IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Cell-Net: Embryonic Cell Counting and Centroid Localization via Residual Incremental Atrous Pyramid and Progressive Upsampling Convolution

REZA MORADI RAD[1], (Student Member, IEEE), PARVANEH SAEEDI[1], (Member, IEEE), JASON AU[2], AND JON HAVELOCK[3]

[1]School of Engineering Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
[2]Pacific Centre for Reproductive Medicine, Burnaby, BC V5B 4X7, Canada
[3]Reproductive Endocrinology and Infertility Special Interest Group, Canadian Fertility Andrology Society, Montreal, QC H8Y 1Y4, Canada

Corresponding author: Reza Moradi Rad (e-mail: rmoradir@sfu.ca).

ABSTRACT In-vitro fertilization (IVF), as the most common fertility treatment, has never reached its maximum potentials. Systematic selection of embryos with the highest implementation potentials is a necessary step towards enhancing the effectiveness of IVF. Embryonic cell numbers and their developmental rate are believed to correlate with the embryo's implantation potentials. In this paper, we propose an automatic framework based on a deep convolutional neural network to take on the challenging task of automatic counting and centroid localization of embryonic cells (blastomeres) in microscopic human embryo images. In particular, the cell counting task is reformulated as an end-to-end regression problem that is based on a shape-aware Gaussian dot annotation to map the input image into an output density map. The proposed *Cell-Net* system incorporates two novel components, residual incremental Atrous pyramid and progressive up-sampling convolution. Residual incremental Atrous pyramid enables the network to extract rich global contextual information without raising the 'grinding' issue. Progressive up-sampling convolution gradually reconstructs a high-resolution feature map by taking into account short- and long-range dependencies. Experimental results confirm that the proposed framework is capable of predicting the cell-stage and detecting blastomeres in embryo images of $1-8$ cell by mean *accuracies* of 86.1% and 95.1%, respectively.

INDEX TERMS Cell Counting, Human Embryonic Cells, IVF, Medical Image Analysis, Deep Learning.

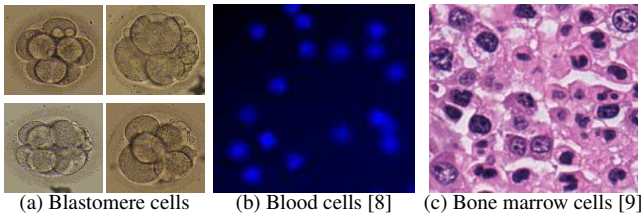
I. INTRODUCTION

ACCORDING to the World Health Organization (WHO), one in every four couples in developing countries suffers from infertility [1]. In-Vitro Fertilization (IVF) is one of the most common infertility treatments that emerged about four decades ago and practiced over one million times annually around the world [2]. Unfortunately, IVF has never reached to its maximum potentials. According to the Canadian Fertility and Andrology Society (CFAC) [3], only 33.1% of embryo transfer cycles led to a clinical pregnancy in Canada in 2017.

In IVF process, the fertilized eggs (refers to as embryos) are cultured for about 5 days inside an incubator to develop into blastocysts. These blastocysts are then subjectively assessed and selected according to their morphological characteristics for implantation. One of the most common embryo quality assessment techniques is pre-implantation genetic screening (PGS). While PGS has an excellent ability to predict non-implanting embryos (negative predictive value 96%) [4], it suffers from a low positive predictive rate (41-57% live birth rates) [4], [5]. Consequently, it is not the best option for embryo quality assessment due to its low positive predictive value and high costs (caused by embryo biopsy and genetic testing). In Canada, only 16.4% of IVF treatment cycles were classified as PGS treatment cycles in 2017 [3]. Therefore, embryo morphological grading remains the most practical method for embryo selection.

Several studies suggest that the timing and the synchronic-

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/ACCESS.2019.2920933, IEEE Access



(a) Blastomere cells (b) Blood cells [8] (c) Bone marrow cells [9]

FIGURE 1: Blastomere cells in human embryo images versus other types of human cells.

ity of the first few cleavages during the early human embryonic development correlate with an embryo's potentials for developing into a healthy baby [6], [7]. Automatic counting and centroid localization of embryonic cells (blastomeres) can provide information about the timing and spatial patterns associated with cell cleavages. Automatic cell counting is also of great interest in other biomedical diagnosis/analysis systems dealing with blood [8]–[12], tumor [13]–[15], and bacterial [15], [16] cells.

Microscopic embryo images are usually acquired by an embryoscope equipped with a digital microscope imaging system that captures images at 5-minute time intervals. Measuring the exact time associated with each cleavage requires processing approximately 576 frames for a single embryo. This measurement, if done manually, is expensive, error-prone, and most importantly impractical. Automating this process, although of great interest, is a challenging task. Ambiguity, partial view due to occlusion and out-of-focus conditions in these images results from the unconstrained transformation of 3D spherically shaped embryos into 2D image planes. In addition, background noise, cell fragmentation, cell transparency, and shape variability make this task even more complicated. Most of these complexities are not observed in other cell-based medical applications, such as tissue cell, blood cell counting (Fig. 1-b) or bone marrow cell counting (Fig. 1-c). In addition, for applications with a high number of cells, under- and over-counting of cells may not affect the accuracy of the outcome significantly. However, under- and over-counting of even one cell in human embryo images can lead to a significant error in assessing an embryo's quality.

In this paper, a modern deep learning based approach with a novel architecture is proposed to automatically count the number of blastomeres and localize their centroids in day 1-3 microscopic human embryo images. It is important to mention that no more than one n-cell (n=1:8) per physical embryo is utilized in the benchmark dataset for a true performance measurement.

II. RELATED WORK

A. DEEP CONVOLUTIONAL NEURAL NETWORKS (DCNNS)

The evolution of neural network architectures for image-to-image translation began in late 2014 by introducing fully Convolutional Neural Network (FCN) [17]. Since FCN architecture does not utilize fully connected layers, it can cope with images of arbitrary sizes. Despite all differences, existing architectures [17]–[25] can be divided into three main classes.

The first class utilizes an encoder-decoder structure. The encoder extracts hierarchical features by gradually scaling down the spatial dimension using pooling layers. The decoder scales up the dense features to reconstruct the original dimension. Models of this class are similar in the encoder part while their decoder design makes them distinct. Some of the most popular architectures in this category include FCN [17], DeconvNet [18], SegNet [19] and U-Net [20].

Methods of the second class extract sparse features in the first place using dilated convolutions. Dilated convolutions increase the field of view without reducing spatial dimensions. DilatedNet [21], RefineNet [22], and DeepLab V2 [23] are among the most popular architectures in this category. These methods developed their own way of aggregating multi-scaled features to generate the final prediction map.

More recently, a third class has emerged with the introduction of PSPNet [24] and DeepLab V3 [25]. These methods introduced the concept of pyramid pooling to capture multi-scale contextual information. This concept is proven to be effective for handling objects at different scales.

B. AUTOMATIC METHODS FOR CELL COUNTING

Conventional approaches to cell counting include a two-stage process where counting is performed following a detection or segmentation phase [26], [27]. Expectedly, the performance of the counting task relies heavily on the effectiveness of the underlying cell detection/segmentation algorithm. Recently, a new class of approaches for cell counting has emerged [8], [15] that perform the task of counting in one step using object density maps. These approaches do not require nor depend on prior knowledge through the detection process. Here, we focus on one-step approaches since the minimum annotation requirement makes them highly advantageous for biomedical applications. These approaches can be divided into two categories:

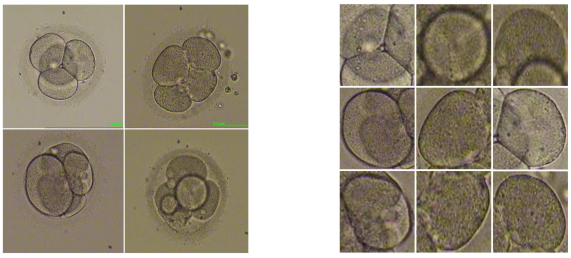
1) Classification Based Approach for Cell Counting

To the best of our knowledge, [28] is the only classification-based counting method for human embryo images. It performs the cell counting task using a multi-label classification approach via AlexNet network [29]. Such an approach assigns the same class label, x, to all images that contain x number of cells (as shown in Fig. 2-a).

2) Regression Based Approach for Cell Counting

Recently, we proposed a regression-based approach [30] for embryonic cell counting by reformulating the task as an end-to-end regression problem. This approach is based on supervised learning and maps the input image into an output cell density map. It undertakes the cell counting task using a Residual Dilated U-Net (RD U-Net) comprised of cascaded dilated convolutional layers and residual blocks.

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/ACCESS.2019.2920933, IEEE Access



(a) Classification approach - 4-cells stage (b) Regression approach - Single cells

FIGURE 2: Day 1-2 human embryo image samples at 1 to 5 cell stages.

Such a regression-based approach alleviates the requirement for balanced training samples by learning how a single cell looks like regardless of its developmental stage (Fig. 2-b).

In this paper, we extend our previous work [30] by introducing a shape-aware Gaussian dot annotation, a content-based loss function, and most importantly a novel DCNN architecture. The main contributions of the proposed approach include:

- Reformulating the task of human embryonic cell counting as an end-to-end regression problem that is trained in a supervised manner using shape-aware Gaussian annotation via a content-based loss function. This approach demonstrated great potentials and can be easily utilized for counting other types of cells such as blood or tumour.
- Proposing two novel components: Residual incremental Atrous pyramid (RIAP) and Progressive Upsampling Convolution (PUC). RIAP efficiently extracts rich global contextual information without raising the 'grinding' issue. PUC gradually reconstructs the high-resolution feature map by aggregating location-aware contextual information. These components can be incorporated into the design of DCNN for other applications such as semantic segmentation.

III. METHODOLOGY

A. EMBRYOS STRUCTURAL ATTRIBUTES

Prior to describing the proposed approach, we detail some of the unique aspects and properties of human embryos to provide some insights regarding some of the choices made for the proposed model. Human embryos possess unique biological attributes that could potentially complicate the task of counting the number of blastomeres inside them. Some of these attributes include:

- *Cell Overlap and Occlusions:* The highly overlapped and densely occupied space inside a human embryo make the task of counting embryonic cells a challenging one. Furthermore, the elliptically shaped overlapping regions between adjacent cells could trigger identifying false-positive cells.
- *Cell Fragmentation and Artifacts:* Fragmentation is defined as the presence of the small portions of cytoplasm that are enclosed by a cell membrane but separated from the nucleus. Human embryos often exhibit some degree of fragmentation that complicates the automatic analysis of these images.
- *Cell Size Variation:* Unlike blood cells which have approximately the same size, embryonic cells may have various sizes. Cells in an 8-cell embryo are smaller than cells in a 2-cell embryo, although of the same importance.

B. PROPOSED MODEL

The block diagram of the proposed *Cell-Net* model, which comprises encoder and decoder parts, is depicted in Fig. 3. In the encoder part of *Cell-Net*, residual incremental Atrous pyramid module is designed following the ResNet-50 to incorporate multi-scale contextual prior. An effective decoder module is created by introducing progressive upsampling convolution to recover fine details and object boundaries. These two novel components are described next.

1) Residual Incremental Atrous Pyramid (RIAP)

Availability of some knowledge on the global context is beneficial to the interpretation of microscopic images [24], [25], [31] for the cell counting task. Context relationship is a crucial factor in handling fragmentation and other artifacts to comprehend the complex nature of embryo images. In addition, scalable receptive field helps to improve the performance on remarkably small or large cells as some visual features become prominent only at a certain scale. *PSP-Net* [24] proposed a pyramid pooling module by applying pooling operations at 4 different scales to capture global context prior. Although [24] extracts rich context features, the pooling operation with striding leads to the information loss at object boundaries. *DeepLabv3* [25] proposed using parallel Atrous convolution with different rates instead of average pooling to capture the global context prior.

Dilated convolution [21] has become popular recently [21], [25], [30], [32]–[34]. Utilizing dilated convolution enlarges the receptive field without introducing addi-

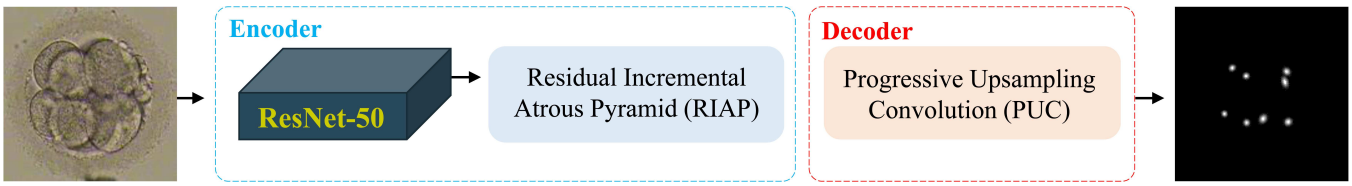


FIGURE 3: The block diagram of the proposed *Cell-Net* model.

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/ACCESS.2019.2920933, IEEE Access

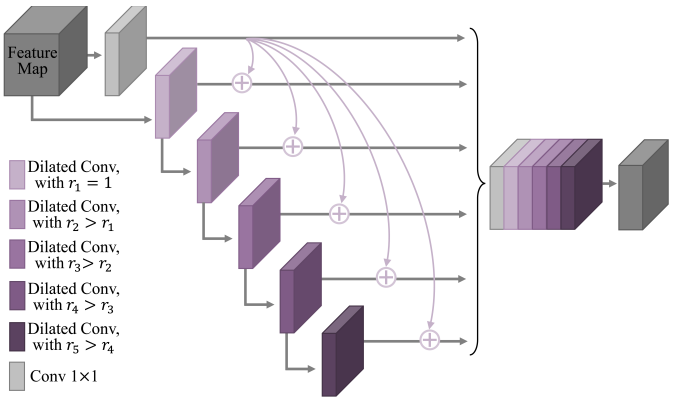


FIGURE 4: Structure of Residual Incremental Atrous Pyramid (RIAP).

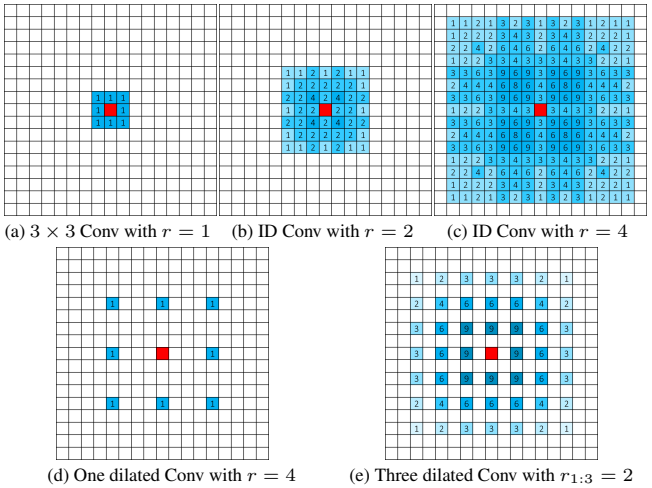


FIGURE 5: The blue shades and numbers on each pixel indicate significance of its contribution to the computation of the kernel response.

tional parameters to the network. Although applying dilated convolution improves the performance in *DeepLabv3* [25], larger dilation rates could lead to a practical problem, known as the 'grinding' issue. Grinding issue occurs when the sampling rate is too large to capture high-frequency content [33]. When applying dilated convolution, we observed that increasing the dilation rate can cause the correlation to fall apart gradually. In practice, when a 3×3 kernel applies to an image region or a feature map, the number of valid weights decreases by increasing the dilation rate. When the dilation rate is large, the number of valid weights reduces to the point where the 3×3 kernel acts as a 1×1 kernel. Fisher et al. [33] applied three policies (removing the max-pooling, adding more layers, and removing residual connections) to address grinding problem in a dilated residual network. More recently, *DeepLabv3* [25] adopted image-level features by global average pooling to overcome the grinding problem. Here, we address the root cause of the grinding problem by a simple yet effective solution.

In a 2-D space, a $s \times s$ dilated convolution between signal F and kernel K with dilation rate r is defined as:

$$(F *_r K)(x,y) = \sum_{m=-t}^{t} \sum_{n=-t}^{t} K(m,n) F(x - r.m, y - r.n) \quad \text{where} \quad t = (s-1)/2 \tag{1}$$

In the dilated convolution, the kernel only visits the signal at every r^{th} location of each dimension. Therefore, from a $s_d \times s_d$ dilated neighbourhood region, where $s_d = (r-1).(s-1)+s$, only $s \times s$ pixels contribute to the computation of the response at the central pixel. The $s \times s$ contributing pixels are all $r-1$ pixels away from each other and have the same distance from the centroid. For example, in a 3×3 dilated kernel with $r = 4$ (Fig. 5-d), only 9 pixels (out of the 81) contribute to the calculation of the kernel response, under-utilizing a substantial $\sim 89\%$ of the information.

Here, we propose a simple yet effective solution, named Residual Incremental Atrous Pyramid (RIAP). RIAP pursues two primary objectives. First, it addresses the grinding issue by allowing every single pixel in the dilated neighborhood to participate in the computation of the kernel response. Second, it further enlarges the receptive field. In RIAP, we set the stage for applying a large dilation rate of 2^j by backing it up with smaller dilation rates of 2^i where $j > i \geq 0$ with residual connections. Particularly, the dilation rate is increased to 2^i at the $(i+1)^{th}$ level of the pyramid, as illustrated in Fig. 4. In ID convolution, not only does each pixel matter but also its contribution is somewhat proportional to its distance from the central pixel. RIAP is computationally efficient with a total of five dilated convolutional layers that are built on the top of each other. Cascaded structure (i.e., instead of parallel) and residual connections are two major difference between RIAP and ASPP in [25].

Figs. 5-a to 5-c depicts the receptive field of the ID convolution. Here, ID has a receptive field of 15×15 (when $r = 4$) which is backed by two convolutions with $r = 2^1$ and $r = 2^0$. Fig. 5-d shows the receptive field of a single dilated convolution with $r = 4$ and Fig. 5-e depicts the receptive field of three cascaded dilated convolutions with $r = 2$. The proposed ID convolution in Fig. 5-c has a wider and enhanced receptive field with the same number of parameters compared to the one in Fig. 5-e.

2) Progressive Upsampling Convolution (PUC)

Information associated with boundary features and texture details could be lost in the absence of a proper up-sampling strategy. In the decoding phase, most state-of-the-art DCNNs simply use either bi-linear upsampling [20] or deconvolution [17], [18] to upscale the downsized dense features and create a final prediction map. Bi-linear upsampling is not learnable, and therefore the deconvolution could suffer from a checkerboard artifact [35]–[37]. Recently, Shi et al. [38] came up with an interesting idea (sub-pixel convolution) to recover resolution in a single-image super-resolution scenarios. The sub-pixel convolution aggregates low-resolution feature maps to reconstruct the high-resolution image. Wang

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/ACCESS.2019.2920933, IEEE Access

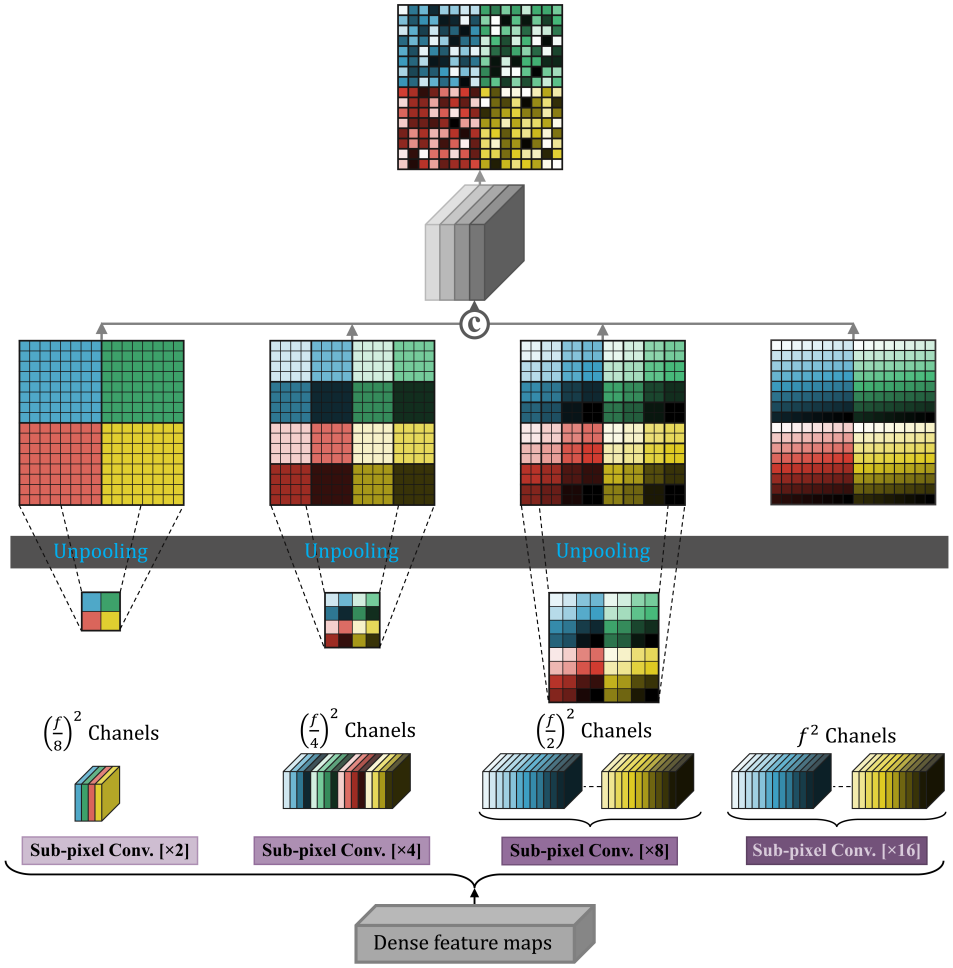


FIGURE 6: Structure of the proposed Progressive Upsampling Convolution (PUC) decoder module.

et al. [34] adopted this idea for upscaling dense feature maps in end-to-end segmentation reconstruction applications. The sub-pixel convolution in [38], however, was originally designed for super-resolution application, where the required upscaling factor was either 2 or 4. However, for end-to-end image processing applications, such as segmentation, usually a much larger upscaling factor is required. For example, ResNet [39], when employed as the encoder, downscales the input by a factor of 32. The main problem is that these pixels are upsampled regardless of their spatial locations, which leads to crucial information loss when reconstructing details and boundaries.

Inspired by [38], we take the idea of sub-pixel convolution one step further and propose the Progressive Upsampling Convolution (PUC) module as illustrated in Fig. 6. Here, the sub-pixel convolution produces a high-resolution image (upsampled by a factor of f) from f^2 low-resolution feature maps. These kernels are activated periodically in the high-resolution space to learn an individual upsampling kernel for locations that are f pixels away from each other. When the upsampling factor f is set to 32, for instance, each kernel learns the upsampling of pixels that are not strongly correlated in most regions. The proposed PUC attempts to reconstruct a high-resolution image in a progressive manner. The reconstruction begins by learning 4 upsampling kernels. It then performs a mirroring action to learn the global context of the high-resolution space. The reconstruction continues by increasing the number of upsampling kernels exponentially and performing the mirroring action. This mechanism enables PUC to capture short- and long-range dependencies between pixels in a high-resolution space.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. DATASET AND GROUND TRUTH

Unfortunately, there is no public dataset for early human embryo images (days $1-3$) in the biomedical field. Here, the first public dataset[1] for human embryo images up to 8-cell stage is introduced and utilized for experimental purposes. This benchmark human embryo dataset comprises 176 images that contain 511 embryonic cells collected at the Pacific Centre for Reproductive Medicine (PCRM). It must be noted that no more than one n-cell ($n = 1:8$) per physical embryo

[1]Dataset is available at: https://vault.sfu.ca/index.php/s/li0M2T3MHw9Vu8r

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/ACCESS.2019.2920933, IEEE Access

TABLE 1: Details of the benchmark embryonic cell dataset.

	No. of	1-cell	2-cell	3-cell	4-cell	5-cell	6-cell	7-cell	8-cell	Total
Train	Images	64	16	10	27	6	2	8	7	140
	Cells	64	32	30	108	30	12	56	56	388
Test	Images	11	5	4	6	3	1	3	3	36
	Cells	11	10	12	24	15	6	21	24	123

is utilized to ensure that there is no bias in the accuracy of the method. These images have been acquired using an Olympus IX71 inverted microscope that employs Nomarski optical enhancement technique (DIC). The training set comprises 140 images (80%) containing 388 embryonic cells. The test set comprises 36 images (20%) containing 123 embryonic cells. Distribution of the data over all cell-stages is provided in Table 1. The Ground Truth (GT) for these images is identified manually by expert embryologists at PCRM. We apply 2D elliptical Gaussian filters (proportional to the elliptical approximation of blastomeres) to the blastomere centroids to create shape-aware Gaussian dot annotation (as illustrated in Fig. 7-b). In addition to our human embryo benchmark dataset, two public datasets are utilized for external evaluation. First, VGG dataset that is introduced in [8] and contains 200 images of simulated bacterial cells from fluorescence-light microscopy. Second, MBM dataset that is introduced in [9] and contains 44 images of bone marrow.

B. IMPLEMENTATION DETAILS

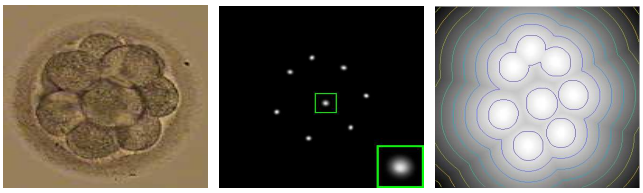
The proposed DCNN models are implemented using an NVIDIA GeForce GTX 1080 Ti with 11-gigabyte memory and 32-gigabyte RAM. The model was trained with 10 mini-batches of size 14 and Adam optimizer [40] with initial learning rate of $9.8e-5$.

Data augmentation: We applied standard data augmentation by randomly performing vertical/horizontal flipping, shear transform with an intensity of 0.1, zooming by a factor within the range of [0.88,1.12] and rotating by an angle within the range of [0°,360°].

Loss function: Predicting cell density map using a regular loss function is not feasible since labels are highly biased in favor of the background class (Fig. 7-b). As black pixels dominate the GT heavily, the network constantly falls into local minima, predicting all-zero image for any input image. In order to direct the learning process to the cells, we applied a Content-Based Mean Squared Error (CBMSE) loss function defined by Eq. 2.

$$CBMSE = \frac{\sum_{t=1}^{n} \left(Pre_t - Tar_t\right)^2 \times M_t}{n} \qquad (2)$$

Here, n is the number of images in the batch and M_t is a content-based attention map that draws the attention of the training task to the most important regions (as illustrated in Fig. 7).



(a) Original sample (b) Gaussian dot annotation (c) Attention map M_t

FIGURE 7: Visualization of the content-based attention map.

C. QUANTITATIVE RESULTS

Performance of the proposed method is evaluated at two levels: image level and cell level. While the first one measures the ultimate success of a cell counting system, the second one provides a more detailed analysis of the performance. The proposed approach is compared against [28], [30] that are the only two methods developed exclusively for blastomere counting. These methods are re-implemented, then trained and tested on the benchmark dataset. Furthermore, to highlight the effectiveness of the proposed architecture, some state-of-the-art architectures are adopted in the proposed shape-aware dot-annotation regression-based framework, including *UNet* [20], *TernausNet* [41], *PSPNet* [24] and *DeepLabv3* [25].

1) The proposed regression-based framework vs the classification-based approaches

Table 2 compares the performance of the proposed regression-based approach with that of the classification-based approach [28]. For a more comprehensive comparison, we extended our experiments by testing other well-established DCNN classification models (in addition to the *AlexNet* [29] used in [28]), including *VGG16* [42], *ResNet50* [43], and *Inception V3* [44]. Table 2 contains the cell-stage prediction accuracy at the image level (i.e., predicting the correct number of cells that exist in an image). A k-fold cross-validation is performed for a more comprehensive evaluation. Since in some categories such as $5-cell$ or $6-cell$, one-third of the data is kept in the test set (as shown in Table 1), there are 3 folds available to perform cross validation.

Overall, the proposed regression-based approach performs significantly better than the classification-based approach, regardless of the underlying DCNN model. There are two main reasons that explain the results in Table 2. First, unlike

TABLE 2: Comparison of the proposed regression approach with the classification approach for cell counting (in %).

	Cell-Stage Prediction Accuracy (in %)		
	$1-3$ stages	$4-8$ stages	Overall
Khan et al. [28]	35.0	18.8	27.8
VGG16 [42]	40.0	18.8	30.6
ResNet50 [43]	40.0	25.0	33.3
Inception V3 [44]	45.0	25.0	36.1
Cell-Net	**95.0**	**75.0**	**86.1**
3-fold cross validation	93.4	68.6	82.4

TABLE 3: Comparison of the proposed *Cell-Net* model with state-of-the-art architectures on cell-stage prediction (in %).

	Cell-Stage Prediction Accuracy (in %)		
	$1-3$ stages	$4-8$ stages	Overall
Baseline UNet [20]	95.0	37.5	69.4
RD UNet [30]	95.0	50.0	75.0
TernausNet [41]	85.0	56.3	72.2
PSPNet [24]	**100**	56.3	80.6
DeepLab V3 [25]	**100**	56.3	80.6
Cell-Net w/o PUC	90.0	62.5	77.8
Cell-Net w/o RIAP	95.0	62.5	80.6
Cell-Net	95.0	**75.0**	**86.1**
3-fold cross validation	93.4	68.6	82.4

TABLE 4: Detailed comparison of the proposed *Cell-Net* model with the state-of-the-art architectures (in %).

	Cell Detection Performance			
	False positive	False negative	True positive	Accuracy
Baseline UNet [20]	4	12	111	87.4
RD UNet [30]	2	11	112	89.6
TernausNet [41]	3	10	113	89.7
PSPNet [24]	0	10	113	91.9
DeepLab V3 [25]	0	8	115	93.5
Cell-Net w/o PUC	2	9	114	91.2
Cell-Net w/o RIAP	1	8	115	92.7
Cell-Net	0	**6**	**117**	**95.1**
3-fold cross validation	1.67	6	117	93.8

TABLE 5: Localization performance comparison between the proposed *Cell-Net* system and state-of-the-art models (in %).

	No. of Miss Localization (out of 123)	No. of Perfect Localization (out of 123)	Euclidean Distance (pixels)
Baseline UNet [20]	28	32	11.3
RD UNet [30]	22	41	10.1
TernausNet [41]	23	37	10.4
PSPNet [24]	14	72	7.5
DeepLab V3 [25]	15	74	7.4
Cell-Net w/o PUC	19	70	7.8
Cell-Net w/o RIAP	16	73	7.5
Cell-Net	**8**	**81**	**6.6**
3-fold cross validation	**10**	**78**	**6.9**

a typical classification problem, images of different cell counts are not independent of each other. For example, miss-classification of a bicycle as a motorcycle is as wrong as the miss-classification of a bicycle as a horse. However, miss-classifying a 2-cell embryo as a 3-cell embryo is not the same as miss-classifying it as a 7-cell embryo. Second, the availability of balanced/adequate training samples for all class categories is necessary to train a CNN model effectively. This, unfortunately, is rarely the case in medical field related applications.

2) The proposed Cell-Net model vs state-of-the-art models

Table 3 and 4 compare the performance of the proposed *Cell-Net* model with the state-of-the-art architectures [20], [24], [25], [41] when employed in the proposed regression based framework. To emphasize the contribution of each of the two RIAP and PUC components, introduced in the proposed *Cell-Net*, two variants of the proposed *Cell-Net* model are implemented. In the first variant, the proposed RIAP component is replaced with a plain dilated pyramid pooling (i.e., as incorporated in *DeepLabV3* [25]). In the second variant, the proposed PUC component was replaced with sub-pixel convolution (i.e., as introduced in [38] and incorporated by [34]).

Table 3 summarizes results at the image (cell-stage prediction) level. In this table, the prediction accuracy is reported for both $4-8$ and $1-8$ cell-stage categories. As shown in this table, the proposed *Cell-Net* model outperforms the state-of-the-art models by a large margin.

Table 4 reports results at the cell level (cell detection performance). Cell detection performance is more discernible than the cell-stage prediction accuracy. Table 4 suggests that the proposed *Cell-Net* model outperforms the state-of-the-art models with an *accuracy* of 95.1%.

3) Localization performance

Table 5 compares the centroid localization performance for detected cells (excluding [28] as it cannot localize cells). Euclidean distance (Eq. 3) between the identified centroids and the corresponding GTs is utilized to measure the localization accuracy. In this table, the number of miss and perfectly localized blastomeres along with the mean Euclidean Distance (ED) are reported. Miss and perfect localization are referred to the cases where ED is greater than 10 pixels and less than 3 pixels, respectively.

$$ED = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_{pi}-x_{gi})^2+(y_{pi}-y_{gi})^2} \qquad (3)$$

Here (x_{pi}, y_{pi}) is the centroid coordinates of the detected blastomere, (x_{gi}, y_{gi}) is the centroid coordinates of the ground truth, and N is the number of blastomeres in the image.

4) Effect of the CBMSE loss function

To highlight the effectiveness of the proposed *CBMSE* loss function, the proposed *Cell-Net* model is trained using a regular *MSE* loss function and the comparison results are reported in Table 6.

5) External validation

External validation is performed on two public datasets, VGG dataset [8] (200 images with an average of 174 ± 64 simulated bacterial cells) and MBM dataset [9] (44 images with an average of 126 ± 33 bone marrow cells). We follow the same training protocol used in [8], [9], where a fixed set of 100 images is reserved for testing while the size of the training and validation sets are varied. Since every image on these datasets contain more than 100 cell samples, a relatively smaller portion of data is sufficient for training and validation purposes. In Table 7, N_t and N_v represents the number of

TABLE 6: Effect of the CBMSE loss function on the prediction accuracy (in %).

	Cell-Stage Prediction		Cell Detection	
	MSE	CBMSE	MSE	CBMSE
Cell-Net	83.3	**86.1**	94.3	**95.1**
3-fold cross validation	79.7	**82.4**	92.9	**93.8**

TABLE 7: Comparison of the test set Mean Absolute Error (MAE) on two external datasets.

	VGG Dataset (200 Images)		MBM Dataset (44 Images)	
	$N_{t,v} = 32$	$N_{t,v} = 100$	$N_{t,v} = 10$	$N_{t,v} = 20$
FCRN [8]	3.4±0.2	2.9±0.2	28.9±22.6	22.2±11.6
Count-ception [9]	2.9±0.5	2.3±0.4	12.6±3.0	10.7±2.5
Cell-Net	**2.7±0.6**	**2.2±0.5**	**11.3±4.8**	**9.8±3.2**

images used for training and validation processes. Sample images from these datasets are depicted in Fig. 1.

D. QUALITATIVE RESULTS

Table 8 displays some sample outputs of the proposed *Cell-Net* model and visually compares them against *Baseline UNet* [20], *TernausNet* [41], *PSPNet* [24], and *DeepLabV3* [25]. The results from 1-cell stages are skipped due to their simplicity and to reserve the space for more complicated cases. The 1^{st} row of Table 8 depicts an example with background floating particles/cells. Both *Baseline UNet* [20] and *TernausNet* [41] mis-interpreted the floating cells as blastomeres. Rows 2 and 8 show cases where the proposed *Cell-Net* model, *PSPNet*, and *DeepLabV3* handle fragmentation by not mis-interpreting it as a blastomere cell, unlike *baseline UNet* and *TernausNet*. The example in the 3^{rd} row depicts a case where two blastomeres overlap. The elliptically shaped overlapping region between neighboring cells triggers false identification of a new blastomere in *Baseline UNet* [20] and *TernausNet* [41] but not in *Cell-Net*. Rows 4, 5, 6, and 7 represent cases where partial view due to occlusion and out-of-focus planes make the blastomeres ambiguous. This is the main area where the proposed *Cell-Net* model delivers superior performance comparing to *PSPNet* and *DeepLabV3* models.

E. COMPUTATIONAL COMPLEXITY

Figure 8 depicts a visual comparison of all the discussed models at both image and cell levels along with their network's parameter sizes. The proposed *Cell-Net* model contains roughly $\sim$ 34 millions parameters. While such a number is larger than some of the earlier models (*Baseline UNet* [20], *RD-UNet* [30], and *TernausNet* [41]) with much lower performance, it is roughly the same as *PSPNet* [24] and $\sim$ 5 millions less than *DeepLab V3* [25].

V. CONCLUSION

In this paper, an automatic framework based on a deep convolutional neural network was proposed to take on the challenging task of automatic counting and centroid localization of blastomeres in microscopic images of early human embryos. In particular, the cell counting task is formulated as an end-to-end regression problem that is based on supervised learning to map the input image into an output density map. The proposed *Cell-Net* system introduced two novel components, residual incremental Atrous pyramid and progressive upsampling convolution. Residual incremental Atrous pyramid extracts rich global contextual information without raising the grinding issue. Progressive upsampling convolution gradually reconstructs the high-resolution feature map by taking into account local and global contextual structures of the scene. Experimental results confirm that the proposed framework is capable of predicting cell-stage and detecting cells by a mean *accuracy* of 86.1% and 95.1%, respectively. Furthermore, experimental results confirmed that the proposed method is capable of localizing blastomere centroids with a mean Euclidean distance error of 6.6 pixels.

REFERENCES

[1] M. N. Mascarenhas, S. R. Flaxman, T. Boerma, S. Vanderpoel, and G. A. Stevens, "National, regional, and global trends in infertility prevalence since 1990: A systematic analysis of 277 health surveys," PLOS Medicine, vol. 9, no. 12, pp. 1–12, 2012.

[2] E. Santos Filho, J. Noble, and D. Wells, "A review on automatic analysis of human embryo microscope images," The open Biomed. Eng. J., vol. 4, pp. 170–177, 2010.

[3] "Canadian Fertility and Andrology Society (CFAS)," http://www.cfas.ca/, (on 08/01/2017).

[4] R. T. Scott Jr, K. Ferry, J. Su, X. Tao, K. Scott, and N. R. Treff, "Comprehensive chromosome screening is highly predictive of the reproductive potential of human embryos: a prospective, blinded, nonselection study," Fertility and Sterility, vol. 97, no. 4, pp. 870–875, 2012.

[5] M. D. Werner, M. P. Leondires, W. B. Schoolcraft, B. T. Miller, A. B. Copperman, E. D. Robins, F. Arredondo, T. N. Hickman, J. Gutmann, W. J. Schillings et al., "Clinically recognizable error rate after the transfer of comprehensive chromosomal screened euploid embryos is low," Fertility and sterility, vol. 102, no. 6, pp. 1613–1618, 2014.

[6] C. C. Wong, K. E. Loewke, N. L. Bossert, B. Behr, C. J. De Jonge, T. M. Baer, and R. A. Reijo Pera, "Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage," Nat. Biotechnol., vol. 28, no. 10, pp. 1115–1121, 2010.

[7] M. Meseguer, J. Herrero, A. Tejera, K. M. Hilligsoe, N. B. Ramsing, and J. Remohi, "The use of morphokinetics as a predictor of embryo implantation," Hum. Reprod., vol. 26, no. 10, pp. 2658–2671, 2011.

[8] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," Comput. Methods in Biomechanics and Biomed. Eng.: Imag. & Vis., pp. 1–10, 2016.

[9] J. Paul Cohen, G. Boucher, C. A. Glastonbury, H. Z. Lo, and Y. Bengio, "Count-ception: Counting by fully convolutional redundant counting," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 18–26.

[10] J. Ge, Z. Gong, J. Chen, J. Liu, J. Nguyen, Z. Yang, C. Wang, and Y. Sun, "A system for counting fetal and maternal red blood cells," IEEE Trans. on Biomed. Eng., vol. 61, no. 12, pp. 2823–2829, Dec 2014.

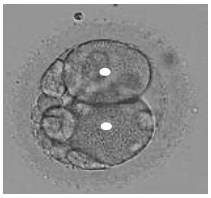
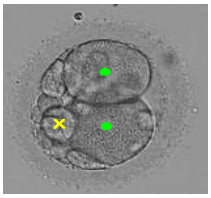
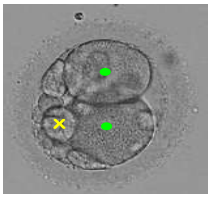
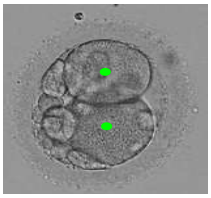
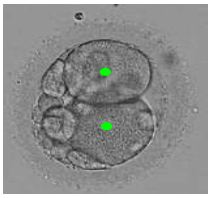
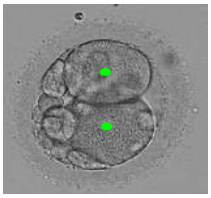
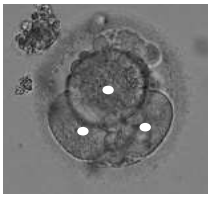
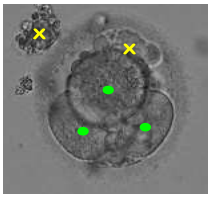
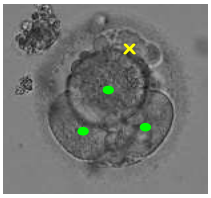
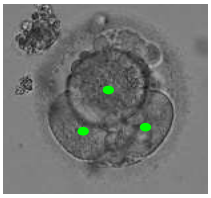
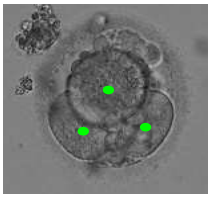
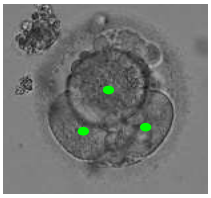
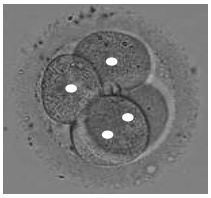
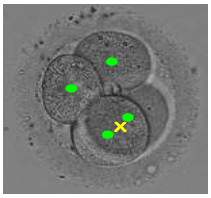
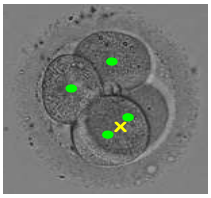
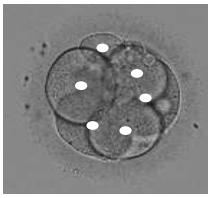
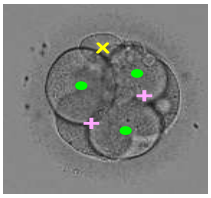
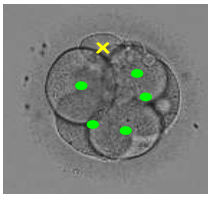
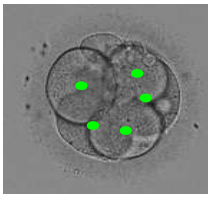
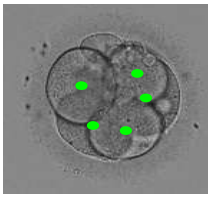
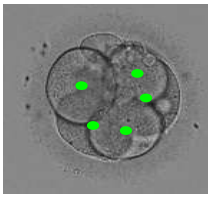
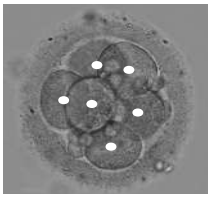
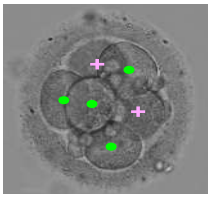
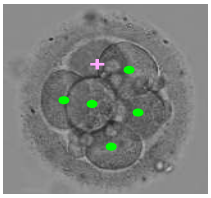
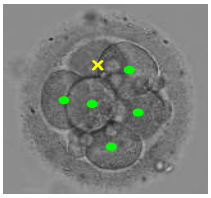
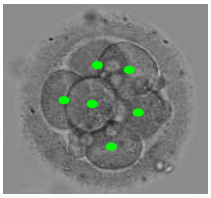
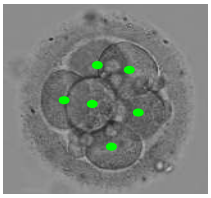
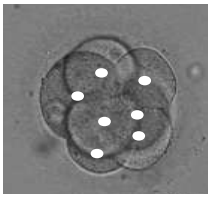
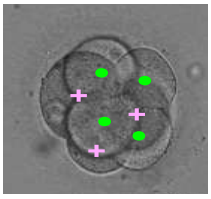
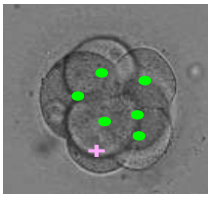
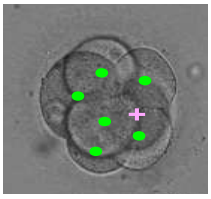
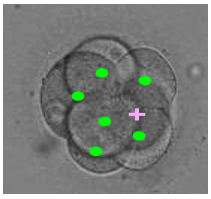
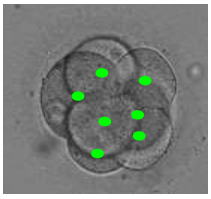
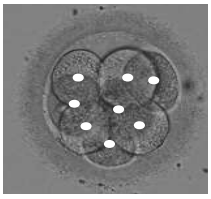
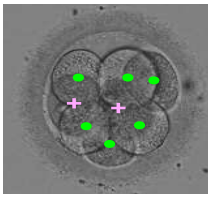
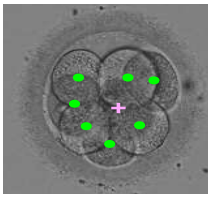
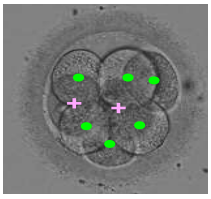
[11] X. Wang, T. Xu, J. Zhang, S. Chen, and Y. Zhang, "So-yolo based wbc detection with fourier ptychographic microscopy," IEEE Access, vol. 6, pp. 51 566–51 576, 2018.

[12] M. Sajjad, S. Khan, Z. Jan, K. Muhammad, H. Moon, J. T. Kwak, S. Rho, S. W. Baik, and I. Mehmood, "Leukocytes classification and segmentation in microscopic blood smear: A resource-aware healthcare service in smart cities," IEEE Access, vol. 5, pp. 3475–3489, 2017.

[13] D. Tellez, M. Balkenhol, I. Otte-Holler, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi, "Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks," IEEE Trans. on Med. Imag., vol. 37, no. 9, pp. 2126–2136, 2018.

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/ACCESS.2019.2920933, IEEE Access

TABLE 8: A qualitative comparison of the *Cell-Net* and the state-of-the-art models (best viewed in color). Here, green dot (•) implies true positive, light-magenta plus (+) highlights false negative, and yellow cross (×) indicates false positive.

	Sample Image with GT	Baseline UNet	TernausNet	PSPNet	DeepLabV3	Cell-Net
Row 1: 2−cell stage						
Row 2: 3−cell stage						
Row 3: 4−cell stage						
Row 4: 5−cell stage						
Row 5: 6−cell stage						
Row 6: 7−cell stage						
Row 7: 8−cell stage						
Row 8: 8−cell stage						

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/ACCESS.2019.2920933, IEEE Access

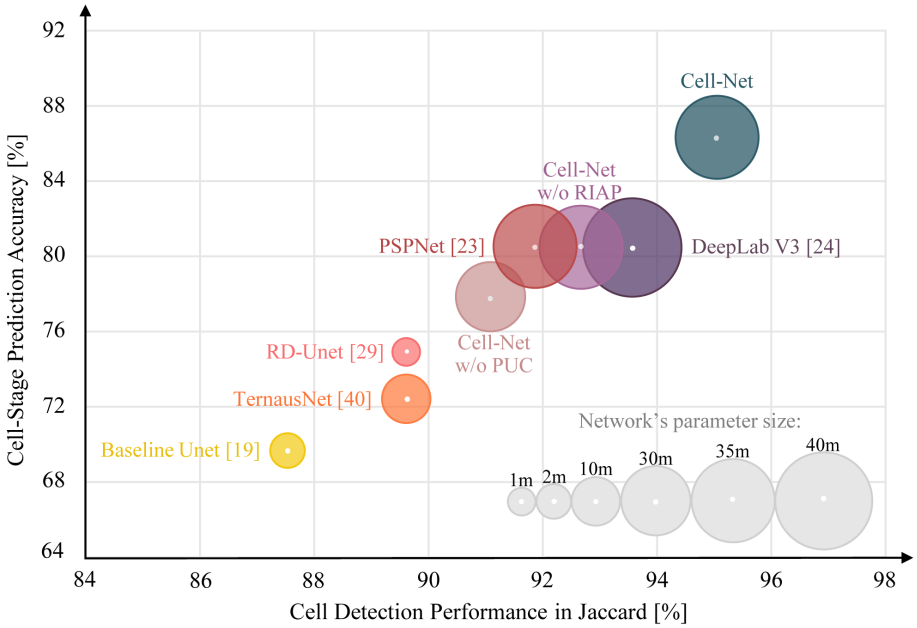


FIGURE 8: Performance versus computational complexity trade-off (network parameter size is proportional to the radius of the circles).

[14] F. Xing, H. Su, J. Neltner, and L. Yang, "Automatic ki-67 counting using robust cell detection and online dictionary learning," IEEE Trans. on Biomed. Eng., vol. 61, no. 3, pp. 859–870, March 2014.

[15] Y. Xue, N. Ray, J. Hugh, and G. Bigras, "Cell counting by regression using convolutional neural network," in Proc. European Conf. on Comput. Vision. Springer, 2016, pp. 274–290.

[16] K. Niitsu, S. Ota, K. Gamo, H. Kondo, M. Hori, and K. Nakazato, "Development of microelectrode arrays using electroless plating for cmos-based direct counting of bacterial and hela cells," IEEE Trans. on Biomed. Circuits and Syst., vol. 9, no. 5, pp. 607–619, Oct 2015.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, 2015, pp. 3431–3440.

[18] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in Proc. IEEE Int. Conf. on Comput. Vision, 2015, pp. 1520–1528.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," arXiv preprint arXiv:1511.00561, 2015.

[20] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. Cham: Springer International Publishing, 2015, pp. 234–241.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in Proc. Int. Conf. on Learning Representations, 2016.

[22] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, vol. 1, no. 2, 2017, pp. 1925–1934.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Trans. on Pattern Anal. and Mach. Intell., vol. 40, no. 4, pp. 834–848, 2018.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, 2017, pp. 2881–2890.

[25] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[26] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "A hybrid approach for multiple blastomeres identification in early human embryo images," Computers in Biology and Medicine, vol. 101, pp. 100 – 111, 2018.

[27] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Learning to detect cells using non-overlapping extremal regions," Med. Image Comput. Comput. Assist. Interv., vol. 15, pp. 348–356, 2012.

[28] A. Khan, S. Gould, and M. Salzmann, "Deep convolutional neural networks for human embryonic cell counting," in Proc. European Conf. on Comput. Vision. Springer, 2016, pp. 339–348.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Info. Process. Syst., 2012, pp. 1097–1105.

[30] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Blastomere cell counting and centroid localization in microscopic images of human embryo," in Proc. IEEE Int. Workshop on Multimedia Signal Process., 2018, pp. 1–5.

[31] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in Advances in Neural Info. Process. Syst., 2016, pp. 4898–4906.

[32] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images," in Proc. IEEE Int. Conf. on Image Process., Oct 2018, pp. 3518–3522.

[33] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks." in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, vol. 2, 2017, p. 3.

[34] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," arXiv preprint arXiv:1702.08502, 2017.

[35] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," Distill, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[36] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in Proc. IEEE Int. Conf. on Comput. Vision, 2017, pp. 4501–4510.

[37] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," arXiv preprint arXiv:1707.02937, 2017.

[38] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, 2016, pp. 1874–1883.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, 2016, pp. 770–778.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[41] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," arXiv preprint arXiv:1801.05746, 2018.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, 2016, pp. 770–778.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. on Comput. Vision and Pattern Recognition, 2016, pp. 2818–2826.