Practical Fluorescence Reconstruction Microscopy for Large Samples and Low-Magnification Imaging

Julienne LaChance[1], Daniel J. Cohen[1,3]*

[1]Department of Mechanical and Aerospace Engineering, Princeton University
[3]Department of Chemical and Biological Engineering, Princeton University

Correspondence should be sent to:
Daniel J. Cohen
danielcohen@princeton.edu
109 Hoyt Laboratory
Princeton University
Princeton, NJ 08544

Abstract

Fluorescence reconstruction microscopy (FRM) describes a class of techniques where transmitted light images are passed into a convolutional neural network that then outputs predicted epifluorescence images. This approach enables many benefits including reduced phototoxicity, freeing up of fluorescence channels, simplified sample preparation, and the ability to re-process legacy data for new insights. However, FRM can be complex to implement, and current FRM benchmarks are abstractions that are difficult to relate to how valuable or trustworthy a reconstruction is. Here, we relate the conventional benchmarks and demonstrations to practical and familiar cell biology analyses to demonstrate that FRM should be judged in context. We further demonstrate that it performs remarkably well even with lower-magnification microscopy data, as are often collected in screening and high content imaging. Specifically, we present promising results for nuclei, cell-cell junctions, and fine feature reconstruction; provide data-driven experimental design guidelines; and provide researcher-friendly code, complete sample data, and a researcher manual to enable more widespread adoption of FRM.

Introduction

Deep learning holds enormous promise for biological microscopy data, and offers especially exciting opportunities for fluorescent feature reconstruction[1–5]**.** Here, fluorescence reconstruction microscopy (FRM) takes in a transmitted light image of a biological sample and outputs a series of reconstructed fluorescence images that predict what the sample would look like had it been labeled with a given series of dyes or fluorescently tagged proteins (Fig. 1A-C) [2,6–10]. FRM works by first training a convolutional neural network (e.g. U-Net) to relate a large set of transmitted light data to corresponding real fluorescence images (the ground truth) for given markers[11–13]. The network learns by comparing its fluorescence predictions to the ground truth fluorescence data and iterating until it reaches a cut off. Once trained, FRM can be performed on transmitted light data without requiring any additional fluorescence imaging. This is a powerful capability and allows FRM to: reduce phototoxicity; free up fluorescence channels for more complex markers; and enable re-processing of legacy transmitted light data to extract new information. In all cases, FRM data are directly compatible with any standard fluorescence analysis software or workflows (e.g. ImageJ plug-ins). Such capabilities are extremely useful, and FRM may eventually become a standard tool to augment quantitative biological imaging once practical concerns are addressed.

However, a number of challenges limit FRM accessibility to the larger biological community. Key among these is the difficulty in relating the abstract accuracy metrics used to score FRM to the practical value of FRM data for actual, quotidian biological analyses such as cell counting or morphological characterization. To better appreciate this, consider first that the quality of FRM is typically assessed using a single numerical metric (*P*) such as the Mean-Squared-Error or Pearson's Correlation Coefficient that typically range from (0,1) or (-1,1), and second that it is

practically impossible to actually reach perfection ($P = 1$). P can be increased closer to 1 either by training with more images, or by using higher resolution magnification (e.g. 40X-100X) to capture finer details. However, increasing P also carries an intrinsic cost in increased wet-lab and computing time. That improving P is expensive and that P cannot be perfect beg the questions of how good is good enough, and good enough for what (Fig. 1D)? For instance, $P = 0.7$ lacks any practical context, and may be quite good enough for a given use case without requiring more work to raise the 'accuracy'. This is why context is extremely important for FRM and why the work we present here focuses on evaluating practical uses of FRM with respect to given P values.

Our goal here is to provide a standardized implementation of FRM and demonstrate its practical performance and limitations for every-day tasks such as nuclear localization and tracking, characterizing cell morphology, cell-cell junction detection and analysis, and re-analyzing legacy data and data collected on different systems (Fig. 1). To further emphasize the use of FRM for routine tasks, we will exclusively focus on those lower magnifications (4X-20X) commonly used in high content imaging and cellular screening in contrast to the focus on higher magnifications in prior studies [9,10]. We hope that the included software we developed and the analyses and comparison data we present will help make FRM more approachable to the broader biological community. To further facilitate this, we have made the entirety of our code and all collected data public as well as providing a full tutorial guide (see Methods and Supplementary Material).

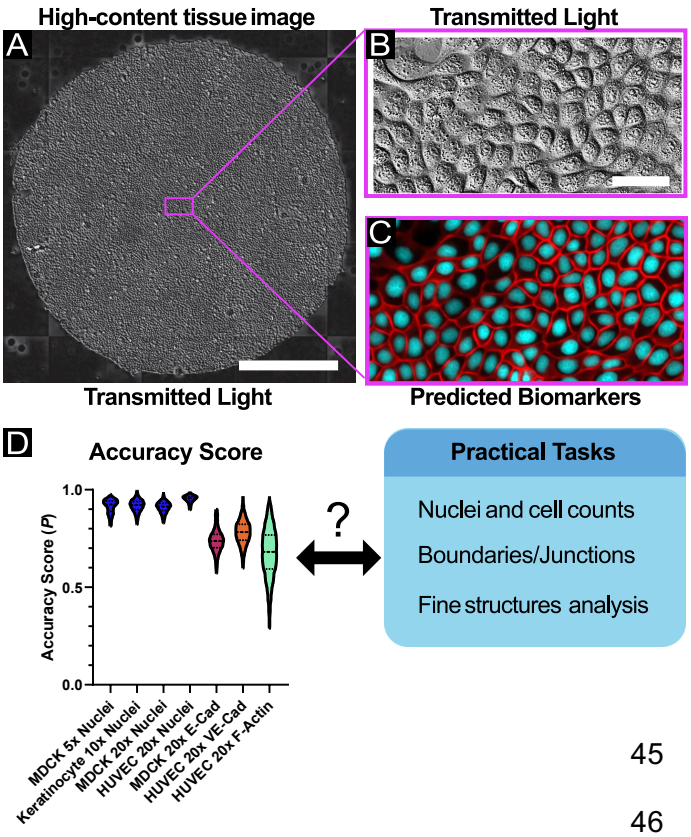


Figure 1. High-content, high-throughput labeling of fluorescent features. (A) Sample large tissue of MDCK cells imaged via transmitted light (DIC). The scale bar represents 1 mm. A sub-region of the large tissue is enlarged in B. (B) A representative image which is given as input to the U-Net processing framework. The scale bar represents 50 μm. (C) The predicted fluorescent features (cell-cell junctions and nuclei) produced by the U-Net, corresponding to the same spatial region as in B. (D) Violin plot of accuracy score results from all experimental datasets. $N > 4400$ raw images and > 20,000 sub-images for all datasets; see Table S1 for summary statistics. Here, the accuracy score (P) represents the Pearson's correlation coefficient between ground truth images containing positive examples of features and their matched reconstructed images. See *Methods.*

Adapting U-Nets for Low Magnification, High-Content FRM

While several FRM methods use computationally complex and expensive networks that rely on Z-stacks of images to capture 2D reconstruction[9], other approaches reconstruct 3D image stacks using a modified U-Net architecture[10]. The U-Net itself is commonly used in machine learning approaches because it is a lightweight convolutional neural network (CNN) which readily captures information at multiple spatial scales within an image, thereby preserving reconstruction accuracy while reducing the required number of training samples and training time. U-Nets, and related deep learning approaches, have found broad application to live-cell

imaging tasks such as cell phenotype classification, feature segmentation[10,14–19], and histological stain analysis[20–23].

Our implementation here provides an archetypal U-Net and framework intended for the cell biology community. Briefly, our workflow is as follows. First, we collected multi-channel training images of cultured cells where each image comprised a transmitted light channel and associated fluorescence channels (labeled using genetically encoded reporters or chemical dyes; see Methods). These images were then broken into 256x256 pix^2 'sub-images' in ImageJ and then input into the network. Such image chopping is necessary for the average researcher to account for the average RAM and graphics cards available on standard workstations. Van Valen et al. examine the trade-off between conv-net receptive field size on feature segmentation accuracy versus training time (and therefore over-fitting of the model): optimization of the receptive field size may lead to further performance gains[14]. These data are then passed through the U-Net network to generate trained weights—the pattern recognition side of the network. Here, the transmitted light images serve as input to the network, which is then optimized to minimize the difference between intensity values of the output predicted images and the intensity values from the ground truth corresponding fluorescence images (e.g. Fig. 1). This process can be extended to full time-lapse video fluorescence reconstruction, making it well suited for high-content live imaging (see Movies S1-4). We have provided all of our code, all raw and processed data, and an extensive researcher manual (DataSpace, GitHub) to encourage exploration of FRM.

As our conventional performance metric, we selected the Pearson's Correlation Coefficient (PCC), which is commonly used in cell biology when comparing the co-localization of two or more proteins, and also used in computer vision to assess spatial-intensity when determining image similarity. However, we observed that naively applying the PCC across our whole dataset skewed the results due to the large number of images containing primarily background

84 (common with high content imaging of oddly shaped or low density samples). This resulted in
85 poor PCC scores as the network tried to reconstruct the pseudo-random background noise. To
86 address this, we report a corrected accuracy score (*P*) representing the PCC of a large subset
87 of images in a given dataset containing positive examples of the feature (nuclei, junctions, etc.)
88 based on an intensity threshold (Fig. S1, see Methods). This approach will improve network
89 performance for datasets containing large amounts of background signal. We also considered a
90 second modified metric, which we call "Segmentation PCC", which reports PCC only on those
91 pixels which exceed an intensity threshold in the ground truth image (Fig. S2, see Methods).
92 However, since many of our key datasets either (1) are used in applications which require
93 background information, such as with nuclei segmentation/tracking, or (2) represent structures
94 on the grayscale spectrum where the background/foreground distinction is highly subjective,
95 such as with MDCK E-cadherin, we found the P metric to be more suited for our analyses. The
96 Segmentation PCC may benefit future researchers in assessing foreground accuracy in
97 alternative FRM applications.

98 To broadly explore the utility of FRM for high-content imaging applications, we captured
99 transmitted light images using 4X, 10X, and 20X air objectives using either Phase Contrast or
100 Differential Interference Contrast (DIC), and collected data across 3 different cell types—renal
101 epithelial cells (MDCK), primary mouse skin keratinocytes (KC), and human umbilical vein
102 endothelial cells (HUVEC). Variable training set sizes were tested to also explore the effect of
103 data set size on 'accuracy'—a key practical aspect of designing an FRM study. The biomarkers
104 we trained against comprised a nuclear dye (Hoechst 33342), an F-actin dye (SiR-Actin) and
105 genetically encoded fluorescence reporters for E-cadherin and VE-cadherin. Traditional
106 Accuracy Scores for each of these are summarized in Fig. 1D and Table S1, and we will next
107 present case studies from each of these data sets before concluding with a discussion of how
108 'accuracy' relates to visual quality to help researchers design experiments for FRM. Specifically,

109 we apply commonly used ‘down-stream’ analyses from high-content screening to both ground
110 truth and FRM predictions to evaluate how closely analysis of a prediction matches that of the
111 ground truth. While we necessarily show representative data here, we provide the statistical
112 distribution for all accuracy scores, as well as complete 2D cytofluorogram heatmaps of ground
113 truth to prediction mapping across all of our data for each biomarker (Fig. S3).
114
115 **Results.**

116 ***Demonstration of FRM for low-magnification nuclear fluorescence reconstruction and***
117 ***analysis***

118 One of the most common computational image processing needs for screening and low-
119 magnification image is nuclei detection or segmentation, which enables cell counting, time-lapse
120 tracking, and statistical analyses of ensemble distribution and geometry. While a variety of
121 traditional image processing approaches exist to extract nuclei from phase or DIC images, such
122 techniques require extensive fine tuning, ultimately only work for certain cell types, and often fail
123 to work at all with DIC images. The most reliable and standardized technique by far is using a
124 vital dye (e.g. Hoechst 33342 or DRAQ) to stain the nuclei. However, Hoechst requires cytotoxic
125 UV illumination while DRAQ (far-red fluorescence characteristics) has been linked to cell cycle
126 alterations due to its chemistry [24–26]. Both dyes also exhibit loss of signal over extended time-
127 lapse imaging. Alternately, genetic reporters such as H2B nuclear labels can be engineered into
128 cells (e.g. transfection, viral addition, etc.), but this adds more overhead, incurs phototoxicity,
129 and still requires a dedicated fluorescence channel for a relatively simple structure (the nucleus)
130 in lieu of a more complex or useful label. Hence, there is a clear practical benefit to fluorescence
131 reconstruction of cell nuclei, especially in time-lapse imaging where freeing up a channel and
132 reducing phototoxicity are each quite valuable. Further, fluorescent reconstruction of nuclei

supports any software or analysis pipeline that might normally be employed with fluorescent nuclei data, meaning that workflows need not be altered to leverage FRM data here.

To validate low-magnification, high-accuracy nuclear FRM, we collected data in both MDCK renal epithelia cells (5X phase contrast, Figs. 2A-D) and primary skin keratinocytes (10X phase contrast, Figs. 2E-H) while using Hoechst to label nuclei and generate our ground truth training data. Representative images are presented as a sequence of phase contrast, nuclear ground truth (green), network predictions (red), and a merged overlay (yellow for a perfect merge). The Accuracy Score (*P*) is included for context, while the statistical distributions of *P* for each cell type are presented in Fig. 2I demonstrating the actual network performance. The performance with the keratinocyte data is particularly striking given how irregular and poorly resolved the cells appear in phase contrast (confounding traditional segmentation).

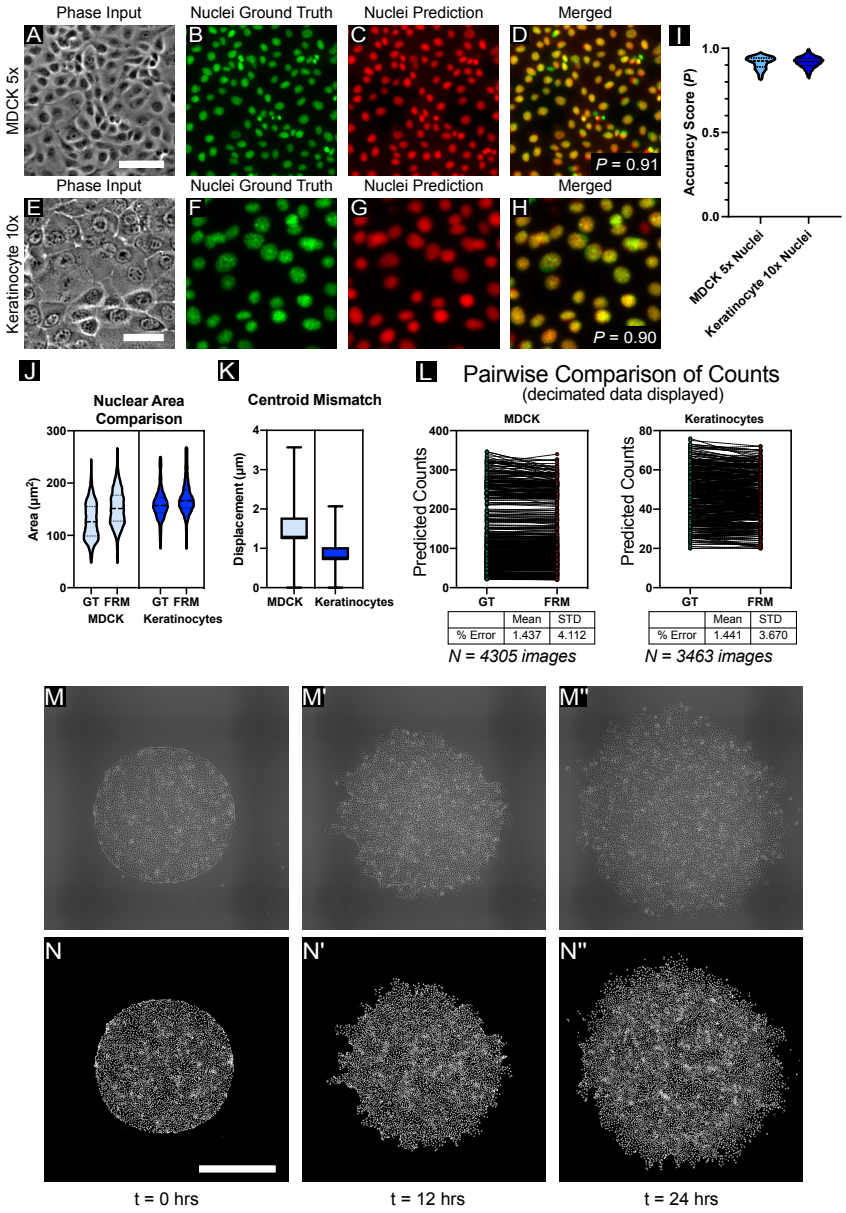


Figure 2: Low-magnification nuclei reconstruction.
(A) Representative transmitted-light image of MDCK cells at 5x magnification, with corresponding: (B) ground-truth nuclei, stained with Hoecscht 33342 and imaged with blue fluorescent light; (C) nuclear prediction produced by the network; and (D) the overlay of (B) and (C) displayed in red and green, respectively. The raw accuracy score between (B) and (C) is given at right. The scale bar is 100 μm. (E) Representative transmitted-light image of keratinocyte cells at 10x magnification, with corresponding (F, G, H) ground truth nuclei image, predicted nuclei, and overlay, respectively. Scale bar is 50 μm. (I) Comparison of the accuracy score distributions across the 5X MDCK and 10X Keratinocyte datasets, $N > 4400$ test images for each dataset (see Table S1). (J,K) A comparison of nuclear area estimations and centroid-centroid displacement estimations, respectively, for the two low-magnification datasets considered here. See *Methods*. (L) Pairwise comparisons of predicted cell counts for MDCK (left) and Keratinocytes (right). Summary statistics and *N* shown below plots. (M-N) Sequence of phase images (M-M") from a time-lapse at 0, 12, and 24 hours of growth, with corresponding nuclear predictions (N-N") respectively. Input data consists of MDCK WT cells imaged at 5x magnification and montaged; the U-Net was applied in a sliding-window fashion to predict small patches of the image in parallel. Scale bar is 1 mm. Higher resolution movies showing the migration dynamics can be seen in Movie S1.

169 The network performs visually well in both cases, with $P \sim 0.9$, but to represent what that means
170 in practice, we quantified disparities in the predictions with respect to nuclear size for geometric
171 accuracy (Fig. 2J) and centroid error to reflect positional accuracy (Fig. 2K). In both nuclear
172 area cases, the U-Net slightly overpredicts area, likely due to slight noise in the predictions
173 blurring the predicted nuclei and effectively increasing area. However, the distributions from the
174 violin plots are quite similar in structure, and the predictions are well within the usable range for
175 practical cell counting and segmentation. With respect to nuclear centroid localization, mean
176 errors span 2 microns (5X MDCK) to 1 micron (10X KCs). The improvement from 5X to 10X can
177 likely be attributed to the resolution increase in the magnification, but in both cases the errors
178 are quite small and more than sufficient for standard cell counting, nuclei tracking, and neighbor
179 distribution analyses. *Finally, we compared basic cell counts using a pair-wise comparison plot*
180 *(decimated), which demonstrated a mean count error over the entire datasets between ground*
181 *truth and FRM predictions of ~1.4% (Fig.2L).* Whether a higher *P* would be beneficial would
182 depend on the specific analysis in question—here, the accuracy is more than sufficient.
183

184 As a final demonstration of the utility of low-magnification reconstruction and nuclear tracking,
185 we input legacy data from a 24 hr time-lapse experiment of the growth dynamics of large
186 epithelia (2.5 mm^2 , 5X) and the network output a reconstructed movie of nuclear dynamics
187 (Figs. 2M-N, and Movie S1) compatible with standard nuclear tracking algorithms (e.g.
188 Trackmate in FIJI). Images were captured every 10 minutes, and previous efforts to perform this
189 experiment using fluorescent imaging of Hoechst resulted in large-scale cell death, hence FRM
190 proved highly effective both as an alternative nuclear labeling approach for large-scale, long-
191 term imaging, and as a means to reprocess pre-existing, legacy datasets.
192

193 ***Reconstructing cell-cell junctions for segmentation and morphology applications***

194 Cell-cell junctions and cellular boundaries in cellular ensembles have implications spanning the
195 epithelial-mesenchyme-transition (EMT), tissue mechanics, and tissue maturation[27–29] and are
196 of broad interest from cellular biophysics to high content screening. However, there are no vital
197 dyes for junctional proteins (e.g. E-cadherin), necessitating either antibodies or genetic

reporters. In the absence of a specific marker, cell boundaries are relatively difficult to accurately segment, especially from DIC images (Fig. 3A), and proxy techniques such as Voronoi tessellation from nuclei data often fail to capture cell shape and organic features such as curved boundaries. Instead, junctions and boundary data most commonly come from biomarkers such as E-cadherin, so we trained our U-Net using MDCK cells stably expressing E-cadherin:RFP (Ecad:dsRed) and imaging with a 20X/0.75NA objective—a well-balanced objective favored for high-content imaging and immersion-free time-lapse imaging.

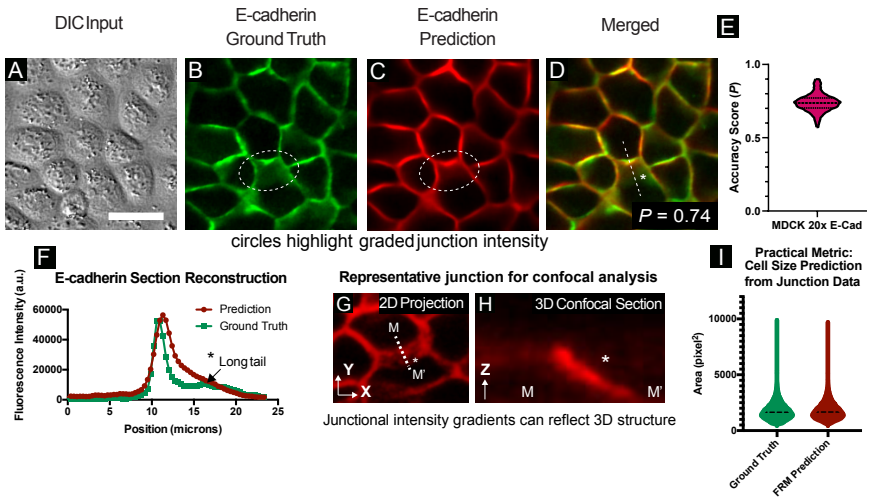


Figure 3: Cell-cell junction reconstruction from DIC data and capturing otherwise invisible morphology (A-D) Images of MDCK WT cells at 20x magnification were processed using a neural network trained to reconstruct cell-cell E-cadherin junctions. Representative ground truth features are shown alongside, and merged with, network predictions. The scale bar is 30 μm. (E) Ensemble statistics for E-cadherin reconstruction; N = 4539 test images, see Table S1. (F) Line sections from identical spatial regions in (B) and (C) highlight the accuracy of predicted fluorescence intensity across cell-cell junctions (normalized to 16-bit histogram). From 2D transmitted light input (A), 3D structures may be detected. (G, H) Representative cell-cell junction and corresponding confocal section, highlighting the relationship between 2D junction signal and 3D features. '*' in (G) is 7 μm above the basal plane. (I) Practical metric assessing estimated cell areas for entire E-cadherin training set (~30,000 individual cells). Junction segmentation was used to calculate cell areas for ground truth and FRM predictions and the distributions of detected cell areas were plotted for comparison. Ground truth mean area was 1813 pix^2 ; FRM predicted mean area was 1791 pix^2.

The U-Net was able to reconstruct E-cadherin junctions with high visual accuracy, as shown in the sequence from Figs. 3A-D. While P = 0.74, the reconstruction is quite spatially accurate, which is unexpected given how difficult it is for humans to detect cell-cell junctions by eye in a DIC image. To better highlight the accuracy and utility of junctional FRM, we explored how the

214 network reconstructed a subtle 3D feature of epithelial junctions where a slanted junction is
215 formed between two cells by one cell pushing slightly under another (the region enclosed in the
216 dashed oval in Figs. 3B-D). Such slanted junctions may indicate a degree of fluidity or direction
217 of migration and are also impossible to discern by eye. We quantified the accuracy of the FRM
218 image by taking a line section perpendicular to this slanted junction (Figs. 3D,E) and comparing
219 the profiles of the ground truth and the FRM image. In this line section, intensity values up to
220 and including the peak value are similar, and intensity values exhibit a graded decay within the
221 slanted junction, indicating that the FRM network is able to capture subtle 3D information from
222 the 2D input image. To emphasize the 3D nature of this feature, a representative Z-section from
223 an E-cadherin junction imaged by scanning confocal is shown in Figs. 3F,G. As a demonstration
224 of a common practical analysis performed using junctional data, we calculated the distribution of
225 cell areas in both ground truth and FRM predictions (see Methods). Shown in Fig. 3I, we
226 observed very similar area distributions with mean calculated cell areas over ~30,000 individual
227 cells varying by ~1%. FRM can again be used for high-fidelity reconstruction during a timelapse,
228 allowing both nuclei and junctions to be predicted throughout long acquisitions (see Movie S2).
229 A large dataset of reconstructed junctions and nuclei can be found in our supplemental data
230 repository (http://doi.org/10.5281/zenodo.3783678, see *Code and Dataset Availability*) for
231 comparison. Overall, our network captures junctional intensity and geometry, both of which are
232 invisible to the eye in the DIC input image.

233 ***Fine structure reconstruction***

234 In practice, high content imaging is inherently a trade-off between throughput and resolution.
235 The more detail we can extract from lower magnification images, the more efficient the imaging
236 and analysis. Here, we demonstrate the practical performance of FRM and a 20X/0.8NA
237 objective to reconstruct fluorescence signatures for several useful sub-cellular markers using
238 HUVEC cells that stably expressed VE-Cadherin:YFP (mCitrine) and were labeled with Hoechst

239 33342 (live nuclear dye) and SiR Actin (infrared live actin dye). Processed timelapse data (see

240 Movie S3) highlights the variation of these fluorescent features given the same input (DIC)

241 image shown in Fig. 4A.

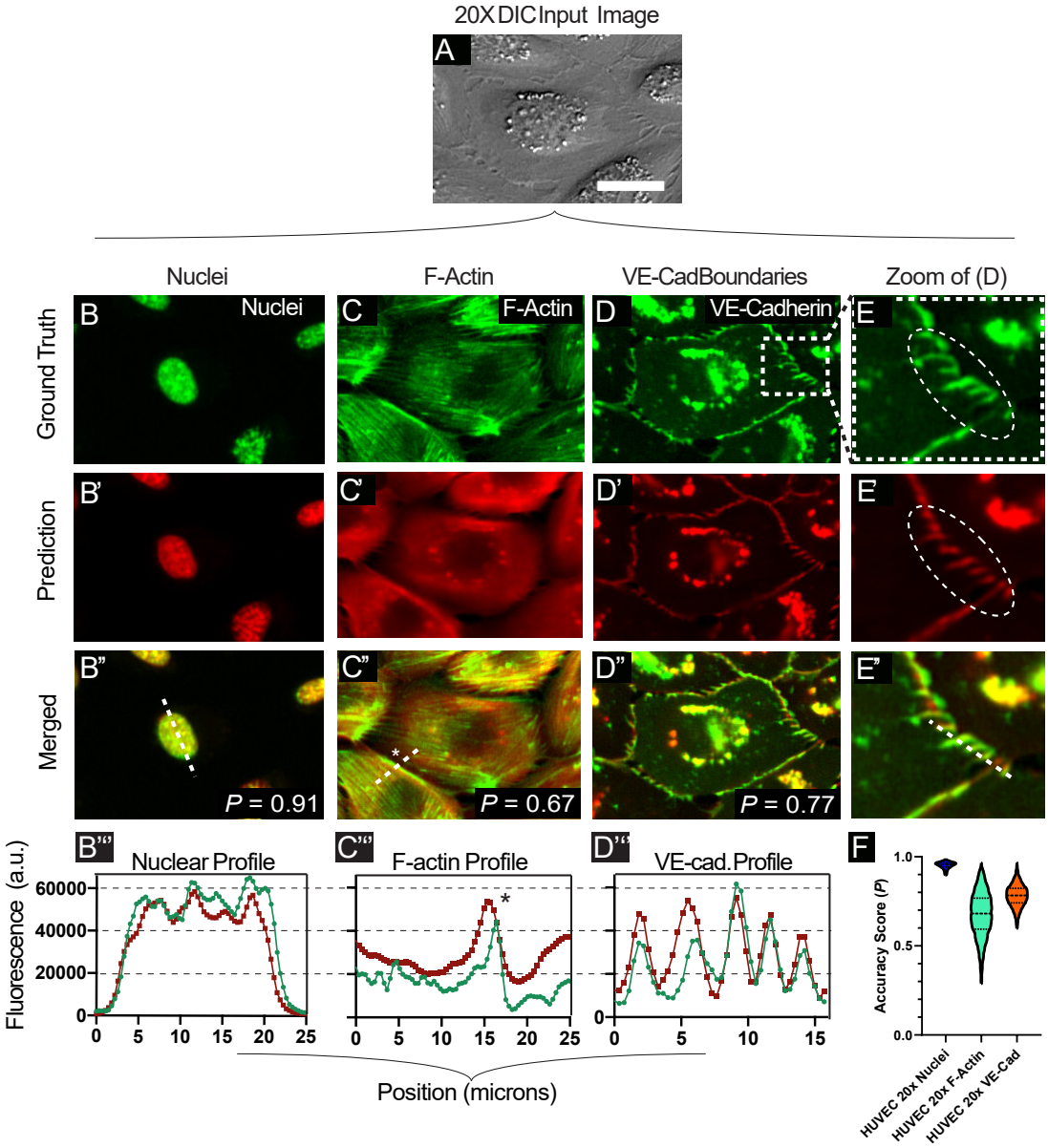


Figure 4: Coarse-to-fine feature reconstruction.
(A) A representative transmitted-light image of HUVEC cells at 20x magnification with its corresponding structures of varying scale (B-D). The scale bar represents 30 μm. (B-D) the relatively large nuclei, the finer VE-Cadherin structures, and thin F-Actin filaments which are not readily resolved by the network. Ground truth fluorescent features are displayed alongside, and merged with, network predictions. (E) displays zoomed-in portions of images shown in (D). Line sections from (B", C", and D") are displayed graphically in (B''', C''', D'''), to enable intensity comparisons across the ground truth and predicted features. (F) summarizes the distribution statistics, clearly showing the uncertainty in F-actin, with tighter reconstruction for nuclei and VE-cadherin. N = 5500+ test images for these datasets.

242 As a baseline, we characterized prediction accuracy for cell nuclei as the nucleus itself is
243 relatively low resolution, but detection of sub-nuclear features requires higher accuracy. The
244 Fig. 4B column demonstrates FRM performance for 20X nuclei including a line section through
245 both the bulk structure and sub-nuclear granules. Visually, the FRM image is quite accurate,
246 and P = 0.91 in this case. The line section easily captures the bulk form of the nucleus, but does
247 not quite capture the texture inside the nucleus, although it does capture the rough form.

248 Next, we trained the network on identifying Actin after first staining HUVECs using the SiR-Actin
249 live imaging dye. Here, the column in Fig. 4C shows significantly reduced performance as the
250 fine F-actin filaments visible in the ground truth fail to be reconstructed in the predictions (P =
251 *0.*67) with the exception of some of the cortical filaments at the very edge of the cells (see the
252 line profile). We hypothesize this is primarily due to fundamental limitations of DIC imaging and
253 the lack of contrast for intracellular F-actin, but it may also be due to the network overprioritizing
254 cortical filaments and the diffuse cytoplasmic signal. However, in practice we found that these
255 FRM data were useful for general cell body detection and potential segmentation analyses due
256 to the relatively homogeneous reconstructed fluorescence in the cytoplasmic space.

257 Finally, we trained the network with VE-Cadherin:YFP data in an attempt to reconstruct not only
258 cellular borders, but also the well-characterized, nano-scale membrane fingers that develop in
259 endothelial cell-cell junctions and indicate the direction of front-rear polarity in each cell [30]. In
260 contrast to the actin performance, FRM proved far more capable here and readily detected both
261 general VE-Cadherin boundaries (Fig. 4D column) and the membrane fingers (Fig. 4E column),
262 although P = 0.77 still seems quite low and likely relates to the network attempting to
263 reconstruct the more variable granules in the center of the cell, which are irrelevant for
264 junctional analyses. While VE-Cadherin protrusions and boundaries are sometimes detectable
265 by eye in DIC (as in Fig. 4A), they are still quite subtle in the best case, and developing a
266 traditional computer vision process to detect and extract them has not been demonstrated,

267 thereby again highlighting the practical value of FRM to reconstruct not only fluorescence, but
268 also key morphological markers that are much easier to analyze in the FRM image than in the
269 DIC image. The statistical accuracy distributions are shown in Fig. 4F, where the spread of the
270 data in F-actin indicates the lack of reliability, while the tighter distributions for nuclei and VE-
271 cadherin indicate more useful reconstructions. Again, the value of FRM depends on the specific
272 question and context, and the decision of whether it is 'good enough' at detecting fine structures
273 rests with the end researcher. For reference, short movies of several markers are presented in
274 Movies S3-5, and a large set of representative images can be found in our supplemental data
275 repository. As a final note, training the network with large and varied datasets enabled it to
276 begin to predict statistically rarer events, such as mitotic divisions, as shown in Movie S6,
277 indicating the benefits of training against time-lapse data where dynamic phenomenon have a
278 higher chance of occurring over time.

279 ***Comparing FRM visual performance to P scores, training set size, and convergence***

280 A key feature of FRM is that its performance can often be increased by collecting more training
281 data, which in turn ought to improve *P*. However, *P* will never be perfect, nor is *P* necessarily
282 the best metric to go by when determining if an FRM image is 'good enough', as clearly the
283 context matters and the key question is 'good enough for what?' Hence, we sought to provide
284 several examples of how the size of the training set affects both *P* and the actual visual
285 accuracy or quality of the resulting FRM predictions.

286 To do this, we first swept through different sizes of training sets for many of the biomarkers
287 presented earlier (see *Methods*). Briefly, we selected random subsets of very large datasets and
288 trained the U-Net from scratch with these subsets. This process was repeated for different
289 fractions of the complete dataset to capture the FRM performance versus training set size. The
290 relationship between *P* and training set size is shown in Figs. 5A,B, where we show both how a

291 single camera images is first chopped into 16 (typically) sub-images for analysis (Fig. 5A), and
292 how *P* and training set size vary relative to either camera images or sub-images (Fig. 5B).
293 These data demonstrate that the rate of change in quantitative quality (*P*) vs. training set size is
294 monotonic across different biomarkers, with more data yielding higher accuracy scores with
295 diminishing returns. To relate these data to a more practical readout, we demonstrate how
296 nuclear count performance depends on training set size using our 5X and 20X MDCK datasets
297 (Figs. 5C,D). These data emphasize both that FRM performance variance goes down with
298 larger training sets, but also that relatively small numbers of training images can produce useful
299 results.

300 For a clearer visual comparison of performance, we also provide FRM results from the reduced
301 training sets for the 20X HUVEC and MDCK data (Figs. 5E,F; respectively). Here, the input and
302 ground truth data are presented alongside representative FRM predictions from networks
303 trained with different numbers of training images (noted below each image). As a single image
304 from the camera is split into sub-images for training, a fractional image (e.g. Fig. 5E, $1/16^{th}$
305 column) implies that the network was trained on just a small crop from a single micrograph.
306 Again, these data all indicate that FRM quality varies directly with the size of the training set, as
307 expected.

308 Further, the actual predicted images shown in Figs. 5E,F offer further nuance because they
309 demonstrate that training the network against even a single image could be sufficient to capture
310 nuclei for the purposes of tracking or segmentation, while just 6 images would be sufficient to
311 capture cell shape and junctional geometry in epithelia assuming the researcher were willing to
312 perform some simple manipulations such as background subtraction. There is an obvious
313 performance increase for both cadherins when the training set comprises several hundred
314 images, but it is difficult to visually detect a difference between nuclei reconstructed from 6 or
315 400 images.

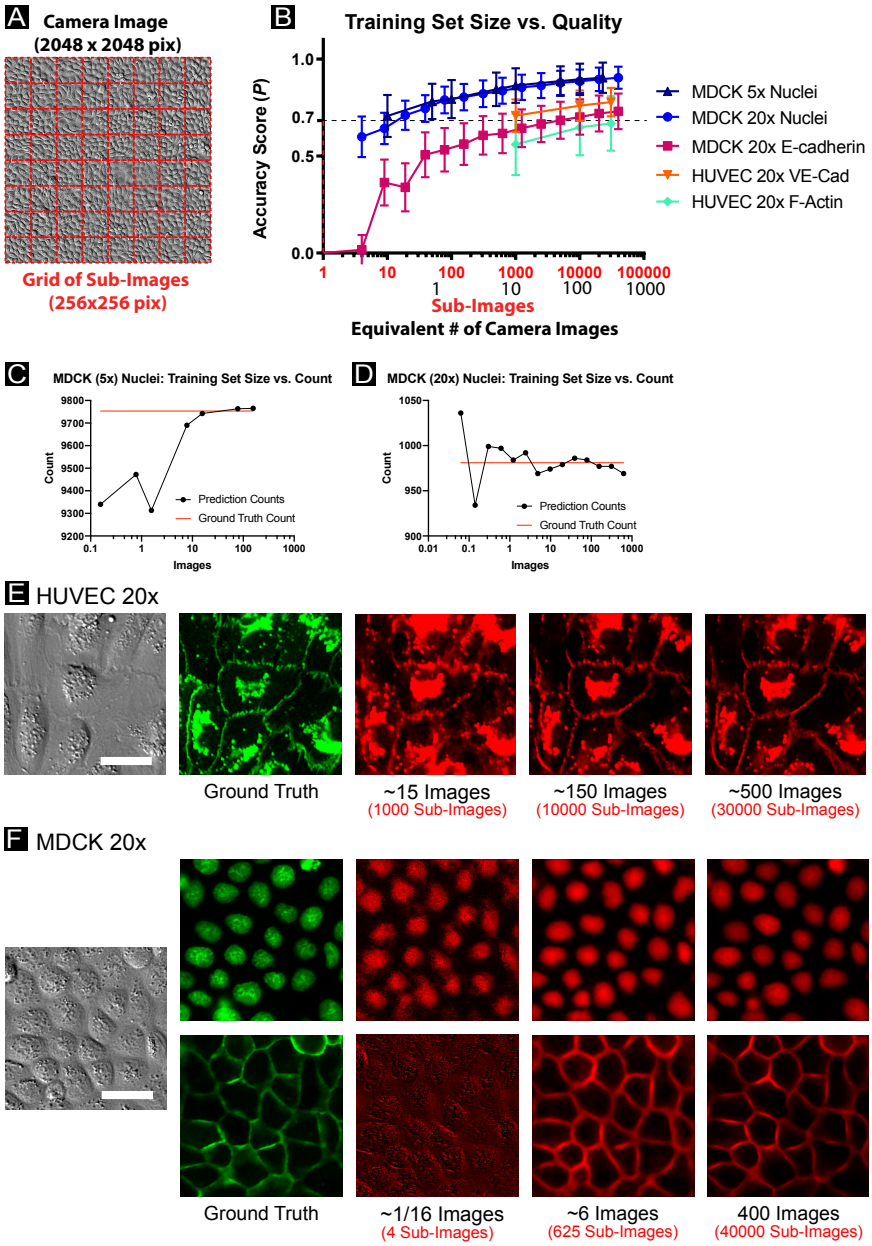


Figure 5: Impacts on prediction accuracy from smaller training sets.
(A) Cropping a sample image into 64 sub-images. (B) A comparison of network prediction accuracy as a function of training set size. The U-Net is trained with the complete dataset as described in Table S1 for each experimental condition. Then, random images representing a fraction of the total training set is used to train a new U-Net from scratch. The average number of cells per sub-image were: MDCK 5x: 107 cells per sub-image; Keratinocyte 10x: 26 cells per sub-image; MDCK 20x: 11 cells per sub-image; HUVEC 20x: 3 cells per sub-image. (C, D) display nuclei counts for representative images of MDCK cells (5x, 20x magnification, respectively) as a function of training set size, with each model independently trained. Practical readouts such as nuclei count may vary widely with small training set size. (E, F) display representative images for the HUVEC 20x dataset and the MDCK 20x dataset, respectively, with predictions shown for various training set sizes. All scale bars represent 30 μm.

318 We next assessed how the number of training epochs (and associated convergence) affected
319 network performance. This is valuable because some researchers may have a great deal of
320 training data but limited computational resources. Here, we present data from several key
321 datasets in Fig. 6 where identical U-Net models were trained for a fixed number of epochs (1, 5,
322 10, 50, 100, or 500 epochs) and the accuracy score (*P*) (Fig. 6A). Additionally, we report the
323 actual training hours (Fig. 6B) based on our specific hardware (NVIDIA Tesla P100 GPU).
324 Again, we relate these data to practical performance using nuclei counts, this time plotting how
325 the distribution of counts evolve as a function of training epochs and relative to the ground truth
326 distribution (Figs. 6C,D). Overall, increasing training epochs correlates with improved
327 performance, as highlighted visually in Figs. 6E,F where the evolution of predictions of nuclei
328 and E-cadherin are presented relative to training epochs. In general, researchers limited by
329 computational bottlenecks may opt to reduce training time as the network converges to 'good
330 enough' results as a function of their practical measure of performance. To emphasize this,
331 consider Fig. 6E, where the dashed circle indicates how regions of higher cell density
332 reconstruct poorly for low training epochs; highlighting the importance of assessing practical
333 model performance on a range of potential biological and imaging conditions.

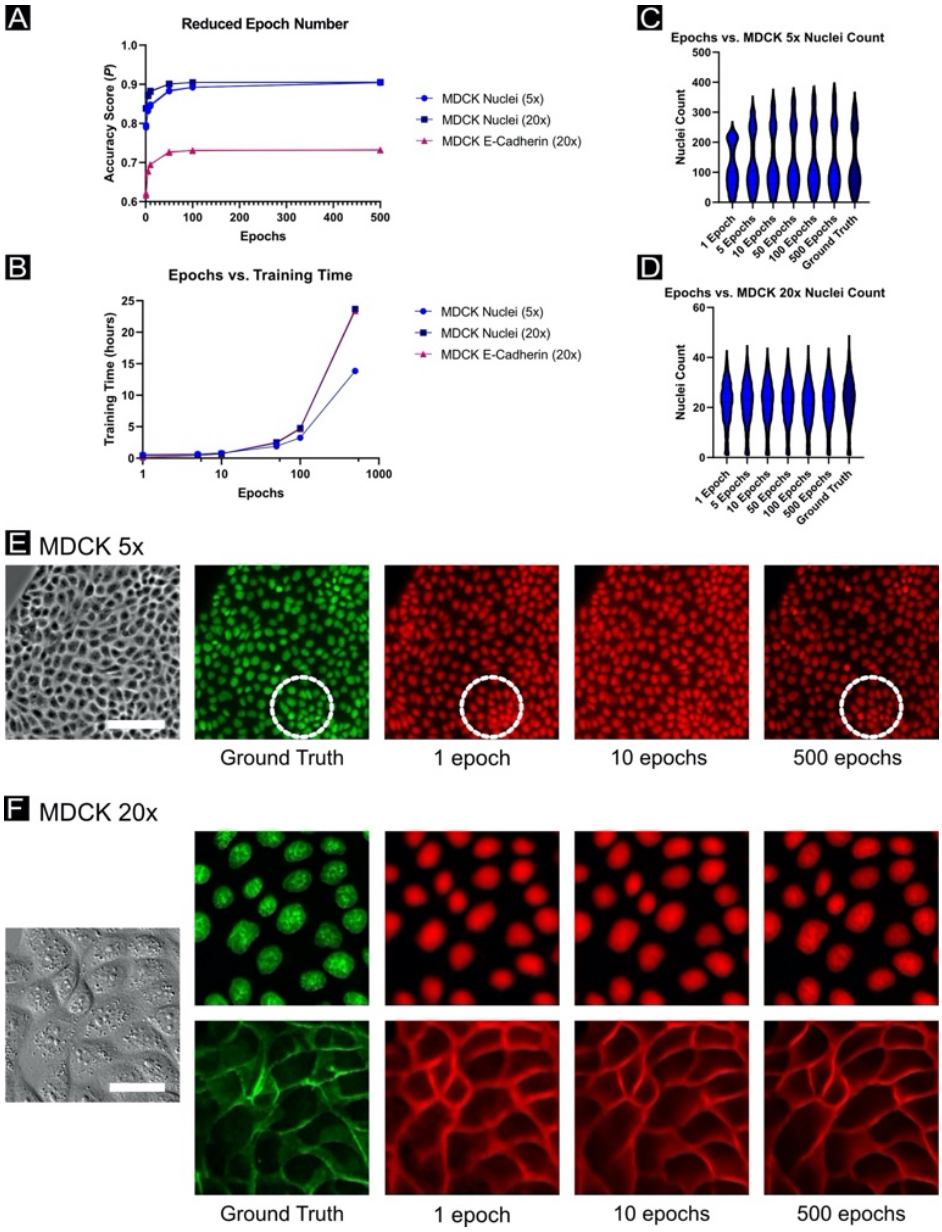


Figure 6: Impacts on prediction accuracy from reduced number of training epochs.
(A) A comparison of network prediction accuracy *P* as a function of the number of training epochs. A standard U-Net model is re-trained from scratch for each experimental condition (1, 5, 10, 50, 100, 500 epochs), with *n*=3 per condition (with accuracy scores overlaid in the plot and visually similar). (B) The time in hours to train the reduced epochs models shown in (A), on an NVIDIA GeForce GTX 1070 Ti GPU. (C, D) display nuclei counts for the segmented prediction images compared to the segmented ground truth image, with predictions produced from the reduced epoch models corresponding to the MDCK datasets at 5x and 20x magnification, respectively. (E, F) display representative images for the MDCK datasets at 5x and 20x magnification, respectively, with predictions shown for various numbers of training epochs. The dotted circle indicates a region of higher cell density; in such regions, practical readouts such as nuclei count from segmentation may vary widely with lower training epoch number. Scale bar in (E) represents 100 µm; scale bar in (F) represents 30 µm.

As a final analysis, we considered improving FRM performance by altering the network architecture. Here, we first compared the standard U-Net to a neural network architecture which was essentially two U-Nets stacked end-to-end with additional residual connections. Such an approach has been shown to improve network depth and performance in other applications[31–33]. Here, however, we observed no benefit to training a deeper network (see Fig. S4). Further, given the significant temporal and computational cost, we advise against its use for this kind of FRM. While no substantial benefit was observed as the result of data augmentation (flip, rotation, zoom) on standard U-Nets trained with large and varied datasets (see Fig. S5), researchers should test data augmentation techniques on their own data to assess accuracy gains prior to widespread application. Alternately, we explored the role of the loss function, testing our Pearson's-based loss function against the traditional Mean-Squared-Error loss function and found no significant difference (Fig. S6; Methods). Hence, we conclude that our minimal U-Net implementation performs well as a foundation for a variety of daily analysis tasks without requiring significant fine tuning.

Discussion

Limitations of existing accuracy metrics and the importance of context

Our data—collected from actual, real-world analyses, highlight the limitations of using traditional accuracy metrics from computer vision for biological image analysis. Specifically, while there is a general relation between an improved *P* Accuracy Score and FRM quality, it is not linear nor intuitive how to determine what is 'good enough' given only a *P* value devoid of context for a specific analysis. Further, and most critically, FRM does not reconstruct images according to human imperatives. The U-Net only optimizes via the specific loss function it has been given (e.g. Mean-Squared-Error or the Pearson's coefficient). What the computer considers 'good' need not match our own assessments of value and quality.

As a practical example, compare the FRM performance for E-cadherin (P = 0.73; Fig. 3) and F-actin (P = 0.67; Fig. 5). While the accuracy metrics differ by < 10%, the FRM of F-actin only detected peripheral actin cables, otherwise blurring all internal features into a homogeneous signal. Nonetheless, even this plainly 'inaccurate' signal could prove useful for cytoplasmic reconstruction and tracking. In stark contrast, the E-cadherin data was much more visually accurate and also captured key quantitative features of the ground truth such as junctional localization and intensity, and even the subtle intensity gradients representing 3D morphology despite having only a slight improvement in P-values. However, a score of 0.73 is far enough from '1' that it is ambiguous in the absence of a specific analysis, which is why FRM must be evaluated in the context of a given question or analysis, since it will never be perfect.

Our work also highlights how FRM and similar machine learning techniques need performance benchmarks and metrics that are both standardized and, critically, context-aware. For instance, although FRM has a relatively low annotation burden (as the fluorescent channel essentially auto-annotates the transmitted light channel) for training purposes, there still remains the problem of how the researcher knows how a given FRM prediction performs against benchmarks of practical value to that researcher. Critically, we demonstrated here that common, but abstract, benchmarks often fail to clearly reflect the practical value of that FRM reconstruction for the average biologist. In other words, what is 'mathematically good' (e.g. high P, etc.) does not adequately reflect what is 'practically useful'. Hence, while FRM is clearly powerful, it is like any other tool and requires that the researcher deploy and assess it intelligently. Establishing clearer benchmarks, challenges, and more approachable performance metrics would be of great value towards transitioning FRM and similar techniques into standard practice.

387

388 ***Practical considerations for training on new, low-magnification data***

389 We specifically targeted the lower-magnification end of the imaging spectrum to explore how
390 well FRM performed at magnifications more commonly used for high content imaging
391 applications such as timelapse studies of very large cellular colonies or massive screens using
392 multiwell plates. Our data indicate that such magnifications can be effectively combined with
393 FRM for applications spanning nuclear tracking, cell-cell junction analysis, and certain fine-
394 structure reconstruction even at just 20X.

395

396 A particular concern for the average researcher of a complex machine learning process is the
397 size of the dataset required as this can impose potentially strenuous experimental demands.
398 However, our characterization of FRM performance vs. data set size again shows the
399 importance of context as relatively few images are needed to get quite accurate nuclei
400 reconstruction, while a greater number of images are needed for junction reconstruction (Fig. 5).
401 However, we also note that our largest training set size comprised at most 500 camera images
402 at 20X (approximately one six-well plate)—something easily obtained with a standard
403 automated microscope, and still compatible with manual capture. Further, a very common
404 approach in machine learning is to 'augment' an image dataset by performing reflections and
405 rotations on images such that the network perceives each augmented image as a different
406 datapoint, thereby virtually increasing the size of the dataset. We did not perform such
407 augmentation here for the sake of simplicity and transparency, which suggests that significantly
408 smaller datasets, if augmented, could still produce good results.

409

410 As a final note, it would be valuable to be able to transfer a trained network, either from one
411 optical system or laboratory group with collaborators who might have a different system, or
412 across different cell types. We evaluated the former by applying a network trained on a Zeiss

413 microscope (5X phase objective, 6.5 μm/pixel camera) to data captured on a Nikon microscope
414 (4X phase objective, 7.3 μm/pixel camera). While we found that applying a scaling factor to
415 make the effective microns/pixel of the Nikon data match the Zeiss data was able to achieve
416 adequate nuclei reconstruction for basic counting, we by far recommend using networks
417 specifically trained on a given optical system in order to improve reliability (see Fig. S7). We
418 also considered the application of trained networks across different cell types; comparable
419 cross-cell-line analyses have previously been performed[34]. Fig. S8 presents such cross-network
420 results indicating sufficient performance for nuclei predictions that researchers unfamiliar with
421 FRM could begin by testing pre-trained networks on their own data. As with cross-platform
422 applications, we recommend researchers train a model using data which best captures the
423 conditions under which they expect the model to be applied- including consistent cell types.
424

425 ***FRM versus machine-learning segmentation approaches***

426 FRM is a quickly developing technology that exists alongside another popular approach where
427 machine learning is used for feature segmentation[14,19,35]. In the latter case, the network is
428 trained to specifically detect 'features' (e.g. nuclei) as binary objects, whereas FRM instead
429 reconstructs the effective fluorescent image of what a fluorescent label against that structure
430 might show. Both are useful techniques, and the best approach depends on the application.
431 However, there are several unique advantages to FRM. First, reconstructing an effective
432 fluorescent image from auto-annotated data (e.g. chemical dyes, antibodies, fluorescent
433 proteins) obviates the need for any manual annotation or pre-processing—often quite time
434 consuming and subjective. This means that an FRM image can be directly incorporated into any
435 existing analysis pipeline intended for fluorescent images, including traditional threshold-based
436 segmentation approaches. Further, more of the original data is preserved in an FRM image,
437 allowing the capture of things such as fluorescence intensity gradients (e.g. Fig. 3), and features
438 that might be lost during traditional binary segmentation.

439

440 ***Concluding remarks***

441 Here, we characterize the value of fluorescence reconstruction microscopy (FRM) for everyday
442 analysis tasks facing researchers working with cell biology. We specifically highlight the need for
443 individual researchers to explore and evaluate FRM in the context of specific research questions
444 rather than accuracy metrics. We also highlight the surprisingly good performance of FRM even
445 with lower magnification imaging or relatively fine structures such as VE-cadherin fingers.
446 Finally, we have made all of our tools and all training datasets publicly available to improve
447 accessibility and provide a starting point for researchers new to FRM to easily explore it for
448 themselves and to eventually build on and improve.

449 **Methods.**

450 ***Tissue culture***

451 MDCK-II (G-type) cells stably expressing E-cadherin:dsRed were cultured in low glucose
452 DMEM. The MDCK-II culture media was supplemented with 10% Fetal Bovine Serum (Atlanta
453 Biological) and penicillin/streptomycin. HUVEC endothelial cells stably expressing VE-
454 cadherin:mCitrine were cultured using the Lonza endothelial bullet kit with EGM2 media
455 according to the kit instructions. Primary murine keratinocytes were isolated from neonatal mice
456 (courtesy of the Devenport Laboratory, Princeton University) and cultured in custom media[36]. All
457 cell types in culture were maintained at 37°C and 5% CO_2 in air.

458 ***Ethics statement***

459 Our study involved standard mammalian cell type the use of which is approved via Princeton
460 IBC committee, Registration #1125-18. MDCK-II cells stably expressing E-cadherin were
461 received from the Nelson laboratory at Stanford University [36], while the wild-type can be
462 obtained through ECACC. HUVEC cells stably expressing VE-cadherin were received from the
463 Hayer laboratory at McGill University [30], originally based on HUVECs provided by Lonza.
464 Primary murine keratinocytes were isolated by the Devenport Laboratory under a protocol
465 approved by the Princeton University IACUC and shared with the Cohen laboratory directly.

466 ***Preparation of training samples***

467 We collected training data using 3.5-cm glass-bottomed dishes coated with an appropriate
468 ECM. To coat with ECM, we incubated dishes with 50 µg/mL in PBS of either collagen-IV
469 (MDCK) or fibronectin (HUVEC, primary keratinocytes) for 30 min 37 °C before washing 3 times
470 with DI water and air drying the dishes.

471 In order to contain a variety of conditions within a single plate to ensure a broad training sample,
472 we placed silicone microwells into the dishes as described in [[37]] at densities from [1-2x10^6
473 cells/mL] which ultimately allowed for single cells, low density confluent monolayers, and high
474 density confluent monolayers to be captured. Silicone microwells consisted of 3x3 arrays of 9
475 mm^2 microwells into which we added 4 µL of suspended cells in media, allowed them to adhere
476 for 30 min in the incubator (6 hrs for keratinocytes), added media and returned them to the
477 incubator overnight prior to imaging. To further ensure variability, several dishes were also
478 randomly seeded with cells for each cell type.

479 ***Fluorescent labeling for ground truth data***

480 We used the live nuclear dye NucBlue (ThermoFisher; a Hoechst 33342 derivative) with a 1 hr
481 incubation for all nuclear labeling. We used SiR-Actin (Spirochrome) at 10 µM for live F-actin
482 labeling in HUVECs. All other labels were genetically encoded reporters as described.

483 ***Image Acquisition***

484 5X MDCK data was collected on a Zeiss (Observer Z1) inverted fluorescence microscope using
485 a 5X/0.16 phase-contrast objective, an sCMOS camera (Photometrics Prime) and controlled
486 using Slidebook (Intelligent Imaging Innovations, 3i). An automated XY stage, a DAPI filter set,
487 and a metal halide lamp (xCite 120, EXFO) allowed for multipoint phase contrast and
488 fluorescent imaging. All epifluorescence imaging was performed using a Nikon Ti2 automated

microscope equipped with a 10X/0.3 phase objective, a 20X/0.75 DIC objective, and a Qi2 sCMOS camera (Nikon Instruments, 14-bit). This same system was also used for the cross-system comparison in Fig. S7. Note that images captured on the Nikon were natively 14-bit, so they were normalized to the 16-bit histogram in ImageJ to make them compatible with standard intensity analyses as would be used by many biologists.

Time-lapse imaging effectively increased dataset size as long as sufficient time was allowed between frames to avoid overfitting in the U-Net. MDCK data was collected at 20 min/frame, while HUVEC and keratinocytes were given 60 min/frame. Standard DAPI, CY5, and YFP filters sets were used. Confocal sections of E-cadherin fluorescence in MDCK cells (Fig. 3) were collected using a Leica SP8 scanning confocal tuned for dsRed excitation/emission.

All imaging was performed at 37 °C with 5% CO2 and humidity control. Exposures varied, but were tuned to balance histogram performance with phototoxic risk. Data with any visible sign of phototoxicity (blebbing, apoptosis, abnormal dynamics) were excluded entirely from training.

Data Pre-Processing and Training

Prior to input to the network, raw images were segmented into 256x256 pixel^2 sub-images, ensuring consistent slicing across the transmitted-light image and the corresponding fluorescent image. The splitting times are reported as a function of image size (see Fig. S9); this plot was generated by cropping the same image into smaller TIFF images, and applying the same splitting procedure to chop the image into 256x256 pixel^2 sub-images. Input images were then normalized by statistics collected across all images in each channel: that is, by subtracting from each image the mean and dividing by the standard deviation. A test-train split was applied, such that a random 20% of the total images were held out to comprise the test set. Additionally, 10% of the training data subset were held out for validation as is standard.

The U-Nets shown in Figure 1 were implemented and trained in Python using the Keras/Tensorflow [38,39] packages, since we find these to be more researcher-friendly for high-level deep learning implementations than alternative open-source deep learning libraries (e.g. PyTorch). Training was performed using the Tensorflow ADADELTA optimizer[40]. In the standard training experiments, the mean squared error (MSE) loss function was applied across pixel intensity values in the predicted images compared to intensity values in the ground truth images. Results from the MSE were contrasted with results from two networks trained to maximize the Pearson's correlation coefficient (PCC). The PCC is commonly used in cell biology for evaluating the colocalization of two fluorescently labeled structures. The PCC loss function was defined, for two intensity datasets R and G as:

$$PCC = \frac{\sum_i (R_i - \bar{R}) \times (G_i - \bar{G})}{\sqrt{\sum_i (R_i - \bar{R})^2 \times \sum_i (G_i - \bar{G})^2}}$$

$$Loss = \frac{(1 - PCC)}{2}$$

Sample training loss plots are provided (Fig. S10), reflecting the use of early stopping during training. That is, when the validation loss did not decrease for 75 epochs, the training process terminated. The training and test set sizes and results are provided for all experimental conditions in Table S1, and timing data as a function of epochs is given for key datasets (see Fig. 6).

For several of the experimental conditions, networks were trained using subsets of the original dataset. To do so, the network architecture was fixed, and networks were trained from scratch by using a random subset of matched input-output image pairs from the original training set. The training setup, including hyperparameters, was unmodified regardless of training set size. The original test set was used to compare results for all training set size tests relative to each experimental condition. Likewise, for several experimental conditions, data augmentation was applied in the form of flipping (both horizontally and vertically), rotation (in increments of 45 degrees), and zoom (up to 20% scale). For flips and zoom, built-in Tensorflow functions were utilized. However, the raw data was re-processed to apply random rotations in order to avoid boundary problems using ImageJ/FIJI, resulting in a modified training set which was five times the size of the original dataset.

Data Testing and Image Processing

The Pearson's correlation coefficient (PCC) of each test set is determined by individually computing the PCC between each predicted image, as output by the network, and its corresponding ground truth image. Additionally, an accuracy score (*P)* based on the PCC was devised to more reliably represent the performance of the network. To determine P, we report the PCC on a subset of the test set which selects for only those test images containing positive examples of features (nuclei, junctions, etc.). We construct this subset by manually determining threshold values to distinguish image intensities indicating the presence of the relevant features versus background noise for each test set. That is, the histograms of a subset of the data containing positive examples of features are plotted, and an approximate lower bound on intensity values is estimated to distinguish the features from the background. Then, the histograms of a subset of the data containing only background are plotted to ensure that the threshold value is adequate to label the images as background-only images. The MatLab function rmoutliers() was utilized to remove outliers when *P* is reported for each condition unless otherwise stated. The segmentation PCC score is determined by using the same intensity threshold used in the *P* analysis, and determining which pixels in each ground truth image exceed that intensity value in order to produce a mask image. Then the standard PCC is reported against the pixels in the ground truth image which correspond to that mask and the pixels in the prediction image which correspond to the mask, and this process is repeated across the entire test set.

For certain experimental conditions, a nuclear count or area comparison was performed between corresponding ground-truth and predicted images. Initially, both pairs of output nuclear images were segmented independently using standard Gaussian blur, auto-thresholding, watershedding, and size exclusion (to exclude clusters) in ImageJ/FIJI, and then outliers were removed using the MatLab function rmoutliers() when nuclear area was reported. We additionally report the centroid-centroid displacement values for the same low-magnification segmented images. The ImageJ/FIJI plugin TrackMate was used to determine displacements between the ground truth and predicted images, as if they were two frames of a video. Standard TrackMate settings were used and outliers were removed using the MatLab function rmoutliers() for reporting.

When intensity plots for line slices are reported, a line is selected as an ROI in ImageJ/FIJI, and intensity values are exported for analysis. Cell areas were calculated using either fluorescence E-cadherin junction data or FRM predictions of the same using a large image containing ~30,000 individual cells. We opted to perform analysis in entirely in FIJI to

better represent what the average researcher might use. We performed the following steps to calculate and compare cell areas. Image pairs were first histogram-normalized to 16-bits. Images were then segmented manually using standard intensity thresholding in FIJI with the same threshold (~30% cut-off). Binary skeletonization was used on both ground truth and FRM images to produce cellular outlines. Outlines were then closed using a single dilation operation. After inverting the images, a standard 'Analyze Particles' in FIJI was used to calculate areas inside the boundaries. These distributions were then plotted to allow comparison.

New large transmitted-light images were processed using a sliding-window technique. We processed a large image by analyzing 256x256 pixel^2 patches of the input image with a stride of 64 pixels in each direction. Additionally, the border of each predicted patch was excluded in the sliding-window process, as features near the patch borders are likely to have lower accuracy (often as a function of cells being cut off). The sliding-window predictions at each pixel were then averaged to produce the final large predicted image. Timelapse movies can be processed on a frame-by-frame basis. If scaling was required as described in Fig. 3, the input was scaled in FIJI and then passed to the network for analysis. On a standard desktop with an NVIDIA GeForce GTX 1070 Ti GPU, the time to read in, process, and save out a single 256x256 pixel^2 image was 0.49 seconds, and the time to read in, split, process, stitch, and save out a 1024x1024 pixel^2 image was 10.3 seconds.

Code and Dataset Availability

All code used for pre-processing data, training the network, testing a trained model, and applying the model to new images, along with an extensive researcher manual and pre-trained weight files can be found at:

https://github.com/CohenLabPrinceton/Fluorescence-Reconstruction .

Additionally, our complete testing datasets, corresponding reconstructed images, and a complete README explaining the file structure, are available through our Zenodo server, which can be found at:

http://doi.org/10.5281/zenodo.3783678

Acknowledgements

Special thanks to Gawoon Shim for assistance with HUVEC and keratinocyte data collection.

606 **Supplementary Materials**

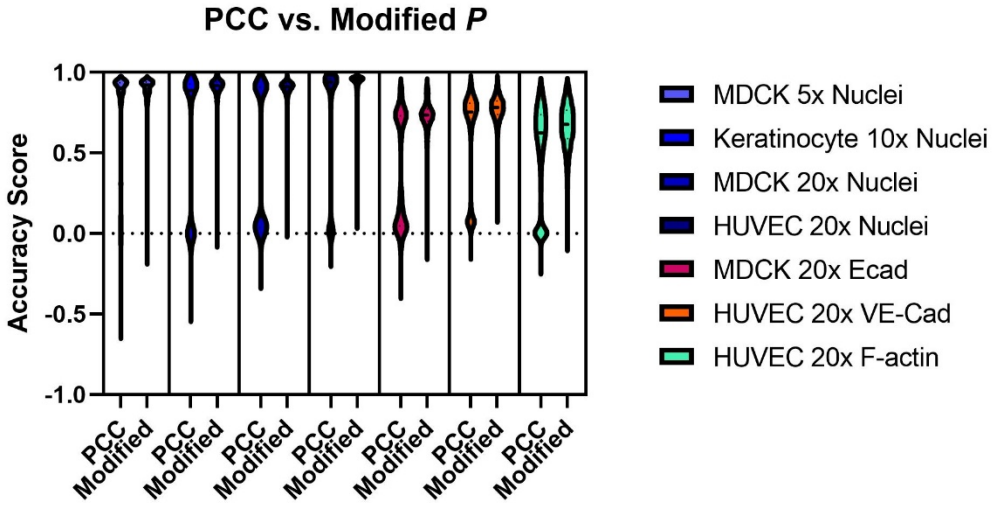
607

Magnification	Cell type	Feature	Training Set Size (sub-images)	Test Set Size	Modified *P:* Reduced Test Set Size	PCC Mean	PCC St.Dev	Modifed *P* Mean	Modified *P* St.Dev
5x	MDCK	Nuclei	22835	5709	4443	0.72529	0.35884	0.90322	0.07801
10x	Keratinocyte	Nuclei	26214	6554	4156	0.57717	0.43833	0.9014	0.093155
20x	MDCK	Nuclei	40000	10000	4556	0.43844	0.43333	0.90327	0.05724
20X	MDCK	E-cadherin	40000	10000	4539	0.37466	0.34172	0.73053	0.091688
20X	HUVEC	Nuclei	30720	7680	5533	0.68936	0.40728	0.93776	0.060886
20X	HUVEC	VE-cadherin	30720	7680	5666	0.60824	0.30398	0.77737	0.074176
20X	HUVEC	F-actin	30720	7680	5820	0.51247	0.30438	0.66808	0.14057

608

609 **Table S1. Accuracy and training statistics for all experimental conditions.**

610



611

612 **Figure S1. A comparison of experimental results in Pearson's Correlation Coefficient**
613 **(PCC), versus a modified accuracy score *P* (see *Methods*)**. To ensure a fairer comparison,
614 outliers were not removed; only intensity thresholding was performed to produce the modified *P*
615 from the PCC. By filtering the PCC results by an intensity threshold in the fluorescent images,
616 we remove low-scoring background images, which bias our accuracy score on the complete
617 dataset. Visual inspection of the plot reveals the low-scoring images as "bumps" near 0.0. Table
618 S1 summarizes the statistics.

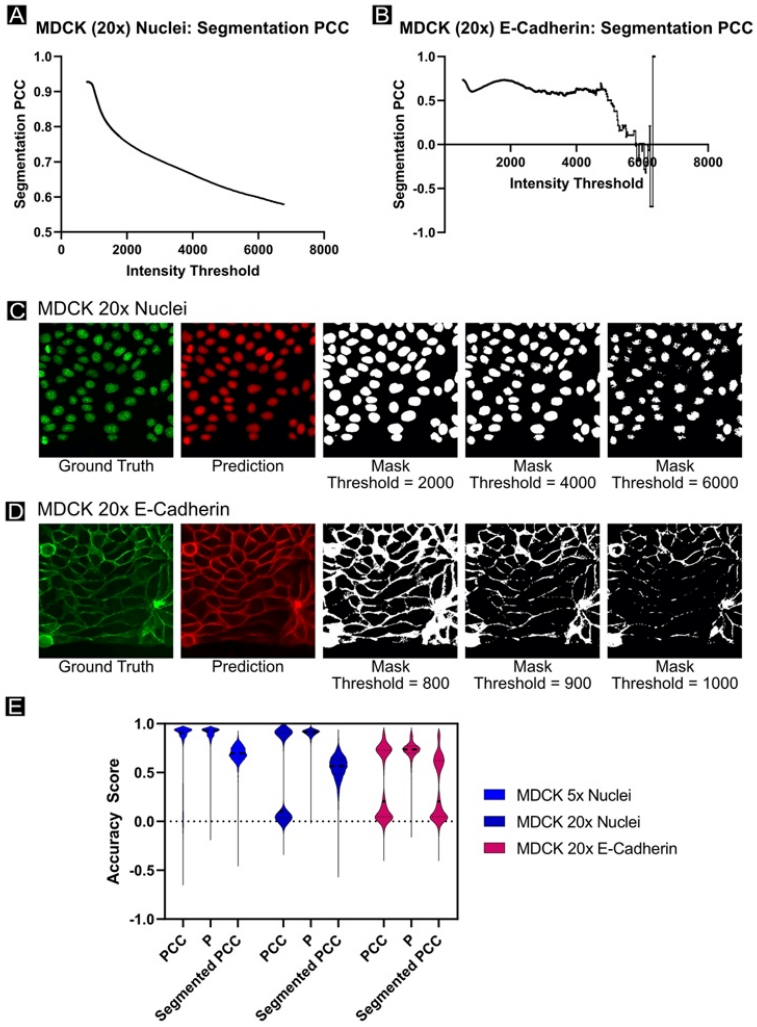


Figure S2. A comparison of the segmentation PCC metric against prior accuracy scores (see *Methods*) on key datasets.
(A, B) The PCC score for the representative "Ground Truth" and "Prediction" images shown in (C, D), respectively, assessed on those ground truth pixels which exceed an intensity threshold value, as the intensity threshold varies. (A) details the segmentation PCC vs. intensity threshold for an MDCK nuclei example, while (B) details this for an MDCK E-cadherin example, both at 20x magnification. The E-cadherin results do not vary smoothly with intensity threshold.
(C, D) Matching ground truth and prediction images for MDCK 20x nuclei and E-cadherin examples, respectively. Segmentation masks are shown corresponding to three different intensity threshold levels. According to the choice of intensity threshold, pixels representing sub-cellular features may be excluded or undesirable background pixels may be included.
(E) Three accuracy scores (PCC, *P*, and segmentation PCC) were compared across three datasets (MDCK nuclei at 5x and 20x magnification, and MDCK E-cadherin at 20x magnification). Outliers were not removed.

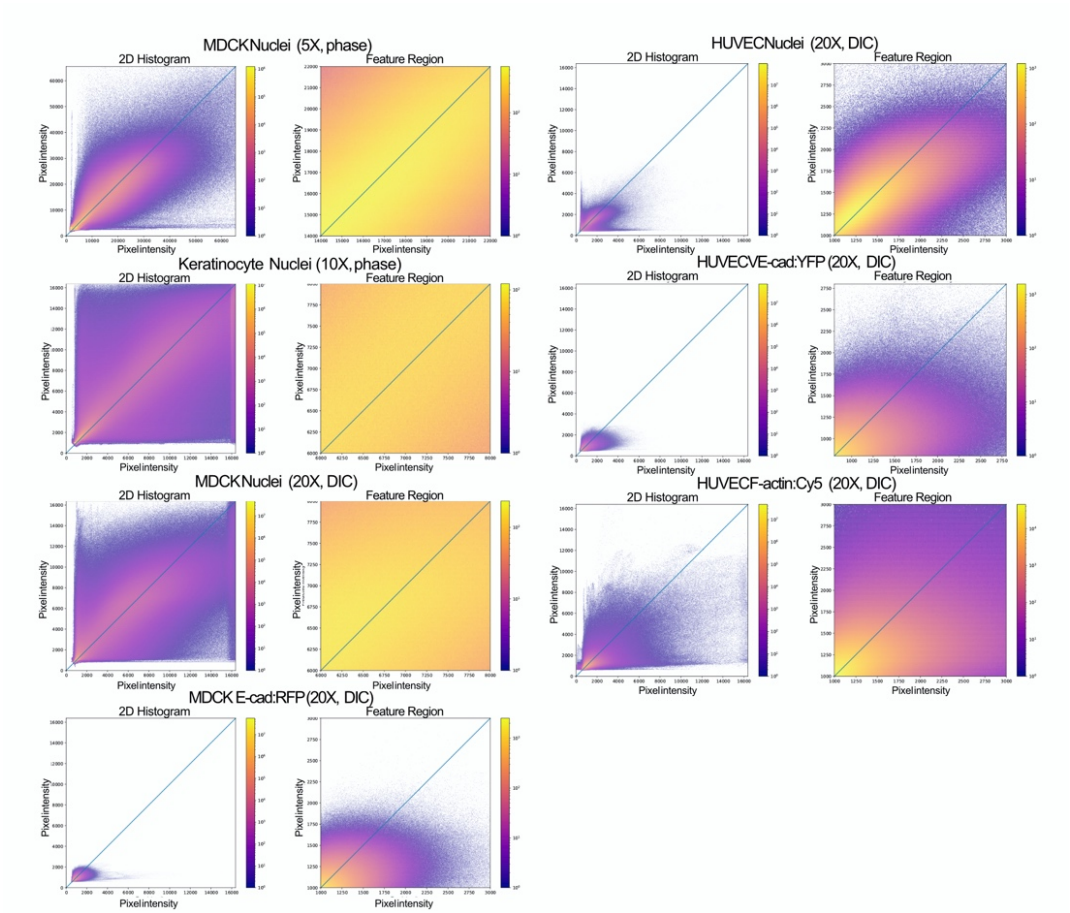


Figure S3. Intensity-intensity heatmaps for all experimental conditions.
(Full 2D Histogram panels) All pixels in the ground truth test set are plotted against all spatially corresponding pixels in the predicted (test) set according to pixel intensity values and colored by density. A perfect prediction would correspond to a heatmap with all values on the 45-degree line. All plots display raw data without any histogram normalization. Axes are representative of camera bit depth (14- or 16-bit), with some biomarkers only filling a portion of the dynamic range.

(Feature Regions panels) Represents a zoomed-in view of the data plotted in the left panel, with the ground-truth axis adjusted to show intensity values above the threshold cutoff value utilized when computing *P* (see *Methods*). This demonstrates the prediction correspondence in the positive-feature region, with low-intensity values (including background noise) excluded- which typically comprise the majority of the test set.

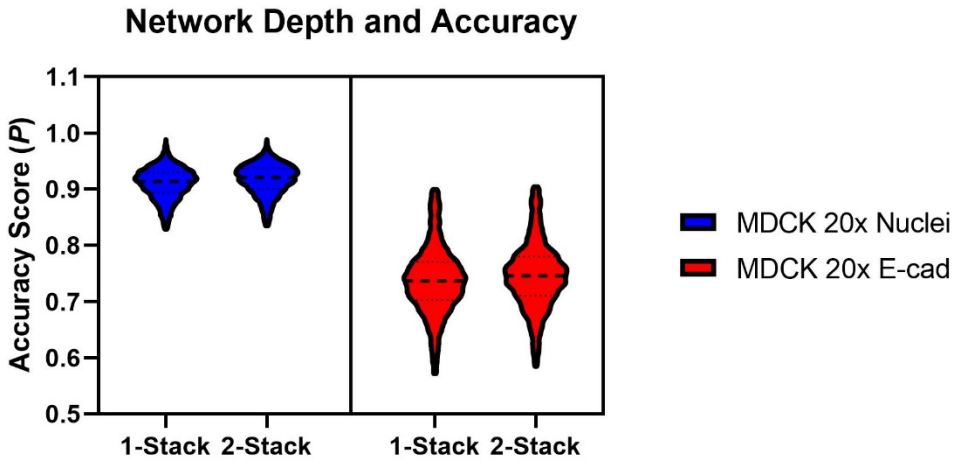


Figure S4. Representative accuracy results for a dataset trained on the standard (1-stack) U-Net, compared to a network comprised of two U-Nets stacked back-to-back, with residual connections (2-stack). Training conditions were otherwise unchanged. Accuracy scores, as reported in terms of the modified *P* (see *Methods*) were comparable. *N* = *1*0,000 sub-image patches for each test condition.

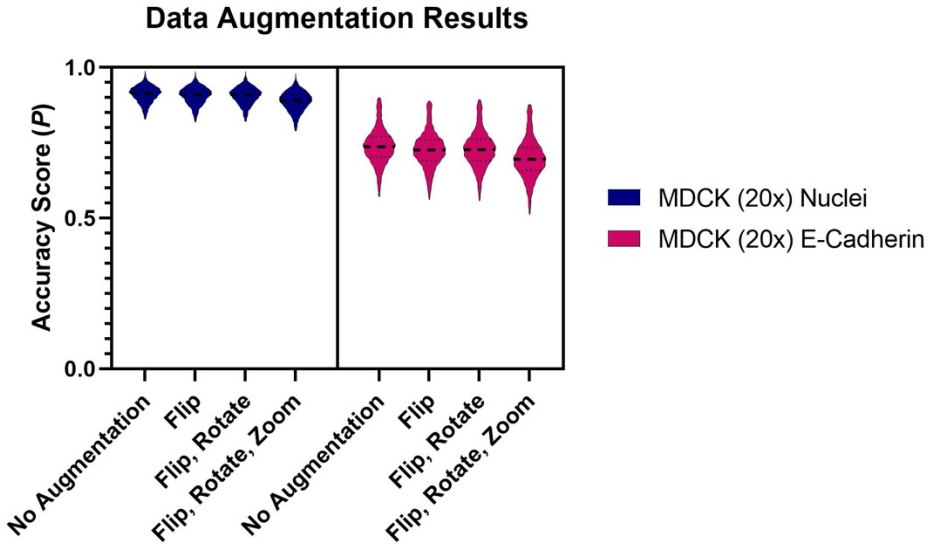


Figure S5. Representative accuracy results for datasets trained on the standard (1-stack) U-Net, with and without data augmentation applied (flip, rotation, zoom) (see *Methods*)**.** Training conditions were otherwise unchanged, and results are reported on the complete unmodified test set for each experimental condition. N = 10,000 sub-image patched for each test condition.

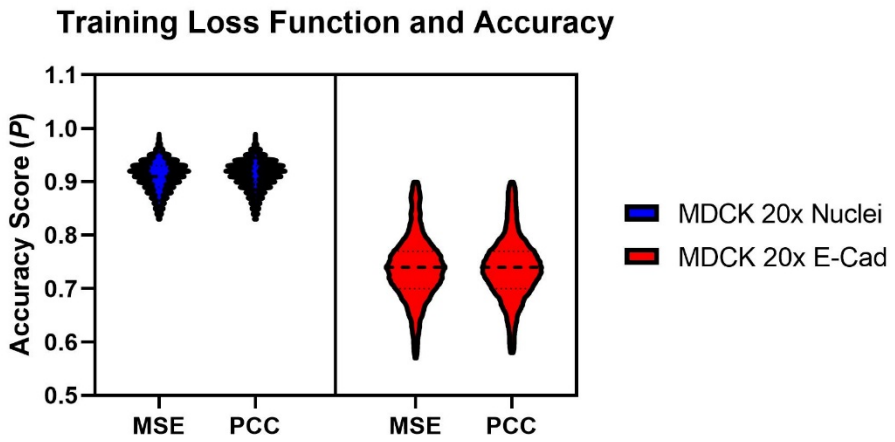


Figure S6. Representative accuracy results for networks trained using the Mean-Squared Error loss function (MSE) compared to the Pearson's Correlation Coefficient loss function (PCC). The neural network architecture and training conditions are the same, with the exception of the choice of loss function. Accuracy scores, as reported in terms of the modified *P* (see *Methods*) were comparable. *N* = *1*0,000 sub-image patches for each test condition.

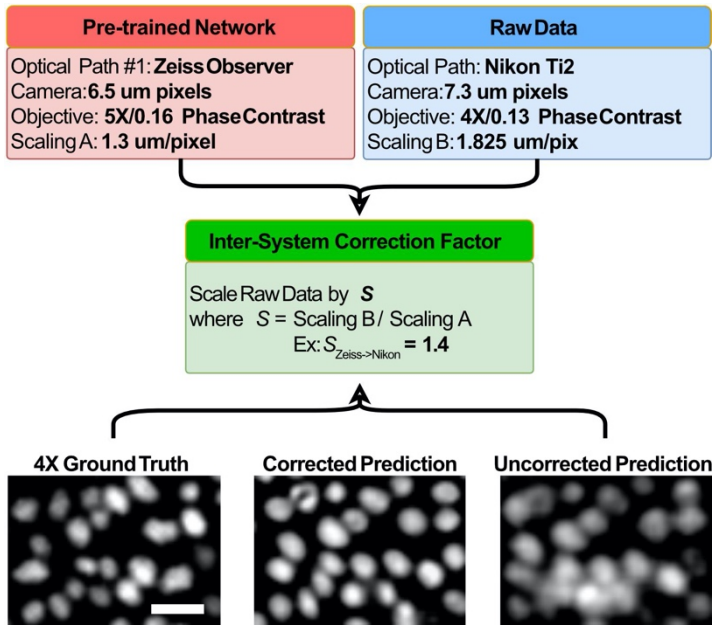


Figure S7. Cross-platform evaluation. A network was trained using data collected at 5x on a Zeiss microscope with a 6.5 µm/pixel camera, and used to process new images collected on a Nikon microscope at 4X with a 7.3 µm/pixel camera. We pre-processed data from the new system (Nikon) by scaling it to match the µm/pixel resolution of data from the original system (Zeiss) using the 'Inter-System Correction Factor' to rescale images in ImageJ. Representative ground truth nuclei from MDCK WT cells images on the Nikon system are shown next to the corresponding cross-platform predictions, resulting from both scaled and unscaled input images. Images were contrast adjusted for reproduction by normalizing the histograms and shifting the lower bound of the histogram up by 1/4 of the dynamic range. The scale bar represents 50 µm.

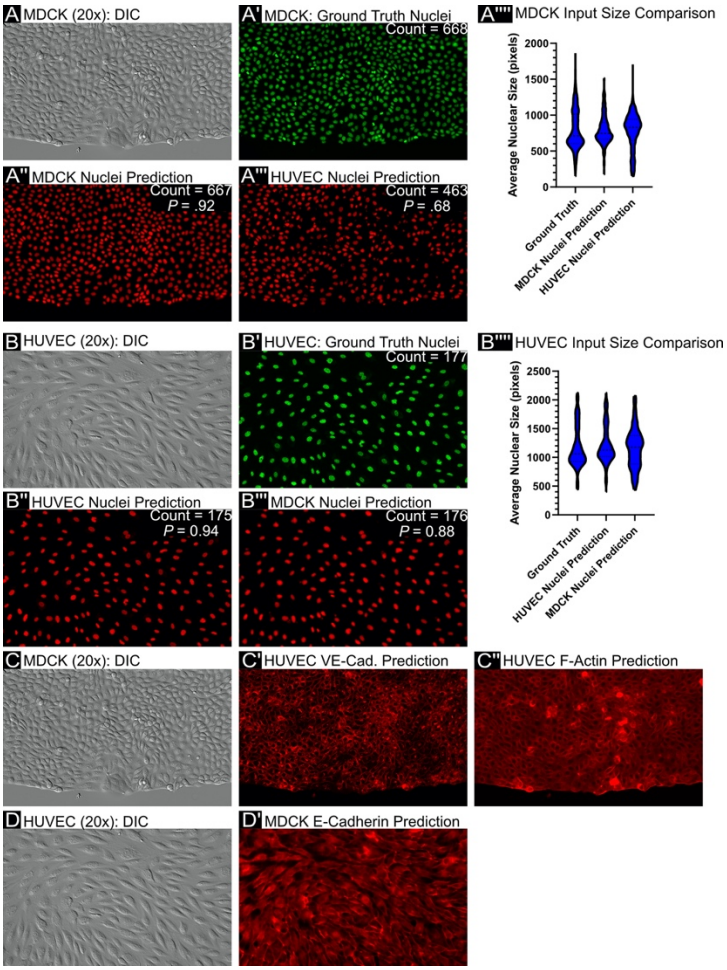


Figure S8. Cross-cell-line evaluation. (A, B) Representative images are shown for either the MDCK cell line or the HUVEC cell line, imaged at 20x magnification. (A', B') For each input image, the ground truth nuclei are shown, (A'', A''', B'', B''') along with predictions from U-Net models fully trained with either the MDCK or HUVEC nuclei data, (A'''', B'''') with corresponding violin plots shown to display average nuclear size, respectively. (C, D) Additionally, the process is repeated for biomarkers lacking a training set across cell types, such as (C', C'') VE-cadherin and F-actin predictions produced from MDCK input, or (D') E-cadherin prediction produced from HUVEC input.

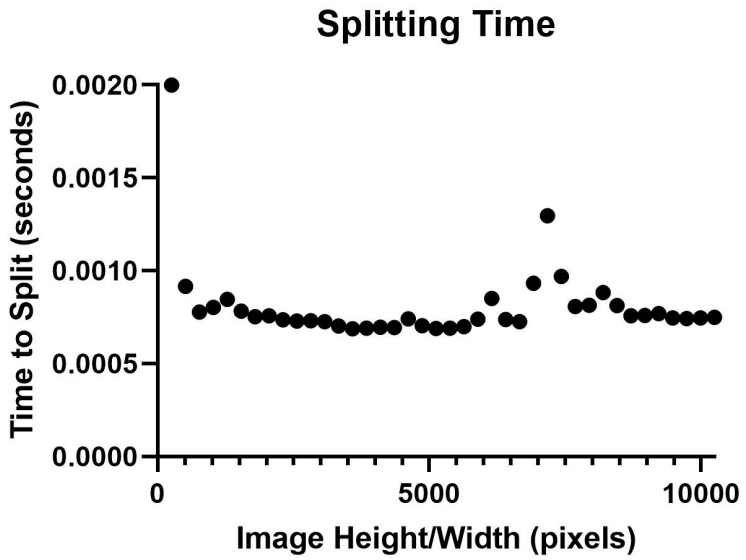


Figure S9. Representative image splitting times by image size. The time in seconds it takes to chop a raw image of equivalent height/width (in pixels) into 256x256 pixel^2 sub-images.

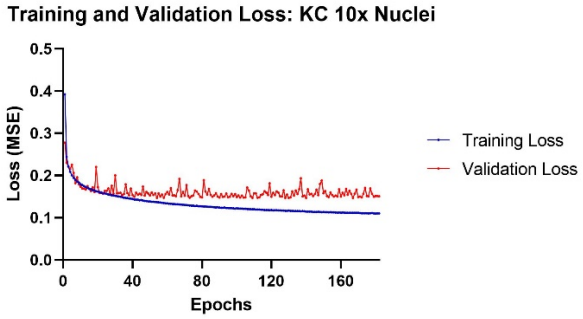


Figure S10. Representative loss functions from the standard U-Net training process. Early stopping was enabled, so that if the validation loss did not decrease within a set number of epochs, the training process terminated.

<u>Movie captions on the following page</u>

Supplementary Movie Captions:

Movie S1. Fluorescence reconstruction microscopy (FRM) on timelapse data.

A phase-contrast timelapse of MDCK cells, imaged at 5x magnification, is shown at left. The center panel displays nuclear predictions produced by the trained U-Net, given individual frames from the phase-contrast timelapse as input. The overlay of the phase-contrast movie and the nuclear predictions is shown at right. Each panel is 0.31 cm x 0.31 cm, and time between frames is 20 minutes. Video compressed for supplement, but raw data available on request.

Movie S2. FRM for high-content screening.

A DIC timelapse movie of MDCK cells, imaged at 20x magnification, is shown at left (top and bottom). The top row displays E-cadherin junctions, while the bottom row displays nuclei. Moving left from right, the second-from-left images are ground truth (actual) fluorescent images of the junctions/nuclei in green, followed by the FRM predictions in red, and finally the merge of the ground truth and predicted images. Predictions are produced by processing the DIC input on the left through a neural network trained on a dataset of matched DIC and fluorescence image pairs. Panel width is approximately 500 μm, and time between frames is 20 minutes.

Movies S3-5. FRM for fine structures.

A DIC timelapse movie of HUVEC cells, imaged at 20x magnification, is shown at left. Movie S3 shows VE-cadherin, Movie S4 shows Nuclei, and Movie S5 shows F-actin. All movies present DIC/Ground Truth/Prediction/Merge from left to right. Individual panel width is 890 μm, with 20 minutes per movie frame.

Movie S6. Mitotic division prediction.

A neural network is able to capture rare events, such as cell divisions, when trained on a sufficiently large and varied dataset. Left panel: a fluorescent timelapse of stained HUVEC nuclei, imaged at 20x magnification. Center panel: the U-Net predictions from DIC images of the same spatial region. Right panel: an overlay of the left and center panels for comparison. Time between frames is 20 minutes.

References.

730 1. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods*
731 **14**, 849–863 (2017).

732 2. Moen, E. *et al.* Deep learning for cellular image analysis. *Nature Methods* **16**, 1233–1246
733 (2019).

734 3. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational
735 biology. *Mol. Syst. Biol.* **12**, 878 (2016).

736 4. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image*
737 *Analysis* **42**, 60–88 (2017).

738 5. Smith, K. *et al.* Phenotypic Image Analysis Software Tools for Exploring and
739 Understanding Big Image Data from Cell-Based Assays. *Cell Systems* **6**, 636–653
740 (2018).

741 6. Belthangady, C. & Royer, L. A. Applications, promises, and pitfalls of deep learning for
742 fluorescence image reconstruction. *Nat. Methods* **16**, 1215–1225 (2019).

743 7. Brent, R. & Boucheron, L. Deep learning to predict microscope images. *Nat. Methods* **15**,
744 868–870 (2018).

745 8. Weigert, M. *et al.* Content-aware image restoration: pushing the limits of fluorescence
746 microscopy. *Nat. Methods* **15**, 1090–1097 (2018).

747 9. Christiansen, E. M. *et al.* In Silico Labeling: Predicting Fluorescent Labels in Unlabeled
748 Images. *Cell* **173**, 792-803.e19 (2018).

749 10. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free
750 prediction of three-dimensional fluorescence images from transmitted-light microscopy.
751 *Nat. Methods* **15**, 917–920 (2018).

752 11. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to
753 document recognition. *Proc. IEEE* **86**, 2278–2323 (1998).

754 12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep*
755 *Convolutional Neural Networks*. (2012).

756 13. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical
757 image segmentation. in *Lecture Notes in Computer Science (including subseries Lecture*
758 *Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351**, 234–241
759 (Springer Verlag, 2015).

760 14. Van Valen, D. A. *et al.* Deep Learning Automates the Quantitative Analysis of Individual
761 Cells in Live-Cell Imaging Experiments. *PLoS Comput. Biol.* **12**, 1–24 (2016).

762 15. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat.*
763 *Methods* **16**, 67–70 (2019).

764 16. Gupta, A. *et al.* Deep Learning in Image Cytometry: A Review. *Cytometry Part A* **95**, 366–
765 380 (2019).

766 17. Caicedo, J. C. *et al.* Evaluation of Deep Learning Strategies for Nucleus Segmentation in
767 Fluorescence Images. *Cytom. Part A* **95**, 952–965 (2019).

768 18. Newby, J. M., Schaefer, A. M., Lee, P. T., Forest, M. G. & Lai, S. K. Convolutional neural
769 networks automate detection for tracking of submicron-scale particles in 2D and 3D.
770 *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9026–9031 (2018).

771 19. Kim, D., Min, Y., Oh, J. M. & Cho, Y. K. AI-powered transmitted light microscopy for
772 functional analysis of live cells. *Sci. Rep.* **9**, (2019).

773 20. Goodfellow, I. J. *et al.* Generative adversarial nets. in *Advances in Neural Information*
774 *Processing Systems* **3**, 2672–2680 (2014).

775 21. Rivenson, Y. *et al.* PhaseStain: the digital staining of label-free quantitative phase
776 microscopy images using deep learning. *Light Sci. Appl.* **8**, (2019).

777 22. Rivenson, Y. *et al.* Virtual histological staining of unlabelled tissue-autofluorescence
778 images via deep learning. *Nat. Biomed. Eng.* **3**, 466–477 (2019).

779 23. De Haan, K., Rivenson, Y., Wu, Y. & Ozcan, A. Deep-Learning-Based Image
780 Reconstruction and Enhancement in Optical Microscopy. *Proc. IEEE* **108**, 30–50 (2020).

781 24. Erba, E., Ubezio, P., Broggini, M., Ponti, M. & D'Incalci, M. DNA damage, cytotoxic effect
782 and cell-cycle perturbation of Hoechst 33342 on L1210 cells in vitro. *Cytometry* **9**, 1–6
783 (1988).

784 25. Durand, R. E. & Olive, P. L. Cytotoxicity, mutagenicity and DNA damage by Hoechst
785 33342. *J. Histochem. Cytochem.* **30**, 111–116 (1982).

786 26. Wlodkowic, D. & Darzynkiewicz, Z. Please do not disturb: Destruction of chromatin
787 structure by supravital nucleic acid probes revealed by a novel assay of DNA-histone
788 interaction. *Cytom. Part A* **73**, 877–879 (2008).

789 27. Bray, M. A. *et al.* Cell Painting, a high-content image-based assay for morphological
790 profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).

791 28. Garcia, M. A., Nelson, W. J. & Chavez, N. Cell–cell junctions organize structural and
792 signaling networks. *Cold Spring Harb. Perspect. Biol.* **10**, (2018).

793 29. Friedl, P. & Mayor, R. Tuning collective cell migration by cell-cell junction regulation. *Cold*
794 *Spring Harb. Perspect. Biol.* **9**, 1–18 (2017).

795 30. Hayer, A. *et al.* Engulfed cadherin fingers are polarized junctional structures between
796 collectively migrating endothelial cells. *Nat. Cell Biol.* **18**, 1311–1323 (2016).

797 31. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition*.
798 (2016).

799 32. Newell, A., Yang, K. & Deng, J. Stacked hourglass networks for human pose estimation.

in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9912 LNCS**, 483–499 (2016).

33. Szegedy, C. *et al. Going deeper with convolutions*. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07-12-June**, (2015).

34. Zaritsky, A. *et al.* Interpretable deep learning of label-free live cell images uncovers functional hallmarks of highly-metastatic melanoma. *bioRxiv* 2020.05.15.096628 (2020). doi:10.1101/2020.05.15.096628

35. Xing, F. & Yang, L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Reviews in Biomedical Engineering* **9**, 234–263 (2016).

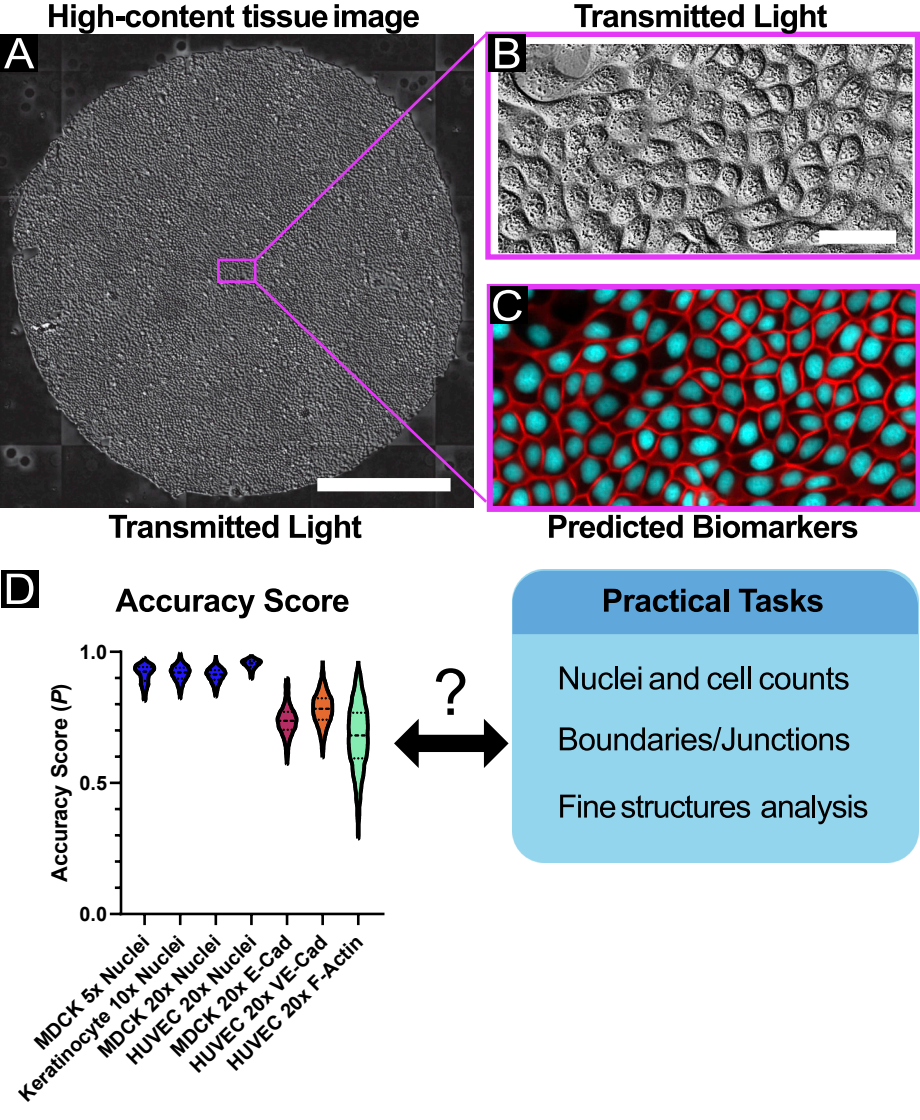
36. Nowak, J. A. & Fuchs, E. Isolation and culture of epithelial stem cells. *Methods Mol. Biol.* **482**, 215–232 (2009).

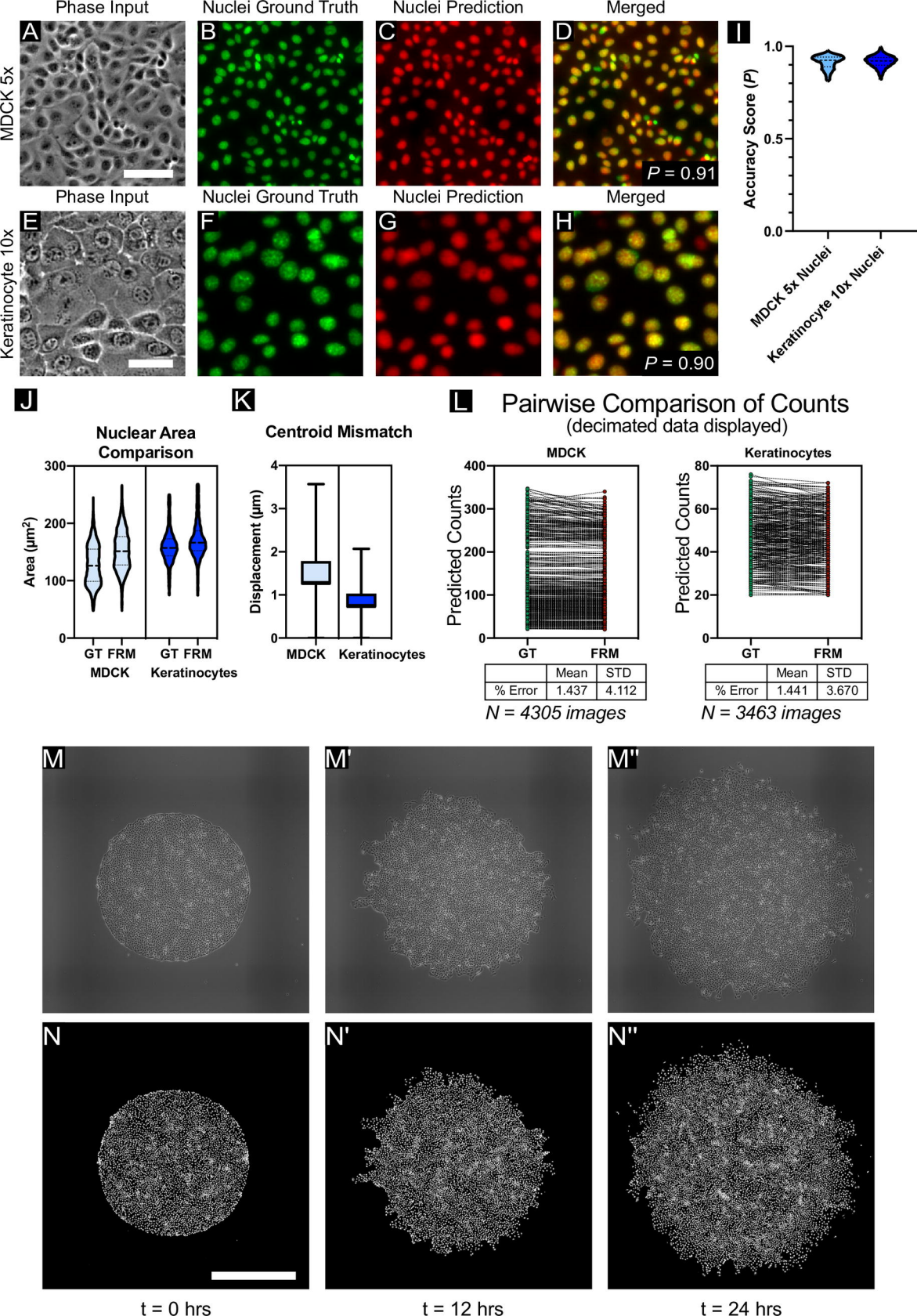
37. Cohen, D. J., Gloerich, M. & Nelson, W. J. Epithelial self-healing is recapitulated by a 3D biomimetic E-cadherin junction. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14698–14703 (2016).

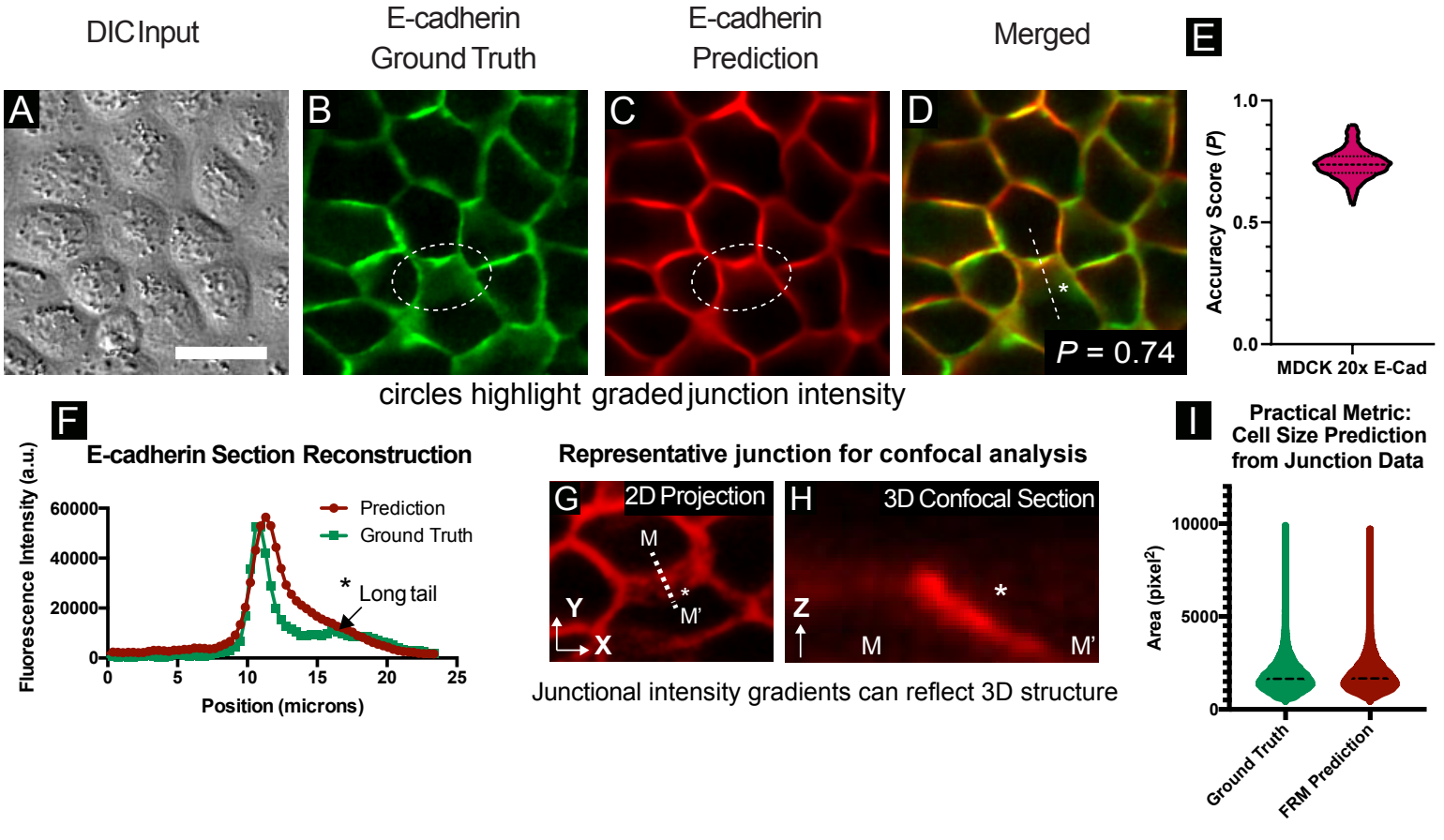
38. Chollet, F. *Keras*. (2015).

39. Weinman, J. J., Lidaka, A. & Aggarwal, S. TensorFlow: Large-scale machine learning. *GPU Comput. Gems Emerald Ed.* 277–291 (2011). doi:1603.04467

40. Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. (2012).


High-content tissue image
A
Transmitted Light
B
C
Transmitted Light
Predicted Biomarkers
D
Accuracy Score
Accuracy Score (P)
1.0
0.5
0.0
MDCK 5x Nuclei
Keratinocyte 10x Nuclei
MDCK 20x Nuclei
HUVEC 20x Nuclei
MDCK 20x E-Cad
HUVEC 20x VE-Cad
HUVEC 20x F-Actin
?
Practical Tasks
Nuclei and cell counts
Boundaries/Junctions
Fine structures analysis


Phase Input
Nuclei Ground Truth
Nuclei Prediction
Merged
MDCK 5x
Keratinocyte 10x
A
B
C
D
E
F
G
H
P = 0.91
P = 0.90
I
Accuracy Score (P)
1.0
0.5
0.0
MDCK 5x Nuclei
Keratinocyte 10x Nuclei
J
Nuclear Area Comparison
Area (μm²)
300
200
100
0
GT FRM
MDCK
GT FRM
Keratinocytes
K
Centroid Mismatch
Displacement (μm)
4
3
2
1
0
MDCK
Keratinocytes
L
Pairwise Comparison of Counts
(decimated data displayed)
MDCK
Predicted Counts
400
300
200
100
0
GT
FRM
Mean STD
% Error 1.437 4.112
N = 4305 images
Keratinocytes
Predicted Counts
80
60
40
20
0
GT
FRM
Mean STD
% Error 1.441 3.670
N = 3463 images
M
M'
M''
N
N'
N''
t = 0 hrs
t = 12 hrs
t = 24 hrs


DIC Input
E-cadherin Ground Truth
E-cadherin Prediction
Merged
A
B
C
D
E
P = 0.74
Accuracy Score (P)
1.0
0.5
0.0
MDCK 20x E-Cad
circles highlight graded junction intensity
F
E-cadherin Section Reconstruction
Fluorescence Intensity (a.u.)
60000
40000
20000
0
Prediction
Ground Truth
* Long tail
0
5
10
15
20
25
Position (microns)
Representative junction for confocal analysis
G
2D Projection
M
M'
Y
X
H
3D Confocal Section
Z
M
M'
Junctional intensity gradients can reflect 3D structure
I
Practical Metric: Cell Size Prediction from Junction Data
Area (pixel²)
10000
5000
0
Ground Truth
FRM Prediction

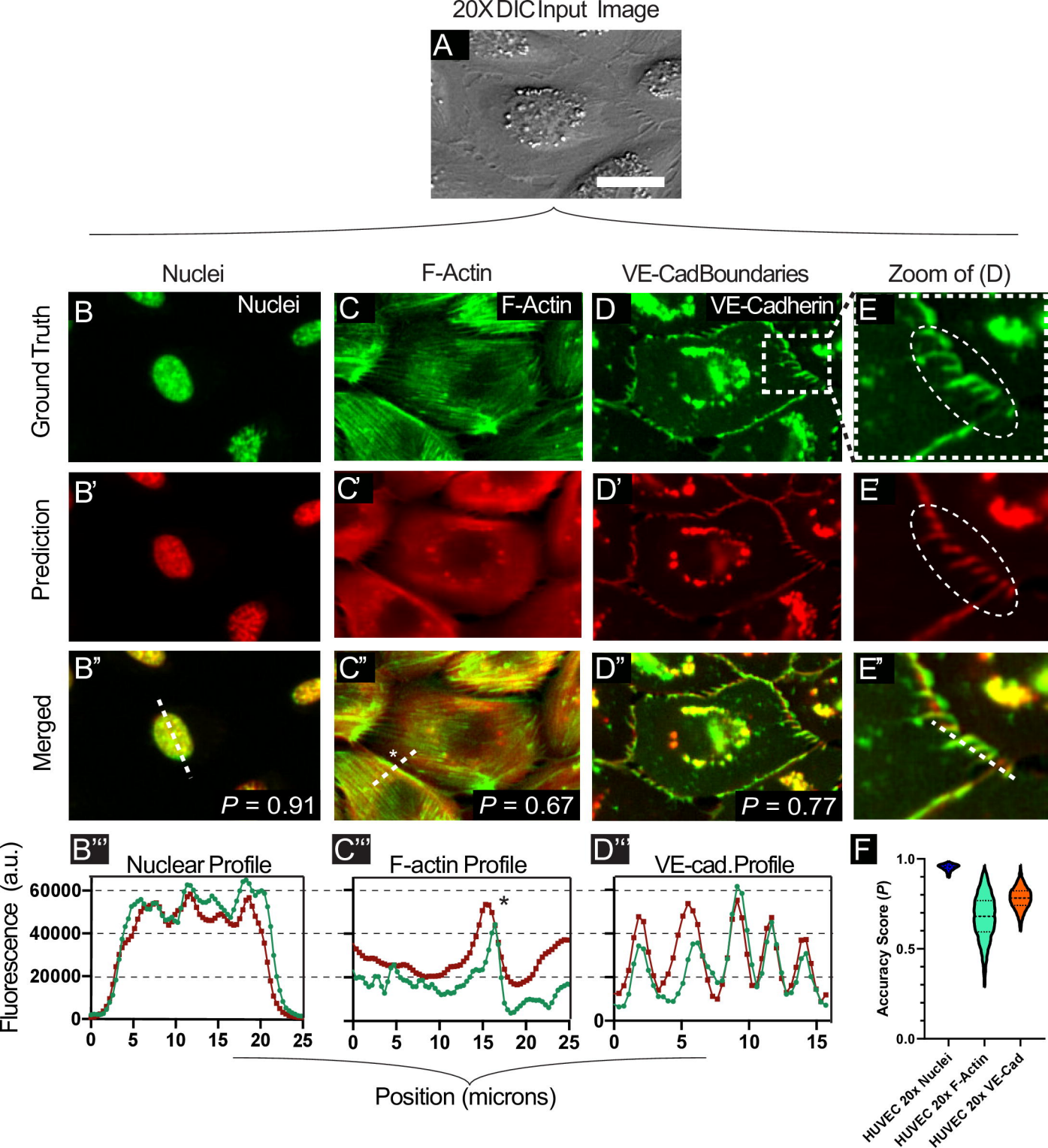

20X DIC Input Image
A
Nuclei
F-Actin
VE-Cad Boundaries
Zoom of (D)
Ground Truth
Prediction
Merged
B
Nuclei
C
F-Actin
D
VE-Cadherin
E
B'
C'
D'
E'
B"
C"
D"
E"
P = 0.91
P = 0.67
P = 0.77
B'''
Nuclear Profile
C'''
F-actin Profile
D'''
VE-cad. Profile
Fluorescence (a.u.)
60000
40000
20000
0
0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15
Position (microns)
F
Accuracy Score (P)
1.0
0.5
0.0
HUVEC 20x Nuclei
HUVEC 20x F-Actin
HUVEC 20x VE-Cad

Fig5_PDF.pdf
1833 x 2625

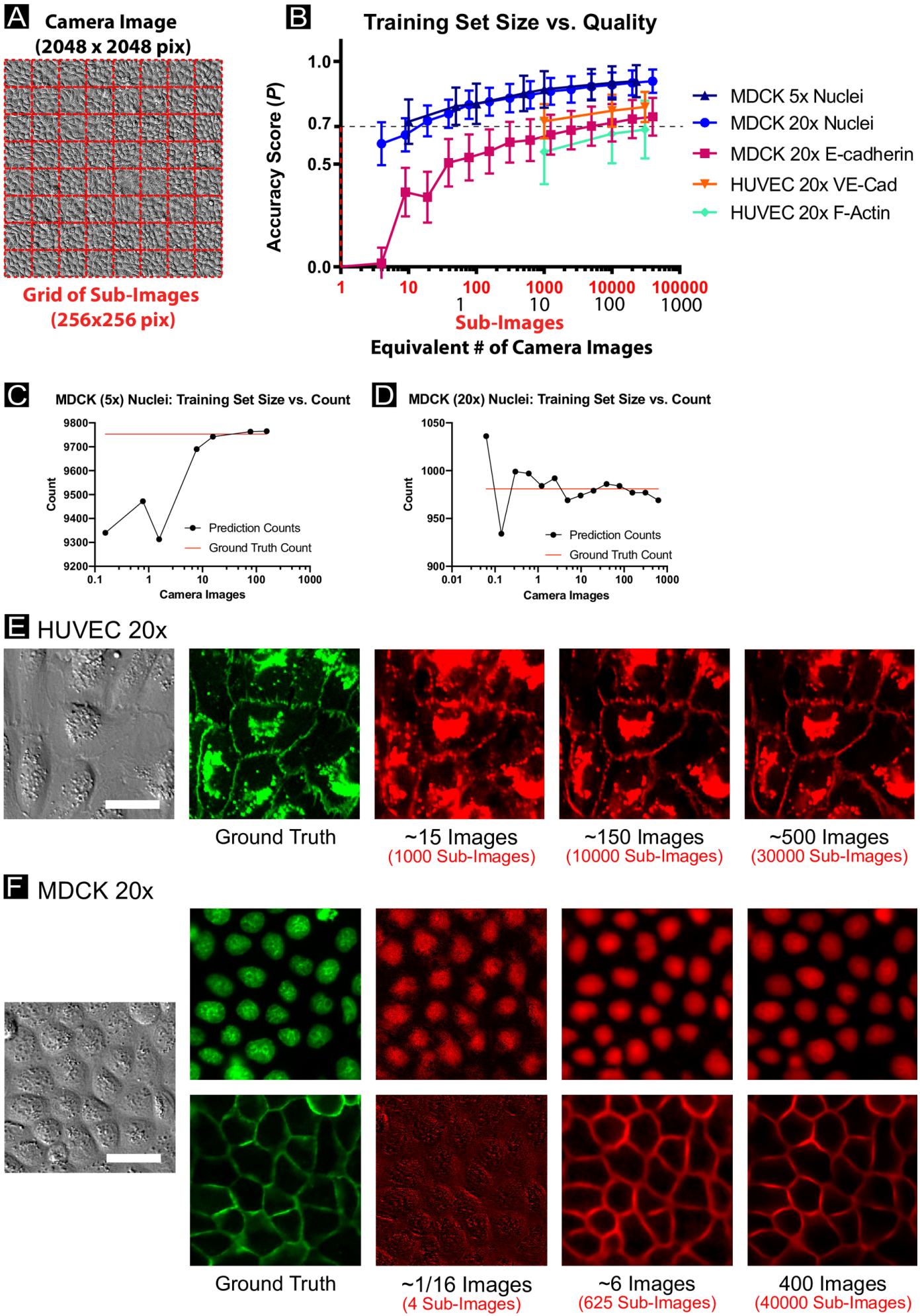
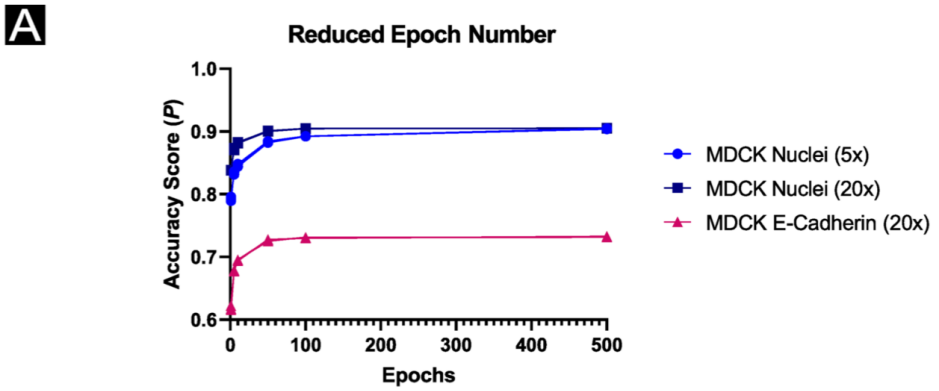
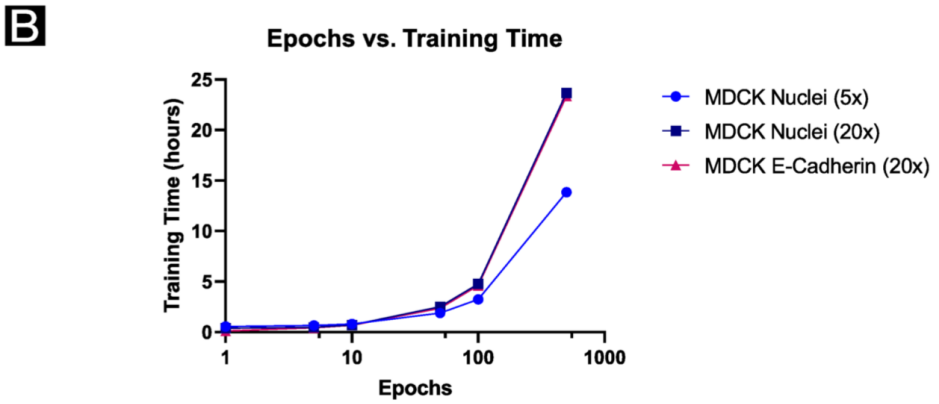
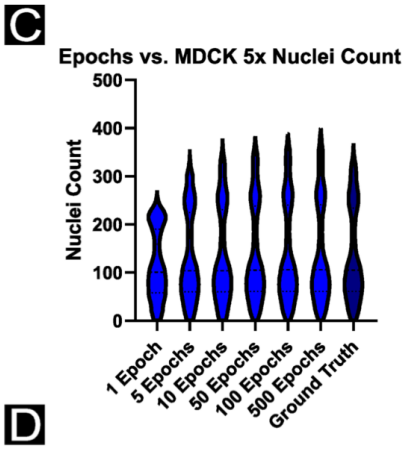
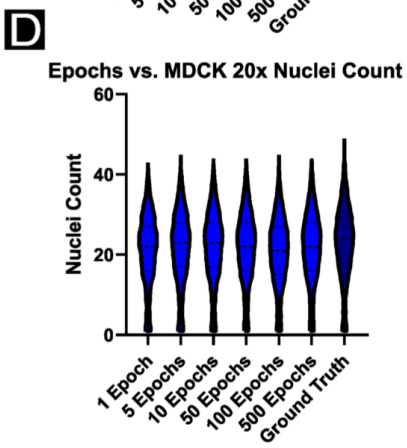

A
Camera Image
(2048 x 2048 pix)
Grid of Sub-Images
(256x256 pix)
B
Training Set Size vs. Quality
Accuracy Score (P)
1.0
0.7
0.5
0.0
1
10
100
1000
10000
100000
1
10
100
1000
Sub-Images
Equivalent # of Camera Images
MDCK 5x Nuclei
MDCK 20x Nuclei
MDCK 20x E-cadherin
HUVEC 20x VE-Cad
HUVEC 20x F-Actin
C
MDCK (5x) Nuclei: Training Set Size vs. Count
Count
9800
9700
9600
9500
9400
9300
9200
0.1
1
10
100
1000
Camera Images
Prediction Counts
Ground Truth Count
D
MDCK (20x) Nuclei: Training Set Size vs. Count
Count
1050
1000
950
900
0.01
0.1
1
10
100
1000
Camera Images
Prediction Counts
Ground Truth Count
E HUVEC 20x
Ground Truth
~15 Images
(1000 Sub-Images)
~150 Images
(10000 Sub-Images)
~500 Images
(30000 Sub-Images)
F MDCK 20x
Ground Truth
~1/16 Images
(4 Sub-Images)
~6 Images
(625 Sub-Images)
400 Images
(40000 Sub-Images)

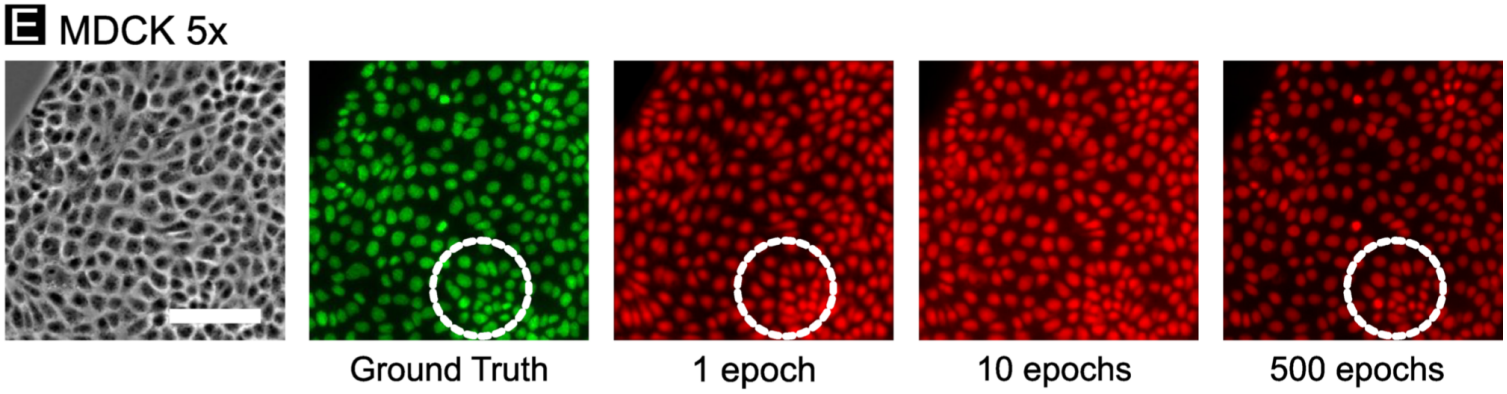
Fig6_pdf.pdf
2008 x 2626

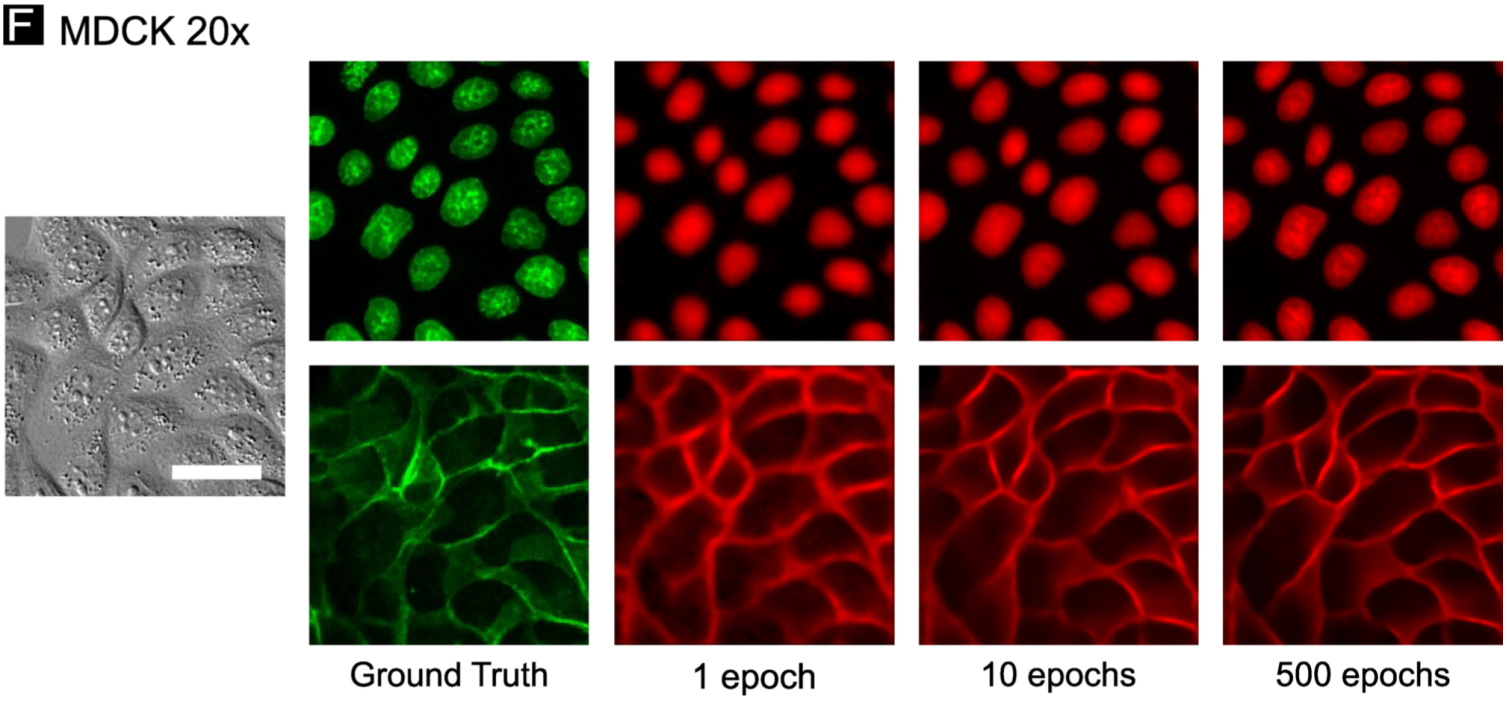

A
Reduced Epoch Number
Accuracy Score (P)
1.0
0.9
0.8
0.7
0.6
0
100
200
300
400
500
Epochs
MDCK Nuclei (5x)
MDCK Nuclei (20x)
MDCK E-Cadherin (20x)


B
Epochs vs. Training Time
Training Time (hours)
25
20
15
10
5
0
1
10
100
1000
Epochs
MDCK Nuclei (5x)
MDCK Nuclei (20x)
MDCK E-Cadherin (20x)


C
Epochs vs. MDCK 5x Nuclei Count
Nuclei Count
500
400
300
200
100
0
1 Epoch
5 Epochs
10 Epochs
50 Epochs
100 Epochs
500 Epochs
Ground Truth


D
Epochs vs. MDCK 20x Nuclei Count
Nuclei Count
60
40
20
0
1 Epoch
5 Epochs
10 Epochs
50 Epochs
100 Epochs
500 Epochs
Ground Truth


E MDCK 5x
Ground Truth
1 epoch
10 epochs
500 epochs


F MDCK 20x
Ground Truth
1 epoch
10 epochs
500 epochs