Exploring Unsupervised Cell Recognition with Prior Self-activation Maps

Pingyi Chen[1,2,3], Chenglu Zhu[2,3], Zhongyi Shui[1,2,3], Jiatong Cai[2,3], Sunyi Zheng[2,3], Shichuan Zhang[1,2,3], and Lin Yang[2,3](✉)

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2] Research Center for Industries of the Future, Westlake University, Hangzhou, China
{chenpingyi,yanglin}@westlake.edu.cn
[3] School of Engineering, Westlake University, Hangzhou, China

Abstract. The success of supervised deep learning models on cell recognition tasks relies on detailed annotations. Many previous works have managed to reduce the dependency on labels. However, considering the large number of cells contained in a patch, costly and inefficient labeling is still inevitable. To this end, we explored label-free methods for cell recognition. Prior self-activation maps (PSM) are proposed to generate pseudo masks as training targets. To be specific, an activation network is trained with self-supervised learning. The gradient information in the shallow layers of the network is aggregated to generate prior self-activation maps. Afterward, a semantic clustering module is then introduced as a pipeline to transform PSMs to pixel-level semantic pseudo masks for downstream tasks. We evaluated our method on two histological datasets: MoNuSeg (cell segmentation) and BCData (multi-class cell detection). Compared with other fully-supervised and weakly-supervised methods, our method can achieve competitive performance without any manual annotations. Our simple but effective framework can also achieve multi-class cell detection which can not be done by existing unsupervised methods. The results show the potential of PSMs that might inspire other research to deal with the hunger for labels in medical area.

Keywords: Unsupervised method · Self-supervised learning · Cell recognition

1 Introduction

Cell recognition serves a key role in exploiting pathological images for disease diagnosis. Clear and accurate cell shapes provide rich details: nucleus structure,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43993-3_54.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 559–568, 2023.
https://doi.org/10.1007/978-3-031-43993-3_54

cell counts, and cell density of distribution. Hence, pathologists are able to conduct a reliable diagnosis according to the information from the segmented cell, which also improves their experience of routine pathology workflow [5,14].

In recent years, the advancement of deep learning has facilitated significant success in medical images [17,18,20]. However, the supervised training requires massive manual labels, especially when labeling cells in histopathology images. A large number of cells are required to be labeled, which results in inefficient and expensive annotating processes. It is also difficult to achieve accurate labeling because of the large variations among different cells and the variability of reading experiences among pathologists.

Work has been devoted to reducing dependency on manual annotations recently. Qu et al. use points as supervision [19]. It is still a labor-intensive task due to the large number of objects contained in a pathological image. With regard to unsupervised cell recognition, traditional methods can segment the nuclei by clustering or morphological processing. But these methods suffer from worse performance than deep learning methods. Among AI-based methods, some works use domain adaptation to realize unsupervised instance segmentation [2,9,12], which transfers the source domain containing annotations to the unlabeled target. However, their satisfactory performance depends on the appropriate annotated source dataset. Hou et al. [10] proposed to synthesize training samples with GAN. It relies on predefined nuclei texture and color. Feng et al. [6] achieved unsupervised detection and segmentation by a mutual-complementing network. It combines the advantage of correlation filters and deep learning but needs iterative training and finetuning.

CNNs with inductive biases have priority over local features of the nuclei with dense distribution and semi-regular shape. In this paper, we proposed a simple but effective framework for unsupervised cell recognition. Inspired by the strong representation capability of self-supervised learning, we devised the prior self-activation maps (PSM) as the supervision for downstream cell recognition tasks. Firstly, the activation network is initially trained with self-supervised learning like predicting instance-level contrastiveness. Gradient information accumulated in the shallow layers of the activation network is then calculated and aggregated with the raw input information. These features extracted from the activation network are then clustered to generate pseudo masks which are used for downstream cell recognition tasks. In the inferring stage, the networks which are supervised by pseudo masks are directly applied for cell detection or segmentation. To evaluate the effectiveness of PSM, we evaluated our method on two datasets. Our framework achieved comparable performance on cell detection and segmentation on par with supervised methods. Code is available at https://github.com/cpystan/PSM.

2 Method

The structure of our proposed method is demonstrated in Fig. 1. Firstly, an activation network U_{ss} is trained with self-supervised learning. After the back-propagation of gradients, gradient-weighted features are exploited to generate the self-activation maps (PSM). Next is semantic clustering where the PSM is

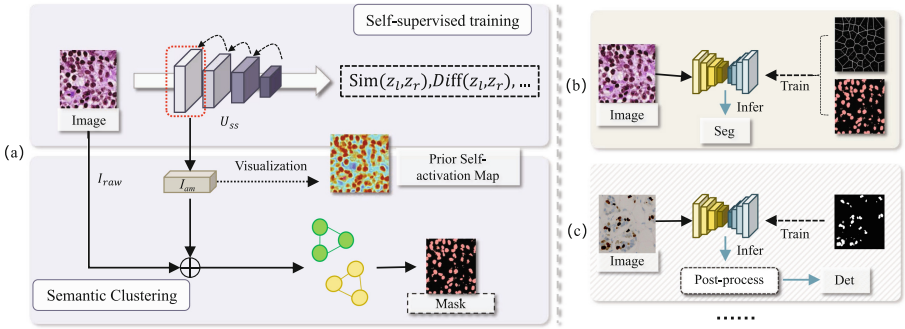

Fig. 1. The framework of our proposed model. (a) The top block shows the activation network which aggregates gradient information to generate self-activation maps. The bottom block presents the work process of semantic clustering. (b–c) The right part is the cell detection and the cell segmentation network which are supervised by the pseudo masks.

combined with the raw input to generate pseudo masks. These pseudo masks can be used as supervision for downstream tasks. Related details are discussed in the following.

2.1 Proxy Task

We introduce self-supervised learning to encourage the network to focus on the local features in the image. And our experiments show that neural networks are capable of adaptively recognizing nuclei with dense distribution and semi-regular shape. Here, we have experimented with several basic proxy tasks below.

ImageNet Pre-training: It is straightforward to exploit the models pre-trained on natural images. In this strategy, we directly extract the gradient-weighted feature map in the ImageNet pre-trained network and generate prior self-activation maps.

Contrastiveness: Following the contrastive learning [3] methods, the network is encouraged to distinguish between different patches. For each image, its augmented view will be regarded as the positive sample, and the other image sampled in the training set is defined as the negative sample. The network is trained to minimize the distance between positive samples. It also maximizes the distance between the negative sample and the input image. The optimization goal can be denoted as:

$$L_{dis}(Z_l, Z_r, Z_n) = diff(Z_l, Z_r) - diff(Z_l, Z_n), \tag{1}$$

where L_{dis} is the loss function. Z_l, Z_r, and Z_n are representations of the input sample, the positive sample, and the negative sample, respectively. In addition, $diff(\cdot)$ is a function that measures the difference of embeddings.

Similarity: LeCun et al. [4] proposed a Siamese network to train the model with a similarity metric. We also adopted a weight-shared network to learn the similarity discrimination task. In specific, the pair of samples (each input and

its augmented view) will be fed to the network, and then embedded as high-dimensional vectors Z_l and Z_r in the high-dimensional space, respectively. Based on the similarity measure $sim(\cdot)$, L_{dis} is introduced to reduce the distance, which is denoted as,

$$L_{dis}(Z_l, Z_r) = -sim(Z_l, Z_r) = diff(Z_l, Z_r). \tag{2}$$

Here, maximizing the similarity of two embeddings is equal to minimizing their difference.

2.2 Prior Self-activation Map

The self-supervised model U_{ss} is constructed by sequential blocks which contain several convolutional layers, batch normalization layers, and activation layers. The self-activation map of a certain block can be obtained by nonlinearly mapping the weighted feature maps A^k:

$$I_{am} = ReLU(\sum_k \alpha_k A^k), \tag{3}$$

where I_{am} is the prior self-activation map. A^k indicates the k-th feature map in the selected layer. α_k is the weight of each feature map, which is defined by global-average-pooling the gradients of output z with regard to A^k:

$$\alpha_k = \frac{1}{N} \sum_i \sum_j \frac{\partial z}{\partial A^k_{ij}}, \tag{4}$$

where i, j denote the height and width of output, respectively, and N indicates the input size. The obtained features are visualized in the format of the heat map which is later transformed to pseudo masks by clustering.

Semantic Clustering. We construct a semantic clustering module (SCM) which converts prior self-activation maps to pseudo masks. In SCM, the original information is included to strengthen the detailed features. It is defined as:

$$I_f = I_{am} + \beta \cdot I_{raw}, \tag{5}$$

where I_f denotes the fused semantic map, I_{raw} is the raw input, β is the weight of I_{raw}.

To generate semantic labels, an unsupervised clustering method K-Means is selected to directly split all pixels into several clusters and obtain foreground and background pixels. Given the semantic map I_f and its N pixel features $F = \{f_i | i \in \{1, 2, ...N\}\}$, N features are partitioned into K clusters $S = \{S_i | i \in \{1, 2, ...K\}\}$, $K < N$. The goal is to find S to reach the minimization of within-class variances as follows:

$$min \sum_{i=1}^{K} \sum_{f_j \in S_i} ||f_j - c_i||^2, \tag{6}$$

where c_i denotes the centroid of each cluster S_i. After clustering, the pseudo mask I_{sg} can be obtained.

2.3 Downstream Tasks

In this section, we introduce the training and inferring of cell recognition models.

Cell Detection. For the task of cell detection, a detection network is trained under the supervision of pseudo mask I_{sg}. In the inferring stage, the output of the detection network is a score map. Then, it is post-processed to obtain the detection result.

The coordinates of cells can be got by searching local extremums in the score map, which is described below:

$$T_{(m,n)} = \begin{cases} 1, & p_{(m,n)} > p_{(i,j)}, \quad \forall (i,j) \in D_{(m,n)}, \\ 0, & otherwise, \end{cases} \tag{7}$$

where $T_{(m,n)}$ denotes the predicted label at location of (m, n), p is the value of the score map and $D_{(m,n)}$ indicates the neighborhood of point (m, n). $T_{(m,n)}$ is exactly the detection result.

Cell Segmentation. Due to the lack of instance-level supervision, the model does not perform well in distinguishing adjacent objects in the segmentation. To further reduce errors and uncertainties, the Voronoi map I_{vor} which can be transformed from I_{sg} is utilized to encourage the model to focus on instance-wise features. In the Voronoi map, the edges are labeled as background and the seed points are denoted as foreground. Other pixels are ignored.

We train the segmentation model with these two types of labels. The training loss function can be formulated as below,

$$\begin{aligned} L = \lambda [y \log I_{vor} + (1-y) \log(1 - I_{vor})] \\ + (1-y) \log(1 - I_{sg}), \end{aligned} \tag{8}$$

where λ is the partition enhancement coefficient. In our experiment, we discovered that false positives hamper the effectiveness of segmentation due to the ambiguity of cell boundaries. Since that, only the background of I_{sg} will be concerned to eliminate the influence of false positives in instance identification.

3 Experiments

3.1 Implementation Details

Dataset. We validated the proposed method on the public dataset of Multi-Organ Nuclei Segmentation (MoNuSeg) [13] and Breast tumor Cell Dataset (BCData) [11]. MoNuSeg consists of 44 images of size 1000 × 1000 with around 29,000 nuclei boundary annotations. BCData is a public large-scale breast tumor dataset containing 1338 immunohistochemically Ki-67 stained images of size 640 × 640.

Table 1. Results on MoNuSeg. According to the requirements of each method, various labels are used: localization (Loc) and contour (Cnt). * indicates the model is trained from scratch with the same hyperparameter as ours.

Methods	Loc	Cnt	Pixel-level		Object-level	
			IoU	F1 score	Dice	AJI
Unet* [20]	✔	✔	0.606	0.745	0.715	0.511
MedT [24]	✔	✔	0.662	0.795	–	–
CDNet [8]	✔	✔	–	–	0.832	0.633
Competition Winner [13]	✔	✔	–	–	–	0.691
Qu et al. [19]	✔	✗	0.579	0.732	0.702	0.496
Tian et al. [22]	✔	✗	0.624	0.764	0.713	0.493
CellProfiler [1]	✗	✗	–	0.404	0.597	0.123
Fiji [21]	✗	✗	–	0.665	0.649	0.273
CyCADA [9]	✗	✗	–	0.705	–	0.472
Hou et al. [10]	✗	✗	–	0.750	–	0.498
Ours	✗	✗	0.610	0.762	0.724	0.542

Evaluation Metrics. In our experiments on MoNuSeg, F1-score and IOU are employed to evaluate the segmentation performance. Denote TP, FP, and FN as the number of true positives, false positives, and false negatives. Then F1-score and IOU can be defined as: $F1 = 2TP/(2TP + FP + FN)$, $IOU = TP/(TP+FP+FN)$. In addition, common object-level indicators such as Dice coefficient and Aggregated Jaccard Index (AJI) [13] are also considered to assess the segmentation performance.

In the experiment on BCData, precision (P), recall (R), and F1-score are used to evaluate the detection performance. Predicted points will be matched to ground-truth points one by one. And those unmatched points are regarded as false positives. Precision and recall are: $P = TP/(TP + FP)$, and $R = TP/(TP + FN)$. In addition, we introduce MP and MN to evaluate the cell counting results. 'MP' and 'MN' denote the mean average error of positive and negative cell numbers.

Hyperparameters. Res2Net101 [7] is adopted as the activation network U_{ss} with random initialization of parameters. The positive sample is augmented by rotation. The weights β are set to 2.5 and 4 for training in MoNuSeg and BCData, respectively. The weight λ is 0.5. The analysis for β and λ is included in the supplementary. Pixels of the fused semantic map will be decoupled into three piles by K-Means. The following segmentation and detection are constructed with ResNet-34. They are optimized using CrossEntropy loss by the Adam optimizer for 100 epochs with the initial learning rate of $1e^{-4}$. The function $diff(\cdot)$ is instantiated as the measurement of Manhattan distance.

Table 2. Results on BCData. According to the requirements of each method, various labels are used: localization (Loc) and the number (Num) of cells.

Methods	Loc	Num	Backbone	P	R	F1 score	MP↓	MN↓
CSRNet [15]	✔	✔	ResNet50	0.824	0.834	0.829	9.24	24.90
SC-CNN [23]	✔	✔	ResNet50	0.770	0.828	0.798	9.18	20.60
U-CSRNet [11]	✔	✔	ResNet50	0.869	0.857	0.863	10.04	18.09
TransCrowd [16]	✗	✔	Swin-Transformer	–	–	–	13.08	33.10
Ours	✗	✗	ResNet34	0.855	0.771	0.811	8.89	28.02

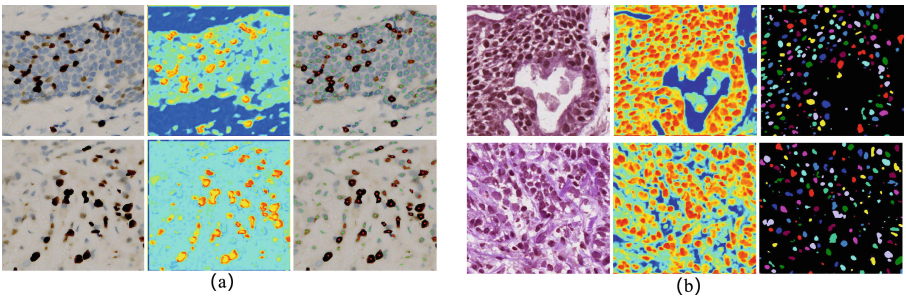

Fig. 2. Visualization. (a) Typical results of multi-class detection, the red dot and green dot represent positive and negative cells respectively. (b) Typical results of cell segmentation. (Color figure online)

3.2 Result

This section includes the discussion of results which are visualized in Fig. 2

Segmentation. In MoNuSeg Dataset, four fully-supervised methods Unet [20], MedT [24], CDNet [8], and the competition winner [13] are adopted to estimate the upper limit as shown in the first four rows of Table 1. Two weakly-supervised models trained with only point annotations are also adopted as the comparison. Compared with the method [22] fully exploiting localization information, ours can achieve better performance without any annotations in object-level metrics (AJI). In addition, two unsupervised methods using traditional image processing tools [1,21] and two unsupervised methods [9,10] with deep learning are compared. Our framework has achieved promising performance because robust low-level features are exploited to generate high-quality pseudo masks.

Detection. Following the benchmark of BCData, metrics of detection and counting are adopted to evaluate the performance as shown in Table 2. The first three methods are fully supervised methods which predict probability maps to achieve detection.

Furthermore, TransCrowd [16] with the backbone of Swin-Transformer is employed as the weaker supervision trained by cell counts regression. By con-

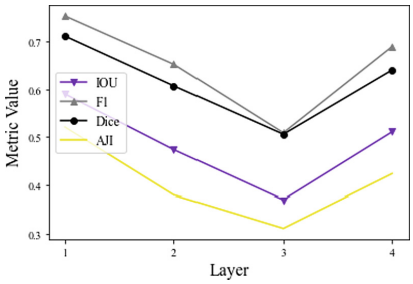

Fig. 3. Analysis of the depth of the extracted layer.

Table 3. The training strategy analysis

Training Strategy	F1	AJI
ImageNet Pretrained	0.658	0.443
Similarity	**0.750**	**0.542**
Contrastiveness	0.741	0.498

trast, even without any annotation supervision, compared to CSRNet [15], NP-CNN [23] and U-CSRNet [11], our proposed method still achieved comparable performance. Especially in terms of MP, our model surpasses all the baselines. It is challenging to realize multi-class recognition in an unsupervised framework. Our method still achieves not bad counting results on negative cells.

Ablation Study. Ablation experiments are built on MoNuSeg. In our pipeline, the activation network can be divided into four layers which consists of multiple basic units including ReLU, BacthNorm, and Convolution. We exploit the prior self-activation maps generated from different depths in the model after training with the same proxy tasks. As shown in Fig. 3, the performance goes down and up with we extracting features from deeper layers. Due to the relatively small receptive field, the shallowest layer in the activation network is the most capable to translate local descriptions

We have also experimented with different types of proxy tasks in a self-supervised manner, as shown in Table 3. We can see that relying on the pre-trained models with external data can not improve the results of subsequent segmentation. The model achieves similar pixel-level performance (F1) when learning similarity or contrastiveness. But similarity learning makes the model performs better in object-level metrics (AJI) than contrastive learning. The high intra-domain similarity hinders the comparison between constructed image pairs. Unlike natural image datasets containing diverse samples, the minor inter class differences in biomedical images may not fully exploit the superiority of contrastive learning.

4 Conclusion

In this paper, we proposed the prior self-activation map (PSM) based framework for unsupervised cell segmentation and multi-class detection. The framework is composed of an activation network, a semantic clustering module (SCM), and the networks for cell recognition. The proposed PSM has a strong capability of learning low-level representations to highlight the area of interest without the

need for manual labels. SCM is designed to serve as a pipeline between representations from the activation network and the downstream task. And our segmentation and detection network are supervised by the pseudo masks. In the whole training process, no manual annotation is needed. Our unsupervised method was evaluated on two publicly available datasets and obtained competitive results compared to the methods with annotations. In the future, we will apply our PSM to other types of medical images to further release the dependency on annotations.

Acknowledgement. This study was partially supported by the National Natural Science Foundation of China (Grant no. 92270108), Zhejiang Provincial Natural Science Foundation of China (Grant no. XHD23F0201), and the Research Center for Industries of the Future (RCIF) at Westlake University.

References

1. Carpenter, A.E., et al.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. **7**, 1–11 (2006). https://doi.org/10.1186/gb-2006-7-10-r100
2. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: towards cross-modality domain adaptation for medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 865–872 (2019)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 539–546. IEEE (2005)
5. Elston, C.W., Ellis, I.O.: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology **19**(5), 403–410 (1991)
6. Feng, Z., et al.: Mutual-complementing framework for nuclei detection and segmentation in pathology image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4036–4045 (2021)
7. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 652–662 (2021)
8. He, H., et al.: CDNet: centripetal direction network for nuclear instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4026–4035 (2021)
9. Hoffman, J., et al.: CyCADA: cycle-consistent adversarial domain adaptation. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1989–1998. PMLR (2018)

10. Hou, L., Agarwal, A., Samaras, D., Kurc, T.M., Gupta, R.R., Saltz, J.H.: Robust histopathology image analysis: to label or to synthesize? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8533–8542 (2019)
11. Huang, Z., et al.: BCData: a large-scale dataset and benchmark for cell detection and counting. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 289–298. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_28
12. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: a domain adaptive representation learning paradigm for object detection. IEEE (2019)
13. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging **36**(7), 1550–1560 (2017)
14. Le Doussal, V., Tubiana-Hulin, M., Friedman, S., Hacene, K., Spyratos, F., Brunet, M.: Prognostic value of histologic grade nuclear components of Scarff-Bloom-Richardson (SBR). An improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. Cancer **64**(9), 1914–1921 (1989)
15. Li, Y., Zhang, X., Chen, D.: CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Liang, D., Chen, X., Xu, W., Zhou, Y., Bai, X.: TransCrowd: weakly-supervised crowd counting with transformers. Sci. China Inf. Sci. **65**, 160104 (2021). https://doi.org/10.1007/s11432-021-3445-y
17. Liu, D., et al.: Nuclei segmentation via a deep panoptic model with semantic feature fusion. In: IJCAI, pp. 861–868 (2019)
18. Naylor, P., Laé, M., Reyal, F., Walter, T.: Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Trans. Med. Imaging **38**(2), 448–459 (2018)
19. Qu, H., et al.: Weakly supervised deep nuclei segmentation using points annotation in histopathology images. In: International Conference on Medical Imaging with Deep Learning, pp. 390–400. PMLR (2019)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Schindelin, J., et al.: Fiji: an open-source platform for biological-image analysis. Nat. Methods **9**(7), 676–682 (2012)
22. Tian, K., et al.: Weakly-supervised nucleus segmentation based on point annotations: a coarse-to-fine self-stimulated learning strategy. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 299–308. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_29
23. Tofighi, M., Guo, T., Vanamala, J.K.P., Monga, V.: Prior information guided regularized deep learning for cell nucleus detection. IEEE Trans. Med. Imaging **38**(9), 2047–2058 (2019)
24. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4