Quantitative Biology 2018, 6(4): 344–358
https://doi.org/10.1007/s40484-018-0156-3

RESEARCH ARTICLE

Characterizing robustness and sensitivity of convolutional neural networks for quantitative analysis of mitochondrial morphology

Xiaoqi Chai[1], Qinle Ba[1] and Ge Yang[1,2,*]

[1] Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2] Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA
* Correspondence: geyang@andrew.cmu.edu

Received March 5, 2018; Revised August 1, 2018; Accepted August 2, 2018

***Background*: Quantitative analysis of mitochondrial morphology plays important roles in studies of mitochondrial biology. The analysis depends critically on segmentation of mitochondria, the image analysis process of extracting mitochondrial morphology from images. The main goal of this study is to characterize the performance of convolutional neural networks (CNNs) in segmentation of mitochondria from fluorescence microscopy images. Recently, CNNs have achieved remarkable success in challenging image segmentation tasks in several disciplines. So far, however, our knowledge of their performance in segmenting biological images remains limited. In particular, we know little about their robustness, which defines their capability of segmenting biological images of different conditions, and their sensitivity, which defines their capability of detecting subtle morphological changes of biological objects.**

***Methods*: We have developed a method that uses realistic synthetic images of different conditions to characterize the robustness and sensitivity of CNNs in segmentation of mitochondria. Using this method, we compared performance of two widely adopted CNNs: the fully convolutional network (FCN) and the U-Net. We further compared the two networks against the adaptive active-mask (AAM) algorithm, a representative of high-performance conventional segmentation algorithms.**

***Results*: The FCN and the U-Net consistently outperformed the AAM in accuracy, robustness, and sensitivity, often by a significant margin. The U-Net provided overall the best performance.**

***Conclusions*: Our study demonstrates superior performance of the U-Net and the FCN in segmentation of mitochondria. It also provides quantitative measurements of the robustness and sensitivity of these networks that are essential to their applications in quantitative analysis of mitochondrial morphology.**

Keywords: convolutional neural network; mitochondrial morphology; image segmentation; robustness; sensitivity

Author summary: Segmentation of mitochondria, the image analysis process of extracting geometry of mitochondria from their images, plays an important role in elucidating their biology. Convolutional neural networks (CNNs), artificial neural networks widely used in artificial intelligence, have achieved great success in segmenting mitochondria. However, little is known about their robustness in segmenting mitochondria under different image conditions and their sensitivity in detecting subtle mitochondrial shape changes. Here we develop a method of using synthesized images to characterize performance of CNNs, specifically FCN and U-Net. Our study demonstrates their superior performance in segmentation of mitochondria and directly quantifies their robustness and sensitivity.

INTRODUCTION

Mitochondria are essential organelles of eukaryotic cells, serving a wide range of important physiological functions such as energy production, metabolic regulation, and stress response [1,2]. Within the intracellular space, they

exhibit remarkably complex and dynamic morphology, which is known to be closely connected with their functions [3]. Quantitative analysis of mitochondrial morphology plays important roles in elucidating mitochondrial functions in cell physiology under normal conditions and in cell pathophysiology in related human diseases [3,4]. Such analysis depends critically on segmentation of mitochondria, the image analysis process that identifies mitochondria and extracts their morphology from images. So far, however, methods that segment mitochondria with high accuracy and reliability remain lacking. This is because mitochondria pose several challenges to image segmentation in practice. First, mitochondria are often densely positioned in space or interconnected into complex networks. To reliably identify and separate neighboring mitochondria within short distances is often a challenge. Second, because of their substantial sizes as well as their diverse spatial configurations, mitochondria often are partially or entirely blurred in images due to defocusing. Consequently, mitochondria in images often lack well-defined boundaries. Third, images of mitochondria can vary substantially in their conditions, such as their signal-to-noise ratios (SNRs) and levels of blurring. Such variations are common in fluorescence microscopy. Although a variety of algorithms have been used for segmentation of mitochondria [5–9], performance of such algorithms often varies substantially in practice because of the challenges.

Recently, convolutional neural networks (CNNs) have achieved great success in challenging image segmentation tasks in several disciplines, such as optical character recognition and semantic image annotation [10,11]. Such networks have also been used successfully for segmentation of biological images [12–14]. In this study, we investigate using CNNs to overcome the challenges of mitochondrial segmentation. So far, however, our knowledge of the performance of CNNs in segmenting biological images remains limited. In particular, we know little about their robustness, which defines their performance in segmenting biological images of different conditions. Given the substantial variations in conditions of mitochondrial images, sufficient robustness is a basic requirement for related segmentation algorithms. Because a key advantage of CNNs is that their parameters are learned automatically in their training [11], it is often assumed that if they are properly trained, they should provide robust performance under different image conditions. But this assumption has not been directly tested. Besides robustness, we also know little about their sensitivity, which defines their performance in detecting subtle morphological changes of biological objects. Given that conventional light microscopes are limited in their resolution to ~200 nm, *i.e.*, the Rayleigh limit, and that mitochondria in images lack well-defined boundaries, small but significant morphological changes of mitochondria often appear as subtle changes in fluorescence microscopy images [15]. Because reliable detection of such changes is essential to applications such as characterization of mitochondrial fusion and fission [15], sufficient sensitivity is another basic requirement for related segmentation algorithms.

In this study, our main goal is to characterize the performance of CNNs in segmentation of mitochondria from fluorescence microscopy images. To this end, we developed a method of using realistic synthetic images to characterize the robustness and sensitivity of CNNs. We started with manual segmentation of real experimental images of mitochondria to extract different instances of mitochondrial geometry. Then the result of manual segmentation was used as ground truth to generate realistic synthetic images of mitochondria by simulating statistical intensity distributions of mitochondrial signal and background noise from real images. To characterize robustness of CNNs, we simulated different image conditions by adding different levels of noise and blurring. To characterize sensitivity of CNNs, we simulated changes in mitochondrial morphology using simple morphological operations such as dilation and erosion. We used the synthetic images to train two widely adopted CNNs: the fully convolutional network (FCN) [16] and the U-Net [17] and characterized their performance on both synthetic and real images. Furthermore, we compared the two networks in performance against the adaptive active-mask (AAM) algorithm [5], which we chose as a representative of high-performance conventional segmentation algorithms. We found that the FCN and the U-Net consistently outperformed the AAM in accuracy, robustness, and sensitivity, often by a significant margin, and that the U-Net provided overall the best performance. In conclusion, our study demonstrates superior performance of the U-Net and the FCN in segmentation of mitochondria. It also provides performance measurements of these networks that are essential to their applications in quantitative analysis of mitochondrial morphology.

RESULTS

Generating realistic synthetic images for characterizing robustness of FCN, U-Net and AAM

To characterize robustness of the FCN, the U-Net, and the AAM in segmentation of mitochondria from fluorescence microscopy images, we generated realistic synthetic images that simulated different levels of SNR and blurring (Supplementary Figure S1), two image conditions that we found in practice to influence segmentation

performance strongly. Briefly, we started with manual segmentation of selected experimental images (Figure 1A). We traced boundaries of individual mitochondria and then converted regions enclosed by the traced boundaries into binary masks (Figure 1B), which we used as ground truth (Materials & Methods). Next, we generated realistic synthetic images from these binary masks. Specifically, we sampled the image intensity distribution of actual mitochondria in experimental images and simulated the distribution at each pixel within

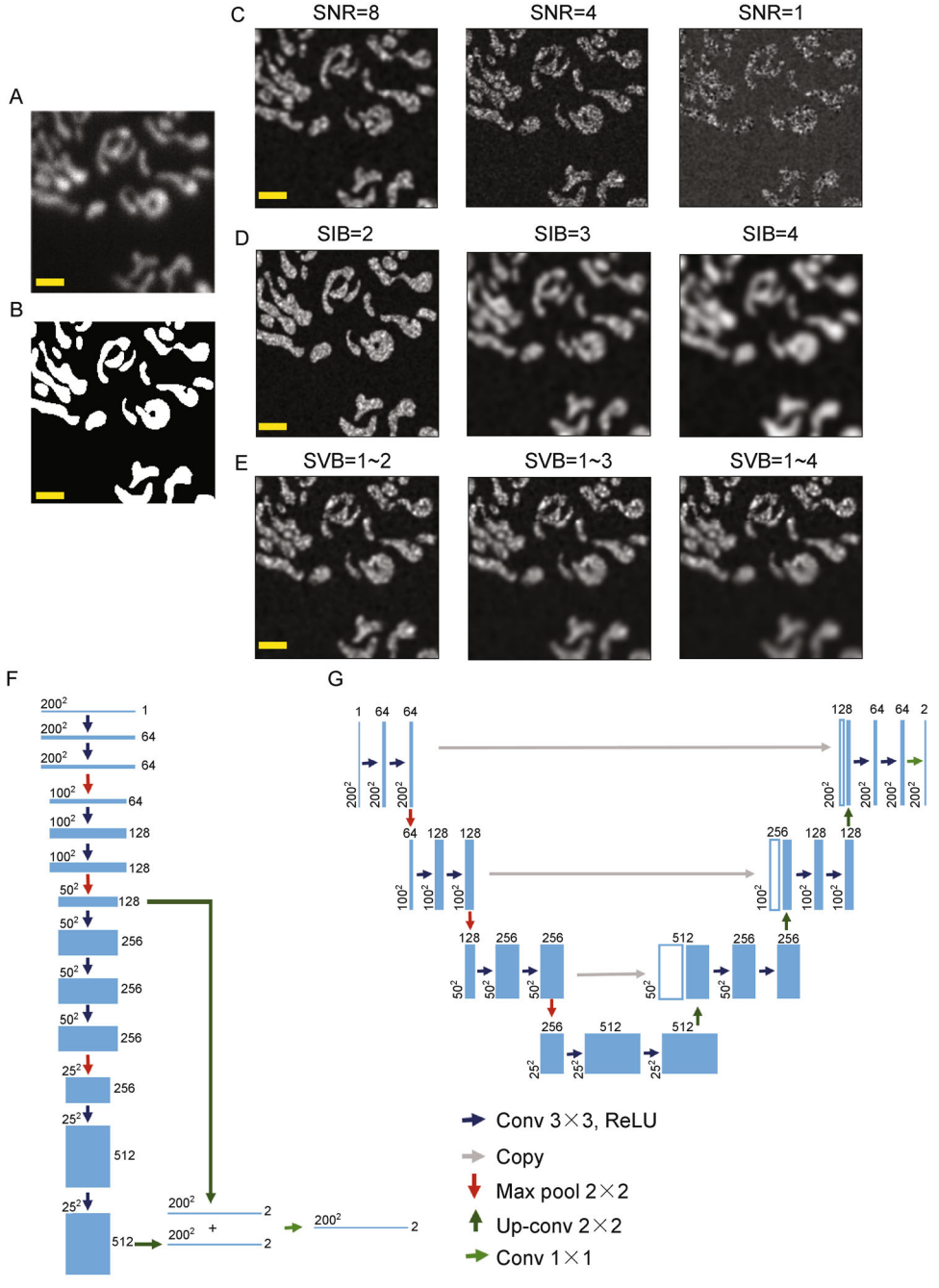

Figure 1. Generating realistic synthetic images for training and testing CNNs. (A–B) A raw image (A) and its corresponding binary mask image (B) generated by manual segmentation. (C) Synthetic images with different SNRs. (D) Synthetic images with different levels of space-invariant blurring (SIB) at SNR = 4. The size (*i.e.*, σ) of the Gaussian kernel applied was listed above each image. (E) Synthetic images with different levels of space-variant blurring (SVB) at SNR = 4. The range of the size (*i.e.*, σ) of the Gaussian kernel applied was listed above each image. (A–E) Scale bars: 2 μm. (F) Architecture of the FCN used in this study. (G) Architecture of the U-Net used in this study.

the regions defined by the binary masks (Materials & Methods). We then simulated different levels of SNR (Figure 1C), space-invariant blurring (Figure 1D), and space-variant blurring (Figure 1E) in the synthetic images, following largely the strategy described in [18] (Materials & Methods).

We chose not to use manual segmentation of real experimental images of mitochondria as ground truth for network training mainly for two reasons. First, it was not feasible to control the level of SNR or blurring in experimental images directly. Furthermore, because the manual segmentation process was laborious and time-consuming, it was not practical to label a large number of experimental images to represent a wide range of image conditions. Second, it was difficult to achieve sufficient consistency in manual segmentation to generate ground truth of real experimental images. This was mainly because boundaries of mitochondria were diffusive, due to low SNRs as well as blurring caused by the limited depth-of-field of the microscope under the high numerical aperture (NA) required for high spatial resolution. Indeed, we found that the level of inconsistency in manual segmentation performed by different investigators was not negligible. Importantly, because the actual ground truth for real images was unknown, using results of manual segmentation as ground truth would unavoidably introduce the inconsistency of manual segmentation into subsequent characterization of robustness and sensitivity. This problem was avoided by using realistic synthetic images because their ground truth was the result of manual segmentation and thus known exactly.

Characterizing robustness of FCN, U-Net and AAM in segmentation of mitochondria

Using the synthetic images generated, we trained the FCN (Figure 1F) [16] and the U-Net (Figure 1G) [17] and characterized their robustness. We also characterized the robustness of the AAM using the same synthetic images to provide a reference for comparing the two networks with high-performance conventional segmentation algorithms not based on CNNs. Overall, our strategy to characterize robustness was to examine how the segmentation accuracy of the two networks and the AAM changed under different image conditions. We quantified segmentation accuracy using two metrics. The first metric was area similarity, which quantifies the level of area overlap between the segmentation result and the corresponding ground truth (Materials & Methods). The second metric was boundary distance, which quantifies the average distance, *i.e.*, deviation, of the boundary of the segmentation result from the boundary of the corresponding ground truth (Materials & Methods). For each condition, defined by the level of SNR and the level of blurring, the FCN and the U-Net were trained then tested using synthetic images generated for that condition (Materials & Methods).

We first examined segmentation accuracy and robustness of the FCN, the U-Net, and the AAM under different SNRs (Figure 2A–2C). All three methods showed higher area similarity and lower boundary distance under higher SNRs, and the U-Net provided overall the best segmentation accuracy under the different SNRs examined. Under SNRs of 4 and 8, the U-Net outperformed the FCN and the AAM in area similarity and boundary distance by a significant margin ($p < 0.001$; Figure 2B–2C), while the FCN and the AAM had statistically the same ($p \geqslant 0.14$) area similarity (Figure 2B). Under an SNR of 8, the FCN had slightly lower ($p = 0.032$) boundary distance than the AAM (Figure 2C). But under an SNR of 4, boundary distances of the FCN and the AAM were the same ($p = 0.23$) (Figure 2C). Under a low SNR of 1, the AAM decreased in area similarity by 28.7% and increased in boundary distance by 111% from an SNR of 4 ($p < 0.001$ in both cases) (Figure 2B and 2C). Although the U-Net and the FCN also showed worsened performance at an SNR of 1 compared to at an SNR of 4, their changes in area similarity and boundary distance were more moderate (Figure 2B and 2C). For simplicity, we calculated the average slope of the metric curve for each method and used it as the quantitative metric of robustness. It should be noted that this metric of robustness was defined under the condition that the networks were trained and tested for each SNR individually (Materials & Methods). For area similarity, the metrics were 0.0112, 0.0236, and 0.0449 for the FCN, the U-Net, and the AAM, respectively. For boundary distance, the metrics were 0.0466, 0.1023, and 0.2162 for the FCN, the U-Net, and the AAM, respectively. Taken together, the results showed that the U-Net significantly outperformed the FCN and the AAM in segmentation accuracy in most cases under the different SNRs examined. The FCN significantly outperformed the AAM in accuracy under a low SNR of 1 but otherwise largely matched AAM in accuracy under an SNR of 4 or 8. Both the U-Net and the FCN substantially outperformed AAM in robustness but the U-Net slightly underperformed the FCN in this aspect.

Next, we examined the segmentation accuracy and robustness of the FCN, the U-Net, and the AAM under different levels of space-invariant blurring (Figure 2A, 2D and 2E). All three methods showed lower area similarity and higher boundary distance under higher levels of blurring. Again, the U-Net provided overall the best segmentation accuracy and significantly ($p \leqslant 0.005$) outperformed the FCN and the AAM in most cases under the different levels of blurring tested. The only exception was at a blurring level of 4 where the U-Net and the FCN

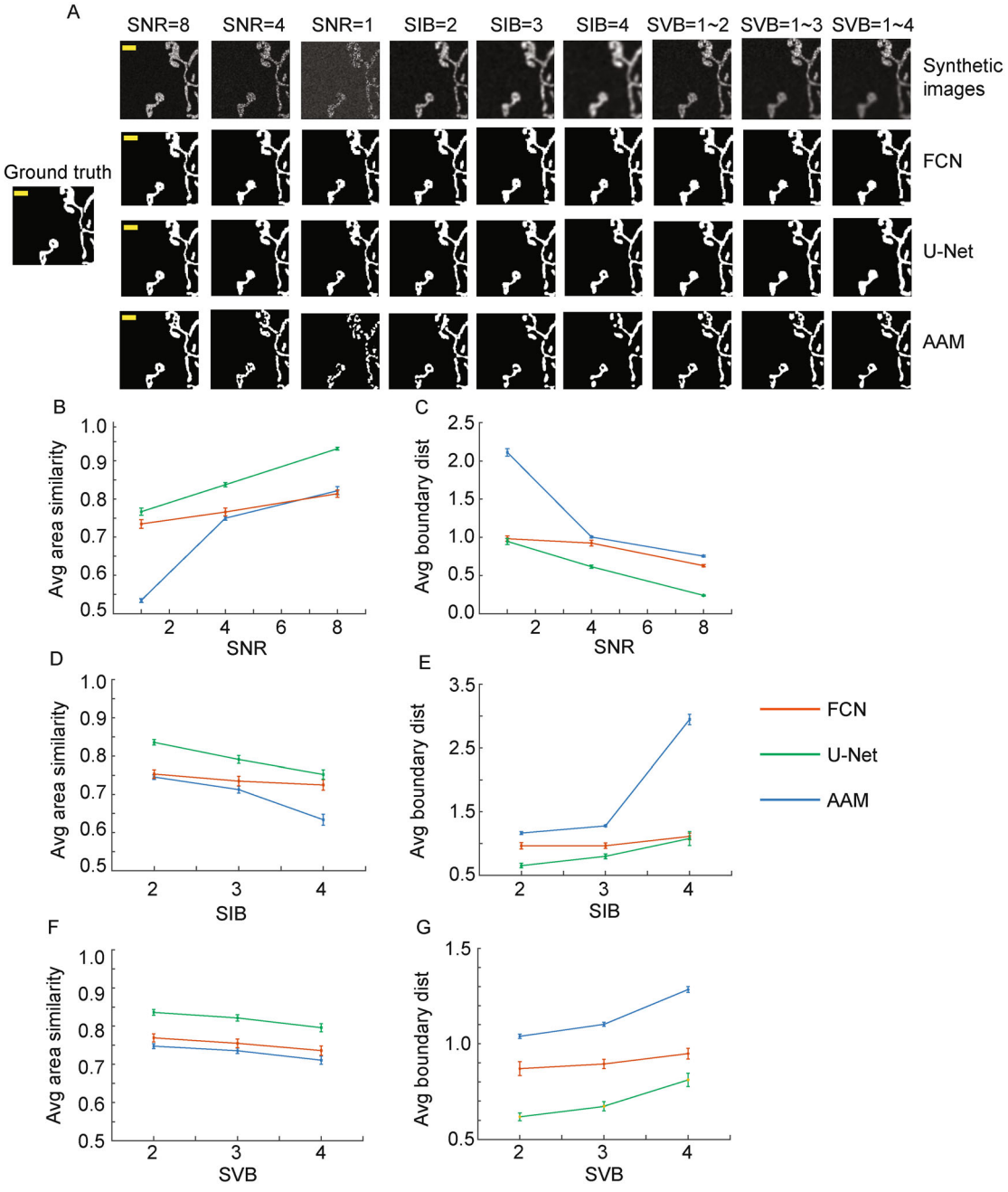

Figure 2. Characterizing robustness of FCN, U-Net, and AAM. (A) Segmentation results using the three methods under different image conditions. Scale bars: 2 μm. (B) Average area similarity under different SNRs. Most of the pairwise differences between the methods under each SNR were statistically significant with $p < 0.001$ except the following: SNR = 8, FCN *vs* AAM, $p = 0.58$; SNR = 4, FCN *vs* AAM, $p = 0.14$; SNR = 1, FCN *vs* U-Net, $p = 0.033$. (C) Average boundary distance under different SNRs. Most of the pairwise differences under each SNR were statistically significant with $p < 0.001$ except the following: SNR = 8, FCN *vs* AAM, $p = 0.032$; SNR = 4, FCN *vs* AAM, $p = 0.23$; SNR = 1, FCN *vs* U-Net, $p = 0.51$. (D) Average area similarity under different levels of space-invariant blurring. Most of the pairwise differences under each level of blurring were statistically significant with $p < 0.001$ except the following: SIB = 2, FCN *vs* AAM, $p = 0.54$; SIB = 3, FCN *vs* AAM, $p = 0.16$; SIB = 4, FCN *vs* U-Net, $p = 0.13$. (E) Average boundary distance under different levels of space-invariant blurring. Most of the pairwise differences under each level of blurring were statistically significant with $p < 0.001$ except the following: SIB = 2, FCN *vs* AAM, $p = 0.11$; SIB = 3, FCN *vs* U-Net, $p = 0.005$; SIB = 4, FCN *vs* U-Net, $p = 0.72$. (F) Average area similarity under different levels of space-variant blurring. Most of the pairwise differences under each level of blurring were statistically significant with $p < 0.001$ except the following: SVB = 1~2, FCN *vs* AAM, $p = 0.088$; SVB = 1~3, FCN *vs* AAM, $p = 0.16$; SVB = 1~4, FCN *vs* AAM, $p = 0.12$. (G) Average boundary distance under different levels of space-variant blurring. Most of the pairwise differences under each level of blurring were statistically significant with $p < 0.001$ except the following: SVB = 1~2, FCN vs AAM, $p = 0.017$; SVB = 1~3, FCN *vs* AAM, $p = 0.0012$; SVB = 1~4, FCN *vs* U-Net, $p = 0.003$. (B–G) All errors bars indicate standard error of the mean (S.E.M.). All comparisons were made using two-sample student's t-test. The number of samples for each data point: $n = 30$.

showed similar ($p \geqslant 0.13$) segmentation accuracy. Under blurring levels of 2 and 3, the FCN matched ($p \geqslant 0.16$) the AAM in area similarity. The FCN also matched ($p = 0.011$) the AAM in boundary distance under a blurring level of 2 but outperformed the AAM significantly ($p < 0.001$) in boundary distance under a blurring level of 3. Under a blurring level of 4, the FCN significantly outperformed the AAM in both area similarity and boundary distance. When the level of blurring increases from 3 to 4, the AAM showed a 12.7% decrease in area similarity and a 131% increase in boundary distance ($p < 0.001$ in both cases), while the FCN and the U-Net showed more moderate changes in both metrics. We again used the average slope of the performance metric curve as the metric of robustness for each method. For area similarity, the metrics were 0.0143, 0.0418, and 0.0559 for the FCN, the U-Net, and the AAM, respectively. For boundary distance, the metrics were 0.0801, 0.2135, and 0.8907 for the FCN, the U-Net, and the AAM, respectively. Taken together, the results showed that the U-Net significantly outperformed the FCN and the AAM in segmentation accuracy in most cases under the different levels of space-invariant blurring examined. The FCN significantly outperformed the AAM in accuracy under a high blurring level of 4 but otherwise largely matched AAM in accuracy under a blurring level of 2 or 3. Both the U-Net and the FCN substantially outperformed AAM in robustness but the U-Net slightly underperformed the FCN in this aspect, similar to the case under different SNRs.

Lastly, we examined the segmentation accuracy and robustness of the FCN, the U-Net, and the AAM under different levels of space-variant blurring (Figure 2A, 2F and 2G). Overall, changes in segmentation performance under the conditions examined followed similar trends as for space-invariant blurring for each method, namely lower area similarity and higher boundary distance under higher levels of blurring. However, because the images were only partially blurred, the performance changes were more moderate compared to the case under space-invariant blurring. For area similarity, the robustness metrics were 0.0166, 0.0200, and 0.0184 for the FCN, the U-Net, and the AAM, respectively. For boundary distance, the robustness metrics were 0.0394, 0.0958, and 0.1226 for the FCN, the U-Net, and the AAM, respectively. Overall, the results showed that the U-Net significantly ($p \leqslant 0.003$) outperformed both the FCN and the AAM in segmentation accuracy under different levels of space-variant blurring. The FCN matched ($p \geqslant 0.088$) the AAM in area similarity but outperformed ($p \leqslant 0.017$) the AAM in boundary distance. The results also showed that both the U-Net and the FCN largely outperformed the AAM in robustness under the conditions examined but the U-Net slightly underperformed the FCN in this aspect, similar to the cases under different SNRs and different levels of space-invariant blurring.

In summary, our results demonstrated that the U-Net consistently outperformed the AAM in segmentation accuracy and robustness, often by a significant margin. The U-Net also outperformed the FCN significantly in segmentation accuracy in most cases but slightly underperformed the FCN in robustness. Lastly, the FCN significantly outperformed the AAM in segmentation accuracy under a low SNR of 1 or a high space-invariant blurring level of 4 but otherwise largely matched the AAM in segmentation accuracy. The quantitative measurements of segmentation accuracy and robustness of the three methods provide information that is essential to their applications in quantitative analysis of mitochondrial morphology.

Characterizing sensitivity of FCN, U-Net, and AAM in segmentation of mitochondria

After characterizing the robustness of the FCN, the U-Net and the AAM, we went on to characterize their sensitivity, which defines their capability of detecting subtle changes of mitochondrial morphology. To this end, we chose to generate synthetic images at an SNR of 4 without additional blurring (Figure 3A) because this was the most representative of the image conditions in our studies. We used image erosion and dilation to simulate contraction and expansion of mitochondrial morphology, respectively (Figure 3A). Specifically, starting from a ground truth image, a synthetic image without deformation was generated first. We refer to this image as the original image. Then, the ground truth image underwent different levels of erosion and dilation to simulate different levels of deformation. For each level of simulated deformation, a synthetic image was generated from the eroded or dilated ground truth image (Figure 3A). The FCN and the U-Net were trained using original images only. They were then tested using different pairs of images. Each pair consisted of an original image and its corresponding synthetic image of a certain level of deformation. In this way, for each pair of testing images, a pair of segmentation result images were generated by each network. The pair of result images were then compared with their corresponding ground truth images (Figure 3A).

To characterize the sensitivity of the three methods in detecting deformation of mitochondrial morphology, we used two metrics. The first metric, which we refer to as deformation area similarity (DAS), quantifies the area similarity between the change in ground truth after erosion/dilation versus the change in segmentation results after erosion/dilation (Materials & Methods). The second metric, which we refer to as deformation distance

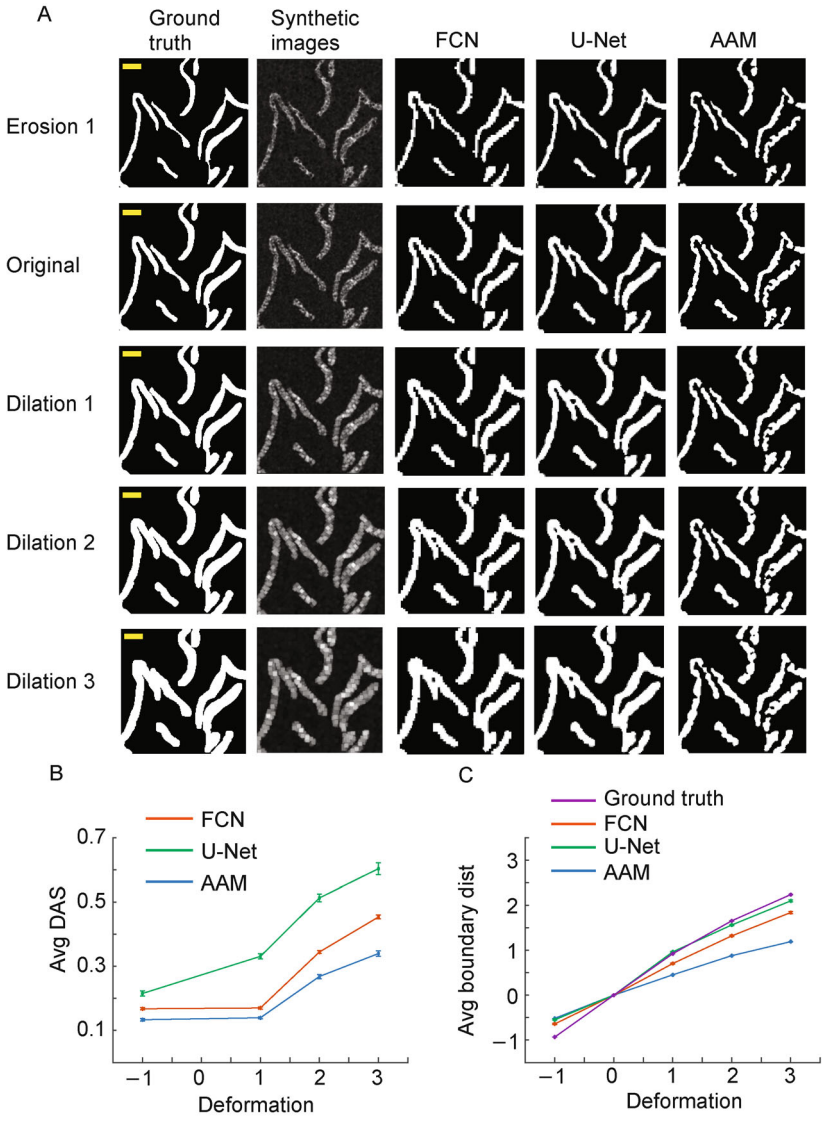

Figure 3. Characterizing sensitivity of FCN, U-Net, and AAM. (A) Segmentation results under different levels of deformation simulated by image erosion and dilation. Scale bars: 2 μm. (B) Average deformation area similarity under different levels of deformation. All the pairwise differences between the three methods under each level of deformation were statistically significant with $p < 0.001$. (C) Average deformation boundary distances under different levels of deformation. The deformation boundary distance is the average distance between the segmentation result of the original image and the segmentation result of the corresponding erosion/dilation image. The ground truth of the deformation boundary distance was shown in purple for comparison. All the pairwise differences between the three methods under each level of deformation were statistically significant with $p < 0.001$. (B and C) All error bars indicate S.E.M. All comparisons were made using two-sample student's t-test. Number of samples for each data point: $n = 30$.

similarity (DDS), is the ratio between the average detected boundary displacement after erosion/dilation and the average boundary displacement in the ground truth after erosion/dilation (Materials & Methods).

We first characterized the sensitivity of the three methods using DAS (Figure 3B). To this end, we simulated contraction of mitochondria using an erosion of 1 pixel and expansion of mitochondria using a dilation of 1–3 pixels. We found that under erosion or dilation of 1 pixel, the DAS was very low, at 0.17 for the FCN for both erosion and dilation, at 0.21 and 0.33 for the U-Net for erosion and dilation, respectively, and at 0.13 for the AAM for both erosion and dilation (Figure 3B). These low rates were due to the fact that the effective pixel size in our study was ~65 nm (Materials & Methods). When the deformation in the ground truth was only ~1 pixel, the

corresponding changes in synthetic images were much wider because of convolution with the microscope point spread function, whose radius was ~200 nm. For all three methods, the DAS increased as the magnitude of dilation increased to 2 and 3 pixels (Figure 3B). This was expected because the increased magnitude of deformation started to approach the radius of the point spread function. Overall, these results indicate that as a sensitivity metric, DAS is more suitable for large deformations. For this reason, we only used the average of DAS at 2 and 3 pixels of dilation to quantify the sensitivity of each method. The metrics were 0.3987, 0.5581 and 0.3035 for the FCN, the U-Net, and the AAM, respectively. Taken together, the results showed that, when characterized using DAS, the U-Net and the FCN consistently and significantly ($p < 0.001$) outperformed the AAM in sensitivity and the U-Net provided the highest sensitivity among the three methods.

We then characterize the sensitivity of the three methods using DDS (Figure 3C). Unlike the DAS, DDS was much less sensitive to the magnitude of deformation. Therefore, we used the average of the DDS over all levels of deformation to quantify the sensitivity of each method. The metrics were 0.7672, 0.8788, and 0.5271 for the FCN, the U-Net, and the AAM, respectively (Figure 3C). Overall, our results showed that, when characterized using DDS, the U-Net and the FCN consistently and significantly ($p < 0.001$) outperformed the AAM in sensitivity, with the U-Net providing the highest sensitivity among the three methods. It is also worth noting that all three methods were able to detect deformation at ~1 pixel, about one third of the diffraction limit of the light microscope.

Characterizing performance of FCN and U-Net trained for multiple image conditions

So far, in characterizing robustness of the FCN and the U-Net, we only trained them for a single image condition, *i.e.*, a specific SNR or a specific level of blurring, and tested them for that condition. However, actual fluorescence microscopy images of mitochondria may vary substantially in their conditions. This raised the question of whether and, if so, how performance of the FCN and the U-Net would be affected if they were used to segment images of conditions different from what they had been trained for. To answer this question, we first trained the networks using synthetic images with an SNR of 8 or 4 and then tested them on synthetic images with an SNR of 8, 4, and 1, respectively (Supplementary Figure S2A–S2D). We found that when the FCN and the U-Net were trained at an SNR of 8 and tested at an SNR of 4 or 1, performance of both networks degraded significantly at striking rates in both area similarity (Supplementary Figure S2A) and boundary distance (Supplementary Figure S2C) compared to their performance at a matching SNR of 8. On average, the mean area similarity of the FCN and the U-Net decreased by ~89.6% and ~91.3%, respectively, while the mean boundary distance of the FCN and the U-Net increased by ~51.2% and ~339.4%, respectively (Supplementary Tables S1–S3). When the FCN and the U-Net were trained at an SNR of 4 and tested at an SNR of 1, their performance also degraded significantly but at reduced rates. The mean area similarity of the FCN and the U-Net decrease by ~36.4% and ~13.6%, respectively (Supplementary Figure S2B), while the boundary distance of the FCN and the U-Net increased by ~38.1% and ~47.3%, respectively (Supplementary Figure S2D) from their performance at a matching SNR of 4. The performance of the FCN and the U-Net trained at an SNR of 4 either remained unchanged or improved slightly when tested at an SNR of 8. Specifically, the mean area similarity of the U-Net increased significantly by ~6.0%, while the mean boundary distance of the FCN and the U-Net decreased significantly by 10.4% and 36.2%, respectively (Supplementary Table S1–S3). Taken together, these results showed that performance of the FCN and the U-Net degraded significantly when trained at a certain SNR and tested at a lower SNR. The degradation was especially pronounced when the networks were trained at a comparatively high SNR of 8.

To further investigate the performance of the FCN and the U-Net in handling images of different conditions, we trained them using a mixture of synthetic images generated for those conditions. Specifically, we first trained the FCN and the U-Net with the same total number of images as previously for single SNRs (Materials & Methods). However, the set of training images consisted of equal numbers of synthetic images with an SNR of 1, 4, and 8, respectively (Figure 4A and 4B). We then tested the networks under an SNR of 1, 4, and 8, respectively. Similarly, we trained the FCN and the U-Net with equal numbers of synthetic images with a space-invariant blurring level of 2, 3, and 4, respectively, and then tested them under these conditions respectively (Figure 4C and 4D). We found that for either the FCN or the U-Net trained for multiple conditions, its performance matched the performance of the FCN or the U-Net trained and, therefore, optimized for single conditions in most of the cases tested (Figure 4A–4D). In some cases (*e.g.*, Figure 4B), the FCN and the U-Net trained for multiple conditions showed slight performance improvement over the FCN and the U-Net trained for single conditions. Taken together, our results demonstrated that when trained properly for multiple conditions, the FCN and the U-Net could simultaneously achieve at all the trained conditions at least the same level of performance as when they were trained for each single condition.

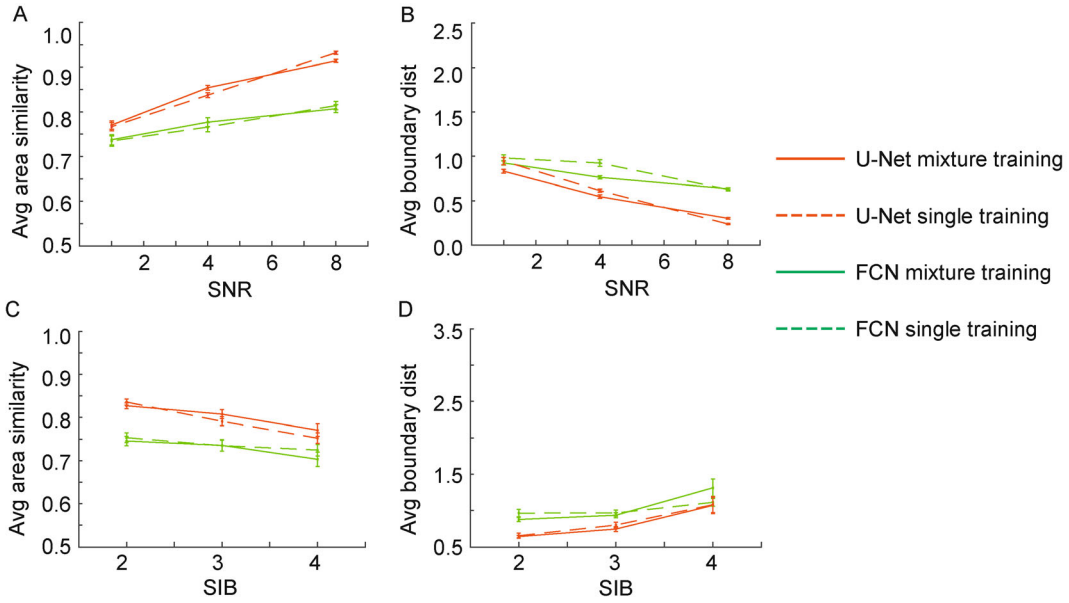

Figure 4. Comparison of performance of FCN and U-Net trained with multiple *vs* single conditions. (A) Average area similarity of FCN and U-Net trained for multiple SNRs (solid lines) *vs* single SNRs (dashed lines). Most of the pairwise differences under each SNR were statistically not significant with $p > 0.1$ except the following: U-Net, single *vs* mixture training, SNR = 4, 0.041. (B) Average boundary distance of FCN and U-Net trained for multiple SNRs (solid lines) *vs* single SNRs (dashed lines). Most of the pairwise differences under SNR were statistically not significant with $p > 0.1$ except the following: FCN SNR = 4, 0.0002. U-Net SNR = 1, 0.015, SNR = 4, 0.017; SNR = 8, 0.00045. (C) Average area similarity of FCN and U-Net trained for multiple levels of blurring (solid lines) versus single levels of blurring (dashed lines). All pairwise differences under each level of blurring were not statistically significant with $p > 0.1$. (D) Average boundary distance of FCN and U-Net trained for multiple levels of blurring (solid lines) versus single levels of blurring (dashed lines). All pairwise differences under each level of blurring were not statistically significant with $p > 0.1$. Error bars indicate S.E.M. Differences were tested using two sample student's t-test. Number of samples for each data point: $n = 30$.

Comparing performance of FCN, U-Net and AAM in segmentation of real images

After characterizing the performance of the FCN, the U-Net, and the AAM on synthetic images, we examined their performance on real experimental images of mitochondria. To this end, we trained the FCN and the U-Net using a mixture of equal numbers of synthetic images with an SNR of 8, 4, and 1, respectively. We then tested them on a group of 16 real images (Figure 5A and 5B; Supplementary Figure S4). The group of images were chosen to represent the different levels of SNRs in our study, with the mean of their SNRs at 7.43 and the standard deviation of their SNRs at 2.05. Because the ground truth was not available, we assessed the segmentation results qualitatively. Overall, although the three methods provided comparable segmentation results, the FCN performed worse than the U-Net and the AAM, as indicated by the irregular boundaries and over-segmentation in its results. Although the U-Net provided qualitatively better results than the FCN, it also showed some over-segmentation, as indicated by the enlargement of segmented mitochondria and merging of neighboring mitochondria. The AAM performed qualitatively better than the U-Net in terms of its lower level of over-segmentation.

To reduce the level of over-segmentation of the U-Net and the FCN, we tested two post-processing strategies. For the first strategy, we adjusted the threshold for pixel classification. In the initial segmentation result of each image, each pixel was associated with a probability defining its likelihood of belonging to the mitochondria or the background. The initial probability threshold was set to be 0.5, the same as the one used previously for synthetic images. An empirical adjustment of the threshold to 0.9 reduced the over-segmentation for both the FCN and the U-Net (Figure 5A and 5B, second and third rows). For the second strategy, we set the initial segmentation results of the FCN and the U-Net as the initial condition for a level-set segmentation [19], which we used for further refinement of the CNN segmentation results (Materials & Methods). This strategy avoided the empirical adjustment of the threshold in the first strategy. The refined results of the FCN and the U-Net were similar. Both showed reduced over-segmentation (Supplementary Figure S3A and S3B). Taken together, the

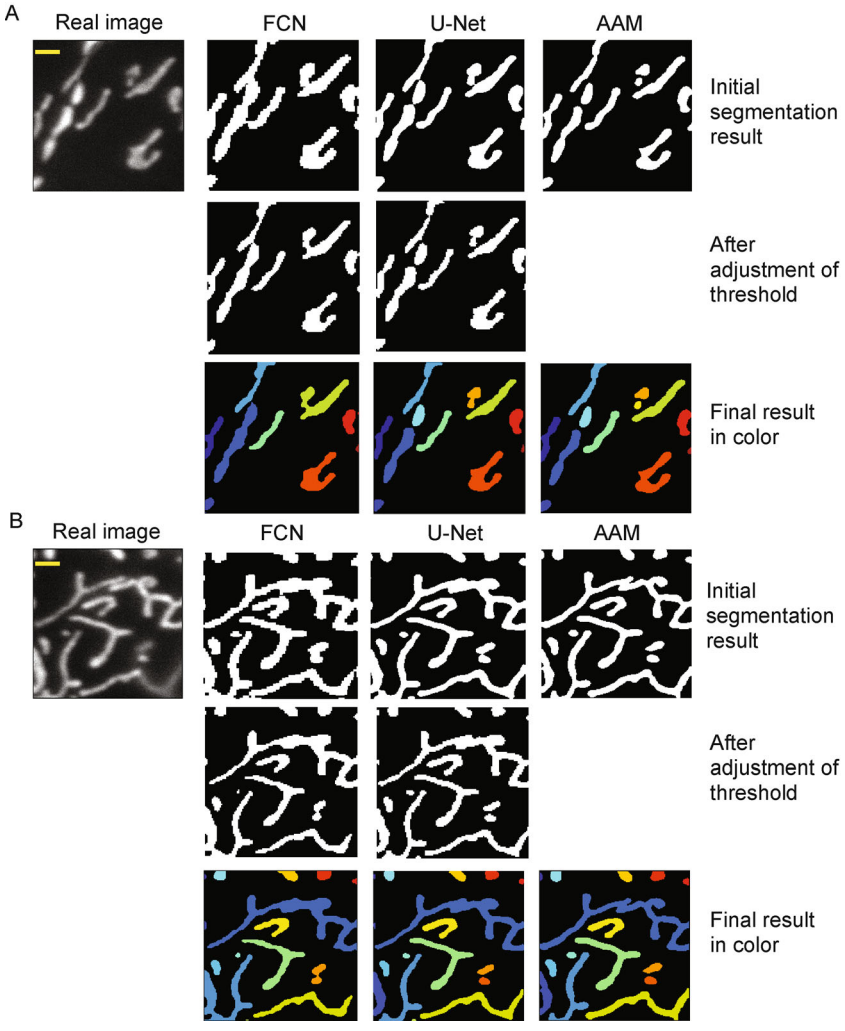

Figure 5. Comparison of segmentation results of real images using FCN, U-Net, and AAM. (A) An example image of mitochondria and the segmentation results using the three methods. Over-segmentation of the FCN and the U-Net (top row) was reduced by adjustment of threshold (middle row). Colors in the bottom row were selected randomly to indicate different mitochondria. (B) Another example, same layout as in (A). (A and B) Scale bars: 2 μm.

results showed that although the FCN and the U-Net provided generally comparable segmentation performance as the AAM on real images, they suffered from different levels of over-segmentation, which could be reduced through post-processing using conventional segmentation techniques such as global thresholding or the level-set.

Examining influence of synthetic image generation on performance of FCN and U-Net

That the FCN and the U-Net suffered from different levels of over-segmentation raised the question of what factors contributed to the over-segmentation. So far, we trained the FCN and the U-Net using synthetic images in which intensities of individual pixels within mitochondria were set independently (Materials & Methods). Although the smooth filtering by the microscope point spread function introduced some spatial correlation between intensities of neighboring pixels, the synthetic images generated in this way still could not fully recapitulate the spatial intensity patterns of mitochondria in real images (Supplementary Figure S5A). This was because intensities of pixels within real mitochondria had stronger spatial correlation and exhibited spatial patterns under the continuity of mitochondrial structures on short scales. To examine the influence of synthetic image generation on the performance of the FCN and the U-Net, we generated another group of synthetic images in which we started with assigning pixel intensities independently as before but increased spatial correlation between intensities of neighboring pixels by smoothing with a Gaussian kernel

with $\sigma = 1$ pixel only within the mitochondrial boundaries. The pixel intensity patterns generated with this additional kernel smoothing appeared to better recapitulate those of actual mitochondria (Supplementary Figure S5B). We then trained the FCN and the U-Net using the new synthetic images and characterized their performance under different levels of added space-invariant blurring (Supplementary Figure S5C–S5H). Overall, the FCN and the U-Net trained with the new synthetic data maintained mostly the same area similarity (Supplementary Figure S5C–S5E) and slightly degraded boundary distance (Supplementary Figure S5F–S5H). In contrast, the AAM showed significantly degraded area similarity and, under strong space-invariant blurring, significantly degraded boundary distance when tested on the new synthetic data (Supplementary Figure S5C–S5H). Taken together, the results showed that although increasing the spatial correlation between intensities of neighboring pixels in synthetic images better recapitulated the spatial intensity patterns of real mitochondria, it did not lead to an improvement in the segmentation performance of the FCN and the U-Net. This was not entirely surprising because although the added smoothing of the Gaussian kernel was restricted to within mitochondria, it could still be seen as an addition to the level of blurring, which was known to cause degradation in both area similarity and boundary distance (Figures 2 and 4). Further studies are required to fully elucidate the influence of synthetic image generation on the performance of the CNNs.

DISCUSSION

In this study, we have developed a method that uses realistic synthetic images to characterize the robustness and sensitivity of convolutional neural networks in segmentation of mitochondria from fluorescence microscopy images. Using this method, we show superior robustness and sensitivity of the FCN and the U-Net for quantitative analysis of mitochondrial morphology. In particular, when trained with a mixture of synthetic images of different levels of SNRs and blurring, the networks were able to accurately segment synthetic images under all the different conditions (Figure 4). Despite the remarkable success of CNNs in challenging image segmentation tasks in different disciplines, our understanding of the mechanisms and theoretical foundations underlying their successes remains limited. By characterizing the two CNNs in terms of their robustness to noise and blurring as well as their sensitivity to small deformation of mitochondria, our study provides insights into the behavior and performance of such networks. Our results not only demonstrate that CNNs provide superior segmentation performance but also show that different network architectures provide differential levels of accuracy, robustness and sensitivity. In particular, because of its multiple cross-layer connections and deeper architecture, the U-Net consistently outperforms the FCN.

Our study also provides insights into the how training of the two CNNs affects their segmentation performance. Specifically, we show that when the FCN and the U-Net are used to segment images of SNRs that do not match what they are trained for, the mismatch can lead to striking degradation in segmentation performance. However, the degradation can be reduced by proper selection of the training condition. We also show that synthetic images with increased spatial correlation better recapitulate the spatial intensity patterns of mitochondria in real images (Supplementary Figure S5A and S5B). However, they can lead to degradation in segmentation performance when used to train the two CNNs. The relation between training and segmentation performance of CNNs remains to be fully elucidated by follow-up studies.

The method we have developed in this study is general and can be used to characterize robustness and sensitivity of different CNNs in different image segmentation applications. In particular, the strategy of utilizing realistic synthetic images for characterizing segmentation performance has two important advantages. First, it makes it possible to exclude the uncertainty of manual segmentation from subsequent characterization of robustness and sensitivity of the networks. Second, it provides direct and flexible control of image conditions, specifically the SNR and the level of blurring in this study. Despite these advantages, a major limitation of our method remains that the synthetic images cannot fully recapitulate real image conditions. In particular, even with increased spatial correlation, the synthetic image still could not fully recapitulate the spatial intensity distribution of real mitochondria, especially at their boundaries. Further studies are required to determine whether this limitation contributes to the over-segmentation of the FCN and the U-Net on real images.

Our study also provides insights into the relation between CNNs and conventional task-specific segmentation algorithms such as the AAM. Because image segmentation is a basic operation in computer vision, a wide variety of task-specific algorithms have been developed. This raises the question regarding the relation between the conventional task-specific segmentation algorithms and CNNs. In this study, we show that the FCN and the U-Net both outperform the AAM in accuracy, robustness and sensitivity by a significant margin under a low SNR or strong blurring. However, the AAM provides performance comparable to the FCN and the U-Net under a medium SNR and moderate blurring. An important advantage of conventional task-specific algorithms such as AAM is that they have a small

number of parameters, making it possible to understand the relation between such parameters and their performance intuitively. In comparison, CNNs often have hundreds of thousands of parameters. Overall, we believe that to combine CNNs with the conventional algorithms such as the AAM is an effective strategy. In this study, we have demonstrated that integration of the FCN and the U-Net with global thresholding or a level-set is an effective strategy to reduce their over-segmentation. Other studies have also demonstrated the effectiveness of including non-network based algorithms into segmentation post-processing [20]. Lastly, follow-up studies are required to determine whether it is feasible to resolve the over-segmentation of the two CNNs using strategies such as customizing the training data and/or modification of the network loss function.

MATERIALS AND METHODS

Collection of experimental images

COS-7 cells were maintained in Dulbecco's modifies Eagle's medium supplemented with 10% fetal bovine serum. Cells, media and serum were purchased from American Type Culture Collection. Mitochondria labeling was performed by transfecting cells with pDsRed2-mito plasmids (Clontech) using the Neon electroporation system (Invitrogen) to express a mitochondrial targeting sequence fused with fluorescent protein DsRed. Following transfection, cells were seeded at 2×10^5 per 3.5 cm in MatTek dishes (MatTek) and incubated for 24–28 h before imaging. Time-lapse movies of mitochondria in live cells were collected on a Nikon Eclipse Ti-E inverted microscope with a CoolSNAP HQ2 camera (Photometric) and a 100×/1.40 NA oil objective lens. The effective pixel size was 0.0645 μm. DsRed-mito was imaged using a TRITC filter set. For live cell imaging, cells were maintained in a Tokai Hit stage incubator at 37°C with 5% CO_2. Time-lapse movies were collected at 5 s per frame. To collect static images of mitochondria, cells labeled with DsRed-mito were fixed 24–28 h post transfection with 4% paraformaldehyde (Sigma) for 10 min and rinsed with PBS (Invitrogen). Cells were then permeabilized with 0.2% Triton X-100 (Sigma) for 5 min, followed by rinsing with PBS. Images was collected under the same microscope setting as for live cell imaging.

Generation of realistic synthetic images

The overall workflow of generating realistic synthetic images to characterize robustness and sensitivity of CNNs is summarized in Supplementary Figure S1. A total of 16 raw images of various sizes were selected from several movies. The images were then segmented manually. Specifically, boundaries of mitochondria were traced using the overlay brush function in ImageJ [21]. Regions enclosed by the traced boundaries were then converted into binary masks, which were used as ground truth in this study. The binary mask images were manually cropped into 105 images of size 200×200. Data augmentation by two flippings (up-down, left-right) and three rotations (90°, 180°, 270°) was used to increase the total number of synthetic images to 630. The intensity distribution of mitochondria from real images was sampled and modeled as a Gamma distribution. Intensities sampled from the Gamma distribution were assigned to each pixel within the binary masks to generate synthetic images without background noise. These synthetic images were then blurred at different levels to simulate defocusing of mitochondria by convolution with a kernel function $K(x,y)$. Specifically, to simulate space-variant blurring, we used the following Gaussian kernel [18]:

$$K(x,y)=A\exp\left(-\frac{1}{2\sigma^2(x,y)}(x^2+y^2)\right), \quad (1)$$

in which A is a normalization constant, $\sigma^2(x,y)$ controls the level of blurring at location (x,y). Space-invariant blurring was simulated by keeping $\sigma^2(x,y)$ constant. $K(x,y)$ was set to be the microscope point spread function if no additional blurring was simulated, in which $\sigma(x,y)$ was set to 1 according to the point spread function of the microscope setting in this study. To simulate space-variant blurring of level 1~M, each binary mask image was divided from the top to the bottom into 4 horizontal bands of equal sizes [18]. Blurring was then applied at the level of $\sigma(x,y)=1,\frac{M}{3},\frac{2M}{3},M$ for the four bands, respectively.

Besides blurring, another image condition simulated was noise at different SNRs. Specifically, additive Gaussian noise $N(\mu_{bg},\sigma_{bg}^2)$ was used so that the signal-to-noise (SNR), defined as $SNR=(\mu_{mito}-\mu_{bg})/\sigma_{bg}$, reached the designed level. Here μ_{mito} is the mean of the mitochondria intensity, μ_{bg} and σ_{bg} are the mean and standard deviation of the background noise, respectively. Putting together blurring and noise addition, the simulated synthetic image I^* generated from the ground-truth image I is defined by

$$I^*(x,y)=K(x,y)\otimes I(x,y)+N(\mu_{bg},\sigma_{bg}^2). \quad (2)$$

Lastly, for simplification, changes of mitochondrial morphology were simulated using image morphology operations including uniform dilation and erosion.

Architectural design of FCN

A FCN was constructed with minor modifications of the

architecture described in Ref. [16]. Specifically, the network consists of 9 convolutional layers with kernels of size 3×3, 3 max pooling layers, and 2 up-sampling layers from the second max pooling layer and the last 3×3 convolutional layer, respectively. Outputs from the two up-sampling layers were fused together. This combination of fine layer and coarse layer is helpful for the network to perform finer scale segmentation [17]. The last layer is a pixel-wise softmax layer that generates the network output. The loss function of this network is the sum of binary pixel-wise cross-entropy between the network output and the ground-truth. The softmax and loss functions are defined as:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}}, \qquad i = 0,1 \tag{3}$$

$$L = -\sum_x \log\sigma(x_1), \tag{4}$$

where x is the network output of a specific pixel, index 0 indicates the pixel is in the background class, while index 1 indicates the pixel is in the object class.

Architectural design of U-Net

A U-Net was constructed with minor modifications of the architecture described in Ref. [17]. The network consists of 14 convolutional layers with kernels of size 3×3. The last convolutional layer has kernels of size 2×2. The structure features symmetric max pooling layers along its contracting path and up-convolutional layers along its expanding path. In addition, the numbers of channels are doubled in every down-sampling step, and halved in every up-sampling step. In this U-shaped structure, three concatenations are made between corresponding expanding path and contracting path. They have been shown to be very important to generate detailed segmentation [17]. The last layer is a pixel-wise softmax layer, similar to the one in FCN to generate the probability map of individual pixels. Again, pixel-wise binary cross-entropy was used as the loss function.

Implementation and training of FCN and U-Net

FCN and U-Net were implemented mostly in Keras (https://github.com/keras-team/keras) and partially in Python on top of Tensorflow [22]. For each image condition, defined by the level of blurring and the SNR, 630 synthetic images were generated. Among them, 500 images were used for training, 100 images were used for validation, and the remaining 30 were used for testing. The CNNs were trained with stochastic gradient descent with a learning rate of 0.01. Training and testing of the two networks were performed using a desktop workstation with 2× Intel Xeon E5503 2.00 GHz CPU, 1× GeForce GTX 1080 GPU, and 32 GB RAM. Images and source codes used to generate all the results in this paper can be downloaded from https://github.com/ccdlcmu/mitosegmentation.

Statistics

Unless specified otherwise, statistical comparisons were made using two-sample student's t-test. Error bars in figures indicate S.E.M. (standard error of the mean).

Characterizing robustness of FCN, U-Net, and AAM on synthetic images

Two different training and testing schemes were used in this study. Under the first scheme, for each simulated image condition, defined by the SNR and the level of blurring, the FCN and the U-Net were trained using the realistic synthetic images generated for that condition. Then they were tested on synthetic images of the same condition (Figures 2 and 3). Under the second scheme, the FCN and the U-Net were trained for multiple conditions using an equal mixture of synthetic images generated for those conditions. Then they were tested for those conditions individually (Figure 4).

The segmentation accuracy was quantified using two metrics. The first metric was area similarity (*AS*) [23], defined by

$$AS(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{5}$$

in which A denotes the ground truth and B denotes the segmentation result. The second metric was average boundary distance (*ABD*), also referred to as boundary distance in this paper for simplification [23]. Let $S(A)$ denote the boundary of ground-truth, and $d(p,S(A))$ denote the shortest distance from a pixel p to $S(A)$

$$d(p,S(A)) = \min_{s_A \in S(A)} \|p - s_A\|, \tag{6}$$

the average symmetric boundary distance is defined by

$$ABD(A,B) = \frac{1}{|S(A)| + |S(B)|} \cdot \left(\sum_{s_A \in S(A)} d(s_A,S(B)) + \sum_{s_B \in S(B)} d(s_B,S(A)) \right). \tag{7}$$

Using these metrics, the robustness of the two networks was tested against different SNRs and different levels of blurring. Furthermore, the two networks were compared

against the adaptive active-mask (AAM) algorithm [5], which was selected as a representative of high-performance segmentation algorithms not based on convolutional neural networks. The MATLAB source code for AAM can be downloaded at https://github.com/ccdlcmu/image-segmentation.

Characterizing sensitivity of FCN, U-Net, and AAM on synthetic images

To quantify the sensitivity of the three methods in detecting deformation of mitochondrial morphology, we used two metrics. The first metric, which we refer to as deformation area similarity (*DAS*), quantifies the area similarity between the change to ground-truth after erosion/dilation and the change to segmentation results after erosion/dilation. It is defined as

$$DAS(A,A_D,B,B_D)=\frac{|(A-A_D)\cap(B-B_D)|}{|(A-A_D)\cup(B-B_D)|}, \quad (8)$$

where A and A_D denote the ground truth before and after deformation simulated by erosion or dilation, B and B_D denotes the pair of segmentation results corresponding to A and A_D, respectively. The second metric, which we refer to as deformation distance similarity (*DDS*), is defined as

$$DDS(A,A_D,B,B_D)=\frac{ABD(B,B_D)}{ABD(A,A_D)}, \quad (9)$$

where $ABD(A,A_D)$ denotes the average boundary distance between A and A_D, and $ABD(B,B_D)$ denotes the average boundary distance between B and B_D. Essentially, *DDS* is the ratio between the detected boundary displacement versus the actual boundary displacement in the ground truth.

Characterizing performance of FCN, U-Net, and AAM on real images

Performance of the two networks and the active-mask algorithm qualitatively was tested using real mitochondrial images. Because the ground-truth was not available, the performance characterization was limited to qualitative visual comparison.

Segmentation post-processing using the level set

From the binary segmentation result images of the FCN or the U-Net, the boundary of each region with 1 pixel in thickness was extracted. All extracted boundaries were then set as the initial condition for the level-set iteration using the parameters recommended in Ref. [19].

ABBREVIATIONS

AAM	adaptive active-mask
CNN	convolutional neural network
DAS	deformation area similarity
DDS	deformation distance similarity
FCN	fully convolutional network
SNR	signal-to-noise ratio

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.1007/s40484-018-0156-3.

ACKNOWLEDGEMENTS

Xiaoqi Chai acknowledges support of a Ji-Dian Liang Graduate Research Fellowship. Qinle Ba acknowledges support of a Bertucci Graduate Research Fellowship. Ge Yang acknowledges support of NSF CAREER grant DBI-1149494 and NSF grant CBET-1804929. The authors would also like to thank Yile Feng and Weicheng Lin for their technical assistance.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xiaoqi Chai, Qinle Ba and Ge Yang declare that they have no conflict of interests.

All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

REFERENCES

1. McBride, H. M., Neuspiel, M. and Wasiak, S. (2006) Mitochondria: more than just a powerhouse. Curr. Biol., 16, R551–R560
2. Nunnari, J. and Suomalainen, A. (2012) Mitochondria: in sickness and in health. Cell, 148, 1145–1159
3. Karbowski, M. and Youle, R. J. (2003) Dynamics of mitochondrial morphology in healthy cells and during apoptosis. Cell Death Differ., 10, 870–880
4. Campello, S. and Scorrano, L. (2010) Mitochondrial shape changes: orchestrating cell pathophysiology. EMBO Rep., 11, 678–684
5. Chen, K. C. J., Yu, Y. Y., Li, R. Q., Lee, H. C., Yang, G. and Kovacevic, J. (2012) Adaptive active-mask image segmentation for quantitative characterization of mitochondrial morphology. In Proceedings of 2012 IEEE International Conference on Image Processing (ICIP), pp. 2033–2036
6. Leonard, A. P., Cameron, R. B., Speiser, J. L., Wolf, B. J., Peterson, Y. K., Schnellmann, R. G., Beeson, C. C. and Rohrer, B. (2015) Quantitative analysis of mitochondrial morphology and membrane potential in living cells using high-content imaging, machine learning, and morphological binning. Biochim. Biophys. Acta, 1853, 348–360

7. Peng, J.-Y., Lin, C.-C., Chen, Y.-J., Kao, L.-S., Liu, Y.-C., Chou, C.-C., Huang, Y.-H., Chang, F.-R., Wu, Y.-C., Tsai, Y.-S., *et al.* (2011) Automatic morphological subtyping reveals new roles of caspases in mitochondrial dynamics. PLoS Comput. Biol., 7, e1002212
8. Iannetti, E. F., Smeitink, J. A. M., Beyrath, J., Willems, P. H. G. M. and Koopman, W. J. H. (2016) Multiplexed high-content analysis of mitochondrial morphofunction using live-cell microscopy. Nat. Protoc., 11, 1693–1710
9. Daniele, J. R., Esping, D. J., Garcia, G., Parsons, L. S., Arriaga, E. A. and Dillin, A. (2017) High-throughput characterization of region-specific mitochondrial function and morphology. Sci. Rep., 7, 6749
10. Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems. pp. 1097–1105
11. LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. Nature, 521, 436–444
12. Sadanandan, S. K., Ranefall, P., Le Guyader, S. and Wählby, C. (2017) Automated training of deep convolutional neural networks for cell segmentation. Sci. Rep., 7, 7860
13. Kraus, O. Z., Ba, J. L. and Frey, B. J. (2016) Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics, 32, i52–i59
14. Xing, F., Xie, Y., Su, H., Liu, F. and Yang, L. (2017) Deep learning in microscopy image analysis: a survey. In IEEE Transactions on Neural Networks and Learning Systems. pp. 1–19
15. Yu, Y., Lee, H.-C., Chen, K.-C., Suhan, J., Qiu, M., Ba, Q. and Yang, G. (2016) Inner membrane fusion mediates spatial distribution of axonal mitochondria. Sci. Rep., 6, 18981
16. Long, J., Shelhamer, E. and Darrell, T. (2015) Fully convolutional networks for semantic segmentation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440
17. Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention. 9351, 234–241
18. Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H. and Yli-Harja, O. (2007) Computational framework for simulating fluorescence microscope images with cell populations. IEEE Trans. Med. Imaging, 26, 1010–1016
19. Li, C., Huang, R., Ding, Z., Gatenby, J. C., Metaxas, D. N. and Gore, J. C. (2011) A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. IEEE Trans. Image Process., 20, 2007–2016
20. Ngo, T. A., Lu, Z. and Carneiro, G. (2017) Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Med. Image Anal., 35, 159–171
21. Schneider, C. A., Rasband, W. S. and Eliceiri, K. W. (2012) NIH Image to ImageJ: 25 years of image analysis. Nat. Methods, 9, 671–675
22. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., *et al.* (2016) TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. pp. 265–283
23. Heimann, T., van Ginneken, B., Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., *et al.* (2009) Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans. Med. Imaging, 28, 1251–1265

