, Yamaguchi M, Ohyama N, Yagi Y. Referenceless image quality evaluation for whole slide imaging. J Pathol Inform 2012;3:9.
3. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, *et al.* Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (Pivotal study). Am J Surg Pathol 2018;42:39-52.
4. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
5. Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: Whole-slide imaging and beyond. Annu Rev Pathol 2013;8:331-59.
6. Stathonikos N, Veta M, Huisman A, van Diest PJ. Going fully digital: Perspective of a Dutch academic pathology lab. J Pathol Inform 2013;4:15.
7. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, *et al.* Artificial intelligence-based breast cancer nodal metastasis detection: Insights into the black box for pathologists. Arch Pathol Lab Med 2019;143:859-68.
8. Montalto MC, McKay RR, Filkins RJ. Autofocus methods of whole slide imaging systems and the introduction of a second-generation independent dual sensor scanning method. J Pathol Inform 2011;2:44.
9. Pantanowitz L, Farahani N, Parwani A. Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives. Pathol Lab Med Int 2015;7:23-33.
10. McKay RR, Baxi VA, Montalto MC. The accuracy of dynamic predictive autofocusing for whole slide imaging. J Pathol Inform 2011;2:38.
11. Ameisen D, Deroulers C, Perrier V, Bouhidel F, Battistella M, Legrès L, *et al.* Automatic Image Quality Assessment in Digital Pathology: From Idea to Implementation. Proc. of IWBBIO; Grenada, Spain, 2014.
12. Zerbe N, Hufnagl P, Schlüns K. Distributed computing in image analysis using open source frameworks and application to image sharpness assessment of histological whole slide images. Diagn Pathol 2011;6 Suppl 1:S16.
13. Lahrmann B, Valous NA, Eisenmann U, Wentzensen N, Grabe N. Semantic focusing allows fully automated single-layer slide scanning of cervical cytology slides. PLoS One 2013;8:e61441.
14. Moles Lopez X, D'Andrea E, Barbot P, Bridoux AS, Rorive S, Salmon I, *et al.* An automated blur detection method for histological whole slide imaging. PLoS One 2013;8:e82710.
15. Walkowski S, Szymas J. Quality evaluation of virtual slides using

methods based on comparing common image areas. Diagn Pathol 2011;6 Suppl 1:S14.

16. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199-210.

17. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.

18. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016;191:214-23.

19. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F, *et al.* Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med 2019;2:48.

20. Chen PH, Gadepalli K, MacDonald R, Liu Y, Nagpal K, Kohlberger T, *et al.* An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nat Med 2019;25:1453-7.

21. Senaras C, Niazi MK, Lozanski G, Gurcan MN. DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning. PLoS One 2018;13:e0205387.

22. Yang SJ, Berndl M, Ando DM, Barch M, Narayanaswamy A, Christiansen E, *et al.* Assessing microscope image focus quality with deep learning. BMC Bioinform 2018;19:1,77.

23. Campanella G, Rajanna AR, Corsale L, Schüffler PJ, Yagi Y, Fuchs TJ. Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology. Comput Med Imaging Graph 2018;65:142-51.

24. Nasse HH. Depth of Field and Bokeh. Carl Zeiss AG. Available from: https://www.lenspire.zeiss.com/photo/app/uploads/2018/04/Article-Bokeh-2010-EN.pdf. [Last accessed on 2019 Oct 18].

25. Wadhwa N, Levoy M, Garg R, Feldman B, Kanazawa N, Carroll R, *et al.* Synthetic depth-of-field with a single-camera mobile phone. ACM Trans Graph 2018;37:1-13.

26. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016.

27. Kokoska S, Zwillinger D. CRC Standard Probability and Statistics Tables and Formulae. Boca Raton, FL, USA: CRC Press; 2000.

28. Liu Y, Gadepalli K, Norouzi M, Dahl G, Kohlberger T, Boyko A, *et al.* Detecting Cancer Metastases on Gigapixel Pathology Images. arXiv 2017; abs/1703.02442.

29. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.

30. Chihara LM, Hesterberg TC. Mathematical Statistics with Resampling and R. 2nd ed. Hoboken, NJ, USA: Wiley; 2018.
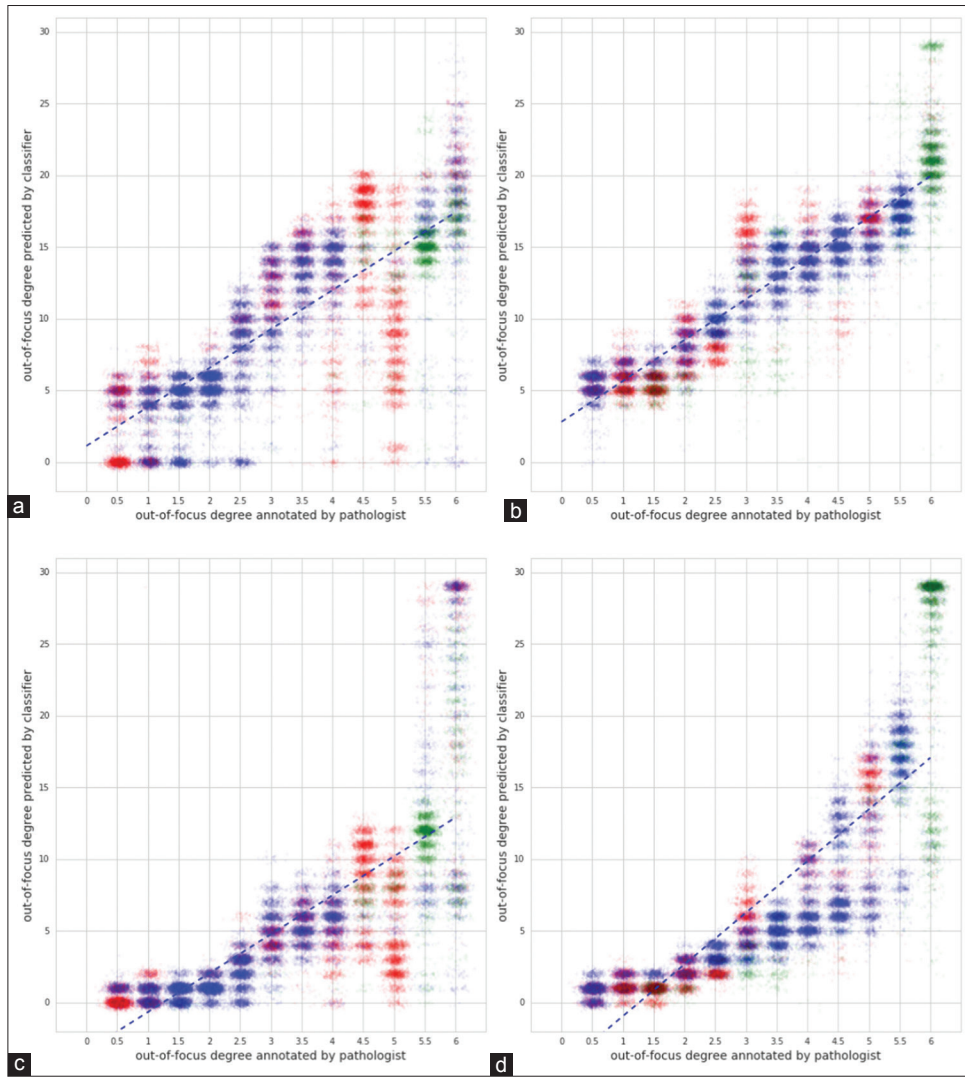
Figure S1: Scatter plots of other configurations of OOF models against pathologist annotations [compare with results in Figure 8]. (a and b) Results of OOF model trained with Gaussian instead of Bokeh blurring [configuration #3 in Table 2]. (c and d) Results of OOF model trained with Bokeh-blurring model but with a linearly increasing kernel size instead of exponential. Left column (a and c) indicates AT2 results; right column (b and d) indicates S360 results
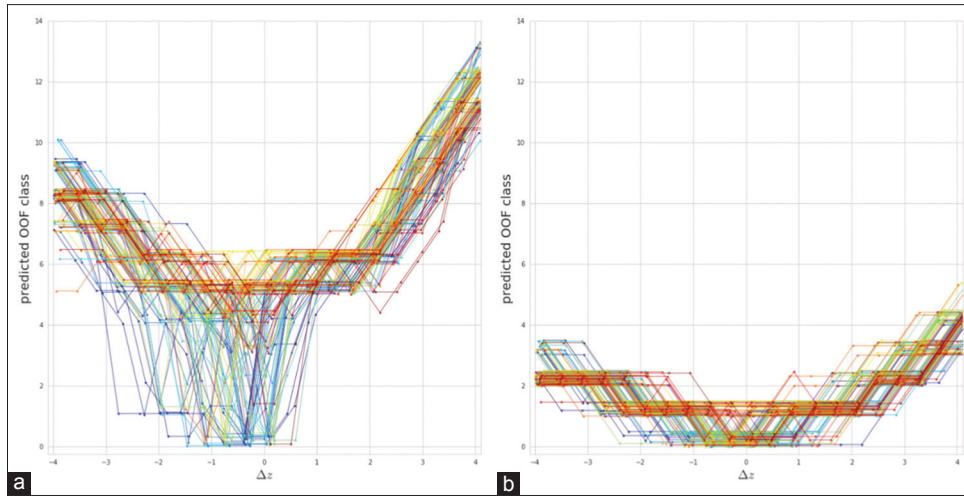
Figure S2: Predicted OOF classes over z for patches sampled from the white rectangle in the z = 0 panel of Figures S3 and S4. (a): Results of configuration trained with Gaussian instead of Bokeh blurring [configuration #3 in Table 2]. (b): Results of configuration trained with Bokeh-blurring but with linearly increasing convolution mask size instead of exponential
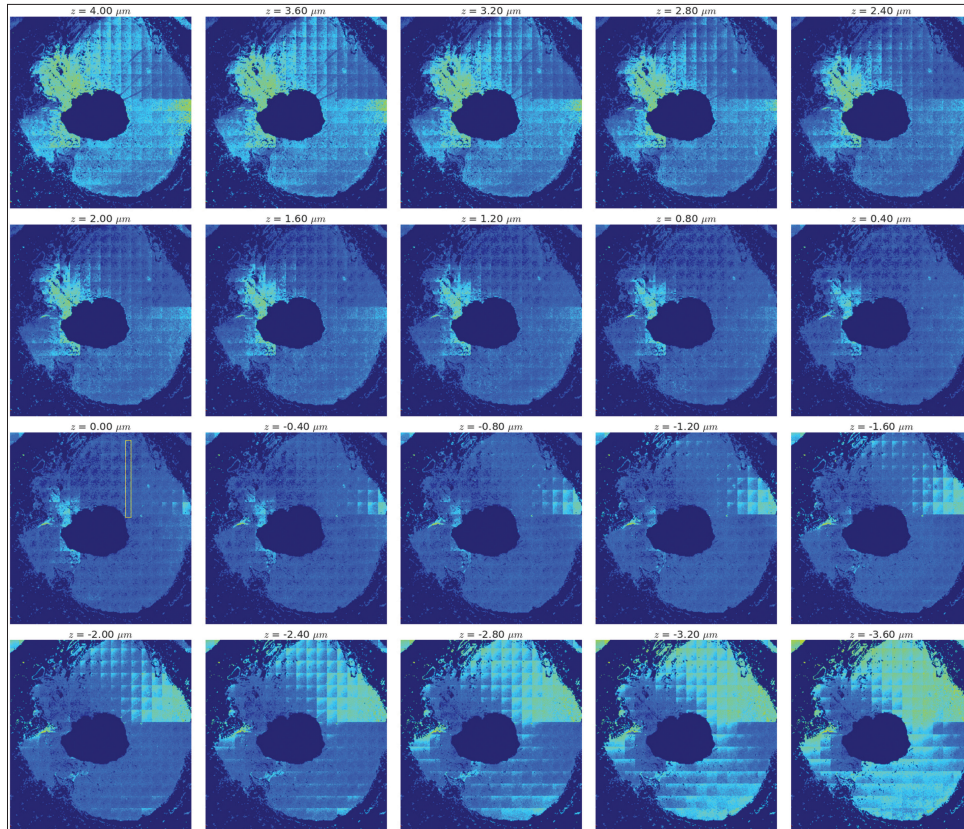

Figure S3: Results of Gaussian-trained configuration on z-stack scans. Corresponds to results shown in Figure S2a
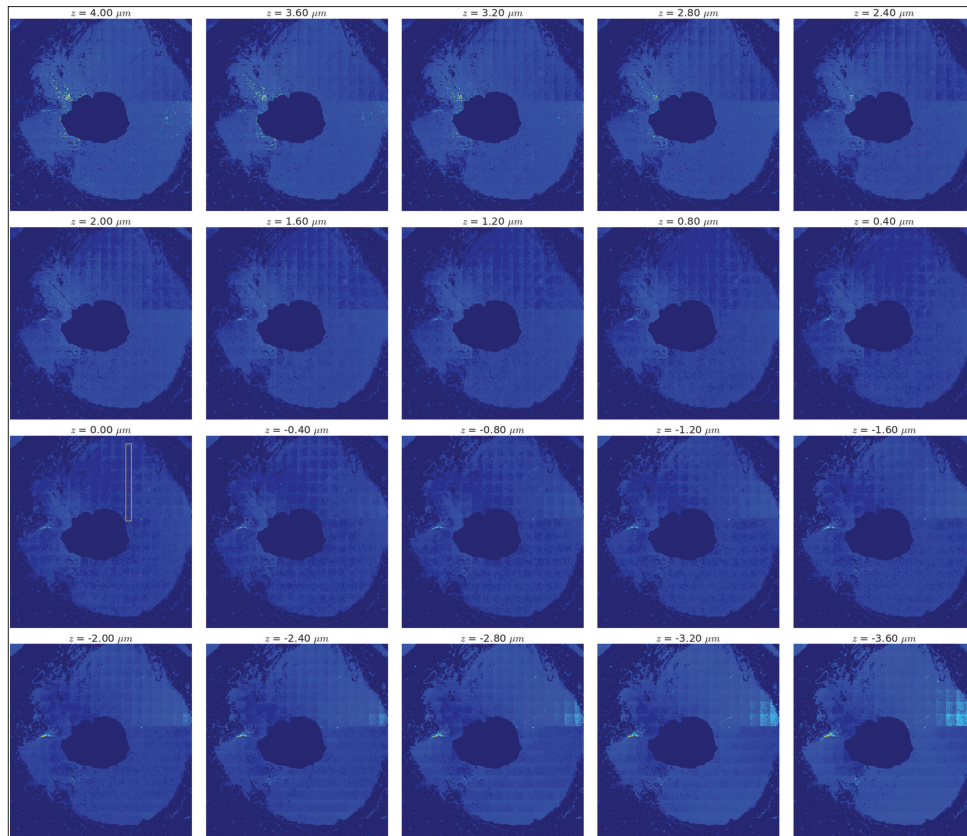
Figure S4: Results of linear kernel size configuration on z-stack scans. Corresponds to results shown in Figure S2b