IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 26 July 2023, accepted 5 August 2023, date of publication 14 August 2023, date of current version 17 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3304993

RESEARCH ARTICLE

MSCA-UNet: Multi-Scale Convolutional Attention UNet for Automatic Cell Counting Using Density Regression

LIKE QIAN[1], WEI QIAN[2], DINGCHENG TIAN[1], YAQI ZHU[1], HENG ZHAO[1], AND YUDONG YAO[3], (Fellow, IEEE)

[1]Research Institute for Medical and Biological Engineering, Ningbo University, Ningbo 315211, China
[2]Department of Electrical and Computer Engineering, The University of Texas at El Paso, El Paso, TX 79968, USA
[3]Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

Corresponding author: Yudong Yao (yaoyudong@nbu.edu.cn)

ABSTRACT The quantification of cell numbers in microscopy images plays a vital role in biomedical research and medical diagnosis. Presently, deep regression networks are widely employed to generate cell density maps, and the number of cells is obtained by integrating the density maps. However, automating cell counting remains challenging due to the variability in cell morphology, the diversity of cell types, and the interference of image backgrounds. This paper aims to address the central question: 'Can we design a robust and efficient deep learning model that can effectively count cells in microscopy images, regardless of these challenges?' To tackle this issue, we propose a novel multi-scale convolutional attention UNet (MSCA-UNet) based on density regression. Compared with other advanced density regression methods, our method introduces two key innovations. Firstly, we employ an MSCA block with multi-scale interaction ability as an encoder component, which, when combined with spatial attention, enhances the extraction of cell details and spatial information. Secondly, the design of the asymmetric UNet allows the encoder to extract more global information and better understand the image. In the meantime, using smaller convolutional kernels and strides in the decoder helps to restore image details and edge information, resulting in improved network performance. Our method outperformed other advanced methods on three publicly available benchmark cell datasets, including the synthetic bacterial (VGG) dataset, the modified bone marrow (MBM) dataset, and the human subcutaneous adipose tissue (ADI) dataset.

INDEX TERMS Automatic cell counting, microscopy images, density map, multi-scale convolutional attention.

I. INTRODUCTION

Image-based cell counting is a crucial aspect of biomedical research and medical diagnosis. Specifically, cell numbers in microscopy images have the potential to predict the presence of diseases, assist physicians in disease staging [1], shed light on cellular and molecular mechanisms [2], [3], and provide valuable information for a multitude of other applications [4], [5], [6]. For instance, a low white blood cell count can indicate susceptibility to various diseases, including malaria, autoimmune diseases, immunodeficiency diseases, blood diseases, and cardiovascular diseases [7], [8], [9]. Manual cell counting in microscopy images is a tedious and time-consuming task, and it is susceptible to subjective errors due to the large number of cells and overlapping distribution in images. Therefore, developing a framework for automated cell counting of different cell types and images is of great value. While several automated cell counting methods have been proposed over the past few decades, an efficient and general automated cell counting framework based on images remains a challenging task due to issues such as variations in cell types, sizes, shapes, overlapping, different

The associate editor coordinating the review of this manuscript and approving it for publication was Chulhong Kim.

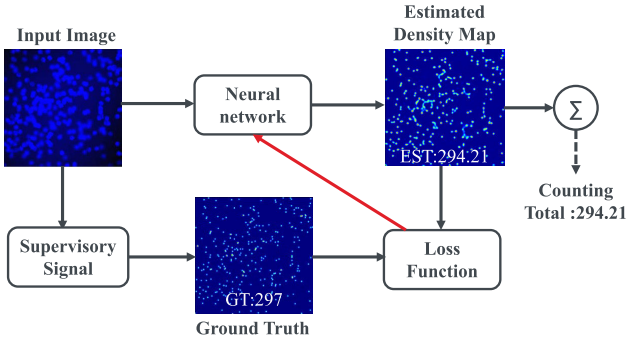

FIGURE 1. Cell counting framework workflow.

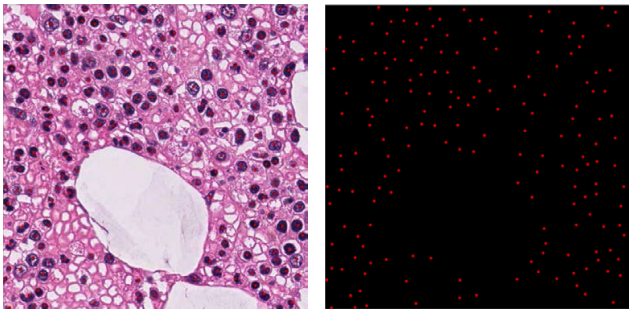

FIGURE 2. Left: Sample image with point annotations from the Modified Bone Marrow (MBM) dataset [19]. Right: Sample label used during training.

staining techniques, diverse image acquisition devices, low image contrast, and background noise interference. Given these challenges, the central question we aim to address in this study is: 'Can we develop a robust and efficient deep learning model that can effectively count cells in microscopy images, handling the variations in cell types, sizes, shapes, staining techniques, imaging devices, image contrast, and background noise?' This question forms the backbone of our research and the motivation for the method proposed in this paper.

In recent years, density regression-based methods have been widely used in cell counting tasks [1], [10], [11], [12], [13]. This method first generates a density map and then calculates the final cell count result by integrating the density map, the workflow of the cell counting framework is illustrated in Fig. 1. This is currently one of the most popular cell counting methods. Advanced models mostly adopt this strategy to generate density maps and integrate them to obtain cell count results. In addition to providing cell count results, this method can also obtain the spatial density distribution of cells. The output density map can provide more supervision information, which is helpful for model convergence. Currently, a point annotation method [10], [11], [14] is used for image-based cell counting benchmarks. This method uses a single pixel to represent each cell, setting the center value of each cell to 1 and the rest of the area to 0. Due to the large number of cells in a single image, fully annotating each cell would require a lot of time. The point annotation method can effectively reduce the workload. As shown in Fig. 2, the center point of each cell is represented by a single pixel, and these point annotations are considered as density maps.

Density map estimation is a pixel-level prediction task that involves dealing with different sizes and shapes of cells, various cell types, and image background interference. These factors pose significant challenges in cell counting. Furthermore, the mapping process from the original image to the density map involves learning labels annotated by sparse points, which can cause the network to prioritize background pixels over foreground pixels. Therefore, the designed network must have the ability to capture spatial details and handle cells of different sizes. To address these challenges, we propose a novel multiscale convolutional attention U-shaped network (MSCA-UNet) based on UNet [15]. We draw inspiration from SegNeXt [16] and utilize the powerful ability of a visual attention network [17] in semantic segmentation. Specifically, we incorporate a powerful encoder, multiscale interaction, and spatial attention to improve the accuracy of density estimation. To handle cells with diverse shapes and scales, we develop a multiscale attention module that can fully extract spatial information from multiscale feature maps. This module enhances the model's ability to represent multiscale features more accurately and establish longer-distance feature dependencies among multiscale channel attention, enabling the network to focus on regions of interest in cells and suppress background noise interference. By integrating multiscale contextual spatial information, the proposed multiscale convolutional attention U-shaped network effectively addresses the challenges of density map estimation. We choose UNet as the regression model because of its success in image segmentation [15], [18]. Furthermore, we combine the advantages of attention and multiscale features to improve the network's performance. With these modifications, our proposed MSCA-UNet can be a promising approach for accurate density map estimation in cell counting.

To examine the effectiveness of our proposed method, we test it on three publicly available datasets. The adipocyte dataset [14], the synthetic VGG dataset [10], and the modified bone marrow (MBM) dataset [20].

Our research has the following contributions.

- We design an asymmetric U-shaped encoder-decoder structure named MSCA-UNet, which integrates UNet with a multi-scale convolution attention module for cell counting. The encoder captures attention from local to global regions, and, in the decoder, global features are upsampled to the input resolution for corresponding pixel-level segmentation prediction.
- We propose an MSCA module to handle changes in cell morphology and capture spatial information of cells.
- We adapt the visual attention network from semantic segmentation to cell counting.

The remainder of this paper is organized as follows. Section II provides a review of relevant literature, while Section III describes our proposed MSCA-UNet method. Section IV presents a description of the datasets utilized in this study, as well as implementation details of the proposed method, and compares its results with other advanced methods. Finally, Section V summarizes the paper and outlines potential future work.

II. RELATED WORK

In this section, we review the work related to cell counting. Cell counting methods can be classified into two categories, detection-based and regression-based.

A. DETECTION-BASED RESEARCH

Detection-based methods employ detectors to locate individual cells in the image and estimate the cell count based on the detected results. Traditional detection-based methods include feature extraction [19], morphological processing [21], multi-curvature cell nucleus contour model [22], a combination of region growing and Markov random field algorithms [23], and the Hough transform [24]. In recent years, with the progress in deep learning, convolutional neural networks have been utilized in various cell detection and counting work [25], [26], [27], [28], [29], [30], [31], [32]. For instance, Falk et al. [25] trained a fully convolutional neural network (UNet) combined with non-maximum suppression to count the number of cells in the image. Arteta et al. [28] introduced a tree-structured discrete graphical model extremal region trees (ERT), that selects and labels a set of non-overlapping regions in the image for detecting overlapping cell instances in microscopy images based on global optimization of classification scores. Zhu et al. [29] developed a cell detection and counting method based on the fully convolutional network (FCN), which can handle different types of cell data and cover most advanced microscopy images, such as bright field, pathology-stained material, and electron microscopy. Xia et al. [32] employed a two-stage detection network (Faster Region-convolutional neural network) to generate region proposals at the feature map level using a heuristic method (selective search) to determine potential cell regions, followed by classification and regression on region proposals, and verified on a leukocyte dataset. Zhang et al. [31] proposed to first use YOLOv3 to detect various cell types in the image and then use density estimation algorithms to count specified cell types, which can achieve higher accuracy than using YOLOv3 alone for detection and counting. These methods highly depend on the accuracy of cell detection results, and cell detection and counting remain challenging tasks due to cell occlusion, shape changes, and image background noise. Moreover, training such models requires individual cell labeling, which is a time-consuming and expensive process for high-density cell images. Therefore, detection-based methods are appropriate for situations where cell distribution is sparse or a small number of cells overlap.

B. REGRESSION-BASED RESEARCH

There are two types of regression-based methods for cell counting, direct counting and density-based counting. The former uses deep learning models to map input images to the number of cells, while the latter generates a density map that is integrated to obtain the cell count.

1) DIRECT REGRESSION

This method only focuses on the regression results of cell numbers and disregards the position information of cells [33], [34], [35]. For instance, Khan et al. [33] used a deep convolutional neural network model to learn cell-related features directly from images, replacing the handcrafted feature selection and achieving end-to-end learning from raw microscopy images to cell numbers. They demonstrated that combining automatic computation of target bounding boxes and conditional random fields (CRF) with temporal information can significantly enhance cell counting performance. However, this method uses a patch-based approach, and before calculating the total number of cells, all image patches must be inferred, which can cause redundant estimates due to patch boundary crossings. Similarly, Xue et al. [34] designed a supervised learning framework using a convolutional neural network and also adopted the small image block idea for cell counting. This method increased the number of training samples, and the total number of cells in the complete image was obtained by summing all the small image blocks. Nevertheless, this method does not solve the problem of redundant estimation caused by patch boundary crossings. Furthermore, the above-mentioned studies did not solve the problem of network error focusing on background noise. Aich and Stavness [35] introduced class activation maps to visualize the feature maps of the final layer of the network and discovered that the network incorrectly focused on some background features while disregarding some regions with indistinct cell features. Learning background regions resulted in counting bias in the network. This problem occurs because weakly supervised regression networks only constrain the image mapping to be as close as possible to the true counting value of the target without displaying information about the target object attributes in the image, causing the network to learn background region features. The authors established a loss function of target position and target counting value through class activation maps to guide the network to learn the target that requires attention and avoid erroneous responses in the background region.

2) DENSITY-BASED REGRESSION

Traditional regression methods for density estimation rely on handcrafted image features for training the model. For instance, Lempitsky and Zisserman [10] proposed using density estimation to learn the linear mapping from local features (such as scale-invariant feature transform features) to the corresponding density map, and then predict the cell count. To simplify the learning of the linear mapping,

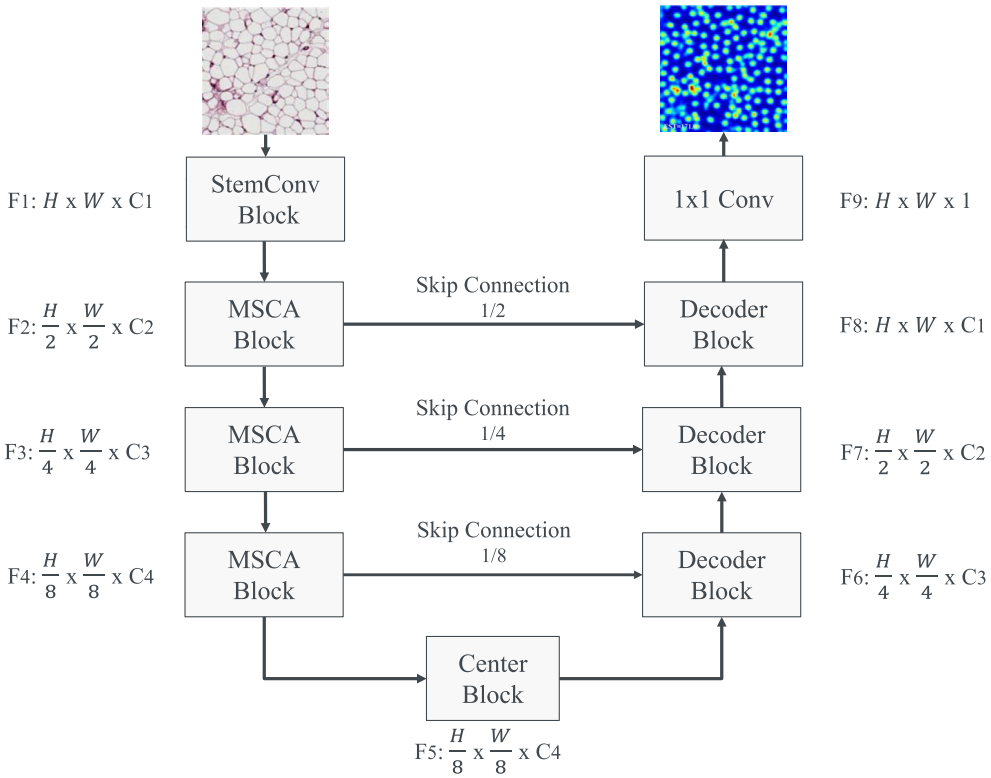

FIGURE 3. MSCA-UNet is composed of an encoder, bottleneck, decoder, and skip connections. The encoder is constructed based on MSCA block, while the bottleneck and decoder are similar to the traditional UNet [35].

Fiaschi et al. [13] developed a structured learning framework of regression random forests to learn the non-linear mapping. This method learns the mapping relationship between all patch features and the relative positions of all objects within the patch, and then generates patch density maps through Gaussian kernel density estimation. Similarly, Pham et al. [36] proposed a structured learning framework of random decision forests to address the density estimation problem and introduced a robust density estimator with three improvements, enhancing accuracy by using crowding priors, increasing estimation speed by using efficient forest reduction methods, and reducing annotation work by using semi-automatic training. The quality of feature extraction methods significantly impacts the performance of these methods.

With the emergence of deep learning neural networks in various fields, researchers have started to use convolutional neural networks to extract features and achieve end-to-end density estimation, replacing models that rely on manually crafted image features [1], [11], [12], [14], [37], [38]. Xie et al. [11] used fully convolutional regression networks (FCRNs) to regress the cell spatial density map of the entire image. With the property of fully convolutional networks, this method can predict density maps of any input image size. By using CNNs to extract image features and output density maps, this work achieved superior performance compared to traditional cell counting methods, especially when dealing with microscope images with severe cell overlap. Cohen et al. [14] combined the ideas of density map estimation [10], fully convolutional network processing [11], and counting everything in receptive fields [39], proposing a regression network that counts cells in image blocks. This method improved accuracy compared to [11], but has the limitation of losing spatial details. He et al. [38] designed a deep supervision density regression network to estimate the number of cells in microscope images. Unlike other density regression methods, this method uses concatenated fully convolutional regression networks (C-FCRN) and employs multi-scale image features to enhance feature extraction. Additionally, they used auxiliary convolutional neural networks (AuxCNNs) to assist in training the intermediate layers of C-FCRN, further improving cell counting performance by learning and supervising the intermediate layers. Jiang and Yu [12] proposed a cell counting network with detail and context paths, where the detail path extractd rich spatial details and the background path obtained multi-scale features using spatial pyramid pooling. They designed a feature fusion module to merge the high-level feature maps of the two paths, achieving superior counting performance. In particular, they also validated the model's generalization ability to other counting datasets using a crowd dataset.

In summary, to address the limitations in the previous work, we propose MSCA-UNet and adopt a density-based regression approach for cell counting, which reduces the

influence of factors such as background noise and cell shape variability in cell images.

III. METHOD

In this section, we introduce the overall framework of our proposed MSCA-UNet network and describe each module and the loss function used.

A. ARCHITECTURE

To address the central question of our research, we designed a robust and efficient deep learning model, MSCA-UNet, which is an asymmetric U-shaped network consisting of an encoder, bottleneck, decoder, and skip connections (as shown in Fig. 3), is designed to handle the inherent challenges in cell counting tasks. The deeper architecture of the encoder compared to the decoder allows for a more detailed extraction of cell features and global image information, which are crucial in cell counting tasks where cell morphologies vary significantly and image backgrounds often present interference. The input image is first passed through a StemConv block in the encoder, which transforms it into a set of high-dimensional feature vectors and captures local features in the image. This transformation establishes a foundation for subsequent feature extraction and task execution. We then employ overlapping block embedding, which divides the image into overlapping blocks of size 3×3 and converts them into sequence embeddings. This method is commonly used in visual Transformers [40] since it maps each block to a low-dimensional vector and preserves spatial structure and local features in the image, allowing the model to better learn the semantic information of the input data. Next, the transformed patch tokens pass through several MSCA blocks to generate hierarchical feature representations. Each MSCA block, as shown in Fig. 4, consists of batch normalization (BN), attention module, and feed-forward network (FFN). The attention module, as shown in Fig. 5(a), is composed of a 1×1 convolution, Gaussian Error Linear Units (GELU) [41] activation function and MSCA module. The FFN, as shown in Fig.5(b), is composed of a 1×1 convolution, depth-wise convolution, and GELU activation function.

High-resolution feature maps contain low-level information that may harm object recognition, while low-resolution feature maps have higher semantic granularity but lack spatial information. For the decoder, we were inspired by the UNet series of studies [15], [25] and designed an asymmetric decoder similar to [15], consisting of convolution and upsampling operations. By using skip connections to fuse features from different layers of the encoder with corresponding layers of the decoder, context information, and multi-scale information are combined to preserve the spatial structure and detail information of the image, compensating for the loss of spatial information caused by downsampling and ultimately improving the network's expressive power and prediction accuracy. Each upsampling resizes the adjacent dimension of the feature map to a resolution of $\times2$, and the final output is a density map with the same size as the original image.

Specifically, our model operates by considering the overall distribution of cells in the density map, rather than the precise location of individual cells. It estimates the total cell count by integrating the values of the predicted density map, rather than pinpointing the exact position of each cell. This design choice was made to optimize the model for estimating total cell counts, which is typically the primary objective in cell counting tasks. However, it means that the model may not always accurately predict the precise locations of individual cells, especially in regions of high cell density. The model prioritizes the overall cell distribution over the exact location of each cell.

B. ENCODER

In the encoder, the input image is first processed using the StemConv block to generate high-dimensional feature vectors for improved data processing in subsequent steps, resulting in a $\mathrm{H}\times\mathrm{W}$ dimensional C_1 vector output. This vector is then input into the MSCA block for feature learning, resulting in a feature dimension that is double that of the input and a resolution that is half of the input. This process is repeated three times in the encoder, which has a structure similar to that of the Vision Transformer [42]. Unlike traditional multi-head self-attention modules, the MSCA block does not use self-attention mechanisms. Instead, it employs a multi-scale convolutional attention module [16] (see Fig. 6), which comprises three components, depth-wise convolution, multi-branch depth-wise stripe convolutions, and 1×1 convolution. These three parts respectively enable the aggregation of local information, the capture of multi-scale context, and the modeling of relationships between different channels. The output of the MSCA module can be expressed as,

$$\mathrm{Att}_m = \mathrm{Conv}_{1\times1}\left(\sum_{i=0}^{3}\mathrm{Scale}_i(\text{DW-}\mathrm{Conv}(F))\right) \quad (1)$$

$$\mathrm{Out} = \mathrm{Att}_m \otimes F. \quad (2)$$

F denotes the input feature, and DW-Conv refers to depth-wise convolution. $Scale_i$, where $i \in 0, 1, 2, 3$, represents the i-th branch. $Scale_0$ is an identity connection, Att_m denotes the attention map and out is the output, $\otimes$ is computed using element-wise matrix multiplication. To reduce the computational cost, two depth-wise separable convolutions are used in each branch, as per [43], to approximate the standard depth-wise convolution with a large kernel.

C. BOTTLENECK

The bottleneck layer in our model is composed of two 3×3 convolutional layers and a ReLU activation function, which aims to compress the multi-scale feature maps from the encoder into a high-dimensional feature vector and then pass it to the decoder for expansion to the original input image size.

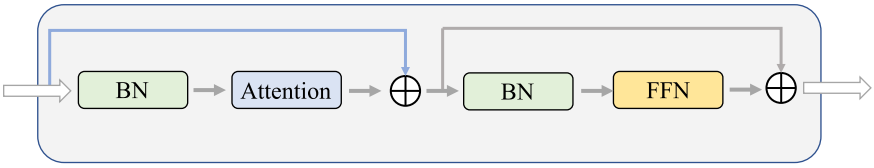

FIGURE 4. The multi-scale convolutional attention block, consists of batch normalization, attention module, and feed-forward network.

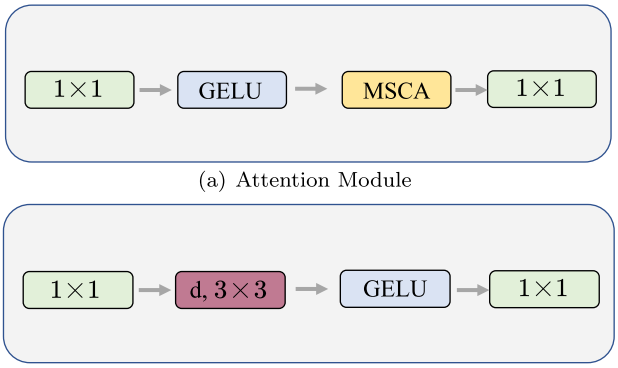

FIGURE 5. The architectures of (a) the attention module, which includes a 1 × 1 convolution, Gaussian Error Linear Units activation function and MSCA module, and (b) the FFN, composed of a 1 × 1 convolution, depth-wise convolution, and GELU activation function.

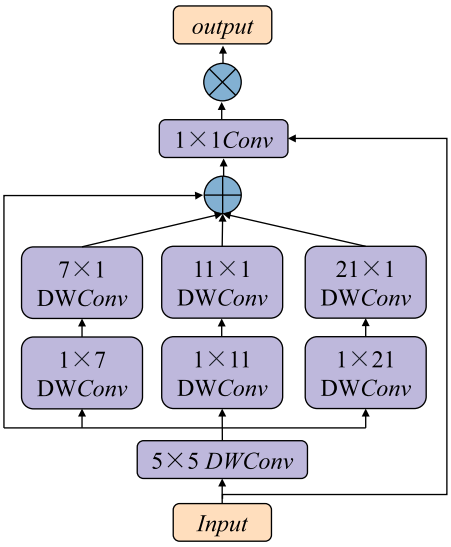

FIGURE 6. The MSCA module, which comprises three components, depth-wise convolution, multi-branch depth-wise stripe convolutions, and 1×1 convolution.

It should be noted that the feature dimension and resolution remain unchanged in this layer.

D. DECODER

Similar to UNet [15], the upsampling module is used to upsample the feature maps from the encoder to the original image resolution via nearest neighbor interpolation. To better recover the spatial information and details of the original image, a 3×3 convolutional layer is employed to fuse the features from the decoder with those from the skip connections of the encoder. The use of skip connections fuses the multiscale features from the encoder with the upsampled features and reduces spatial detail loss caused by downsampling, connecting low-level features with high-level features. Finally, the cascaded and upsampled features have a consistent channel dimension.

E. LOSS FUNCTION

We use the mean squared error (MSE) loss to train the network and evaluate its performance, with the following formula,

$$loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 \tag{3}$$

here, $\hat{y}_i$ represents the predicted value of the i-th sample, y_i represents the true value of the i-th sample, N represents the number of samples. A smaller $loss_{MSE}$ indicates a smaller difference between the model prediction and the true value, i.e., higher model accuracy. Therefore, we use $loss_{MSE}$ as the optimization objective to minimize the difference between the predicted value and the true value, and obtain better prediction performance.

F. DISCUSSION OF THE MULTI-SCALE APPROACH

We introduce MSCA-UNet, an approach effective in handling cellular imagery across a broad range of scales and complexities. Whether the task necessitates high-resolution imaging to capture miniature structures and cellular details, or lower resolutions suffice, our method proves effective. MSCA-UNet employs a multi-scale approach, which enhances both the depth and breadth of the network, thereby bolstering the model's representational capacity. In this study, we enhance prediction accuracy by performing multi-scale extraction and feature integration on the encoder, which allows the network to learn a more comprehensive and diverse set of feature representations. The efficacy of this method is corroborated by other research. For instance, Gudhe et al. [44] proposed a multi-level dilated residual deep neural network that successfully captures local and contextual features, and performs effective segmentation of lesions or tumors across multiple biomedical imaging modalities. Therefore,

TABLE 1. Details of the three datasets.

Dataset	Images	Image size	Count statistics	Type
VGG [34]	200	256×256	174±64	Synthetic
MBM [17]	44	600×600	126±33	Real
ADI [33]	200	150×150	165±44	Real

multi-scale methods not only enhance the predictive accuracy of the model but also improve the model's flexibility, robustness, and generalization capabilities.

IV. EXPERIMENTS

In this section, we discuss the datasets used, implementation details of our network training, experimental configuration, and evaluation metrics. We also compare our proposed cell counting network, MSCA-UNet, with state-of-the-art methods and conduct an ablation study to analyze the impact of model scale on performance.

A. DATASETS

In this study, we used three distinct microscopy image datasets to evaluate the performance of the MSCA-UNet: the synthetic bacteria (VGG) dataset [10], the modified bone marrow (MBM) dataset [20], and the human adipose tissue (ADI) dataset [14]. We selected these datasets due to their extensive variety and complexity, encapsulating a broad spectrum of cell types, cell morphologies, and image backgrounds. Moreover, there exist additional datasets, such as the Dublin Cell Counting (DCC) dataset [45] and the mouse blastocyst (MBC) dataset [46]. The DCC dataset encompasses a diverse range of cell types and counts, thereby providing an excellent basis for examining the scalability and adaptability. Similarly, the MBC dataset, with its 3D context, offers an ideal platform for testing the model's capability in processing volumetric data. However, in the scope of this study, our emphasis was placed on handling 2D images, and hence, the datasets chosen, VGG, MBM, and ADI, better reflect the challenges our model is designed to address.

Table 1 shows the details for these datasets. Image size is represented by pixel, count statistics indicate the average cell count and corresponding variance for each image, as Fig. 7 displays example images from the three datasets.

1) VGG CELL

The VGG dataset was created by Lempitsky and Zisserman [10] using the method proposed by Lehmussola et al. [47]. It comprises 200 synthetic images, each with a size of 256×256 pixels and an average of 174±64 cells per image. These synthetic images simulate bacterial cells in fluorescent microscope images and exhibit characteristics such as cell overlap, shape variability, defocus blur, and halo effects, closely resembling real-world microscopy images.

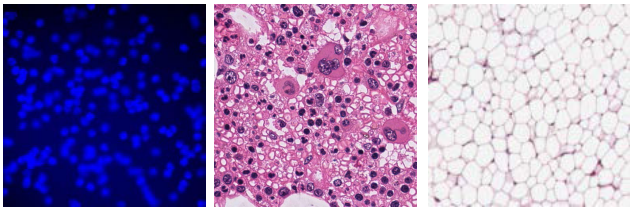

FIGURE 7. The example images of the three datasets used in this study, from left to right, are VGG [10], MBM [20], and ADI [14].

2) MBM CELL

The MBM dataset was modified by Cohen et al. [14] from the dataset published by Kainz et al. [20]. It includes 44 hematoxylin-eosin-stained microscope images of human bone marrow tissue from 8 different patients. Each image has a size of 600×600 pixels and an average of 126±33 cells per image. The images in this dataset have an uneven background and a wide variety of cell shapes, representing a challenging scenario for cell counting automation.

3) ADI CELL

The ADI dataset is a human subcutaneous adipose tissue dataset [14] constructed by the genotype-tissue expression (GTEx) consortium [48]. It contains a total of 200 images, each with a size of 150×150 pixels and an average of 165±44 cells per image. Human subcutaneous adipocytes exhibit high morphological diversity, with closely interconnected and variable shapes and sizes, posing a significant challenge for automated cell counting.

B. IMPLEMENTATION DETAILS

1) PREPROCESSING

Our network, originally designed as a fully convolutional network, can accommodate input images of varied dimensions. However, it is noted that our network's encoder comprises three pooling operations, necessitating the input image size to be a multiple of eight. Therefore, we partitioned the MBM cell dataset image into four blocks of 304 × 304 resolution from its original size of 600 × 600. Moreover, we padded the edges of the ADI dataset images from 150×150 to 152×152.

However, in the case of extremely high-resolution images, such as those spanning 4096 × 4096 pixels and with cell densities exceeding 10,000 cells, additional pre-processing steps may be required. Specifically, handling such large-scale images might require their division into smaller patches, each processed independently before collating the results. This consideration arises from the escalated computational demands required for these larger, more densely populated images. Hence, users should keep these factors in mind while applying our model to their specific research tasks.

2) DATA AUGMENTATION

During the network training phase, we applied random horizontal and vertical flipping strategies as a form of

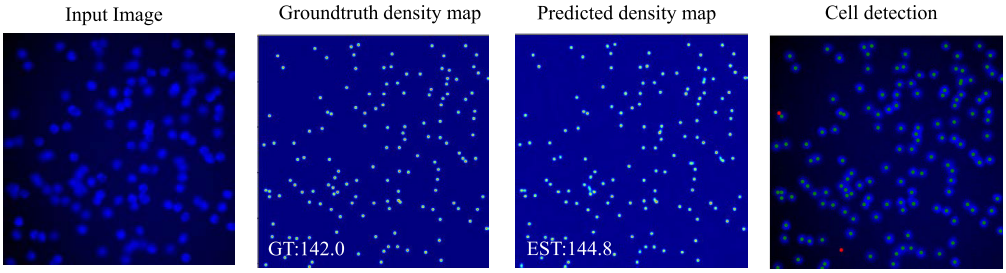

FIGURE 8. Density estimation results of samples in the VGG dataset. Ground truth count: 142.0, predicted:144.8.

data augmentation to increase the variety of our training set and make the model more robust to different cell orientations. To prevent network overfitting, we implemented two regularization strategies in the encoder: DropPath [49] and Dropout [50]. These strategies helped us to mitigate the model's over-reliance on specific features by randomly setting the output features of some neurons to zero during training. This not only reduced the network's complexity but also increased its robustness and generalization ability, allowing it to perform more accurately on unseen data. The specific rates for Dropout and DropPath used were 0.1 and 0.1, respectively. We found that these values effectively balanced the need for model complexity and the risk of overfitting.

3) OPTIMIZATION

We utilized the Adam optimizer [51] to train the network, which combines the ideas of momentum gradient descent and adaptive learning rate, allowing for faster convergence and avoiding getting trapped in local optima. The weight decay value of the Adam optimizer was set to 0.001. Specifically, when training the VGG dataset, the weight decay of the optimizer was set to 0.0001. To facilitate hyperparameter optimization, we utilized the Wandb platform https://wandb.ai/. Specifically, we used Wandb to search for optimal values of hyperparameters such as learning rate, horizontal and vertical flip probabilities, and training epochs by setting search spaces for each hyperparameter. The learning rate interval was set to [0.005, 0.002], and the probability of random horizontal and vertical flipping was set to [0, 1]. The training epoch interval was set to [300, 800]. Then, we used the Bayesian optimization algorithm to search and select the best hyperparameter values in the hyperparameter space, guided by the validation results. Finally, we selected the model generated with the best hyperparameter values as our training result.

C. PERFORMANCE EVALUATION METRICS

We use the mean absolute error (MAE) as the evaluation metric, which is the most commonly used counting performance evaluation metric,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|T_i - P_i| \tag{4}$$

where N is the number of test images, and T_i and P_i are the true and predicted cell counts in the i-th image, respectively. The MAE represents the average absolute error between the true and estimated cell counts of all test images. A lower MAE value indicates higher cell counting accuracy.

D. EXPERIMENT CONFIGURATIONS

For each experiment, we randomly and equally select images from the dataset as training and testing samples, and repeat the experiment 10 times. The experimental results are reported as the mean and variance of the mean absolute error evaluation metric. It is important to note that the network is trained using a pre-trained model on the ImageNet dataset [52], and we compare our method with state-of-the-art techniques on each dataset. Finally, to demonstrate the effectiveness of our proposed method, we conduct ablation studies.

E. COMPARISON WITH OTHER STATE-OF-THE-ART METHODS

In our comparative analysis, we have the following considerations in model selection. First, our focus was on contemporary, state-of-the-art models which have demonstrated exemplary performance in tasks that parallel our research objectives. Additionally, we selected classical models with a proven track record in cell counting tasks. The availability of the models' implementation and the feasibility of their reproduction were key determinants in our final selection. Therefore, we provide a robust, comprehensive, and equitable comparison between our proposed MSCA-UNet and existing methods in the field. Table 2, Table 3, and Table 4, present the comparative results of our method against other state-of-the-art approaches on the VGG, MBM, and ADI datasets. On the VGG dataset, our method outperforms advanced techniques such as CCF proposed by Jiang and Yu [53] and Two-Path Net [12]. The sample prediction results are illustrated in Fig. 8. On the MBM dataset, our method demonstrates comparable performance to the leading method, SAU-Net [54], which integrates a self-attention module to enhance the network's focus on the foreground of the image and improve its performance. SAU-Net reuses low-level details and encodes global information to obtain richer spatial details. The sample prediction results are shown in Fig. 9. On the ADI

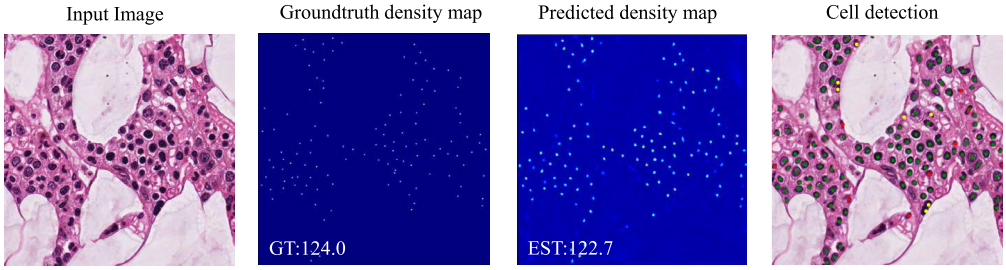

FIGURE 9. Density estimation results of samples in the MBM dataset. Ground truth count: 124.0, predicted:122.7.

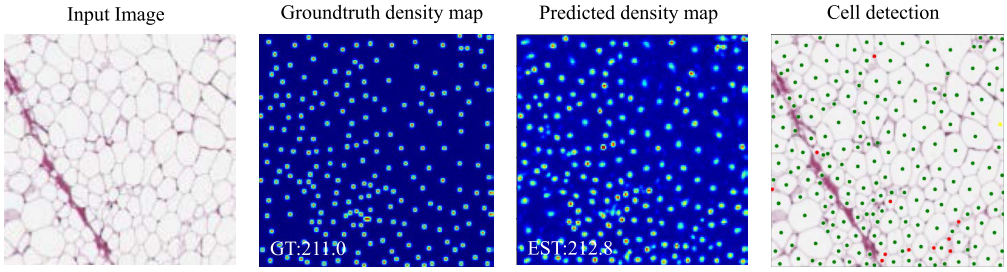

FIGURE 10. Density estimation results of samples in the ADI dataset. Ground truth count: 211.0, predicted:212.8.

TABLE 2. Comparison results on VGG dataset.

Method	$N_{train} = 8$	$N_{train} = 16$	$N_{train} = 32$
Lempitsky et al. [10]	4.9±0.7	3.8±0.2	3.5±0.2
FCRN-A [11]	3.9±0.5	3.4±0.2	2.9±0.2
Count-ception [14]	3.9±0.4	2.9±0.5	2.4±0.4
CCF [53]	2.9±0.2	2.8±0.1	2.6±0.1
Two-Path Net [12]	N/A	2.6±0.2	2.3±0.2
MSCA-UNet(Proposed)	**2.4±0.3**	**2.1±0.2**	**2.0±0.2**

TABLE 3. Comparison results on MBM dataset.

Method	$N_{train} = 5$	$N_{train} = 10$	$N_{train} = 15$
Count-ception [14]	12.6±3.0	10.7±2.5	8.8±2.3
SAU-Net [54]	N/A	N/A	**5.7±1.2**
Two-Path Net [12]	8.2±1.1	**6.9±0.9**	6.0±0.6
MSCA-UNet(Proposed)	**8.0±0.9**	7.1±0.8	5.8±0.7

dataset, our proposed method achieves superior performance. We attribute this to our method's ability to capture spatial details and the interaction of multiscale features, allowing the network to effectively handle challenging problems such as tightly connected, variable in shape and size. An example test case is provided in Fig. 10. It is important to note that our model operates by generating predicted density maps, which are then integrated to produce cell numbers. Therefore, the performance of our model may lack the ability to detect the precise location of individual cells, although it possesses the ability to accurately estimate the overall cell number.

TABLE 4. Comparison results on ADI dataset.

Method	$N_{train} = 10$	$N_{train} = 25$	$N_{train} = 50$
Count-ception [14]	25.1±2.9	21.9±2.8	19.4±2.2
SAU-Net [54]	N/A	N/A	14.2±1.6
Two-Path Net [12]	13.8±0.7	11.6±0.4	10.6±0.3
MSCA-UNet(Proposed)	**11.5±1.2**	**10.5±1.0**	**9.8±0.7**

F. ABLATION STUDY

To demonstrate the importance of the multi-scale convolutional attention (MSCA) module, we conducted an ablation study on the ADI dataset, which is known to exhibit a wide range of cell shape variations and thus better reflects the importance of multi-scale interaction. We adopted the setting used by Guo et al. in VAN [17] and replaced the multiple branch convolutions in MSCA with a single convolution using a large kernel, which we referred to as single-scale convolutional attention(SSCA) module. In addition, we also used U-Net to demonstrate that improvements to the encoder contribute to better cell counting performance. The experimental results are presented in Table 5. By utilizing the MSCA module to further enhance the performance of the model, the counting error of ADI was reduced from 11.0 ± 1.1 to 9.8 ± 0.7.

G. CELL DETECTION

We follow the approach proposed by Xie et al. [11], acquiring cell detection results by identifying local maxima on the density map. In the detection results, green dots signify True Positives (TP), red dots signify False Positives (FP), and

TABLE 5. Ablation study on the impact of MSCA.

Method	MAE
UNet	15.3±2.1
MSCA-UNet w/ SSCA	11.0±1.1
MSCA-UNet w/ MSCA	9.8±0.7

TABLE 6. The comparison results of the cell detection on VGG dataset.

Batch size	Precision (%)	Recall (%)	F1-Score
8	100±0.0	91.25±0.46	0.9532±0.0035
16	100±0.0	90.59±1.06	0.9491±0.0051
32	100±0.0	91.75±0.58	0.9538±0.0073

TABLE 7. The comparison results of the cell detection on MBM dataset.

Batch size	Precision (%)	Recall (%)	F1-Score
5	87.87±0.32	87.27±0.25	0.8733±0.0039
10	88.89±0.90	88.10±0.54	0.8868±0.0067
15	90.11±0.57	89.32±0.45	0.8942±0.0034

TABLE 8. The comparison results of the cell detection on ADI dataset.

Batch size	Precision (%)	Recall (%)	F1-Score
10	97.22±0.77	85.76±0.46	0.9124±0.0073
25	97.81±0.56	85.10±1.11	0.9080±0.0072
50	98.16±0.29	85.40±0.81	0.9128±0.0068

yellow dots signify False Negatives (FN) (refer to Fig. 8, Fig. 9, Fig. 10).

Precision, Recall, and F1-Score are adopted as the metrics of cell detection performance, which are given as:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

Precision measures how many of all coordinates predicted as cells actually exist in the ground truth image. Recall measures how many of all the cells in the ground truth image are accurately predicted. The F1-Score is the harmonic mean of precision and recall, considering both metrics simultaneously. TP represents instances where our model correctly predicts positives. In other words, instances where the predicted point is within a designated radius of the actual coordinates and is indeed a real cell instance (we establish that each actual cell coordinate corresponds uniquely to a predicted coordinate). FP signifies instances where our model inaccurately predicts positives. This means instances where the predicted point and actual coordinate are within the designated radius, but it is not an actual cell instance. FN represents instances where our model inaccurately predicts negatives. This indicates instances where the predicted point and actual coordinate exceed the designated radius, but the instance is a real cell. Notice that, considering the average cell size in our dataset, we decided to set the radius to 10 pixels. This value was determined through a series of trials and evaluations.

We conducted cell detection on three datasets using models trained with different batch sizes to validate the detection performance of the models, as shown in Table 6, Table 7, and Table 8. The best results our models achieved on the VGG, MBM, and ADI datasets are as follows: a precision of 100±0.0%, a recall of 91.75±0.58%, and an F1-Score of 0.9538±0.0073; a precision of 90.11±0.57%, a recall of 89.32±0.45%, and an F1-Score of 0.8942±0.0034; a precision of 98.16±0.29%, a recall of 85.40±0.81%, and an F1-Score of 0.9128±0.0068, respectively.

We found that our model performs exceptionally well on the VGG dataset, where all predicted cell samples were indeed actual cells, although some actual cells were not detected. On the MBM dataset, the key metrics such as Precision, Recall, and F1-Score were generally lower. We believe that this might be due to the smaller size of the MBM dataset and the sparsity of cells, leading to the model not being adequately trained on this dataset. On the ADI dataset, while the Precision score was high, we observed that the model tends to predict the same cell multiple times, which might be the reason for the lower Recall score. Overall, our experimental results indicate that our model has a high probability of correctly identifying real cells in its predictions, but issues of missed detections and incorrect cell predictions persist. This implies that there could be biases in actual cell counting, necessitating further optimization of our model in future work to mitigate these problems.

V. CONCLUSION

This paper presents a novel asymmetric U-shaped encoder-decoder, named MSCA-UNet, for cell counting. Our proposed method outperforms existing methods and can handle various types of cell counting tasks, even in situations with complex cell structures and high background noise. It is suitable for tasks with large cell shape variation, complex structures, and background noise interference. Experiments on three public counting benchmarks demonstrate that MSCA-UNet has good performance and generalization ability.

Despite the high accuracy and reliability of density estimation-based cell counting methods in our research, there are still limitations and challenges in practical applications. For example, counting errors may occur when cells are very sparse or very dense. Additionally, counting irregularly shaped cells such as neurons may also be challenging. Therefore, we suggest that future research should continue to explore and optimize density estimation-based cell counting methods, including developing more accurate density estimation algorithms and combining them with other methods such as morphological analysis to achieve more precise cell counting.

REFERENCES

[1] Z. Wang and Z. Yin, "Cell counting by a location-aware network," in *Proc. 12th Int. Workshop Mach. Learn. Med. Imag. (MLMI)*, Strasbourg, France. Cham, Switzerland: Springer, Sep. 2021, pp. 120–129.

[2] S.-C. Zhang, M. Wernig, I. D. Duncan, O. Brüstle, and J. A. Thomson, "In vitro differentiation of transplantable neural precursors from human embryonic stem cells," *Nature Biotechnol.*, vol. 19, no. 12, pp. 1129–1133, Dec. 2001.

[3] A. Mukherjee, N. A. Repina, D. V. Schaffer, and R. S. Kane, "Optogenetic tools for cell biological applications," *J. Thoracic Disease*, vol. 9, no. 12, pp. 4867–4870, Dec. 2017.

[4] A. Vizcaíno, H. Sánchez-Cruz, H. Sossa, and J. L. Quintanar, "Neuron cell count with deep learning in highly dense hippocampus images," *Expert Syst. Appl.*, vol. 208, Dec. 2022, Art. no. 118090.

[5] V. Bisutti, A. Vanzin, A. Toscano, S. Pegolo, D. Giannuzzi, F. Tagliapietra, S. Schiavon, L. Gallo, E. Trevisi, R. Negrini, and A. Cecchinato, "Impact of somatic cell count combined with differential somatic cell count on milk protein fractions in Holstein cattle," *J. Dairy Sci.*, vol. 105, no. 8, pp. 6447–6459, Aug. 2022.

[6] M. Redetzky, A. Rabenstein, B. Seidel, E. Brinksmeier, and H. Wilhelm, "The influence of cell counts, cell size, EPS and microbial inclusions on the lubrication properties of microorganisms," *Prod. Eng.*, vol. 9, no. 2, pp. 149–159, Apr. 2015.

[7] F. E. McKenzie, W. A. Prudhomme, A. J. Magill, J. R. Forney, B. Permpanich, C. Lucas, R. A. Gasser Jr., and C. Wongsrichanalai, "White blood cell counts and malaria," *J. Infectious Diseases*, vol. 192, no. 2, pp. 323–330, Jul. 2005.

[8] M. A. Bonilla and J. S. Menell, "Disorders of white blood cells," in *Lanzkowsky's Manual of Pediatric Hematology and Oncology*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 209–238.

[9] B. D. Horne, J. L. Anderson, J. M. John, A. Weaver, T. L. Bair, K. R. Jensen, D. G. Renlund, and J. B. Muhlestein, "Which white blood cell subtypes predict increased cardiovascular risk?" *J. Amer. College Cardiol.*, vol. 45, no. 10, pp. 1638–1643, May 2005.

[10] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, 2010, pp. 1–9.

[11] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Comput. Methods Biomechanics Biomed. Eng., Imag. Vis.*, vol. 6, no. 3, pp. 283–292, May 2018.

[12] N. Jiang and F. Yu, "A two-path network for cell counting," *IEEE Access*, vol. 9, pp. 70806–70815, 2021.

[13] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 2685–2688.

[14] J. P. Cohen, G. Boucher, C. A. Glastonbury, H. Z. Lo, and Y. Bengio, "Count-ception: Counting by fully convolutional redundant counting," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 18–26.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[16] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," 2022, *arXiv:2209.08575*.

[17] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.

[18] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[19] C. Sommer, L. Fiaschi, F. A. Hamprecht, and D. W. Gerlich, "Learning-based mitotic cell detection in histopathological images," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 2306–2309.

[20] P. Kainz, M. Urschler, S. Schulter, P. Wohlhart, and V. Lepetit, "You should use regression to detect cells," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 276–283.

[21] S. Chen, M. Zhao, G. Wu, C. Yao, and J. Zhang, "Recent advances in morphological cell image analysis," *Comput. Math. Methods Med.*, vol. 2012, pp. 1–10, Jan. 2012.

[22] B. Pang, L. Zhou, W. Zeng, and X. You, "Cell nuclei detection in histopathological images by using multi-curvature edge cue," in *Proc. 7th Int. Conf. Comput. Intell. Secur.*, Dec. 2011, pp. 1095–1099.

[23] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 642–653, Mar. 2010.

[24] T. Mouroutis, S. J. Roberts, A. A. Bharath, and G. Alusi, "Compact Hough transform and a maximum likelihood approach to cell nuclei detection," in *Proc. 13th Int. Conf. Digit. Signal Process.*, vol. 2, 1997, pp. 869–872.

[25] T. Falk et al., "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019.

[26] S. He, J. Zheng, A. Maehara, G. Mintz, D. Tang, M. Anastasio, and H. Li, "Convolutional neural network based automatic plaque characterization for intracoronary optical coherence tomography images," *Proc. SPIE*, vol. 10574, pp. 800–806, Mar. 2018.

[27] C. Liu, D. Li, and P. Huang, "ISE-YOLO: Improved squeeze-and-excitation attention module based YOLO for blood cells detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 3911–3916.

[28] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Detecting overlapping instances in microscopy images using extremal region trees," *Med. Image Anal.*, vol. 27, pp. 3–16, Jan. 2016.

[29] R. Zhu, D. Sui, H. Qin, and A. Hao, "An extended type cell detection and counting method based on FCN," in *Proc. IEEE 17th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2017, pp. 51–56.

[30] M. Lucidi, M. Marsan, D. Visaggio, P. Visca, and G. Cincotti, "Microscopy direct *Escherichia coli* live/dead cell counting," in *Proc. 20th Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2018, pp. 1–4.

[31] D. Zhang, P. Zhang, and L. Wang, "Cell counting algorithm based on YOLOv3 and image density estimation," in *Proc. IEEE 4th Int. Conf. Signal Image Process. (ICSIP)*, Jul. 2019, pp. 920–924.

[32] T. Xia, R. Jiang, Y. Q. Fu, and N. Jin, "Automated blood cell detection and counting via deep learning for microfluidic point-of-care medical devices," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 646, Oct. 2019, Art. no. 012048.

[33] A. Khan, S. Gould, and M. Salzmann, "Deep convolutional neural networks for human embryonic cell counting," in *Computer Vision—ECCV 2016 Workshops*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 339–348.

[34] Y. Xue, N. Ray, J. Hugh, and G. Bigras, "Cell counting by regression using convolutional neural network," in *Computer Vision—ECCV 2016 Workshops*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 274–290.

[35] S. Aich and I. Stavness, "Improving object counting with heatmap regulation," 2018, *arXiv:1803.05494*.

[36] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: CO-voting uncertain number of targets using random forest for crowd density estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3253–3261.

[37] N. Jiang and F. Yu, "Multi-column network for cell counting," *OSA Continuum*, vol. 3, no. 7, pp. 1834–1846, 2020.

[38] S. He, K. T. Minn, L. Solnica-Krezel, M. A. Anastasio, and H. Li, "Deeply-supervised density regression for automatic cell counting in microscopy images," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101892.

[39] S. Seguí, O. Pujol, and J. Vitrià, "Learning to count with deep object features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 90–96.

[40] X. Ma, H. Wang, C. Qin, K. Li, X. Zhao, J. Fu, and Y. Fu, "A close look at spatial modeling: From attention to convolution," 2022, *arXiv:2212.12552*.

[41] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[43] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1743–1751.

[44] N. R. Gudhe, H. Behravan, M. Sudah, H. Okuma, R. Vanninen, V.-M. Kosma, and A. Mannermaa, "Multi-level dilated residual network for biomedical image segmentation," *Sci. Rep.*, vol. 11, no. 1, p. 14105, Jul. 2021.
[45] M. Marsden, K. McGuinness, S. Little, C. E. Keogh, and N. E. O'Connor, "People, penguins and Petri dishes: Adapting object counting models to new visual domains and object types without forgetting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8070–8079.
[46] N. Saiz, K. M. Williams, V. E. Seshan, and A.-K. Hadjantonakis, "Asynchronous fate decisions by single cells collectively ensure consistent lineage composition in the mouse blastocyst," *Nature Commun.*, vol. 7, no. 1, p. 13463, Nov. 2016.
[47] A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja, "Computational framework for simulating fluorescence microscope images with cell populations," *IEEE Trans. Med. Imag.*, vol. 26, no. 7, pp. 1010–1016, Jul. 2007.
[48] J. Lonsdale et al., "The genotype-tissue expression (GTEx) project," *Nature Genet.*, vol. 45, no. 6, pp. 580–585, 2013.
[49] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 646–661.
[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[53] N. Jiang and F. Yu, "A cell counting framework based on random forest and density map," *Appl. Sci.*, vol. 10, no. 23, p. 8346, Nov. 2020.
[54] Y. Guo, O. Krupa, J. Stein, G. Wu, and A. Krishnamurthy, "SAU-Net: A unified network for cell counting in 2D and 3D microscopy images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 4, pp. 1920–1932, Jul. 2022.

LIKE QIAN is currently pursuing the master's degree with Ningbo University, China. His research interests include cell counting and image processing.

WEI QIAN is currently a Professor with The University of Texas at El Paso. His research interests include biomedical imaging and image processing. He is a fellow of the American Institute for Medical and Biological Engineering (AIMBE).

DINGCHENG TIAN is currently pursuing the master's degree with Ningbo University, China. His research interests include computer vision and medical image processing.

YAQI ZHU is currently pursuing the master's degree with Ningbo University, China. Her research interests include segmentation of cardiac ultrasound images and videos.

HENG ZHAO is currently pursuing the master's degree with Ningbo University, China. His research interests include deep learning and medical image denoising.

YUDONG YAO (Fellow, IEEE) is currently a Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, USA. His research interests include deep learning and medical imaging processing. He is a fellow of the American Institute for Medical and Biological Engineering (AIMBE).