Journal of Intelligent & Fuzzy Systems 46 (2024) 4895–4907
DOI:10.3233/JIFS-235138
IOS Press

PmcaNet: Pyramid multiscale channel attention network for electron microscopy image segmentation

Kaihan Gao[a,b], Yiwei Ju[c], Shuai Li[c,d], Xuebing Yang[a], Wensheng Zhang[a,e] and Guoqing Li[a,b,*]
[a]*State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[b]*University of Chinese Academy of Sciences, Beijing, China*
[c]*National Center for Electron Microscopy in Beijing, School of Materials Science and Engineering, The State Key Laboratory of New Ceramics and Fine Processing, Key Laboratory of Advanced Materials (MOE), Tsinghua University, Beijing, China*
[d]*Focus e-Beam Technology (Beijing) Co., Ltd., Beijing, China*
[e]*College of Computer Science, Nankai University, Tianjin, China*

Abstract. Recent advances in high-throughput electron microscopy (EM) have revolutionized the examination of microstructures by enabling fast EM image generation. However, accurately segmenting EM images remains challenging due to inherent characteristics, including low contrast and subtle grayscale variations. Moreover, as manually annotated EM images are limited, it is usually impractical to utilize deep learning techniques for EM image segmentation. To address these challenges, the pyramid multiscale channel attention network (PmcaNet) is specifically designed. PmcaNet employs a convolutional neural network-based architecture and a multiscale feature pyramid to effectively capture global context information, enhancing its ability to comprehend the intricate structures within EM images. To enable the rapid extraction of channel-wise dependencies, a novel attention module is introduced to enhance the representation of intricate nonlinear features within the images. The performance of PmcaNet is evaluated on two general EM image segmentation datasets as well as a homemade dataset of superalloy materials, regarding pixel-wise accuracy and mean intersection over union (mIoU) as evaluation metrics. Extensive experiments demonstrate that PmcaNet outperforms other models on the ISBI 2012 dataset, achieving 87.85% pixel-wise accuracy and 73.11% mean intersection over union (mIoU), while also advancing results on the Kathuri and SEM-material datasets.

Keywords: Electron microscopy, image segmentation, convolutional neural network, multiscale feature pyramid

1. Introduction

The preliminary material property evaluation phase entails a comprehensive analysis of the surface of materials. Optical microscopes have traditionally been the primary instruments used for visually examining microscopic objects. In contrast, Electron microscopy (EM) provides superior resolution compared to light-based imaging, as it uses an electron beam to capture high-resolution images, allowing observation of objects with wavelengths smaller than light. It is achieved by leveraging an electron beam to irradiate a solid substance and detecting scattered electrons to generate high-resolution EM images [1]. To bridge the gap between the microscopic and macroscopic scales, image segmentation algorithms

*Corresponding author. Guoqing Li, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. E-mail: guoqing.li@ia.ac.cn.

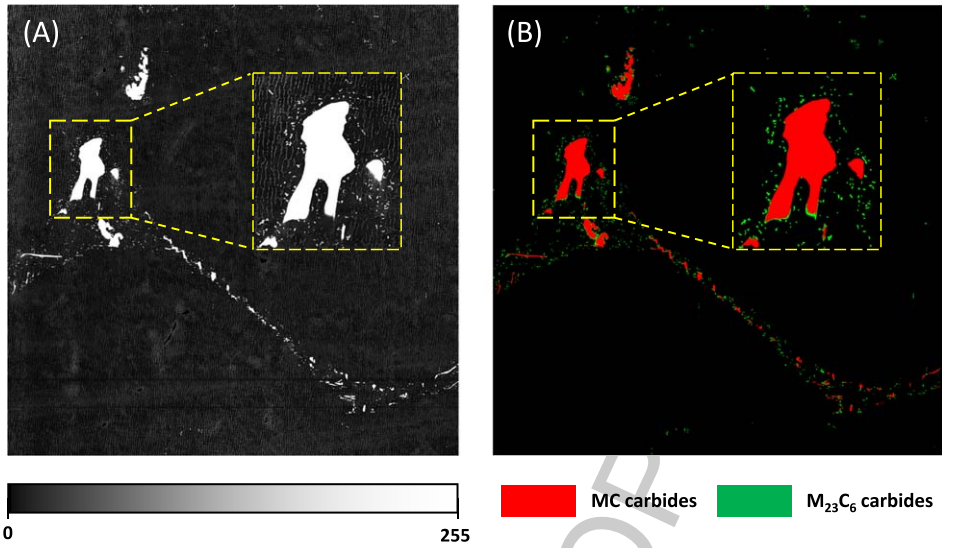

Fig. 1. An illustration of the EM image segmentation challenges, which shows the original image from our homemade SEM-material dataset (A) and the ground truth (B). The region of interest is annotated using a yellow dash box. The absence of color information and the presence of class imbalance are challenging issues.

are employed to analyze EM images and extract information about the material's surface or near-surface structure. This research paradigm has found widespread application in various fields, including medicine [2], biology [3], and other relevant domains. It plays a crucial role in advancing fundamental disciplines.

With advancements in electron microscopy imaging technology, these techniques now generate high-resolution EM images with data volumes ranging from gigabytes to terabytes [4]. Figure 1 depicts our homemade SEM-material dataset alongside the corresponding ground truth, which was derived from an EM scanning result of a high-temperature alloy material. There is a significant difference in pixel distribution between EM images and natural images. EM images, in contrast to natural images, exhibit minimal color information, significant scale variations, and a pronounced class imbalance [5]. This observation implies that methods of segmentation designed primarily for natural images may struggle to produce promissing results when applied to EM images. Consequently, segmenting EM images with volumes up to terabytes presents a formidable challenge. Recently, several machine learning techniques, such as ilastik [6], TrakEM2 [7] and Microscopy Image Browser [8], have been proposed as potential solutions to this issue. While these methods have demonstrated satisfactory inference metrics, deep learning models offer further advancements through their enhanced feature extraction capabilities, enabling them to better fit the data.

Recently, various deep learning models have successfully been applied to segment EM images. An illustrative example is U-net [9], which enhances the architecture of fully convolutional networks (FCNs) [10] by incorporating upsampling in a reverse expansive path along with skip connections. This module facilitates improved mask output and fosters a comprehensive learning framework. Moreover, EM-net [11], a scalable deep convolutional neural network that extends the U-net [9] architecture, was specifically developed for segmenting EM images. This study demonstrates the ability to achieve effective learning outcomes with a limited set of ground truth samples in the context of 2D EM image segmentation tasks.

However, due to the high resolution of EM images, it is necessary to divide them into smaller patches during model training. Unfortunately, this partitioning inevitably results in a reduction in available contextual information [12]. Additionally, the limited spatial range of convolutional filters restricts the incorporation of comprehensive global information from the image. This limitation is particularly significant for image segmentation tasks that heavily rely on the utilization of global context. Furthermore, the process of providing ground truth samples for electron microscopy datasets presents considerable challenges due to the low contrast exhibited by these

samples. According to a study [13], manual data segmentation or annotation costs an average of $10 per μm^3, which might amount to thousands of dollars for a relatively large data volume. The absence of annotated data poses a significant obstacle for EM image segmentation tasks.

To address these limitations, a pyramid multiscale channel attention network (PmcaNet) is proposed for enhancing EM image segmentation. In contrast to other methods, PmcaNet distinguishes itself by leveraging pyramid multiscale feature extraction and attention mechanisms to capture contextual relationships spanning various feature scales. This approach enables effective learning from a limited number of ground truth samples and concurrently enhances the accuracy of 2D EM image segmentation. Furthermore, the paper introduces the lightweight adaptive channel attention (LACA) module to reduce computational complexity while ensuring performance. By extracting attention vector of each channel, the LACA module captures intricate nonlinear structures present in EM images with fewer computations. The incorporation of multiscale contextual information in PmcaNet, achieved by harnessing multiscale attention information to enhance features, contributes to an improved accuracy in segmenting larger objects within EM images.

In summary, the following four contributions are offered by this work: (i) A novel EM dataset named SEM-material is proposed, focusing specifically on superalloy materials. (ii) A new architecture named PmcaNet is specifically developed to enhance the utilization of multiscale attention value and context information. (iii) A lightweight adaptive channel attention (LACA) module is designed to improve the understanding of the complex structures present in EM images and reduce computational complexity while ensuring efficient learning. (iv) The effectiveness of PmcaNet is extensively evaluated on our homemade SEM-material dataset as well as two general EM datasets, ISBI 2012 [14] and Kathuri [15]. Extensive experiments reveal that PmcaNet achieves competitive results in EM image segmentation.

The subsequent sections are organized as follows. In Section 2, related work concerning natural image segmentation and electron microscopy image segmentation is explored. Section 3, describes each module of our approach. The experiments and analysis are presented in Section 4. Section 5 introduces the examination of the individual roles of each module and an assessment of the model's convergence. Finally, Section 6 concludes this paper.

2. Related works

2.1. Natural image segmentation

Computer vision researchers have predominantly focused on natural images, which capture ordinary scenes from the natural environment and exhibit abundant and high-quality content. In this domain, deep learning models based on encoder-decoder architectures have emerged as the prevailing approach, with particular prominence given to fully convolutional networks (FCNs) [10]. Several methods have been proposed to enhance the receptive field of convolutional networks, including the use of dilated or atrous convolutions [16]. Initial approaches [17–20] involved a sequence of consecutive convolutions followed by spatial pooling to generate dense predictions. For instance, PSPNet [21] employed spatial pyramid pooling to capture contextual information at multiple scales, while DeepLabv3+ [22] integrated atrous spatial pyramid pooling to achieve an efficient encoder-decoder architecture. OneFormer [23] employs a backbone and pixel decoder to extract multi-scale features from the input image, yielding a more densely packed feature representation. Recently, many studies have explored the integration of attention mechanisms into encoder feature maps to replace coarse pooling. Sophisticated techniques have been proposed to enhance channel-wise dependencies [24–26], and spatial attention mechanisms [27–29] have been integrated with other techniques to improve long-range dependency acquisition. SegViT [30] introduces an Attention-to-Mask (ATM) decoder module, which harnesses the spatial information within the attention map to generate mask predictions for each category. However, due to the substantial disparities in grayscale distribution and other characteristics between electron microscopy images and natural images, achieving satisfactory results on EM images through the application of natural image segmentation methods is challenging.

2.2. Electron microscopy image segmentation

Electron microscopy involves collecting electrons that reflect from the imaged specimen surface, resulting in EM images [11]. The pixel distribution observed in EM images differs significantly from that of natural images. In recent years, deep learning models have made notable advancements in addressing the challenges of semantic segmenta-

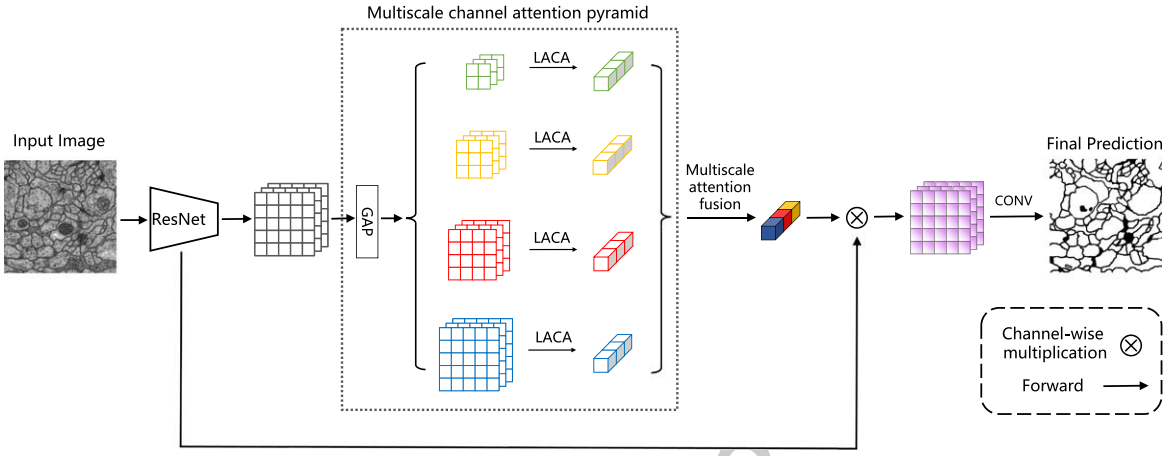

Fig. 2. An overview of the proposed PmcaNet. The flow of information is indicated by the black arrows. Within the architecture, "GAP" denotes a global average pooling layer, and "LACA" signifies lightweight adaptive channel attention.

tion in EM images, surpassing traditional approaches that relied on manually engineered features [31–33]. Previous research [34, 35] on scanning electron microscopy (SEM) images of sandstone and transmission electron microscopy (TEM) images of mouse brain slices has provided evidence supporting this phenomenon. Oztel et al. [36] introduce a deep convolutional neural network that incorporates a sliding window strategy and subsequent postprocessing steps to enhance its performance. MitoNet [37] introduced a generalizable model for segmenting individual mitochondria across volume electron microscopy datasets. Nonetheless, these methods have not effectively tackled the challenges associated with the absence of edge information and incomplete object segmentation in EM images. We contend that the acquisition of supplementary global contextual information is imperative for addressing these issues.

3. Method

3.1. Overview

Figure 2 provides an overview of the proposed model. The backbone network of PmcaNet is based on the ResNet-50 model, which incorporates dilated convolutions to expand the receptive field. The choice of the ResNet-50 model as the backbone network for PmcaNet was made to strike a balance between model capacity and computational efficiency, which is particularly well-suited for our segmentation task. The final feature map from the backbone network is then fed into the multiscale channel attention pyramid module. This module captures global contextual dependencies across different geographical dimensions. Distinguished from alternative approaches, within the multiscale channel attention pyramid module, the lightweight adaptive channel attention (LACA) module is utilized to compute channel attention on feature maps at multiple scales. The channel attention vector is then produced by the multiscale attention fusion module, which incorporates attention values from different dimensions. Finally, the low-level features collected from the backbone network are fused with the enhanced high-level features, generating the final features used for pixel-level mask prediction.

3.2. Multiscale channel attention pyramid

In semantic segmentation tasks with complex scenes, integrating multiscale information has shown significant performance improvement. Incorporating attention mechanisms allows the model to selectively focus on important regions, enhancing the segmentation results for objects of varying dimensions. Hence, it is necessary to independently derive attention of different scales and utilize it to enhance features. This methodology optimizes the use of contextual information from lower-level feature maps, improving segmentation performance.

The multiscale channel attention pyramid module extracts feature maps applying the same pyramid pooling parameters as PSPNet [21]. By performing pooling operations with kernel sizes of 1×1, 2×2, 3×3, and 6×6, feature maps at different scales are generated, capturing information at various scales.

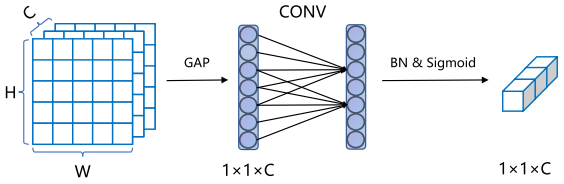

Fig. 3. The specifics of our lightweight adaptive channel attention (LACA) module. In the diagram, "GAP" corresponds to global average pooling, "CONV" represents 1D convolution, and "BN" stands for batch normalization.

The lightweight adaptive channel attention (LACA) module is then applied to collect the global priors for each layer to gather attention values covering the entire image and those covering half of the image and small portions. Finally, the adaptive fusion of attention maps yields contextual information that encompasses a broader range of semantic details.

3.3. Lightweight adaptive channel attention (LACA)

Channel attention mechanisms, leveraging squeeze and activation operations, play a crucial role in facilitating feature integration and recalibration. The lightweight adaptive channel attention (LACA) module is designed to obtain channel attention values, enabling the capture of channel dependencies and enhancing the feature representation capability. Figure 3 provides an overview of the zooming process, while the LACA module can be represented as follows:

$$Atten = \text{sigmoid}\left(\text{BN}\left(\text{Conv}_k(\text{GAP}(F))\right)\right), \quad (1)$$

where F represents the input feature map. The LACA module aggregates the convolutional features by diminishing their dimensions using global average pooling (GAP). Then, a 1D convolution operation is performed with a kernel size k. Finally, the attention value is generated utilizing a sigmoid function following batch normalization (BN), to achieve a steady activation value distribution throughout training. Using 1D rapid convolution for channel calibration is suggested to minimize network parameters and prevent dimension reduction. While the size of the convolution kernel k can be manually determined, this method lacks generalization and may not apply to diverse samples.

Recently, Qilong Wang et al. [25] have proposed a mapping relationship between the number of feature map channels C and the convolution kernel size k to better represent the correlation between high-dimensional and low-dimensional channels:

$$C = \phi(k). \quad (2)$$

The linear relationship is acknowledged as the most fundamental mapping function. However, the presence of sparse feature relationships can pose challenges in representing the interconnections between channels. Hence, the employment of nonlinear functions to facilitate the mapping process is taken into consideration. Exponential functions are particularly favored for their effectiveness in integrating information across multiple channel dimensions. Therefore, exponential functions are used as the mapping function:

$$\phi(k) = \alpha e^k + \beta. \quad (3)$$

Then, the size of the convolution kernel can be adaptively determined by the channels C:

$$k = \left|\ln\left(\frac{C-\beta}{\alpha}\right)\right|_{odd}, \quad (4)$$

where $|X|_{odd}$ represents the closest odd number to X. Choosing an odd number for the convolution kernel k is inspired by the prevalent approach in the field of image processing. In this study, α and β are consistently set to the values 2 and 1, respectively, throughout all experiments. The nonlinear transformation can facilitate the amplification of long-distance interactions in high-dimensional channels, while short-distance interactions are more prominent in low-dimensional channels.

3.4. Multiscale attention fusion (MAF)

After obtaining attention information at multiple spatial scales, a straightforward additive fusion technique can be used to derive multiscale attention weights. However, it is important to note that the relative importance of different hierarchical attention modules may vary and should not be presumed to be equal. Determining the weights experimentally can be computationally expensive and may not guarantee generalization. Weizhen Wang et al. [38] have suggested employing an adaptive learning strategy to determine the magnitudes of the weights. Similarly, in this study, the weight vector is derived by applying a 1×1 convolution operation to the various hierarchical level outputs:

$$w = \text{softmax}(\text{concat}(\text{C1D}(Atten_i))), \quad (5)$$

where $Atten_i$ represents attention information from the four pyramid levels, and i can take values of 0, 1, 2, or 3. Attention weights are acquired using a 1×1 convolution operation called C1D, producing a weight for each input. The value of the weight is learned through the training process. After combining the channel attention information from different levels, the softmax operation is used to generate the weight vector w. The final multiscale attention information, denoted as $Atten$, is obtained by combining the channel attention information with the weight vector:

$$Atten = \sum_{i=0}^{3} w_i Atten_i. \tag{6}$$

Finally, the attention information is fused with the low-level feature F as follows:

$$Out = Atten \otimes F. \tag{7}$$

where $\otimes$ represents the channel-wise multiplication operarion between the low-level feature F and the high-level feature $Atten$. The feature map Out is then upsampled to the original image resolution using transposed convolution. The output is further passed through a 2D convolutional layer with a kernel size of 1 for pixel-by-pixel classification, generating the final segmentation mask.

3.5. Training and inference

The sample imbalance issue in EM images is considered during the model training process. To address this bottleneck, Dice loss [39] is utilized as the loss function. The Dice loss encourages the model to allocate more attention to foreground regions that may have sparse features, effectively addressing the challenge posed by class imbalance:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2}, \tag{8}$$

where N represents the number of pixels, p_i denotes the predicted probability of the i-th pixel, and g_i denotes the ground truth of the i-th pixel. The Dice loss function is implemented to mitigate the negative effects of the unbalanced distribution of foreground and background samples. This technique demonstrates high efficacy in improving segmentation results, particularly in scenarios involving small foreground regions like EM images.

Additionally, the multiclass cross-entropy (CE) loss [40] is employed to evaluate the pixel-wise classification error:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} g_i \log p_i. \tag{9}$$

Consequently, PmcaNet is trained by minimizing the overall loss objective:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{Dice}, \tag{10}$$

where λ_1 and λ_2 are tradeoff parameters that control the weights of the two losses. In this paper, λ_1 and λ_2 are set to 1 and 3, respectively.

4. Experiments

4.1. Datasets and metrics

Figure 4 illustrates some samples from the three datasets ISBI 2012 [14], Kathuri [15], and SEM-material used in the experiments, with details as follows:

ISBI 2012. This dataset is from the 2012 ISBI EM segmentation challenge [14]. The training dataset of this competition consists of 30 stacked EM images of successive sections of the ventral nerve cord of a Drosophila larva in its first instar, with binary labels representing cell membranes and neuronal cell bodies. The test dataset has the same dimensions as the training dataset, comprising 30 images with a resolution of 512×512.

Kathuri. It consists of scanning electron microscopy (SEM) data of the left ventricular myocytes of three mice [15]. The dataset consists of two classes: cardiac muscle mitochondria and the background. Each image in this dataset has a resolution of 1334×1334. The training set includes 96 images, the validation set has 32 images, and the testing set consists of 32 images.

SEM-material. The homemade SEM-material dataset contains 130 images generated through scanning electron microscopy (SEM) at a resolution of 8192×8192. The data were collected using a single-beam high-throughput SEM, Navigator-100 (Focus e-Beam Technology, Beijing), equipped with a direct electron detector and a finely tuned deflection system. It allows the capture of SE (secondary electron) and BSE (backscattered electron) images simultaneously at a speed of up to 2×100 megapixels/second [4]. The SEM-material dataset consists of two distinct types of superalloy carbides used for segmentation: primary Ta/Hf-rich *MC* carbides and secondary Cr/Re-rich $M_{23}C_6$ carbides. Due to GPU

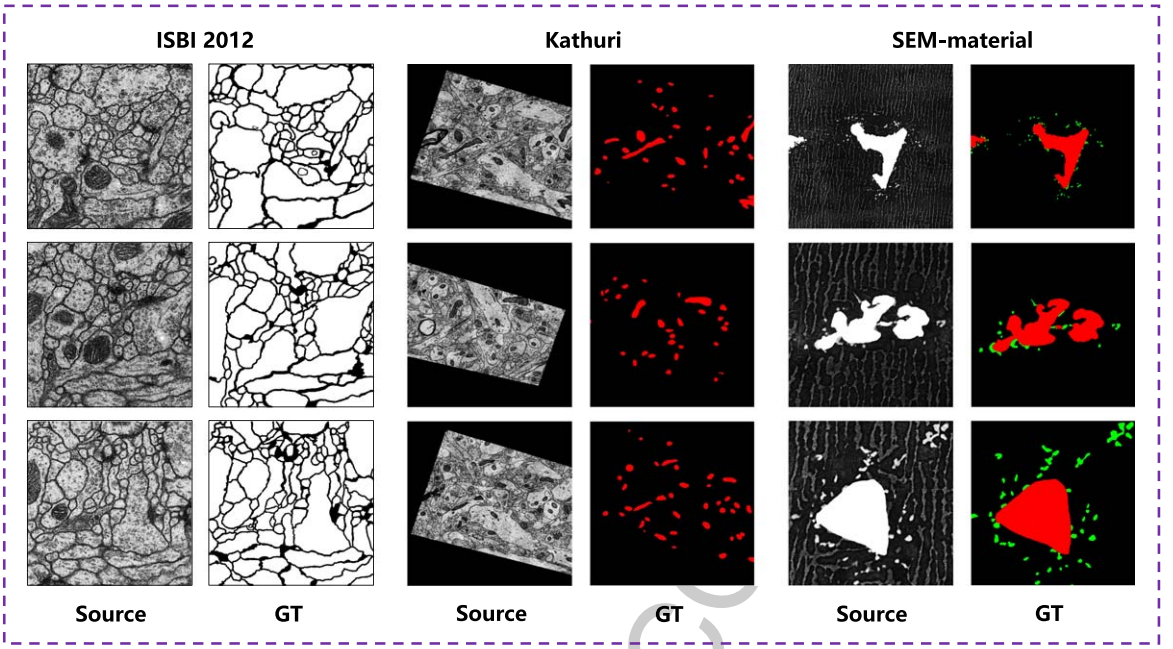

Fig. 4. Some examples of ISBI 2012, Kathuri, and SEM-material datasets. 'Source' indicates the original images, while 'GT' represents the corresponding ground truth.

memory limitations, the images were divided into patches of size 2048×2048 before training. After the division process, the dataset was divided into training, validation, and test sets containing 1248, 416, and 416 images, respectively.

Metrics. For evaluation, the mean intersection over union (mIoU) and pixel-wise accuracy (pixel acc.) were adopted to measure the similarity between the predicted results, the ground truth, and the proportion of correctly classified pixels to all pixels.

4.2. Implementation details

The backbone network used for PmcaNet is ResNet-50, pre-trained on the ImageNet-1k dataset. During training, the data augmentation pipeline from the MMSegmentation [41] library is applied. This pipeline includes random horizontal flipping, random cropping, and random resizing with a scale between 0.5 and 2.0. Stochastic gradient descent (SGD) [42] is employed for training the models with a fixed momentum of 0.9. Additionally, the 'poly' learning rate schedule is utilized, defined as $\gamma = \gamma_0 \left(1 - \frac{N_{iter}}{N_{total}}\right)^{0.9}$, where N_{iter} represents the current iteration number, and N_{total} represents the total number of iterations. In our study, a standard holdout validation methodology is employed. The hyperparameters are set up as follows for transfer learning and fine-tuning on different datasets:

a) ISBI 2012: By default, the initial learning rate is set to 0.01, the weight decay is set to 0.9, the crop size is set to 128×128, and the batch size is set to 16. If not supplied, the training iterations default at 40K.

b) Kathuri: By default, the initial learning rate is set to 0.005, the weight decay is set to 0.9, the crop size is set to 512×512, and the batch size is set to 8. If not supplied, the training iterations default at 20K.

c) SEM-material: By default, the initial learning rate is set to 0.005, the weight decay is set to 0.9, the crop size is set to 512×512, and the batch size is set to 8. If not supplied, the training iterations default at 40K.

All experiments are conducted on a workstation equipped with 8 NVIDIA A40 48G GPU cards.

4.3. Results analysis

A comprehensive evaluation is conducted to compare the performance of PmcaNet with the latest models on three distinct datasets: ISBI 2012, Kathuri, and SEM-material. The objective of this evaluation is to assess the effectiveness of PmcaNet in electron microscopy image segmentation. Table 1 presents the results, demonstrating the efficacy of the proposed approach in accomplishing the segmentation tasks.

Table 1
Comparison with state-of-the-art methods. 'pixel acc.' refers to pixel-wise accuracy, while 'mIoU' stands for mean intersection over union

Method	ISBI 2012		Kathuri		SEM-material		Param.
	pixel acc.	mIoU	pixel acc.	mIoU	pixel acc.	mIoU	
DeepLabv3+ [22] [CVPR'2018]	70.62	41.39	98.47	75.01	99.51	42.12	43.58M
EM-net [11] [ICPR'2021]	86.30	69.69	98.30	78.91	99.61	52.46	39.06M
Segmentor [27] [ICCV'2021]	86.41	70.70	98.36	74.59	99.53	43.01	102.50M
SegViT [30] [NIPS'2022]	76.88	58.91	98.46	74.82	99.62	56.68	96.75M
PIDNet [43] [CVPR'2023]	72.18	53.83	97.92	76.14	99.64	62.73	**11.65M**
PmcaNet(Ours)	**87.85**	**73.11**	**98.97**	**83.99**	**99.66**	**68.39**	32.79M

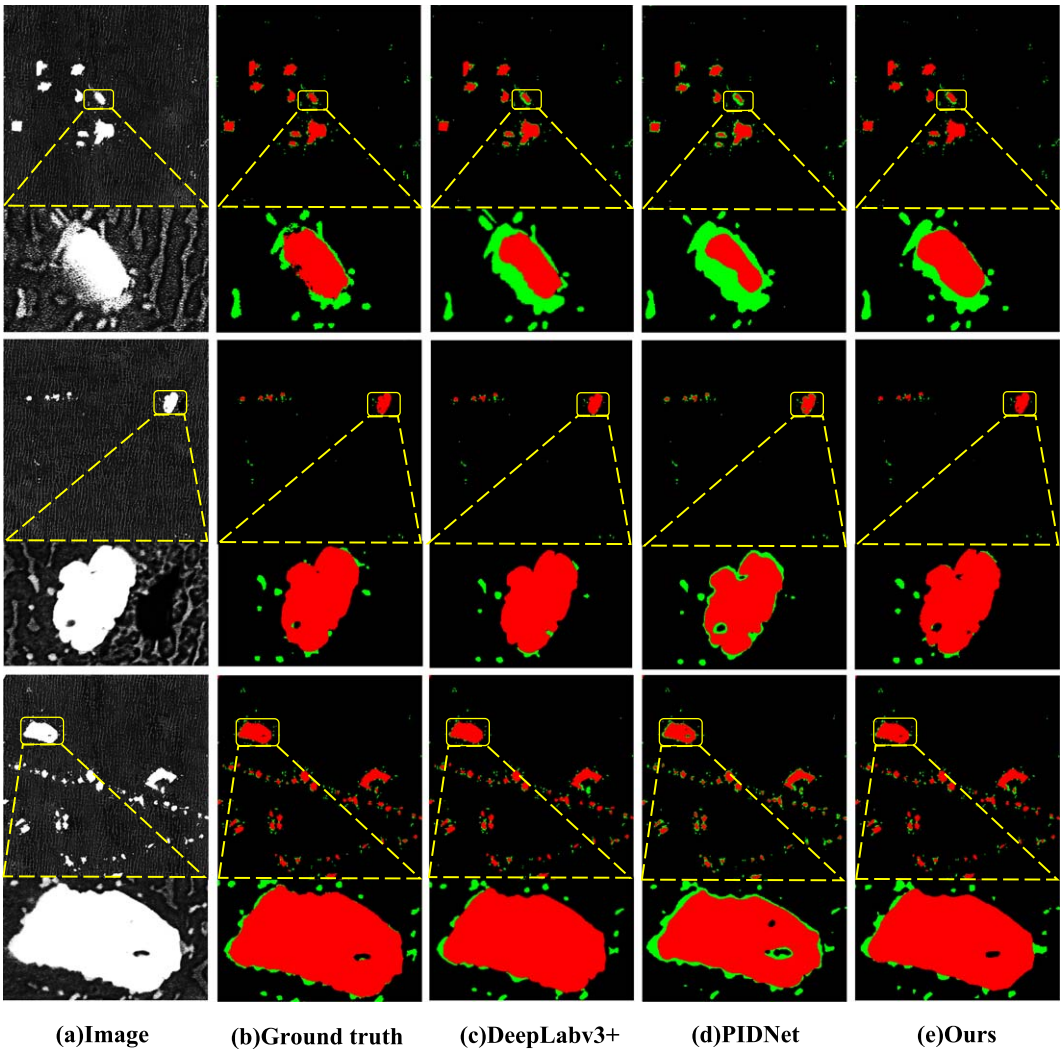

Fig. 5. Visual improvements on the SEM-material dataset. PmcaNet produces more accurate and detailed results.

On the ISBI 2012 dataset, PmcaNet achieves a pixel-wise accuracy (pixel accuracy) of 87.85% and a mean intersection over union (mIoU) of 73.11%, outperforming the other models. Similarly, on the Kathuri dataset, PmcaNet performs the best with a pixel accuracy of 98.97% and a mIoU of 83.99%. These outcomes illustrate the comparable segmentation performance of the proposed method for electron microscopy images.

Furthermore, when evaluated on the homemade SEM-material dataset, which presents challenges such as larger image sizes, cluttered backgrounds, and sparser foreground information, PmcaNet demonstrates exceptional performance. It achieves a pixel

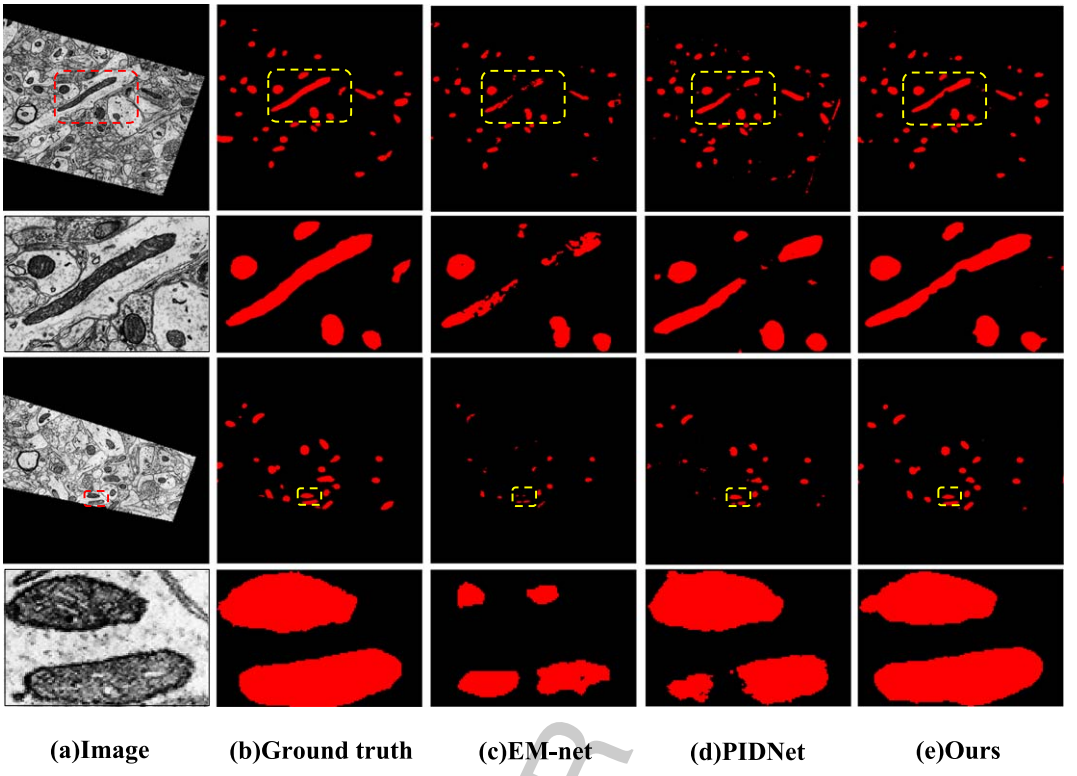

Fig. 6. Qualitative results of the proposed PmcaNet and other methods on the Kathuri dataset.

accuracy of 99.66% and a mIoU of 68.39%, surpassing other models evaluated in this study. Regarding the number of parameters, it is worth noting that while PIDNet [43] has fewer parameters, its performance is relatively lower. This accomplishment can be ascribed to the integration of a multiscale attention network within the primary decoder. This integration allows for the utilization of diverse information at various resolutions, aiding in the capture of comprehensive contextual details and enhancing feature representation.

4.4. Visualization of segmentation results

This section presents the comparative results of PmcaNet on various datasets. The test results from the SEM-material dataset and the Kasthuri dataset are visualized in Figs. 5 and 6, respectively. Figure 5 shows the results of PmcaNet, which indicate improved precision and greater object completeness on the SEM-material dataset. Notably, segmenting the primary phase (highlighted in red) and hole structures inside the primary phase (highlighted in black) indicates some improvement. In contrast, DeepLabv3+ [22] and PIDNet [43] encounter difficulties that result in gaps, causing misclassification of pixels.

In Fig. 6, a selection of results from the Kasthuri dataset's test set is shown. The results obtained by PmcaNet exhibit more complete objects and demonstrate superior performance in terms of contours and details, as compared to both EM-net [11] and PIDNet [43]. This improvement can be attributed to PmcaNet's effective incorporation of rich contextual information and comprehensive global information. The utilization of such information enables PmcaNet to effectively handle category boundaries, which is crucial for EM image segmentation tasks that lack color information and heavily rely on texture information.

5. Discussion

5.1. Ablation study

To conduct a comprehensive analysis of the methodology, a series of ablation experiments are devised to evaluate the efficacy of the network modules.

Table 2
Ablation study of multiscale channel attention pyramid

Method	Pixel Acc.	mIoU
ResNet50+MAX(w/o pyramid)	99.48	53.32
ResNet50+AVE(w/o pyramid)	99.61	55.85
ResNet50+MAX(w/ pyramid)	99.65	66.74
ResNet50+AVE(w/ pyramid)	**99.66**	**68.31**

Analysis of the multiscale channel attention pyramid. Experiments are conducted using either a single global feature or four-level feature pooling, comparing maximum and average pooling. The results in Table 2 indicate that average pooling exhibits superior performance compared to max pooling. Additionally, the incorporation of pyramid parsing for pooling leads to enhanced performance compared to relying solely on global pooling. In the context of the SEM-material dataset, attention pyramid and average pooling techniques result in a model performance of 99.66% pixel acc. and 68.31% mIoU score. These results provide evidence for the efficacy of these pooling strategies in segmenting electron microscopy images.

Effectiveness of the lightweight adaptive channel attention module. The efficacy of the LACA module is substantiated by conducting experiments that incorporate various attention modules, such as SE and ECA. Including attention modules in the baseline model led to performance enhancement, as shown in Table 3. Significantly, the LACA module exhibits the most remarkable improvement, emphasizing its capacity to enhance the model's attention to detail and overall performance. This serves as evidence of the pivotal role played by the LACA in our network.

Effectiveness of the multiscale attention fusion module. Experiments are carried out utilizing three fusion methods: simple additive fusion, elementwise multiplication (also known as the hardam product), and a novel multiscale attention fusion (MAF) approach. Table 4 shows the advantageous impact of the MAF module on the PmcaNet model, resulting in enhanced attention calibration and achieving the highest performance with a pixel accuracy of 99.66% and mIoU of 68.34%. This suggests that the MAF module achieves optimal fusion weights through a training process and acquires the ability to perform spatial filtering on predictions at each level. Consequently, the preservation of relevant information for combination enhances the network's accuracy in making predictions, substantiating the MAF module's critical role within our network.

Effectiveness of combining two loss functions. During training, the model's performance stagnates in the early epoch when using a single loss, specifically the cross-entropy loss. To address this issue, a weighted combination of cross-entropy loss and Dice loss is introduced as the overall loss Eq. (10). Different combinations of λ_1 and λ_2 are tested, and Table 5 shows that $\lambda_1 = 1$ and $\lambda_2 = 3$ achieve the best performance of 99.66% pixel-wise accuracy and 68.39% mIoU. It is demonstrated that the integration of cross-entropy loss and Dice loss plays a key role in our work. Therefore, in the experiments, λ_1 is set to 1, and λ_2 is set to 3.

5.2. In-depth view of PmcaNet

Moving forward, the complexity and convergence time of different methods will be analyzed to highlight the advantages of PmcaNet in terms of model complexity and convergence speed.

Figure 7 compares the parameter count and mIoU performance of various methodologies used for the SEM-material dataset analysis. Among them,

Table 3
Evaluating the effectiveness of different channel attention modules

Method	SE [24]	ECA [25]	LACA	Pixel Acc.	mIoU
PmcaNet(w/o attention)	×	×	×	99.62	65.49
PmcaNet	✓	×	×	99.64	67.83
PmcaNet	×	✓	×	99.64	67.51
PmcaNet	×	×	✓	**99.66**	**68.38**

Table 4
Ablation study with different attention fusion modules. 'Addition' represents that the fusion method is the form of addition, 'Multiplication' stands for bitwise multiplication and 'MAF' represents multiscale attention fusion

Method	Addition	Multiplication	MAF	Pixel Acc.	mIoU
PmcaNet	✓	×	×	99.65	68.14
PmcaNet	×	✓	×	99.65	67.94
PmcaNet	×	×	✓	**99.66**	**68.34**

Table 5
Performance comparison with or without diversity loss. λ_1 and λ_2 represent the weight of cross-entropy loss and Dice loss, respectively

Method	λ_1	λ_2	Pixel Acc.	mIoU
PmcaNet	1	0	99.24	51.20
PmcaNet	1	1	99.60	55.68
PmcaNet	1	2	99.48	40.29
PmcaNet	1	3	**99.66**	**68.39**
PmcaNet	1	4	99.48	44.16
PmcaNet	1	5	98.14	34.95

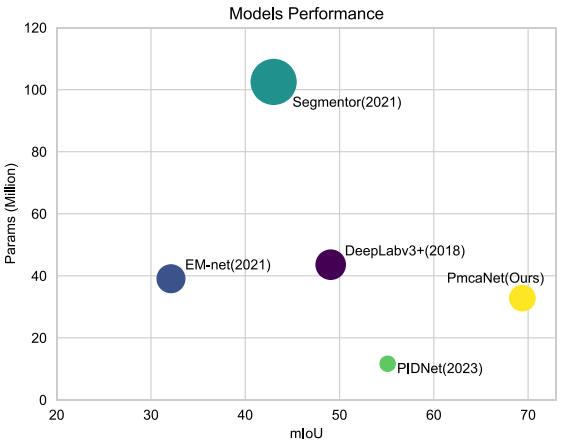

Fig. 7. Comparison of model performance and number of parameters on SEM-material. PmcaNet achieves improved performance with minimal additional parameters.

PIDNet requires the fewest parameters, while segmentor [27] requires the greatest number. PmcaNet, with 11.65 million parameters, has a slightly higher count than PIDNet [43] but significantly fewer than segmentor [27] and SegViT [30], which utilizes a ViT-based encoder with 102 million parameters. Furthermore, PmcaNet has a comparable number of parameters to other convolutional neural network-based models, such as EM-net [11] and DeepLabv3+ [22]. Despite this, it should be noted that PmcaNet does not necessitate significantly more computational resources during a single forward pass when compared to other methods. This allows PmcaNet to achieve superior performance without incurring a significant increase in the parameter count.

To assess the training progress of all models, the mean intersection over union (mIoU) and loss metrics on a validation set are depicted graphically in Fig. 8. Notably, the validation results indicate that PmcaNet exhibited the lowest validation loss and the highest mIoU score, thus emphasizing its superior performance and faster convergence speed. These advantages enable PmcaNet to deliver outstanding results in a shorter training time compared to more complex models.

5.3. *Superiority of PmcaNet compared to existing techniques*

In this paper, the problem of challenging the segmentation of electron microscopy (EM) images due to inadequate contrast and grayscale approximation is examined. To enhance the precision of EM image segmentation and address the issue of insufficient data sets, a novel multi-scale channel attention pyramid is employed. Diverging from other techniques, this structure comprehensively extracts local and global information from EM images by capturing attention values at various scales. This approach is pivotal for representing intricate nonlinear characteristics within the images. Additionally, an electron microscopy (EM) image data set of high-temperature

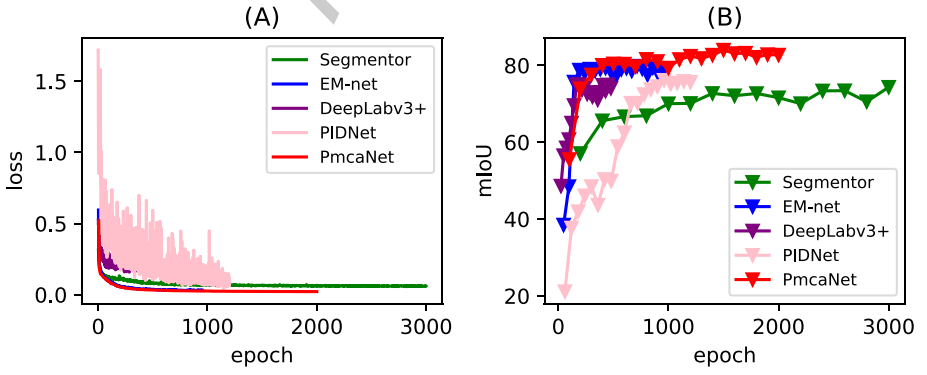

Fig. 8. An illustration of the tendency of the loss (A) and mIoU (B) for different methods during training for mitochondria segmentation on the Kathuri dataset.

alloy materials, known as SEM-material, is presented. The provision of more datasets is imperative for the advancement of the field.

6. Conclusion

This paper introduces PmcaNet, a novel model aimed at enhancing the utilization of contextual information across multiple scales in low-contrast electron microscopy images. The key contribution of this research lies in the proposal of a multiscale channel attention pyramid, which effectively integrates semantic context from various scales. Furthermore, a lightweight adaptive channel attention module is introduced to capture channel dependencies and enhance the representation capacity of features. Ablation studies were carried out to determine the effectiveness of the proposed modules, which were designed to address the challenge of identifying approximate grayscale pixels under low contrast conditions in electron microscopy image segmentation. These advancements have significantly reduced the difficulty of this task by improving the use of multiscale information. According to the results, it is evident that PmcaNet produces promising results for electron microscopy image segmentation.

Additionally, the issue of lacking available data for electron microscopy (EM) image segmentation is addressed in this study by introducing a novel dataset named SEM-material, which may contribute to alleviating data scarcity in the EM image segmentation domain. In the future, to achieve better performance and further reduce computing complexity, more efficient self-attention methods will be employed.

Acknowledgement

The research was supported by the National Science Foundation of China (No. U22B2048).

References

[1] Y. Ju, S. Li, X. Yuan, L. Cui, A. Godfrey, Y. Yan, Z. Cheng, X. Zhong and J. Zhu, A macro-nano-atomic–scale high-throughput approach for material research, *Science Advances* **7**(49) (2021), eabj8804.

[2] F. Yuan, Z. Zhang and Z. Fang, An effective CNN and transformer complementary network for medical image segmentation, *Pattern Recognition* **136** (2023), 109228.

[3] H. Jiang, R. Zhang, Y. Zhou, Y. Wang and H. Chen, DoNet: Deep De-overlapping Network for Cytology Instance Segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15641–15650.

[4] S. Li and W. He, Scanning electron microscope objective lens system and method for specimen observation, *US Patent* **11** (2021), 075,056.

[5] F. Lateef and Y. Ruichek, Survey on semantic segmentation using deep learning techniques, *Neurocomputing* **338** (2019), 321–348.

[6] C. Sommer, C. Straehle, U. Koethe and F.A. Hamprecht, Ilastik: Interactive learning and segmentation toolkit, in: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 230–233.

[7] A. Cardona, S. Saalfeld, J. Schindelin, I. Arganda-Carreras, S. Preibisch, M. Longair, P. Tomancak, V. Hartenstein and R.J. Douglas, TrakEM2 software for neural circuit reconstruction, *PloS One* **7**(6) (2012), e38011.

[8] I. Belevich, M. Joensuu, D. Kumar, H. Vihinen and E. Jokitalo, Microscopy image browser: A platform for segmentation and analysis of multidimensional datasets, *PLoS Biology* **14**(1) (2016), e1002340.

[9] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[10] J. Long, E. Shelhamer and T. Darrell, Fully convolutional networks for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[11] A. Khadangi, T. Boudier and V. Rajagopal, EM-net: Deep learning for electron microscopy image segmentation, in: *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 31–38.

[12] Y. Wu and S. Misra, Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix, *IEEE Geoscience and Remote Sensing Letters* **17**(7) (2019), 1144–1147.

[13] M. Helmstaedter, K.L. Briggman, S.C. Turaga, V. Jain, H.S. Seung and W. Denk, Connectomic reconstruction of the inner plexiform layer in the mouse retina, *Nature* **500**(7461) (2013), 168–174.

[14] I. Arganda-Carreras, S.C. Turaga, D.R. Berger, D. Cireşan, A. Giusti, L.M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J.M. Buhmann, et al., Crowdsourcing the creation of image segmentation algorithms for connectomics, *Frontiers in Neuroanatomy* **9** (2015), 142.

[15] B. Glancy, L.M. Hartnell, D. Malide, Z.-X. Yu, C.A. Combs, P.S. Connelly, S. Subramaniam and R.S. Balaban, Mitochondrial reticulum for cellular energy distribution in muscle, *Nature* **523**(7562) (2015), 617–620.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, *arXiv preprint arXiv:1412.7062* (2014).

[17] C. Farabet, C. Couprie, L. Najman and Y. LeCun, Learning hierarchical features for scene labeling, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8) (2012), 1915–1929.

[18] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou and P.E. Barbano, Toward automatic phenotyping of developing embryos from videos, *IEEE Transactions on Image Processing* **14**(9) (2005), 1360–1371.

[19] P. Pinheiro and R. Collobert, Recurrent convolutional neural networks for scene labeling, in: *International Conference on Machine Learning*, 2014, pp. 82–90.

[20] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng and S.-M. Hu, Segnext: Rethinking convolutional attention design for semantic segmentation, *arXiv preprint arXiv:2209.08575* (2022).

[21] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[23] J. Jain, J. Li, M. Chiu, A. Hassan, N. Orlov and H. Shi, One-Former: One Transformer To Rule Universal Image Segmentation, in: *Proceeings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[24] J. Hu, L. Shen and G. Sun, Squeeze-and-excitation networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[25] W. Qilong, W. Banggu, Z. Pengfei, L. Peihua, Z. Wangmeng and H. Qinghua, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11534–11542.

[26] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi and A. Agrawal, Context encoding for semantic segmentation, in: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[27] R. Strudel, R. Garcia, I. Laptev and C. Schmid, Segmenter: Transformer for semantic segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[28] Z. Zhu, M. Xu, S. Bai, T. Huang and X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 593–602.

[29] T.-D. Truong, N. Le, B. Raj, J. Cothren and K. Luu, Fredom: Fairness domain adaptation approach to semantic scene understanding, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19988–19997.

[30] B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei and C. Shen, SegViT: Semantic segmentation with plain vision transformers, *Advances in Neural Information Processing Systems* **35** (2022), 4971–4982.

[31] P.F. Felzenszwalb, R.B. Girshick, D. McAllester and D. Ramanan, Object detection with discriminatively trained part-based models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(9) (2009), 1627–1645.

[32] J. Carreira and C. Sminchisescu, Constrained parametric mincuts for automatic object segmentation, in: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3241–3248.

[33] Y. Yang, S. Hallman, D. Ramanan and C.C. Fowlkes, Layered object models for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(9) (2011), 1731–1743.

[34] A. Bihani, H. Daigle, J.E. Santos, C. Landry, M. Prodanović and K. Milliken, Mudrocknet: Semantic segmentation of mudrock sem images through deep learning, *Computers & Geosciences* **158** (2022), 104952.

[35] Q. Yu, Z. Xiong, C. Du, Z. Dai, M.R. Soltanian, M. Soltanian, S. Yin, W. Liu, C. Liu and C. Wang, Identification of rock pore structures and permeabilities using electron microscopy experiments and deep learning interpretations, *Fuel* **268** (2020), 117416.

[36] I. Oztel, G. Yolcu, I. Ersoy, T. White and F. Bunyak, Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network, in: *2017 IEEE International Conference on Bioinformatics and Biomedicine*, 2017, pp. 1195–1200.

[37] B. Glancy, Mitonet: A generalizable model for segmentation of individual mitochondria within electron microscopy datasets, *Cell Systems* **14**(1) (2023), 7–8.

[38] W. Wang, S. Wang, Y. Li and Y. Jin, Adaptive multi-scale dual attention network for semantic segmentation, *Neurocomputing* **460** (2021), 39–49.

[39] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu and Y. Pan, Rethinking dice loss for medical image segmentation, in: *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 851–860.

[40] Y. Ho and S. Wookey, The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling, *IEEE Access* **8**(8) (2020), 2169–3536.

[41] M. Contributors, MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, (2020).

[42] H. Robbins and S. Monro, A stochastic approximation method, *The Annals of Mathematical Statistics* (1951), 400–407.

[43] J. Xu, Z. Xiong and S.P. Bhattacharyya, PIDNet: A real-time semantic segmentation network inspired by pid controllers, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19529–19539.