Practical segmentation of nuclei in brightfield cell images with neural networks trained on fluorescently labelled samples

Dmytro Fishman¹ | Sten-Oliver Salumaa¹ | Daniel Majoral¹ | Tõnis Laasfeld¹,² | Samantha Peel³ | Jan Wildenhain³ | Alexander Schreiner⁴ | Kaupo Palo⁴ | Leopold Parts¹,⁵

¹ Department of Computer Science, University of Tartu, Narva Str 20, Tartu 51009, Estonia
² Chair of Bioorganic Chemistry, Institute of Chemistry, University of Tartu, Ravila, Estonia
³ Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
⁴ PerkinElmer Cellular Technologies Germany GmbH, Hamburg, Germany
⁵ Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

Correspondence
Leopold Parts.
Email: leopold.parts@ut.ee

Authors Dmytro Fishman and Sten-Oliver Salumaa contributed equally to this work.

Abstract
Identifying nuclei is a standard first step when analysing cells in microscopy images. The traditional approach relies on signal from a DNA stain, or fluorescent transgene expression localised to the nucleus. However, imaging techniques that do not use fluorescence can also carry useful information. Here, we used brightfield and fluorescence images of fixed cells with fluorescently labelled DNA, and confirmed that three convolutional neural network architectures can be adapted to segment nuclei from the brightfield channel, relying on fluorescence signal to extract the ground truth for training. We found that U-Net achieved the best overall performance, Mask R-CNN provided an additional benefit of instance segmentation, and that DeepCell proved too slow for practical application. We trained the U-Net architecture on over 200 dataset variations, established that accurate segmentation is possible using as few as 16 training images, and that models trained on images from similar cell lines can extrapolate well. Acquiring data from multiple focal planes further helps distinguish nuclei in the samples. Overall, our work helps to liberate a fluorescence channel reserved for nuclear staining, thus providing more information from the specimen, and reducing reagents and time required for preparing imaging experiments.

KEYWORDS
brightfield, image analysis, deep learning, nuclear segmentation

1 INTRODUCTION

Much of our understanding of cells has been derived from microscopy experiments. Images informed the very definition of the word ‘cell’ to describe their structure,¹ led to the discovery of organelles like nuclei² and mitochondria,³ and are currently central to phenotyping individual cells.⁴ Advances in instrumentation have enabled scalable...
approaches to visualise and quantify cell features, which are widely used in basic research and drug development.1

High throughput fluorescence cell imaging collects information from a small number of partially overlapping emission spectra. Each of these can correspond to the expression of a protein to reflect its abundance and localisation in the cell, or a stain that has concentrated in areas with particular biochemical properties. The former are typically used to study individual genes, while the latter usually mark entire organelles. Popular staining choices are Hoechst dyes for DNA (and thereby, cell nuclei), mitotracker for mitochondria, phalloidin for actin, concanavalin A for lectins, and wheat germ agglutinin for membranes.6 The formed structures are easily visually distinguished and can be accurately detected with automated computational methods.7,8

While cell stains are useful, there are drawbacks to using them. Only a limited number of non-overlapping stains can be used on one specimen. This is a bottleneck in experiments where stains are used to record the layout of organelles as a reference for measuring individual proteins, as each organelle takes up a channel that could be used to study a gene. The process of staining itself takes substantial time and resources. If cells need to be fixed, it destroys the sample and therefore precludes recording dynamic information, such as response to a drug.9 It is also possible to stain live cells, but such reagents tend to be reactive and leak from the organelles, making interpretation of other effects difficult. Common live imaging alternatives are cell line constructs with fluorescent labelling that either need to be purchased or laboriously engineered. Therefore, reducing the number of objects that need to be labelled is of great practical importance.

Brightfield images provide complementary information about a sample. This imaging modality records white light transmission properties and is therefore not specific to any particular structure. As nuclei contain densely packed DNA, photons passing through them take an optical path different from the surrounding cytoplasm, which can in principle be used to identify their location. If feasible, using brightfield images for detecting nuclei would enable experiments with an additional free channel for staining, or possibility for the temporal acquisition of live cells without perturbing their state.

The main barrier for widespread use of brightfield images for nuclear segmentation is the difficulty of automated image analysis. This is a complicated task even for humans, and as classical computational approaches have not been sufficiently accurate, fluorescent staining with its clean nuclear signal has been preferred. However, recent breakthroughs in training multilayer neural networks using abundant data have led to impressive performance on image analysis challenges,10 and generated excitement for their use on biological data as well.11 Further optimisations on the process, such as employing data augmentation12 have also relaxed the requirements on dataset size.

As a result, deep learning methods have rapidly pervaded into analysing images of cells and tissues.13–15 Nuclear segmentation has a venerable history in histopathology,16,37 and a flurry of recent work has effectively utilised convolutional neural networks on this task.18–21 In cell microscopy, deep learning was first confirmed to be useful for this purpose for fluorescence images,22 and was later thoroughly evaluated under different training regimes and performance metrics.23 For brightfield data, semi-automated ground truth from the fluorescence channel was successfully used to train a network to detect nuclei in a stack of brightfield acquisitions.24 Expanding beyond nuclei, deep convolutional architectures have further been trained to predict fluorescent signals from multiple organelles,25–27 a pre-trained U-Net28 has been included as an ImageJ plugin for analysing various modalities,29 and a generic online segmentation solution has been proposed.30

In spite of this technical progress, segmenting nuclei from brightfield cell images is not yet a standard part of an imaging workflow. When setting up a new project, it is not obvious which of the proposed segmentation architectures is most practical to use, and if existing solutions can be applied out of the box. If model training or fine-tuning is required, there is little guidance on how much data need to be collected for a new experimental setting, even if theoretical31 and domain-specific results32 exist. Instead of annotating brightfield images for training directly, it can also be relatively simple to gather both fluorescence and brightfield images for a subset of the samples on which to train models that can then be used on the remaining unstained specimens. Similarly, it is straightforward to gather brightfield data from multiple focal planes, which in principle captures the light phase information, but it is not clear if this aids segmentation. This lack of understanding of practical considerations motivates testing alternative models as well as data collection strategies to optimise the imaging and analysis process.

Here, we investigate the capacity of deep learning methods to segment nuclei from brightfield images of fixed and fluorescently labelled cells obtained using different microscopes, cell lines, and combinations of focal planes. The ground truth is extracted from fluorescence images in a semi-automatic fashion, and used to train networks on both modalities. We compare the performance of three convolutional neural network architectures that have been previously used for segmentation tasks to find out which one works best in practice, and apply the highest performing one on a panel of images from seven different cell
2 METHODS

All addressed research questions rely on pre-processed microscopy datasets, a set of deep learning architectures, and approaches for training and evaluation. We describe the common methods first (Sections 2.1-2.3), followed by the study setup with question-specific details (Section 2.4).

2.1 Data collection

To evaluate model and transfer learning quality, we used a semi-automatically curated dataset of seven cell lines for which both fluorescent readout of a DNA-binding dye and a brightfield measurement were acquired. The exact same field of view was acquired with both fluorescence and brightfield modality, and the images overlaid without additional registration. Seven different cell lines (mouse fibroblasts (NIH/3T3), canine kidney epithelial cells (MDCK), human cervical adenocarcinoma (HeLa), human breast adenocarcinoma (MCF7), human lung carcinoma (A549), human hepatocellular carcinoma (HepG2) and human fibrosarcoma (HT1080)) were seeded each into 48 wells of a CellCarrier-384 Ultra microplate (PerkinElmer, #6057300). The following cell numbers were seeded per well and cell line – NIH/3T3: 7000, MDCK: 7000, HeLa: 5000, MCF7: 10,000, A549: 12,000, HepG2: 20,000, HT1080: 10,000. On the following day, the cells were fixed using 3.7% formaldehyde solution (Sigma Aldrich, #252549) and nuclei stained with 10 μg/mL Hoechst 33342 (Thermo Fisher, #H3570). Brightfield and fluorescence images were acquired on an Opera Phenix™ high-content screening system (PerkinElmer) using a 20× water immersion objective in confocal mode to capture fluorescence and brightfield images. This acquisition configuration gives images of 1080×1080 pixels stored losslessly at 16 bits per pixel per channel, with each pixel corresponding to a square of 0.59 μm × 0.59 μm. A total of 432 fields of view were captured for each cell line, for a total of 3024 images. We refer to these data as the seven cell lines dataset.

2.2 Training data generation

Ground truth generation – pixels. Ground truth pixel assignments were created semi-automatically by Harmony, PerkinElmer proprietary image analysis software. Nuclei Detection method C was chosen for robustness with respect to size and fluorescence signal contrast variation of nuclei thanks to local thresholding. The method uses local thresholding to identify candidate objects in fluorescence images, followed by approaches to split stuck nuclei combined with morphological opening, closing and filling. To improve ground truth quality, we applied background flatfield correction to the fluorescent image. The segmentation quality was inspected manually across different cell lines and measurements. We used the binary pixel assignments to nucleus or background for all training.

Ground truth generation – objects. To evaluate object detection quality, we further processed the pixel masks to extract objects. The Nuclei Detection method C splits touching nuclei combined with morphological opening, closing and filling. We used the object calls for evaluation only, and did not train the models explicitly to detect objects.

Training and test set selection. The 3024 images from seven cell lines were divided into three independent sets: training set with 2016 images, validation set with 504 images, and test set with 504 images, with different representations from the seven lines (Table S1). All training and evaluation took place on this single split of images for both datasets. We did not apply additional data augmentation.

2.3 Performance metrics

Pixelwise metrics. We used 0.5 threshold on the model output probability maps to classify pixels, and counted the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) pixels in each image. We report

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)
\]

\[
\text{F1 – score} = \frac{(2 \times TP)}{(2 \times TP + FP + FN)}, \quad (2)
\]

\[
\text{Precision} = \frac{TP}{(TP + FP)} \quad (3)
\]

\[
\text{Recall} = \frac{TP}{(TP + FN)} \quad (4)
\]
and

\[
\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)
\]

We also refer to pixel error, computed as \(1.0 – \text{F1-score}\).

Objectwise metrics. First, we binarised the prediction map, and detected objects using the Scipy ndimage.label method.\(^{35}\) We filtered out small objects (area less than 50 pixels). We then generated an intersection over union matrix (‘IoU matrix’) with the ground truth, which stores the ratio of the intersection pixel count and the union pixel count of each pair of predicted and ground truth objects. We applied a 0.5 cutoff to this matrix to identify predicted objects that match a ground truth one well, and designated these as true positives. As predicted objects cannot overlap, at most one predicted nucleus can exhibit IoU of more than 0.5 with a true nucleus. Analogously, we computed the number of false-negative and false-positive cells in the image, and calculated the relevant metrics. It is important to note that specificity and accuracy could not be estimated on the cell level, as they require the number of true negative cells to be known. For a ground truth nucleus that overlaps with multiple predicted objects, the IoU was calculated for all of them, and the largest value picked as the IoU. To obtain the per-image IoU score, we averaged the IoU scores of all ground truth objects in the image.

Splits and merges. We used the IoU matrix described above and applied a threshold of 0.1 on it. For each ground truth object in the resulting matrix we counted the number of intersecting predicted objects. If there were more than one of these intersections then we registered a split for this object – a ground truth object split between multiple predictions with IoU values being at least 0.1. Analogously, for each predicted object we registered a merge if it was overlapping with more than one ground truth object.

2.4 Study setup

All computational experiments were performed on the High Performance Computing cluster of University of Tartu equipped with NVIDIA GPU Tesla V100 with 16 gigabytes of memory, and Intel(R) Xeon(R) Platinum 8160 CPU of 2.10GHz.

2.4.1 Nuclei segmentation from brightfield and fluorescence images with different networks in the A549 cell line

We considered three popular neural network architectures for image analysis – DeepCell,\(^{22}\) U-Net\(^{28}\) and Mask R-CNN,\(^{36}\) as well as two previously established methods of NucleAIzer,\(^{30}\) and out-of-the-box U-Net plugin for ImageJ.\(^{29}\) The three selected architectures represent distinct classes of deep learning approaches to image segmentation. While U-Net uses the information about all input pixels to segment the image in an end-to-end fashion, DeepCell iteratively builds up a segmentation by classifying individual pixels based on a small surrounding area. Finally, Mask R-CNN considers proposal regions, which are accepted or rejected as a whole, and the accepted regions are further binary segmented.

U-Net model and training. The U-Net architecture used (Figure 1A) was based on the original publication\(^{28}\) and consisted of contracting and expansive paths of 1.3 million neurons in total. The contracting part was built using a series of three 3x3 convolutional filters, followed by rectified linear unit activation and 2x2 max pooling operation. Each step in the expansive path contained an upsampling operation, series of convolutional layers followed by rectified linear unit and concatenations with dimension-matched outputs from contracting layers. The model was trained for a maximum of 200 training epochs. Each epoch consisted of 200 × 16 = 3200 inputs in batches of size 16. Each image in the batch was a 288 × 288 pixel random crop from a different full-sized training image. The Adam optimisation algorithm\(^{37}\) was used with a learning rate of 0.0002 for updating network weights with respect to binary cross entropy loss after each forward pass. Upon completion of an epoch, validation loss was calculated on validation images (80 images of 288 × 288 pixels in case of A549 cell line, 504 images when training with all cell lines). The learning rate was decreased by a factor of 10 each time validation loss stopped improving for 10 consecutive epochs. If the number of consecutive epochs without improvement reached 40, the training process was terminated. The model with lowest loss on validation images across epochs was selected for testing. In our experiments, training was terminated for all models before reaching the maximum number of 200 epochs.

DeepCell model and training. The DeepCell architecture\(^{22}\) consisted of four layers of 64 convolutional filters with rectified linear unit activation, each followed by 2x2 max pooling, and a deeply connected layer (Figure 1B). The model was trained for a maximum of 200 training epochs. Each training epoch consisted of 264 × 4096 = 1,081,344 inputs in batches of size 4096. Each input was a 31x31 pixel patch randomly cropped from a training image. Adam optimisation algorithm was used with a learning rate of 0.001 with respect to categorical cross-entropy loss to update the weights after each forward pass. Upon completion of an epoch, validation loss was calculated on 249,654 individual 31x31 pixel patches from validation images. Learning rate was decreased by a factor of 10 each time validation loss stopped improving for 10...
consecutive epochs. If the number of consecutive epochs without improvement reached 40, the training process was terminated. The model with lowest loss on validation images across epochs was selected for testing.

**Mask R-CNN model and training.** The Mask R-CNN architecture (Figure 1C) was repurposed from Ref. (36), and updated to accept 512×512 pixel images as input. The model was trained for a maximum of 200 training epochs. Each epoch consisted of 600 steps with batch size of 1 due to RAM restrictions. Each image in a given step was a 512×512 pixel image randomly cropped from a training image. Stochastic gradient descent was used with a learning rate of 0.001 (as in the Mask R-CNN paper) and binary cross-entropy loss for mask prediction branch to update weights after each forward pass. Upon completion of an epoch, validation loss was calculated on 50 patches of 512×512 pixels from validation images. Learning rate was decreased by a factor of 5 when validation loss stopped improving for 10 consecutive epochs. The model with lowest loss on validation images across epochs was selected for testing.

**Applying ImageJ U-Net plugin.** As object segmentation from bright-field microscopy images has been attempted previously, we compared both the performance and convenience of the earlier approaches with the currently proposed approach. For this, we identified two deep learning-based approaches. The first approach uses a pre-trained U-net architecture proposed for cell counting, detection and morphometry.\(^{29}\) We downloaded and installed the freely available pre-trained network as well as the corresponding ImageJ plugin (ImageJ U-net available at [https://lmb.informatik.uni-freiburg.de/resources/opensource/unet/](https://lmb.informatik.uni-freiburg.de/resources/opensource/unet/)) as described in the instructions and used the software to provide the segmentation for the A549 cell line images.

**Applying NucleAIzer.** A deep learning framework for nucleus segmentation using image style transfer, NucleAIzer,\(^ {30}\) was used to segment images of seven cell lines dataset to provide a comparison to the existing tools. The webserver was accessed via a publicly available url: [https://www.nucleaizer.org/](https://www.nucleaizer.org/). Four test set images from the seven cell lines dataset were uploaded using 'Upload images' form on the home page. These images consisted of two pairs of cell lines (A549 and HepG2) and modalities (fluorescence and brightfield). The 'General – nuclei' regime was selected, and the resulting segmentations downloaded.

**Evaluation.** Segmentation of nuclei from brightfield images was performed using ground truth derived from fluorescence modality described earlier. Due to the DeepCell approach taking hundreds of seconds per image to predict, we trained all three network architectures to segment nuclei from brightfield images only on data from the A549 cell line. Models were evaluated on held out A549 test data according to the pixel-level and object-level performance metrics described above. No quantitative evaluation of the model performance was carried out for NucleAIzer and ImageJ plugin, as they clearly did not generalise to our data.
2.4.2 | U-Net performance on diverse cell lines

For each of the seven cell lines, we trained a separate U-Net model on their brightfield image data with fluorescence-derived ground truth using the same architecture and parameter settings as for the A549 line. We also trained a single model on a training dataset combined across all cell lines. In parallel, to provide a reference for achievable model quality, we trained the same models using fluorescence image data with fluorescence-derived ground truth. All resulting models were validated using cell line specific and combined validation sets, respectively, of appropriate imaging modality during training, and tested using cell line specific test sets.

2.4.3 | Identifying common errors

We inspected results on the held-out test dataset manually, and subjectively categorised the pixel and object errors into groups with shared properties.

2.4.4 | Computational considerations of brightfield segmentation

*Training on individual cell lines.* For each of the seven cell lines, we constructed nine training sets (1, 2, 4, 8, 16, 32, 64, 128 and 256 images; 7x9 = 63 datasets total). We trained a U-Net model on each. Resulting models were validated and tested using cell line-specific validation and test sets.

*Training on all cell lines jointly.* We constructed nine training sets, each containing a fixed number of images from each of the seven cell lines (1, 2, 4, 8, 16, 32, 64, 128 and 256 images per cell line; dataset sizes 7, 14, 28, 56, 112, 224, 448, 896 and 1792 images, respectively). We trained a U-Net model on each, and validated on a joint validation set.

*Leave-one-out training.* For each cell line, we constructed nine training sets, each containing a fixed number of images from each of the other six cell lines (1, 2, 4, 8, 16, 32, 64, 128 and 256 images per cell line; dataset sizes 6, 12, 24, 48, 96, 192, 384, 768 and 1536 images). We trained a U-Net model on each, and validated on the target cell line specific validation set described above.

2.4.5 | Influence of number of focal planes

*Nine focal planes data set.* To understand the benefits of multiple focal planes, we used data acquired using another microscope. The LNCaP cell line (sourced from ATCC) was grown in standard medium, plated to 384-well imaging plate (Cell Carrier Ultra, Perkin Elmer), fixed using formaldehyde, stained using DRAQ5 fluor (Abcam) to label nuclear DNA, and imaged on the CellVoyager 7000 (Yokogawa) using 20× objective in both confocal mode to capture fluorescence images, and non-confocal brightfield mode, in 9 z-planes using 1 μm steps. This acquisition configuration gives images of 2556x2156 pixels stored losslessly at 16 bits per pixel per channel, with each pixel corresponding to a square of 0.325 μm x 0.325 μm. A total of 784 fields of view were captured, each at 9 focal planes.

*Training.* The 784 images of the multi-plane dataset were divided into three independent sets: training set with 628 images, validation set with 78 images, and test set with 78 images. We modified the U-Net structure to accept 9 input focal planes. We then trained the network with 9 different combinations of planes: keeping all the input planes identical at first and then gradually replacing them, creating new input combinations. This allows for a direct evaluation of the influence of adding an extra input plane, as the number of model parameters does not change. The planes for the experiment were added in the following order: 1, 6, 0, 4, 5, 7, 3, 8, 2. The sequence was iteratively constructed by starting with plane 1 only, training each unused plane at every step, and selecting the best-performing plane to be added next. We kept the number of times every input focal plane was used as balanced as possible; for example, for three input focal planes, we used each one three times. Each plane combination was trained in 5 independent runs.

3 | Results

We set out to evaluate deep convolutional neural networks for brightfield nucleus segmentation with paired fluorescence acquisition for ground truth generation (example image pairs in Figure S1). As a first step, we semi-automatically generated segmentations of the fluorescence images (Section 2), and confirmed that these were of high quality (pixelwise F1-score with manual annotation 89%-91%, Figure S2). We then proceeded to use these as the ground truth segmentation masks.

3.1 | Deep learning methods can effectively segment cell nuclei from brightfield images in A549 cell line monocultures

We first trained the DeepCell, Mask R-CNN and U-Net models to segment brightfield images from the A549 cell line. To develop intuition, we started by examining
whether the outcomes they produce are qualitatively similar (Figure 2A). U-Net uses a model where the posterior probabilities of nearby pixels are coupled, allowing trading sensitivity for specificity while maintaining overall object shape. In contrast, DeepCell classifies individual pixels, resulting in unclear boundaries, and speckle noise. Finally, Mask R-CNN considers proposal regions, which are accepted or rejected as a whole, and the accepted regions further binary segmented. This results in sharp decision boundaries, and entire regions collectively assigned to foreground or background. As a comparison, we also used NucleAIzer and ImageJ U-Net plugin to segment an example image, but as it was clear that these pre-trained models do not transfer to our data, we excluded them from further consideration (Figure S3).

Next, we assessed the performance of the different methods quantitatively. While the brightfield data modality is visually challenging, all approaches performed better than a random classifier (Figure 2B). Scores obtained for all test images for each pair of models under standard 0.5 posterior probability cutoff are shown in Figure S4. U-Net achieved the best accuracy overall on the A549 cell line, with an area under receiver operating characteristic (AUROC) of 0.98, and accuracy of 96%. Mask R-CNN had a lower AUROC of 0.89, but similar accuracy of 96%, and DeepCell was less accurate (90%, 0.89 AUROC).

While our primary aim was to assign pixels to nuclei, we also obtained individual objects using each of the network architectures on the A549 cell line (Section 2), and compared their overlap with the semi-automatically generated annotations from Harmony. The relative pixel-level performance of U-Net on different cell lines was mirrored in both fluorescence (objectwise F1-score 0.90-0.98) and brightfield images (objectwise F1-score 0.63-0.91, Figure 2C), and the object-level performance for an image was well-predicted by pixel-level outcome (Pearson’s $R = 0.95$, $y = x$; area under the curve for each method provided in brackets. (C) U-Net segmentation error ($1.0 – F1$ score; $y$-axis) for different cell lines ($x$-axis) using pixelwise (top) and objectwise (bottom) measures for models trained and applied on brightfield (red) and fluorescence (blue) channels, and models trained on one cell line at a time (light) as well as on all data (dark) of the corresponding acquisition type, using fluorescence-derived ground truth.

**Figure 2** Neural networks can identify nuclei in brightfield images. (A) Example segmentation. A patch of a brightfield image from the A549 cell line (first image) annotated with ground truth nucleus boundaries (white, second image) determined from fluorescence data, as well as with predicted object boundaries for the three considered network architectures of DeepCell (green, third image), U-Net (red, fourth image), and Mask R-CNN (blue, fifth image). (B) Neural networks successfully segment nuclei from brightfield images of the A549 cell line. Pixel-level true positive rate ($y$-axis) and false positive rate ($x$-axis) at different score cutoffs for U-Net (red), DeepCell (green), and Mask R-CNN (blue). Grey dashed line: $y = x$; area under the curve for each method provided in brackets. (C) U-Net segmentation error ($1.0 – F1$ score; $y$-axis) for different cell lines ($x$-axis) using pixelwise (top) and objectwise (bottom) measures for models trained and applied on brightfield (red) and fluorescence (blue) channels, and models trained on one cell line at a time (light) as well as on all data (dark) of the corresponding acquisition type, using fluorescence-derived ground truth.
Figure S5). Mask R-CNN explicitly detects individual objects (see below), and performed better than U-Net for limiting the number of splits and merges of objects compared to ground truth. However, due to segmenting full objects, it recalled fewer nuclei, so the average overlap between objects was similar for these two models (Figure S6).

3.2 U-Net segments brightfield nuclei well in a range of cell lines

We then tested the U-Net model on the data from all cell lines together. To establish a reference to the model trained on brightfield images, we trained the same architecture to segment fluorescence images from which the ground truth was derived. Brightfield channel performance varied across cell lines (accuracy 0.89-0.97, and F1-score 0.76-0.86), while fluorescent images were consistently well segmented (accuracy 0.98-1.00, and F1-score 0.95-0.99, Figure 2C). This bias in performance between data modalities is not surprising, as the brightfield signal is more convoluted, and the ground truth is derived from fluorescence readout. HepG2 cells were the most difficult to segment in the brightfield channel, likely due to the high cell density in images, which was positively correlated with per-image error (Figure S7).

We next trained the best performing U-Net architecture on images from each cell line and channel separately. The individual cell line models had 0.5% lower error compared to the network trained on all data (17.4%-17.9% average pixel error, Figure 2C; Tables S2 and S3) for the brightfield channel. Fluorescence channel segmentation error increased slightly in the individual cell line models, but the difference was small in absolute terms (0.05% on average; Tables S4 and S5).

3.3 Common features of prediction errors in brightfield segmentation using U-Net

We inspected the outputs of all test images for the U-Net model and made the following observations about various types of mistakes in segmentation. The recall of ground truth pixels for each nucleus was bimodal, with some nuclei completely missed, and others with incorrect boundaries (Figure S8). The missing nuclei were either appearing in cells with unusual morphology, or having very low signal (Figure S9). Nuclei densely packed in the fluorescent channel, likely often reflecting partially overlapping cells, were more difficult to delineate correctly (Figure S7). In the most extreme cases, entire layers of cells were stacked, and while pure fluorescence emission is unchanged in such 3D configurations, the light captured in brightfield images takes a very different path compared to a monolayer. When the nucleus was detected, but the boundaries did not match, the possible causes were a different focal plane resulting in a shift, or mitosis that splits the stained DNA into two objects.

3.4 Computational considerations of brightfield segmentation

Speed, memory consumption, and data volume requirements for training are important practical considerations. All tested models had tens of thousands of parameters, and take up between 5Gib and 7Gib of memory (15-175 Mib on disk). Segmenting an image with size 1080 × 1080 pixels takes 0.28 seconds for U-net and 6.64 seconds for Mask R-CNN, while the patching approach required by DeepCell renders its application impractical, with 159 seconds per image using a GPU.

Substantial amounts of data are required to train deep neural networks. We first tested models trained on data from a single cell line at a time. On average, we achieved performance within 6% of overall optimum with 32 full micrographs for each line, with moderate variability between the training sets (Figure 3A, B). This suggests that extensive manual labelling is required to train a new brightfield segmentation model.

Next, we tested a model trained on many cell lines. Again, we varied the number of training images, and evaluated on held out data. The performance on any single cell line was improved by including the same number of images from other cell lines during training (Figure 3B, C), even when hundreds of training images were available for the target line. Conversely, removing images corresponding to the target line from a large diverse training set increased error on that cell line (Figure S10), suggesting that the trained networks do not generalise to previously unseen cell lines. However, if only a small number of up to eight training images was available to train a model for a line, an alternative model trained on the same number of images from each of the other six lines outperformed it (Figures 3B, D, S10). These results indicate that a pool of several annotated images or a large diverse training set is required for optimal performance.

3.5 Data from multiple focal planes improves segmentation

Alternative focal planes contain different information about the sample, and it is not obvious which one is best
Segmentation performance depends on training set size. Pixel-level segmentation error (1.0 − pixel F1 score; y-axis) for increasing number of training images per cell line (x-axis). (A) U-Net trained on A549 cell line only, performance measured on individual test images (light lines), and their average (dark line and markers). (B) As (A), but averages only for models trained separately on the seven different cell lines. (C) As B, but for models trained on a training set that includes the same fixed number of images from each cell line. (D) As C, without including the tested line in the training set. The total number of training images is seven times larger than x-axis value in (C), and six times larger than x-axis value in (D).

FIGURE 3

FISHMAN ET AL.

First, we tested whether segmentation performance depends on the focal plane. We trained U-Net on all available data on each acquired plane, and evaluated on held out data. All planes had sufficient information to identify nuclei, but the central ones (planes 2-5) had a higher error compared to the rest (0.847 vs 0.861 pixel-level F1, Figure 4B; 0.761 vs 0.786 object-level F1, Figure 4C). One possible reason for this is the better contrast of sub-nuclear structures, such as nucleoli, in other planes (Figure 4A). Next, we considered whether using multiple input planes improves performance. We augmented the input layer of U-Net to accept a higher-dimensional input, leaving the rest of the architecture unchanged. Two planes improved the performance over a single one (Figure 4D, E), while including information from additional ones gave diminishing returns. Overall, the single plane segmentation performance on this dataset and cell line is within the range of the previously analysed seven cell lines that were imaged on another instrument by independent operators, and could be improved upon by an additional acquisition.

4 | DISCUSSION

We have demonstrated that multiple different neural network architectures can be trained to accurately segment nuclei from images of diverse cell lines acquired in fluorescence and brightfield channels. It is beneficial to have brightfield data from multiple focal planes, and annotations for dozens of training images to achieve optimal performance. Large training datasets from multiple different cell lines lead to more robust performance in segmenting to use for segmentation. Further, multiple acquisitions are feasible, as the brightfield images are quick to capture, and do not damage the specimen. Therefore, we next generated a dataset of prostate cancer cells (LNCaP) images with nine different focal planes in the brightfield channel (Figure 4A).

First, we tested whether segmentation performance depends on the focal plane. We trained U-Net on all available data on each acquired plane, and evaluated on held out data. All planes had sufficient information to identify nuclei, but the central ones (planes 2-5) had a higher error compared to the rest (0.847 vs 0.861 pixel-level F1, Figure 4B; 0.761 vs 0.786 object-level F1, Figure 4C). One possible reason for this is the better contrast of sub-nuclear structures, such as nucleoli, in other planes (Figure 4A). Next, we considered whether using multiple input planes improves performance. We augmented the input layer of U-Net to accept a higher-dimensional input, leaving the rest of the architecture unchanged. Two planes improved the performance over a single one (Figure 4D, E), while including information from additional ones gave diminishing returns. Overall, the single plane segmentation performance on this dataset and cell line is within the range of the previously analysed seven cell lines that were imaged on another instrument by independent operators, and could be improved upon by an additional acquisition.

4 | DISCUSSION

We have demonstrated that multiple different neural network architectures can be trained to accurately segment nuclei from images of diverse cell lines acquired in fluorescence and brightfield channels. It is beneficial to have brightfield data from multiple focal planes, and annotations for dozens of training images to achieve optimal performance. Large training datasets from multiple different cell lines lead to more robust performance in segmenting
nuclei. They help both in cases when only a few annotated images are available for a new cell line, but also when data are plentiful. Altogether, nuclei can be consistently segmented from both brightfield and fluorescence images regardless of the cell line. This suggests that online batch learning will increase model performance over time.

The U-Net model was not only the best performing approach according to the pixelwise AUC metric (Figure 2B) but also when considering the combination of speed, ease of training and performance. In our tests, it was about 25 times faster than Mask R-CNN, and about 500 times faster than DeepCell, making it usable in an online mode along with data acquisition. Its output is a smooth probability map that can be used as input to other methods for further refinement. This way, it does not suffer from the sharp thresholds of Mask R-CNN. Perhaps notably, the training process was subjectively easier using this architecture. The models trained faster, and were less fragile, with small changes having little effect on the performance. A moderate number of training images, which can be annotated with ground truth objects obtained from fluorescence signal, is sufficient to train a U-Net.

Upon manual inspection of all test images, nearly all the pixel assignment errors can subjectively be ascribed to a small number of interpretable causes. Many of these are due to fundamental discrepancies in ground truth data from fluorescence. For example, cells undergoing mitosis will identify with two ground truth nuclei from the Hoechst DNA stain, but as a single nucleus from a bright-field image. The acceptable error level depends on the context of the application. A single endpoint assay with a relatively crude read-out and stringent post hoc size filters can tolerate large false negative detection rates and still provide usable signal. A long time-course that relies on accurate cell tracking over thousands of frames cannot allow losing too many trajectories, however. In many real-world applications, the small segmentation discrepancies we discuss next are often negligible, due to their minor contribution to the overall experimental variation.

One error mode we identified could be ascribed to a mismatch between acquisitions in the fluorescence and bright-field channels. This could happen for the entire image, but such global shift is usually corrected for by the microscope control software. Another reason could be a physical shift of the nucleus during the intervening time, for example due to failed adhesion to the plate surface. While this introduces noise into the training or fine-tuning datasets, the majority of the cells were segmented accurately, and
therefore it does not diminish the overall value of the approach, especially as there is no reasonable alternative.

The second major error mode we observed was likely due to densely packed, or even overlapping nuclei. In the limit of tissue slices multiple cells deep, or thick multilayers, brightfield microscopy is not effective. The extent to which dense cultures can be delineated into individual nuclei remains an open question. The solutions we presented here capture what is feasible using these models and the given training dataset. Further advances from image analysis of overlapping objects as applied in domains from self-driving cars to face recognition, or principled generative models of the data could in principle be employed to attempt to reconstruct signal from obscured objects. In any case, additional dedicated data collection and annotation is required to understand the limitations of, and provide a general solution to the problem of overlapping cells.

The demands on computational resources and datasets are set by the context of the method application, and the quality required. In general, more training data leads to better results, but is expensive to generate. While we could achieve good performance using U-net, many labelled images are required. A practical approach to using the network is to stain nuclei during the assay setup phase in at least one well, acquire images from many fields of view, and use these to perform the much simpler task of fluorescence-based segmentation. This annotation can then be used to generate a ground truth dataset that would serve to train or fine-tune a model for application on the experiment at scale.

We observed a substantial decrease in both pixel and object error upon adding a second input focal plane into the model, while the improvement after including further planes was more modest. The reasons for this could be multiple. First, it could be that the benefit is the same as of multiple independent captures from the same plane, reducing noise. Alternatively, as the different focal planes in brightfield cell imaging capture different distortions of light, observing a variety of them could help to identify true structure in the signal, or even to create extended focus images, which could help resolve overlapping objects in three dimensions. Finally, two acquisitions theoretically enable capturing the phase of light, which can be informative for organelle segmentation. While we do not attempt to explicitly reconstruct or segment extended focus or phase images in this work, we hypothesise that salient aspects of the relevant signal are implicitly used in multi-plane models to improve segmentation.

Cell imaging is ultimately about finding and characterising objects. Our focus was the very first pixel-level classification step (also known as semantic segmentation) rather than object detection (instance segmentation). While standard algorithms exist to separate a semantic segmentation into individual objects, this task can be done using a range of approaches, including deep learning, and remains an active area of research. We reported strong concordance of pixel-level and object-level segmentation performance, as well as high quality of object detection. We made no attempt to explicitly delineate individual objects, so therefore both the object segmentation evaluation dataset and our predictions include clumps of nuclei, which lowers individual object concordance. Nevertheless, to test performance on cell counting, a popular task for non-invasive imaging, we calculated the correlation of detected and ground truth cells per image, and observed excellent correlation (Pearson’s $R = 0.996$, Figure S11). A more detailed analysis of object detection methods remains outside the scope of this work.

We propose a practical solution for training a neural network by gathering a reasonable amount of training data that can be manually inspected, using a pre-trained model, and ideally taking advantage of acquisitions in two focal planes. While there remains scope for advances in transfer learning and separating densely packed or overlapping cells, using deep learning for segmenting brightfield images is already practical.

ACKNOWLEDGEMENTS

Author contributions: Performed model comparisons: DF, SOS. Performed model evaluation on multiple cell lines: DF, SOS. Performed model evaluation on multiple focal planes: DF, DM. Generated training data: SP, JW, AS, KP, TL. Supervised project: KP, LP. Wrote paper: DF, SOS, LP.

Grant support: Estonian Research Council (IUT34-4) and Wellcome (206194) to LP. Estonian Research Council (Project No 2014–2020.4.01.16-0271, PRG1095; PSG59) and Estonian Centre of Excellence in IT (EXCITE) (TK148) to DF, SOS; Estonian Research Council (ERA-NET TRANSCAN-2 (BioEndoCar)) to DF.

CONFLICT OF INTEREST

SP and JW were employees of AstraZeneca while performing this work. AS and KP were employees of PerkinElmer while performing this work. All other authors declare no conflict of interest.

ORCID

Leopold Parts https://orcid.org/0000-0002-2618-670X

REFERENCES

1. Altmann, R. (1894). *Die Elementarorganismen und ihre Beziehungen zu den Zellen*. Leipzig: Veit.
2. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, 12, 878.
3. Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartford, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., & Carpenter, A. E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols, 11*, 1757–1774. http://doi.org/10.1038/nprot.2016.105

4. Brown, R. (1833). On the organs and mode of fecundation in Orchideae and Asclepiadaceae. *Transactions of the Linnean Society of London, 16*, 685–738.

5. Caicedo, J. C., Roth, J., Goodman, A., Becker, T., Karhohs, K. W., Broisin, M., Molnar, C., McQuin, C., Singh, S., Theis, F. J., & Carpenter, A. E. (2019). Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytoometry A*, 95, 952–965.

6. Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?

7. Christiansen, E. M., Yang, S. J., Ando, D. M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O’Neil, A., Shah, K. K., Lee, A. K., Goyal, P., Fedus, W., Poplin, R., Esteva, A., Berndl, M., Rubin, L. L., Nelson, P., & Finkbeiner, S. (2018). In silico labeling: Predicting fluorescent labels in unlabeled images. *Cell*, 173, 792–803, e719.

8. Du, S. S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R., & Singh, A. (2018). How many samples are needed to estimate a convolutional or recurrent neural network?.

9. Eschen, C., Blaaberg, S., Winther, O., & Paulsen, R. R. (2018). Semantic segmentation of cell nuclei and cytoplasms in microscopy images. openreview.net

10. Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C., Walsh, S., Saltukoglu, D., Tay, T. L., Prinz, M., Palme, K., Simons, M., Dieter, I., Brox, T., & Ronneberger, O. (2019). U-Net: Deep learning for cell counting, detection, and morphometry. *Nature Methods, 16*, 67–70.

11. Fan, H., & Zhou, E. (2016). Approaching human level facial landmark localization by deep learning. *Image and Vision Computing*, 47, 27–35.

12. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

13. Holland, R., Szkalalisty, A., Toth, T., Tasnadi, E., Molnar, C., Mathe, B., Grexa, I., Molnar, J., Balind, A., Gorbe, M., Kovacs, M., Migh, E., Goodman, A., Balassa, T., Koos, K., Wang, W., Caicedo, J. C., Bara, N., Kovacs, F., Paavolainen, L., Danka, T., Kriston, A., Carpenter, A. E., Smith, K., & Horvath, P. (2020). nucelAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Systems*, 10, 453–458.

14. Hooke, R. (1665). *Micrographia: Or, some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon*.

15. Irshad, H., Veillard, A., Roux, L., & Racoeauceau, D. (2014). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. *IEEE Reviews in Biomedical Engineering*, 7, 97–114.

16. Isherwood, B., Timpson, P., McGhee, E. J., Anderson, K. I., Canel, M., Serrels, A., Brunton, V. G., & Carragher, N. O. (2011). Live cell in vitro and in vivo imaging applications: Accelerating drug discovery. *Pharmaceutics*, 3, 141–170.

17. Jones, W., Alasoo, K., Fishman, D., & Parts, L. (2017). Computational biology: Deep learning. *Emerging Topics in Life Sciences*, 1, 257–274.

18. Kask, P., Palo, K., Hinnah, C., & Pommerenke, T. (2016). Flat field correction for high-throughput imaging of fluorescent samples. *Journal of Microscopy*, 263, 328–340.

19. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv. https://arxiv.org/abs/1412.6980

20. Li, L., Zhou, Q., Voss, T. C., Quick, K. L., & LaBarbera, D. V. (2016). High-throughput imaging: Focusing on in drug discovery in 3D. *Methods*, 96, 97–102.

21. Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33, 170–175.

22. Mendelsohn, M. L., Kolman, W. A., Perry, B., & Prewitt, J. M. S. (1965). Computer analysis of cell images. *Postgraduate Medicine*, 38, 567–573.

23. Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., & Van Valen, D. (2019). Deep learning for cellular image analysis. *Nature Methods*, 16, 1233–1246.

24. Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology* (vol. 2, pp. 80–203).

25. Naylor, P., Lae, M., Regal, F., & Walter, T. (2017). Nuclei segmentation in histopathology images using deep neural networks. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*.

26. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collin, F., & Johnson, G. R. (2018). Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15, 917–920.

27. Pärnmaa, T., & Parts, L. (2017). Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning, *G3*, 7, 1385–1392. http://doi.org/10.1534/g3.116.033654

28. Rohban, M. H., Singh, S., Wu, X., Berthet, J. B., Bray, M. - A., Shrestha, Y., Varelas, X., Boehm, J. S., & Carpenter, A. E. (2017). Systematic morphological profiling of human gene and allele function via Cell Painting. *Elife*, 6. http://doi.org/10.7554/elife.24060

29. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Computer Vision and Pattern Recognition*, 234–241.

30. Sadanandam, S. K., Ranefall, P., Guyader, S. Le, & Wühlby, C. (2017). Automated training of deep convolutional neural networks for cell segmentation. *Scientific Reports, 7*, 7860.

31. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6 (1), http://doi.org/10.1186/s40537-019-0197-0

32. Sirinukunwattana, K., Ahmed Raza, S. E., Tsang, Y. - W., Sneed, D. R. J., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35, 1196–1206. http://doi.org/10.1109/tmi.2016.2525803

33. Sornapudi, S., Stanley, R. J., Stoecker, W. V., Almubarak, H., Long, R., Antani, S., Thoma, G., Zuna, R., & Frazier, S. R. (2018).
Deep learning nuclei detection in digitized histology images by superpixels. *Journal of Pathology Informatics*, 9, 5.

34. Usaj, M. M., Styles, E. B., Verster, A. J., Friesen, H., Boone, C., & Andrews, B. J. (2016). High-content screening for quantitative cell biology. *Trends in Cell Biology*, 26, 598–611. [http://doi.org/10.1016/j.tcb.2016.03.008](http://doi.org/10.1016/j.tcb.2016.03.008)

35. Van Valen, D. A., Kudo, T., Lane, K. M., Macklin, D. N., Quach, N. T., DeFelice, M. M., Maayan, I., Tanouchi, Y., Ashley, E. A., & Covert, M. W. (2016). Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLOS Computational Biology*, 12, e1005177.

36. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy, C. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, 261–272. [http://doi.org/10.1038/s41592-019-0686-2](http://doi.org/10.1038/s41592-019-0686-2)

37. Xie, L., Qi, J., Pan, L., & Wali, S. (2020). Integrating deep convolutional neural networks with marker-controlled watershed for overlapping nuclei segmentation in histopathology images. *Neurocomputing*, 376, 166–179.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

---

**How to cite this article**: Fishman D, Salumaa S-O, Majoral D, et al. Practical segmentation of nuclei in brightfield cell images with neural networks trained on fluorescently labelled samples. *Journal of Microscopy*, 2021;1–13. [https://doi.org/10.1111/jmi.13038](https://doi.org/10.1111/jmi.13038)