Automated Semantic Segmentation of Red Blood Cells for Sickle Cell Disease

Mo Zhang*, Student Member, IEEE, Xiang Li*, Member, IEEE, Mengjia Xu*,
and Quanzheng Li, Member, IEEE *Joint first authors

***Abstract*—Red blood cell (RBC) segmentation and classification from microscopic images is a crucial step for the diagnosis of sickle cell disease (SCD). In this work, we adopt a deep learning based semantic segmentation framework to solve the RBC classification task. A major challenge for robust segmentation and classification is the large variations on the size, shape and viewpoint of the cells, combining with the low image quality caused by noise and artifacts. To address these challenges, we apply deformable convolution layers to the classic U-Net structure and implement the deformable U-Net (dU-Net). U-Net architecture has been shown to offer accurate localization for image semantic segmentation. Moreover, deformable convolution enables free-form deformation of the feature learning process, thus making the network more robust to various cell morphologies and image settings. dU-Net is tested on microscopic red blood cell images from patients with sickle cell disease. Results show that dU-Net can achieve highest accuracy for both binary segmentation and multi-class semantic segmentation tasks, comparing with both unsupervised and state-of-the-art deep learning based supervised segmentation methods. Through detailed investigation of the segmentation results, we further conclude that the performance improvement is mainly caused by the deformable convolution layer, which has better ability to separate the touching cells, discriminate the background noise and predict correct cell shapes without any shape priors.**

***Index Terms*—RBC, automated semantic segmentation, sickle cell disease, U-Net, deformable convolution**

I. INTRODUCTION

SICKLE cell disease (SCD) is an inherited blood disorder, where SCD patients have abnormal hemoglobin that can cause normal disc-shaped red blood cells (RBCs) to distort and generate heterogeneous shapes. Differences in cell morphology between healthy and pathological cells make it possible to perform image-based diagnosis, which is very important for faster and more accurate diagnosis of potential SCD. Image-based analysis of SCD is capable of performing highly specific and sensitive sickle and normal erythrocyte classification through cell shape statistics [1]. Most of the works reported in previous literatures are based on multi-stage workflow, including steps such as image preprocessing, cell segmentation, feature extraction, and single cell classification using various machine learning models [2-5]. In recent years, Convolutional Neural Network (CNN) has received increasing attention in the field of computer vision, as it can automatically learn the low-to-high level features from images, providing more robust and generalized representation of objects comparing with handcrafted features. CNN has also been applied for image analysis of SCD RBCs, which follows the similar multi-stage workflow: constant-sized single cell patches are extracted from raw cellular specimen images and fed into the network for obtaining its label (i.e. cell type) [6, 7]. For extracting cell patches, manual segmentation [8] and ground truth bounding box methods [9] have been applied. Dong et al. applied three well-known CNN models of LeNet-5, AlexNet, and GoogLeNet on simulation studies for malaria-infected cell classification [10]. In [11], Xu et al. applied a 10-layer CNN on single cell patches with size normalization for SCD classification and achieved a mean accuracy of 89%.

In practice, there exists multiple challenges in automatic RBC segmentation and classification, including: 1) There are many stains and artifacts spreading on the microscopic images, some of them share similar features with cells. 2) Some cells are partially or entirely blurred. At the same time, touching and overlapping cells can be commonly found in the images. The combined effect from these two makes it difficult to perform cell recognition even for humans. 3) Pixel-wise labels are highly unbalanced. In the current dataset, pixel proportion between the background and four RBC types is roughly 240:11:2:1:1. 4) Large inter-patient variation exits on global image conditions such as illumination and color hue. 5) Large inter-cell variation exists for the same type of RBC in cellular morphology such as size, shape, texture, and pose.

While some of the challenges such as touching and overlapping cells have been discussed and addressed in

This paper was submitted on November 25, 2019. Quanzheng Li is supported in part by the National Institutes of Health under Grant R01AG052653. This work is supported by the computational resources of MGH/BWH Center for Clinical Data Science. We gratefully acknowledge the help of algorithm development from the Data Integration, Visualization, and Exploration (DIVE) Laboratory of Department of Computer Science & Engineering, Texas A&M University.

Mo Zhang is with the Center for Data Science in Health and Medicine, Center for Data Science, Peking University, Beijing 100871, China; and Laboratory for Biomedical Image Analysis, Beijing Institute of Big Data Research, Beijing 100871, China (e-mail: zhangmo007@pku.edu.cn).

Xiang Li and Quanzheng Li are with the MGH/BWH Center for Clinical Data Science, Boston, MA 02115, USA (e-mail: xli60@mgh.harvard.edu; li.quanzheng@mgh.harvard.edu).

Mengjia Xu is with the Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China (e-mail: mengjia_xu1@hotmail.com).

previous cell segmentation works [12], fully automatic and accurate segmentation of cells is still an unsolved problem. Further, for multi-stage cell segmentation and classification models, the performance of classification relies heavily on the previous steps such as image preprocessing and cell segmentation. One problem of such decoupled framework is the decreased robustness to different image settings, as parameter tuning is usually needed for preprocessing and segmentation steps. Also, overlapping and touching cells that were not accurately identified in the segmentation results will not be useful (and sometimes harmful) for later classification models. For supervised classification, those failed segmentation cases have to be abandoned from the training set [11].

To address the above challenges, we formulate the problem of SCD diagnosis based on RBC images as an end-to-end, pixel-wise semantic segmentation task. Semantic segmentation is of particular interest in the field of computer vision, as it can make dense prediction for meaningful comprehending of scenes. In recent years deep learning models have been widely used for semantic segmentation [13, 14]. Among deep learning models, Fully Convolutional Network (FCN) [15], which replaces fully connected layers with convolutional layers, has the capability of obtaining a prediction map with the same size as the input image, making it particularly suitable for semantic segmentation. In the field of biomedical image analysis, a popular FCN-based network structure is U-Net [16] which adds skip connection for more accurate localization.

In this work, we extend the U-Net architecture by adding deformable convolution [17] to implement the deformable U-Net (dU-Net) model for RBC semantic segmentation, specifically to address the challenge of inter-cell variation and needs for spatial invariance. It has been shown in the literature that deformable convolution is more robust, as it can accommodate geometric variations by learning an adaptive, data-driven receptive field [17]. dU-Net is trained and tested on a multi-institutional RBC microscopic image database consisting of both healthy and pathological populations. We perform both binary segmentation (i.e. cell detection) and multi-class semantic segmentation experiments, then evaluate the performance of dU-Net using multiple metrics. We then compare the performance of dU-Net with both unsupervised (region growing, Ilastik) and supervised state-of-the-art image segmentation methods (U-Net, PSPNet, DeeplabV3+).

The paper is organized as follows. Section II introduces the related work and Section III describes the proposed method in detail. In Section IV, experimental settings are presented, including data description, experimental setup and evaluation criteria. In Section V, we evaluate the proposed method through three kinds of experiments. Then, we discuss the methods and conclude the paper in Section IV.

II. RELATED WORK

A. Solutions on Cell Segmentation

In the field of biomedical imaging, cell segmentation plays a critical role in quantitative analysis. Most of the cell segmentation methods can be divided into two categories: unsupervised and supervised. For unsupervised cell segmentation, classical methods including region growing [18], watershed transformation [19], active contours [20], gradient flow [21] as well as semi-automatic methods [22] were used. For supervised cell segmentation, deep learning methods such as CNN and FCN have been applied [23-25], which have outperformed other supervised segmentation methods.

CNN has been used to discriminate different blood cells by using patches centered at each pixel as the input [26]. Although this work can obtain a semantic segmentation map by aggregating labels of each pixel, the sliding window approach for patch extraction lowers the resolution of segmentation (limited by stride of extraction) and susceptible to patch size selection. Yuexiang Li et al. utilized a FCN-based model to perform HEp-2 cell semantic segmentation, yet their focus was the whole specimen image classification by assigning label of the largest population to an entire image [24]. More recently, some works adopted various FCN-based frameworks to deal with blood cell segmentation, such as U-Net [27], SegNet [28,29] and FCN-AlexNet [30]. However, the issue of large variations in cellular morphology still remains to be solved, making it necessary to apply deformable convolutions.

B. Solutions for Spatial Invariance

For tasks in computer vision, one of the core challenges is the presence of enormous geometric transformations and spatial variations in object pose, shape, and scale. In addition to the data augmentation techniques which enriches the training set with those variations [31], transformation-invariant representations such as SIFT (scale invariant feature transform) [32] are also used to incorporate the deformations. Spatial Transformer Networks (STN) proposed by Jaderberg et al., provide a learnable module within the network to achieve spatial invariance [33]. Based on STN, Li et al. recently proposed Dense Transformer Networks (DTN), which restores spatial correspondence between inputs and outputs through a dense transformation [34]. For biomedical image analysis, STN has been employed in [35] for cell differentiation by deriving cell localization information. However, STN uses a global parametric transformation which only works on the whole feature map, making it unable to capture more sophisticated local information needed by dense prediction. Instead, deformable convolution [17] enhances spatial invariance by reforming the fixed receptive field of traditional convolution unit. Deformable convolution has been applied on biomedical image analysis as well, including our preliminary work [36] and the subsequent related work [37]. To perform blood vessel segmentation, the method in [37] utilized a portion of deformable convolutional layers in the U-Net architecture, which deals with a simple, foreground/background segmentation task. However, in this work, our model replaces all the convolutional layers in U-Net with deformable convolutions, providing multi-class, semantic segmentation results directly from the input RBC image.

III. METHODOLOGY

In this work we propose deformable U-Net (dU-Net) which

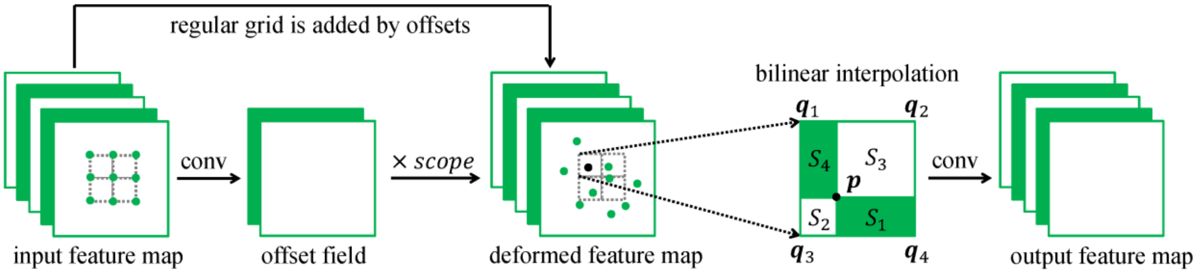

Fig. 1. Illustration of deformable convolution, showing how the square sampling locations are adaptively deformed into irregular shape. In the figure "conv" represents standard convolution kernel, and parameter *scope* is used to regulate the scale of the sampling region.

replaces regular convolution with deformable convolution throughout the U-Net structure, in order to overcome the limitation of regular square receptive field thus enhancing the network capability of dealing with object shape transformations. Traditional convolution kernel is defined with fixed shape and size to sample the input feature map on a regular grid. For example, grid $\mathcal{R}$ for a 3×3 convolution kernel is $\mathcal{R}=\{(-1,-1),(-1,0),\cdots,(1,0),(1,1)\}$. For each pixel $\boldsymbol{p}_0$ on the output feature map $\boldsymbol{y}$, the convolution operation can be expressed as:

$$\boldsymbol{y}(\boldsymbol{p}_0)=\sum_{\boldsymbol{p}_n\in\mathcal{R}}\boldsymbol{w}(\boldsymbol{p}_n)\cdot\boldsymbol{x}(\boldsymbol{p}_0+\boldsymbol{p}_n)\,, \tag{1}$$

where $\boldsymbol{y}(\boldsymbol{p}_0)$ denotes the value of pixel $\boldsymbol{p}_0$ on the output feature map, $\boldsymbol{x}(\boldsymbol{p}_0+\boldsymbol{p}_n)$ denotes the value of pixel $\boldsymbol{p}_0+\boldsymbol{p}_\mathrm{n}$ on the input feature map, and $\boldsymbol{w}(\boldsymbol{p}_n)[n=1,2,\ldots,9]$ are weight parameters of the 3×3 kernel.

In contrast, deformable convolution adds extra offsets $\Delta\boldsymbol{p}_n[n=1,2,\ldots,9]$ (for a 3×3 kernel) to the regular sampling grid $\mathcal{R}$, thus Eq. (1) becomes:

$$\mathbf{y}(\boldsymbol{p}_0)=\sum_{\boldsymbol{p}_n\in\mathcal{R}}\boldsymbol{w}(\boldsymbol{p}_n)\cdot\boldsymbol{x}(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n)\,. \tag{2}$$

Offsets $\Delta\boldsymbol{p}_n$ (each $\boldsymbol{p}_0$ has 9 corresponding offsets) are learned from data and used to adjust $\boldsymbol{p}_n$, making the sampling grid more suitable for specific task comparing to a uniform square kernel.

As offset $\Delta\boldsymbol{p}_n$ is typically fractional, coordinate $\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n$ may not lie exactly on the input regular grid, in such case, the value of fractional coordinate needs to be interpolated from integer coordinates on the input grid. The term $\boldsymbol{x}(\boldsymbol{p}_0+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n)$ in Eq. (2) is implemented by bilinear interpolation:

$$\boldsymbol{x}(\boldsymbol{p})=\sum_{\boldsymbol{q}}f(q_x,p_x)\cdot f(q_y,p_y)\cdot\boldsymbol{x}(\boldsymbol{q}), \tag{3}$$

where $\boldsymbol{p}=\boldsymbol{p_0}+\boldsymbol{p}_n+\Delta\boldsymbol{p}_n$ enumerates an arbitrary fractional location on the input feature map, $\boldsymbol{q}$ denotes all integer locations on the input feature map, p_x (p_y) denotes x and y-coordinate of $\boldsymbol{p}$, the function f is defined as:

$$f(q,p)=max(0,1-|q-p|). \tag{4}$$

From the definition it can be seen that $\boldsymbol{x}(\boldsymbol{p})$ is only related with the four integer coordinates $\boldsymbol{q}_i[i=1,2,3,4]$ adjacent to $\boldsymbol{p}$, as the interpolation kernel $f(q_x,p_x)\cdot f(q_y,p_y)$ assigns 0 for other pixels. A sample illustration for bilinear interpolation is shown in Fig. 1, where the black dot is the fractional location $\boldsymbol{p}$ and $S_i[i=1,2,3,4]$ are the areas of rectangles generated by the four nearest integer coordinates $\boldsymbol{q}_i[i=1,2,3,4]$. Thus Eq. (3) is equivalent to Eq. (5), based on the concept of "area of rectangle":

$$\boldsymbol{x}(\boldsymbol{p})=\sum_{i=1}^{4}\boldsymbol{x}(\boldsymbol{q}_i)\cdot S_i\,, \tag{5}$$

and more references can be found in [33] and [34].

As shown in Fig. 1, the detailed procedure of deformable convolution starts with an additional classic convolution with activation function TANH to learn offset field from the input feature map, which are then normalized to $[-1,1]$. The offset field has the same height and width with the input feature map while its number of channels is $2N$ $(N=|\mathcal{R}|)$. Second, the offset field is multiplied by parameter $scope$ which adjusts the scale of receptive field and then added to the regular grid $\mathcal{R}$ to obtain new sampling locations (each coordinate on the offset field has N pairs of offsets corresponding to regular grid $\mathcal{R}$). Finally, values of the irregular sampling coordinates are computed via bilinear interpolation as in Eq. (3) to obtain the deformed feature map. Standard convolution is then applied on the deformed feature map to get the output feature map. Through learning the offset field, deformable convolution can sample the input feature map in a more flexible and dense way, thus making it more adaptive to geometric transformations in object shape and scale [17].

The main architecture of dU-Net is shown in Fig. 2, consisting of the encoder path and the decoder path. In the encoder path, each layer contains two 3×3 deformable convolutions followed by 2×2 max pooling operation with a stride of 2, which doubles the number of channels and halves the resolution of the input feature map for down-sampling. The encoder ends with two 3×3 deformable convolutions called bottom layers. In the right decoder path, each layer contains one 3×3 deconvolution followed by two 3×3 deformable convolutions, which halves the channel number and doubles the resolution of the input feature map for up-sampling, except for the last layer where the label map is predicted. Both the encoder and decoder paths contain three layers, and the skip connection between encoder and decoder path helps preserving contextual information for better localization [16]. Implementation details are illustrated in Supplementary I.

IV. EXPERIMENTAL METHODS

A. *Data Description*

dU-Net is tested on the latest public SCD RBC image dataset [11]. We use 314 raw microscopy images from 5 different SCD

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3000484, IEEE Journal of Biomedical and Health Informatics

patients for experiments, and a total number of around 3000 cells are involved. The original blood sample is collected from UPMC (University of Pittsburgh Medical Center) and MGH (Massachusetts General Hospital). Raw images (resolution is 1920×1080) are preprocessed by removing the blank margins on left and right sides and resizing to the same size of 512×512.

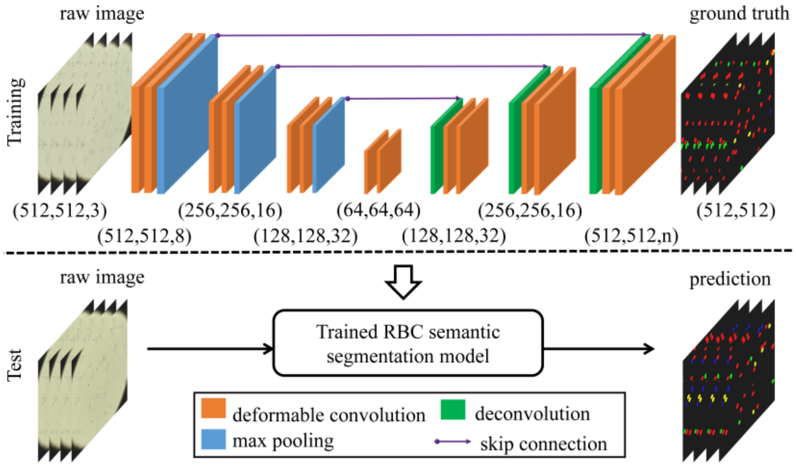

Fig. 2. Architecture of the dU-Net. Dimensions (height, width, channel) of the feature maps in each layer are shown beneath the networks. dU-Net takes the raw images of size (512, 512, 3) as input and generates prediction maps of the same resolution. In the first layer, the number of channels is set to 8. In the last layer, *n* is the number of classes for the semantic segmentation task.

In our previous work [11], eight SCD RBC categories are defined: Discocytes (Dic), Oval (Ovl), Reticulocytes (Ret), Elongated (El), Stomatocyte (Sto), Echinocytes (Ech), Granular (Grl) and Sickle (Sk). Based on these definitions, we merge some of RBC patterns according to their image characteristics such as shape and texture, resulting in four RBC categories: 1) Dic+Ovl, 2) El+Sk, 3) Grl, and 4) Ret+Sto+Ech. Fig. 3 illustrates some sample images in each category, showing that major difference among the first three categories is shape, and their internal textures are nearly indistinguishable. In contrast, RBCs of the forth category have relatively messy texture with diverse shape appearances. Note that the integration of categories is reasonable and meaningful for clinical diagnosis, as the pathological changes of RBC are staged. For example, Dic+Ovl represent the relatively healthy RBCs in the early stage, while El+Sk are sick RBCs in the most severe cases. Additionally, RBC regions and their labels (as four categories) are manually annotated by the data provider.

B. Experimental Setup

The performance of dU-Net is evaluated based on the following three experiment settings:

1. Binary RBC segmentation: differentiating RBC from background. We employ 5-fold cross validation to obtain a reliable model evaluation. The original 314 RBC images are randomly partitioned into 5 subsets with size of 63,63,63,63 and 62 respectively. During each experiment, we use four subsets to train the model and the remaining subset for testing. Also, we compare the performance of dU-Net with Ilastik software, region growing, U-Net, PSPNet [38] and DeeplabV3+ [39].

2. Multi-class RBC semantic segmentation: differentiating the four sub-types of SCD RBC as previously defined, as well as detecting cells from background. We employ the same 5-fold cross validation scheme.

3. Binary RBC segmentation under different data sizes: The setting of this experiment is to investigate the impact of data size on the network performance. Consequently, we perform six experiments sharing the same testing data (containing 64 images), while their training data contains 10, 20, 40, 80, 160, and 250 samples respectively.

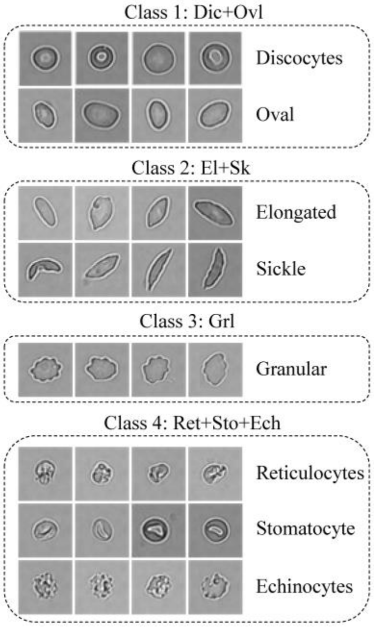

Fig. 3. Definition of the four categories of RBC. Note that these RBC examples are from the previous work [11]. Compared to class 1, the other three types of RBCs have more variations in shape and texture, which makes the detection/classification more complicated.

C. Evaluation Criteria

Pixel-level evaluation: three types of indices including Dice Coefficient, Jaccard Index, and Hausdorff Distance are calculated at pixel level. The detailed implications of these indices are explained in Supplementary II.

Cell-level Evaluation: We define three indices at cell level for performance evaluation: Error I, Error II, and Error III by manually checking the prediction results. To ensure the credibility of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3000484, IEEE Journal of Biomedical and Health Informatics

manually counting, each sample is handled by three experts and finally we take the average for result. For binary segmentation, Error I rate measures the number of cases where the model fails to separate touching cells, which is a common yet important challenge for tasks such as cell counting. Error II rate measures the number of artifacts which are recognized as cells by the model, as small dirty spots are commonly found in microscopic images. Error III rate measures the number of incomplete segmented cells in the prediction map, as it is desirable to maintain the object integrity. For multi-class semantic segmentation, the above cell-wise evaluation measurements work in the same way. Except for Error III, where in multi-class case it measures the number of cells that are identified to more than one classes (i.e. at least one third of the total area of a cell is labeled as a different class).

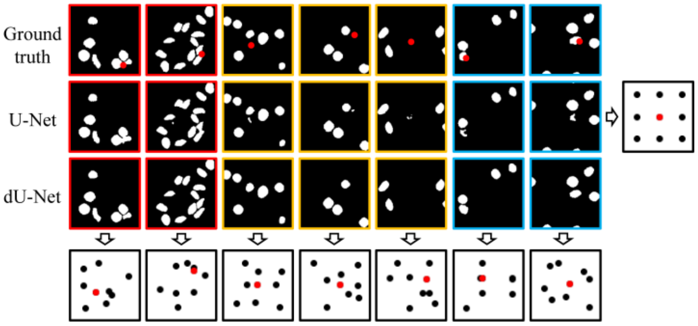

Fig. 4. Some patches of segmentation results of U-Net and dU-Net for binary RBC segmentation. Visualizations of sampling locations in both standard convolution (in U-Net) and deformable convolution (in dU-Net) are displayed. Center pixels in these kernels are colored as red. In red squares, dU-Net separates the touching cells perfectly; in yellow squares, dU-Net correctly identifies background noise as negative labels; and in blue squares, dU-Net maintains the integrity of segmented cells. The whole prediction maps of different methods are shown in Fig. 3 of Supplementary.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. *Binary RBC Segmentation*

Table I reports the pixel-level performance evaluations for the task of binary RBC segmentation using six methods (Ilastik, region growing, U-Net, PSPNet, DeeplabV3+ and dU-Net). These quantitative indicators are computed by averaging over the five experiments, and the index after "±" means the standard deviation on the testing set. Results show that the proposed dU-Net outperforms the other five approaches in all metrics, and deep learning models achieve superior performance compared with unsupervised methods. In Supplementary III, confusion matrices of U-Net and dU-Net are used to explore the advantage of dU-Net, and we find that dU-Net can produce more accurate predictions for cell boundaries.

TABLE I
PIXEL-LEVEL PERFORMANCE FOR BINARY SEGMENTATION.

Method	Dice	Jaccard	Hausdorff
Ilastik	0.652±0.004	0.486±0.003	110.133±5.013
Region Growing	0.663±0.009	0.500±0.009	123.150±3.578
U-Net	0.960±0.001	0.923±0.003	36.918±5.853
PSPNet	0.952±0.002	0.908±0.004	35.077±2.755
DeeplabV3+	0.964±0.001	0.932±0.002	41.149±6.833
dU-Net	**0.965±0.002**	**0.933±0.002**	**30.453±4.693**

TABLE II
CELL-LEVEL BINARY SEGMENTATION PERFORMANCE.

Method	Error I	Error II	Error III
U-Net	50.0±8.1	20.0±9.1	29.8±12.2
dU-Net	**31.8±3.6**	**6.0±2.1**	**13.4±2.4**

As we aim at analyzing the function of deformable convolution, cell-level performance is only evaluated for U-Net and dU-Net as listed in Table II. Results show that all three types of error (touching cell, false cell identification caused by noise/artifact and incomplete segmentation) can be reduced by using dU-Net (by 36%, 70% and 55% respectively). Lowered Error I rate indicates that dU-Net can separate the touching RBCs, and the sample cases can be found in Fig. 4 (red squares). Splitting of touching objects into individual instances is an important task and has been discussed in various literatures using unsupervised techniques [12]. While deep learning algorithms are in general less capable of dealing with such cases due to its uniform kernel over the whole feature map, dU-Net overcomes this limitation thanks to its deformable operations on convolutional filters, providing a feasible solution for the crowding problem in neural networks [40]. Lowered Error II rate indicates that dU-Net is more robust to background noise (e.g. dirties, halos, etc.) presented in microscopic images. As also visualized in Fig. 4 (yellow squares), U-Net mislabels background artifacts as cells while dU-Net makes the correct negative predictions. According to our observation, those artifacts usually have smaller size than real cells. Therefore, the performance difference can be possibly caused by the deformable kernel in dU-Net which learns more spatial features to capture the size information of objects. Finally, lowered Error III rate indicates that dU-Net is more capable of obtaining integrated cell-level predictions without any shape priors, which are also visualized in blue squares in Fig. 4. In the figure, the indistinct cells with low contrast to the background, as well as those cells located at the periphery of microscope visual field with low illumination can only be fully identified by dU-Net, where other methods fail to recover the entire cells. It is difficult to fully segment those cells with only texture information, which can also be observed from the incomplete cell segmentation results of U-Net. We thus conclude that dU-Net can incorporate more local structure/shape information to ensure the integration and smoothness of segmentation results.

In Fig. 4, we investigate the relationship between the segmentation performance and the sampling kernel, as the only difference between U-Net and dU-Net is the extra deformable kernel used in dU-Net. For U-Net, we visualize its fixed square sampling locations. For dU-Net, we visualize its data-dependent transformative sampling regions. It can be observed that deformable convolution has more flexible receptive fields with various deformed shapes learned from surroundings of the center pixel, allowing it to be more adaptive to complicated segmentation cases, such as cell boundary, touching cells, image artifacts and blurred regions.

TABLE III
PIXEL-LEVEL PERFORMANCE FOR MULTI-CLASS SEGMENTATION.

	Method	Background	Class 1	Class 2	Class 3	Class 4	Average
Dice	U-Net	0.996±0.001	0.867±0.010	0.547±0.047	0.469±0.058	0.620±0.040	0.700±0.012
	PSPNet	0.996±0.001	0.858±0.016	0.591±0.045	0.496±0.054	0.649±0.011	0.720±0.010
	DeeplabV3+	**0.997±0.000**	0.873±0.008	0.569±0.038	0.514±0.031	0.615±0.025	0.714±0.005
	dU-Net	0.997±0.001	**0.880±0.007**	**0.652±0.045**	**0.515±0.088**	**0.657±0.032**	**0.740±0.016**
Jaccard	U-Net	0.993±0.001	0.773±0.015	0.426±0.040	0.396±0.037	0.507±0.040	0.619±0.012
	PSPNet	0.991±0.002	0.759±0.023	0.470±0.045	0.413±0.056	0.537±0.016	0.637±0.010
	DeeplabV3+	**0.994±0.000**	0.782±0.012	0.443±0.029	0.429±0.031	0.502±0.024	0.630±0.005
	dU-Net	0.993±0.001	**0.791±0.010**	**0.536±0.039**	**0.446±0.060**	**0.551±0.031**	**0.664±0.011**
Hausdorff	U-Net	8.511±0.626	90.895±3.329	172.621±18.000	150.007±18.336	150.934±8.530	112.238±8.744
	PSPNet	7.744±0.712	87.287±2.899	153.079±20.213	131.534±18.880	**131.089±16.236**	100.917±5.514
	DeeplabV3+	**7.456±0.324**	91.408±4.881	175.874±19.370	**124.704±17.493**	145.161±14.071	108.921±8.458
	dU-Net	8.303±0.835	**85.907±1.861**	**152.595±17.132**	128.222±17.112	142.383±9.444	**100.250±7.867**

B. Multi-class RBC Semantic Segmentation

RBC semantic segmentation for 5 classes (4 classes of cells and background) is performed by dU-Net, U-Net, PSPNet and DeeplabV3+. Corresponding quantitative results are listed in Table III. dU-Net outperforms the other methods in the averaged accuracy of all the five classes. We observed that certain cells in SCD RBC dataset are small in sizes, thus their features cannot be well captured by ASPP module with large dilation rate (DeepLabV3+), nor by the pyramid pooling module (PSPNet). This observation can explain why dU-Net performs better, as the deformable convolution can preserve more fine details by choosing the optimal receptive field for the specific task. This is especially obvious for RBCs belonging to the categories 2-4, which have relatively larger variations in cell shape but smaller sample sizes. In practice, imbalanced data is common, where minor classes with small sample sizes can play vital roles for the diagnosis.

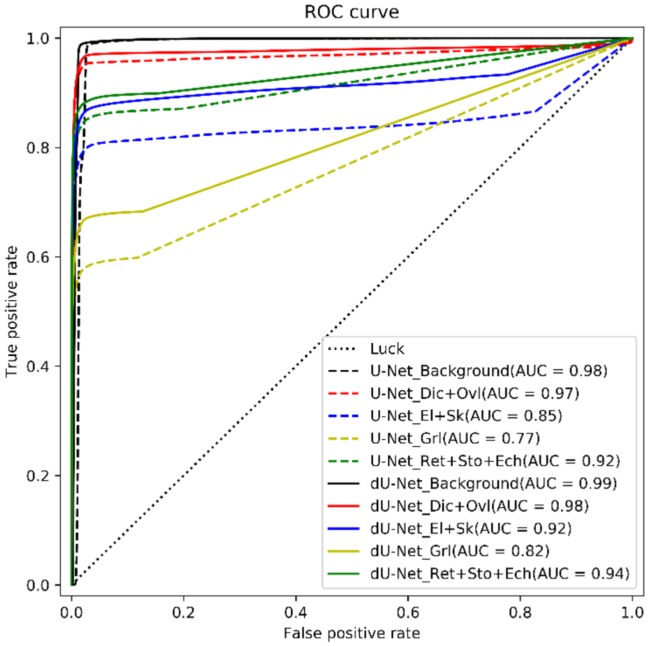

Fig. 5. ROC curves of multi-class semantic segmentation. Colored dashed curves represent results of U-Net, and colored solid curves represent results of dU-Net. Different colors represent different classes to segment.

In addition, ROC curve of the U-Net and dU-Net are shown in Fig. 5. All the AUC values of dU-Net are greater than 0.8, where U-Net obtains 0.77 AUC for segmenting class 3. Moreover, compared to U-Net, dU-Net achieves 7% and 5% AUC improvement for class 2 and class 3 respectively. By analyzing the confusion matrices (in Fig. 4 of Supplementary), we find that RBCs of classes 1-3 are difficult to differentiate due to their similar textures, but dU-Net makes more accurate predictions for these classes. As the major difference among these confusing classes is shape, it demonstrates that dU-Net is more capable of capturing more morphological variations and learning corresponding representative features without any priors , which accounts for its higher AUCs for class 2 and 3.

In Table IV we report the cell-level performance evaluation for multi-class RBC semantic segmentation using Error I, II and III defined previously. Results show that dU-Net is superior in predicting more integrated RBCs, while U-Net identifies more cells as multiple classes simultaneously. This can also be observed in Fig. 6 (highlighted in green blocks). Assigning multiple labels to the same instance is a common challenge for semantic segmentation [34] yet preserving cell integrity is an important task especially for later analyses such as cell counting.

TABLE IV
CELL-LEVEL MULTI-CLASS SEGMENTATION PERFORMANCE.

Method	Error I	Error II	Error III
U-Net	50.0±8.1	20.0±9.1	29.8±12.2
dU-Net	**31.8±3.6**	**6.0±2.1**	**13.4±2.4**

C. Binary RBC Segmentation using Different Sizes of Training Data

In this section, we train both U-Net and dU-Net with 6 different sizes of training data (10, 20, 40, 80, 160 and 250 images), and test the trained networks on the same testing data (64 images). Results are shown in Fig. 7, where dU-Net outperforms U-Net in every experiment measured in Dice coefficient. While both networks perform better with larger size

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3000484, IEEE Journal of Biomedical and Health Informatics

of training data, we find that dU-Net can achieve similar performance to U-Net with only half of the data (limited to the sample quantity the last “train250” experiment does not double the size of “train160”.). Since dU-Net employs much more parameters than U-Net due to the extra deformation layers, fewer training samples requirements of dU-Net indicate that dU-Net can utilize much more information from the training data than U-Net. This conclusion is consist with the discussion of network generalizability in [41], where it has been shown that patterns shared by training samples is more useful than mere sample size for training the network. So here we hypothesize that the ability of a network to capture those shared patterns (e.g. because of extra deformable operations) is also more important than simply increasing the training data size.

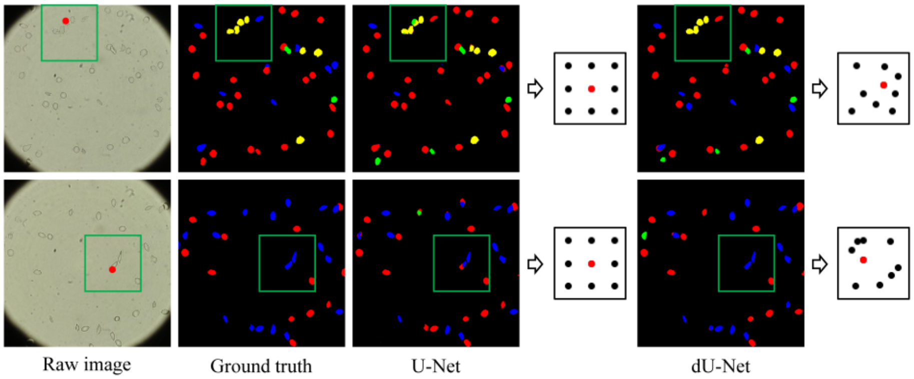

Fig. 6. Prediction maps of U-Net and dU-Net for multi-class RBC semantic segmentation. Different colors in the map represent different RBC types: red (Dic+Ovl), blue (El+Sk), yellow (Grl), and green (others). We highlight the cells that dU-Net maintains its shape integrity while U-Net does not in green blocks. We also display the sampling locations centered at the key pixels (red points in the raw image).

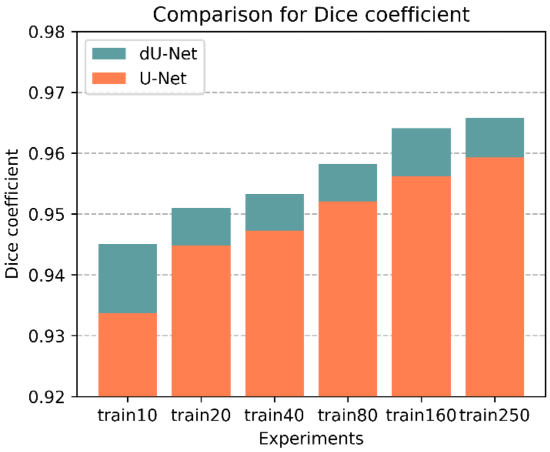

Fig. 7. Performance comparisons (measured in Dice coefficient) between U-Net and dU-Net for binary RBC segmentation using different sizes of training data. All the experiments use the same testing set (64 images), while “train10” represents the experiment using 10 images for training.

VI. CONCLUSION AND DISCUSSION

In this work, we apply an improved U-Net framework (dU-Net) for automated SCD RBC semantic segmentation. Experimental results show that dU-Net achieves better performance than other segmentation models. It also achieves similar classification results comparing with works in [11], which was performed on the pre-identified cell patches. Further, through detailed investigation of segmentation results, we find that dU-Net is more robust to the variations in cell size, texture and shape, which is reflected in its ability of correctly segmenting the cell boundary, separating touching cells, discriminating noise objects from real cells and ensuring the integrity of cell shapes. Overall, in this work we show that dU-Net is a robust method for SCD detection and diagnosis. Moreover, its capability of learning discriminative features from limited training samples makes it especially a suitable solution for biomedical image analysis, as it is usually difficult to collect sufficient biomedical images with annotations for training a complex model. Also, we are working on the implementation of post-processing steps for further exploration of the prediction maps, providing detailed statistics such as cell counts, density and average area, for better clinical decision supporting.

REFERENCES

[1] E. J. van Beers, L. Samsel, L. Mendelsohn, R. Saiyed, K. Y. Fertrin, C. A. Brantner, M. P. Daniels, J. Nichols, J. P. McCoy, and G. J. Kato, “Imaging flow cytometry for automated detection of hypoxia-induced erythrocyte shape change in sickle cell disease,” *American journal of hematology*, vol. 89, no. 6, pp. 598-603, 2014.

[2] N. Theera-Umpon, and S. Dhompongsa, “Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 3, pp. 353-359, 2007.

[3] H. Lee, and Y.-P. P. Chen, “Cell morphology based classification for red cells in blood smear images,” *Pattern Recognition Letters*, vol. 49, pp. 155-161, 2014.

[4] M. I. Razzak, and S. Naz, "Microscopic Blood Smear Segmentation and Classification Using Deep Contour Aware CNN and Extreme Machine Learning." pp. 801-807.

[5] R. Tomari, W. N. W. Zakaria, M. M. A. Jamil, F. M. Nor, and N. F. N. Fuad, “Computer Aided System for Red Blood Cell Classification in Blood Smear Image,” *Procedia Computer Science*, vol. 42, pp. 206-213, 2014.

[6] L. Zhang, L. Le, I. Nogues, R. M. Summers, S. Liu, and J. Yao, “DeepPap: Deep Convolutional Networks for Cervical Cell Classification,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1633-1643, 2017.

[7] Z. Gao, L. Wang, L. Zhou, and J. Zhang, “HEp-2 Cell Image Classification With Deep Convolutional Neural Networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 2, pp. 416-428, 2017.

[8] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Human Epithelial Type 2 cell classification with convolutional neural networks." pp. 1-6.

[9] L. Hongwei, H. Hao, W. Zheng, X. Xiaohua, and J. Zhang, "HEp-2 specimen classification via deep CNNs and pattern histogram." pp. 2145-2149.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.3000484, IEEE Journal of Biomedical and Health Informatics

[10] Y. Dong, Z. Jiang, H. Shen, W. D. Pan, L. A. Williams, V. V. B. Reddy, W. H. Benjamin, and A. W. Bryan, "Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells." pp. 101-104.

[11] M. Xu, D. P. Papageorgiou, S. Z. Abidi, M. Dao, H. Zhao, and G. E. Karniadakis, “A deep convolutional neural network for classification of red blood cells in sickle cell anemia,” *PLOS Computational Biology,* vol. 13, no. 10, pp. e1005746, 2017.

[12] F. Xing, and L. Yang, “Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review,” *IEEE Reviews in Biomedical Engineering,* vol. 9, pp. 234-263, 2016.

[13] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, “A review of semantic segmentation using deep neural networks,” *International Journal of Multimedia Information Retrieval,* vol. 7, no. 2, pp. 87-93, 2018.

[14] H. Zhu, F. Meng, J. Cai, and S. Lu, “Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation,” *Journal of Visual Communication and Image Representation,* vol. 34, pp. 12-27, 2016.

[15] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 4, pp. 640-651, 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham, 2015, pp. 234-241.

[17] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks." pp. 764-773.

[18] J. M. Chassery, and C. Garbay, “An Iterative Segmentation Method Based on a Contextual Color and Shape Criterion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-6, no. 6, pp. 794-800, 1984.

[19] M. E. Plissiti, C. Nikou, and A. Charchanti, "Watershed-based segmentation of cell nuclei boundaries in Pap smear images." pp. 1-4.

[20] F. Zamani, and R. Safabakhsh, "An unsupervised GVF snake approach for white blood cell segmentation based on nucleus." p. 1.

[21] G. Li, T. Liu, A. Tarokh, J. Nie, L. Guo, A. Mara, S. Holley, and S. T. C. Wong, “3D cell nuclei segmentation based on gradient flow tracking,” *BMC Cell Biology,* vol. 8, no. 1, pp. 40, 2007.

[22] X. Li, Z. Zhou, P. Keller, H. Zeng, T. Liu, and H. Peng, "Interactive exemplar-based segmentation toolkit for biomedical image analysis." pp. 168-171.

[23] A. S. Aydin, A. Dubey, D. Dovrat, A. Aharoni, and R. Shilkrot, "CNN Based Yeast Cell Segmentation in Multi-modal Fluorescent Microscopy Data." pp. 753-759.

[24] Y. Li, L. Shen, and S. Yu, “HEp-2 Specimen Image Segmentation and Classification Using Very Deep Fully Convolutional Network,” *IEEE Transactions on Medical Imaging,* vol. 36, no. 7, pp. 1561-1572, 2017.

[25] L. Yang, Y. Zhang, I. H. Guldner, S. Zhang, and D. Z. Chen, "3D Segmentation of Glial Cells Using Fully Convolutional Networks and k-Terminal Cut." pp. 658-666.

[26] L. Xiang, W. Li, X. Xiaodong, and H. Wei, "Cell classification using convolutional neural networks in medical hyperspectral imagery." pp. 501-504.

[27] K. De Haan, C. Koydemir H, Y. Rivenson, et al. "Automated screening of sickle cells using a smartphone-based microscope and deep learning." *arXiv: 1912.05155*, 2019.

[28] T. Tran, O. Kwon, K. Kwon, S. Lee, and K. Kang, "Blood Cell Images Segmentation using Deep Learning Semantic Segmentation." pp. 13-16.

[29] M. Shahzad, I. Umar A, A. Khan M, et al. "Robust Method for Semantic Segmentation of Whole-Slide Blood Cell Microscopic Images." *Computational and Mathematical Methods in Medicine*, 2020: 1-13.

[30] A. Sadafi, M. Radolko, I. Serafeimidis, et al. "Red Blood Cells Segmentation: A Fully Convolutional Network Approach." pp. 911-914, 2018.

[31] L. Perez, and J. Wang, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning,” *arXiv:1712.04621*, 2017.

[32] D. G. Lowe, "Object recognition from local scale-invariant features." pp. 1150-1157 vol.2.

[33] M *:1705.08881*, 2017.

[35] M. Aubreville, M. Krappmann, C. Bertram, R. Klopfleisch, and A. Maier, “A Guided Spatial Transformer Network for Histology Cell Differentiation,” *arXiv:1707.08525*, 2017.

[36] M. Zhang, X. Li, M. Xu, and Q. Li, "RBC Semantic Segmentation for Sickle Cell Disease Based on Deformable U-Net." pp. 695-702.

[37] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, “DUNet: A deformable network for retinal vessel segmentation,” *Knowledge-Based Systems,* vol. 178, pp. 149-162, 2019.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network." pp. 2881-2890.

[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation."

[40] A. Volokitin, G. Roig, and T. Poggio, “Do Deep Neural Networks Suffer from Crowding?,” in Neural Information Processing Systems, 2017.

[41] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A Closer Look at Memorization in Deep Networks."

Mo Zhang is a PhD student at the Center for Data Science, Academy for Advanced Interdisciplinary Studies, Peking University. She obtained her bachelor’s degree in mathematics from the School of Science at the China University of Petroleum (Beijing). Her research interests are artificial intelligence in medicine, machine learning, computer vision and medical image analysis.

Xiang Li is an Instructor at the Department of Radiology, Massachusetts General Hospital, Harvard Medical School. He obtained his PhD degree from the Department of Computer Science at the University of Georgia. He obtained his bachelor’s degree in Automation from the School of Electronic and Electric Engineering at Shanghai Jiaotong University. His research focuses on developing AI solutions for medical applications.

Mengjia Xu is currently a Postdoctoral Associate at the following institutes: 1) McGovern Institute for Brain Research, Massachusetts Institute of Technology; 2) Department of Radiology, Massachusetts General Hospital, Harvard Medical School; and 3) Peking University. She received her PhD degree in Computer Science from Northeastern University of China. During her PhD, she worked at Brown University for two years as a visiting PhD student in the Division of Applied Mathematics and previously a full-time software engineer intern at Neusoft for two years. Her main research interests are to develop data-driven machine learning methods for medical image analysis in real-world applications.

Quanzheng Li is an Associate Professor of Radiology at Massachusetts General Hospital, Harvard Medical School. He is also the director of the Center for Advanced Medical Computing and Analysis, a core faculty of Gordon Center for Medical Imaging, and the scientific director of the MGH/BWH Center for Clinical Data Science. He received his bachelor's degree from Zhejiang University, M.S. degree from Tsinghua University, and his PhD degree in Electrical Engineering from the University of Southern California. His research interests include image reconstruction and analysis in PET, SPECT, CT and MRI, and data science in health and medicine.