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 which 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, current FRM benchmarks are abstractions that are difficult to relate to how valuable or trustworthy an FRM prediction 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 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 the code, sample data, and user manual to enable more widespread adoption of FRM.

Introduction

1   Deep learning holds enormous promise for biological microscopy data, and offers

2   especially exciting opportunities for fluorescent feature reconstruction[1–5]. Here, fluorescence

3   reconstruction microscopy (FRM) takes in a transmitted light image of a biological sample and

4   outputs a series of reconstructed fluorescence images that predict what the sample would look

5   like had it been labeled with a given series of dyes or fluorescently tagged proteins (Fig. 1A-C)

6   [2,6–10]. FRM works by first training a convolutional neural network (e.g. U-Net) to relate a large set

7   of transmitted light data to corresponding real fluorescence images (the ground truth) for given

8   markers[11–13]. The network learns by comparing its fluorescence predictions to the ground truth

9   fluorescence data and iterating until it reaches a cut off. Once trained, FRM can be performed

10   on transmitted light data without requiring any additional fluorescence imaging. This is a

11   powerful capability and allows FRM to: reduce phototoxicity; free up fluorescence channels for

12   more complex markers; and enable re-processing of legacy transmitted light data to extract new

13   information. In all cases, FRM data are directly compatible with any standard fluorescence

14   analysis software or workflows (e.g. ImageJ plug-ins). Such capabilities are extremely useful,

15   and FRM may eventually become a standard tool to augment quantitative biological imaging

16   once practical concerns are addressed.

17

18   However, a number of challenges limit FRM accessibility to the larger biological community. Key

19   among these is the difficulty in relating the abstract accuracy metrics used to score FRM to the

20   practical value of FRM data for actual, quotidian biological analyses such as cell counting or

21   morphological characterization. To better appreciate this, consider first that the quality of FRM is

22   typically assessed using a single numerical metric (P) such as the Mean-Squared-Error or

23   Pearson's Correlation Coefficient that typically range from (0,1) or (-1,1), and second that it is

24   practically impossible to actually reach perfection ($P = 1$). P can be increased closer to 1 either

25   by training with more images, or by using higher resolution magnification (e.g. 40X-100X) to

26   capture finer details. However, increasing P also carries an intrinsic cost in increased wet-lab

27   and computing time. That improving P is expensive and that P cannot be perfect beg the

28   questions of how good is good enough, and good enough for what (Fig. 1D)? For instance, $P =$

29   0.7 lacks any practical context, and may be quite good enough for a given use case without

30   requiring more work to raise the 'accuracy'. This is why context is extremely important for FRM

31   and why the work we present here focuses on evaluating practical uses of FRM with respect to

32   given P values.

33

34   Our goal here is to provide a standardized implementation of FRM and demonstrate its practical

35   performance and limitations for every-day tasks such as nuclear localization and tracking,

36   characterizing cell morphology, cell-cell junction detection and analysis, and re-analyzing legacy

37   data and data collected on different systems (Fig. 1). To further emphasize the use of FRM for

38   routine tasks, we will exclusively focus on those lower magnifications (4X-20X) commonly used

39   in high content imaging and cellular screening in contrast to the focus on higher magnifications

40   in prior studies [9,10]. We hope that the included software we developed and the analyses and

41   comparison data we present will help make FRM more approachable to the broader biological

42   community. To further facilitate this, we have made the entirety of our code and all collected

43   data public as well as providing a full tutorial guide (see Methods and Supplementary Material).
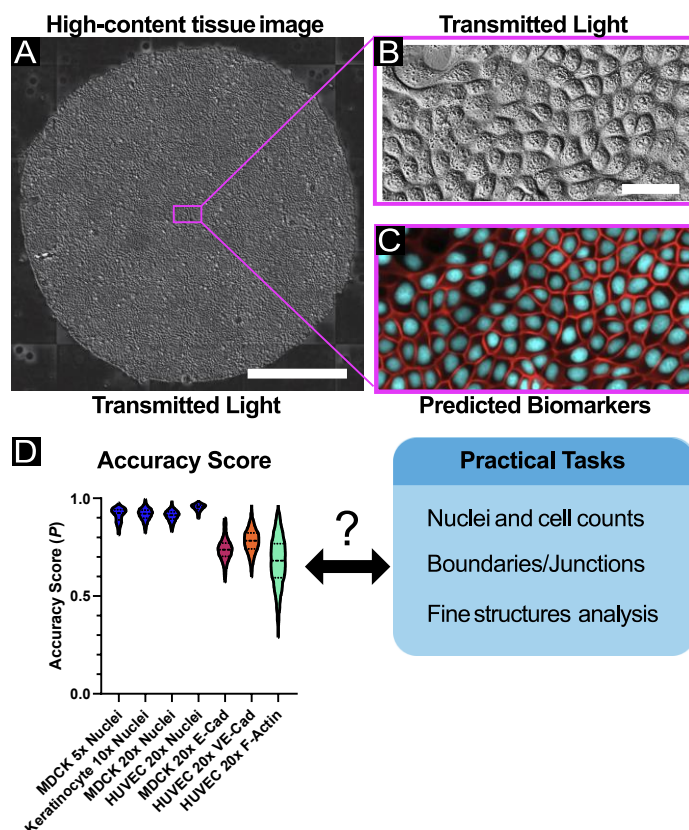
44

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 for all datasets; see Table S1 for summary statistics. However, such accuracy metrics are not necessarily indicative of useful feature reconstructions for many practical applications.

45

46

47

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

48

49   While early FRM methods used computationally complex and expensive networks that relied on

50   Z-stacks of images to capture 2D reconstruction[9], more recently this has been adapted to

51   reconstruct 3D image stacks using a modified U-Net architecture[10]. The U-Net itself is

52   commonly used in machine learning approaches because it is a lightweight convolutional neural

53   network (CNN) which readily captures information at multiple spatial scales within an image,

54   thereby preserving reconstruction accuracy while reducing the required number of training

55   samples and training time. U-Nets, and related deep learning approaches, have found broad

56   application to live-cell imaging tasks such as cell phenotype classification, feature

57   segmentation[10,14–19], and histological stain analysis[20–23].

58

59 Our implementation here provides an archetypal U-Net and framework intended for the cell

60 biology community. Briefly, our workflow is as follows. First, we collected multi-channel training

61 images of cultured cells where each image comprised a transmitted light channel and

62 associated fluorescence channels (labeled using genetically encoded reporters or chemical

63 dyes; see Methods). These images were then broken into 256x256 pix^2 sub-images in ImageJ

64 and then input into the network. Such image chopping is necessary for the average user to

65 account for the average RAM and graphics cards available on standard workstations. These

66 data are then passed through the U-Net network to generate trained weights—the pattern

67 recognition side of the network. Here, the transmitted light images serve as input to the network,

68 which is then optimized to minimize the difference between intensity values of the output

69 predicted images and the intensity values from the ground truth corresponding fluorescence

70 images (e.g. Fig. 1). This process can be extended to full time-lapse video fluorescence

71 reconstruction, making it well suited for high-content live imaging (see Movies S1-4). We have

72 provided all of our code, all raw and processed data, and an extensive user manual

73 (DataSpace, GitHub) to encourage exploration of FRM.

74

75 As our conventional performance metric, we selected the Pearson's Correlation Coefficient

76 (PCC), which is commonly used in cell biology when comparing the co-localization of two or

77 more proteins, and also used in computer vision to assess spatial-intensity when determining

78 image similarity. However, we observed that naively applying the PCC across our whole dataset

79 skewed the results due to the large number of images containing primarily background

80 (common with high content imaging of oddly shaped or low density samples). This resulted in

81 poor PCC scores as the network tried to reconstruct the pseudo-random background noise. To

82 address this, we report a corrected accuracy score (P) representing the PCC of a large subset

83 of images in a given dataset containing positive examples of the feature (nuclei, junctions, etc.)

84   based on an intensity threshold (Fig. S1, see Methods). This approach will improve network

85   performance for datasets containing large amounts of background signal.

86   To broadly explore the utility of FRM for high-content imaging applications, we captured

87   transmitted light images using 4X, 10X, and 20X air objectives using either Phase Contrast or

88   Differential Interference Contrast (DIC), and collected data across 3 different cell types—renal

89   epithelial cells (MDCK), primary mouse skin keratinocytes (KC), and human umbilical vein

90   endothelial cells (HUVEC). Variable training set sizes were tested to also explore the effect of

91   data set size on 'accuracy'—a key practical aspect of designing an FRM study. The biomarkers

92   we trained against comprised a nuclear dye (Hoechst 33342), an F-actin dye (SiR-Actin) and

93   genetically encoded fluorescence reporters for E-cadherin and VE-cadherin. Traditional

94   Accuracy Scores for each of these are summarized in Fig. 1D and Table S1, and we will next

95   present case studies from each of these data sets before concluding with a discussion of how

96   'accuracy' relates to visual quality to help researchers design experiments for FRM. While we

97   necessarily show representative data here, we provide the statistical distribution for all accuracy

98   scores, and encourage exploration of our provided datasets.

99

100

101   **Results.**

102   ***Demonstration of FRM for low-magnification nuclear fluorescence reconstruction and***

103   ***analysis***

104

105   One of the most common computational image processing needs for screening and low-

106   magnification image is nuclei detection or segmentation, which enables cell counting, time-lapse

107   tracking, and statistical analyses of ensemble distribution and geometry. While a variety of

108   traditional image processing approaches exist to extract nuclei from phase or DIC images, such

109   techniques require extensive fine tuning, ultimately only work for certain cell types, and often fail

110   to work at all with DIC images. The most reliable and standardized technique by far is using a

111   vital dye (e.g. Hoechst 33342 or DRAQ) to stain the nuclei. However, Hoechst requires cytotoxic

112   UV illumination while DRAQ (far-red fluorescence characteristics) has been linked to cell cycle

113   alterations due to its chemistry [24–26]. Both dyes also exhibit loss of signal over extended time-

114   lapse imaging. Alternately, genetic reporters such as H2B nuclear labels can be engineered into

115   cells (e.g. transfection, viral addition, etc.), but this adds more overhead, incurs phototoxicity,

116   and still requires a dedicated fluorescence channel for a relatively simple structure (the nucleus)

117   in lieu of a more complex or useful label. Hence, there is a clear practical benefit to fluorescence

118   reconstruction of cell nuclei, especially in time-lapse imaging where freeing up a channel and

119   reducing phototoxicity are each quite valuable. Further, fluorescent reconstruction of nuclei

120   supports any software or analysis pipeline that might normally be employed with fluorescent

121   nuclei data, meaning that workflows need not be altered to leverage FRM data here.

122

123   To validate low-magnification, high-accuracy nuclear FRM, we collected data in both MDCK

124   renal epithelia cells (5X phase contrast, Figs. 2A-D) and primary skin keratinocytes (10X phase

125   contrast, Figs. 2E-H) while using Hoechst to label nuclei and generate our ground truth training

126   data. Representative images are presented as a sequence of phase contrast, nuclear ground

127   truth (green), network predictions (red), and a merged overlay (yellow for a perfect merge).

128   The Accuracy Score (P) is included for context, while the statistical distributions of P for each

129   cell type are presented in Fig. 2I demonstrating the actual network performance. The

130   performance with the keratinocyte data is particularly striking given how irregular and poorly

131   resolved the cells appear in phase contrast (confounding traditional segmentation).

132

133   The network performs visually well in both cases, with $P \sim 0.9$, but to represent what that means

134   in practice, we quantified disparities in the predictions with respect to nuclear size for geometric

135   accuracy (Fig. 2J) and centroid error to reflect positional accuracy (Fig. 2K). In both nuclear

136   area cases, the U-Net slightly overpredicts area, likely due to slight noise in the predictions

137   blurring the predicted nuclei and effectively increasing area. However, the distributions from the

138   violin plots are quite similar in structure, and the predictions are well within the usable range for

139   practical cell counting and segmentation. With respect to nuclear centroid localization, mean

140   errors span 2 microns (5X MDCK) to 1 micron (10X KCs). The improvement from 5X to 10X can

141   likely be attributed to the resolution increase in the magnification, but in both cases the errors

142   are quite small and more than sufficient for standard cell counting, nuclei tracking, and neighbor

143   distribution analyses. Whether a higher P would be beneficial would depend on the specific

144   analysis in question—here, the accuracy is more than sufficient.

145

146   As a final demonstration of the utility of low-magnification reconstruction and nuclear tracking,

147   we input legacy data from a 24 hr time-lapse experiment of the growth dynamics of large

148   epithelia (2.5 mm^2 , 5X) and the network output a reconstructed movie of nuclear dynamics

149   (Figs. 2L-M, and Movie S1) compatible with standard nuclear tracking algorithms (e.g.

150   Trackmate in FIJI). Images were captured every 10 minutes, and previous efforts to perform this

151   experiment using fluorescent imaging of Hoechst resulted in large-scale cell death, hence FRM

152   proved highly effective both as an alternative nuclear labeling approach for large-scale, long-

153   term imaging, and as a means to reprocess pre-existing, legacy datasets.
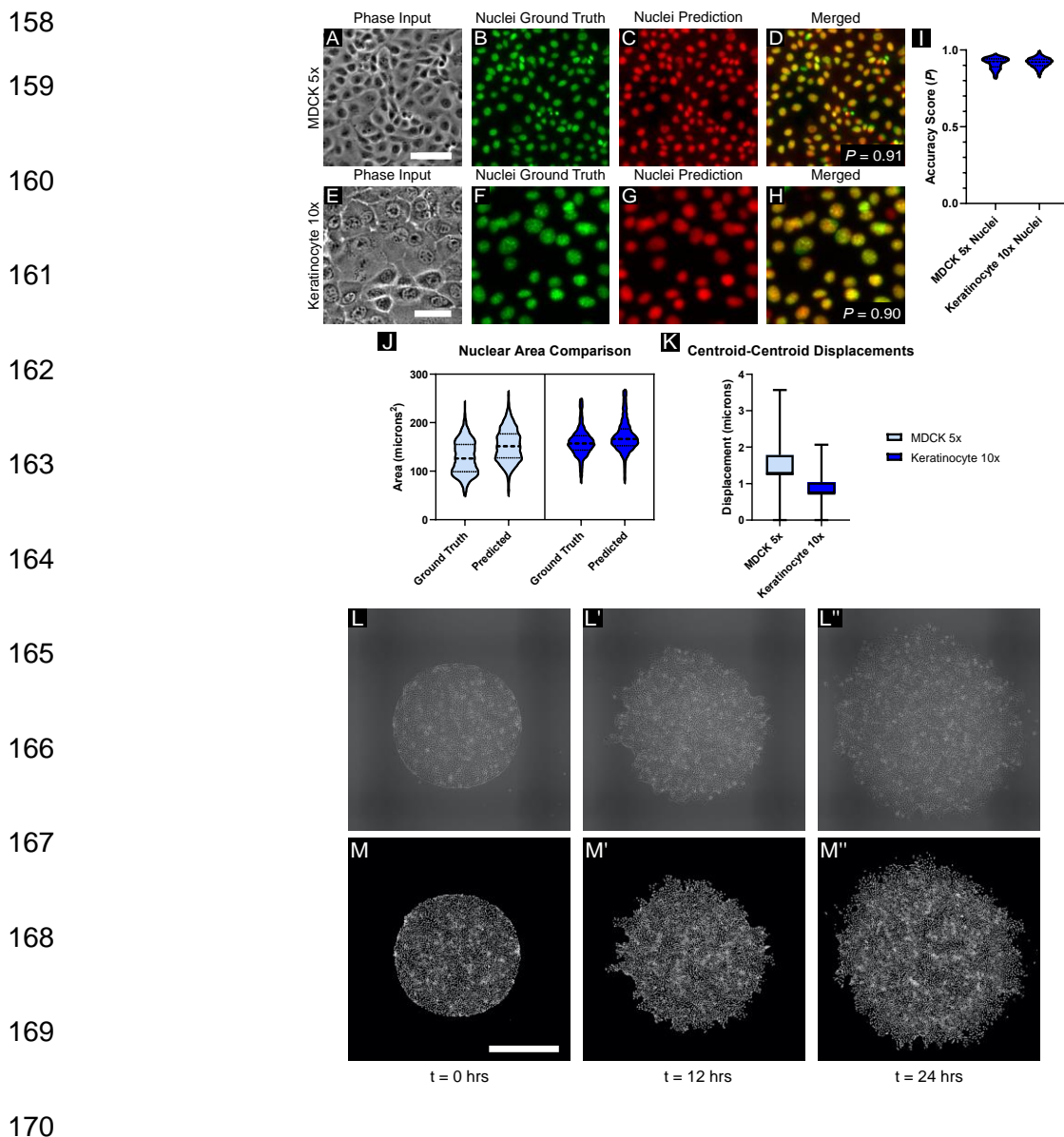
154

155

156

157

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. The accuracy score is given at right. The 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. The *n* is approximately 3500 for both plots. See *Methods*. (L-M) Sequence of phase images (L-L") from a time-lapse at 0, 12, and 24 hours of growth, with corresponding nuclear predictions (M-M") 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. The scale bar is 1 mm.

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

172   Cell-cell junctions and cellular boundaries in cellular ensembles have implications spanning the

173   epithelial-mesenchyme-transition (EMT), tissue mechanics, and tissue maturation[27–29] and are

174   of broad interest from cellular biophysics to high content screening. However, there are no vital

175   dyes for junctional proteins (e.g. E-cadherin), necessitating either antibodies or genetic

176   reporters. In the absence of a specific marker, cell boundaries are relatively difficult to

177   accurately segment, especially from DIC images (Fig. 3A), and proxy techniques such as

178   Voronoi tessellation from nuclei data often fail to capture cell shape and organic features such

179   as curved boundaries. Instead, junctions and boundary data most commonly come from

180   biomarkers such as E-cadherin, so we trained our U-Net using MDCK cells stably expressing E-

181   cadherin:RFP (Ecad:dsRed) and imaging with a 20X/0.75NA objective—a well-balanced

182   objective favored for high-content imaging and immersion-free time-lapse imaging.

183   The U-Net was able to reconstruct E-cadherin junctions with high visual accuracy, as shown in

184   the sequence from Figs. 3A-D. While $P = 0.74$, the reconstruction is quite spatially accurate,

185   which is unexpected given how difficult it is for humans to detect cell-cell junctions by eye in a

186   DIC image. To better highlight the accuracy and utility of junctional FRM, we explored how the

187   network reconstructed a subtle 3D feature of epithelial junctions where a slanted junction is

188   formed between two cells by one cell pushing slightly under another (the region enclosed in the

189   dashed oval in Figs. 3B-D). Such slanted junctions may indicate a degree of fluidity or direction

190   of migration and are also impossible to discern by eye. We quantified the accuracy of the FRM

191   image by taking a line section perpendicular to this slanted junction (Figs. 3D,E) and comparing

192   the profiles of the ground truth and the FRM image. In this line section, intensity values up to

193   and including the peak value are similar, and intensity values exhibit a graded decay within the

194   slanted junction, indicating that the FRM network is able to capture subtle 3D information from

195   the 2D input image. To emphasize the 3D nature of this feature, a representative Z-section from

196   an E-cadherin junction imaged by scanning confocal is shown in Figs. 3F,G. FRM can again be

197   used for high-fidelity reconstruction during a timelapse, allowing both nuclei and junctions to be

198   predicted throughout long acquisitions (see Movie S2). Overall, our network captures junctional

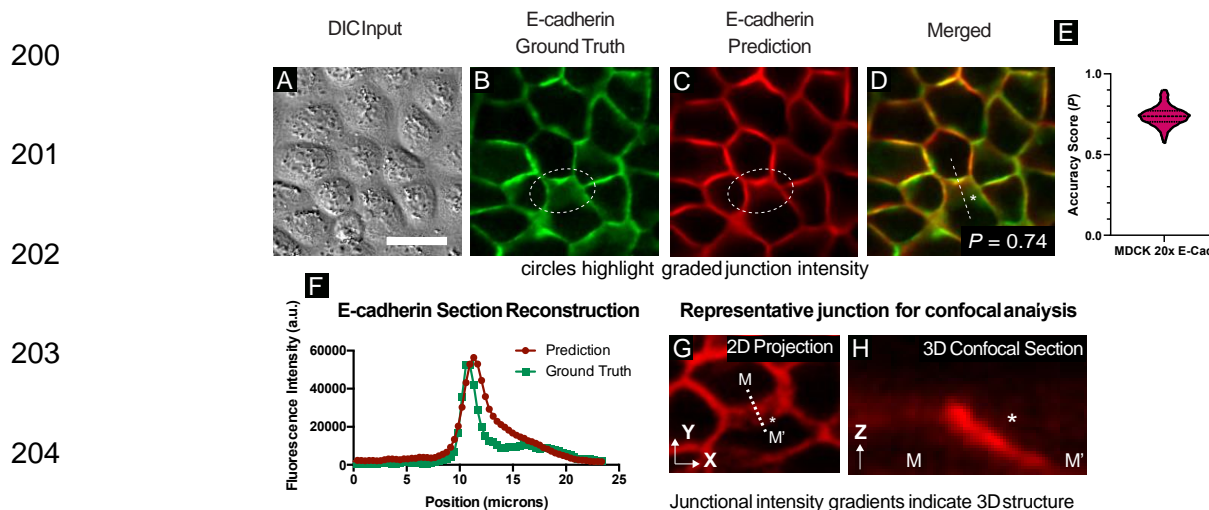199   intensity and geometry, both of which are invisible to the eye in the DIC input image.

200

circles highlight graded junction intensity

Junctional intensity gradients indicate 3D structure

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 the cell-cell junctions. From 2D transmitted light input (A), 3D structures may be approximated. (G, H) Representative cell-cell junction and corresponding confocal section, highlighting the

205

Fine structure reconstruction

207   In practice, high content imaging is inherently a trade-off between throughput and resolution.

208   The more detail we can extract from lower magnification images, the more efficient the imaging

209   and analysis. Here, we demonstrate the practical performance of FRM and a 20X/0.8NA

210   objective to reconstruct fluorescence signatures for several useful sub-cellular markers using

211   HUVEC cells that stably expressed VE-Cadherin:YFP (mCitrine) and were labeled with Hoechst

212   33342 (live nuclear dye) and SiR Actin (infrared live actin dye). Processed timelapse data (see

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

214   image shown in Fig. 4A.

215   As a baseline, we characterized prediction accuracy for cell nuclei as the nucleus itself is

216   relatively low resolution, but detection of sub-nuclear features requires higher accuracy. The

217   Fig. 4B column demonstrates FRM performance for 20X nuclei including a line section through

218   both the bulk structure and sub-nuclear granules. Visually, the FRM image is quite accurate,

219   and $P = 0.91$ in this case. The line section easily captures the bulk form of the nucleus, but does

220   not quite capture the texture inside the nucleus, although it does capture the rough form.

221   Next, we trained the network on identifying Actin after first staining HUVECs using the SiR-Actin

222   live imaging dye. Here, the column in Fig. 4C shows significantly reduced performance as the

223   fine F-actin filaments visible in the ground truth fail to be reconstructed in the predictions ($P =$

224   0.67) with the exception of some of the cortical filaments at the very edge of the cells (see the

225   line profile). We hypothesize this is primarily due to fundamental limitations of DIC imaging and

226   the lack of contrast for intracellular F-actin, but it may also be due to the network overprioritizing

227   cortical filaments and the diffuse cytoplasmic signal. However, in practice we found that these

228   FRM data were useful for general cell body detection and potential segmentation analyses due

229   to the relatively homogeneous reconstructed fluorescence in the cytoplasmic space.

230   Finally, we trained the network with VE-Cadherin:YFP data in an attempt to reconstruct not only

231   cellular borders, but also the well-characterized, nano-scale membrane fingers that develop in

232   endothelial cell-cell junctions and indicate the direction of front-rear polarity in each cell [30]. In

233   contrast to the actin performance, FRM proved far more capable here and readily detected both

234   general VE-Cadherin boundaries (Fig. 4D column) and the membrane fingers (Fig. 4E column),

235   although $P = 0.77$ still seems quite low and likely relates to the network attempting to

236   reconstruct the more variable granules in the center of the cell, which are irrelevant for

237   junctional analyses. While VE-Cadherin protrusions and boundaries are sometimes detectable

238   by eye in DIC (as in Fig. 4A), they are still quite subtle in the best case, and developing a

239   traditional computer vision process to detect and extract them has not been demonstrated,

240   thereby again highlighting the practical value of FRM to reconstruct not only fluorescence, but

241   also key morphological markers that are much easier to analyze in the FRM image than in the

242   DIC image. The statistical accuracy distributions are shown in Fig. 4F, where the spread of the

243   data in F-actin indicates the lack of reliability, while the tighter distributions for nuclei and VE-

244   cadherin indicate more useful reconstructions. Again, the value of FRM depends on the specific

245   question and context, and the decision of whether it is 'good enough' at detecting fine structures

246   rests with the end user. For reference, short movies of several markers are presented in Movies

247   S3-5. Additionally, training the network with large and varied datasets enable it to begin to

248   predict statistically rarer events, such as mitotic divisions, as shown in Movie S6.
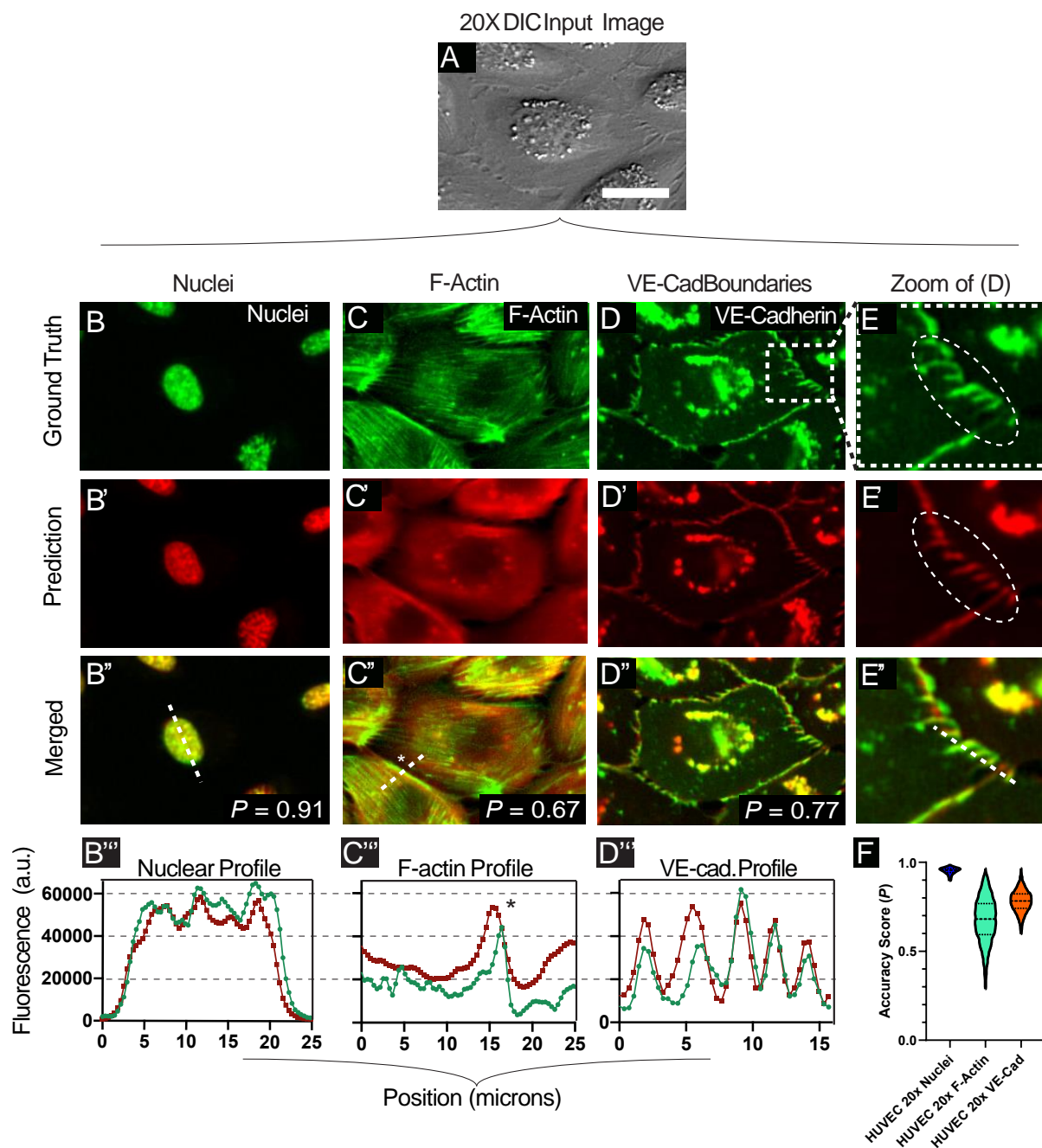

249

250

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.

251

252 ***Comparing FRM visual performance to P scores, training set size, and network***

253 A key feature of FRM is that its performance can often be increased by collecting more training

254 data, which in turn ought to improve *P*. However, *P* will never be perfect, nor is *P* necessarily

255 the best metric to go by when determining if an FRM image is 'good enough', as clearly the

256 context matters and the key question is 'good enough for what?' Hence, we sought to provide

257 several examples of how the size of the training set affects both *P* and the actual visual

258 accuracy or quality of the resulting FRM predictions.

259 To do this, we first swept through different sizes of training sets for many of the biomarkers

260 presented earlier (see *Methods*). Briefly, we selected random subsets of very large datasets and

261 trained the U-Net from scratch with these subsets. This process was repeated for different

262 fractions of the complete dataset to capture the FRM performance versus training set size. The

263 relation between *P* and training set size is shown in Fig. 5A to give a sense of how the

264 quantitative accuracy progresses and eventually plateaus. In contrast, we also provide FRM

265 results from the fractional training sets for the 20X HUVEC and MDCK data (Figs. 5B,C;

266 respectively). Here, the input and ground truth data are presented alongside representative

267 FRM predictions from networks trained with different numbers of training images (noted below

268 each image). As a single image from the camera (2048x2048 pixels) is first split into 256x256

269 pixel sub-images for training, a fractional image (e.g. Fig. 5C, 1/16th column) implies that the

270 network was trained on just a small crop from a single micrograph. Broadly speaking, and as

271 expected, these data all indicate that FRM quality varies directly with the size of the training set,

272 as expected.

273 However, Fig. 5A demonstrates that the rate of change in quantitative quality (*P*) vs. training set

274 size is neither linear nor is it uniform across different biomarkers. Further, the actual predicted

275 images shown in Figs. 5B,C offer further nuance because they demonstrate that training the

276   network against even a single image would be sufficient to capture nuclei for the purposes of

277   tracking or segmentation, while just 6 images would be sufficient to capture cell shape and

278   junctional geometry in epithelia assuming the researcher were willing to perform some simple

279   manipulations such as background subtraction. There is an obvious performance increase for

280   both cadherins when the training set comprises several hundred images, but it is difficult to

281   visually detect a difference between nuclei reconstructed from 6 or 400 images.

282   An alternate way to improve FRM would be to alter the U-Net architecture. Here, we first

283   compared the standard U-Net to a neural network architecture which was essentially two U-Nets

284   stacked end-to-end with additional residual connections. Such an approach has been shown to

285   improve network depth and performance in other applications[31–33]. Here, however, we observed

286   no benefit to training a deeper network (see Fig. S2). Further, given the significant temporal and

287   computational cost, we advise against its use for this kind of FRM. Alternately, we explored the

288   role of the loss function, testing our Pearson's-based loss function against the traditional Mean-

289   Squared-Error loss function and found no significant difference (Fig. S3; Methods). Hence, we

290   conclude that our minimal U-Net implementation performs well as a foundation for a variety of

291   daily analysis tasks without requiring significant fine tuning.

292

293

294
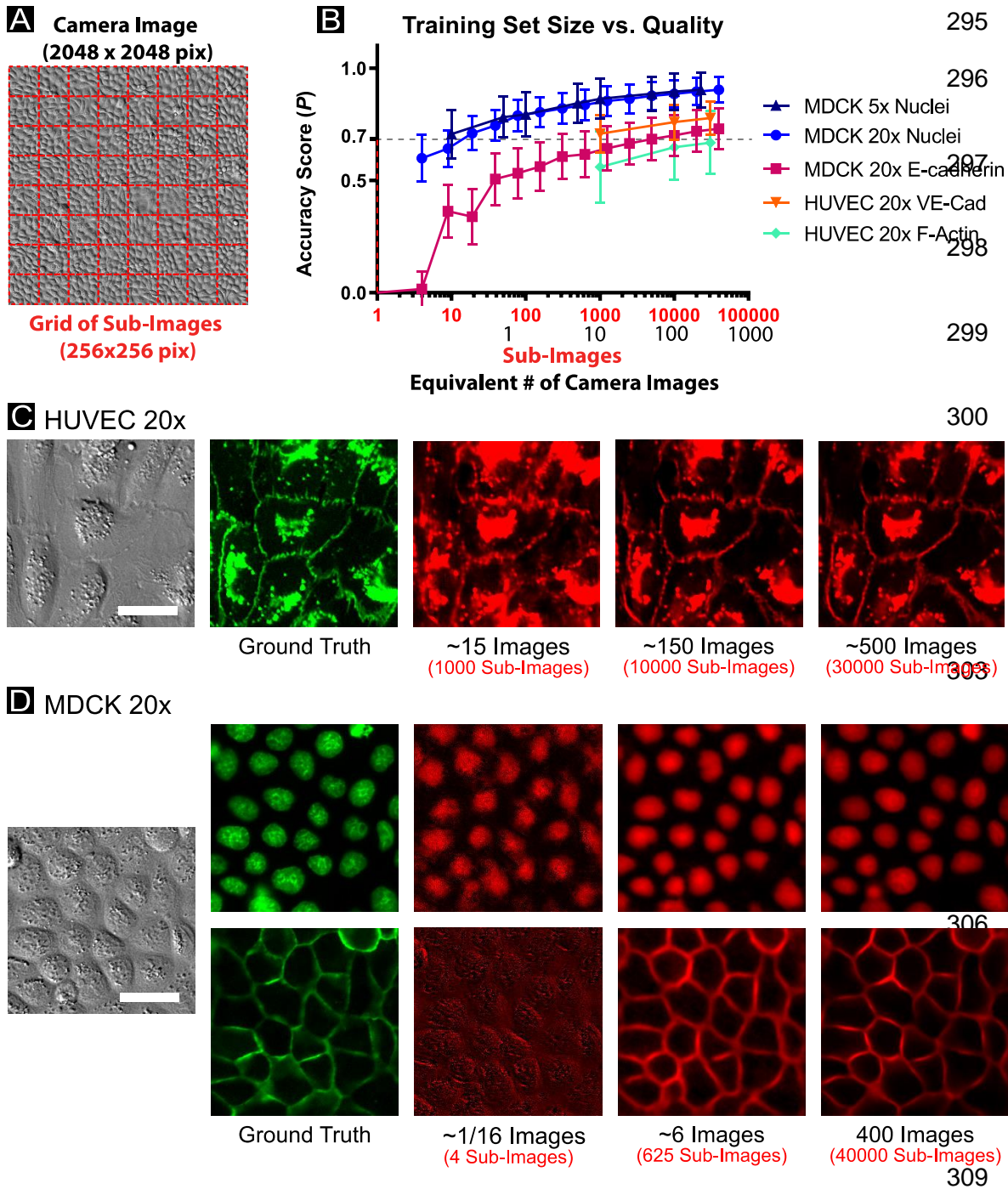
Figure 5: Impacts on prediction accuracy from smaller training sets.
(A) Cropping a sample image into 64 sub-images. (B) A comparison of network prediction accuracy as a function of training set size. The U-Net is trained with the complete dataset as described in Table S1 for each experimental condition. Then, random images representing a fraction of the total training set is used to train a new U-Net from scratch. (C, D) display representative images for the HUVEC 20x dataset and the MDCK 20x dataset, respectively, with predictions shown for various training set sizes. This type of analysis may assist users in collecting enough data for their task-specific quality requirements. All scale bars represent 30 μm.

311   **Discussion**

312

313   ***Limitations of existing accuracy metrics and the importance of context***

314   Our data—collected from actual, real-world analyses, highlight the limitations of using traditional

315   accuracy metrics from computer vision for biological image analysis. Specifically, while there is

316   a general relation between an improved P Accuracy Score and FRM quality, it is not linear nor

317   intuitive how to determine what is 'good enough' given only a P value devoid of context for a

318   specific analysis. Further, and most critically, FRM does not reconstruct images according to

319   human imperatives. The U-Net only optimizes via the specific loss function it has been given

320   (e.g. Mean-Squared-Error or the Pearson's coefficient). What the computer considers 'good'

321   need not match our own assessments of value and quality.

322

323   As a practical example, compare the FRM performance for E-cadherin ($P = 0.73$; Fig. 3) and F-

324   actin ($P = 0.67$; Fig. 5). While the accuracy metrics differ by < 10%, the FRM of F-actin only

325   detected peripheral actin cables, otherwise blurring all internal features into a homogeneous

326   signal. Nonetheless, even this plainly 'inaccurate' signal could prove useful for cytoplasmic

327   reconstruction and tracking. In stark contrast, the E-cadherin data was much more visually

328   accurate and also captured key quantitative features of the ground truth such as junctional

329   localization and intensity, and even the subtle intensity gradients representing 3D morphology

330   despite having only a slight improvement in P-values. Yet despite that, a score of 0.73 is far

331   enough from '1' that it is ambiguous in absence of a specific analysis, which is why FRM must

332   be evaluated in the context of a given question or analysis.

333

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

335   We specifically targeted the lower-magnification end of the imaging spectrum to explore how

336   well FRM performed at magnifications more commonly used for high content imaging

337   applications such as timelapse studies of very large cellular colonies or massive screens using

338   multiwell plates. Our data indicate that such magnifications can be effectively combined with

339   FRM for applications spanning nuclear tracking, cell-cell junction analysis, and certain fine-

340   structure reconstruction even at just 20X.

341

342   A particular concern for the average user of a complex machine learning process is the size of

343   the dataset required as this can impose potentially strenuous experimental demands. However,

344   our characterization of FRM performance vs. data set size again shows the importance of

345   context as relatively few images are needed to get quite accurate nuclei reconstruction, while a

346   greater number of images are needed for junction reconstruction (Fig. 6). However, we also

347   note that our largest training set size comprised at most 500 camera images at 20X

348   (approximately one six-well plate)—something easily obtained with a standard automated

349   microscope, and still compatible with manual capture. Further, a very common approach in

350   machine learning is to 'augment' an image dataset by performing reflections and rotations on

351   images such that the network perceives each augmented image as a different datapoint,

352   thereby virtually increasing the size of the dataset. We did not perform such augmentation here

353   for the sake of simplicity and transparency, which suggests that significantly smaller datasets, if

354   augmented, could still produce good results.

355

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

357   FRM is a quickly developing technology that exists alongside another popular approach where

358   machine learning is used for feature segmentation[14,19,34]. In the latter case, the network is

359   trained to specifically detect 'features' (e.g. nuclei) as binary objects, whereas FRM instead

360   reconstructs the effective fluorescent image of what a fluorescent label against that structure

361   might show. Both are useful techniques, and the best approach depends on the application.

362   However, there are several unique advantages to FRM. First, reconstructing an effective

363   fluorescent image from auto-annotated data (e.g. chemical dyes, antibodies, fluorescent

364   proteins) obviates the need for any manual annotation or pre-processing—often quite time

365   consuming and subjective. This means that an FRM image can be directly incorporated into any

366   existing analysis pipeline intended for fluorescent images, including traditional threshold-based

367   segmentation approaches. Further, more of the original data is preserved in an FRM image,

368   allowing the capture of things such as fluorescence intensity gradients (e.g. Fig. 3), and features

369   that might be lost during traditional binary segmentation.

370

371   ***Concluding remarks***

372   Here, we characterize the value of fluorescence reconstruction microscopy (FRM) for everyday

373   analysis tasks facing researchers working with cell biology. We specifically highlight the need for

374   individual researchers to explore and evaluate FRM in the context of specific research questions

375   rather than accuracy metrics. We also highlight the surprisingly good performance of FRM even

376   with lower magnification imaging or relatively fine structures such as VE-cadherin fingers.

377   Finally, we have made all of our tools and all training datasets publicly available to improve

378   accessibility and provide a starting point for researchers new to FRM to easily explore it for

379   themselves and to eventually build on and improve.

380 **Methods.**

381 *Tissue culture*

382 MDCK-II (G-type) cells stably expressing E-cadherin:dsRed were cultured in low glucose
383 DMEM. The MDCK-II culture media was supplemented with 10% Fetal Bovine Serum (Atlanta
384 Biological) and penicillin/streptomycin. HUVEC endothelial cells stably expressing VE-
385 cadherin:mCitrine were cultured using the Lonza endothelial bullet kit with EGM2 media
386 according to the kit instructions. Primary murine keratinocytes were isolated from neonatal mice
387 (courtesy of the Devenport Laboratory, Princeton University) and cultured in custom media[35]. All
388 cell types in culture were maintained at 37°C and 5% CO_2 in air.

389 *Preparation of training samples*

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

394 In order to contain a variety of conditions within a single plate to ensure a broad training sample,
395 we placed silicone microwells into the dishes as described in [[36]] at densities from [1-2x10^6
396 cells/mL] which ultimately allowed for single cells, low density confluent monolayers, and high
397 density confluent monolayers to be captured. Silicone microwells consisted of 3x3 arrays of 9
398 mm^2 microwells into which we added 4 µL of suspended cells in media, allowed them to adhere
399 for 30 min in the incubator (6 hrs for keratinocytes), added media and returned them to the
400 incubator overnight prior to imaging. To further ensure variability, several dishes were also
401 randomly seeded with cells for each cell type.

402 *Fluorescent labeling for ground truth data*

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

406 *Image Acquisition*

407 5X MDCK data was collected on a Zeiss (Observer Z1) inverted fluorescence microscope using
408 a 5X/0.16 phase-contrast objective, an sCMOS camera (Photometrics Prime) and controlled
409 using Slidebook (Intelligent Imaging Innovations, 3i). An automated XY stage, a DAPI filter set,
410 and a metal halide lamp (xCite 120, EXFO) allowed for multipoint phase contrast and
411 fluorescent imaging.

412 All epifluorescence imaging was performed using a Nikon Ti2 automated microscope equipped
413 with a 10X/0.3 phase objective, a 20X/0.75 DIC objective, and a Qi2 sCMOS camera (Nikon
414 Instruments). Time-lapse imaging effectively increased dataset size as long as sufficient time
415 was allowed between frames to avoid overfitting in the U-Net. MDCK data was collected at 20
416 min/frame, while HUVEC and keratinocytes were given 60 min/frame. Standard DAPI, CY5, and
417 YFP filters sets were used. Confocal sections of E-cadherin fluorescence in MDCK cells (Fig. 3)
418 were collected using a Leica SP8 scanning confocal tuned for dsRed excitation/emission.

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

422 ***Data Pre-Processing and Training***

423 Prior to input to the network, raw images were segmented into 256x256 pixel^2 sub-images,
424 ensuring consistent slicing across the transmitted-light image and the corresponding fluorescent
425 image. The images were then normalized by statistics collected across all images in each
426 channel: that is, by subtracting from each image the mean and dividing by the standard
427 deviation. A test-train split was applied, such that a random 20% of the total images were held
428 out to comprise the test set. Additionally, 10% of the training data subset were held out for
429 validation as is standard.

430   The U-Net style architecture shown in Figure 1 was trained using TensorFlow[37] and the
431 ADADELTA optimizer[38]. In the standard training experiments, the mean squared error (MSE)
432 loss function was applied across pixel intensity values in the predicted images compared to
433 intensity values in the ground truth images. Results from the MSE were contrasted with results
434 from two networks trained to maximize the Pearson's correlation coefficient (PCC). The PCC is
435 commonly used in cell biology for evaluating the colocalization of two fluorescently labeled
436 structures[39]. The PCC loss function was defined, for two intensity data sets R and G, as:

437
$$PCC = \frac{\sum_i (R_i - \bar{R}) \times (G_i - \bar{G})}{\sqrt{\sum_i (R_i - \bar{R})^2 \times \sum_i (G_i - \bar{G})^2}}$$

438
$$Loss = \frac{(1 - PCC)}{2}$$

439 Sample training loss plots are provided (Fig. S4), reflecting the use of early stopping during
440 training. That is, when the validation loss did not decrease for 75 epochs, the training process
441 terminated. The training and test set sizes and results are provided for all experimental
442 conditions in Table S1.

443   For several of the experimental conditions, networks were trained using subsets of the
444 original dataset. To do so, the network architecture was fixed, and networks were trained from
445 scratch by using a random subset of matched input-output image pairs from the original training
446 set. The training setup, including hyperparameters, was unmodified regardless of training set
447 size. The original test set was used to compare results for all training set size tests relative to
448 each experimental condition.

449 ***Data Testing and Image Processing***

450   The Pearson's correlation coefficient (PCC) of each test set is determined by individually
451 computing the PCC between each predicted image, as output by the network, and its
452 corresponding ground truth image. Additionally, an accuracy score (*P*) based on the PCC was
453 devised to more reliably represent the performance of the network. To determine P, we report
454 the PCC on a subset of the test set which selects for only those test images containing positive
455 examples of features (nuclei, junctions, etc.). We construct this subset by manually determining
456 threshold values to distinguish image intensities indicating the presence of the relevant features
457 versus background noise for each test set. That is, the histograms of a subset of the data
458 containing positive examples of features are plotted, and an approximate lower bound on

459   intensity values is estimated to distinguish the features from the background. Then, the
460   histograms of a subset of the data containing only background are plotted to ensure that the
461   threshold value is adequate to label the images as background-only images. The MatLab
462   function rmoutliers() was utilized to remove outliers when P is reported for each condition.

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

472   When intensity plots for line slices are reported, a line is selected as an ROI in
473   ImageJ/FIJI, and intensity values are exported for analysis.
474   New large transmitted-light images were processed using a sliding-window technique.
475   We processed a large image by analyzing 256x256 pixel^2 patches of the input image with a
476   stride of 64 pixels in each direction. Additionally, the border of each predicted patch was
477   excluded in the sliding-window process, as features near the patch borders are likely to have
478   lower accuracy (often as a function of cells being cut off). The sliding-window predictions at
479   each pixel were then averaged to produce the final large predicted image. Timelapse movies
480   can be processed on a frame-by-frame basis. If scaling was required as described in Fig. 3, the
481   input was scaled in FIJI and then passed to the network for analysis.

482   ***Code and Dataset Availability***

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

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

487   Additionally, our complete testing datasets, along with corresponding reconstructed images, are
488   available through our Zenodo repository, which can be found at:

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

490   **Acknowledgements**

Supplementary Materials

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		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		VE-cadherin	30720	7680	5666	0.60824	0.30398	0.77737	0.074176
20X		F-actin	30720	7680	5820	0.51247	0.30438	0.66808	0.14057

Table S1. Accuracy and training statistics for all experimental conditions.
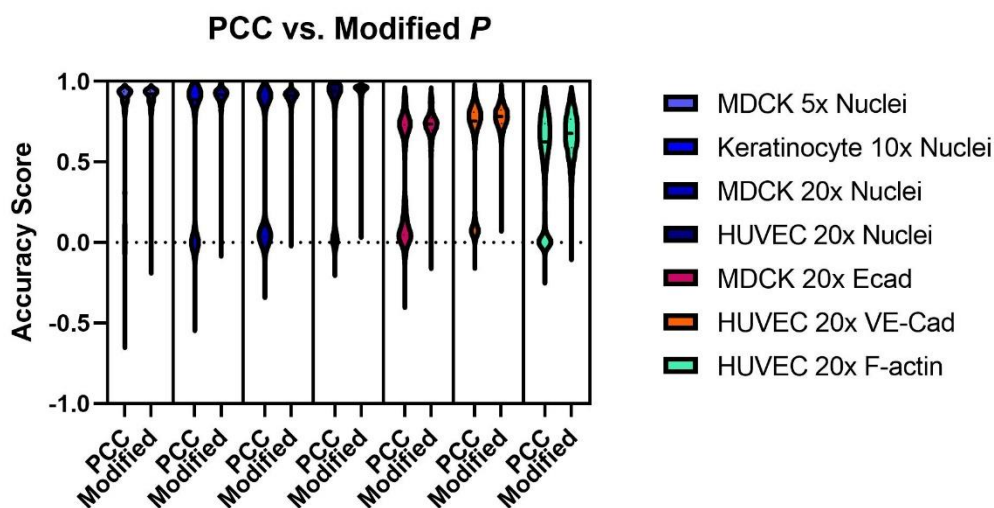
Figure S1. A comparison of experimental results in Pearson's Correlation Coefficient (PCC), versus a modified accuracy score P (see *Methods*). To ensure a fairer comparison, outliers were not removed; only intensity thresholding was performed to produce the modified P from the PCC. By filtering the PCC results by an intensity threshold in the fluorescent images, we remove low-scoring background images, which bias our accuracy score on the complete dataset. Visual inspection of the plot reveals the low-scoring images as "bumps" near 0.0.
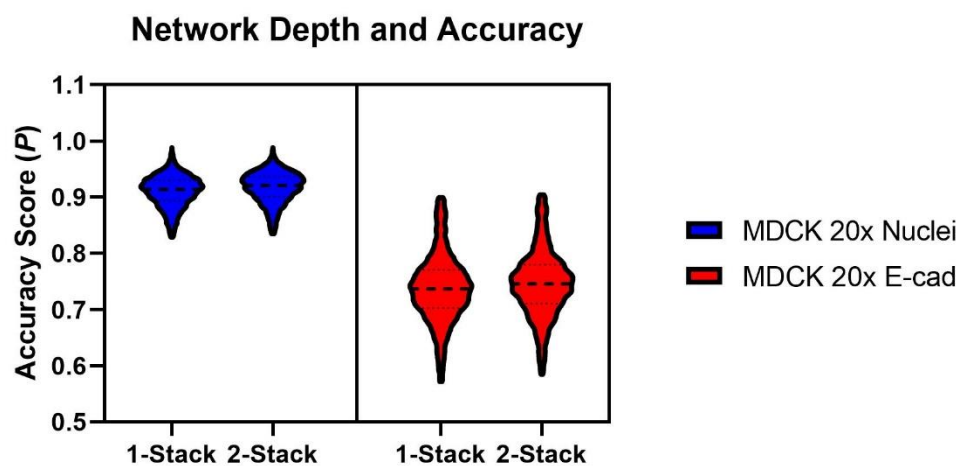
509
510

Network Depth and Accuracy

511

512 **Figure S2.** Representative accuracy results for a dataset trained on the standard (1-stack) U-
513 Net, compared to a network comprised of two U-Nets stacked back-to-back, with residual
514 connections (2-stack). Training conditions were otherwise unchanged. Accuracy scores, as
515 reported in terms of the modified P (see *Methods*) were comparable.

516

517

Training Loss Function and Accuracy
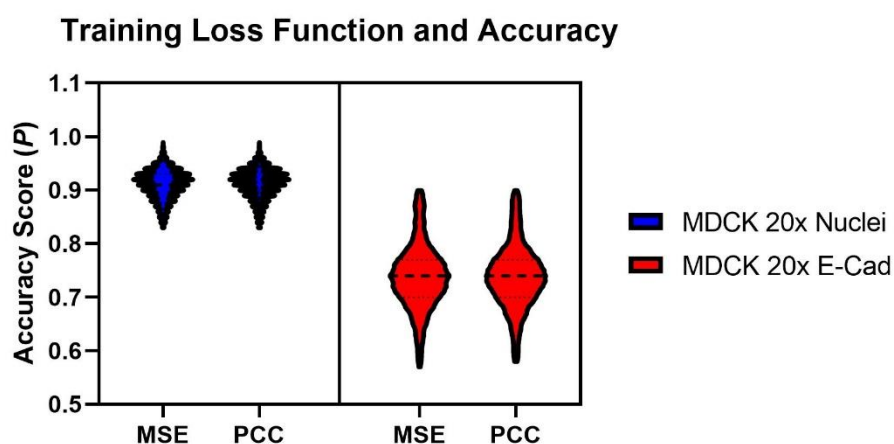
518

519 **Figure S3.** Representative accuracy results for networks trained using the Mean-Squared Error
520 loss function (MSE) compared to the Pearson's Correlation Coefficient loss function (PCC). The
521 neural network architecture and training conditions are the same, with the exception of the
522 choice of loss function. Accuracy scores, as reported in terms of the modified P (see *Methods*)
523 were comparable.

524

525

526
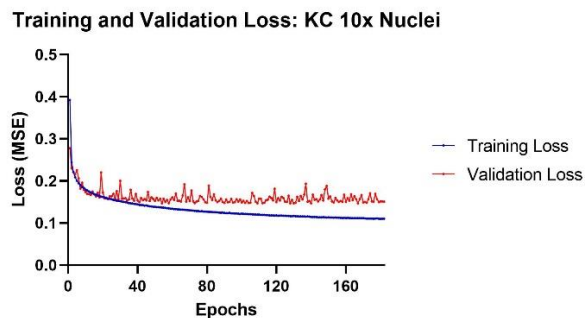
527
528
529 **Figure S4.** Representative loss functions from the standard U-Net training process. Early
530 stopping was enabled, so that if the validation loss did not decrease within a set number of
531 epochs, the training process terminated.

532

533

534

535

536

537

538   **Movie captions on the following page**

539

540

541

542

543

544

545

546

547   **Supplementary Movie Captions:**

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

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

554   **Movie S2. FRM for high-content screening.**

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

562   **Movies S3-5. FRM for fine structures.**

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

567   **Movie S6. Mitotic division prediction.**

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

573
574

References.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

31. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition*.

(2016).

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.   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).

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

36.   Cohen, D. J., Gloerich, M. & Nelson, W. J. Epithelial self-healing is recapitulated by a 3D biomimetic E-cadherin junction. (2016). doi:10.1073/pnas.1612208113

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

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

39.   Dunn, K. W., Kamocka, M. M. & McDonald, J. H. A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology - Cell Physiology* **300**, (2011).