

Difference-Deformable Convolution With Pseudo Scale Instance Map for Cell Localization

Chengyang Zhang¹, Jie Chen¹, Bo Li¹, Min Feng¹, Yongquan Yang¹, Qikui Zhu², and Hong Bu¹

Abstract—Cell localization still faces two unresolved challenges: 1) the dramatic variations in cell morphology, coupled with the heterogeneous intensity distribution of lightly stained cells; 2) existing cell location maps lack scale information, resulting in insufficient supervision for point maps and inaccurate supervision for density maps. 1) To address the first challenges, we introduce a novel gradient-aware and shape-adaptive Difference-Deformable Convolution (DDConv), which enhances the model's robustness to color by leveraging gradient information while adaptively adjusting the shape of the convolutional kernel to tackle the substantial variability in cell morphology. 2) To overcome the issue of unreasonable location maps, we propose the Pseudo-Scale Instance (PSI) map, which can adaptively provide the corresponding scale information for each cell to realize accurate supervision. We analyze and evaluate DDConv and the PSI map in three challenging cell localization tasks. In comparison to existing methods, our proposed approach significantly enhances localization performance, setting a new benchmark for the cell localization task.

Index Terms—Cell localization, location map, deformable-difference convolution.

I. INTRODUCTION

ACCURATELY locating the precise position and spatial layout of each cell is a crucial yet formidable task with

Manuscript received 24 May 2023; revised 24 September 2023; accepted 25 October 2023. Date of publication 6 November 2023; date of current version 5 January 2024. This work was supported in part by the 1-3-5 Project for Disciplines of excellence under Grant ZYGD18012 and in part by the Technological Innovation Project of Chengdu New Industrial Technology Research Institute under Grant 2017-CY02—00026-GX. (Chengyang Zhang and Jie Chen contributed equally to this work.) (Corresponding authors: Qikui Zhu; Hong Bu.)

This work did not involve human subjects or animals in its research. Chengyang Zhang and Bo Li are with the Department of Pathology and Institute of Clinical Pathology, West China Hospital, Sichuan University, Chengdu 610041, China, and also with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: cy_zhang@emails.bjut.edu.cn; bo_li@emails.bjut.edu.cn).

Jie Chen, Yongquan Yang, and Hong Bu are with the Department of Pathology and Institute of Clinical Pathology, West China Hospital, Sichuan University, Chengdu 610041, China (e-mail: jzc-jedu@foxmail.com; remy_yang@foxmail.com; hongbu@scu.edu.cn).

Min Feng is with the Department of Pathology and Institute of Clinical Pathology, West China Hospital, Sichuan University, Chengdu 610041, China, and also with the Department of Pathology, West China Second University Hospital of Sichuan University/Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Chengdu 610041, China (e-mail: huaxipathfm@163.com).

Qikui Zhu is with the Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: qikuizhu@163.com).

Our code is available at <https://github.com/ChyaZhang/DDConv-PSI>. Digital Object Identifier 10.1109/JBHI.2023.3329542

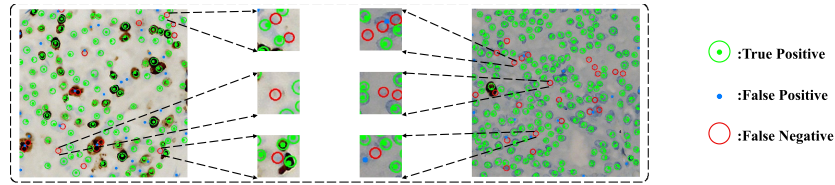
broad applications in biological research [1]. Previous studies have made significant progress. For example, Alam et al. [2] use a modified You Only Look Once (YOLO) network to automate the detection and counting of red blood cells, white blood cells, and platelets. Huang et al. [3] design a network based on Congested Scene Recognition Network (CSRNet) that regresses density maps to locate positive and negative cells in breast cancer pathological sections. However, two challenges remain undressed, which significantly affect the accuracy of cell counting.

The large variability in cell morphology, coupled with the heterogeneous intensity distribution of lightly stained cells, presents the first unsolved challenge, as shown in Fig. 1(a). In order to tackle the challenge of disparities in morphology, Tofghi et al. [4] propose a Tunable Shape Prior Convolutional Neural Network (TSP-CNN). TSP-CNN incorporates shape priors, which are customized to match the intricate and diverse cell shapes in images. However, the fixed shape priors of this approach restrict its applicability to other scenarios. To address the issue of the heterogeneous intensity distribution of lightly stained cells, Li et al. [5] propose a multi-scale difference convolution module. This approach enhances the model's robustness to cell color in images. However, difference convolution amplifies the edge information of cells in the feature map, which to some extent exacerbates the interference of cell shape to the model. Current methods aim to address cell morphology and staining heterogeneity separately but face significant challenges, limiting accurate cell localization and counting. Hence, accurately identifying and localizing lightly stained and morphologically diverse cells remains an unexplored area with the potential to improve cell localization and counting performance.

Unreasonable location maps are the second unsolved challenge. Typically, existing location maps can be roughly divided into two categories: Density maps [3] and Point maps [6], [7]. Density maps reflect the density of cells in different regions, and point maps are binary images that contain disks with cell centers as shown in Fig. 1(b) and (c). Although the two location maps show effectiveness in cell localization and counting, both of them come with several clinical drawbacks. Regarding the density map, it primarily involves two key issues: 1) Inability to avoid the overlapping challenge. Density maps generated based on Gaussian convolution inevitably encounter the issue of overlap in dense regions, resulting in the model being guided in the wrong direction, ultimately affecting its coverage and accuracy. 2) Complex post-processing. Density maps need to be calculated by the local maxima algorithm to obtain the specific location of each cell. When cells are not uniformly distributed, it is difficult

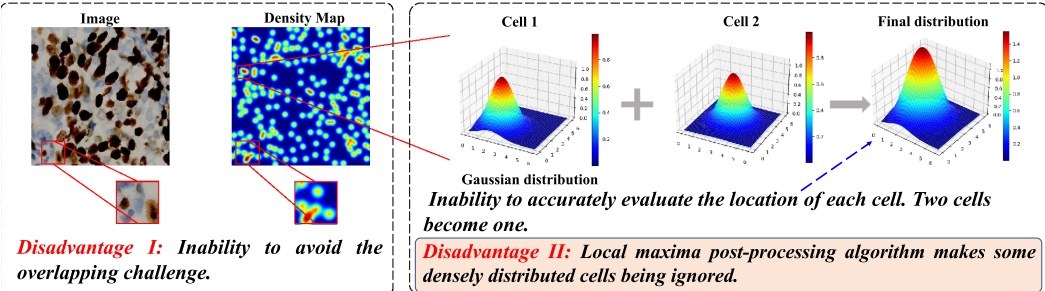
Existing automatic cell localization methods are still facing two unresolved challenges.

Challenge I: Inability to effectively address the large variability in cell shape and heterogeneous intensity distribution of lightly stained cells, which is the bottleneck restricting the accurate counting of cells.

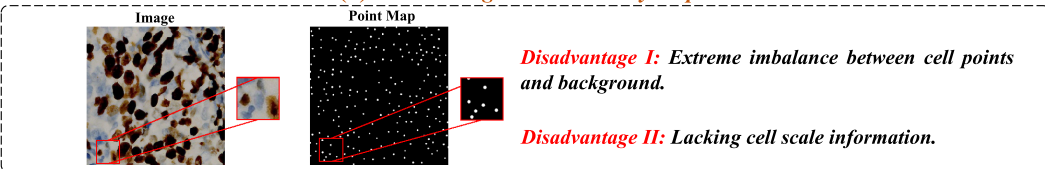


(a) Challenges in imaging

Challenge II: Unreasonable location map.



(b) Disadvantages in the density map



(c) Disadvantages in the point map

Fig. 1. Existing automatic cell localization methods are still facing two unresolved challenges. Challenge I: The large variability in cell size and shape, coupled with the heterogeneous intensity distribution of lightly stained cells, presents the first unsolved challenge. Challenge II: Unreasonable location map is the second unsolved challenge.

to define the range for calculating the local maxima, resulting in some cells being ignored. In contrast, point maps use very small disks to represent cells. Although point maps avoid the overlapping challenge, which also brings several drawbacks. 1) Extreme imbalance. In point maps, there is an extreme imbalance between point pixels and background pixels. The large negatives can be overwhelming and make up the majority of the loss, resulting in the gradient being dominated by large negative values and ignoring the cell information. 2) Lacking cell scale information. Point maps can hardly represent cell scale information, which makes the model lose sensitivity to cell size and reduces the performance of the model. Hence, there is an urgent need for an accurate and reasonable location map manner.

To overcome the challenge of the large variability of cell morphology and heterogeneous intensity distribution, we propose a novel Difference-Deformable Convolution (DDConv) (Fig. 2). DDConv is gradient-aware and shape-adaptive, which enables the model to focus on the gradient information of cells for extracting the edge information of lightly stained cells and meanwhile adaptively adjust the shape of the convolutional kernel for overcoming the challenge of the large variability in cell shape. Specifically, DDConv employs eight filters for feature representation learning, and each of the filters is used to calculate the differences in one of the eight directions. This

operation preserves neighboring activation difference information for determining edges and corners and makes DDConv gradient-aware. Furthermore, to enable filters to handle the large variability in cell morphology, we adopt deformable filters inside DDConv. Based on the advanced deformable filters, our DDConv is adaptively adjusted according to the cell's morphology during feature representation learning.

To address the issue of unreasonable location maps, we introduce a novel concept called the Pseudo-Scale Instance (PSI) map. As illustrated in Fig. 2, within the PSI map, each cell is treated as an individual connected circular domain with a scale-related radius. The PSI map dynamically computes the scale information and associates it with the annotation of each cell. In comparison to existing location maps, our innovative PSI map offers two notable advantages: 1) Computational efficiency: Instead of relying on uncertain probability distributions, PSI assigns a specific scale value to each cell. This approach effectively mitigates computational challenges encountered in density maps. 2) Scale awareness: The utilization of scale information in PSI enhances the model's sensitivity to the size and shape of diverse cells. This, in turn, helps prevent overlapping issues and mitigates extreme imbalances in the data. Overall, our PSI map represents a valuable enhancement in constructing location maps by incorporating scale information, addressing

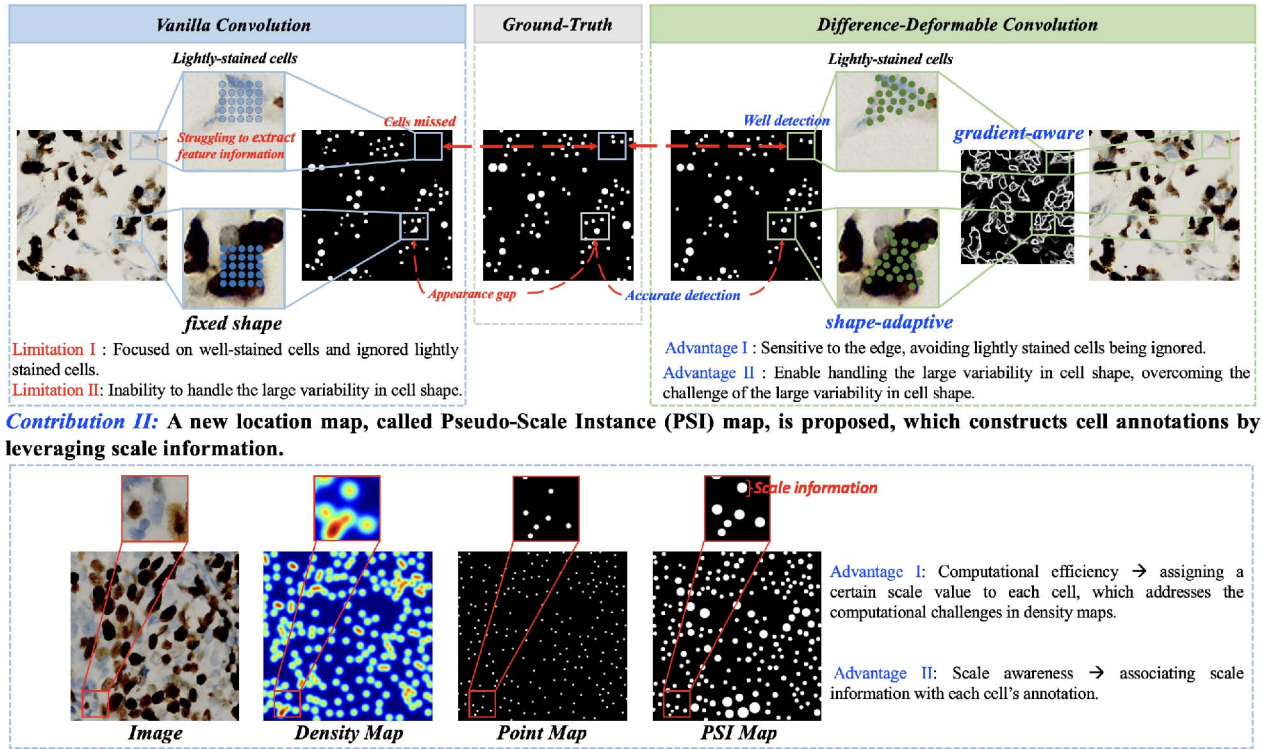


Fig. 2. Our novel DDConv is gradient-aware and shape-adaptive, which enables the model to focus on the gradient information of cells for extracting the edge information of lightly stained cells and meanwhile adaptively adjust the shape of the convolutional kernel for overcoming the challenge of the large variability in cell shape. Our novel method constructs location maps by leveraging scale information, offering two advantages: 1) Computational efficiency. 2) Scale awareness.

computational efficiency concerns, and fostering scale awareness in the model.

To assess the advancements brought by DDConv and PSI, three challenging cell localization datasets are utilized for thorough comparisons and evaluations [3], [8], [9]. Extensive experimental results demonstrate that the scale information provided by our PSI contributes significantly to the performance improvement in localization tasks. Further, to validate our DDConv module, we introduce the DCLNet model built upon this module. Comprehensive comparative experiments and ablation studies indicate that this model outperforms the state-of-the-art methods. Our primary contributions encompass:

- 1) Our novel DDConv is gradient-aware and shape-adaptive, which enables extracting the edge information of lightly stained cells for overcoming the challenge of the heterogeneous intensity distribution in lightly stained cells and meanwhile adaptively adjusting the shape of the convolutional kernel for overcoming the challenge of the large variability in cell shape.
- 2) Our new PSI map enables adaptively computing the scale information and associating it with each cell's annotation, which addresses the computational inability challenge in density maps and advanced makes the model sensitive to the size of cells.
- 3) Due to the synergistic integration of the PSI map and DDConv, our novel approach has outperformed existing methods and set a new performance benchmark.

II. RELATED WORKS

In this section, we briefly describe the current research status of cell localization and counting using CNN, mainly including methods based on detection, density map, and point map. Additionally, related work on deformable convolution is also reviewed.

A. Detection-Based Methods

Detection-based cell localization methods typically generate multiple candidate regions in the image and then classify and filter each region to identify those that truly contain cells [10], [11]. Typically, Alam et al. [2] use a modified YOLO network to automate the detection and counting of red blood cells, white blood cells, and platelets. Ma et al. [12] propose an abnormal cell detection network based on Mask R-CNN, which integrates different features using an attention mechanism to improve detection performance. Du et al. [13] design a method based on Retinanet in the state of Super Depth of Field (SDoF) to achieve high precision detecting of leucorrhea components by the SDoF feature aggregation module.

Detection-based methods perform well in scenarios with sparse target distribution. However, the performance tends to degrade as cell density increases. Moreover, bounding box-level annotation is often relatively expensive. Therefore researchers now commonly use point-based annotation methods.

B. Density Map-Based Methods

To better utilize the spatial information, most current cell localization and counting works are based on density maps. For instance, Pan et al. [14] introduce a multiscale fully convolutional neural network for density map regression. The network can detect small single cells as well as large and overlapping cells. To address the scarcity of datasets in the cell localization and counting field, Sirinukunwattana et al. [15] propose a spatially constrained convolutional neural network and release a dataset named UW. Recently, Huang et al. [3] propose a large-scale dataset called BCDData for cell counting, localization, and classification, and design a network based on CSRNet that regresses density maps.

These works have promoted the development and application of cell localization and counting to some degree. However, density map-based methods cannot effectively use the scale information of cells. In some irregularly shaped cells, a cell may be marked with multiple localization points, resulting in more false positives. Therefore, some researchers begin to solve cell localization and counting by generating pseudo-segmentation maps (point maps).

C. Point Map-Based Methods

Nowadays, some researchers generate point maps based on point labels to better utilize the scale information of cells in images. For example, Hagos et al. [6] propose an Inception-v3-based neural network and use point maps to supervise its training. Raza et al. [7] combine point labels with mapping filters to generate artificial pseudo-labels for training convolutional neural networks.

Compared with density maps, point maps can reflect the scale information of cells to some extent. However, most existing point maps use circles of uniform size to represent each cell. Since the size of each cell in real cell images varies, it is highly unreasonable to use circles of the same size to represent all cells. In addition to manually generating pseudo-labels, some researchers [8], [16] have used instance segmentation to simultaneously perform cell localization in cell classification tasks. However, instance segmentation annotation is often expensive and does not provide labels directly related to cell localization and counting. Consequently, using instance segmentation datasets for cell localization and counting is not cost-effective when cell classification is not needed.

D. Deformable Convolution

In the field of image segmentation [17], [18], the target object is often irregular in shape. Fixed-shaped convolution kernels tend to perform poorly on such targets. To achieve more accurate segmentation results, researchers introduce deformable convolution [19] to segmentation tasks. Generally, Huang et al. [20] design a feature alignment module based on deformable convolutions. This module learns pixel offsets and inserts them into the FPN structure to contextually align upsampled features. In medical image segmentation, there are also related studies on deformable convolution. Xie et al. [21] propose a feature

fusion module for medical image segmentation. The module consists of feature attention selection, cross-offset generation, and deformable convolution layers to alleviate the ambiguous semantic information between the encoder and decoder. Furthermore, deformable convolution has also been used in cell detection. Li et al. [22] insert deformable convolution into the FPN structure and extend the Faster R-CNN model for automatic detection of cervical squamous epithelial cells in liquid-based cytology.

The emergence of deformable convolution has effectively improved the problem of misaligned contextual features in segmentation tasks, thereby improving the accuracy of segmentation results. In this paper, we introduce deformable convolution into the field of cell localization and combine it with difference convolution [23].

III. METHOD

In this section, we detail the DDConv and the generation of the PSI map. During the training phase, we design a network called DDConv-based Cell Localization Network (DCLNet) to learn the mapping relationship from cell images to the PSI maps. To optimize the model parameters, the loss between the output image and the PSI map is calculated.

A. Difference-Deformable Convolution

Our novel Difference-Deformable Convolution (DDConv) is gradient-aware and shape-adaptive, which enables the model sensitive to lightly stained cells and meanwhile adaptively adjust the offsets of the convolutional sampling points. Specifically, various cells have diverse shapes, which causes deviations in shape from the circular annotations. Additionally, it is inevitable to have lightly stained parts of the cell due to variations in staining techniques, scoring methods, and selection of scoring regions. Lightly stained cells are the bottleneck restricting the accurate detection of cells due to low contrast and blur boundaries. Although existing convolution methods can extract high-level semantic features from cells, they cannot capture cells with irregular shapes due to the intrinsic locality of the convolution operator and lack of sensitivity to lightly stained cells.

To overcome the above challenges, a novel Difference-Deformable Convolution is proposed. As shown in Fig. 3, our DDConv adaptively adjusts the offsets of the convolutional sampling points during feature extraction and enables extracting gradient information of cell edges [24] for overcoming the challenge of lightly stained cells. Formally, the vanilla convolution can be represented as

$$y(p_0) = \sum_{p_n \in R} w(p_n) \times x(p_0 + p_n), \quad (1)$$

where p_0 denotes the central position of the local receptive field R , p_n represents the relative position of each value from R to p_0 , and $w(p_n)$ is a learnable parameter. x and y is the input and output feature map respectively.

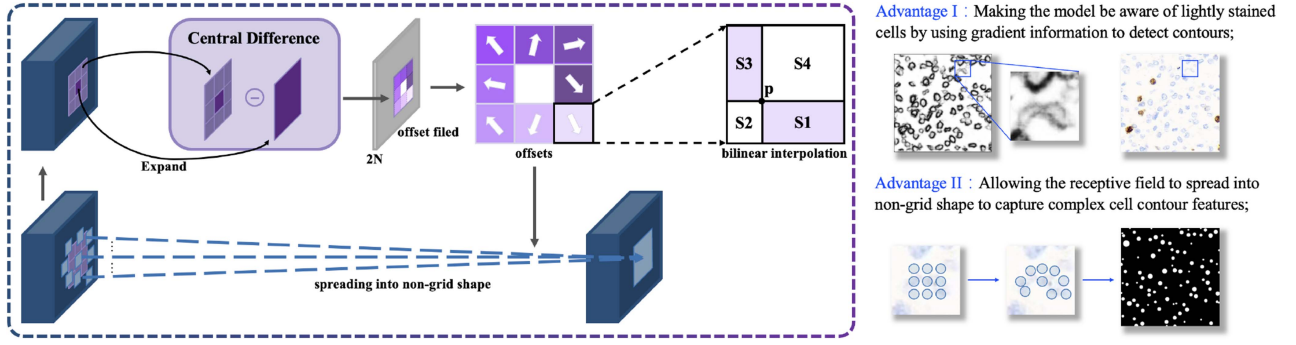


Fig. 3. Illustration of DDConv. First, the gradient information of the image is extracted by central difference convolution. Then, the offset of each sampling point of the convolutional kernel is calculated by the extracted gradient information. Finally, each sampling point of DDConv is convolved with the pixels on the image that have been mapped by the offset.

The definition of DDConv can be represented as

$$y_D(p_0) = \sum_{p_n \in R} w(p_n) \times BI(x(p_0 + p_n + \Delta p_n)). \quad (2)$$

Compared with vanilla convolution, our DDConv enables adjusting the scope of the convolution operation through a learnable parameter Δp_n , which is updated by back-propagation during the training process. Δp_n is generated by convolving the input feature map with another convolution (central difference here) and it is usually a decimal number. After adding Δp_n , the sampling is on the irregular and offset locations $p_n + \Delta p_n$. As the offset Δp_n is typically fractional, $x(p_0 + p_n + \Delta p_n)$ is implemented via Bilinear Interpolation (BI):

$$BI(x(p)) = \sum_q \max(0, 1 - |q_x - p_x|) \cdot \max(0, 1 - |q_y - p_y|) \cdot x(q), \quad (3)$$

where p denotes the fractional location ($p = p_0 + p_n + \Delta p_n$), q spread out all integral spatial locations in the input feature map x . The formula $\max(0, 1 - \dots)$ is the restriction that the interpolated point will not be more than 1 pixel away from the domain point. After BI, p is uniquely determined.

In the DDConv, Δp_n is generated by the central difference convolution, which can be expressed as

$$\Delta p_n = \sum_{p_n \in R} w(p_n) \times (x(p_0 + p_n) - x(p_0)). \quad (4)$$

That is, each value $x(p_0 + p_n)$ in the local receptive field R is subtracted from its centroid $x(p_0)$ to form local gradient information. Meanwhile, considering that the vanilla convolution can bring stronger semantic information, the offset calculation can be expressed as

$$\Delta p_n = \sum_{p_n \in R} w(p_n) \times x(p_0 + p_n) + \theta \left(-x(p_0) \times \sum_{p_n \in R} w(p_n) \right), \quad (5)$$

where θ is a hyperparameter that controls the ratio between vanilla convolution and difference convolution. The results of

applying vanilla convolution and DDConv to cell images are shown in Fig. 2. It can be observed that the segmentation results obtained by applying DDConv are closer to the annotated images than those obtained by vanilla convolution.

B. Pseudo Scale Instance Map

To address the computational inability challenge in existing maps and advanced makes the model sensitive to the scale information of cells, we propose a new cell location map, called Pseudo Scale Instance (PSI) map. Compared with point maps that use fixed-radius circles to represent each cell, PSI maps use circles of different sizes to represent each cell. Since the annotation of the instance segmentation dataset includes the boundary of each cell, it is easy to obtain the scale information of each cell and generate PSI maps accordingly. We first generate PSI maps using an instance segmentation dataset and perform a preliminary experiment to validate the effectiveness of the PSI map. The process of generating PSI maps using an instance segmentation dataset is shown in Fig. 4(a). Then, considering that existing cell localization datasets [3], [4], [15] often do not contain information about the scale of cells, which limits the further promotion of PSI map. To address this issue, we aim to add scale information to existing point-annotated datasets. Moreover, the annotation cost of instance segmentation datasets is often very expensive, while point labels are much easier to obtain. To this end, a scale-giving method that can provide scale information for datasets without scale information is introduced, as shown in Fig. 4(b).

1) Scale Validation: According to Fig. 4(a), the *Instance extraction & Distance transform analysis* is performed to generate PSI maps from an instance segmentation dataset. Then, a preliminary experiment is deployed to verify the performance enhancement of the PSI map, which is shown in Section IV-E.

Instance extraction & Distance transform analysis: In this step, each instance in an image is extracted from the annotation and represented using circles of different sizes, with position labels added to generate PSI maps, as shown in Algorithm 1. The total number of instances in an image is denoted as N , and each instance is saved as a separate image, referred to as *ins map*. Therefore, one cell image corresponds to N *ins maps*. For

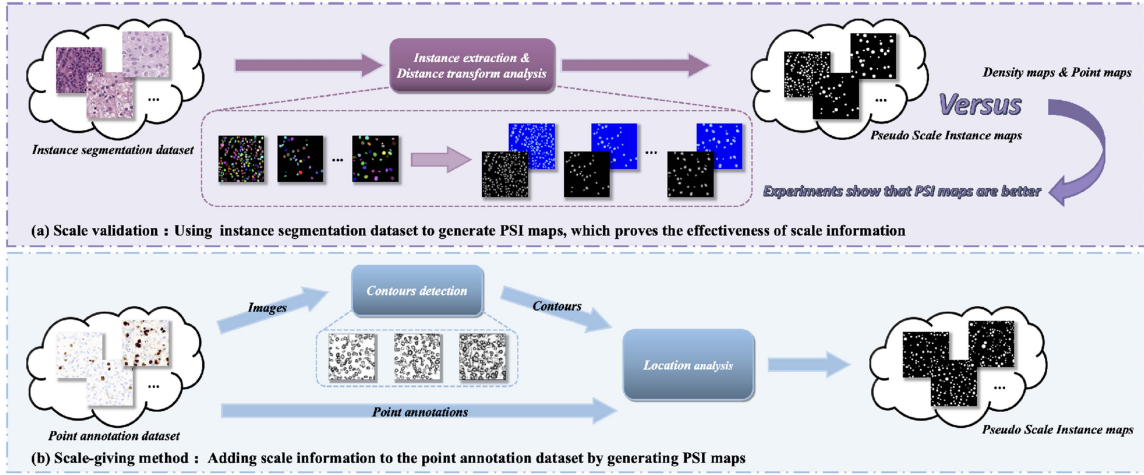


Fig. 4. (a) Presents the process of scale verification experiment and generating PSI maps using instance segmentation datasets. The results compared with density maps and point maps demonstrate that scale information can improve cell localization performance. (b) Illustrates the process of the scale-giving method, which mainly consists of contour detection and location analysis. This method can be used to generate PSI maps using point annotation datasets and comprehensively incorporate scale information into the cell localization task.

Algorithm 1: Instance Extraction & Distance Transform Analysis.

Input: The *instance segmentation annotations* per image.

Output: PSI map.

```

1  $N$  = number of instances per image;
2 for  $i$  in  $N$  do
3   save instance as ins map;
4 end
5  $\text{PSI\_map} = \text{zeros\_like}(\text{ins map})$ ;
6 for  $i$  in  $N$  do
7    $\text{dt} = \text{distanceTransform}(\text{ins map})$ ;
8    $\text{point, dist} = \text{minMaxLoc}(\text{dt})$ ;
9    $\text{circle}(\text{PSI\_map}, \text{point}, \text{radius} = \text{dist})$ ;
10 end

```

each *ins map*, the distanceTransform function in OpenCV is performed on the unique connected domain it contains, obtaining the centroid of the domain and the set of distances from the centroid to the boundaries of the connected domain. The instance segmentation image, distance map, and corresponding heatmap are shown in Fig. 4. Finally, a circle is drawn on the PSI map using the centroid as the center and the maximum value in the set of distances as the radius, which can be obtained by the minMaxLoc function in OpenCV. After iterating through all N *ins maps*, the PSI map corresponding to the image can be obtained.

2) Scale-Giving Method: According to Fig. 4(b), the *Contours detection* and the *Location analysis* are deployed to introduce the scale information to existing cell localization datasets without scale information.

Contour detection: Firstly, it is significant to detect the contour of each cell to obtain the scale information. There are many methods for boundary detection, such as Sobel, Laplace, Canny,

Algorithm 2: Location Analysis.

Input: The *contours & point annotations* per image.

Output: PSI map.

```

1  $\text{PSI\_map} = \text{zeros\_like}(\text{image})$ ;
2  $M = \text{dot annotations per image}$ ;
3 for  $(x, y)$  in  $M$  do
4    $\text{min\_dist} = 0.5 * \min(\text{Euclidean}((x, y), (x', y')), (x', y') \in M)$ ;
5    $\text{dist} = \text{pointPolygonTest}((x, y), \text{contour})$ ;
6   if  $\text{dist} > 0$  then
7      $\text{circle}(\text{PSI\_map}, (x, y), \text{radius} = \min(\text{min\_dist}, \text{dist}))$ ;
8   else
9      $\text{dist} = \text{mean}(\sum_{k=1}^{M-1} \text{dist}^{(k)})$ ;
10     $\text{circle}(\text{PSI\_map}, (x, y), \text{radius} = \min(\text{min\_dist}, \text{dist}))$ ;
11  end
12 end

```

and other operators. In order to be able to detect both lightly stained and well stained cells, we use a difference convolution-based network [25] to extract cell contours.

Location analysis: After finding the contour of each cell, the PSI map is generated based on the contours and the point annotation. Specifically, we calculate the distance between each point and contour by using the pointPolygonTest function in OpenCV. If the distance is greater than 0, the point and contour are supposed to be successfully matched. Then, the corresponding PSI map is generated through location analysis, as shown in Algorithm 2. Suppose there are M annotation points in an image. We traverse these M points and calculate the half distance between the currently traversed point and the nearest point, denoted as *min dist*. If the current point is matched with a contour, the distance is calculated from the current point to the matched

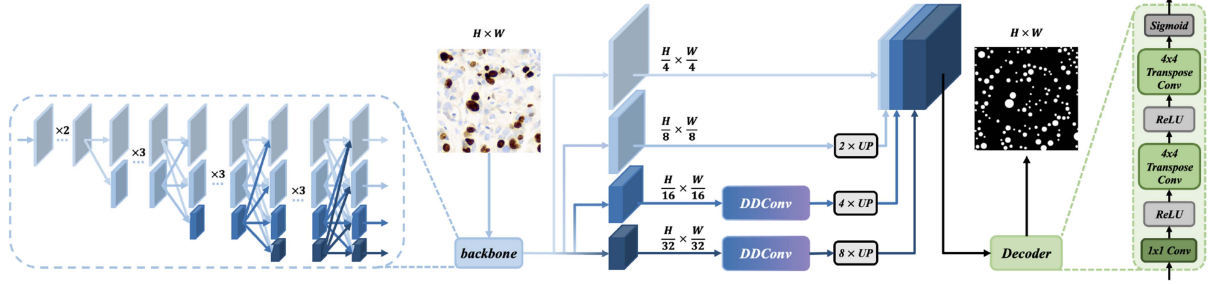


Fig. 5. Illustration of DCLNet. The backbone consists of 4 stages, and each stage will add a branch. Different-sized features will fuse at the end of each stage and finally output 4 features. After that, two of the four output feature maps with smaller sizes are fed into the DDConv. Then, all the feature maps are upsampled to the same size and concatenated along the channel dimension.

contour, which is denoted as *dist*. Using the current point as the center of the circle, a circle is drawn on the PSI map with a radius equal to the smaller value between *dist* and *min dist*. This circle represents the corresponding cell at that coordinate, and the radius of the circle reflects the cell's scale information. If the current point doesn't match any contour, it means that the corresponding contour hasn't been well detected. Then, the average radius of all the circles corresponding to the other cells is signified as *dist*, and the smaller value between *dist* and *min dist* equals the radius of the circle.

C. DCLNet for Cell Localization

Based on the advance of DDConv, we propose a new cell localization network, called DCLNet, as shown in Fig. 5. DCLNet mainly consists of a backbone and two DDConv. The backbone consists of 4 stages and achieves both strong semantic information and accurate location information by parallelizing multiple branches of resolution and continuously interacting with information between different branches. The reason for choosing the decoder instead of the encoder is that the decoder is mainly responsible for patching up the image. The decoder restores high-dimensional features to low-dimensional images, perfects the geometric shapes of objects in the process, and compensates for the detail loss caused by the pooling layers in the encoder, making the results closer to the annotation. The network structure diagram is shown in Fig. 5. The cell image is first passed through two 3×3 convolutions to obtain a feature map of size $\frac{H}{4} \times \frac{W}{4}$, which is then fed into the backbone. Subsequently, the backbone outputs four feature maps with sizes of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$, respectively. Then, the two smaller feature maps are sent to the DDConv separately, and all feature maps are upsampled to the size of $\frac{H}{4} \times \frac{W}{4}$ and concatenated in the channel dimension. Finally, the feature map is reduced to one channel and the size is restored to $H \times W$ through the decoder, and the location map is obtained after post-processing.

IV. EXPERIMENTS

A. Datasets and Implementation

1) *Datasets*: In the experiment, we convert the annotation information of three datasets, including an instance

segmentation dataset and two point annotation dataset, into PSI maps. We briefly describe these two datasets as follows.

The nuclei grading dataset [8] consists of 1,000 Haematoxylin and Eosin (H&E) stained images with a resolution of 512×512 . This dataset contains 70,945 annotated cell nuclei, including 16,652 endothelial nuclei and 54,293 tumor nuclei. The tumor regions are selected by two experienced pathologists from 150 ccRCC and 50 pRCC WSIs and scanned at $40\times$ objective magnification. In this dataset, the training, validation, and testing sets contain 700, 200, and 100 cell images, respectively.

BCData [3] is a large Ki-67 staining dataset for cell localization and counting, containing 1,338 breast tumor cell images. The original WSIs are scanned at $40\times$ magnification (~ 0.2239 microns/pixel). The cropped images have a uniform resolution of 640×640 , with a total of 181,074 annotated cells. In this dataset, there are 803, 133, and 402 images in the training, validation, and testing set, respectively. It is worth mentioning that this experiment is implemented based on <https://openi.pcl.ac.cn/xuf01/ki67>, represented as U-CSRNet^x in Table VI.

CoNIC [9] is the current largest publicly available nuclei-level dataset in computational pathology. It contains around half a million labeled nuclei. The dataset consists of H&E stained histology images at $20\times$ objective magnification (~ 0.5 microns/pixel) from 6 different data sources. For each image, an instance segmentation and a classification mask is provided. To make the cells in a single image of the three different datasets similar in size, we cropped the datasets to 256×256 resolution images. The dataset is split into 4,807 images (images without cells are not included). We set 2,837 images from the CoNIC dataset as the training set, 501 images as the validation set, and 1,469 images as the test set.

2) *Implement Details*: During the training process, random horizontal flipping and random scaling are implemented. The scaling ratio ranges from 0.8 to 1.2. The experiments are conducted on an Nvidia GeForce RTX 3090 (~ 24 GB), with a batch size of 4 and a learning rate set to $1e-4$. After 200 iterations, the learning rate is decayed to $1e-5$, and the total epoch is set to 800. The AdamW optimizer is used to optimize the network. To train the proposed network, the standard mean squared error loss function is chosen in the experiment.

3) *Evaluation Metrics*: To evaluate the performance of cell localization and counting, separate metrics for localization and counting are required. A match is considered successful when

TABLE I
PERFORMANCE COMPARISON OF MAINSTREAM METHODS ON BCData DATASET, NUCLEI GRADING DATASET, AND CoNIC DATASET

Method	BCData					Nuclei Segmentation					CoNIC				
	F1(%)	Pre(%)	Rec(%)	MAE	RMSE	F1(%)	Pre(%)	Rec(%)	MAE	RMSE	F1(%)	Pre(%)	Rec(%)	MAE	RMSE
Inception U-Net [26]	80.0	85.7	76.5	34.4	45.6	87.9	89.8	86.2	6.4	9.1	77.4	82.8	72.6	39.4	52.8
MPViT-base [27]	82.4	86.9	78.8	26.5	35.7	88.8	90.1	87.5	5.6	7.8	79.6	84.5	75.2	27.3	36.7
Unext [28]	84.9	89.6	79.7	30.3	37.3	88.4	90.1	87.0	7.5	10.5	79.4	84.9	74.5	33.4	42.8
RepVGG-B3 [29]	85.2	87.6	82.9	19.9	27.3	89.8	90.8	88.8	5.0	6.8	81.3	81.9	80.7	20.4	25.7
VGG16+FPN [30]	85.6	85.3	85.9	<u>17.2</u>	23.3	90.4	89.8	90.1	<u>4.6</u>	<u>6.1</u>	80.4	84.9	76.3	26.9	34.8
PVT-L [31]	85.8	87.2	84.6	20.8	28.3	89.3	89.6	88.9	5.4	7.2	81.4	84.8	78.3	19.5	26.7
U-Net [32]	86.0	86.2	85.9	17.6	23.9	89.9	92.2	87.7	5.7	7.9	80.8	85.2	76.9	25.3	32.9
DAE-Former [33]	86.4	87.6	85.1	17.3	<u>22.5</u>	89.8	89.9	89.6	5.0	6.9	80.2	85.7	75.8	26.8	35.3
TransUnet [34]	86.3	85.9	86.6	17.8	23.2	89.1	93.4	84.8	6.2	8.8	81.9	84.6	79.4	32.5	42.5
UCSRNet [3]	86.7	87.9	85.6	18.2	24.8	89.1	93.5	85.1	7.9	10.6	81.0	83.8	78.3	26.4	34.5
DenseNet-201 [35]	85.5	87.0	84.1	20.6	26.7	89.0	90.0	87.9	7.1	9.7	81.1	84.8	77.6	27.9	36.5
ResNet-101 [36]	85.8	87.2	84.4	18.5	25.1	89.3	90.2	88.5	5.2	6.9	80.1	85.3	75.5	30.3	39.0
Efficient-Unet [37]	86.1	87.5	84.7	20.1	27.2	90.0	90.1	89.4	4.8	6.5	82.5	84.9	80.3	22.6	30.3
HoVer-Net [38]	85.0	87.8	82.4	27.4	37.9	89.1	90.1	88.2	7.4	10.3	<u>83.6</u>	85.0	82.3	19.4	<u>24.8</u>
HRNet [39]	<u>87.2</u>	86.5	87.8	18.4	24.4	<u>90.5</u>	89.7	91.3	4.7	6.4	83.1	85.7	80.7	<u>19.2</u>	25.9
DCLNet	87.8	87.9	87.7	16.0	21.2	90.9	91.7	90.1	4.5	5.8	84.4	85.4	83.3	18.1	22.6

F1, Pre, and Rec reflect localization performance, the higher the better. MAE and RMSE reflect counting performance, the lower the better. Bold and underlined indicate the best and second-best performance, respectively.

the distance between the given predicted point and the true point is less than a threshold value. In this paper, the threshold is set to the radius of each mask in PSI maps.

Localization Metrics: To accurately evaluate the matching relationship between predicted cell points and ground truth, we use F1 score, precision, and recall to assess the localization performance of the model. They are defined as

$$Recall = \frac{TP}{TP + FN}, \quad (6)$$

$$Precision = \frac{TP}{TP + FP}, \quad (7)$$

$$F1 = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}, \quad \beta = 1, \quad (8)$$

where TP , FP , and FN represent the values of true positive, false positive, and false negative.

Counting Metrics: In this paper, instead of directly regressing the number of cells, we obtain the results by counting the connected regions in the output image. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to evaluate the counting performance of the model, which are defined as

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i^{gt} - \hat{y}_i|, \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |y_i^{gt} - \hat{y}_i|^2}, \quad (10)$$

where N represents the total number of samples in the validation or test set, y_i^{gt} is the ground truth boundary, and \hat{y}_i is the predicted value for the i th sample.

B. Comparison With State-of-the-Art Methods

Experimental results on the BCData dataset, nuclei grading dataset, and CoNIC dataset from two aspects, quantitative and

qualitative performances. Table I shows the quantitative result of DCLNet and state-of-the-art methods. From Table I, we can notice that our DCLNet achieves the best localization performance on all datasets, which demonstrates that our DCLNet has a more effective cell feature representation learning ability. Fig. 6 demonstrates several visualization results of images in BCData and the nuclei grading dataset generated by various methods. From Fig. 6, we can notice that with the utility of the PSI map, all methods can mostly overcome the challenges of low sensitivity in the scale of cells. However, existing state-of-the-art methods still suffer from the challenge of ignoring lightly stained cells, which leads to a decline in localization performance. Compared with state-of-the-art methods, our DCLNet obtained more accurate cell localization, which demonstrates that DDConv is sensitive to cells with various shapes. Additionally, regardless of whether the cell images are well stained with Ki-67 or H&E, DCLNet almost always maintains the best performance, demonstrating the robustness of the DDConv to lightly stained cells.

C. Ablation Studies on DDConv

1) Comparison With Other Convolutions: To demonstrate the superiority of our novel DDconv, we also compare DDconv with other convolutions including vanilla convolution and deformable convolution. Fig. 7 shows the location maps generated by baseline with vanilla convolution, deformable convolution, and DDconv. In addition to the qualitative comparisons, Table II shows the quantitative comparison of the baseline with three convolutions on the CoNIC dataset and BCData.

From Fig. 7, two points are summarized. 1) Deformable convolution and DDConv can make the cell shapes closer to the ground truth, which addresses of challenge of variations in cell shapes. 2) Compared to deformable convolution, the DDConv can better optimize the cell shapes and detect more lightly stained cells at the same time. It can also be seen from the

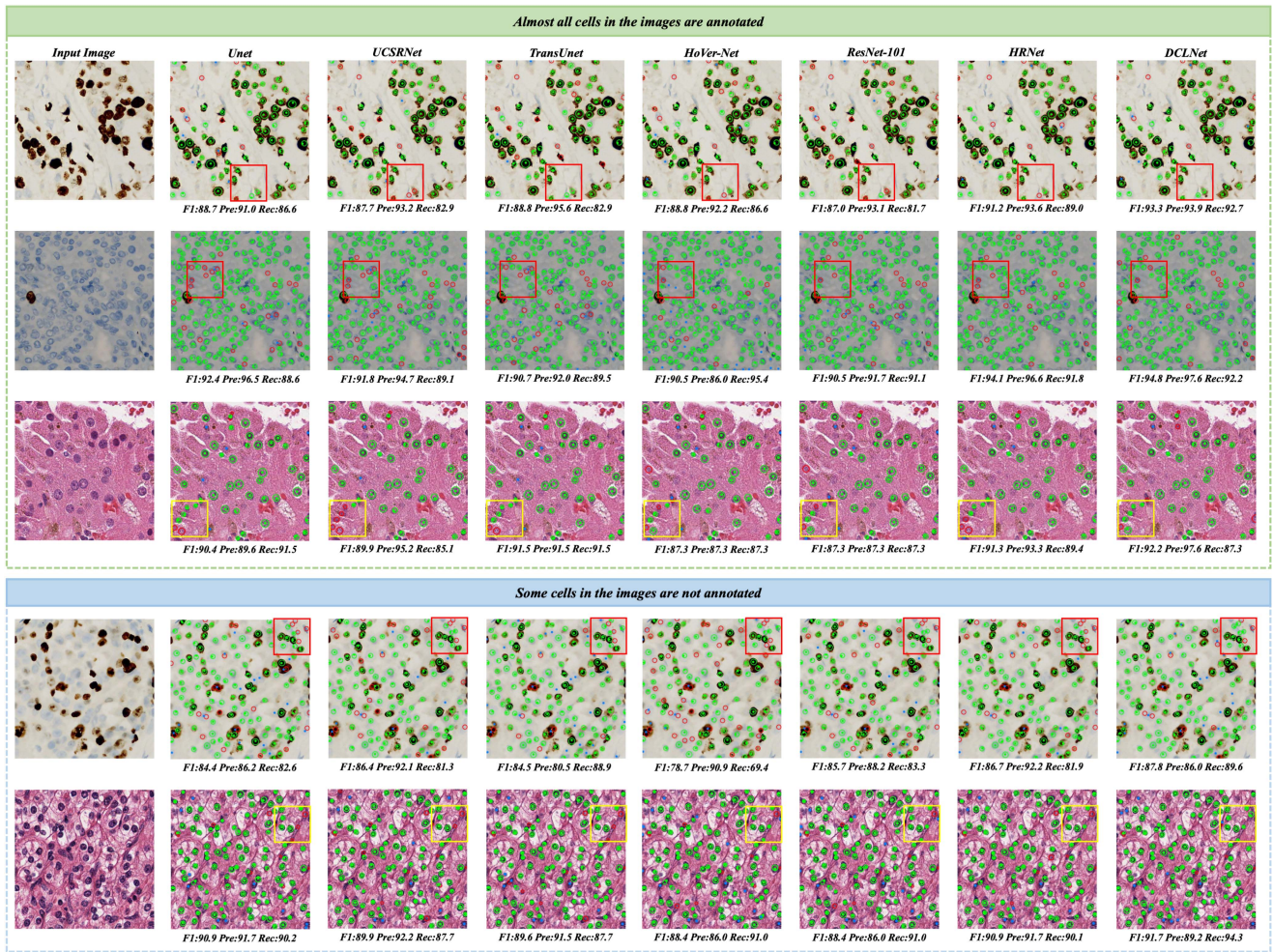


Fig. 6. Some typical visualization results of six popular methods and the proposed DCLNet. The green and blue points denote true positive (TP) and false positive (FP), respectively. The green and red circles are the ground truth of each cell. The red and yellow boxes highlight some representative comparisons. The images of the first, second, and fourth rows originated from BCData, and the images of the third and fifth rows originated from the nuclei grading dataset.

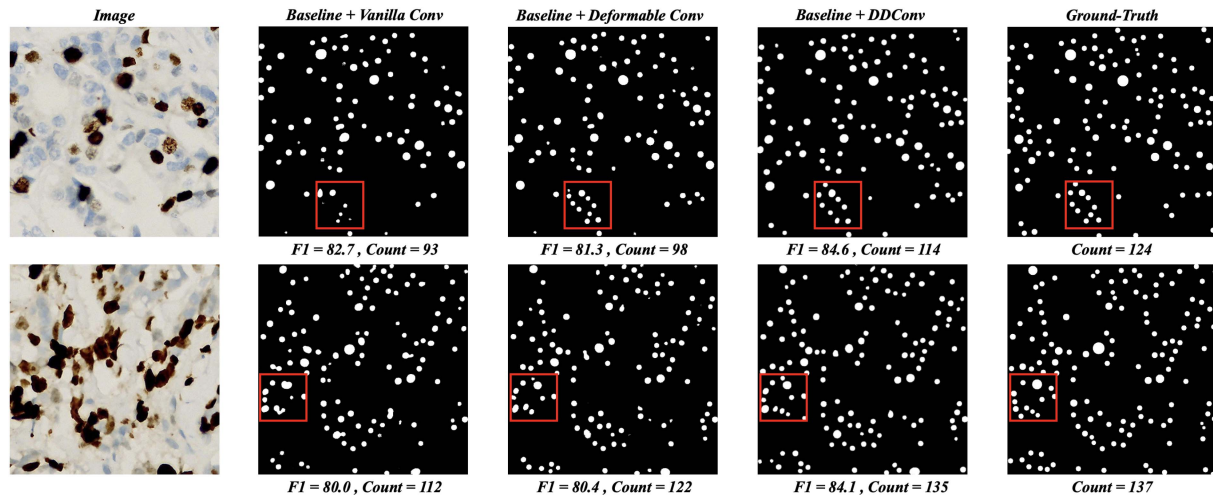


Fig. 7. Comparison of location maps between baseline with vanilla convolution, baseline with deformable convolution, and baseline with DDConv. The sample images originated from BCData.

TABLE II
QUANTITATIVE COMPARISON OF VANILLA CONVOLUTION, DEFORMABLE CONVOLUTION, AND DD CONV ON THE CoNIC DATASET

Method	Conv Type	BCData			CoNIC		
		F1(%)	MAE	RMSE	F1(%)	MAE	RMSE
baseline	Vanilla Conv	87.2	18.4	24.4	83.1	19.2	25.9
	Deformable Conv	87.4	17.8	23.1	83.8	19.0	24.1
	DDConv	87.8	16.0	21.2	84.4	18.1	22.6

The bold values indicate the best performance.

TABLE III
IMPACT OF THE NUMBER AND LOCATION OF DD CONVS ON CELL LOCALIZATION PERFORMANCE

Stage0	Stage1	Stage2	Stage3	F1(%)	MAE	RMSE
✓	✓	✓	✓	87.3	16.7	21.6
	✓	✓	✓	87.4	16.6	22.3
		✓	✓	87.8	16.0	21.2
			✓	87.7	16.2	21.4
✓	✓	✓		86.7	19.5	25.7
✓	✓			87.0	16.8	21.8
✓				86.9	17.8	23.3

The bold values indicate the best performance.

quantitative results in Table II that DDConv achieves the best localization and counting performance on both datasets. This demonstrates that DDConv has better localization performance than vanilla convolution and deformable convolution on images with different staining patterns and cell types.

2) Settings of DDConv: To investigate the effect of the number and position of the DDConv on localization and counting performance, an ablation experiment is performed on the number and position of the DDConv on BCData. The backbone outputs four feature maps of different sizes. According to the feature map size from large to small, the branches are designated as stage 0, stage 1, stage 2, and stage 3, respectively. The experimental results are listed in Table III. It can be seen that the best localization and counting performance is achieved when the DDConv is added to stage 2 and stage 3.

When the hyperparameter θ in (5) is 0, the difference convolution degenerates into the traditional vanilla convolution, which leads to the loss of relative gradient information. Therefore, an ablation study on θ is carried out to explore its impact on cell localization and counting performance. As shown in Fig. 8, we select DCLNet to experiment on the test set of BCData. The best localization and counting performance is achieved when θ is 0.7. When the θ is 0, the difference convolution degenerates into vanilla convolution, and the model performance significantly decreases.

D. Computational Cost

The number of parameters and the computational cost of a method affect its practical application a lot. Therefore, several mainstream methods are chosen to compare with our DCLNet, including ResNet-101 [36], U-Net [32], HRNet [39], and HoVer-Net [38], and the results are shown in Table IV. We use 512×512 resolution images from the CoNIC dataset as input to measure

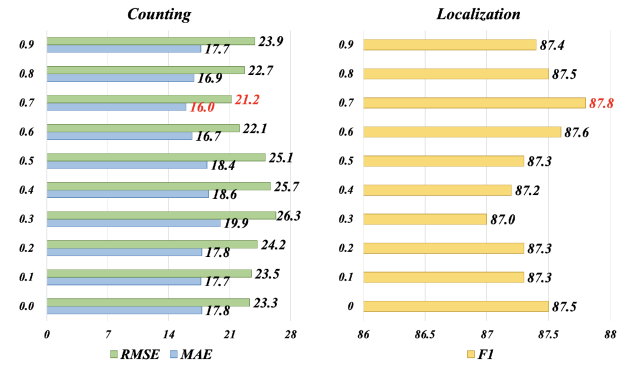


Fig. 8. Impact of different values of θ on cell localization and counting performance.

TABLE IV
COMPARISON OF THE COMPUTATIONAL COST OF SEVERAL REPRESENTATIVE METHODS

Method	ResNet-101	U-Net	HRNet	HoVer-Net	DCLNet
Params(M)	46.6	14.9	66.6	33.6	68.3
FLOPs(G)	116.1	141.4	142.4	628.4	143.1

We measure GFLOPs with 512×512 resolution images as input.

TABLE V
PRE-EXPERIMENTS ARE CONDUCTED ON AN INSTANCE SEGMENTATION DATASET USING THE DENSITY MAP, POINT MAP, AND PSI MAP

Method	Map Type	F1(%)	MAE	RMSE
baseline	Density Map	85.6	4.9	6.9
	Point Map	83.6	10.1	14.3
	PSI Map	90.5	4.7	6.4

The results show that the PSI map with scale information achieves the best performance in both localization and counting.

the GFLOPs of all methods. As indicated by the data in Table IV, DCLNet improves the localization performance on the CoNIC dataset by 5.4%, 4.5%, 1.6%, and 1%, respectively. Meanwhile, the GFLOPs of DCLNet is 23%, 1.2%, and 0.5% higher than ResNet-101, U-Net, and HRNet. In addition, the GFLOPs of DCLNet is 77% lower than HoVer-Net. Based on the above analysis, we can conclude that the performance improvement ratio brought by DCLNet is greater than the increase in GFLOPs.

E. Impact of PSI

1) Advancement of PSI: As mentioned in Section III-B1, we first perform the pre-experiment on the nuclei grading dataset [8], and the results are listed in Table V. As can be seen from Table V, under the same model, the PSI map outperforms the density map and point map in both localization and counting performance, proving that scale information can bring performance improvement to the cell localization task.

To further validate the advantage of the PSI map, experiments are conducted on existing cell localization datasets with point annotations, i.e., BCData. Huang et al. [3] use a strategy of separately predicting positive and negative cells in their experiments. This strategy of separately predicting negative and positive cells avoids the problem of large color variations in cells, but it also

TABLE VI
COMPARISON OF LOCALIZATION PERFORMANCE USING DENSITY MAP AND PSI MAP ON BCData

Method	Label	Positive Localization			Negative Localization			Average Localization		
		F1(%)	Pre(%)	Rec(%)	F1(%)	Pre(%)	Rec(%)	F1(%)	Pre(%)	Rec(%)
SC-CNN [15]	Point	79.8	77.0	82.8	77.8	73.4	82.9	78.8	75.2	82.9
CSRNet [40]	Point	82.9	82.4	83.4	81.4	80.9	81.9	82.2	81.7	82.6
U-CSRNet [3]	Point	86.3	86.9	85.7	85.2	84.4	86.0	85.7	85.6	85.9
U-CSRNet [×]	Point	85.0	84.7	84.4	84.5	84.6	84.7	84.8	84.7	84.6
U-CSRNet	Circle	86.4	87.5	85.3	85.8	86.5	85.1	86.1	87.0	85.2
U-CSRNet*	Point	-	-	-	-	-	-	85.2	85.4	84.9
U-CSRNet*	Circle	-	-	-	-	-	-	86.7	87.9	85.6

The point indicates the use of density maps, and the circle indicates the use of PSI maps. Those with asterisks are unified predictions, which do not distinguish between negative and positive cells, while the rest are classified predictions.

The bold values indicate the best performance.

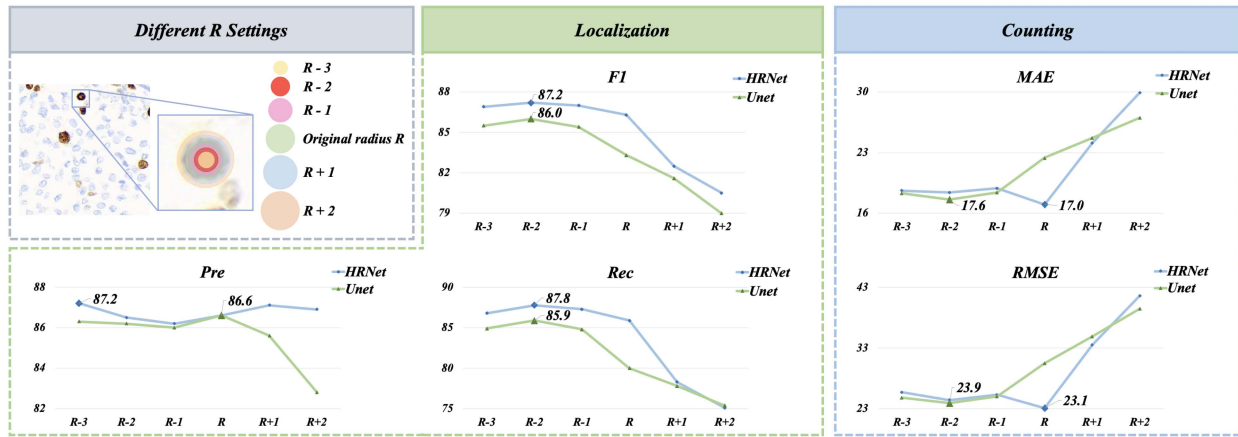


Fig. 9. Comparison of changing the radius R of the masks in the PSI map on cell localization and counting performance. The experimental results are derived from the BCData dataset.

raises two issues: 1) due to differences in staining techniques and different standards for classifying negative and positive cells in different departments, independently predicting models is greatly restricted in practical applications; 2) when only one type of cell needs predicting during the model training process, other types of cells will interfere with the target cells. Therefore, we predict all cells simultaneously, and the results are shown with an asterisk (*) at the bottom of Table VI.

As shown in Table VI, the performance comparison of the PSI map and density map in separately and uniformly predicting negative and positive cells is listed. The following three points can be obtained. 1) When predicting negative and positive cells separately, using PSI maps can improve the localization performance of both negative and positive cells. Compared with the results achieved by U-CSRNet[×], using PSI maps increases the average F1 score, accuracy, and recall by 1.3, 2.3, and 0.6, respectively. 2) When predicting negative and positive cells uniformly, using PSI maps can further improve the cell localization performance, with F1 score, accuracy, and recall increasing by 1.5, 2.5, and 0.7. 3) Using PSI maps can improve the localization performance of cells and greatly increase the accuracy, regardless of whether negative and positive cells are predicted separately or uniformly.

2) *Influence of the Scale*: As mentioned in Section III, the masks of each cell in the PSI map are independent. To further

investigate the influence of the scale of masks on cell localization performance, we conduct an ablation experiment on the radius of each mask in the PSI map. We choose six different radius sizes, with the radius of the mask generated by the method mentioned in Section III used as the reference, denoted as R . Then, the original mask is eroded by 1–3 pixels and dilated by 1–2 pixels. The experiment is conducted on BCData, and the results are shown in Fig. 9. For localization, Unet and HRNet achieve the best F1 and Rec at $R-2$. For counting, Unet achieves the minimum MAE and RMSE at $R-2$, while HRNet achieves the optimal performance at R . However, the localization performance of HRNet at R is much lower than that at $R-2$, so the radius size of all masks applied in this paper is set to $R-2$.

V. CONCLUSION

In this paper, we addressed two unresolved challenges in cell localization. Our novel gradient-aware and shape-adaptive Difference-Deformable Convolution (DDConv) can extract the edge information of lightly stained cells to overcome the challenge of lightly stained cells. meanwhile, adaptively adjusting the shape of the convolutional kernel to overcome the challenge of the large variability in cell shape. And, for the first time, we

propose a more effective and computationally accurate Pseudo-Scale Instance map for cell localization, which addresses three challenges: 1) the inability to avoid the overlapping challenge; 2) the complex post-processing algorithm to locate each cell; 3) the lacking cell scale information. Through extensive experiments, it is demonstrated that the proposed method has significant advantages over other methods and achieves the latest SOTA level. Therefore, we believe that the PSI map and DDConv are expected to become important tools in the field of cell image processing, with broad application prospects in life science research and medical diagnosis, among other fields.

REFERENCES

- [1] Y. Chen, D. Liang, X. Bai, Y. Xu, and X. Yang, "Cell localization and counting using direction field map," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 359–368, Jan. 2022.
- [2] M. M. Alam and M. T. Islam, "Machine learning approach of automatic identification and counting of blood cells," *Healthcare Technol. Lett.*, vol. 6, no. 4, pp. 103–108, 2019.
- [3] Z. Huang et al., "BCData: A large-scale dataset and benchmark for cell detection and counting," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 289–298.
- [4] M. Tofghi, T. Guo, J. K. Vanamala, and V. Monga, "Prior information guided regularized deep learning for cell nucleus detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2047–2058, Sep. 2019.
- [5] B. Li, J. Chen, H. Yi, M. Feng, Y. Yang, and H. Bu, "Exponential distance transform maps for cell localization," 2023, *arXiv:22275958.v3*.
- [6] Y. B. Hagos, P. L. Narayanan, A. U. Akarca, T. Marafioti, and Y. Yuan, "ConCORDe-Net: Cell count regularized convolutional neural network for cell detection in multiplex immunohistochemistry images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 667–675.
- [7] S. E. A. Raza et al., "Deconvolving convolutional neural network for cell detection," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 891–894.
- [8] Z. Gao et al., "Nuclei grading of clear cell renal cell carcinoma in histopathological image by composite high-resolution network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 132–142.
- [9] S. Graham et al., "Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 684–693.
- [10] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1313–1321, May 2016.
- [11] H. Kutlu, E. Avcı, and F. Özyurt, "White blood cells detection and classification based on regional convolutional neural networks," *Med. Hypotheses*, vol. 135, 2020, Art. no. 109472.
- [12] B. Ma, J. Zhang, F. Cao, and Y. He, "MACD R-CNN: An abnormal cell nucleus detection method," *IEEE Access*, vol. 8, pp. 166658–166669, 2020.
- [13] X. Du et al., "SDoF-Net: Super depth of field network for cell detection in leucorrhea micrograph," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1229–1238, Mar. 2022.
- [14] X. Pan et al., "Cell detection in pathology and microscopy images with multi-scale fully convolutional neural networks," *World Wide Web*, vol. 21, pp. 1721–1743, 2018.
- [15] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.
- [16] A. Mahbod, G. Schaefer, G. Dorffner, S. Hatamikia, R. Ecker, and I. Ellinger, "A dual decoder U-Net-based model for nuclei instance segmentation in hematoxylin and eosin-stained histological images," *Front. Med.*, vol. 9, 2022, Art. no. 978146.
- [17] Q. Zhu, Y. Wang, L. Yin, J. Yang, F. Liao, and S. Li, "SelfMix: A self-adaptive data augmentation method for lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 683–692.
- [18] Q. Zhu, Y. Wang, B. Du, and P. Yan, "OASIS: One-pass aligned atlas set for medical image segmentation," *Neurocomputing*, vol. 470, pp. 130–138, 2022.
- [19] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
- [20] S. Huang, Z. Lu, R. Cheng, and C. He, "FaPN: Feature-aligned pyramid network for dense image prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 864–873.
- [21] J. Xie, R. Zhu, Z. Wu, and J. Ouyang, "FFUNet: A novel feature fusion makes strong decoder for medical image segmentation," *IET Signal Process.*, vol. 16, no. 5, pp. 501–514, 2022.
- [22] X. Li, Z. Xu, X. Shen, Y. Zhou, B. Xiao, and T.-Q. Li, "Detection of cervical cancer cells in whole slide images using deformable and global context aware faster RCNN-FPN," *Curr. Oncol.*, vol. 28, no. 5, pp. 3585–3601, 2021.
- [23] Z. Yu et al., "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5295–5305.
- [24] X. Yang et al., "A novel image deep learning-based sub-centimeter pulmonary nodule management algorithm to expedite resection of the malignant and avoid over-diagnosis of the benign," *Eur. Radiol.*, pp. 1–14, 2023, doi: [10.1007/s00330-023-10026-2](https://doi.org/10.1007/s00330-023-10026-2).
- [25] Z. Su et al., "Pixel difference networks for efficient edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5117–5127.
- [26] N. S. Punn and S. Agarwal, "Inception U-Net architecture for semantic segmentation to identify nuclei in microscopy cell images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–15, 2020.
- [27] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7287–7296.
- [28] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2022, pp. 23–33.
- [29] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style convnets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13733–13742.
- [30] J. Gao, T. Han, Y. Yuan, and Q. Wang, "Learning independent instance maps for crowd localization," 2020, *arXiv:2012.04164*.
- [31] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
- [32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
- [33] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "DAE-Former: Dual attention-guided efficient transformer for medical image segmentation," in *Proc. Int. Workshop Predictive Intell. Med.*, 2022, pp. 83–95.
- [34] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
- [35] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [37] M. Jahanifar, N. Z. Tajeddin, N. A. Koohbanani, and N. M. Rajpoot, "Robust interactive semantic segmentation of pathology images with minimal user input," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 674–683.
- [38] S. Graham et al., "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, 2019, Art. no. 101563.
- [39] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
- [40] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1091–1100.