An Ensemble Learning and Slice Fusion Strategy for Three-Dimensional Nuclei Instance Segmentation

Liming Wu* Alain Chen* Paul Salama†
Kenneth W. Dunn‡ Edward J. Delp*
* Video and Image Processing Laboratory (VIPER), Purdue University, West Lafayette, Indiana
‡ Division of Nephrology, School of Medicine, Indiana University, Indianapolis, Indiana
† Department of Electrical and Computer Engineering, Indiana University-Purdue University, Indianapolis, Indiana

Abstract

Automated microscopy image analysis is a fundamental step for digital pathology and computer aided diagnosis. Most existing deep learning methods typically require post-processing to achieve instance segmentation and are computationally expensive when directly used with 3D microscopy volumes. Supervised learning methods generally need large amounts of ground truth annotations for training whereas manually annotating ground truth masks is laborious especially for a 3D volume. To address these issues, we propose an ensemble learning and slice fusion strategy for 3D nuclei instance segmentation that we call Ensemble Mask R-CNN (EMR-CNN) which uses different object detectors to generate nuclei segmentation masks for each 2D slice of a volume and propose a 2D ensemble fusion and a 2D to 3D slice fusion to merge these 2D segmentation masks into a 3D segmentation mask. Our method does not need any ground truth annotations for training and can inference on any large size volumes. Our proposed method was tested on a variety of microscopy volumes collected from multiple regions of organ tissues. The execution time and robustness analyses show that our method is practical and effective.

1. Introduction

Optical microscopes have been widely used for imaging microscopic organisms and are important for the understanding of subcellular structures and disease diagnosis [1, 2]. The advances in digital fluorescence microscopy enabled multi-channel high resolution 3D imaging by using a diffraction-limited laser beam which can image deeper subcellular tissue structures [3, 4]. The 3D volumes generated by fluorescence microscopy need to be quantitatively analyzed to obtain useful information [5]. Manual assessment of large-scale microscopy volumes is laborious and time-consuming.

Deep learning-based methods have shown significant performance for computer vision tasks such as image classification, object localization and segmentation [6]. One of the major challenges in analyzing 3D microscopy volumes is to accurately delineate the boundary of individual nuclei having high intraclass variability and densely overlapping distributions [7, 8]. Many deep learning methods for image segmentation such as encoder-decoder-based networks typically require post-processing such as watershed or morphological operations to separate touching objects which may result in unstable results [9, 10]. Large computational resources such as large amounts of GPU memory are also necessary. One solution to reduce computational complexity is to process the volume as 2D slices and then merge the results to form a 3D nuclei segmentation. Robustly merging or fusing the 2D nuclei segmentations without compromising the overall accuracy remains challenging since the 3D nuclei information is not learned properly. Ensemble learning has been widely used to increase the overall robustness of segmentation approaches as well as increasing the segmentation accuracy by integrating the voting results from different networks [11]. However, supervised learning methods require large amounts of annotated training samples to achieve accurate results. Due to the lack of large amounts of ground truth data, quantitative analysis for some applications needs to be conducted without ground truth annotation for training supervised learning models [12].

In this paper, we describe an ensemble method for 3D nuclei segmentation, known as Ensemble Mask R-CNN (EMR-CNN), that is based on a collection of Mask R-CNN models with different network architectures for detecting and segmenting 3D nuclei in fluorescence microscopy volumes. We propose a weighted 2D mask fusion technique for aggregating 2D detection results from different Mask R-CNN networks to achieve more accurate and robust 2D results. We describe a 2D to 3D slice fusion method for merging segmentation results from 2D slices to a 3D volume using an unsupervised clustering method. By using ensemble learn-

ing, we demonstrate that our method achieves high nuclei detection accuracy compared to other methods we examined. In addition, we use Generative Adversarial Networks (GANs) to generate synthetic 3D microscopy volumes for training our EMR-CNN. Therefore our approach does not require any hand annotated ground truth for training which will be more useful when the hand annotated data is limited.

2. Related Work

Many methods have been reported in the literature for nuclei segmentation. They can generally be divided into five categories: threshold-based methods, clustering methods, energy-based methods, region-based methods, and machine learning methods [13].

The thresholding methods such as Otsu's method try to determine a threshold that separates the foreground and background pixels by minimizing the intraclass intensity variance [14]. The typical region-based method is watershed [15], which treats the grayscale image as a topographic landscape with ridges and valleys, and the watershed transform is used to build barriers on the ridges to separate water source from different regions. The use of Otsu's thresholding and watershed has been a popular combination for microscopy image segmentation [16, 17, 18]. Energy-based methods known as "Active contours" seek an equilibrium between the foreground object and background pixels by iteratively moving a deformable spline to minimize an energy function [19, 20].

Unsupervised learning techniques such as k-means, agglomerative hierarchical clustering (AHC), fuzzy C-means (FCM), and mean shift clustering have been used to split touching nuclei [21, 22, 23, 24]. These clustering methods explore the structure of nuclei and aggregate the pixels with similar features into different nuclei instances. The methods described above have been implemented and integrated as ImageJ plugins as well as other open source tools for quantitative microscopy image analysis [25, 17, 20, 26].

More recently, deep learning-based methods have shown promising results for cellular image analysis [27]. The encoder-decoder networks such as U-Net [28] and SegNet [29] have been used in microscopy image segmentation [30, 31, 32, 33], which demonstrate significant improved performance over classical image segmentation techniques. To segment different nuclei instances, post-processing such as watershed or morphological operation with proper parameter tuning is required [34]. To address this issue, the network has been modified to learn the centroid and boundary information of individual nucleus using a voting mechanism [35] or vector gradient map [36]. Similarly, [37] models the nuclei as ellipsoids and directly estimates the centroids and radii of the nuclei for instance segmentation. Alternatively, instead of directly segmenting an entire image, top-down approaches such as Region Proposal Networks (RPN) first identify the regions of interest (RoIs) and then segment the RoIs to obtain instance segmentations [38].

Ensemble learning techniques have been used to improve the overall detection performance by combining multiple diverse detectors, which can compensate for the errors generated by individual detectors [39]. An ensemble Mask-aided R-CNN described in [40] uses a graph clique voting method for improving the detection performance. In [41] an ensemble learning method based on CNNs and random forest (RF) for blood vessel segmentation is presented. Similarly, in [42] a transform modal ensemble learning for breast tumor segmentation is described. In addition, [43] described a cross-modality fusion and feature learning level ensemble learning for multimodal medical image segmentation and demonstrated the superiority of feature fusion over network output fusion such as voting. A weighted boxes fusion method was described for aggregating detected bounding boxes from different object detection models [44].

For segmenting 3D microscopy volumes, directly using 3D CNN networks can generally obtain more accurate results since the 3D information is utilized. However, 3D methods require 3D annotated ground truth for training, which is difficult to obtain in practice, especially for 3D microscopy volumes. To address these issues, many 2D to 3D methods have been introduced, which first perform the segmentation on the x-, y-, and z-direction of a volume and then fuse the results. In [45] 3D vector gradients are estimated by averaging 2D vector gradients from a modified 2D U-Net from three different directions of a volume, and followed by a clustering method to group the pixels to 3D masks. Similarly, [46] uses majority voting to combine 2D segmentation results obtained from SegNet into a 3D segmentation. However, accurately and robustly aggregating these 2D segmentation into 3D masks remains challenging.

To obtain satisfactory segmentation results, deep learning methods typically need large amounts training images with corresponding ground truth annotations. Manually delineating 3D nuclei contours or even 2D nuclei contours is laborious in a microscopy volume even for an expert. To address these issues, data augmentation methods including elastic deformation [47], random intensity correction, and spatial transformation are commonly used [48]. However, these methods require an adequate number of existing ground truth images. Learning-based data augmentation techniques such as generative adversarial networks (GANs) [49] can generate synthetic data without ground truth images. In [32] and [50], nuclei segmentation masks are generated by modeling nuclei as 3D ellipsoids and 3D non-ellipsoids using Bézier curves. In [31], a synthetic microscopy image generation method that is based on a modified CycleGAN [51], known as Sp-CycleGAN, uses the binary nuclei segmentation masks to generate corresponding synthetic microscopy images. In this paper we will use GANs to generate synthetic microscopy volumes for training our proposed method.

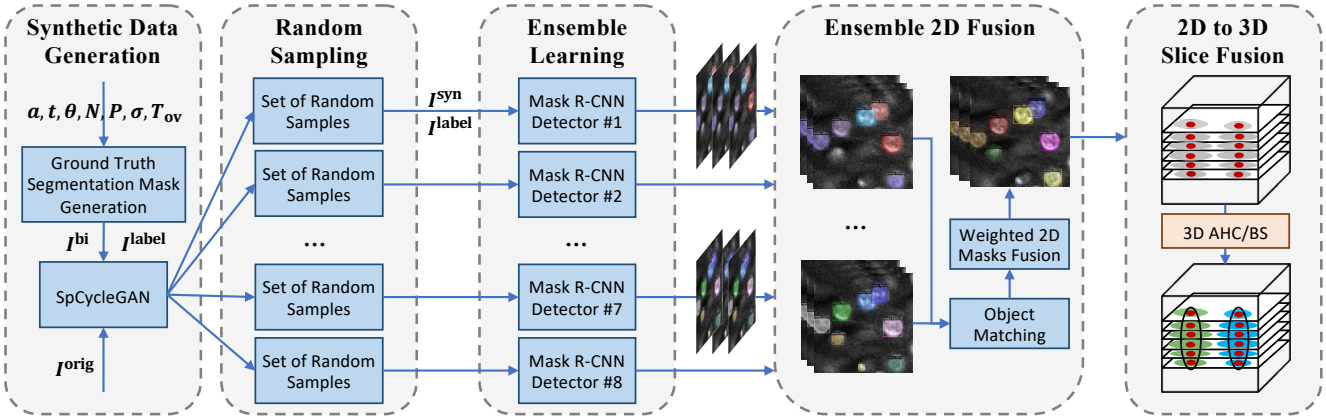

Figure 1. Overview of the proposed EMR-CNN for 3D nuclei instance segmentation using ensemble learning and slice fusion

3. PROPOSED METHOD

In this section we describe the proposed Ensemble Mask R-CNN (EMR-CNN). We denote I as a 3D volume with dimension $X \times Y \times Z$. I_{z_j} is the j-th slice of a volume on the z direction where $j \in \{1, ..., Z\}$, and I_z denotes all 2D slices along the z direction of I. Also, I^{orig} denotes the original microscopy volumes. I^{bi} and I^{label} denote the binary segmentation masks and corresponding labels.

Let E be a collection or ensemble of 2D object detectors where m_i is a detector in E and $i \in \{1, ..., M\}$. Given a 2D microscopy image $I^{\text{orig}}_{z_j}$ and an object detector m_i, the detection results are denoted as $\text{Det}^{m_i,z_j} = \{\text{Det}^{m_i,z_j}_1, ..., \text{Det}^{m_i,z_j}_n\}$ where $n = N^{m_i,z_j}_{\text{det}}$ is the total number of detected objects in $I^{\text{orig}}_{z_j}$ from m_i, and each detection result $\text{Det}^{m_i,z_j}_d = \{\text{Seg}^{m_i,z_j}_d, \text{Ctr}^{m_i,z_j}_d, \text{Pr}^{m_i,z_j}_d\}$ consists of a segmentation mask Seg^{m_i,z_j}_d, an object centroid Ctr^{m_i,z_j}_d and a confidence score Pr^{m_i,z_j}_d. Seg^{m_i,z_j}_d is a binary image of size $X \times Y$ where segmented pixels are highlighted with intensity 1. Ctr^{m_i,z_j}_d is a 2D coordinate indicating the centroid of the segmentation mask Seg^{m_i,z_j}_d, and $\text{Pr}^{m_i,z_j}_d \in (0, 1)$ is the detection confidence score of the segmentation mask.

Also, let $\text{Det}^{m,z_j} = \{\text{Ctr}^{m,z_j}, \text{Seg}^{m,z_j}, \text{Pr}^{m,z_j}\}$ be the 2D detection results on $I^{\text{orig}}_{z_j}$ for all detectors in E. Our goal is to take all the detection results and fuse them together. Let $\text{Det}^{z_j} = \{\text{Ctr}^{z_j}, \text{Seg}^{z_j}, \text{Pr}^{z_j}\}$ be the fused 2D detection result using ensemble 2D fusion method described in Section 3.3. Let Det^{z} represent a set of all 2D fused results for I^{orig}, and $\text{Det} = \{\text{Seg}, \text{Ctr}, \text{Pr}\}$ is the final 3D results fused from Det^{z}. Next we take the fused 2D results Det^{z} and merge them to form our final 3D detection Det using a 2D to 3D slice fusion method described in Section 3.4.

3.1. Synthetic Data Generation

As we indicated earlier it is very difficult to obtain manually annotated microscopy volumes due to the tedious nature of the annotation process. We use a data augmentation process that consists of generating synthetic 3D microscopy volumes using GANs [51, 31]. As shown in Figure 1, the synthetic data generation module includes 3D ground truth nuclei segmentation masks generation and synthetic microscopy volume generation.

Ground truth segmentation mask generation. We first generate synthetic 3D segmentation masks of nuclei that serve as the ground truth for the training data. Our approach is different from previous approaches described in [32, 50] in that we model each candidate nucleus as a deformed 3D ellipsoid parameterized by three parameters $\boldsymbol{a}$, $\boldsymbol{t}$, and $\boldsymbol{\theta}$, where $\boldsymbol{a} = (a_x, a_y, a_z)$ defines three axis lengths of an ellipsoid, $\boldsymbol{t} = (t_x, t_y, t_z)$ is the spatial translation that defines the location of a nucleus, and $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$ is the spatial rotation that defines the orientation of a nucleus. The parameters $\boldsymbol{a}$, $\boldsymbol{t}$ and $\boldsymbol{\theta}$ are randomly generated in a range shown in Table 1 for each candidate nucleus. These candidate nuclei are recursively added to an empty 3D volume I^{label} with an incremental unique intensity k that is used to distinguish different nuclei instances. The total number of nuclei is set to N, and the overlapping voxels of two nuclei must be less than T_{ov}.

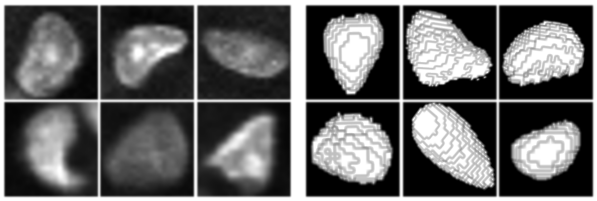

Figure 2. Real nuclei (left) and deformed ellipsoids generated using elastic transform (right)

In the data we used for our experiments, we observed many nuclei are not strictly ellipsoids but more like "deformed" ellipsoids (Figure 2). To model these type of nuclei, we further deform the binary ellipsoids using an elastic transform [52]. Specifically, given a volume I of size $X \times Y \times Z$ that needs to be deformed, we first generate a random coarse displacement, which is a matrix of size $3 \times P \times P \times P$,

sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$. Then a displacement vector field I^{vec} was generated by interpolating the coarse displacement from size $3 \times P \times P \times P$ to size $3 \times X \times Y \times Z$ by cubic spline interpolation [53]. Then, I^{vec} is used to shift the voxels in I to obtain the deformed volume I^{bi} (Figure 2).

Microscopy volume generation. We use the SpCycleGAN described in [31, 51] to generate the synthetic 3D microscopy volume. As shown in Figure 3, for SpCycleGAN training, we use all XY focal planes of unpaired original microscopy volumes I^{orig} and synthetic binary segmentation masks I^{bi}. Once the SpCycleGAN is trained, we use I^{bi} as input to SpCycleGAN to generate a corresponding synthetic microscopy volume I^{syn}. We generate the volume slice by slice and stack the results together.

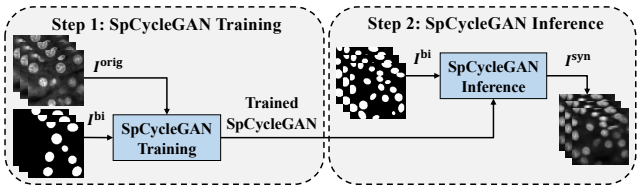

Figure 3. 3D synthetic microscopy volume generation using SpCycleGAN

SpCycleGAN consists of 5 networks G, F, H, D_A and D_B where G is the generator that translates the image from binary domain to microscopy domain. Similarly, F translates the image from microscopy domain to the binary domain. D_A and D_B are two discriminators that are used to discriminate whether a given image is real or synthetic. The additional segmentation network H has the same architecture as F along with spatial constrained loss $\mathcal{L}_{\text{sc}}$ were introduced in [31] to preserve the spatial shift of objects in the generated images. The entire objective loss function of SpCycleGAN is shown in Equation 1.

$$\begin{aligned}\mathcal{L}(G, F, H, D_A, D_B) &= \mathcal{L}_{\text{GAN}}(G, D_A, I^{\text{bi}}, I^{\text{orig}}) \\ &+ \mathcal{L}_{\text{GAN}}(F, D_B, I^{\text{orig}}, I^{\text{bi}}) \\ &+ \lambda_1 \mathcal{L}_{\text{cycle}}(G, F, I^{\text{orig}}, I^{\text{bi}}) \\ &+ \lambda_2 \mathcal{L}_{\text{sc}}(G, S, I^{\text{orig}}, I^{\text{bi}}) \qquad (1)\end{aligned}$$

where $\mathcal{L}_{\text{sc}}$ is the spatial constrained loss defined as a L_2 norm shown in Equation 2.

$$\mathcal{L}_{\text{sc}}(G, S, I^{\text{orig}}, I^{\text{bi}}) = \mathbb{E}_{I^{\text{bi}}}[||H(G(I^{\text{bi}})) - I^{\text{bi}}||_2] \qquad (2)$$

3.2. Ensemble Mask R-CNN: EMR-CNN

In order to increase the robustness and accuracy of nuclei segmentation, we propose a simple but effective method that trains a collection of M different but similar Mask R-CNN detectors implemented by [54]. The details of the training are described in Section 4.1. Our method includes ensemble 2D fusion that is able to fuse the 2D detection results from all detectors, and a 2D to 3D slice merging method that merges the detection results from fused 2D slices to 3D volumes.

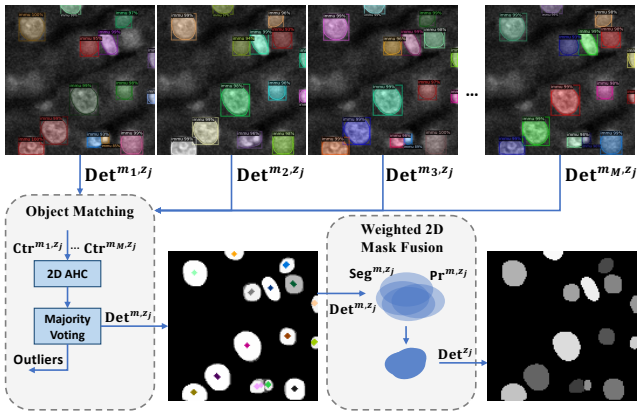

Figure 4. Overview of weighted 2D mask fusion

3.3. Ensemble 2D Fusion

We propose a weighted 2D mask fusion method for fusing 2D detection results from different detectors into a final 2D detection. This is an extension of the weighted boxes fusion described in [55]. Instead of simply using Non-Maximum Suppression (NMS), we use the estimated confidence scores to compute the fused 2D segmentation mask which is represented by a probability map. A fused mask with new confidence score is generated. Given the detection results of a 2D slice $I^{\text{orig}}_{z_j}$ from M detectors, the goal is to fuse Seg^{m,z_j} and Pr^{m,z_j} into Seg^{z_j} and Pr^{z_j}.

Object matching. For an object, we need to identify all objects detected with the M detectors and then fuse the results. We use the agglomerative hierarchical clustering (AHC) with average linkage criterion to match the same segmented object in the segmentation masks Seg^{m,z_j}. Specifically, we use Ward's minimum variance implemented with Lance–Williams dissimilarity (L) [56] to determine which segmentation mask to be merged at each iteration. The Lance–Williams dissimilarity measures the similarity between an existing cluster and newly merged cluster. AHC will treat each sample as one cluster initially and merge the most similar sample pair based on the Lance–Williams dissimilarity until all samples are merged as one cluster. To evaluate the clustering performance, we define the mean intracluster distance $a(i)$, and the mean intercluster distance $b(i)$ for a given number of clusters k in Equation 3.

$$\begin{aligned}a(i) &= \frac{1}{n_c - 1} \sum_{i,j \in C_c, i \neq j} d(i,j) \\ b(i) &= \frac{1}{k-1} \sum_{q, q \neq c} \left[\frac{1}{n_q} \sum_{i \in C_c, j \in C_q} d(i,j) \right] \qquad (3)\end{aligned}$$

For a given segmentation mask centroid i and its cluster C_c, $a(i)$ measures the intracluster distance between i and

other samples j within the same cluster. Similarly, $b(i)$ measures the intercluster distance between i and other clusters C_q. n_c is the number of elements in cluster C_c. Note that $d(i,j)$ is the Euclidean distance between two centroids. Finally, the Silhouette Coefficient (SC) is used to determine the ideal number of clusters $N_{z_j}^{\text{cluster}}$. As shown in Equation 4, $\text{SC}_{z_j}(k)$ is the Silhouette Coefficient for the j-th slice given k as the number of clusters.

$$\text{SC}_{z_j}(k) = \frac{1}{n} \sum_{i \in \text{Ctr}^{m,z_j}} \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4}$$

Our objective function can be formulated in Equation 5 to find the number of clusters $k \in \{k_{\min}, ..., k_{\max}\}$ with the highest Silhouette Coefficient.

$$N_{z_j}^{\text{cluster}} = \underset{k}{\text{argmax}}\, \text{SC}_{z_j}(k), k \in \{k_{\min}, ..., k_{\max}\} \tag{5}$$

where $k_{\min}$ and $k_{\max}$ shown in Equation 6 are the minimum and maximum number of detection for all detectors and σ^2 is the variance of the number detections for all M methods. Note that AHC requires there to be at least 2 potential clusters for unclustered centroids.

$$\begin{aligned} k_{\min} &= \max\left(\min_{i \in \{1,...,M\}} (N_{\text{det}}^{m_i,z_j}) - \sigma^2, 2\right) \\ k_{\max} &= \max_{i \in \{1,...,M\}} (N_{\text{det}}^{m_i,z_j}) + M + \sigma^2 \end{aligned} \tag{6}$$

The majority voting mechanism is used to increase the robustness of the fusion by removing false positive detections or outliers. As shown in Equation 7, if the number of detections for an object in cluster C_r is less than $M/2$, the corresponding detection results will be removed.

$$\text{Det}^{m,z_j} = \text{Det}^{m,z_j} - \bigcup_{d \in C_r} \text{Det}_d^{m,z_j} \tag{7}$$

Weighted 2D mask fusion. To fuse the 2D segmentation masks from the M detectors, we propose weighted 2D mask fusion that takes the matched objects as input and outputs the fused 2D detections with corresponding confidence scores. This is shown in Equation 8.

$$\begin{aligned} \text{Seg}_w^{z_j} &= \frac{\sum_{i \in C_w} \text{Pr}_i^{m,z_j} \text{Seg}_i^{m,z_j}}{\sum_{i \in C_w} \text{Pr}_i^{m,z_j}} > 0.5 \\ \text{Pr}_w^{z_j} &= \frac{1}{|C_w|} \sum_{i \in C_w} \text{Pr}_i^{m,z_j} \end{aligned} \tag{8}$$

where C_w is the w-th cluster within which the detection results are matched. Then $\text{Ctr}_w^{z_j}$ is the center of mass of $\text{Seg}_w^{z_j}$.

3.4. 2D to 3D Slice Fusion

As described above, for a given an image slice $I_{z_j}^{\text{orig}}$ from the original volume, we obtain M 2D detentions and fuse them to form a 2D fused slice result $\text{Det}^{z_j} = \{\text{Seg}^{z_j}, \text{Ctr}^{z_j}, \text{Pr}^{z_j}\}$. Our 2D to 3D slice fusion method merges 2D fused detections from all slices Det^z into a 3D detection $\text{Det} = \{\text{Seg}, \text{Ctr}, \text{Pr}\}$. As shown in Figure 1, we use two 2D to 3D slice fusion approaches, both based on the spatial location of the 2D centroid of the fused slices. The first approach known as blob-slice (BS) [57, 17] merges each 2D fused segmentation from the top slice to the bottom slice based on a predefined Euclidean distance of the centroid. The second approach uses agglomerative hierarchical clustering (AHC) described in Equation 3, 4, and 5 to cluster Ctr^z. The confidence scores of the final 3D fused segmented objects are the average of confidence scores of corresponding 2D segmentations within the same cluster.

Slice merging. Since EMR-CNN operates on different slices of a volume independently without knowing the 3D nuclei structures in the z-direction, it may fail to detect the nuclei in a slice due to artifacts or the effect of point spread function. This results in a single segmented nucleus containing two or more disjoint connected components. We propose a technique for merging these disconnected components. For the 3D fused segmentation of a single nucleus Seg_d, suppose a 2D fused segmentation $\text{Seg}_r^{z_j}$ is missing on the j-th slice, and suppose the 2D fused segmentation for its neighbor slices are $\text{Seg}_p^{z_{j-1}}$ and $\text{Seg}_q^{z_{j+1}}$, then the missing segmentations are given by the intersection of its two neighbor segmentations ($\text{Seg}_r^{z_j} = \text{Seg}_p^{z_{j-1}} \cap \text{Seg}_q^{z_{j+1}}$).

4. EXPERIMENTAL RESULTS

Datasets. In our experiments, we use three microscopy datasets denoted as $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ for evaluation. The data is from various regions of a rat kidney using confocal fluorescent microscopy with fluorescence label (Hoechst 33342 stain). Original microscopy data were collected by Malgorzata Kamocka and Michael Ferkowicz at the Indiana Center for Biological Microscopy [58]. Microscopy $\mathcal{D}_1$ consists of one volume of $X \times Y \times Z = 128 \times 128 \times 64$ voxels, $\mathcal{D}_2$ consists of 16 volumes of size $128 \times 128 \times 32$ voxels, and $\mathcal{D}_3$ consists of 4 volumes of size $128 \times 128 \times 40$ voxels. These datasets were manually annotated using ITK-SNAP [59]. For training, we generate synthetic data for training EMR-CNN and other comparison methods. The synthetic $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are generated using 3 different trained SpCycleGANs (Figure 3). Each type of synthetic dataset consists of 50 volumes of size $128 \times 128 \times 128$.

4.1. Experimental Setup

The training of SpCycleGAN requires unpaired I_z^{orig} and I_z^{bi} images. We first generate 54 binary segmentation masks

I^{bi} for $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ respectively, using the method described in Section 3.1. The axes lengths, translation distances, and rotation angles are randomly selected from uniform distributions having ranges $r_{x,y,z}$, $t_{x,y,z}$, and $\theta_{x,y,z}$, which are provided in Table 1, respectively. For synthetic microscopy volume generation, shown in Figure 3, we use all XY focal planes of 4 sub-volumes of I^{orig} along with 4 volumes of I^{bi} for training each SpCycleGAN model, respectively. We use the trained SpCycleGANs and the remaining 50 binary volumes to generate corresponding I^{syn} volumes for each dataset. Example images of the synthetic volumes are shown in Figure 5. The SpCycleGAN parameters are the same settings as described in [23, 31].

Table 1. Parameters for generating synthetic binary segmentation volumes. $r_{x,y,z}$, $t_{x,y,z}$, $\theta_{x,y,z}$ are randomly generated for each nucleus, and P, σ, N and T_{ov} are predefined values based on actual microscopy volumes

Data	$r_{x,y,z}$	$t_{x,y,z}$	$\theta_{x,y,z}$	T_{ov}	N	P	σ
$\mathcal{D}_1$	(4, 8)	(1,128)	(0,2π)	5	400	0	0
$\mathcal{D}_2$	(10, 14)	(1,128)	(0,2π)	10	40	0	0
$\mathcal{D}_3$	(8, 10)	(1,128)	(0,2π)	300	400	4	4

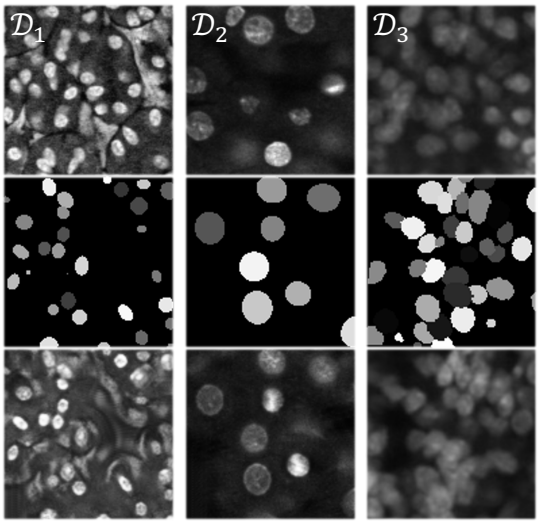

Figure 5. Original microscopy image (first row), synthetic nuclei segmentation mask (second row), and corresponding synthetic microscopy image (third row) for each dataset

EMR-CNN training and inference. In our experiments, we trained 3 EMR-CNNs each consisting of M Mask R-CNN detectors that are initialized with different network architectures randomly chosen from ResNet-50, ResNet-101 [60], ResNeXt-101 [61], Feature Pyramid Network [62], and deep CNN layers [63]. All the networks are pretrained on the MSCOCO dataset [64] and are available at [54]. The networks are then retrained on a subset of the synthetic microscopy images and tested on actual microscopy volumes. The training subset is randomly chosen from the entire synthetic image training set. All ensemble models are trained on 4 TITAN XP GPUs in parallel with a base learning rate of $2.5e^{-4}$ and batch size of 2 for 2000 iterations. The remaining parameters, which are Mask R-CNN detector parameters, are set to their default values [54]. In addition, since the training images are of size 128×128 while the testing images can be arbitrarily large, directly inferencing on those large volumes may not generate accurate detection results (see Figure 6 left column) because the objects are downscaled. To address this, we propose a divide-and-conquer inference scheme that partitions the input volume to multiple $128 \times 128 \times 128$ sub-volumes with a 16-pixel border overlap (see Figure 6 right column). After the EMR-CNN inferencing stage, partial objects that lie on the overlapping boundaries of each partition are reconstructed based on their overlapping regions. Specifically, two objects on the partition boundaries that overlap by more than 10 pixels are merged into one object. Figure 6 middle column demonstrates how a naïve divide-and-conquer inferencing that does not utilize overlapping sub-volumes will result in segmentation errors at the border of the inference windows.

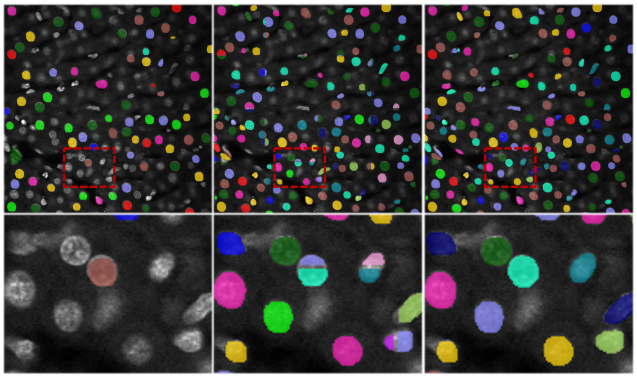

Figure 6. Segmentation results of EMR-CNN+AHC (M=4) on large microscopy $\mathcal{D}_2$. Left column: direct inference on each slice. Middle column: naïve segmentation based on non-overlapping partitions. Right column: results using our proposed divide-and-conquer inference method.

4.2. Quantitative Evaluation

To evaluate the accuracy of our method, we define that a ground truth nucleus is successfully detected if the Intersection-over-Union (IoU) between this ground truth nucleus and a detected nucleus is greater than an IoU threshold t. Then we count the True Positive (TP) detections (number of ground truth nuclei that are successfully detected), the False Positive (FP) detections (number of objects that are falsely detected as nuclei), and the False Negative (FN) detections (number of ground truth nuclei that are not detected). We use the F_1 score ($\text{F}_1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$) for object-based evaluation. In addition, we adopt the widely used Average Precision (AP), known as the PR curve [67], and the mean Average Precision (mAP) metrics used by the VOC PASCAL [68] and MSCOCO evaluation benchmarks [64].

Table 2. The object-based evaluation results using Average Precision (AP), Mean Average Precision (mAP), and mean F_1 scores (mF_1)

Methods	Microscopy $\mathcal{D}_1$				Microscopy $\mathcal{D}_2$				Microscopy $\mathcal{D}_3$			
	$AP_{.25}$	$AP_{.45}$	mAP	mF_1	$AP_{.25}$	$AP_{.45}$	mAP	mF_1	$AP_{.25}$	$AP_{.45}$	mAP	mF_1
3D Watershed	61.51	35.12	48.87	68.31	67.58	51.20	60.30	76.36	22.33	4.62	11.96	30.22
CellProfiler [25]	50.46	27.35	37.79	59.41	64.26	48.66	57.52	75.09	20.46	2.85	10.15	27.75
3D Squassh [20]	12.39	6.19	9.21	23.51	66.79	53.76	61.61	78.15	0.59	0.36	0.49	4.73
VTEA [17]	45.52	38.25	42.71	61.64	64.69	45.90	57.76	74.11	25.91	10.44	17.89	38.45
VNet [65]	77.31	61.78	70.29	82.41	53.67	34.97	45.88	64.61	37.03	12.58	23.74	41.73
3D U-Net [66]	75.56	62.97	69.63	81.81	66.89	48.14	58.12	74.62	36.89	12.40	25.19	43.39
DeepSynth [30]	81.57	75.59	79.51	87.19	76.42	51.86	66.19	79.42	34.98	10.91	22.83	41.77
Cellpose [45]	81.70	81.70	81.70	89.47	71.92	64.61	68.92	80.74	45.29	17.01	30.43	51.41
EMR-CNN+BS, M=8	82.31	69.60	76.64	86.44	77.81	70.31	75.50	85.68	53.18	33.93	45.53	64.90
EMR-CNN+AHC, M=8	**93.19**	**89.46**	**91.04**	**94.95**	**82.37**	**75.71**	**80.13**	**88.15**	**68.26**	**47.65**	**59.61**	**71.05**

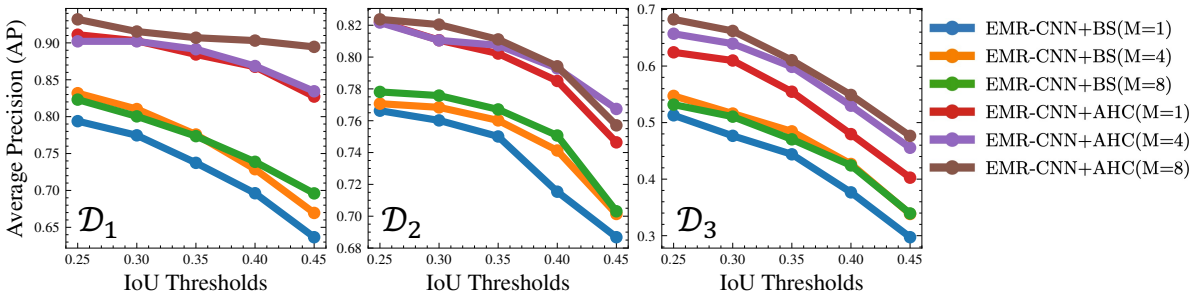

Figure 7. Average Precision (AP) under different IoU thresholds for microscopy $\mathcal{D}_1 - \mathcal{D}_3$. BS and AHC specifies the blob slice or agglomerative hierarchical clustering used for slice fusion module. M represents number of detectors in an ensemble.

To reduce evaluation bias given the variability of the segmentation results for different testing data, we use a moving average Intersection-over-Union (IoU) threshold ranging from 0.25 to 0.45 in 0.05 increments ($t \in T_{\text{IoUs}} = \{0.25, 0.3, ..., 0.45\}$). To this end, we define AP_t and $F_1(t)$ as the AP and F_1 score evaluated under IoU threshold t, respectively. We then define $\text{mAP} = \frac{1}{|T_{\text{IoUs}}|} \sum_{t \in T_{\text{IoUs}}} AP_t$ and $mF_1 = \frac{1}{|T_{\text{IoUs}}|} \sum_{t \in T_{\text{IoUs}}} F_1(t)$ as the mean AP and F_1 score among all the IoU thresholds, respectively. The evaluation results of our proposed technique and other compared methods on the actual datasets $\mathcal{D}_1 - \mathcal{D}_3$ are given in Table 2. Figure 7 shows the AP of EMR-CNN using BS and AHC (see Section 3.4) with a different number of detectors in an ensemble. All segmentation results and evaluation criteria were verified by a biologist.

4.3. Running Time Analysis

The running time of our approach and compared methods are given in Table 3. The times reported are based on the training and testing of $\mathcal{D}_2$. The compared deep learning methods were trained for 200 epochs on 4 TITAN XP GPUs using the same synthetic data for training EMR-CNN. Our approach takes significantly less training time because each Mask R-CNN model is trained with a subset of the entire training set for only 2000 iterations (2 images/iteration). EMR-CNN and Cellpose take longer time in inferencing since they have to run multiple batches for one volume. The processing of EMR-CNN includes ensemble 2D fusion and 2D to 3D slice fusion. The processing for VNet, UNet, and DeepSynth includes 3D watershed and morphological operations to split touching nuclei.

Table 3. Running time analysis. Training: total training time (hours), Inference: model inference time (seconds/volume), Processing: Pre-and Post-processing time (seconds/volume).

Methods	Microscopy $\mathcal{D}_2$		
	Training	Inference	Processing
CellProfiler	-	-	30.00 s
Squassh	-	-	30.00 s
3D Watershed	-	-	4.90 s
VTEA	-	-	2.00 s
VNet	3.07 h	0.24 s	3.11 s
UNet	5.13 h	0.29 s	3.24 s
DeepSynth	2.53 h	0.19 s	3.33 s
Cellpose	4.63 h	4.48 s	0.47 s
EMR-CNN+BS, M=8	0.20 h	1.01 s	1.38 s
EMR-CNN+AHC, M=8	0.20 h	1.01 s	1.68 s

4.4. Robustness Analysis

Due to the randomly initialized networks and randomly sampled training set, each detector in the ensemble may produce different False Positive segmentation results for a given image slice. In order to test the stability and robustness

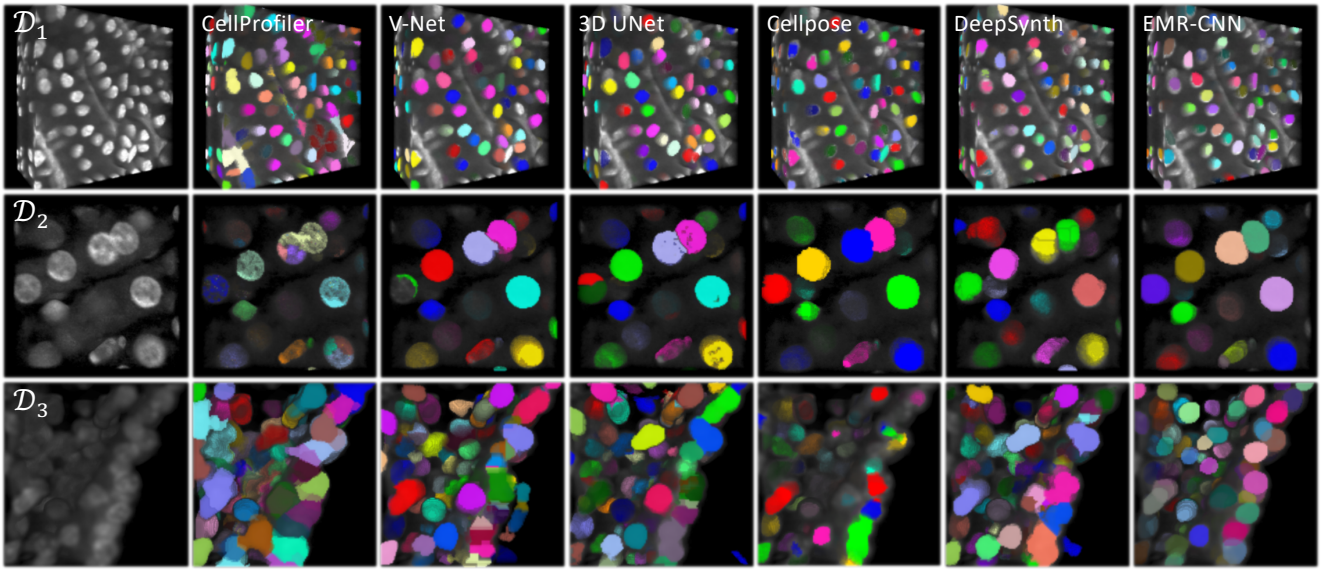

Figure 8. Example testing sub-volumes from original microscopy images and the nuclei instance segmentation results for compared methods. The last column is the segmentation results of our EMR-CNN + AHC, M=8

of our system, we conducted two experiments. In the first experiment, we randomly add to each detection Det^{m_i,z_j}, a different number N (ranging from 1 to 7) of false positive segmentations of a round shape mask with radius $R = 4$ having a random confidence score between 0.7 and 1.0 (see Figure 9 (left)). In the second experiment, we randomly add $N = 2$ false positive segmentations with different radii R ranging from 2 to 8 (see Figure 9 (right)). As the figure indicates, the false positive outliers were removed by our weighted 2D mask fusion with only small losses incurred in the accuracy metrics.

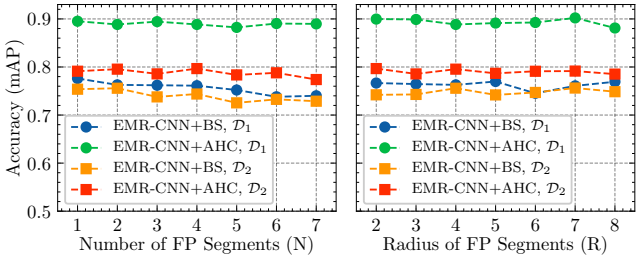

Figure 9. Robustness analysis of proposed method by randomly adding false positive segmentations to each detected image slice

4.5. Discussion

Our proposed method was compared with both traditional methods and deep learning-based methods. The traditional methods such as 3D Watershed, CellProfiler, 3D Squash, and VTEA require tuning of many parameters and may work well with one dataset but fail for others. Deep learning-based techniques such as 3D U-Net and DeepSynth are all 3D CNN-based models that directly work on 3D data, whereas Cellpose uses a 2D to 3D reconstruction. In contrast, for the proposed EMR-CNN technique, which is based on Mask R-CNN [54], we utilized two different 2D to 3D slice fusion approaches: Blob-Slice (BS) [57, 17] and Agglomerative Hierarchical Clustering (AHC) [23], with $M = 1, 4, 8$ detectors in an ensemble. As shown in Table 2, the proposed EMR-CNN + AHC approach outperforms the other techniques, based on the AP, mAP, and mF_1 metrics. As expected, the ensemble with $M = 8$ detectors achieves higher accuracy than $M = 4$ and $M = 1$ for both BS and AHC strategies.

5. CONCLUSION

In this paper, we described an ensemble learning and slice fusion strategy for 3D nuclei segmentation. The proposed method uses a weighted 2D mask fusion technique to fuse 2D segmentation masks from different object detectors as well as unsupervised clustering for combining 2D segmentation masks into a 3D segmentation mask. The evaluation results indicate that our approach is stable and robust in the presence of false detections or outliers as well as outperforms some recent 3D CNN-based methods. Moreover, our method does not need any ground truth annotations for training and can inference on any large size volumes. Code has been made available at: `http://skynet.ecn.purdue.edu/~micro/emrcnn/emrcnn_release.zip`

6. Acknowledgments

This work was partially supported by a George M. OBrien Award from the National Institutes of Health under grant NIH/NIDDK P30 DK079312 and the endowment of the Charles William Harrison Distinguished Professorship at Purdue University. The authors have no conflicts of interest.

References

[1] M. W. Davidson and M. Abramowitz, "Optical microscopy," *Encyclopedia of imaging science and technology*, vol. 2, no. 1106-1141, p. 120, 2002. 1

[2] R. H. Webb, "Confocal optical microscopy," *Reports on progress in physics*, vol. 59, no. 3, p. 427, March 1996. 1

[3] W. Denk, J. H. Strickler, and W. W. Webb, "Two-photon laser scanning fluorescence microscopy," *Science*, vol. 248, no. 4951, pp. 73–76, April 1990. 1

[4] K. Dunn, R. Sandoval, K. Kelly, P. Dagher, G. Tanner, S. Atkinson, R. Bacallao, and B. Molitoris, "Functional studies of the kidney of living animals using multicolor two-photon microscopy," *American Journal of Physiology-Cell Physiology*, vol. 283, no. 3, pp. C905–C916, September 2002. 1

[5] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized medical imaging and graphics*, vol. 31, no. 4-5, pp. 198–211, March 2007. 1

[6] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, February 2018. 1

[7] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Medical image analysis*, vol. 33, pp. 170–175, July 2016. 1

[8] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4550–4568, October 2018. 1

[9] S. Wolf, Y. Li, C. Pape, A. Bailoni, A. Kreshuk, and F. A. Hamprecht, "The semantic mutex watershed for efficient bottom-up semantic instance segmentation," *European Conference on computer vision*, pp. 208–224, August 2020, Glasgow, Scotland. 1

[10] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5221–5229, July 2017, Honolulu, Hawaii. 1

[11] Y. Cao, T. A. Geddes, J. Y. H. Yang, and P. Yang, "Ensemble deep learning in bioinformatics," *Nature Machine Intelligence*, vol. 2, no. 9, pp. 500–508, August 2020. 1

[12] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1016–1025, April 2019. 1

[13] K. Y. Win, S. Choomchuay, K. Hamamoto, and M. Raveesunthornkiat, "Comparative study on automated cell nuclei segmentation methods for cytology pleural effusion images," *Journal of healthcare engineering*, vol. 2018, September 2018. 2

[14] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979. 2

[15] S. Beucher, "The watershed transformation applied to image segmentation," *Scanning Microscopy*, vol. 1992, no. 6, p. 28, 1992. 2

[16] J. M. Sharif, M. F. Miswan, M. A. Ngadi, M. S. H. Salam, and M. M. A. Jamil, "Red blood cell segmentation using masking and watershed algorithm: A preliminary study," *Proceedings of the International Conference on Biomedical Engineering*, pp. 258–262, 2012. 2

[17] S. Winfree, S. Khan, R. Micanovic, M. T. Eadon, K. J. Kelly, T. A. Sutton, C. L. Phillips, K. W. Dunn, and T. M. El-Achkar, "Quantitative three-dimensional tissue cytometry to study kidney tissue and resident immune cells," *Journal of the American Society of Nephrology*, vol. 28, no. 7, pp. 2108–2118, July 2017. 2, 5, 7, 8

[18] N. M. Sobhy, N. M. Salem, and M. E. Dosoky, "A comparative study of white blood cells segmentation using otsu threshold and watershed transformation," *Journal of Biomedical Engineering and Medical Imaging*, vol. 3, no. 3, p. 15, July 2016. 2

[19] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, Feburary 2001. 2

[20] A. Rizk, G. Paul, P. Incardona, M. Bugarski, M. Mansouri, A. Niemann, U. Ziegler, P. Berger, and I. F. Sbalzarini, "Segmentation and quantification of subcellular structures in fluorescence microscopy images using squassh," *Nature Protocols*, vol. 9, no. 3, pp. 586–596, March 2014. 2, 7

[21] O. Sarrafzadeh and A. M. Dehnavi, "Nucleus and cytoplasm segmentation in microscopic images using k-means clustering and region growing," *Advanced biomedical research*, vol. 4, August 2015. 2

[22] M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 233–241, March 2011. 2

[23] L. Wu, S. Han, A. Chen, P. Salama, K. W. Dunn, and E. J. Delp, "RCNN-SliceNet: A Slice and Cluster Approach for Nuclei Centroid Detection in Three-Dimensional Fluorescence Microscopy Images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3750–3760, June 2021, Nashville, TN. 2, 6, 8

[24] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2405–2414, November 2006. 2

[25] C. McQuin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, W. Wiegraebe, S. Singh, T. Becker, J. C. Caicedo, and A. E. Carpenter, "Cellprofiler 3.0: Next-generation image processing for biology," *PLoS biology*, vol. 16, no. 7, pp. e2 005 970–1–17, July 2018. 2, 7

[26] C. S. Bjornsson, G. Lin, Y. Al-Kofahi, A. Narayanaswamy, K. L. Smith, W. Shain, and B. Roysam, "Associative image analysis: a method for automated quantification of 3D multi-parameter images of brain tissue," *Journal of Neuroscience Methods*, vol. 170, no. 1, pp. 165–178, May 2008. 2

[27] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. V. Valen, "Deep learning for cellular image analysis," *Nature methods*, vol. 16, no. 12, pp. 1233–1246, May 2019. 2

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, pp. 231–241, November 2015, Munich, Germany. 2

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, December 2017. 2

[30] K. W. Dunn, C. Fu, D. J. Ho, S. Lee, S. Han, P. Salama, and E. J. Delp, "Deepsynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data," *Scientific Reports*, vol. 9, no. 1, pp. 18 295–18 309, December 2019. 2, 7

[31] C. Fu, S. Lee, D. J. Ho, S. Han, P. Salama, K. W. Dunn, and E. J. Delp, "Three dimensional fluorescence microscopy image synthesis and segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2302–2310, June 2018, Salt Lake City, UT. 2, 3, 4, 6

[32] D. J. Ho, C. Fu, P. Salama, K. W. Dunn, and E. J. Delp, "Nuclei segmentation of fluorescence microscopy images using three dimensional convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 834–842, June 2017. 2, 3

[33] L. Wu, "Biomedical image segmentation and object detection using deep convolutional neural networks," M.S. dissertation, Purdue University, Hammond, IN, May 2019. 2

[34] Q. D. Vu, S. Graham, T. Kurc, M. N. N. To, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, J. K. Cramer, T. Zhao, *et al.*, "Methods for segmentation and classification of digital microscopy tissue images," *Frontiers in bioengineering and biotechnology*, vol. 7, p. 53, April 2019. 2

[35] K. Dijkstra, J. van de Loosdrecht, L. R. B. Schomaker, and M. A. Wiering, "Centroidnet: A deep neural network for joint object localization and counting," *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 585–601, September 2018, Dublin, Ireland. 2

[36] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical Image Analysis*, vol. 58, p. 101563, December 2019. 2

[37] D. J. Ho, D. M. Montserrat, C. Fu, P. Salama, K. W. Dunn, and E. J. Delp, "Sphere estimation network: three-dimensional nuclei detection of fluorescence microscopy images," *Journal of Medical Imaging*, vol. 7, no. 4, pp. 1 – 16, August 2020. 2

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017. 2

[39] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018. 2

[40] P. Zhang, X. Li, and Y. Zhong, "Ensemble mask-aided r-cnn," *Proceedings of the 2019 Challenge on Endoscopy Artefacts Detection*, pp. 6154–6162, April 2019, Venice, Italy. 2

[41] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang, "Hierarchical retinal blood vessel segmentation based on feature and ensemble learning," *Neurocomputing*, vol. 149, pp. 708–717, 2015. 2

[42] P. Tang, X. Yang, Y. Nan, S. Xiang, and Q. Liang, "Feature pyramid nonlocal network with transform modal ensemble learning for breast tumor segmentation in ultrasound images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 12, pp. 3549–3559, July 2021. 2

[43] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, January 2019. 2

[44] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104117, March 2021. 2

[45] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, October 2021. 2, 7

[46] C. Fu, D. J. Ho, S. Han, P. Salama, K. W. Dunn, and E. J. Delp, "Nuclei segmentation of fluorescence microscopy images using convolutional neural networks," *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pp. 704–708, April 2017, Melbourne, Australia. 2

[47] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis." *Proceedings of the International Conference on Document Analysis and Recognition*, vol. 3, no. 2003, 2003. 2

[48] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019. 2

[49] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, January 2018. 2

[50] A. Chen, L. Wu, S. Han, P. Salama, K. W. Dunn, and E. J. Delp, "Three dimensional synthetic non-ellipsoidal nuclei volume generation using bezier curves," *Proceedings of the IEEE International Symposium on Biomedical Imaging*, April 2021, Nice, France. 2, 3

[51] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2242–2251, October 2017, Venice, Italy. 2, 3, 4

[52] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 958–963, August 2003, Edinburgh, UK. 3

[53] S. McKinley and M. Levine, "Cubic spline interpolation," *College of the Redwoods*, vol. 45, no. 1, pp. 1049–1060, 1998. 4

[54] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019. 4, 6, 8

[55] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, pp. 104 117–104 122, March 2021. 4

[56] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion?" *Journal of Classification*, vol. 31, no. 3, pp. 274–295, October 2014. 4

[57] A. Santella, Z. Du, S. Nowotschin, A. K. Hadjantonakis, and Z. Bao, "A hybrid blob-slice model for accurate and efficient detection of fluorescence labeled nuclei in 3d," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–13, November 2010. 5, 8

[58] "Indiana Center for Biological Microscopy." [Online]. Available: http://web.medicine.iupui.edu/icbm/ 5

[59] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, July 2006. 5

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016, Venice, Italy. 6

[61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, July 2017, Honolulu, HI. 6

[62] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, July 2017, Honolulu, HI. 6

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, December 2012. 6

[64] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," *European conference on computer vision*, pp. 740–755, September 2014, Zurich, Switzerland. 6

[65] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *International Conference on 3D Vision*, pp. 565–571, October 2016. 7

[66] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D u-net: Learning dense volumetric segmentation from sparse annotation," *Medical Image Computing and Computer-Assisted Intervention*, vol. 9901, pp. 424–432, October 2016. 7

[67] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," *Proceedings of the international conference on Machine learning*, pp. 233–240, June 2006, Pittsburgh, PA. 6

[68] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, January 2015. 6