

QUALITY ASSESSMENT OF SYNTHETIC FLUORESCENCE MICROSCOPY IMAGES FOR IMAGE SEGMENTATION

Yile Feng^{1†}, Xiaoqi Chai^{1†}, Qinle Ba¹, and Ge Yang^{1,2,3,4}

¹Department of Biomedical Engineering; ²Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA;

³National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; ⁴School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

†: Authors contributed equally

ABSTRACT

Synthetic images are widely used in image segmentation for algorithm training and performance assessment. Recently, advances in image synthesis techniques, especially generative adversarial networks (GANs), have made it possible to generate fluorescence microscopy images with remarkably realistic appearance. However, intuitive and specific metrics to assess the quality of these images remain lacking. Here, we propose three quality metrics that quantify the fidelity of the foreground signal, the background noise, and blurring, respectively, of synthesized fluorescence microscopy images. Using these metrics, we examine images of mitochondria synthesized by two representative GANs: pix2pix, which requires paired training data, and CycleGAN, which does not require paired training data. We find that both networks generate realistic images and achieve similar fidelity in reproducing background noise and blurring of real images. However, CycleGAN achieves significantly higher fidelity than pix2pix in reproducing intensity patterns of real mitochondria. When used to train the U-Net for segmentation, images synthesized by both networks achieve performance on par with real images. Overall, we have developed a method to assess quality of synthetic fluorescence microscopy images and to evaluate their training performance in image segmentation. The quality metrics proposed are general and can be used to assess fluorescence microscopy images synthesized by different methods.

Index Terms— Quality assessment, synthetic image, fluorescence microscopy, generative adversarial network, image segmentation

1. INTRODUCTION

Synthetic images are used extensively in computational analysis of biological images in applications such as image segmentation [1] and feature tracking [2]. In image segmentation, synthetic images are often used for training algorithms because their ground truth is known and therefore, requires no manual labelling. Another advantage of synthetic images is that their conditions such as signal-to-noise ratios (SNRs) often can be controlled [3,

4]. Despite these advantages, performance of synthetic images is fundamentally defined by the level of fidelity they achieve in mimicking real images.

In synthesizing fluorescence microscopy images, traditional methods often rely on explicit extraction of conditions such as SNRs and blurring of real images and matching such conditions in synthesized images [3, 4]. However, physical and chemical properties of fluorescence image formation pose unique challenges to this strategy of image synthesis. In particular, the high numerical aperture required for high-resolution fluorescence microscopy also leads to small depth-of-field [5]. Consequently, objects or regions within images may become partially or completely blurred, resulting in weakened boundaries of varying levels of diffusiveness (see e.g. Fig. 1).

Furthermore, minimizing photobleaching and/or phototoxicity often results in low and varying signal-to-noise ratios (SNRs) [6], which also complicate image conditions. Variations in image formation conditions, especially in live cell imaging, may further complicate image conditions. Overall, complex conditions of real fluorescence microscopy images make it very difficult, if not impossible, to fully emulate real images through the strategy of explicit extraction and matching of image conditions.

Recently, deep learning networks based image synthesis techniques, especially generative adversarial networks (GANs) [7], have succeeded in generating synthetic fluorescence microscopy images with remarkably realistic appearance [8, 9]. GANs consist of two competing subnetworks: a generator and a discriminator [7]. The generator aims to synthesize images as similar to real images as possible, while the discriminator aims to differentiate between synthesized images and real images as

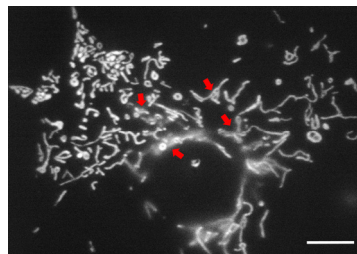


Figure 1. Mitochondria in a COS-7 cell. Arrows point to examples of blurred regions. Scale bar: 10 μm .

much as possible. Because these networks learn and match image conditions implicitly in image synthesis, they avoid the limitations of explicit image condition extraction and matching. However, although images synthesized by these network bear strong visual resemblance to real fluorescence images, our understanding of their properties and performance remains limited. In particular, current evaluation measures for GANs [10] are developed primarily for images of natural and social scenes and do not take into account the distinct properties of fluorescence microscopy images. Furthermore, as we will show later in the paper, traditional image quality metrics such as SSIM (structural similarity index measure) and PSNR (peak-signal-to-noise-ratio) [11, 12] lack the sensitivity in differentiating between synthesized images. So far, intuitive and specific metrics to assess quality of synthetic fluorescence microscopy images remain lacking.

In this study, we address this deficiency by proposing a group of three quality metrics, which take into account the distinct properties of fluorescence image formation and quantify the fidelity of the foreground signal, the background noise, and blurring, respectively, of synthetic images. Using these metrics, we examine images of mitochondria synthesized using two representative GANs: pix2pix [13], which requires paired training data, and CycleGAN [14], which does not require paired training data. To further assess the fidelity of the synthetic images, we test their performance in training convolutional neural networks (CNNs) for segmentation of real fluorescence microscopy images. Because performance of CNNs in image segmentation depends critically on the quality of their training images [4], testing synthesized images in training CNNs for image segmentation provides an integrative and stringent approach to assess their fidelity.

Figure 2 summarizes the overall workflow of our study. We manually extracted geometry of mitochondria of real images into binary masks and then filled in these masks with realistic spatial intensity patterns using GANs. We then calculated the quality metrics using different pairs of real (i.e. reference) and synthetic images. Furthermore, we trained the U-Net [15] using real and synthetic images, respectively, and compared their accuracy in segmentation of real images. Overall, our study makes two main research First, it develops intuitive, specific, and general-purpose metrics for assessing quality of synthetic fluorescence microscopy images. Such metrics make it possible not only to differentiate between visually realistic

images synthesized by different GANs but also to evaluate quality of images synthesized by different methods. Second, our study reveals distinct properties of images synthesized by different GANs, especially their training performance for segmentation of real fluorescence microscopy images.

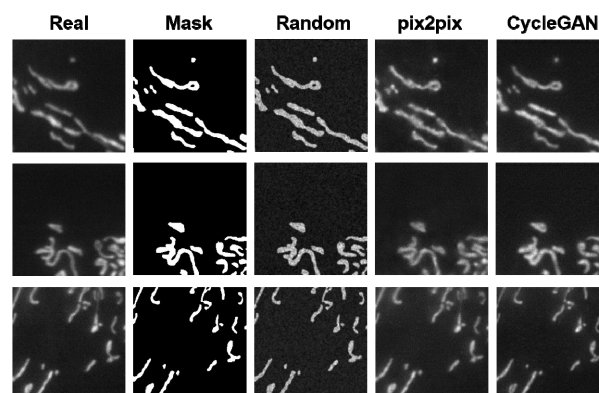


Figure 3. Comparison of real and different synthesized images. Mask: binary mask images. Random: pixel intensity in binary masks assigned randomly from the ensemble mitochondrial intensity distribution of real images.

2. METHODS

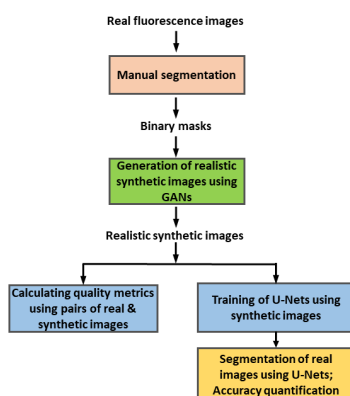
2.1 Image collection and mask generation

Mitochondria are essential organelles of eukaryotic cells, serving a wide variety of important cellular functions [16]. Within the intracellular space, they exhibit complex and dynamic morphology. Quantitative characterization of morphology of mitochondria based on image segmentation is important to elucidating their physiology and pathophysiology [17]. Raw images of mitochondria (Fig. 1) were collected from cultured COS-7 cells transfected with pDsRed2-mito plasmids (Clontech) to express a mitochondrial targeting sequence fused with fluorescent protein DsRed. Image collection was performed on a Nikon Eclipse Ti-E inverted microscope with a CoolSNAP HQ2 camera (Photometric) and a 100×/1.40NA oil objective lens. The effective pixel size was 0.0645 μm . Manual labeling of individual mitochondria (see e.g. Figure 3, second column) was performed using ITK-SNAP (<http://www.itksnap.org/pmwiki/pmwiki.php>).

2.2 Assessing image quality using traditional metrics

Figure 3 shows representative results of the image synthesis workflow in Figure 2. Specifically, the second column shows the binary masks from manual labeling, while the third column shows images synthesized by filling each pixel within the binary masks by an intensity drawn randomly from the ensemble image intensity distribution of real mitochondria, as described in further details in [4]. For simplicity, we label images synthesized by this strategy as “random”. Images synthesized by pix2pix and CycleGAN are shown in the fourth and fifth columns, respectively. We refer to the similarity of a synthesized image to its

Figure 2. Overview of workflow



corresponding real image as the fidelity of the synthesized image. To characterize and compare fidelity of the images synthesized by different methods, we first used traditional metrics, specifically SSIM, PSNR, and NCC (normalized cross correlation) [11, 12]. The result is shown in Table 1. Because differences between images synthesized by random intensity assignment (column 3) and by GANs (columns 4 & 5) are visually clear, the result indicates that traditional metrics lack the specificity and sensitivity in differentiating between the two groups of synthetic images.

Table 1. Selected quality metrics of synthetic images

	SSIM		PSNR		NCC	
	MEA*	STD*	MEA	STD	MEA	STD
Random	0.991	0.009	57.82	3.806	0.980	0.009
pix2pix	0.996	0.006	60.91	3.705	0.989	0.007
CycleGAN	0.993	0.008	58.65	3.571	0.983	0.009

*MEA: mean; STD: standard deviation. Each case was calculated from the same 78 pairs of real and synthetic images.

2.3 A quality metric for foreground signal

We reason that images synthesized by random intensity assignment and GANs may be better differentiated if we only compare the intensity distributions within mitochondria, i.e. within binary masks. To this end, we propose the signal fidelity measure (SFM), for which the normalized correlation coefficient is computed only within actual mitochondria defined by the binary masks, denoted as Ω . If we denote the real images as $I(x, y)$ and its corresponding synthetic images as $\hat{I}(x, y)$, SFM is defined as.

$$SFM(I, \hat{I}) = \frac{\int_{\Omega} I(x, y) \hat{I}(x, y) dx dy}{\sqrt{\int_{\Omega} I(x, y)^2 dx dy \int_{\Omega} \hat{I}(x, y)^2 dx dy}}, \quad (1)$$

where $\bar{I}$ and $\bar{\hat{I}}$ are the mean intensities of the real image and the synthetic image, respectively, within Ω .

2.4 A quality metric for background noise

We define the noise fidelity measure (NFM) as the Sorensen distance [18] between the probability density function (pdf) of the background noise in the real image $I(x, y)$ and the corresponding pdf of the background noise in the synthetic image $\hat{I}(x, y)$.

$$NFM(I, \hat{I}) = \frac{\int_0^L |pdf_{I_N}(x) - pdf_{\hat{I}_N}(x)| dx}{\int_0^L pdf_{I_N}(x) dx + \int_0^L pdf_{\hat{I}_N}(x) dx}, \quad (2)$$

where L denotes the maximum level of the background noise and pdf_{I_N} and $pdf_{\hat{I}_N}$ denote the pdf of background

noise in the real image I , and the synthetic image $\hat{I}$, respectively. The measure equals 0 when the two pdfs have complete overlap, i.e. are identical, and equals 1 when the two pdfs have no overlap.

2.5 A quality metric for blurring

We define the blurring fidelity measure (BFM) as the ratio between the total level of blurring in the synthetic image and the total level of blurring in the real image.

$$BFM(I, \hat{I}) = \frac{\|S(I(x, y))\|_{L^2}}{\|S(\hat{I}(x, y))\|_{L^2}}, \quad (3)$$

where S is an image sharpness operator. In this study, we used the Laplacian operator so that the BFM took the following specific form

$$BFM(I, \hat{I}) = \frac{\int_{\Omega} \left(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \right)^2 dx dy}{\int_{\Omega} \left(\frac{\partial^2 \hat{I}}{\partial x^2} + \frac{\partial^2 \hat{I}}{\partial y^2} \right)^2 dx dy}, \quad (4)$$

where the operator was implemented as $\begin{bmatrix} 1 & 4 & 1 \\ 4 & 20 & 4 \\ 1 & 4 & 1 \end{bmatrix}$ [19].

3. RESULTS

All the computation was performed on a workstation with two Intel Xeon E5503 CPU, 32GB memory, and one Nvidia GTX 1080 GPU card. Custom code can be downloaded from <https://github.com/XiaoqiChai/mitoGAN>. Implementation of pix2pix and CycleGAN at <https://github.com/tjwei/GANotebooks> was used.

3.1 Training of pix2pix, CycleGAN and U-Net

For both pix2pix and CycleGAN, a total of 500 images were used for training, 50 for validation, and 78 for testing, and the numbers of training epochs were 800 and 200, respectively. Adam optimizer [20] was used for training pix2pix, with a learning rate of 2×10^{-4} , $\beta_1 = 0.5$, and the weight μ for the regularizer equals 100. Adam optimizer with the same parameters was also used for training CycleGAN, with the weight of cycle consistency loss equals 10. For the U-Net, a total of 600 images were used for training, 50 for validation, and 78 for testing. The number of training epoch was 50. Stochastic gradient descent was used, with a learning rate of 0.01.

3.2 Comparison of foreground signal fidelity

Figure 4 compares the foreground signal quality in SFM of the images synthesized by random intensity assignment, pix2pix, and CycleGAN. Both GANs achieved significantly better foreground fidelity than random intensity assignment, with CycleGAN significantly outperformed pix2pix. This may be explained by the fact

that CycleGAN is trained with all mitochondrial intensity distributions while pix2pix is trained only with paired data. However, none of the three methods achieved a mean SFM greater than 0.5, a limitation to be overcome by further improvement in GANs.

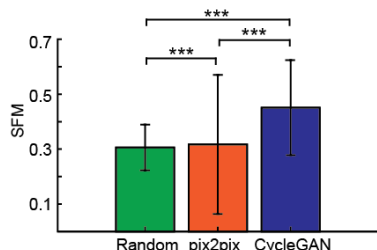


Figure 4. Comparison of foreground fidelity of synthesized images. Error bars indicate standard deviation. Comparison was made using Kolmogorov-Smirnov tests. **, p-value < 0.01, *** p-value < 0.001. Sample size for each group is 78.

3.3 Comparison of background noise fidelity

Figure 5A compares the background noise fidelity in NFM of the images synthesized by the three methods. There was no significant difference between the background noise fidelity of pix2pix and CycleGAN. However, both significantly outperformed the “random” images, whose background was generated as Gaussian white noise [4]. Further analysis revealed that background synthesized by GANs exhibits weak patterned noise (results not shown due to space limitation).

3.4 Comparison of blurring fidelity

Figure 5B compares the blurring fidelity in BFM of the synthetic images. There was no significant difference between CycleGAN and pix2pix. However, both significantly outperformed random intensity assignment.

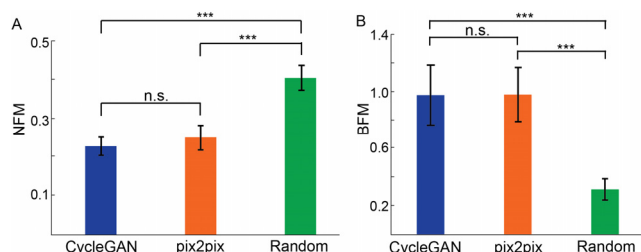


Figure 5. Comparison of background and blurring fidelity of synthesized images. (A) Background fidelity. Error bars indicate standard error of the mean (SEM). (B) Blurring fidelity. Error bars indicate standard deviation. (A-B) Comparison was made using two-sample student's t-test. Same p-value notations as in Fig. 4. n.s.: not significant. Sample size: (A) 33, (B) 78.

3.5 Performance of synthesized images in training U-Nets for segmentation of real mitochondria images

To further compare the quality of the synthetic images, we used them to train U-Nets and then used the trained U-Nets to segmentation real images of mitochondria (Fig. 6). The segmentation performance is evaluated by intersection over union (IoU). For reference, we also trained U-Nets on segmentation images (the “real” group in Fig. 7). We used manual segmentation results as our ground truth. The results (Fig. 7) show that images synthesized by pix2pix and CycleGAN significantly outperformed images

with random intensity patterns and were on par with manually segmented real images in training performance.

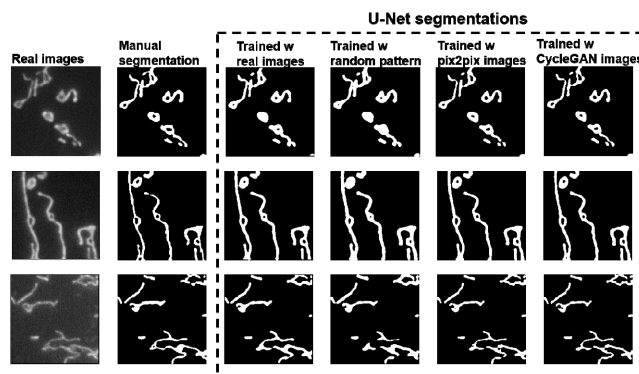


Figure 6. Segmentation results after training with synthesized images.

4. DISCUSSION

Fluorescence microscopy images have at least two properties that distinguish them from natural and social images. First, because of specificity of fluorescence signal generation, fluorescence images often include substantial fractions of areas filled with noise and/or defocused background because of absence of fluorophores. To synthesize high fidelity images, especially those with low SNRs, high fidelity in synthesized image background is essential. Second, blurring due to defocusing is inherent to fluorescence image formation. To achieve high fidelity in mimicking blurring in synthetic images is also essential. Motivated by these distinct properties of fluorescence microscopy images, we propose quality metrics that specifically assess the fidelity of their background noise and blurring. These metrics can better differentiate synthesized images than traditional metrics such as SSIM and PSNR (compare Fig. 4&5 vs Table 1). However, our study has its limitations, which we are addressing in ongoing research. First, our analysis of blurring fidelity only quantifies the total level of blurring, not its spatial distribution. Second, our metrics may leave out certain aspects of the synthetic images. For example, the PSNR in Table 1 shows that pix2pix outperforms CycleGAN. Third, our metrics require binary masks, i.e. segmentation of reference images.

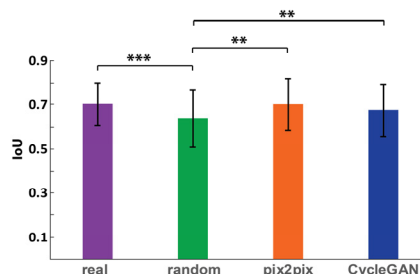


Figure 7. Comparison of training performance of synthetic vs real images. Error bars: standard deviation. Comparison was made using two-sample student's t-test. Sample size = 78 for each group.

ACKNOWLEDGEMENT

This study was supported in part by US National Science Foundation Grant CBET-1804929.

REFERENCES

- [1] E. Meijering, "Cell segmentation: 50 years down the road," *IEEE Signal Processing Magazine*, vol. 29, pp. 140-145, 2012.
- [2] N. Chenouard, I. Smal, F. de Chaumont, M. Maška, I. F. Sbalzarini, Y. Gong, *et al.*, "Objective comparison of particle tracking methods," *Nature Methods*, vol. 11, pp. 281-289, 2014.
- [3] A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja, "Computational framework for simulating fluorescence microscope images with cell populations," *IEEE Transactions on Medical Imaging*, vol. 26, pp. 1010-1016, 2007.
- [4] X. Chai, Q. Ba, and G. Yang, "Characterizing robustness and sensitivity of convolutional neural networks for quantitative analysis of mitochondrial morphology," *Quant. Biol.*, vol. 6, pp. 344-358, 2018.
- [5] S. Inoue and K. R. Spring, *Video Microscopy: The Fundamentals*, 2nd ed.: Springer, 2012.
- [6] J. W. Lichtman and J.-A. Conchello, "Fluorescence microscopy," *Nature Methods*, vol. 2, pp. 910-919, 2005.
- [7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al.*, "Generative adversarial nets," *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pp. 2672-2680, 2014.
- [8] C. Fu, S. Lee, D. J. Ho, S. Han, P. Salama, K. W. Dunn, *et al.*, "Three dimensional fluorescence microscopy image synthesis and segmentation," *Proc. 2018 IEEE International Conference on Computer Vision and Pattern Recognition, Workshop on Computer Vision for Microscopy Image Analysis*, 2018.
- [9] A. Osokin, A. Chessel, R. E. C. Salas, and F. Vaggi, "GANs for biological image synthesis," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2252-2261.
- [10] A. Borji, "Pros and cons of GAN evaluation measures," *arXiv preprint*, p. arXiv:1802.03446, 2018.
- [11] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366-2369.
- [12] D. Kundu, L. K. Choi, A. C. Bovik, and B. L. Evans, "Perceptual quality evaluation of synthetic pictures distorted by compression and transmission," *Signal Processing: Image Communication*, vol. 61, pp. 54-72, 2018.
- [13] J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967-5976, 2017.
- [14] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242-2251, 2017.
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention, Pt Iii*, vol. 9351, pp. 234-241, 2015.
- [16] H. M. McBride, M. Neuspiel, and S. Wasiak, "Mitochondria: more than just a powerhouse," *Current Biology*, vol. 16, pp. R551-R560.
- [17] S. Campello and L. Scorrano, "Mitochondrial shape changes: orchestrating cell pathophysiology," *EMBO Reports*, vol. 11, pp. 678-684, 2010.
- [18] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, pp. 300-307, 2007.
- [19] Y. Yao, B. Abidi, N. Doggaz, and M. Abidi, "Evaluation of sharpness measures and search algorithms for the auto-focusing of high-magnification images," in *Proc. SPIE 6246, Visual Information Processing XV*, 2006.
- [20] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. International Conference on Learning Representations*, 2014.