

PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Automated chondrocyte viability analysis of articular cartilage based on deep learning segmentation and classification of two-photon microscopic images

Fan, Hongming, Xu, Pei, Le, Michael, Hsu, Jennifer, Chen, Xun, et al.

Hongming Fan, Pei Xu, Michael Le, Jennifer Hsu, Xun Chen, Yang Li, Zhao Zhang, Bruce Gao, Shane Woolf, Tong Ye, "Automated chondrocyte viability analysis of articular cartilage based on deep learning segmentation and classification of two-photon microscopic images," Proc. SPIE 11966, Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XXIX, 119660C (2 March 2022); doi: 10.1117/12.2609880

SPIE.

Event: SPIE BiOS, 2022, San Francisco, California, United States

Automated chondrocyte viability analysis of articular cartilage based on deep learning segmentation and classification of two-photon microscopic images

Hongming Fan^a, Pei Xu^b, Michael Le^a, Jennifer Hsu^{a,b}, Xun Chen^a, Yang Li^a, Zhao Zhang^a, Bruce Gao^a, Shane Woolf^c, Tong Ye^{*a,d}

^aDepartment of Bioengineering, Clemson University, Clemson, South Carolina, USA; ^bSchool of Computing, Clemson University, Clemson, South Carolina, USA; ^cDepartment of Orthopaedics & Physical Medicine, Medical University of South Carolina, Charleston, South Carolina, USA;

^dDepartment of Regenerative Medicine and Cell Biology, Medical University of South Carolina, Charleston, South Carolina, USA.

ABSTRACT

Chondrocyte viability is an important measure to consider when assessing cartilage health. Dye-based cell viability assays are not suitable for in vivo or long-term studies. We have introduced a non-labeling viability assay based on the assessment of high-resolution images of cells and collagen structure using two-photon stimulated autofluorescence and second harmonic generation microscopy. By either the visual or quantitative assessment, we were able to differentiate living from dead chondrocytes in those images. However, both techniques require human participation and have limited throughputs. Throughput can be increased by using methods for automated cell-based image processing. Due to the poor image contrast, traditional image processing methods are ineffective on autofluorescence images produced by nonlinear microscopes. In this work, we examined chondrocyte segmentation and classification using Mask R-CNN, a deep learning approach to implement automated viability analysis. It has been demonstrated an 85% accuracy in chondrocyte viability assessment with proper training. This study demonstrates that automated and highly accurate image analysis is achievable with the use of deep learning methods. This image processing approach can be helpful to other imaging applications in clinical medicine and biological research.

Keywords: nonlinear optical microscopy, autofluorescence, second harmonic generation, deep learning, viability

1. INTRODUCTION

As one of the most important inventions in the field of optical microscopy, two-photon fluorescence microscopy (TPFM)¹ has become an indispensable tool in biomedical research. Compared to confocal microscopy, a technology of equal importance, TPFM has its advantages mainly due to the nonlinearity of signals generated by ultrashort pulses generally in the near-infrared region. It is commonly accepted that TPFM can achieve a higher signal-to-noise ratio, similar spatial resolution, less photobleaching volume, and deeper penetration depth than confocal microscopy. These advantages make TPFM a more suitable technique for imaging thick or living tissues². Using almost the same hardware, TPFM can include second harmonic generation (SHG) imaging, a useful modality for imaging fibrillar collagens, myosin, and microtubules without dye labeling.

[*ye7@clemson.edu](mailto:ye7@clemson.edu); phone 843 876 2314; fax 843 792 6626

It is well known that two-photon excitation autofluorescence (TPAF) originated from measuring signals from reduced forms of nicotinamide adenine dinucleotide (NADH) or nicotinamide adenine dinucleotide phosphate (NADPH) and flavin proteins (FPs) to study cell metabolism³. Recently, our lab has developed a chondrocyte viability assay using TPAF and SHG imaging⁴. The chondrocyte is the only cell type that exists in the articular cartilage and is responsible for maintaining the homeostasis of the extracellular matrix (ECM). Chondrocyte death⁵ may lead to degradation of the ECM, causing osteoarthritis, which affects over 30.8 million people in the U.S alone and is a world-leading cause of disability. As such, chondrocyte viability, which is defined as a percentage of living cells over the total number of cells in a sample, is regarded as an important indicator in the assessment of cartilage health. We have demonstrated that live or dead cells exhibit different TPAF intensity in NAD(P)H and FPs channels; live cells show stronger signals in the NAD(P)H channel than dead cells do, while the dead cells show stronger signals in the FPs channel. We have also demonstrated that both visual and quantitative assessments yield over 90 percent in specificity and sensitivity. In the visual evaluation, the color appearance of a cell in merged TPAF/SHG pseudo-color images was used to determine the viability status. In the quantitative evaluation, thresholding of the normalized autofluorescence ratios of cells was used. Owing to the nonlabeling and nondestructive nature of this method, chondrocyte viability may also be measured on patient cartilage in vivo. However, this TPAF-based assay requires human involvement and, thus, the throughput is low. Therefore, automated methods suitable for processing cell images and evaluating chondrocyte viability are needed to improve the throughput.

The automated methods should be designed to accurately identify live and dead cells in an image and count their respective numbers. Typically, such an image analysis involves two main tasks: to recognize individual cell regions (segmentation) and to classify them as live or dead cells (classification). In previous studies, we have successfully demonstrated a deep learning strategy for the chondrocyte viability analysis⁶. A U-Net⁷ model was developed for cell segmentation and Convolution Neural Network (CNN)-based models were developed for two independent classifications that determined the living and total cell populations, respectively. The reason for using two classification networks was to improve the accuracy since perfect segmentation of each cell was difficult to achieve. Although the accuracy of 90% was achieved in the viability analysis, this method has an obvious disadvantage related to the involvement of multiple neural networks and multiple steps to achieve the viability result. Since each neural network needed its own ground truth, model training, and evaluation processes, the workload for building the models was high. To build a unified and efficient network for both segmentation and classification, we hypothesize that Mask R-CNN⁸, a state-of-art deep CNN, can meet our needs in the viability analysis. Mask R-CNN⁸ is an instance segmentation algorithm and a multitasking network that can perform object detection, target instance segmentation, and target keypoint detection. Mask R-CNN is an extension of region-based CNN (R-CNN) and Faster R-CNN, and works by adding a branch for predicting an object mask (Region of Interest) in parallel with the existing branches for bounding box recognition. Advantages of Mask R-CNN include its simplicity and intuitive structure for training, higher performance than almost all existing, single-model entries on every task, and flexibility to be adapted in different applications. Mask R-CNN provides high speed and high accuracy in classification and instant segmentation, which is beneficial imaging processing tasks in the field of microscopy.

In this report, we first describe a customized multi-channel two-photon microscope built with the measurement of the chondrocyte viability in mind. This microscope can simultaneously acquire two channels of fluorescence (NAD(P)H and FAD, respectively) and one channel of SHG images. We then introduce a newly developed deep learning method for the chondrocyte viability analysis using Mask R-CNN. The preliminary result shows that the accuracy can reach 90% with a single neural network.

2. METHODS AND MATERIALS

2.1 Home-built four-channel desktop two-photon microscope

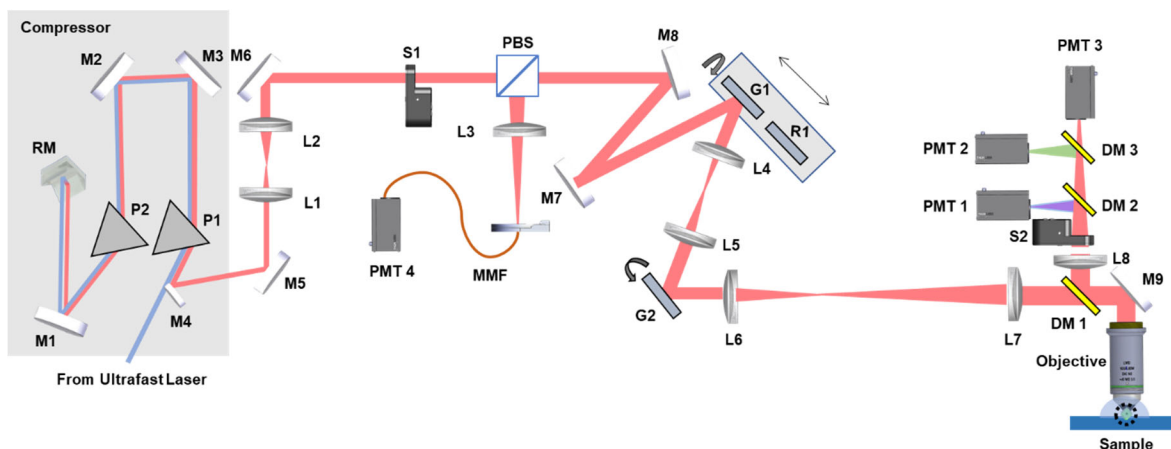


Figure 1. The schematic diagram of the homebuilt two-photon fluorescence/SHG microscope. PBS, polarizing beamsplitter; L1~7: lens; M1~8: mirrors; PMT1~4: photomultiplier; G1~2: galvanometer scanner; R1: resonant galvanometer; DM1~3: dichroic mirror; S1~2: shutters; MMF: multimode fiber; P1~2: dispersion compensation prism pair.

As shown in Figure 1, the home-built two-photon microscope was equipped with three recording channels: two for TPAF of NAD(P)H and FPs, respectively, and one for the SHG signal. In addition, one near-IR reflective confocal channel was implemented for recording back-scattering signals from samples. All 4 channels acquire images simultaneously. A femtosecond Ti:Sapphire laser (Chameleon Ultra II, Coherent Inc.) was equipped as the excitation light source. To maintain the ultrashort pulse length (about 140 femtosecond) of the excitation pulses, the laser beam passed a dispersion compensation system before entering the microscope. The dispersion compensation system consisting of a pair of SF10 prisms provided negative group velocity dispersion (GVD)⁹ to compensate the positive GVDs brought by all optical components and tissues through which the excitation beam traveled. After the dispersion compensation system, a telescope composed of L1 (AC204-030-B-ML, Thorlabs) and L2 (AC204-060-B-ML, Thorlabs) was used to collimate and resize the beam. A pair of galvanometer scanning mirrors (G1 and G2) (GVS011, Silver-Coated Mirror, Thorlabs) was used for imaging at slow scanning speeds (up to 160 Hz). A resonant scanner (SC-30, Electro-Optical Products Corp.) that sat on a translation stage with G1 could be switched in to replace G1 for imaging at a fast scanning speed (8 kHz). Between the two galvanometer scanning mirrors (G1 and G2), a telescope composed of L4 (AC204-050-B-ML, Thorlabs) and L5, (AC204-075-B-ML, Thorlabs) was used to relay and expand the beam. Being deflected by galvanometers, the laser beam passed through the scan lens (VIS-NIR Achromatic Lenses, 60 mm, #49-379, Edmund Optics) and the tube lens (AC508-250-B-ML, Thorlabs) and was incident to the imaging objective (CFI LWD Plan Fluorite, 16x, NA0.8, Nikon). The emitted fluorescence from samples was collected by the same 16x objective and was separated by a dichroic mirror DM1 (FF735-Di01, Semrock) from the excitation beam. Three photomultiplier tubes, PMT1 (PMTSS, Thorlabs), PMT2 (PMT2101, Thorlabs), and PMT3 (PMT2101, Thorlabs) were used to detect TPAF and SHG with an appropriate set of dichroic mirrors and bandpass filters. For the chondrocyte viability study, PMT1, 2, and 3 were assigned to detect signals of SHG, NAD(P)H, and FPs, respectively. To increase the collection efficiency, emission from samples was first collected by an achromatic lens L8 (#49-391, Edmund Optics) and was then focused onto each PMT by aspheric lenses (ACL2520-A,

ACL25416U-A, and ACL25416U-A, Thorlabs). Two dichroic mirrors (FF397-Di01-25x36 and FF470-Di01-25x36, Semrock) and three bandpass filters (FF01-370/36-25, FF01-442/42-25, and FF01-607/70-25, Semrock) placed in front of the three PMTs were used to sort out appropriate signals to each channel. Two sets of data acquisition systems were installed in a PXI chassis (NI PXIe-1620Q, National Instruments) for taking the video rate (resonant galvanometer) and low frame rate (galvanometers) imaging, respectively. For the video-rate image acquisition, a high-speed digitizer (NI 5732, National Instruments) mounted on a Field Programmable Gate Array (FPGA) adapter module (PXIe 7961R, National Instruments) provided a signal acquisition externally triggered by the reference frequency from the ultrafast laser pulses (80 Hz) so that a sample per pulse was implemented for imaging acquisition. For the low frame rate imaging, a multifunction data acquisition board (PXI 6115, National Instruments) provided waveforms to control the XY scanner and acquired 4 channels of data simultaneously. The back-scattering signal was descanned and picked off by a polarizing beam splitter PBS. A multi-mode fiber with a NA 0.25 worked as a pinhole and directed light to a PMT(PMM02, Thorlabs) to detect the signal. SciScan (Scientifica, UK)¹⁰ written in LabVIEW 2014 32-bit (National Instruments, Austin, TX) was used to control the scanning and acquired images.

2.2 Sample preparation and imaging

Porcine hind knee joints were obtained from a local slaughterhouse. The joint cavity was opened leaving the cartilage surface exposed. The cartilage samples were harvested using 5 mm (ID) sample corers (18035-05, Fine Science Tools) and were stored in Dulbecco's phosphate-buffered saline (DPBS, Corning) at 4 °C in a fridge. For imaging, samples were placed in Petri dishes or 3D printed sample holders and immersed in DPBS. Samples were imaged with the homebuilt multi-channel TPF/SHG microscope described in the previous section. For imaging porcine cartilage, the excitation wavelength was tuned to 740 nm so that autofluorescence of NAD(P)H and FPs in chondrocytes could be excited simultaneously. The two fluorescence channels have bandpass filters at the ranges of 400 – 480 nm (Channel 2) and 530 – 670 nm (Channel 3), corresponding to the NAD(P)H and FP emissions, respectively. The SHG of collagen was recorded in Channel 1 with a bandpass of 352-388nm. The three-channel two-photon microscope allowed us to acquire NAD(P)H, FP, and SHG images simultaneously. Image stacks were typically acquired consisting of more than 30 512x512 pixel images from the surface to deeper layers at a depth interval of 2 μ m/step.

2.3. Ground truth of segmentation and classification

For visual assessment of cell viability, ImageJ (Fiji)¹¹ was used to form a single RGB color image by merging acquired three-channel images, with red (R) assigned to the FPs channel, green (G) NAD(P)H channel, and blue (B) to the SHG channel. Figure 2 shows typical TPAF and SHG images and the corresponding RGB color image, with the scale bar representing 10 μ m. The raw images were not subjected to any image processing or enhancement during this image merging process. To establish the pre-training data set, we used the Python-based environment Labelme¹² to label individual chondrocytes. We visually assess the viability of each cell in the image by evaluating the brightness and color appearance and classify the cells into two categories: live and dead². Cells that were bright and green were categorized as live, while cells that were dim and red were categorized as dead. Each cell was outlined by multi-point line segments and labeled with a dead or live status in Labelme. The resulting training set was used as the ground truth to train the Mask R-CNN model.

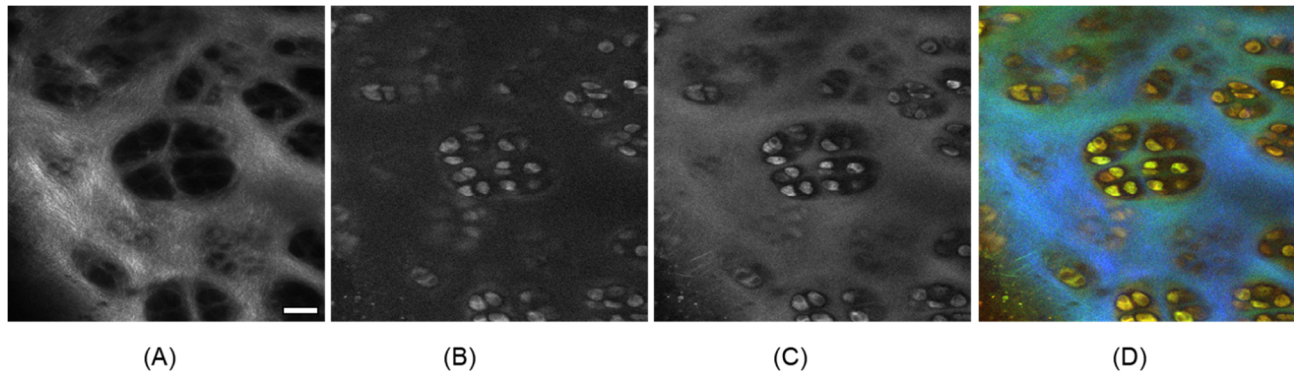


Figure 2. Typical TPAF/SHG images acquired by the homebuilt three-channel two-photon/SHG microscope. (A) SHG; (B) FP channel; (C) NAD(P)H channel; (D) Merged. Scale: 10 μm .

2.4 Deep learning segmentation and classification with Mask R-CNN

Building on the Faster R-CNN framework¹², Mask R-CNN uses the same two stages architecture as Faster R-CNN, but adds full connected network (FCN)¹³ for semantic segmentation of each property box of Fast R-CNN¹⁴. As shown in Figure 3, the Mask R-CNN algorithm processes images in two stages. The first stage obtains a feature map and an alternate ROI through the ResNet⁴ and FCN algorithm. For our application, images with a size of 256×256 pixels are fed into a pre-trained neural network (e.g., ResNet) to obtain corresponding feature maps. ResNet uses cross-layer connections to make training easier, and the FCN algorithm is a classic semantic segmentation algorithm that accurately segments targets in images. By setting a predetermined ROI for each point in the FCN's feature map, multiple candidate ROIs are obtained. Then, these candidate ROIs are fed into the RPN network for binary classification (foreground or background) and (Bounding Box) regression⁸ to filter out some of the candidate ROIs. The network recommended by the RPN region is used to help the network recommend the region of interest. The second stage is to perform the ROIAlign⁷ on these remaining ROIs. That is, the original image and the pixel of the feature map have corresponded, and the feature map is corresponding to the fixed feature. ROI Align⁸ is a proposed way of region feature aggregation in Mask R-CNN, which is a good solution to the problem of region mismatch caused by two quantization processes, i.e., ResNet and RPN) in ROI Pooling operation. In CNN, a small part of the image (local receptive area) is used as the input of the lowest layer of the hierarchy, and the information is transmitted to different layers; in turn, each layer passes a digital filter to obtain the most salient features of the observation data. This method can capture salient features of observations that are invariant to translation, scaling, and rotation because the local receptive regions of the image allow neurons or processing units to access the most fundamental features, such as oriented edges or points. Finally, these ROI values form a vector as the input to a traditional neural network for classification. In parallel, the above obtained ROIs undergo classification (N-class classification), (Bounding Box) regression¹⁵, and mask generation. Here, each ROI corresponds to only one object, and an FCN operation is performed on each ROI. Mask R-CNN uses parallel computation to classify and generate masks for each ROI. Such an operation can improve the operation speed of the Mask R-CNN model. The outcome of the analysis included images with cells that were classified, segmented, and categorized as live and dead cells.

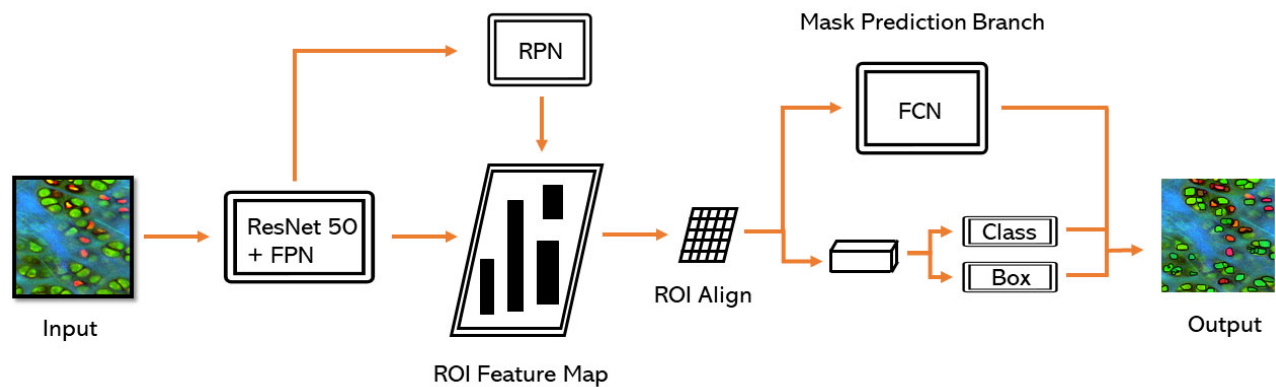


Figure 3. Schematics of the architecture of Mask R-CNN for chondrocyte segmentation and classification. ResNet: residual networks; FPN: feature pyramid network; RPN: region proposal network; ROI: region of interest.

2.5 Chondrocyte viability assessment

Chondrocyte viability is defined as the percentage of live cells to the total number of cells. With perfect cell segmentation, viability is easily calculated after cell classification, requiring only two categories, i.e., live and dead. Once the number of live and total cells is determined by the classification process, the chondrocyte viability of a cartilage sample can be determined using equation (1). Figure 3 represents the workflow and results of the proposed automated chondrocyte assessment⁶.

$$\text{Viability} = \frac{\sum_{i=1}^N N^{(i)}_{\text{Live}}}{\sum_{i=1}^N N^{(i)}_{\text{Total}}} \quad (1)$$

where i is the index of images that represents the sample, with a maximum value of N ; N_{Live} and N_{Total} are the number of viable cells and the total number of cells in an image, respectively. In this study, we used one image per sample for the viability analysis, i.e. $N=1$.

2.6 Evaluation of the performance of segmentation and classification

The F1-score⁴ is a commonly accepted measure for evaluating the performance of a classifier. It is the harmonic mean the precision and recall, with a maximum of 1 and a minimum of 0. The F1-score is defined as,

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

The precision and recall are defined as follows,

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

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

where TP, FP, and FN stand for the true positive, false positive, and false negative, respectively.

The root mean square error (RMSE) is used to assess the overall outcome of the viability assessment and is defined as,

$$RMSE = \sqrt{\sum_{i=1}^N (R_G - R_P)^2 / N} \quad (5)$$

where R_G is the ground truth as manually defined, and R_P is the model prediction. The root square error (RSE) of a single image evaluated in Figure 5 is calculated by,

$$RSE_i = \sqrt{(R_{Gi} - R_{Pi})^2} \quad (6)$$

where R_{Gi} is the viability determined by the ground truth and R_{Pi} is the viability predicted by the model.

3. RESULTS AND DISCUSSION

The Mask R-CNN model was trained by 800 images, which consisted of 200 acquired 256×256 pixels images and 600 transformed images by operations such as rotation, flipping, and zooming. To evaluate the trained Mask R-CNN, we randomly selected a total of 30 images with a size of 256×256 pixels from 30 different image stacks. Two typical such images are shown in Figure 4 (A) and (B). Their segmentation and classification results are shown in Figure 4(C) and (D), respectively. The segmented cells are circled by black lines and the circled areas are filled with either red or green colors with a different color saturation level. The red color indicates a dead cell, while the color saturation level represents a network output weight value that is below 0.5 (threshold). The green color indicates a live cell, while the color saturation level represents a network output weight value that is over 0.5 (threshold). Note that not all cells in an image are segmented because the program those cells are in the other depth due to their low signal levels. This implies that 3D analysis may provide more accurate results.

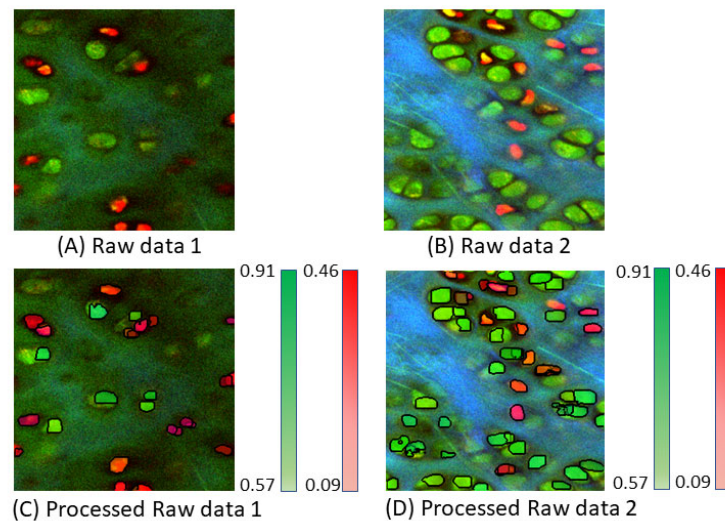


Figure 4. Two typical three-channel cartilage images and their segmentation and classification results using the Mask R-CNN. (A) and (C) are three-channel merged raw images acquired by the homebuilt two-photon microscope. (B) and (D) are the outputs from the Mask R-CNN.

The Mask R-CNN achieves a high precision in the classification of live/dead chondrocytes. With the 30 evaluation images, the F1-score reaches 0.90 ± 0.04 . The RSE of the chondrocyte viability in corresponding images is found to be 0.15 ± 0.06 , which means an average 85% accuracy in the viability analysis. Figure 5 shows the statistical result of the RSE of the 30

images. The current accuracy using Mask R-CNN is lower than the best result that we could achieve using the previously reported multiple networks⁶, partly due to the smaller training set, 600 (Mask RCNN) vs 2000 (multiple networks).

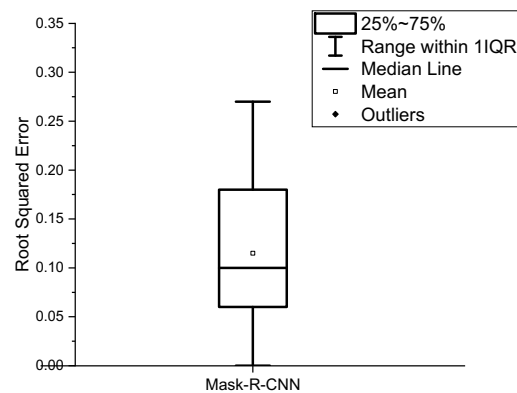


Figure 5. The RSE of the chondrocyte viability according to the classification results of the Mask R-CNN.

Besides Mask R-CNN, there are other similar networks, such as Yolo⁴, MaskLab⁴, etc., which use more layers for image segmentation/detection and classification. Deeper networks may improve performance and quantitative accuracy but increase computational costs and are highly dependent on the computing power of processors (CPUs, single GPUs, or multiple GPUs). Since acquired images are 3D images in general, 3D chondrocyte segmentation and classification may provide a more accurate viability analysis. Mask R-CNN may be used to build the network for 3D microscopic image processing.

4. CONCLUSION

In this paper, the Mask R-CNN algorithm demonstrates the ability to analyze cell viability with high precision through a multitasking network of target recognition in unlabeled chondrocyte images obtained from the home-built two-photon fluorescence microscopy. Regarding the segmentation and classification of chondrocytes, the method using Mask R-CNN could achieve 85% accuracy. Mask R-CNN algorithm simplifies the computational model, makes the algorithm easy to use, and improves the operation speed by using parallel computation. The segmentation and classification strategies described in this report no longer require multiple complex models and the trained model can be used to process images of similar shapes in different types of chondrocytes without additional training. This greatly improves the range of use of this automated imaging processing method.

5. ACKNOWLEDGEMENT

This research was in part supported by South Carolina IDeA Networks of Biomedical Research Excellence (SC INBRE), a National Institutes of Health (NIH) funded center (Award P20 GM103499), MTF Biologics Extramural Research Grant, South Carolina Translation Research Improving Musculoskeletal Health (TRIMH), an NIH funded Center of Biomedical Research Excellence (Award P20GM121342), Clemson University's Robert H. Brooks Sports Science Institute (RHBSSI) Seed Grant, a grant from the National Science Foundation (1539034), and the National Science Foundation EPSCoR

6. REFERENCES

1. Denk, W., Strickler, J.H., and Webb, W.W., "Two-photon laser scanning fluorescence microscopy," *Science* 248(4951), 73–76 (1990).
2. Helmchen, F., and Denk, W., "Deep tissue two-photon microscopy," *Nature Methods* 2(12), 932–940 (2005).
3. Georgakoudi, I., and Quinn, K.P., "Optical Imaging Using Endogenous Contrast to Assess Metabolic State," *Annual Review of Biomedical Engineering* 14(1), 351–367 (2012).
4. Li, Y., Chen, X., Watkins, B., Saini, N., Gannon, S., Nadeau, E., Reeves, R., Gao, B., Pelligrini, V., et al., "Nonlabeling and quantitative assessment of chondrocyte viability in articular cartilage with intrinsic nonlinear optical signatures," *Experimental Biology and Medicine* 245(4), 348–359 (2020).
5. Kühn, K., D'Lima, D.D., Hashimoto, S., and Lotz, M., "Cell death in cartilage," *Osteoarthritis and Cartilage* 12(1), 1–16 (2004).
6. Chen, X., Li, Y., Wyman, N., Zhang, Z., Fan, H., Le, M., Gannon, S., Rose, C., Zhang, Z., et al., "Deep learning provides high accuracy in automated chondrocyte viability assessment in articular cartilage using nonlinear optical microscopy," *Biomedical Optics Express* 12(5), 2759–2772 (2021).
7. Ronneberger, O., Fischer, P., Brox, T., U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer2015, pp. 234-241.
8. He K., Gkioxari G., Dollar P., and Girshick R., "Mask R-CNN," in 2017 ' IEEE International Conference on Computer Vision (ICCV), December 2017, pp. 2980-2988.
9. Fork, R.L., Cruz, C.H.B., Becker, P.C., and Shank, C.V., "Compression of optical pulses to six femtoseconds by using cubic phase compensation," *Optics Letters* 12(7), 483–485 (1987).
10. Ward, G., Murray, A., and Wilms, C., "Open-Source Software for Controlling Two-Photon Laser Scanning Microscopes," *Microscopy Today* 25(5), 12–17 (2017).
11. Rasband, W.S., ImageJ: Image processing and analysis in Java, *Astrophysics Source Code Library*2012, 2:378.
12. Russell B, Torralba A, Murphy K, Freeman W. T., Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision* 2007.
13. Ren S., He K., Girshick R., Faster R-CNN: Towards real-time object detection with region proposal networks. In *TPAMI*2017.
14. Long, J., Shelhamer, E., Darrell, T., Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*2015, pp. 3431-3440.
15. Girshick R. Fast R-CNN. In *ICCV* 2015.
16. Bell S., Zitnick C. L., Bala K., and Girshick R. Inside outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR* 2016.