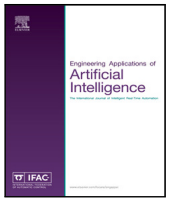
Engineering Applications of Artificial Intelligence 129 (2024) 107634

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Lite-UNet: A lightweight and efficient network for cell localization

Li Bo [a], Zhang Yong [a,*], Ren Yunhan [b], Zhang Chengyang [a], Yin Baocai [a]

[a] Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing Institute of Artificial Intelligence, Beijing University of Technology, Beijing, 100124, China

[b] Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Dataset link: https://github.com/Boli-trainee/Lite-UNet

Keywords:
Cell localization
Difference convolution
Gradient aggregation
Ghost_CBAM
Graph correlation attention

ABSTRACT

Cell localization constitutes a fundamental research domain within the realm of pathology image analysis, with its core objective being the precise identification of cell spatial coordinates. The task has always involved the challenge of large color variations among cells, uneven distribution, and overlapping borders. Furthermore, in realistic cell localization scenarios, the existing state-of-the-art methods suffer from high computational costs and slow inference times, which severely reduce the efficiency of computer-assisted. To tackle the above issues, a lightweight and efficient cell localization model named Lite-UNet is proposed. Specifically, the Lite-UNet encompasses three pivotal modules. Firstly, we introduce a gradient aggregation module grounded in difference convolution. This module effectively mitigates the challenge posed by extensive color variations among cells by adeptly leveraging gradient information. Secondly, we propose an efficient plug-and-play graph correlation attention module, which optimizes the feature representation capabilities by encoding higher-order feature associations. Finally, we design a lightweight Ghost_CBAM module that alleviates the difficulty of uneven cell distribution while forming the base module of the Lite-UNet. Extensive experiments show that our Lite-UNet is capable of locating cells in images quickly and accurately, thus further improving the efficiency of computer-assisted medicine.

1. Introduction

The observation of cells is a crucial resource for human exploration of the microscopic biological world. Cell analysis has long been an essential field and a challenging research topic in medical image analysis. In this field, the purpose of the cell localization task is to precisely locate the specific position of each cell center in the image. This task plays an integral role in a host of medical scenarios, and the localization effect will have a direct influence on the subsequent image analysis (Chen et al., 2022; Asha et al., 2023). Manually locating cells in images is tedious, time-consuming, and expensive, for which reason researchers have been working to develop algorithms that can automatically locate them. However, the large color variations among cells, uneven distribution, and overlapping borders in the cell images make this work extremely challenging.

Traditional cell localization methods (Suryani et al., 2015; Kainz et al., 2015; Li et al., 2021) use image processing algorithms such as thresholding, Canny edge detection, and color recognition filtering to identify cells. However, these methods do not perform well in the case of overlapping cells or large color variations among cells. With the development of artificial intelligence techniques, especially the popularity of deep learning techniques, methods to directly predict the location and number of cells using deep learning techniques have shown encouraging results. Currently, deep learning-based cell localization methods are rapidly emerging as new adjuncts to diagnosis and treatment. Many convolutional neural network (CNN)-based methods (Xie et al., 2018a; Li et al., 2023; Dijkstra et al., 2018) achieve excellent cell localization performance owing to the powerful nonlinear fitting ability of CNN.

The current paradigm commonly used in the field of cell localization is to regress the location map and then obtain the exact location of the cell. Specifically, the cell images are first mapped to the corresponding location maps using a CNN-based approach, and then the location maps are post-processed to obtain information on the location and number of cells. The location maps mentioned here refer to maps that can be used for cell localization, such as probability maps (Sirinukunwattana et al., 2016), density maps (Huang et al., 2020; Tofighi et al., 2019), directional field maps (Chen et al., 2021a), pseudo scale instance map (Zhang et al., 2023), and exponential distance transform maps (Li et al., 2023a). For example, (Xie et al., 2018b) use a fully convolutional neural network to regress the cell density map. Since the network is a fully convolutional design, prediction can be performed on a random size input image, thus enabling end-to-end training for efficiency. Notably, the key to cell localization is the quality of the

* Corresponding author.
E-mail address: zhangyong2010@bjut.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.engappai.2023.107634
Received 9 January 2023; Received in revised form 5 October 2023; Accepted 27 November 2023
Available online 6 December 2023

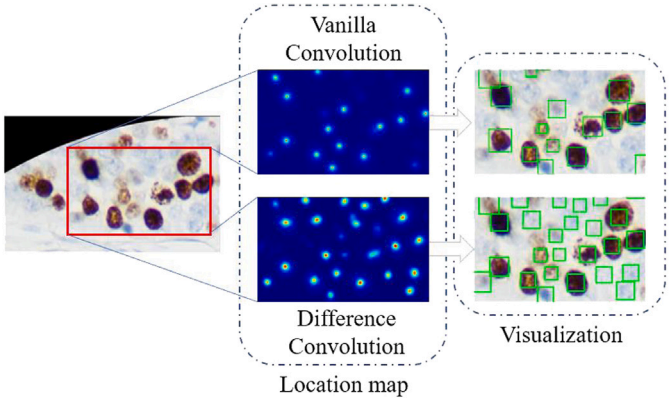


Fig. 1. The main difficulty in cell localization: the large color variations among cells. Left: original cell image with large color variations among cells; middle: location map generated from the original cell image based on different convolution models; right: visualization results of the location map.

generated location map, and a location map that accurately reflects the cell location can significantly improve the localization results.

According to our observations, location maps often have trouble accurately reflecting cell location information. The core reason is that the color of the cells varies too much, resulting in too large a gap in the corresponding responses of the cells in the location maps, and eventually the light-colored cells are often ignored. As shown in Fig. 1, when vanilla convolution-based models predict cell images with large color variations, the light-colored cells in the images respond extremely weakly in the generated location maps, resulting in existing localization models tending to ignore the light-colored cells, which greatly reduces the localization performance of the models. It is worth mentioning that Huang et al. (2020) annotate and predict lighter-colored and dark-colored cells separately, cleverly avoiding the difficulty of large color variations. However, this strategy introduces more cumbersome operations and more expensive manual annotation costs, which is not conducive to further expansion of the application. Ideally, all colors of cells should correspond to the same degree of response in the location map. Cells with large color variations correspond to large differences in pixel values, yet ultimately all cells need to be mapped to the same response in the location map, thus creating a contradiction. In this regard, we expect the model to reduce the color gap between cells, i.e., to reduce the gap between the absolute values of cell pixels. Considering that difference convolution has a similar effect, which mitigates the large gap in cell color to some extent by performing local difference computation on the cell pixels, thus capturing the gradient information between neighboring pixels. To this end, we propose difference convolution-based gradient aggregation module to alleviate this problem by enhancing the gradient information among pixels.

The current field of cell localization also suffers from uneven cell distribution and high model computational cost. For the tricky problem of uneven cell distribution, Guo et al. (2021) use the self-attention module to capture the interconnections among features. However, the process of calculating the association weights of global features of an image by the self-attention module is expensive, and they use the more costly U-Net (Ronneberger et al., 2015) as the backbone, which further increases the computational cost of the model. The high number of parameters and computational effort of the model limits its application, especially in the light of the increasing popularity of mobile applications and edge computing deployments. In addition, too slow inference speed reduces the enjoyment and productivity of healthcare professionals. To solve these issues, a new Ghost_CBAM module is proposed that combines the CBAM attention module, which has an exceedingly small computational cost, with the Ghost module. This module effectively implements model compression while alleviating the problem of uneven cell distribution. In this paper, a lightweight, high-performance cell localization model Lite-UNet is built based on this module.

Given the powerful modeling capability of CNN, the above CNN-based models have been capable to achieve strong cell localization performance. However, the CNN suffers from two problems: first, it uniformly transform local features in images by convolutional operations, ignoring the correlation among features. Secondly, the receptive fields in CNN are restricted to local areas, which are difficult to adapt to the complex topology of the scene. Therefore, some researchers (Defferrard et al., 2016; Atwood and Towsley, 2016; Welling and Kipf, 2016; Li et al., 2023b) propose graph convolutional network (GCN) to model the complex associations in the scenes. Compared with traditional CNN methods, GCN can encode the graph structure of the input data and continuously learn and aggregate relevant information from a full-graph perspective to better characterize the features. For example, the uneven distribution of cells and the variety of cell shapes are incompatible with the receptive fields of the CNN, yet the topology of graph convolutional networks can better fit this non-uniform distribution. Accordingly, this paper proposes a graph correlation attention module that can improve the performance of cell localization by encoding higher-order association information among features to guide the distribution of features.

In summary, the contributions of this paper are summarized as follows.

- In this paper, we propose a lightweight and efficient cell localization model that achieves competitive performance at a very low cost and can improve the efficiency of computerized medical assistance.
- A novel gradient aggregation module based on difference convolution is proposed to effectively alleviate the problem of large color variations among cells.
- An attention-based Ghost_CBAM module is designed to effectively alleviate the problem of uneven cell distribution while achieving model compression, and a lightweight high-performance cell localization model Lite-UNet is built based on this module.
- A graph correlation attention module is proposed, which encodes higher-order associations among features for better representation.

2. Related works

In this section, we briefly describe the current state of research on CNN-based cell localization and counting tasks, which can be broadly classified into detection-based localization methods and map-based regression methods.

2.1. Detection-based localization methods

This method aims to determine the specific location of each cell in the image as well as the overall number by the paradigm of detection. For fast and efficient detection of blood cells in microscopic images, Shakarami et al. (2021) propose a detector based on YOLOV3 (Redmon and Farhadi, 2018). They expand the sensory domain by stacking the null convolution and make the model lighter by using depth-separable convolution and finally achieve better detection accuracy on the BCCD dataset. Alam and Islam (2019) also design a YOLO-based cell detector for automatic identification and counting of red blood cells, blood cells, and platelets. In addition, Kutlu et al. (2020) propose a deep learning and migration learning-based approach for automatic leukocyte detection from smear images.

The detection-based cell localization methods have good detection performance in the case of sparse cells, but as the cell density increases, the problem of intercellular occlusion becomes more severe, leading to a rapid decline in the performance of these detection models. As a result, it is now common for researchers to use a map-based regression paradigm to calculate and localize cells.

2.2. Map-based regression methods

To solve the problem of mutual occlusion caused by dense cell density, some researchers have proposed regression localization and counting methods based on probability maps, density maps, directional field maps, or exponential distance transform maps. The researchers (Tofighi et al., 2019; Huang et al., 2020; Li et al., 2021) use CNN models to predict the probability/density maps corresponding to the images, and then consider the local maxima in them as the centroids of the cells. Tofighi et al. (2019) propose a spatially constrained convolutional neural network for cell center detection. They predict the probability of a patch in each image becoming the center of the cell nucleus, and then aggregated the above results to form a probability map to locate the specific location of the cell. Huang et al. (2020) regress the cell density map by adding a transposed convolutional layer on top of CSRNet (Li et al., 2018). Falk et al. (2019) implement U-Net-based cell localization, and provide an ImageJ plugin that enables non-machine learners to analyze their data. Guo et al. (2021) establish a unified 2D and 3D cell counting framework based on the U-Net segmentation network and self-attention mechanism. Similarly, Mao et al. (2021) decompose the cell localization task into two subtasks to be processed: first mapping the raw pathology image to a binary mask, and then mapping the binary mask to a density mask in the second subtask, resulting in improved performance. However, the above-mentioned density map characterization methods are difficult to accurately express the cell localization information, so researchers further propose the directed field map (Chen et al., 2021a). The directional field map defines a directional field at each pixel of a cell region such that adjacent pixels are oriented in opposite directions, but the directional fields of the same cell region all point to the same center. However, the direction field map is susceptible to mutual overlap between cells. Therefore, Li et al. (2023a) propose an exponential distance strategy to optimize the cell localization map so that the cells are independent of each other, which is a promising solution to the problem of cell overlap. In addition to the above work, in order to further extend the cell localization task while adding more comparison methods, we also introduce a number of popular models in the field of image analysis, including Attention U-Net (Oktay et al., 2018), HRNet (Wang et al., 2020), TransUNet (Chen et al., 2021b), MPViT (Lee et al., 2022), Swin Transformer (Liu et al., 2021), and ConvNeXt (Liu et al., 2022). Among them, Attention U-Net, HRNet, and TransUNet can be directly used for localization tasks, while MPViT, Swin Transformer, and ConvNeXt need to follow HRNet's stacking approach to be used for localization tasks, i.e., stacking feature maps at multiple stages.

The above works have largely contributed to the development and application of the field of cell localization and counting. However, the above work does not take into account the computational cost of the models, and they tend to use computationally expensive models for localization and counting, and thus have high parameter counts, large computational effort, and very long inference times, each of which limits their application in real-world scenarios.

3. The model

The lightweight and efficient Lite-UNet model for cell localization proposed in this paper is shown in Fig. 2, which illustrates the dimensional changes of the feature maps at each stage. The input to the model is the cell image and the output is the corresponding cell location map. The model consists of 3 main modules: first, a Gradient Aggregation (GA) module based on difference convolution, which serves as a front-end to the encoding part to fully extract and aggregate the gradient and semantic information among features; second, to alleviate the problem of uneven cell distribution while obtaining a lightweight and high-performance model, we replace the convolution module in the network with the Ghost_CBAM module; finally, the Graph Correlation Attention (GCA) module that encodes the higher-order association relations among features is placed at the semantic information-rich end of the encoding. We will explain them separately in the following.

3.1. Gradient aggregation module

As shown in Fig. 1, the main challenge faced in the field of cell analysis is the large color variations among cells, which causes models to often ignore light-colored cells in the images. Specifically, the local values corresponding to cells of different colors are similar in the location map, so it is difficult for the model to learn a uniform mapping representations. From the localization results, the model tends to ignore the lighter-colored cells, which leads to the degradation of localization performance. In this regard, we propose a gradient aggregation module based on difference convolution, which captures the gradient information between neighboring pixels by locally computing the difference between cell pixels, while reducing the difference in pixel values between cells of different colors. Accordingly, this paper proposes a GA module based on difference convolution, which can improve the utilization of gradient information by the model and alleviate the problem of the large color variations among cells.

The vanilla convolution commonly used by researchers can be expressed as

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

Where p_0 denotes the central position of the local receptive field R, p_n denotes the relative position of each value in the R to p_0, and $w(p_n)$ is the learnable parameter. Correspondingly, the difference convolution (Yu et al., 2020) can be expressed as

$$y(p_0) = \sum_{p_n \in R} w(p_n) \times (x(p_0 + p_n) - x(p_0)), \tag{2}$$

That is, each value of the local receptive field is subtracted from the value of its centroid (hence also known as central difference convolution), which is used to form the local gradient information. Meanwhile, considering that vanilla convolution can bring stronger semantic information, the final difference convolution is defined as

$$y(p_0) = \sum_{p_n \in R} w(p_n) \times x(p_0 + p_n) + \theta(-x(p_0) \times \sum_{p_n \in R} w(p_n)), \tag{3}$$

where θ is an artificially designed trade-off parameter, which will be followed by a more in-depth discussion and ablative experimental studies in the experimental session.

Considering that difference convolution weakens the semantic information while enhancing the gradient information, we only use it as the front-end of feature extraction. Given the varying scales of cells, we expect the GA module to extract gradient information at different scales. To this end, inspired by Szegedy et al. (2015, 2016), the GA module based on difference convolution designed in this paper is shown in Fig. 3. It has four branches, each of which contains a difference convolution block to enhance the gradient information of the features. Different branches have different depths so that features at different levels can be extracted. The refined feature maps are obtained by overlaying different branches and adjusting the number of channels using the 1×1 block.

3.2. Attention-based Ghost_CBAM module

The cells in the image are often unevenly distributed, which contradicts the strategy of most methods to treat all localities in the image equally. Specifically, existing models do not give more attention to cell-dense regions and are computationally expensive. Therefore, we cleverly combine CBAM (Woo et al., 2018) and Ghost (Han et al., 2020) modules to get the Ghost_CBAM module, which has the advantages of being lightweight and efficient.

As shown in Fig. 4, the Ghost module enhances the feature map F_1 obtained by primary convolution using a cheap convolution operation, which aims to exploit an extremely small computational cost in

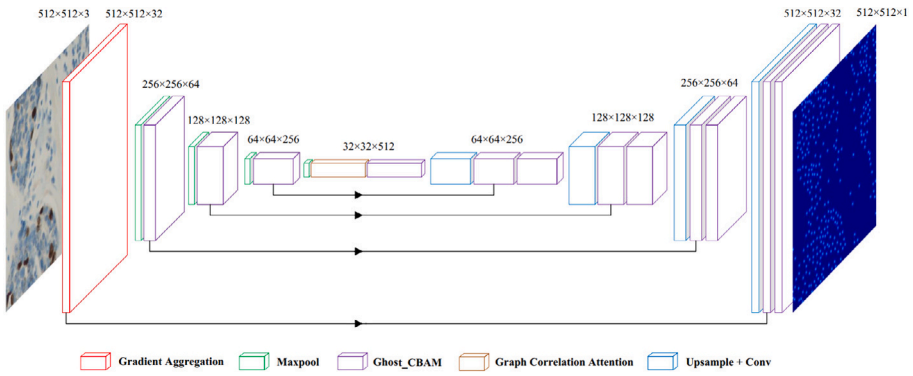


Fig. 2. The overall framework of Lite-UNet. To visualize the variability of the model, we represent the module together with a feature map, e.g., the gradient aggregation module represents the feature map output by the module. The model consists of 3 main components: a gradient aggregation module based on difference convolution used to aggregate gradient information in the encoding phase; the Ghost_CBAM module for compressing the model and alleviating the problem of uneven cell distribution; a graph correlation attention module that improves cell localization performance by encoding higher-order associations among features.

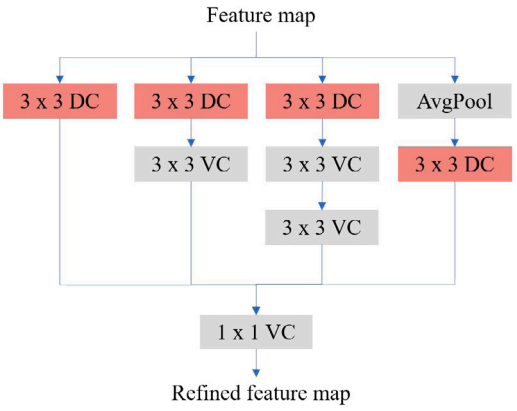


Fig. 3. Gradient aggregation module based on difference convolution. DC denotes difference convolution, VC denotes traditional vanilla convolution, and 3 × 3 denotes the size of the convolution kernel.

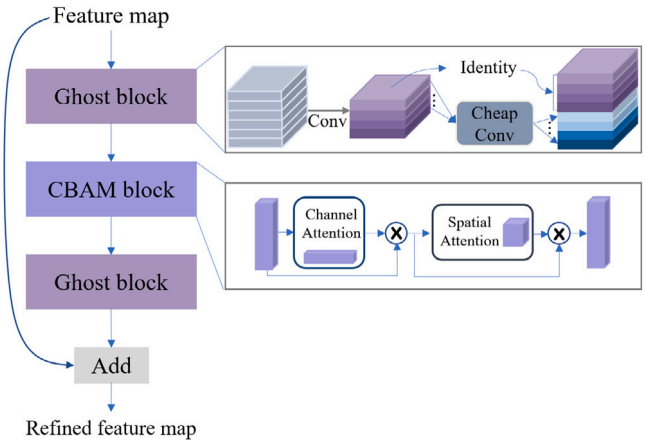


Fig. 4. Illustration of the attention-based Ghost_CBAM module. The module consists of two main submodules: the Ghost module enhances the performance of the model by augmenting the feature maps obtained by convolution using a cheap linear operation; The CBAM module learns attention weights sequentially along the channel and spatial dimensions and then multiplies the weights with the input feature maps for adaptive feature optimization.

exchange for better performance. The primary convolution process is as follows:

$$F_1 = F_0 * f, \tag{4}$$

where $f \in \mathbb{R}^{c\times k\times k\times m}$ is the utilized filters, F_0 is the input data, and F_1 is the intrinsic feature maps generated by primary convolution. Next, we perform a series of cheap linear operations on each intrinsic feature in F_1 to generate a ghost feature map:

$$y_{ij} = \Phi_{i,j}(y'_i), \quad \forall i = 1, \ldots, m, j = 1, \ldots, s, \tag{5}$$

where y'_i is the intrinsic feature map in F_1 and $\Phi_{i,j}$ is the jth liner operation for generating the jth ghost feature map y_{ij}. With the above operations, we can obtain $n = m \times s$ feature maps $F_2 = [y_{11}, y_{12}, \ldots, y_{ms}]$ as the refined data.

Then, the CBAM module learns attention weights sequentially along the channel M_c and spatial M_s dimensions and then multiplies the weights with the input feature map F_2 for adaptive feature optimization, which is calculated as follows

$$\begin{aligned} F_3 &= M_c(F) \otimes F_2, \\ F_4 &= M_s(F_3) \otimes F_3, \end{aligned} \tag{6}$$

where $\otimes$ denotes element-wise multiplication, F_2 is the input feature maps, and F_4 is the output feature maps after adaptive optimization. As shown in Fig. 2, we replace most of the convolution modules in U-Net with the Ghost_CBAM module, which greatly reduces the number of parameters and the computation cost of the model. Specifically, as shown in Fig. 2, we replace all modules in U-Net except the upsampled

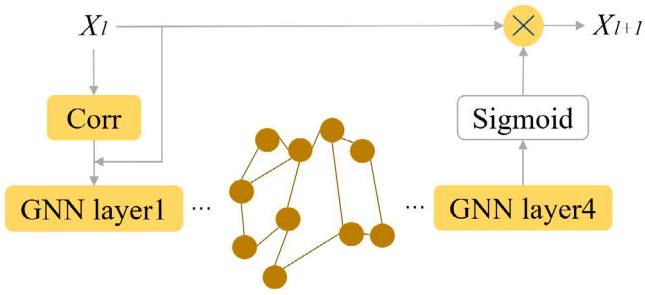


Fig. 5. The overall processing flow of the GCA module.

convolution with the Ghost_CBAM module, which achieves a significant reduction in computational cost while keeping the performance of the model basically unchanged.

3.3. Graph correlation attention module

The common paradigm of existing relevance-based graph information embedding methods (Wu et al., 2020; Wang and Gupta, 2018) is to encode global features in an aggregated manner and then update

node features. However, there are two major problems with the above methods in the process of updating: (1) Features with weak associations are selectively filtered, making it difficult for some local features to be updated; (2) The optimized features are directly used for subsequent operations, which may have the problem of feature over-smoothing. Therefore, in this paper, a new GCA module is proposed to capture the association relationships among features to better guide the distribution of features.

We use a generic GCN ($G = (V, E)$) to model the associations among features and iteratively optimize them, where V is the set of all nodes and E is the set of all edges. Specifically, given a feature map as $F \in \mathbb{R}^{C\times H\times W}$, in order to generate a graph, the grid of $C\times1\times1$ is treated as a node V_i. For each $(V_i, V_j) \in E$, the correlation score between node V_i and V_j is denoted by $Corr_{ij} = f(F_i, F_j)$, where F_i and F_j both belong to $\mathbb{R}^c$ and are the feature vectors of two nodes. As with most methods, we use the inner product between features to measure their similarity. To better adaptively characterize the correlation between features, we add a linear transformation operation before the feature inner product operation, denoted as

$$f(F_i, F_j) = (W_i F_i) \times (W_i F_j)^T. \tag{7}$$

For a more comprehensive update of local features, we do not filter for correlations. Using the obtained $Corr_{ij}$ to update the features X_l, the optimized features X_{l+1} are obtained:

$$X_{l+} = Dropout(ReLU(Corr_{ij} X_l W_l)). \tag{8}$$

Considering that the optimized features X_{l+} are too smooth, we adopt them as attention weights to optimize the features:

$$X_{l+1} = Sigmoid(X_{l+}) \times X_l. \tag{9}$$

Finally, the features optimized by the GCA module are obtained, and the whole operation is shown in Fig. 5.

4. Experiments and analysis

4.1. Datasets and experimental details

In the field of cell localization, the popular publicly available datasets include BCData (Huang et al., 2020), Seg_data, and PSU (Tofighi et al., 2019). The BCData dataset has the largest sample size and is the main dataset used in this paper. The following is a brief description of these datasets.

BCData (Huang et al., 2020) is a large-scale breast tumor cell dataset for Ki-67 cell localization and enumeration. It covers 1,338 images with a resolution size of 640 × 640 and 181,074 annotated tumor cells. The dataset is exceedingly close to a real cell localization scenario and has the following characteristics: (1) Diversity of tumor cell distribution density; (2) Different positive rates of tumor cells in the images; (3) The tumor cells vary in size and shape, and the cell borders are indistinct. The dataset is divided into a training set, a validation set, and a test set with a number ratio of 803:133:402.

Seg_Data (Gao et al., 2021) dataset consists of 1000 H&E stained image patches with a resolution of 512 × 512, containing a total of 70,945 labeled cell nuclei, each with an instance segmentation mask and a classification mask. To use this dataset for cell localization, we derived the centroids of the cells by means of a connected domain algorithm. The resulting dataset (which we call Seg_Data) can be used for cell localization tasks.

PSU (Tofighi et al., 2019) dataset contains 120 images of colonic tissue from 12 pigs with a resolution of 612 × 452. It covers 25,462 annotated cells, and the overall tone of the image is dark. We adopt 84 images as the training set and 36 images as the validation set.

Experimental details: During both the training and testing phases, we maintain a uniform image resolution of 512 × 512. This choice is informed by the relatively consistent image sizes within the aforementioned datasets. Regarding the design details of the model, as illustrated in Fig. 2, the overall structure of Lite-UNet is similar to that of U-Net, including four downsampling and upsampling stages. However, the key difference lies in the extensive optimization applied to the modules within Lite-UNet. Notably, there are significant variations in the sizes of convolution kernels: (1) In the GA module, all convolution kernels have a size of 3; (2) The convolution kernel sizes in the Ghost_CBAM module are more intricate, the Ghost sub-module uses combinations of 3 and 1, while the CBAM sub-module employs combinations of 7 and 1; (3) The Graph correlation attention module does not involve convolution operations. (4) For the upsampling paired with convolution module, the convolution kernel size is set to 3. For the training process, the settings are as follows: we employ the mean squared error loss as the optimization objective, use the Adam optimizer (Kingma and Ba, 2014) to update network parameters, set the model's learning rate to 1e-4, utilize a batch size of 12, and apply data augmentation through horizontal flipping. All experiments were conducted on Ubuntu 18.04, using an NVIDIA Tesla P100 GPU (~16 GB). Lastly, we plan to make both the experimental code and processed datasets publicly accessible on GitHub at the following repository: Lite-UNet.

4.2. Evaluation criteria

The task of this paper is cell localization while counting the number of cells, so the evaluation criteria contain localization criteria and counting criteria.

Localization criteria: To accurately evaluate the matching relationship between prediction points and cell truth points, we use Precision, Recall, and F1 score to evaluate the localization performance.

$$Precision = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Positive}, \tag{10}$$

$$Recall = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative}, \tag{11}$$

$$F1 = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \quad \beta = 1 \tag{12}$$

, where True Positive indicates a successful match when the distance between a given predicted point and the true value point is less than a threshold σ. The design of σ is closely related to the accuracy required for cell localization in the task, and a fixed threshold of two levels ($\sigma = 5$, $\sigma = 10$) is chosen in this paper to evaluate the performance of the model. The smaller the threshold value is, the tighter the localization accuracy is.

Counting criteria: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to evaluate the counting performance.

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |y_i - \hat{y}_i|, \tag{13}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}, \tag{14}$$

where m is the number of cell images, y_i is the number of ground truth in ith image, and $\hat{y}_i$ is the predicted number derived from this image by cell localization.

In addition, to objectively analyze the computational cost of the model, we use three commonly used metrics: the number of parameters (Params in M), computational complexity Giga Floating-point Operations Per second (GFLOPs), and inference time (in ms). The number of model parameters is the total number of parameters to be trained in the network model, reflecting the spatial complexity of the model. The computational complexity refers to the number of floating-point operations, i.e., the number of floating-point operations per second. Furthermore, we test the inference time of the model on the GPU, considering that the number of parameters and the amount of computation are difficult to accurately reflect the inference speed of the model in real application scenarios.

Table 1
Quantitative comparison of localization and counting performance of different models on the BCData test dataset. For more visualization, the extreme values in the different groups are bolded separately, as well as in the table below.

Methods	Counting	Localization (5)	Localization (10)
	MAE/RMSE↓	F1/Pre/Rec (%)↑	F1/Pre/Rec (%)↑
U_CSRNet	18.1/23.8	73.8/73.7/74.0	85.6/85.4/85.7
MPViT	22.3/29.2	75.3/79.3/71.6	85.5/90.1/81.4
Attention U-Net	19.6/24.9	77.1/74.9/**79.5**	86.5/84.0/**89.1**
TransUNet	17.7/23.3	77.3/76.5/78.1	86.9/86.1/87.8
Swin Transformer	**17.1**/**23.1**	78.1/77.8/78.4	87.1/86.7/87.4
HRNet	18.5/24.7	79.2/80.1/78.3	87.3/88.4/86.3
ConvNeXt	18.9/26.0	**79.3**/**82.1**/76.7	**87.4**/**90.5**/84.6
U-Net (Baseline)	24.9/33.4	**76.7**/**81.7**/72.1	85.7/80.7/**91.4**
Lite-UNet (Our)	**18.1**/**24.3**	76.5/77.0/**76.1**	**86.3**/**86.3**/86.4

Table 2
Comparison of computational cost of different models.

Methods	Computational cost		
	Params↓	GFLOPs↓	Speed↓
U_CSRNet	**16.30**	**109.48**	**150**
MPViT	57.98	189.51	199
Attention U-Net	34.88	266.27	168
TransUNet	93.19	128.68	192
Swin Transformer	203.20	280.08	268
HRNet	66.58	142.12	195
ConvNeXt	89.72	225.53	172
U-Net (Baseline)	34.53	261.87	118
Lite-UNet (Our)	**1.30**	**17.50**	**92**

Table 3
Quantitative comparison of localization and counting performance of different models on the Seg_Data test dataset.

Methods	Counting	Localization (5)	Localization (10)
	MAE/RMSE↓	F1/Pre/Rec (%) ↑	F1/Pre/Rec (%) ↑
U_CSRNet	7.5/8.8	81.8/78.5/85.5	88.2/84.6/**92.1**
MPViT	6.2/7.8	82.0/82.5/81.6	88.5/88.4/88.7
Attention U-Net	**4.6**/**6.0**	83.4/83.6/83.2	89.7/89.8/89.5
TransUNet	5.0/7.0	83.6/84.0/83.3	90.0/90.3/89.6
Swin Transformer	6.0/7.9	82.4/80.2/84.6	89.4/87.0/91.9
HRNet	4.9/6.9	**85.6**/**87.3**/83.9	**90.4**/**92.3**/88.7
ConvNeXt	5.8/7.1	83.7/82.9/84.4	89.4/88.5/90.4
W-Net	– / –	85.0/83.0/**88.0**	– / – / –
U-Net (Baseline)	5.8/7.9	83.7/**86.7**/80.8	89.4/**92.6**/86.4
Lite-UNet (Our)	**5.2**/**7.4**	**84.4**/85.8/**83.1**	**89.6**/91.3/**86.9**

Table 4
Quantitative comparison of localization and counting performance of different models on the PSU dataset.

Methods	Counting		Localization	
	MAE↓	RMSE↓	F1↑ (σ=5)	F1↑ (σ=10)
U_CSRNet	32.1	38.2	54.5	80.4
MPViT	36.6	44.7	61.1	81.0
Attention U-Net	31.4	36.8	63.5	82.1
TransUNet	40.1	49.4	58.9	80.1
Swin Transformer	**26.6**	**32.1**	62.2	82.1
HRNet	27.6	32.4	**66.1**	**83.5**
ConvNeXt	27.4	33.5	63.5	82.8
U-Net (Baseline)	**32.6**	38.5	**61.0**	**81.0**
Lite-UNet (Our)	33.7	**38.4**	60.2	79.6

4.3. Comparative experiments and analysis

Given that this paper is devoted to design a lightweight and efficient cell localization model, we not only test the localization and counting performance of the model, but also calculate the computational cost, including the number of parameters(in M), computational complexity(in GFLOPs), and inference speed(in ms). First, we compare the localization and counting performance of multiple models on BCData (the largest cell localization task dataset), as shown in Table 1. It is worth mentioning that our comparison models include U_CSRNet (Li et al., 2018), MPViT (Lee et al., 2022), Attention U-Net (Oktay et al., 2018), TransUNet (Chen et al., 2021b), Swin Transformer (Liu et al., 2021), HRNet (Wang et al., 2020), W-Net (Mao et al., 2021), and ConvNeXt (Liu et al., 2022). Referring to the operation of concatenating the output feature maps of multiple phases in HRNet, we also performed the same operation on the above models to ensure that the inputs and outputs are of the same size. As can be seen, the Lite-UNet proposed in this paper shows some improvement over our baseline model U-Net, and is also competitive with many recently popular models. Second, we show the comparison of the computational cost of different models in Table 2. Combined with Table 1, we can conclude that the Lite-UNet model in this paper achieves a very competitive performance while significantly reducing the computational cost. Specifically, compared with U-Net, Lite-UNet achieves similar performance with only 3.7% of the original number of parameters, 6.7% of the computation, and 75% of the inference time.

In addition, in order to verify the performance of the models more comprehensively, we also compare them on the datasets Seg_data (Gao et al., 2021) and PSU (Tofighi et al., 2019), as shown in Tables 3 and 4. It can be seen that Lite-UNet outperforms the baseline model U-Net in Table 3 and is not far from other popular models. However, according to Table 4, it can be seen that the model proposed in this paper performs poorly in the PSU dataset. According to our observation, the main reason is that the overall image in this dataset is very dark and the cell colors are more uniform, which is not compatible with the gradient aggregation module proposed in this paper. Specifically, in the BCData and Seg_data datasets, the image background is white and the cell color disparity is large, and the proposed GA module is able to enhance the robustness of the model for cell color recognition. However, in the PSU dataset, the image background is black and the cell color is uniformly very light blue, which makes it difficult for the GA module to work. For more visualization of the effect of cell localization in our model, we perform tests on some typical cell images, as shown in Fig. 6.

4.4. Ablation experiments

In order to objectively assess the contribution of each module, we conduct experiments on each module, as shown in Table 5. We use U-Net as the baseline and keep iterating the model to finally obtain Lite-UNet. First, most of the convolutional modules in U-Net are replaced by Ghost_CBAM modules, and the number of channels is reduced, resulting in a significant reduction in the computational cost of the model, where the number of parameters and FLOPs of the model is reduced to 3.6% and 6.7% of the original model, respectively. However, benefiting from the Ghost_CBAM module's mitigation of the uneven cell distribution and the clever use of feature maps, the model's localization and counting performance does not deteriorate significantly but rather improves at a threshold of 10. Next, the GA module based on difference convolution is used to replace the first coding module in the original model to resolve the large color differences in the cell images. Experimental results show that the module is able to significantly improve the localization and counting performance of the model with almost no increase in computational cost, with a significant 4% improvement in localization performance. Finally, the GCA module optimizes the high-level features in the U-Net structure to better guide the distribution of features by capturing their correlation relationships. The addition of this module significantly improves the localization performance of the model by 8% for a threshold value of 5.

As shown in Eq. (3), the main advantage of difference convolution is to enhance the gradient information, where the trade-off parameter θ determines the ratio of gradient information to semantic information. It

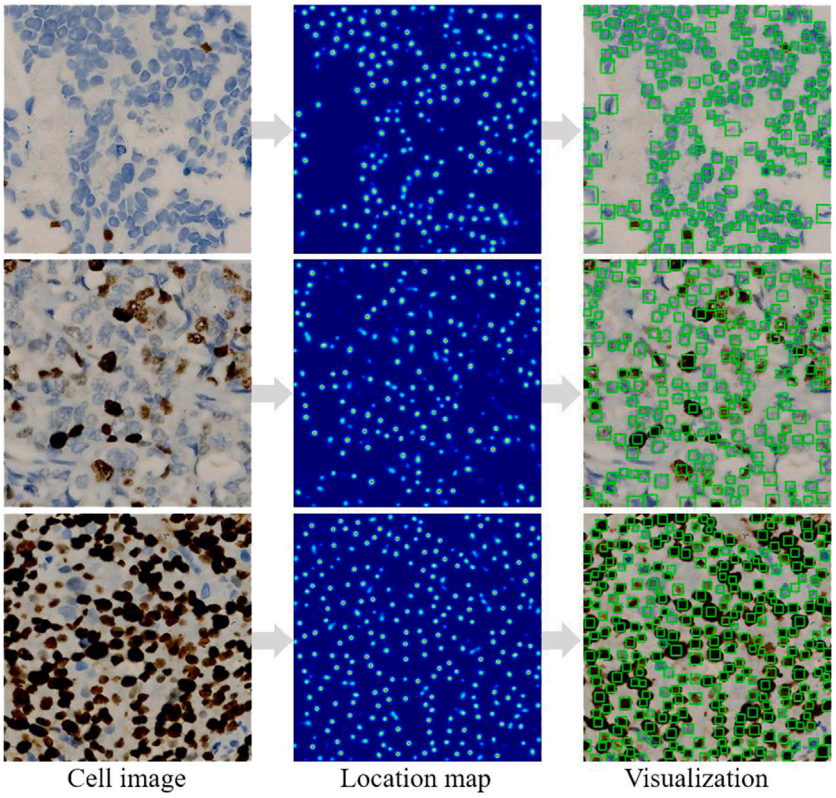
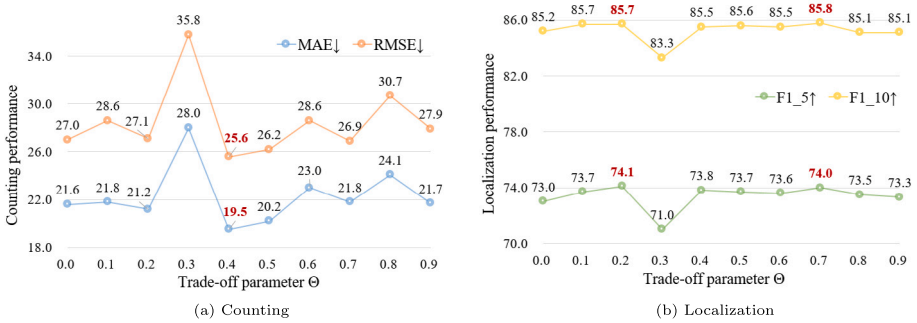


Fig. 6. Visualization results of the cell localization effect of our Lite-UNet.



(a) Counting (b) Localization

Fig. 7. Ablation experiments for the trade-off parameter θ of the difference convolution. Here, Figures A and B represent dissection experiments for counting and localization performance, respectively. When the $\theta = 0$, the difference convolution degenerates to the traditional vanilla convolution.

Table 5
Ablation experiments on the BCData test dataset. The computational cost, counting, and localization performance are shown, respectively.

Methods	Computational cost			Counting		Localization	
	Params↓	GFLOPs↓	Speed ↓	MAE↓	RMSE↓	F1 (σ=5)↑	F1 (σ=10)↑
U-Net	34.53	261.87	118	24.9	33.4	76.7	85.7
+Ghost_CBAM	**1.25**	**16.80**	102	19.5	25.8	75.2	85.8
+GA	**1.25**	**16.80**	**89**	18.3	25.1	75.7	86.2
+GCA (Our)	1.30	17.50	92	**18.1**	**24.3**	**76.5**	**86.3**

is worth noting that when $\theta = 0$, the proportion of gradient information drops to 0, and the gradient aggregation module degenerates to the traditional feature aggregation module. As can be seen from Fig. 7, the trade-off parameter θ has a significant influence on both the localization and counting performances of the model. The best counting performance is achieved when $\theta = 0.4$. The localization performance is the best when the $\theta = 0.2$ or 0.7.

5. Conclusion

In this paper, a lightweight and efficient cell localization model based on U-Net is proposed. Our Lite-UNet consists of three main components: (1) the gradient aggregation module, which can effectively utilize the multi-scale gradient information of features to enhance the robustness of the model to cell color changes; (2) the Ghost_CBAM

module, which can significantly compress the computational cost of the model without losing a large amount of accuracy; (3) the graph correlation attention module, which can improve the localization performance by learning the higher-order correlations among features to optimize features. Comprehensive experiments show that our Lite-UNet is capable of quickly and accurately localizing cells in images with competitive performance. Compared to existing models, the Lite-UNet can be deployed to more medical scenarios where computational power resources are not abundant, further increasing the available scope of computer medical assistance. In the future, we will design an end-to-end, lightweight and high-performance cell localization model based on graph neural networks.

CRediT authorship contribution statement

Li Bo: Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Zhang Yong:** Project administration, Supervision, Writing – review & editing. **Ren Yunhan:** Investigation, Methodology, Conceptualization. **Zhang Chengyang:** Investigation, Validation, Methodology. **Yin Baocai:** Resources, Funding acquisition, Supervision.

Declaration of competing interest

The authors declared that they have no conflicts of interest to this work.

Data availability

The processed dataset and code are posted on https://github.com/Boli-trainee/Lite-UNet.

Acknowledgments

The research project is supported by the National Key R&D Program of China (No. 2021ZD0111902), NSFC (No. 62072015, U21B2038, U19B2039), R&D Program of Beijing Municipal Education Commission (No. KZ202210005008).

References

Alam, M.M., Islam, M.T., 2019. Machine learning approach of automatic identification and counting of blood cells. Healthc. Technol. Lett. 6 (4), 103–108.

Asha, S., Gopakumar, G., Subrahmanyam, G.R.S., 2023. Saliency and ballness driven deep learning framework for cell segmentation in bright field microscopic images. Eng. Appl. Artif. Intell. 118, 105704.

Atwood, J., Towsley, D., 2016. Diffusion-convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, Vol. 29.

Chen, J., Chen, W., Zeb, A., Zhang, D., 2022. Segmentation of medical images using an attention embedded lightweight network. Eng. Appl. Artif. Intell. 116, 105416.

Chen, Y., Liang, D., Bai, X., Xu, Y., Yang, X., 2021a. Cell localization and counting using direction field map. IEEE J. Biomed. Health Inf. 26 (1), 359–368.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021b. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proceedings of the Advances in Neural Information Processing Systems, Vol. 29.

Dijkstra, K., Loosdrecht, J., Schomaker, L.R., Wiering, M.A., 2018. Centroidnet: A deep neural network for joint object localization and counting. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 585–601.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al., 2019. U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods 16 (1), 67–70.

Gao, Z., Shi, J., Zhang, X., Li, Y., Zhang, H., Wu, J., Wang, C., Meng, D., Li, C., 2021. Nuclei grading of clear cell renal cell carcinoma in histopathological image by composite high-resolution network. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 132–142.

Guo, Y., Krupa, O., Stein, J., Wu, G., Krishnamurthy, A., 2021. SAU-net: A unified network for cell counting in 2d and 3d microscopy images. IEEE/ACM Trans. Comput. Biol. Bioinform.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C., 2020. Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1580–1589.

Huang, Z., Ding, Y., Song, G., Wang, L., Geng, R., He, H., Du, S., Liu, X., Tian, Y., Liang, Y., et al., 2020. Bcdata: A large-scale dataset and benchmark for cell detection and counting. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 289–298.

Kainz, P., Urschler, M., Schulter, S., Wohlhart, P., Lepetit, V., 2015. You should use regression to detect cells. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 276–283.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kutlu, H., Avci, E., Özyurt, F., 2020. White blood cells detection and classification based on regional convolutional neural networks. Med. Hypotheses 135, 109472.

Lee, Y., Kim, J., Willette, J., Hwang, S.J., 2022. Mpvit: Multi-path vision transformer for dense prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7287–7296.

Li, B., Chen, J., Yi, H., Feng, M., Yang, Y., Bu, H., 2023a. Exponential distance transform maps for cell localization. TechRxiv.

Li, B., Huang, H., Zhang, A., Liu, P., Liu, C., 2021. Approaches on crowd counting and density estimation: a review. Pattern Anal. Appl. 24, 853–874.

Li, Y., Zhang, X., Chen, D., 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1091–1100.

Li, B., Zhang, Y., Zhang, C., Piao, X., Yin, B., 2023. Multi-scale Hypergraph-based Feature Alignment Network for Cell Localization. http://dx.doi.org/10.36227/techrxiv.22680877.v1.

Li, B., Zhang, Y., Zhang, C., Piao, X., Yin, B., 2023b. Hypergraph association weakly supervised crowd counting. ACM Trans. Multimed. Comput. Commun. Appl.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Mao, A., Wu, J., Bao, X., Gao, Z., Gong, T., Li, C., 2021. W-Net: A two-stage convolutional network for nucleus detection in histopathology image. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. pp. 2051–2058.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Shakarami, A., Menhaj, M.B., Mahdavi-Hormat, A., Tarrah, H., 2021. A fast and yet efficient YOLOv3 for blood cell detection. Biomed. Signal Process. Control 66, 102495.

Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1196–1206.

Suryani, E., Wiharto, W., Polvonov, N., 2015. Identification and counting white blood cells and red blood cells using image processing case study of leukemia. arXiv preprint arXiv:1511.04934.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Tofighi, M., Guo, T., Vanamala, J.K., Monga, V., 2019. Prior information guided regularized deep learning for cell nucleus detection. IEEE Trans. Med. Imaging 38 (9), 2047–2058.

Wang, X., Gupta, A., 2018. Videos as space-time region graphs. In: Proceedings of the European Conference on Computer Vision. pp. 399–417.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3349–3364.

Welling, M., Kipf, T.N., 2016. Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z., 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: Proceedings of the European Conference on Computer Vision. pp. 322–339.

Xie, W., Noble, J.A., Zisserman, A., 2018a. Microscopy cell counting and detection with fully convolutional regression networks. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 6 (3), 283–292.

Xie, W., Noble, J.A., Zisserman, A., 2018b. Microscopy cell counting and detection with fully convolutional regression networks. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 6 (3), 283–292.

Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G., 2020. Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5295–5305.

Zhang, C., Chen, J., Li, B., Feng, M., Yang, Y., Zhu, Q., Bu, H., 2023. Difference-deformable convolution with pseudo scale instance map for cell localization. IEEE J. Biomed. Health Inf. 1–12. http://dx.doi.org/10.1109/JBHI.2023.3329542.