Cell counting via Attentive Recognition Network

Xiangyu Guo[1] · Jinyong Chen[1] · Guisheng Zhang[1] · Guofeng Zou[1] · Qilei Li[2] · Mingliang Gao[1*] ·

Received: date / Accepted: date

Abstract Accurately inferring the number of cells in a biomedical image is a fundamental yet challenging task for disease diagnosis. The early manual cell counting methods are time-consuming and prone to errors. With the advent of deep learning, convolutional neural network (CNN)-based cell counting has become a mainstream method. Despite the outstanding performance of these methods, the complex tissue background in medical images still hinders the improvement of counting performance. In this paper, we propose an attentive recognition network (ARNet) to resolve the problem. Specifically, it is composed of five convolution blocks and a channel attention (CA) module. The convolution blocks are employed to extract the basic features, and the CA module is introduced to suppress the complex background. Experimental results on cell counting datasets have verified that the proposed method outperforms the mainstream methods.

Keywords Healthcare · Cell counting · Attention mechanism · Convolutional neural network

1 Introduction

Cell counting, the purpose of which is inferring the number of cells in a given medical image, occupies a significant status in microscopy medical images analysis. The amount of cells can reflect the presence of disease [21], assist in determining tumor types [4], and aid in learning cellular and molecular genetic mechanisms [20]. The early manually cell counting is inefficient

✉ Mingliang Gao
E-mail: mlgao@sdut.edu.cn

1 School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China.

2 School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom.

and prone to error. Realizing high throughput processing by this means in real medical applications is not realistic. Therefore, it is desirable to design an efficient model to for accurate cell counting.

The early work implemented cell counting task by detection and regression. Generally, the counting by detection aims to discover the centroid position of cells and count them to predict the final result [17]. These methods require a high accuracy of centroid labelling, and perform unsatisfactory in medical images with occlusions. The regression-based methods [18, 25] tend to learn a mapping from the input to cell count. These methods can handle more complex scenes than the detection-based methods. Nevertheless, they only output the cell counts, and cannot provide the cell position information, which is also very valuable for medical image analysis [11]. Due to the powerful feature extraction and inferring ability of CNNs in computer vision[**? ?**], predicting a density map from the given image has been a mainstream solution [3, 10]. With the advance of attention mechanism, the object counting task has developed rapidly in many fields, *i.e.*, crowd counting [8, 26], vehicle counting [16, 27]. The attention-based counting methods minimizes the influence of irrelevant information by re-adjusting parameters in different dimensions [7]. Although the above methods have enhanced the performance of cell counting to some extent, the low contrast, variance in cell size and the complex tissue background in medical microscopic images still hinder the further improvement of the counting accuracy.

To mitigate the adverse effects of the complex tissue background, we propose the attentive recognition network (ARNet). It consists of five convolution blocks for extracting the low-level features. Especially, the first four blocks are employed to extract the basic features, and the last blocks adopts dilated convolution layers to enlarge the receptive field. Followed that, a channel attention (CA) module is introduced to suppress the background clutter. In a nutshell, the contributions of this paper are as follows.

1. We propose an ARNet to enhance the counting performance in cell images with complex tissue background.

2. We introduce a CA module to adjust the weight along the channel dimension, which aims to reduce the background weight and increase the foreground weight.

3. We carry extensive experiments to demonstrate the performance of ARNet. Meanwhile, ablation study is performed to prove the effectiveness of the individual components in the proposed model.

The remainder of this paper is arranged as follows. The related work is reviewed in Section 2. The proposed method is introduced in detail in Section 3. The details of the experiment and the conclusion of this paper are provided in Section 4 and 5, respectively.

2 Related work

In this section we describe three types of cell counting methods, *i.e.*, counting by detection, counting by regression and counting by density estimation.

Previous works complete the cell counting by the means of detection [1, 2] and regression[15, 18]. The counting by detection methods employ a detector to locate the cells and then sums up the cells as a count. Arteta *et al.* [2] proposed a tree-structured discrete graphical model to detect all the cells in microscopy images. The detection-based methods are difficult to deal with occlusion and shape variations. Therefore, the counting by regression methods are proposed to enhance the counting performance, which learn a mapping from the medical image to a count. Lempitsky *et al.* [15] built a general learning-based model for object counting tasks, *i.e.*, crowd counting and cell counting.

Recently, the development of CNN has further enhanced the accuracy of cell counting. Cohen *et al.* [5] built a redundant counting method to replace that predicting density map, which is helpful to address the complicated object counting task. Falk *et al.* [6] employed the U-Net to simultaneously complete the task of cell counting, detection and morphometric. Xie *et al.* [23] developed two parallel fully convolution regression networks, namely FCRN-A and FCRN-B to improve efficiency in an end-to-end manner. He *et al.* [11] built an auxiliary CNN model to boost the main regression model. Furthermore, the attention mechanism [22, 7] has inspired numerous researchers to use it to improve network performance. Guo *et al.* [9] incorporated a self-attention module to U-Net to enhance cell detection accuracy in microscopic images. Jiang *et al.* [12] built a weighted channel module to tackle the occlusions in cell counting. The module takes the residual connection and is easy to be plugged into any CNNs.

3 The proposed method

Given a medical microscopic image $X \in \mathbb{R}^{H \times W}$, a density map $Y \in \mathbb{R}^{H \times W}$ can be regressed by the CNN. The number of cells is obtained by integrating the pixels on the density map. The process can be represented by,

$$N = \int Y = \int F(X; \theta), \tag{1}$$

where the $F(X; \theta)$ denotes a function of density regression, in which θ represents the parameters to be learned by the network.

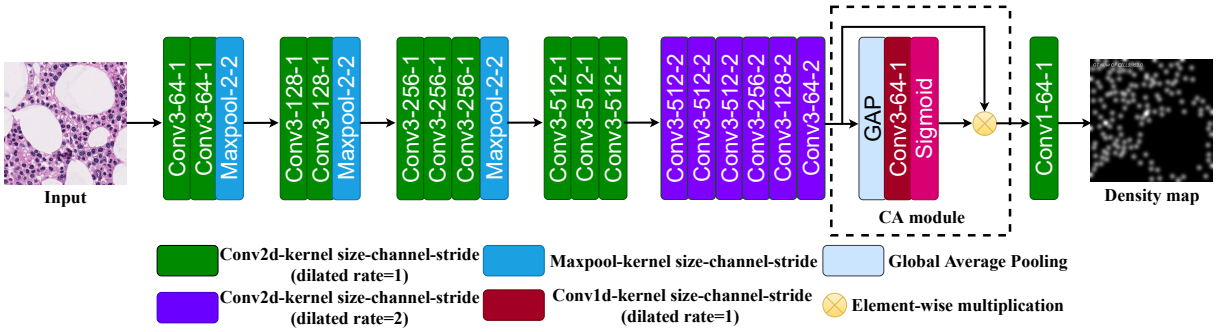

Fig. 1: The architecture of the proposed ARNet for cell counting.

3.1 Network design

The architecture of the proposed ARNet is depicted in Fig 1. It consists of five convolution blocks for extracting basic features and a channel attention (CA) module to cope with the tissue background. The first four blocks are composed of convolution layers with dilated rate of 1 and max pooling layers, which are adopted to extract the cell features with small size. The convolution layers with dilated rate of 2 in the final block are employed to enlarge the receptive field so that the network can capture cells with a large scale.

To address the problem of complex background in the medical image, the CA module [22] is introduced. In the last layer of the ARNet, a convolution layer with kernel size of 1×1 is employed to reduce the channels to 1 and output the density map. In a nutshell, the proposed ARNet is formulated as,

$$M = Conv_{1\times1}(f_{ca}(Block_i(\mathbf{X}))), i \in \{1,2,3,4,5\}, \tag{2}$$

where the f_{ca} denotes a function of channel attention. $Block_i$ represents the operation of the five convolution blocks.

3.2 Channel attention module

The purpose of channel attention is to select the targets (cells in this paper) by recalibrating the weight of each channel [7] adaptively. As shown in Fig. 1, the CA module contains three operations, *i.e.*, global average pooling (GAP), one-dimensional convolution layer and an activation function.

The GAP operation reduces the input 2D image to a 1D array for subsequent parameter adjustments. The fast 1D convolution operation is performed to produce the channel weights. The size of the convolution kernel is an adjustable parameter, which affects the final channel weight generation. The issue will be discussed in detail in the ablation study. Finally, we choose

the sigmoid activation function to output the optimized weight, which can perform element-wise multiplication with the initial input features and output the refined feature map. The map can suppress the side effects of background clutter and emphasize the areas where cells exist. Briefly, the CA module can be represented by the following formula,

$$P = M \otimes \mathrm{Sig}(\mathrm{Conv1d}(\mathrm{GAP}(M_i))), \tag{3}$$

where P represents the refined feature map, and M is the extracted basic feature map. Conv1d denotes a fast 1-dimension convolutional operation. Sig denotes the Sigmoid function.

3.3 Loss function

The MSE function is adopted as the loss function to optimize the network. It can minimize the Euclidean distance between the ground truth and the prediction. The function is formulated as,

$$Loss = \frac{1}{\mathrm{N}} \left\| y - \hat{y} \right\|_2^2, \tag{4}$$

where N denotes the number of test images. y and $\hat{y}$ denote the estimated and the ground truth values, respectively.

3.4 Ground truth generation

The ground truth map M_{gt} is generated by adopting a Gaussian kernel G_{σ_i} convolving a delta function:

$$M_{gt} = \sum_{i=1}^{H} \delta(x - x_i) * G_{\sigma_i}(x), \sigma_i = \beta \bar{d}_i, \tag{5}$$

where H denotes the number of cell annotations and x refers to the position pixel. σ_i represents the variance of the kernel, and the $\delta(x - x_i)$ depicts a target cell.

4 Experiment and analysis

4.1 Datasets

4.1.1 Synthetic bacterial cells

The synthetic bacterial cells (VGG) dataset was established by [15]. It includes 200 fluorescent microscopy images with resolution of 256×256. Because the dataset is synthetic, the cells almost have the same size, but they cluster together and are heavily obscured. Annotations are

automatically executed during the data construction process, which makes the labels error-free. Samples of VGG dataset are present in Fig 2.

Table 1: The information of VGG and MBM datasets.

Dataset	VGG	MBM
Resolution	256×256	600×600
Train/Validation/Test	50/50/100	15/15/14
Number of cell	174±64	126±33
Image type	synthetic	real

4.1.2 Modified bone marrow cells

The modified bone marrow (MBM) cell counting dataset contains 44 RGB microscopy images (600×600), which are collected from [13]. More specifically, the images are obtained from 11 Hematoxylin-Eosin images (1200×1200) of bone marrow tissue, which are from eight different patients. Each of the image is cut into four images of the same size. The labeled cells in this dataset have large-scale variation and non-uniform background, which make the counting task more troublesome. Fig 2 provides some examples of the MBM dataset. More details are reported in the Table 1.

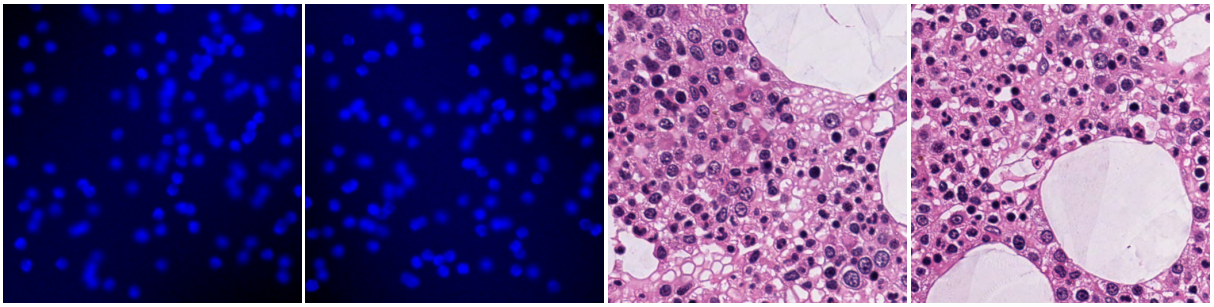

Fig. 2: Examples of VGG and MBM datasets. The left two images are from VGG dataset and the right two images take from MBM dataset.

4.2 Implementation details

All experiments are performed based on the PyTorch framework and are implemented on two NVIDIA 3080 GPUs. For data augmentation, we adopt random cropping, horizontal and vertical flipped. To avoid the overfitting, we set several dropout layers in the end of the network. At the training stage, Adam optimizer [14] with a learning rate of 1e-5 is employed to optimize the network. The weight decay is set as 1e-3. The batch size is set as 8 for VGG dataset, and 4 for MBM dataset. The maximum number of training epochs is set as 1000.

4.3 Evaluation metrics

We choose the mean absolute error (MAE) and its related standard deviations (STD) as the evaluation metrics. They are formulated as,

$$MAE = \frac{1}{S}\sum_{i=1}^{S}\left|C_{gt_i} - C_{est_i}\right|, \tag{6}$$

$$STD = \sqrt{\frac{1}{S-1}\sum_{i=1}^{S}\left(\left|C_{gt_i} - C_{est_i}\right| - MAE\right)^2}, \tag{7}$$

where S represents the number of test samples. C_{gt_i} and C_{est_i} denote the ground truth and estimated cell counts of the i-th images, respectively. The lower MAE and STD indicate that the model has better counting accuracy and counting stability.

4.4 Comparative analysis

We compare and analyze the proposed ARNet with other cell counting methods. The comparison results are reported in Table 1.

Table 2: The results of comparison on VGG and MBM datasets. Boldface indicates the best performance.

Methods	VGG		MBM	
	MAE	STD	MAE	STD
U-Net [6]	27.8	25.5	48.0	19.0
ResNet-152[25]	7.5	2.2	-	-
StructRegNet [24]	9.8	8.7	12.8	8.6
Mask R-CNN [10]	36.9	19.7	44.4	14.2
Marsden'method[18]	-	-	20.5	3.5
FCRN-A [23]	2.9	**0.2**	21.3	9.4
FCRN [19]	2.8	2.5	8.5	7.6
ARNet(Ours)	**2.7**	2.2	**5.0**	**3.2**

On VGG dataset, one can see that the ARNet ranks first with an MAE score of 2.7, which achieves 6.8% gains compared with the second-best method FCRN-A [23]. The STD metric is also competitive, ranking second with a score of 2.2, indicating that the counting stability of the ARNet on the VGG dataset could be further improved. It is a remarkable fact that the proposed method has the same STD score as the ResNet-152 [25], but the MAE score is 64% lower than it.

On the MBM dataset, the proposed ARNet scores 5.0 and 3.2 in MAE and STD, which outperform all the reported methods. Compared with FCRN [19], it improves by 41.1% and

57.9% in MAE and STD, respectively. Furthermore, the Marsden'method[18] also adopts the GAP operation to compress the image to generate new weights. However, there is still a big gap in MAE (75.6%).

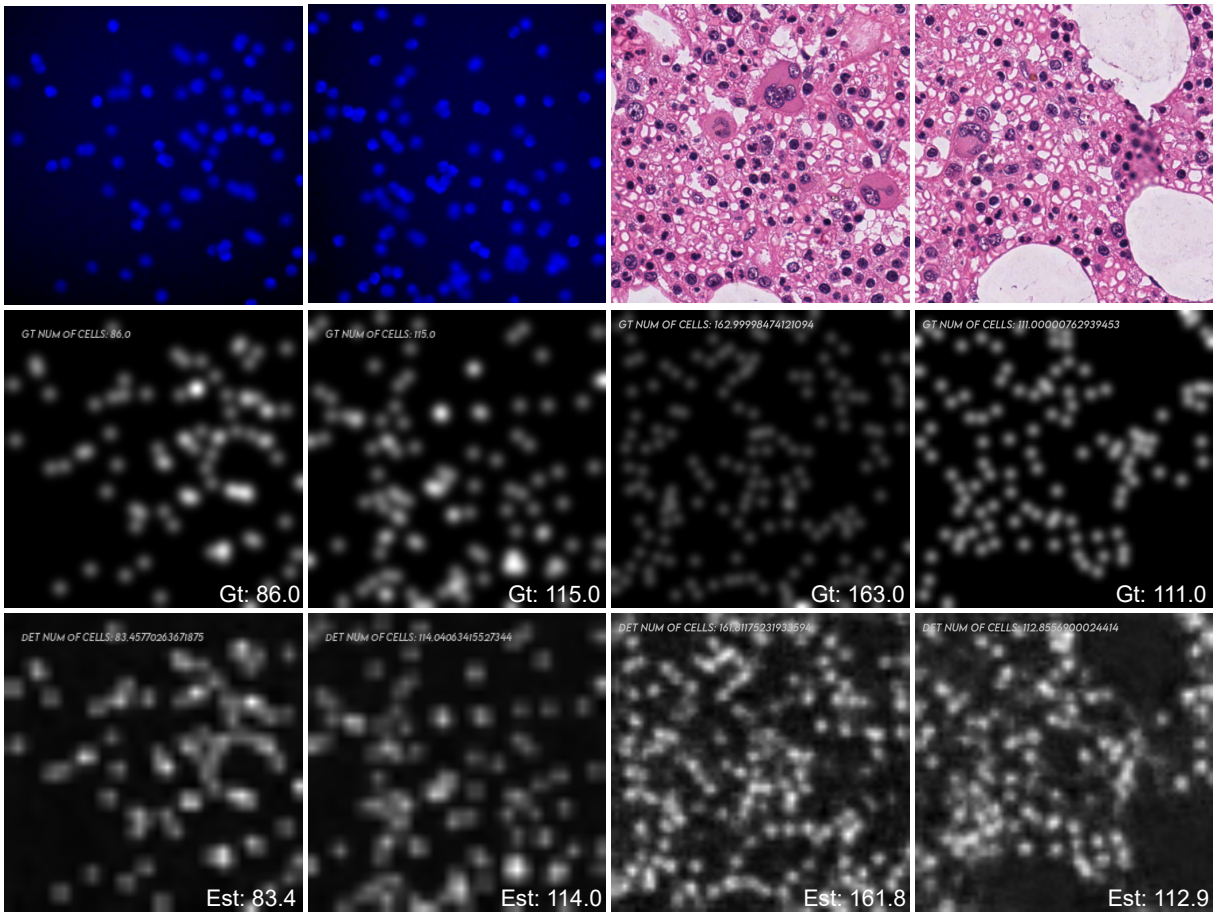

Fig. 3: Visualized results of the VGG and MBM datasets.

4.5 Ablation study

Ablation studies are carried out on MBM dataset to explore the benefits of the CA module in this section. The experimental configuration items are as follows.

'baseline' represents only five convolution blocks are employed as the basic network.

'baseline + CA (k=n)' means adding the CA module to the baseline, and the kernel size of the 1d convolution is set to n.

The ablation experimental results are tabulated in Table 3. The item 'baseline' scores 9.6 and 7.6 in MAE and STD, respectively. It obtains the worst performance compared with other items. We can observe that the CA module is beneficial to boosting the counting accuracy and stability. Specifically, the score of item 2 (k=3) is better than the item 3 (k=5) in MAE, while the opposite is true for STD. By contrast, the item 4 achieves the best performance in both

Table 3: The results of comparison on VGG and MBM datasets. Boldface indicates the best performance.

Methods	MAE	STD
baseline	9.6	7.6
baseline+ CA(k=3)	6.6	5.1
baseline+ CA(k=5)	5.8	5.9
baseline+ CA(k=1)	5.0	3.2

MAE and STD. The reason for this phenomenon is that the size of cells in medical images is generally small, and the use of large convolution kernel will lead to the loss of details and thus fail to achieve the optimal results.

5 Conclusion

In this paper, we propose an ARNet for accurate cell counting. It includes five basic convolution blocks for extracting low-level features, and a CA module to deal with the complex tissue background in a biomedical image. The CA module aims to re-adjust the weight along the channel dimension through GAP, 1D convolution and activation function. The feature map refined by the CA module can minimize the weights of background and improve the weight of foreground. Experimental results prove that the ARNet outperforms the mainstream methods in cell images with background clutter.

Acknowledgements This work is supported in part by the National Natural Science Foundation of Shandong Province (Nos. ZR2021QD041 and ZR2020MF127).

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arteta, C., Lempitsky, V.S., Noble, J.A., Zisserman, A.: Learning to detect cells using non-overlapping extremal regions. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **15 Pt 1**, 348–56 (2012)
2. Arteta, C., Lempitsky, V.S., Noble, J.A., Zisserman, A.: Detecting overlapping instances in microscopy images using extremal region trees. Medical image analysis **27**, 3–16 (2016)

3. Ciampi, L., Carrara, F., Amato, G., Gennaro, C.: Counting or localizing? evaluating cell counting and detection in microscopy images. In: VISIGRAPP (2022)
4. Coates, A.S., Winer, E.P., Goldhirsch, A., Gelber, R.D., Gnant, M., Piccart-Gebhart, M.J., Thürlimann, B., Senn, H.: Tailoring therapies—improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015. Annals of Oncology **26**, 1533 – 1546 (2015)
5. Cohen, J.P., Boucher, G., Glastonbury, C.A., Lo, H.Z., Bengio, Y.: Count-ception: Counting by fully convolutional redundant counting. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) pp. 18–26 (2017)
6. Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Bosco, C.D., Walsh, S., Saltukoglu, D., Tay, T.L., Prinz, M., Palme, K., Simons, M., Diester, I., Brox, T., Ronneberger, O.: U-net: deep learning for cell counting, detection, and morphometry. Nature Methods **16**, 67–70 (2018)
7. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. arXiv preprint arXiv:2111.07624 (2021)
8. Guo, X., Gao, M., Zhai, W., Shang, J., Li, Q.: Spatial-frequency attention network for crowd counting. Big data (2022)
9. Guo, Y., Stein, J.L., Wu, G., Krishnamurthy, A.K.: Sau-net: A universal deep network for cell counting. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (2019)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 386–397 (2020)
11. He, S., Minn, K.T., Solnica-Krezel, L., Anastasio, M.A., Li, H.: Deeply-supervised density regression for automatic cell counting in microscopy images. Medical image analysis **68**, 01892 (2021)
12. Jiang, N., Yu, F.: Cell counting with channels attention. 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP) pp. 494–498 (2020)
13. Kainz, P., Urschler, M., Schulter, S., Wohlhart, P., Lepetit, V.: You should use regression to detect cells. In: MICCAI (2015)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations (ICLR) (2015)
15. Lempitsky, V.S., Zisserman, A.: Learning to count objects in images. In: NIPS (2010)

16. Li, W., Wang, Z., Wu, X., Zhang, J., Peng, Q., Li, H.: Codan: Counting-driven attention network for vehicle detection in congested scenes. Proceedings of the 28th ACM International Conference on Multimedia (2020)
17. Liu, F., Yang, L.: A novel cell detection method using deep convolutional neural network and maximum-weight independent set. In: MICCAI (2015)
18. Marsden, M.A., McGuinness, K., Little, S., Keogh, C.E., O'Connor, N.E.: People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 8070–8079 (2018)
19. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. CoRR **abs/1312.6120** (2014)
20. Solnica-Krezel, L.: Conserved patterns of cell movements during vertebrate gastrulation. Current Biology **15**, R213–R228 (2005)
21. Venkatalakshmi, B., Thilagavathi, K.: Automatic red blood cell counting using hough transform. 2013 IEEE CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES pp. 267–271 (2013)
22. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11531–11539 (2020)
23. Xie, W., Noble, J.A., Zisserman, A.: Microscopy cell counting and detection with fully convolutional regression networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **6**, 283 – 292 (2018)
24. Xie, Y., Xing, F., Shi, X., Kong, X., Su, H., Yang, L.: Efficient and robust cell detection: A structured regression approach. Medical Image Analysis **44**, 245–254 (2018)
25. Xue, Y., Ray, N., Hugh, J.C., Bigras, G.: Cell counting by regression using convolutional neural network. In: ECCV Workshops (2016)
26. Zhai, W., Gao, M., Anisetti, M., Li, Q., Jeon, S., Pan, J.: Group-split attention network for crowd counting. Journal of Electronic Imaging (2022)
27. Zhang, J., Qiao, J., Wu, X., Li, W.: Vehicle counting network with attention-based mask refinement and spatial-awareness block loss. Proceedings of the 29th ACM International Conference on Multimedia (2021)