Pattern Recognition 149 (2024) 110260

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-scale hypergraph-based feature alignment network for cell localization

Bo Li, Yong Zhang [*], Chengyang Zhang, Xinglin Piao, Yongli Hu, Baocai Yin

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

ARTICLE INFO

Dataset link: GitHub (https://github.com/Boli-trainee/MHFAN)

Keywords:
Cell localization
Feature alignment
Hypergraph neural network
Multi-scale hypergraph attention
Stepwise adaptive fusion

ABSTRACT

Cell localization in medical image analysis is a challenging task due to the significant variation in cell shape, size and color. Existing localization methods continue to tackle these challenges separately, frequently facing complications where these difficulties intersect and adversely impact model performance. In this paper, these challenges are first reframed as issues of feature misalignment between cell images and location maps, which are then collectively addressed. Specifically, we propose a feature alignment model based on a multi-scale hypergraph attention network. The model considers local regions in the feature map as nodes and utilizes a learnable similarity metric to construct hypergraphs at various scales. We then utilize a hypergraph convolutional network to aggregate the features associated with the nodes and achieve feature alignment between the cell images and location maps. Furthermore, we introduce a stepwise adaptive fusion module to fuse features at different levels effectively and adaptively. The comprehensive experimental results demonstrate the effectiveness of our proposed multi-scale hypergraph attention module in addressing the issue of feature misalignment, and our model achieves state-of-the-art performance across various cell localization datasets.

1. Introduction

The primary objective of cell localization is to precisely determine the position of each cell in an image. This task holds significant applications in the medical domain, including embryo counting for infertility treatment and malignant cell detection for cancer diagnosis. Moreover, precise cell localization is a fundamental initial step for further medical image analysis, such as cell segmentation [1] and the identification of subcellular localization of protein signals [2]. As depicted in Fig. 1(a), the considerable variability in cell shape, size, and color poses a major challenge to the cell localization task. To overcome these challenges, existing methods [3,4] typically rely on the location maps paradigm for cell localization. In this paradigm, the localization network predicts a location map, which is then post-processed to derive the number and spatial coordinates of the cells. The location map, also known as the Ground Truth (GT), is illustrated in Fig. 1(b).

In order to tackle the challenge of disparities in cell size and shape, Tofighi et al. [3] propose a Tunable Shape Prior Convolutional Neural Network (TSP-CNN). This model incorporates shape priors, which are customized to match the intricate and diverse cell shapes in images. By including a trainable and optimized shape prior layer, this model significantly improves cell localization performance. However, the fixed shape priors of this approach restrict its applicability to other scenarios. To address the issue of significant variations in cell color, Li et al. [4] propose a multi-scale difference convolution module. The module applies difference convolution [5] to cell localization, using difference operations to mitigate pixel differences between cells of varying color depths. This approach enhances the model's robustness to cell color in images. However, difference convolution amplifies the edge information of cells in the feature map, which to some extent exacerbates the interference of cell shape to the model.

Existing cell localization methods are still limited to independently addressing the significant variation in cell shape, size, and color, leading to multiple problems that interfere with each other and severely limit performance. To address this issue, we first propose defining the problem uniformly as the challenge of feature misalignment between cell images and GT.

Specifically, Fig. 1(a) displays a pathological image with the corresponding GT shown in Fig. 1(b). The pixel distributions of the color-marked cells (Fig. 1(c)) and GT (Fig. 1(d)) are shown in the subfigure below, with the horizontal axis representing the distance from the center point of the cell and the vertical axis representing the relative pixel values. The three curves in Fig. 1(c) illustrate the feature distributions of three cells corresponding to the color-marked points in Fig. 1(a), where the starting point on the left side of the curve corresponds to the cell center, and the feature distribution extending along the positive X-axis from the cell center is depicted as shown in the curve. For example, the center of the red cell is relatively dark in Fig. 1(a), resulting in a relatively small pixel value at the beginning of the red curve in

[*] Corresponding author.
E-mail addresses: bo_li@emails.bjut.edu.cn (B. Li), zhangyong2010@bjut.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.patcog.2024.110260
Received 31 May 2023; Received in revised form 29 December 2023; Accepted 4 January 2024
Available online 8 January 2024

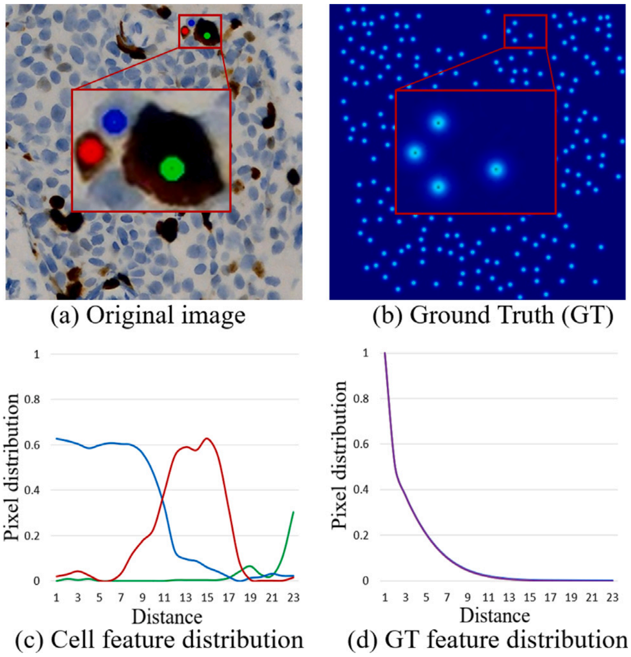

Fig. 1. Mapping cell images to ground truth is challenging due to the problem of feature misalignment. This is due to the complex and diverse pixel distribution caused by the drastic changes in cell size, shape and color, which conflicts with the uniform feature distribution in GT. The three curves in Fig. 1(c) illustrate the feature distributions of three cells corresponding to the color-marked points in Fig. 1(a), where the starting point on the left side of the curve corresponds to the cell center, and the feature distribution extending along the positive X-axis from the cell center is depicted as shown in the curve. In Fig. 1(a), we have labeled only three out of four cells in the enlarged area for the sake of brevity.

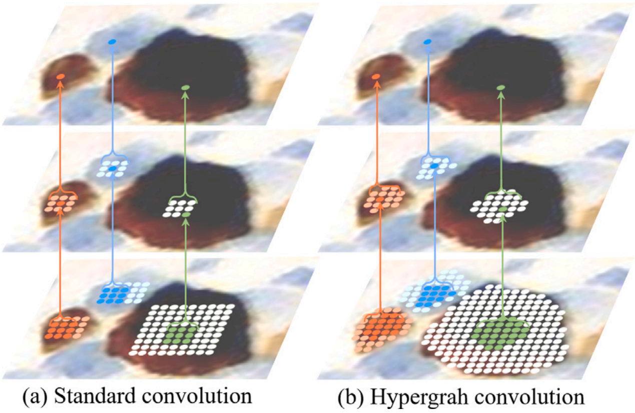

Fig. 2. Comparison of feature sampling process between standard convolution (a) and hypergraph convolution (b). Due to the design limitations of standard convolution, including the convolution kernel and the local receptive field, its ability to explore relationships between features is restricted, making it challenging to effectively aggregate features from different spatial extents. In contrast, hypergraph convolution can adaptively aggregate features from different ranges by utilizing hyperedges.

Fig. 1(c). The pixel values increase significantly when the distance from the center point exceeds the cell range and reaches the white background, but then decrease again as the curve reaches the edge of the black cell at a distance of about 20 pixels. However, the pixel distribution corresponding to the GT in Fig. 1(d) uniformly decreases from the center of each cell, leading to a feature misalignment issue.

Existing CNN-based methods are constrained in addressing the feature misalignment problem due to the convolutional kernel design, which limits the ability to capture correlations between different regions. Standard convolution presents two obvious drawbacks: (1) the use of convolution kernels for uniform mapping of local regions hinders the exploration of feature correlation; (2) the design of local receptive fields limits the feature aggregation capability. Fig. 2(a) illustrates the limitations of standard convolution in dealing with complex-shaped objects, where the use of fixed-shaped convolution kernels results in either insufficient or excessive convolution. To overcome these limitations, researchers proposed deformable convolution [6,7], which allows convolution kernels to deform within a certain range to adapt to more flexible target features. However, deformable convolution introduces high computational complexity due to the added deformation networks and parameters, which significantly increases the model's computational requirements and parameter count. Additionally, the deformation network and parameters are susceptible to the input image's deformation, leading to inaccurate convolutional kernel sampling positions and suboptimal model performance for distorted or deformed input images.

Ideally, it is desirable for the model to continually aggregate the features surrounding the cell towards its central point, without resorting to learning supplementary parameters to achieve feature alignment between the image and GT. Remarkably, we discover that the above-mentioned solution approach bears a striking resemblance to the node aggregation characteristics observed in graph neural networks [8,9]. However, owing to the pairwise connections only existing between nodes in graph, it is challenging to model the wide-ranging features surrounding the cell. Therefore, we introduce hypergraph neural networks [10–12] to the cell localization task for the first time. As shown in Fig. 2(b), the hypergraph convolution adaptively aggregates features to cell centroids by continuously learning to aggregate of neighboring nodes, which in turn enables feature alignment with the GT. In addition, given that traditional similarity measures based on Euclidean distance are difficult to distinguish cells and their similar backgrounds, this paper designs a multi-metric approach to construct the attention matrix. Further, since varying sizes of localization targets lead to different ranges of features to be aggregated, we design a multi-scale hypergraph to enable the model to adaptively aggregate features at different scales, thus achieving feature alignment.

The objective of this paper is to address the significant variations in cell shape, size, and color encountered in cell localization tasks. Firstly, we reframe these challenges as a feature misalignment issue between cell images and location maps. Secondly, we introduce a novel feature alignment model that effectively tackles this problem in cell localization. To this end, we propose a multi-scale hypergraph attention module that captures features of neighboring nodes at different scales. This method allows for the alignment of features between cell images and location maps, resulting in enhanced accuracy of cell localization. Moreover, we introduce a stepwise adaptive fusion module that efficiently reassigns weights to features for optimal fusion. Adequate experimental results show that our model achieves state-of-the-art performance on a variety of cell localization datasets.

In brief, this paper makes the following contributions:

- This paper innovatively addresses the challenges stemming from significant variations in cell shape, scale, and color by reframing them as a feature misalignment problem between cell images and location maps, thereby presenting a unified solution to these complexities.
- We propose an innovative multi-scale hypergraph attention module that achieves feature alignment through the adaptive aggregation of features across various scale ranges.
- The proposed model achieves state-of-the-art performance on multiple cell localization datasets and reveals great potential.

The rest of paper is structured as follows: Section 2 provides a review of related works, while Section 3 presents the proposed method in detail. In Section 4, the experimental results and analysis are presented.

Section 5 discusses the hypergraph module, and Section 6 presents the conclusion and future outlook.

2. Related works

This section presents a comprehensive overview of related works that are pertinent to the contributions of this paper. Initially, we introduce existing works that aim to tackle the issue of feature misalignment in cell localization and counting tasks, including problems with cell shape, scale, and color. Subsequently, we discuss feature alignment approaches that are frequently utilized in medical image analysis. Finally, we introduce relevant works on hypergraph neural networks.

2.1. Feature alignment in cell localization

To overcome the challenges posed by significant variations in cell shapes, sizes, and colors typically observed in pathological images, researchers have invested considerable effort in developing novel approaches. For instance, Tofighi et al. [3] have proposed a convolutional neural network model called TSP-CNN to address the issue of diverse cell shapes and sizes. This model incorporates shape priors, with the assistance of domain experts, via a trainable shape prior layer that aims to closely match the complex and varied shapes of cells in cell images. Meanwhile, Chen et al. [13] have addressed the issue of significant color variations in cells by defining a direction field for each pixel and setting the direction field of background pixels to zero. This improves the model's ability to accurately recognize light-colored cells. Moreover, Li et al. [4] have fully exploited gradient information at different scales by aggregating multi-scale difference convolutions [5], thereby enhancing the model's robustness to color. To distinguish foreground and background more effectively, Guo et al. [14] have introduced a self-attentive module based on the U-Net [15] network to improve localization and counting performance. In addition, Li et al. [16] designed a lightweight and efficient cell localization network, Lite-UNet, which improves performance while saving computational cost.

The aforementioned works have significantly improved the accuracy of cell localization. Nevertheless, such works are often customized to address specific challenges, which may restrict their applicability in a more general context. Moreover, the coexistence of multiple challenges may exacerbate the issue by interfering with each other, leading to further declines in localization performance.

2.2. Feature alignment for other fields in medical image analysis

In the field of medical image analysis, feature alignment strategies are typically based on deformable convolution. For instance, Huang et al. [17] have introduced a feature alignment module based on deformable convolution [6,7] to align features from different levels in FPN structures. Similarly, Xie et al. [18] have employed feature alignment strategies in medical image segmentation, utilizing deformable convolution layers to address ambiguous semantic information. Deformable convolution is also utilized in DefED-Net, a deformable encoder–decoder network proposed by Tao et al. [19] for liver and liver tumor segmentation. In the domain of cell detection, Li et al. [20] have extended the Faster R-CNN [21] model by incorporating deformable convolution into the FPN structure to detect cervical cancer cells in whole slide images of Pap smear specimens. Despite the effectiveness of deformable convolution in various image analysis tasks, they are still limited by inherent problems, such as high computational costs and the challenge of determining optimal parameters.

These works utilizing deformable convolution described above have largely addressed the issue of feature misalignment. However, it is associated with two primary drawbacks. Firstly, additional deformation networks and parameters are introduced, leading to a significant increase in the model's computational complexity and parameter count. Secondly, deformable convolutions can be sensitive to image deformation, resulting in inaccurate sampling positions of convolution kernels and ultimately impairing the model's performance.

2.3. Hypergraph neural network

Convolutional Neural Networks (CNNs) have been widely used in various fields, but their applicability is limited when applied to non-grid structured data. To overcome this limitation, Graph Neural Networks [8,22] (GNNs) have been introduced to perform message passing and pooling operations on nodes for processing non-grid structured data. However, GNNs struggle with high-order feature relationships. Recently, the HyperGraph Neural Network [10,23,24] (HGNN) has emerged as a promising alternative to GNNs, taking advantage of the hypergraph structure to handle high-order relationships. Hypergraphs are capable of connecting multiple nodes to a group of hyperedges, allowing for the representation of complex relationships between multiple nodes simultaneously. Currently, hypergraphs are better used in several fields, including segmentation [25], heterogeneous data [26] and re-identification [27].

To the best of our knowledge, HGNN has not been applied in object localization tasks, encompassing cell localization, crowd localization, and vehicle localization. It is worth noting that in such tasks, the location map often exhibits a pixel distribution that extends outward from the center, which aligns well with the node aggregation feature of HGNNs.

3. Our method

The Multi-scale Hypergraph-based Feature Alignment Network (MHFAN) consists of three main modules, as illustrated in Fig. 3. These modules include: (i) the backbone network, which extracts multi-scale features from the input pathological image; (ii) the Multi-scale Hypergraph Attention (MHA) module, which adaptively aggregates features in the vicinity of nodes; and (iii) the Stepwise Adaptive Fusion (SAF) module, which adaptively fuses information from different feature levels. First, the pathological image is fed into the backbone network, which generates four feature maps of different levels. Then, the MHA module optimizes each feature map by adaptively aggregating feature information near nodes. Finally, the SAF module fuses features from various levels to create a fused feature map, which is used to generate the predicted location maps via up-sampling and convolution.

3.1. Backbone network

Cell localization task often involve small and lightly colored cells, thus requiring a backbone network that can provide fine-grained features. Several existing backbone networks provide multiple stages of features, including VGG-16 [28], ConvNeXt-Base [29], and HRNet-W48 [30]. VGG-16 employs consecutive 3x3 convolutional kernels to implement deep networks and thus improve performance. ConvNeXt-Base integrates various techniques from existing convolutional networks, such as large convolutional kernels, deep separable convolution, inverse bottleneck design, to achieve the highest accuracy at the time on ImageNet. HRNet-W48 maintains high-resolution features and constant interaction between features at different levels, making it well-suited for pixel-level prediction tasks. To demonstrate the generality of our method, we conducted experiments with these three backbone networks, which are detailed in the experimental section.

3.2. Multi-scale hypergraph attention module

After extracting multi-scale features from the HRNet backbone, we utilize the Multi-scale Hypergraph Attention (MHA) module to optimize each feature map. Specifically, this module partitions the image into multiple local regions, where each region acts as a node in a hypergraph. By employing a learnable similarity metric, we generate a correlation matrix that represents the correlation between the nodes. Next, we create multi-scale hyperedges based on this matrix, with each hyperedge capturing feature information at different scales. To enhance

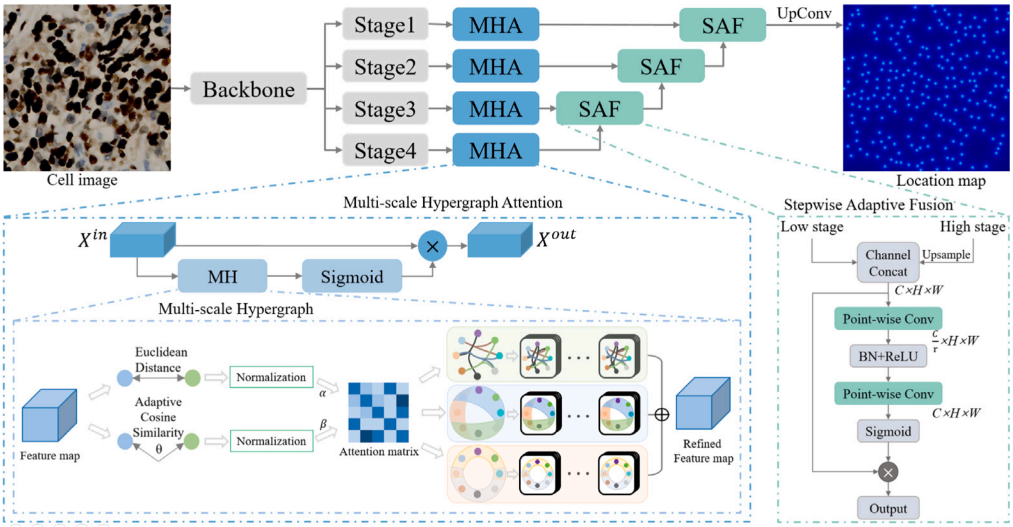

Fig. 3. The overall framework of the Multi-scale Hypergraph-based Feature Alignment Network (MHFAN) model. The model mainly include three modules: (i) the backbone network, which extracts multi-scale features from the input pathological image; (ii) the Multi-scale Hypergraph Attention (MHA) module, which adaptively aggregates features in the vicinity of nodes; and (iii) the Stepwise Adaptive Fusion (SAF) module, which adaptively fuses information from different feature levels.

the expression of node features, we use hypergraph convolution to adaptively aggregate features around the nodes at multiple scales. This results in a more accurate distribution of the cell centroid and enables feature alignment between the cell images and the location maps.

The feature map, denoted by $F \in \mathbb{R}^{C\times H\times W}$, is partitioned into $C \times 1 \times 1$ pixel blocks that serve as nodes. A correlation matrix is then constructed based on learnable similarity metrics, which forms the basis for constructing hyperedges in subsequent stages. Typically, a smaller Euclidean distance between node features in an image indicates a stronger association between the nodes. Therefore, we compute the similarity between two nodes $\mathcal{V}_i$ and $\mathcal{V}_j$ using the Euclidean distance. However, in certain cell localization scenarios, differentiating cells from similar backgrounds using only Euclidean distance can be challenging, especially when light-colored cells are indistinguishable from their surroundings, which are highly similar. To address this issue, we propose a new approach based on weighted cosine similarity. Unlike Euclidean distance, cosine similarity considers only the angle between vectors and is not affected by the absolute size of feature vectors, making it more suitable for high-dimensional feature spaces. Moreover, cosine similarity is more sensitive to measuring similar features, which can alleviate confusion between certain cells and their similar backgrounds. To improve the model's ability to capture the interplay between cells and their surrounding background, we build upon previous work [31,32] by allowing the model to learn the similarities between node features autonomously. To achieve this, we introduce learnable parameters, resulting in the cosine similarity between nodes being represented as a learnable quantity. The learnable cosine similarity is computed using the following equation:

$$Sim(f_i, f_j) = \frac{(f_i W_i)(f_j W_j)^T}{\|f_i W_i\|_2 \cdot \|f_j W_j\|_2}, \tag{1}$$

where W_i and W_j are learnable weights, f_i is feature vector of $\mathcal{V}_i$, and $\|\cdot\|_2$ denotes the L_2 norm.

In this study, we propose a method to measure the similarity between node features by combining the Euclidean distance M_{dis} and the learnable cosine similarity M_{sim}. However, we note that the discrepancy between M_{dis} and M_{sim} can be disproportionately large due to the varying similarity measures of node features, resulting in the loss of the weighting effect on one or both sides. To address this issue, we normalize M_{dis} and M_{sim} using the Softmax function. Specifically, we apply Softmax normalization to each row of both matrices, converting them into probability distributions. This ensures that both matrices have the same value domain and the weighting effect of either matrix is preserved. Finally, we combine the normalized matrices to obtain the correlation matrix $Co_M at$ using the following formula:

$$Co\_Mat = \alpha \times M_{dis} + \beta \times M_{sim}, \tag{2}$$

where α and β are hyperparameters used to adjust the ratio of the distance matrix and the similarity matrix, which would be discussed in Section 5.1.

Algorithm 1 Hypergraph Construction

```
Input: Embedding $X$
Function: Construct multi-scale hypergraph incidence matrix $H$ based on Euclidean distance and learnable cosine similarity.
for F in $X$ do
    $M_{tmp}$ = Eu_dis(F)
    $M_{dis}$.append($M_{tmp}$)
end for
Generate $M_{sim}$ according to Eq. (1)
$M_{dis}$ = Softmax($M_{dis}$)
$M_{sim}$ = Softmax($M_{sim}$)
$Co\_Mat = \alpha \times M_{dis} + \beta \times M_{sim}$
for i in range(len($Co\_Mat$)) do
    $H_{tmp1}$ = Construct_H_with_KNN($Co\_Mat$[i], k=K1)
    $H_{tmp2}$ = Construct_H_with_KNN($Co\_Mat$[i], k=K2)
    $H_{tmp3}$ = Construct_H_with_KNN($Co\_Mat$[i], k=K3)
    $H1$.append($H_{tmp1}$)
    $H2$.append($H_{tmp2}$)
    $H3$.append($H_{tmp3}$)
end for
Output: Multi-scale hypergraph incidence matrix $H1$, $H2$, $H3$
```

After computing the correlation matrix Co_Mat, we propose a methodology for constructing a multi-scale hypergraph that captures feature relationships at varying scales. We determine the number of neighboring nodes and specific values based on some previous works [33,34] and sufficient experiments, as described in Section 5.1. Specifically, we construct hyperedges using the K1, K2, and K3 neighboring nodes and the corresponding correlation matrix Co_Mat. Algorithm 1 outlines the overall process of constructing the multi-scale hypergraph. Here, $Eu\_dis(F)$ refers to the Euclidean distance calculation between each node in the feature map F, which is first converted to $F \in \mathbb{R}^{HW \times C}$ for a stage of the backbone network. More specifically, $Eu\_dis(F)$ calculates the distance between each row in the matrix F, returning the matrix of $M_{tmp} \in \mathbb{R}^{HW \times HW}$, which represents the distance relationship between each node. Subsequently, we apply Softmax to each row of the matrices M_{dis} and M_{sim}, converting all values into probabilities and ensuring the two matrices have the same value domain without losing the weighting effect of either. Finally, the Construct_H_with_KNN function is used to identify the K neighboring nodes that are most adjacent to the current node and thus construct the super-edge, as detailed in Section 5.1.

We employ hypergraph convolutional structures to propagate features within hypergraphs of different scales. This approach combines the features of each node with those of its neighbors through hyperedge connections using a convolutional method, resulting in the generation of new node features and adaptive aggregation of multi-scale features. Ultimately, we observe that the features optimized by the hypergraph convolutional network tend to become overly smooth, leading to a significant loss of feature discriminability [35]. To mitigate this issue, we propose a novel approach where hypergraph output features are used as attention weights to fine-tune feature optimization. This technique improves both the interpretability and generalization performance of the model. Specifically, the original features are optimized using the hypergraph-optimized features as attention. The process is illustrated as follows:

$$X^{out} = Sigmoid(MH(X^{in})) \cdot X^{in}. \tag{3}$$

3.3. Stepwise adaptive fusion module

After the optimization with the MHA module, the features from the four stages must be aligned with the features located at different levels within the location map. To achieve the adaptive selection of the most relevant features across multiple hierarchical levels, we introduce a Stepwise Adaptive Fusion (SAF) module. This module is designed with a layer-by-layer approach, which allows for the effective selection of significant features and improves the overall performance of the model.

In our model, the SAF module is utilized thrice to merge features at various hierarchical levels. As depicted in the SAF module diagram in the lower right corner of Fig. 3, the SAF module receives two input feature maps from different levels. Firstly, we up-sample the higher-level feature map to the same resolution as the lower-level feature map. Notably, to retain the geometric information of the original image, we set the $align\_corners$ parameter to $True$ during the up-sampling process, aligning the pixel values of the original four corners with the up-sampled pixel values. Secondly, we concatenate the features from the two stages by channels and employ a channel transformation operation akin to the Squeeze-and-Excitation module to capture the significance of various channels. Subsequently, we multiply the sigmoid-activated features with the original concatenated features element-wise to adaptively weigh the channel information. The entire process is illustrated as follows:

$$X^{out} = X^{in} \cdot Sigmoid(SE([X^{Low}, Up(X^{High})])), \tag{4}$$

$$SE = W_2 \cdot ReLU(W_1 X), \tag{5}$$

where $Up(\cdot)$ refers to the up-sampling operation, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. Finally, we feed the output features X^{out} into the subsequent SAF module with the features from the next stage, achieving another round of adaptive feature fusion. After three rounds of feature fusion, we perform a straightforward up-sampling convolution operation, similar to the one used in U-Net [15], to resize the output features to the original image size.

4. Experiments and analysis

4.1. Datasets and experimental details

In the field of cell localization and counting, the BCData [36] and PSU [3] datasets, consisting of 1338 and 120 images respectively, are currently the primary public datasets. To further validate the effectiveness of our proposed method, we have transformed the cell instance segmentation datasets [37,38] into cell localization datasets, named ccRCC Grading and CoNIC, respectively. In the subsequent sections, we provide a comprehensive description of all the datasets.

The **BCData** [36] is a large-scale dataset for localizing Ki-67 cells in breast tumor cells. It comprises 1,338 images with a resolution of 640 × 640 and 181,074 annotated tumor cells. The tumor cells exhibit diversity in size and shape, and the staining depth of the cells varies widely.

The **ccRCC Grading** [37] dataset is originally created for grading the nuclei of ccRCC cells. It includes 1000 H&E stained image patches with a resolution of 512 × 512, containing a total of 70,945 labeled nuclei, each with an instance segmentation mask and a classification mask. In order to repurpose this dataset for cell localization, we utilized a segmentation map-based union domain analysis algorithm to derive the centroids of the cells. The resulting dataset, which we call ccRCC Grading, can be used for cell localization tasks.

The **CoNIC** [38,39] dataset is currently the largest dataset used for nucleus segmentation in histopathological images stained with Hematoxylin and Eosin (HE). This dataset comprises 4981 samples, each with a resolution of 256 × 256 pixels, and each sample has two types of annotations: nucleus segmentation and classification. It stands as one of the most extensive datasets available, containing approximately 500,000 labeled cell nuclei. Consistent with the processing of the ccRCC Grading [37] dataset, to adapt this dataset for cell localization tasks, we first transformed it into a localization dataset. Subsequently, we split the dataset into a training set, consisting of samples from [0, 4000], and a test set, consisting of samples from (4000, 4981], for validation purposes.

The **PSU** [3] dataset is comprised of 120 images of colonic tissue from 12 pigs, with a resolution of 612 × 452 and 25,462 annotated cells. The images in this dataset are generally darker in tone. We utilized the first 90 images in the dataset for training, and the remaining 30 images for validation purposes.

The images are uniformly resized to a resolution of 512 × 512 during both the training and testing phases, since the image sizes in the aforementioned datasets do not vary significantly. The training process is configured as follows: the loss function is selected as mean squared error, the network parameters are optimized using Adam [40] as the optimizer with a learning rate of 1e−4, a batch size of 6 is set, and horizontal flip is employed for data augmentation. The experiments are conducted on Ubuntu 18.04 with an NVIDIA Tesla P100 (∼16 GB). Lastly, the experimental code and processed datasets will be publicly available at GitHub (https://github.com/Boli-trainee/MHFAN).

4.2. Evaluation criteria

The objective of this paper is cell localization combined with cell counting. Therefore, the evaluation criteria encompass both localization and counting criteria.

Localization criteria: Following previous works [4,16], we utilize F1-score, precision, and recall to assess the localization performance of the model. Among these, F1-score serves as a comprehensive metric that balances both precision and recall, standing out as a pivotal indicator for comparison in this paper. A successful localization match is determined when the distance between a predicted point and its true counterpart is below a threshold σ. The model's performance is evaluated using a fixed threshold at two levels (σ = 5, 10), where a smaller threshold value signifies a more stringent localization accuracy.

Table 1
Quantitative comparison of localization and counting performance of different models on the BCData test set. The best performance in this table is marked in **bold**, the second performance is marked with an underline. The arrow ↓ indicates that the larger the indicator the better the performance, and the arrow ↑ indicates that the smaller the indicator the better the performance. F1 scores are percentages, and the percent sign is omitted in this paper. The same markings are used for all subsequent tables.

Methods	Counting MAE/RMSE↓	Localization(5) F1/Pre/Rec(%) ↑	Localization(10) F1/Pre/Rec(%) ↑
UNext	20.4/27.0	68.9/68.8/69.0	82.4/82.3/82.5
U_CSRNet	18.1/23.8	73.8/73.7/74.0	85.6/85.4/85.7
MPViT	22.3/29.2	75.3/79.3/71.6	85.5/90.1/81.4
U-Net	24.9/33.4	76.7/81.7/72.1	85.7/80.7/**91.4**
Lite-UNet	18.1/24.3	76.5/77.0/76.1	86.3/86.3/86.4
Attention U-Net	19.6/24.9	77.1/74.9/79.5	86.5/84.0/89.1
TransUNet	17.7/23.3	77.3/76.5/78.1	86.9/86.1/87.8
Swin Transformer	17.1/23.1	78.1/77.8/78.4	87.1/86.7/87.4
VGG16	17.7/23.2	79.2/78.6/79.8	86.9/86.3/87.5
HRNet	18.5/24.7	79.2/80.1/78.3	87.3/88.4/86.3
CMUNeXt	16.6/22.9	77.1/77.1/77.2	86.2/86.3/86.2
CMU-Net	19.7/25.1	78.0/75.5/80.7	86.5/83.7/89.5
ConvNeXt-Base	18.9/26.0	79.3/**82.1**/76.7	87.4/**90.5**/84.6
M_HRNet	19.9/25.6	79.3/80.2/78.3	87.2/88.3/86.2
DAE-Former	17.9/24.0	79.6/79.0/80.2	88.1/87.4/88.8
Ours(VGG16)	17.1/22.7	80.0/80.1/79.9	87.7/87.9/87.6
Ours(ConvNeXt)	17.2/23.0	80.2/79.8/80.5	88.2/89.0/87.4
Ours(HRNet)	**16.2**/**21.2**	**81.8**/**82.1**/**81.5**	**88.5**/88.9/88.1

Table 2
Quantitative comparison of localization and counting performance of different models on the BCData val set.

Methods	Counting MAE/RMSE↓	Localization(5) F1/Pre/Rec(%) ↑	Localization(10) F1/Pre/Rec(%) ↑
UNext	22.8/28.8	67.5/66.4/68.5	81.9/80.6/83.2
U_CSRNet	19.8/25.3	75.7/74.3/77.0	86.8/85.3/88.3
MPViT	21.5/27.7	74.4/76.5/72.4	85.8/88.3/83.5
U-Net	22.3/28.3	75.6/73.1/78.2	86.3/83.4/89.3
Attention U-Net	23.4/28.5	75.1/71.9/78.5	86.1/82.5/**90.0**
TransUNet	23.6/29.6	76.2/79.1/73.5	86.3/**89.6**/83.2
Swin transformer	19.9/25.8	76.1/74.7/77.5	86.8/85.2/88.3
VGG16	19.9/25.8	77.7/76.0/79.3	86.6/84.8/88.5
HRNet	22.2/27.8	77.0/76.5/77.5	86.5/86.0/87.1
CMUNeXt	19.7/25.1	75.3/75.7/74.8	84.6/85.1/84.0
CMU-Net	20.5/16.2	76.4/76.3/76.4	85.7/85.6/85.8
ConvNeXt-Base	20.6/27.2	77.6/78.7/76.5	87.2/88.5/86.1
M_HRNet	20.7/26.5	79.6/**79.9**/79.4	88.0/88.3/87.7
DAE-Former	20.5/26.9	77.9/76.7/79.1	87.6/86.2/88.9
Ours(VGG16)	19.0/24.5	78.1/77.2/79.0	87.2/86.3/88.2
Ours(ConvNeXt)	19.3/24.9	78.0/76.7/79.3	87.5/86.1/88.9
Ours(HRNet)	**18.1**/**22.3**	**80.0**/79.0/**81.1**	**88.3**/87.1/89.5

Counting criteria: Following previous works [4,36], Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are utilized to evaluate the counting performance of the model. MAE is more sensitive to the magnitude of the prediction error because it takes into account the absolute value of the error. RMSE, on the other hand, is measured by calculating the root mean square of the squared difference between the predicted and true values, which means that it is more sensitive to larger errors. The smaller the value of both, the better the counting performance of the model.

4.3. Comparative experiments and analysis

We first evaluate the performance of our MHFAN model on the test and validation sets of the BCData dataset [36]. The quantitative comparison results are presented in Tables 1 and 2. It is evident from the results that our MHFAN model outperforms existing models in terms of both cell localization and counting performance. These models include UNext [41], U-CSRNet [36], MPViT [42], U-Net [15], Lite-UNet [16], Attention U-Net [43], TransUNet [44], Swin Transformer [45], Hover-Net [46], CMU-Net [47], CMUNeXt [48], M_HRNet [4], W-Net [49], and DAE-Former [50]. Since the size of the feature map output from some models in the comparison model does not meet the localization task specification, such as Swin Transformer, we upsampled its multi-level output according to the stitching method in HRNet, and then stitched them together to obtain the final output predicted map. It is worth mentioning that we conducted experiments to investigate the dependence of the proposed MHA module on different backbone networks. We built models based on three backbone networks: VGG16 [28], HRNet [30], and ConvNeXt-Base [29]. The primary reason for choosing the VGG16 and HRNet backbone is its adaptability and universality. It not only provides the four-stage features required as shown in Fig. 4 but is also widely adopted as the backbone network in many current works [51–53]. Additionally, ConvNeXt is currently a popular high-performance backbone. For comparison, we evaluated the performance of localization using only the backbone network. The results demonstrate that our modules consistently achieve significant performance improvements, regardless of the specific backbone network used.

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

Methods	Counting MAE/RMSE↓	Localization(5) F1/Pre/Rec(%) ↑	Localization(10) F1/Pre/Rec(%) ↑
UNext	6.1/7.5	80.7/78.9/82.4	88.4/86.5/90.5
U_CSRNet	7.5/8.8	81.8/78.5/85.5	88.2/84.6/**92.1**
MPViT	6.2/7.8	82.0/82.5/81.6	88.5/88.4/88.7
U-Net	5.8/7.9	83.7/86.7/80.8	89.4/92.6/86.4
Lite-UNet	5.2/7.4	84.4/85.8/83.1	89.6/91.3/86.9
Attention U-Net	**4.6**/**6.0**	83.4/83.6/83.2	89.7/89.8/89.5
TransUNet	5.0/7.0	83.6/84.0/83.3	90.0/90.3/89.6
Swin Transformer	6.0/7.9	82.4/80.2/84.6	89.4/87.0/91.9
VGG16	5.3/6.8	83.9/82.8/84.9	89.4/87.8/90.9
HRNet	4.9/6.9	85.6/87.3/83.9	90.4/92.3/88.7
CMUNeXt	5.2/7.5	82.6/81.4/83.8	88.7/89.2/88.2
CMU-Net	5.7/7.3	83.7/83.4/83.9	89.8/89.4/90.2
Hover-Net	5.7/7.8	83.9/84.2/83.6	89.0/88.3/89.8
ConvNeXt-Base	5.8/7.1	83.7/82.9/84.4	89.4/88.5/90.3
DAE-Former	5.8/7.5	83.6/84.5/82.6	88.9/90.1/87.7
W-Net	-/-	85.0/83.0/**88.0**	-/-/-
MHFAN(Ours)	4.7/6.4	**86.6**/**88.1**/85.1	**91.2**/**92.8**/89.6

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

Methods	Counting MAE/RMSE↓	Localization(5) F1/Pre/Rec(%) ↑	Localization(10) F1/Pre/Rec(%) ↑
UNext	25.1/33.0	71.6/76.9/66.9	77.6/83.5/72.6
U_CSRNet	17.8/24.7	72.6/73.7/71.5	78.7/80.0/77.5
MPViT	19.5/25.5	74.0/74.7/73.3	79.7/80.4/78.9
U-Net	20.0/24.9	77.7/81.9/73.9	82.9/87.4/78.8
Attention U-Net	**14.3**/**19.7**	79.0/79.9/78.1	83.9/84.9/82.9
TransUNet	20.0/27.5	74.1/76.8/71.7	80.5/83.3/77.8
Swin Transformer	24.1/30.8	76.3/81.4/71.8	80.8/86.2/76.0
HRNet	18.0/28.4	77.8/76.9/**78.7**	82.4/81.5/**83.4**
CMUNeXt	18.9/24.2	77.6/82.0/73.7	82.8/87.4/78.6
CMU-Net	19.2/26.1	78.5/80.4/76.6	83.3/86.4/80.2
Hover-Net	20.4/25.1	80.3/**85.2**/75.4	83.9/**89.4**/79.1
DAE-Former	20.1/25.3	77.4/81.9/73.4	83.2/87.9/78.9
M_HRNet	14.5/21.0	78.6/80.5/76.8	83.9/85.9/82.0
MHFAN(Ours)	15.5/20.7	**81.6**/84.9/78.5	**85.3**/88.7/82.0

Our model uses the multi-scale hypergraph module to adaptively aggregate the features of neighboring nodes in different ranges, making the distribution of features near the cell center more reasonable. Specifically, by utilizing point-spread-based location maps for supervision, the hypergraph facilitates continuous learning of feature aggregation, allowing for better alignment between cell images and corresponding location maps. As shown in Table 1, compared to the existing suboptimal model DAE-Former, our model(HRNet) improves the localization performance by 2.2% at a threshold of 5 and the counting performance

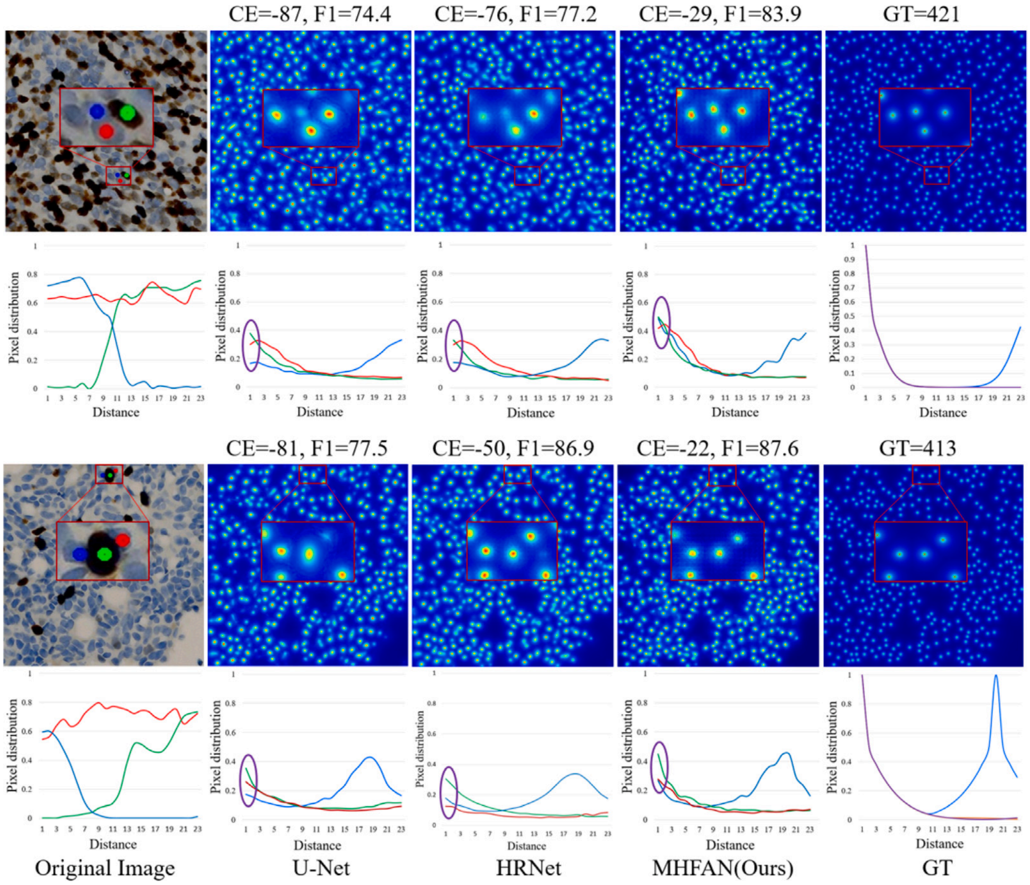

Fig. 4. Visual comparison of feature alignment capability among different localization models. Following the description in Fig. 1, this figure consists of four rows of images. The first and third rows represent the location map predictions of various models based on pathological images, while the second and fourth rows depict the feature distribution of three cells marked with corresponding colors from the first and third rows, originating from the cell center (the starting point of the curve) and extending along the positive X-axis. From the purple ellipses in the images, it can be observed that our MHFAN model significantly enhances the feature values at the cell centers (indicated by the higher relative position of MHFAN's ellipses compared to other models). This results in more rational localization maps, thereby enhancing the precision of model localization. Furthermore, as the MHFAN model here is built upon the HRNet backbone, additional ablation experiments in this figure demonstrate the effectiveness of our module. Additionally, "CE" denotes count error. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by 1.7 of MAE and 2.8 of RMSE. Further, experimental results also demonstrate that our model(HRNet) outperforms the HRNet baseline, achieving a localization performance improvement of 2.6% and 1.2% at thresholds of 5 and 10, respectively. Additionally, the MAE and RMSE in counting performance have improved by 2.3 and 3.5 than baseline, respectively.

Furthermore, we also conducted evaluations on ccRCC Grading, CoNIC and PSU datasets, with the results presented in Tables 3–5, respectively. In these several datasets, our approach has demonstrated significant advantages. Taking the largest dataset, CoNIC, as an example, compared to the second-best method, M_HRNet, our MHFAN achieved a remarkable increase in F1 scores by 3.0% and 1.4% at localization thresholds of 5 and 10, respectively. Compared to the baseline model, HRNet, the improvements were 3.8% and 2.9%, respectively. These results highlight the state-of-the-art performance achieved by our model across multiple datasets.

To clearly demonstrate the feature alignment capabilities of our MHFAN(HRNet-based) model, we visually compare the feature alignment effects in the location maps predicted by different models (including U-Net and HRNet), as depicted in Fig. 4. The figure consists of 4 rows and 5 columns, following a similar layout to Fig. 1. The first row presents the cell image alongside the corresponding location maps generated by different models. The second row showcases the feature distributions of three cell center points spreading along the positive x-axis within the image or location maps. In order to showcase the MHFAN model's robustness in capturing variations in cell shape, color, and size, we specifically select cells labeled in blue (small size), red (lighter color), and green (darker color and huge size). By distinguishing cells with different colors, we plot the pixel distributions of these cells in the second row. The third and fourth rows follow the same structure. From the purple ellipses in the images, it can be observed that our MHFAN model significantly enhances the feature values at the cell centers (indicated by the higher relative position of MHFAN's ellipses compared to other models). This results in more rational localization maps, thereby enhancing the precision of model localization. Furthermore, as the MHFAN model here is built upon

Table 5
Quantitative comparison of localization and counting performance of different models on the PSU dataset.

Methods	Counting		Localization	
	MAE↓	RMSE↓	F1(5)↑	F1(10)↑
UNext	43.2	52.1	59.2	78.1
U_CSRNet	32.1	38.2	54.5	80.4
MPViT	36.6	44.7	61.1	81.0
U-Net	32.6	38.5	61.0	81.0
Lite-UNet	33.7	38.4	60.2	79.6
Attention U-Net	31.4	36.8	63.5	82.1
TransUNet	40.1	49.4	58.9	80.1
Swin Transformer	**26.6**	**32.1**	62.2	82.1
HRNet	27.6	32.4	66.1	83.5
DAE-Former	28.2	33.4	66.3	84.3
MHFAN(Ours)	28.9	33.9	**70.7**	**86.4**

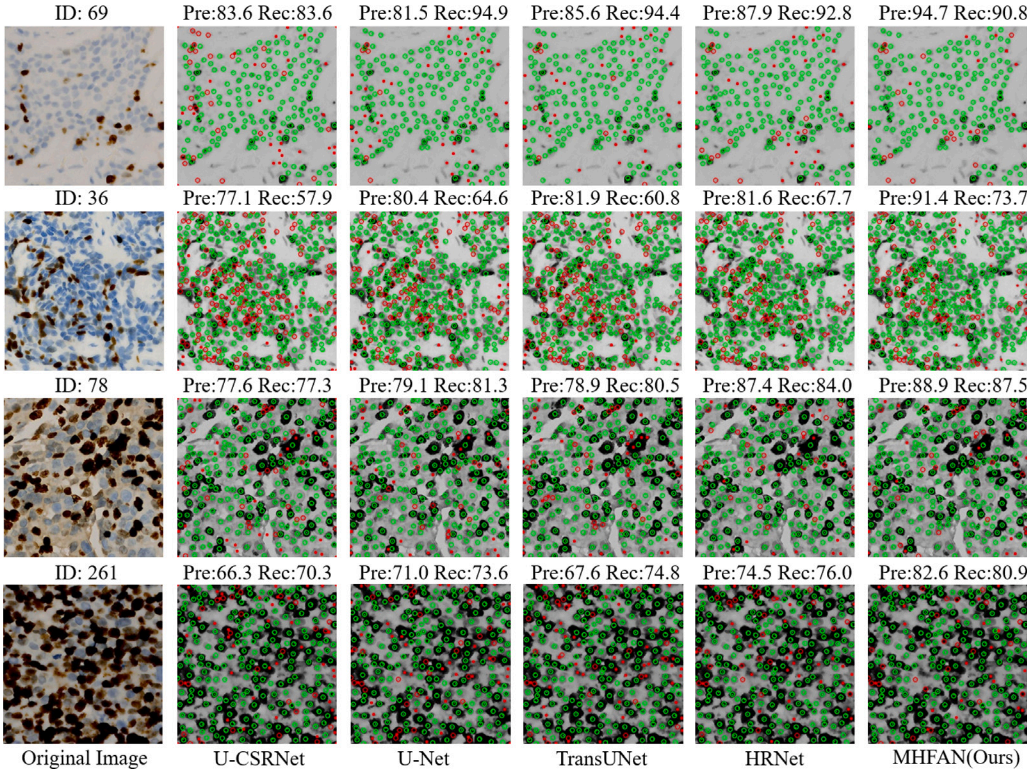

Fig. 5. Visualization of cell localization performance of multiple models on the BCData test dataset. Where the circles indicate Ground Truth and the dots indicate predicted results. Green indicates accurate prediction (TP) and red indicate wrong prediction (red circles for FN and red dots for FP). Therefore, the less red in the figure, the higher the localization performance. For easier reading, images are transformed to gray-scale data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the HRNet backbone, additional ablation experiments in this figure demonstrate the effectiveness of our module.

Lastly, for visual comparison of the localization results, we present the cell localization outcomes of multiple models on the BCData test dataset in Fig. 5. The figure exhibits four sets of representative cell images, encompassing significant variations in cell shape, size, and color. It is evident that our method consistently demonstrates outstanding localization performance across diverse cell scenarios.

4.4. Ablation experiments

To further confirm the contribution of each module and the corresponding hyperparameters, we conducted a large number of experiments based on the BCData dataset. Based on the HRNet backbone network, we first added the MHA module on the Stage 4 branch with the lowest resolution, and then added the Stage 3,2,1 branches respectively. Then, we conducted sufficient experiments on each combination to explore the pairing of different branches and modules to the maximum extent possible. Finally, the multi-stage features were fed into the SAF module for adaptive fusion. The step-by-step process of module ablation is presented in Table 6.

Based on the findings presented in Table 6, several significant conclusions can be drawn. Firstly, the incorporation of the hypergraph attention module results in a substantial enhancement in the model's localization performance. Specifically, when the MHA module is added solely to branch 4, there is a notable improvement of 1.9% (σ = 5) and 1.5% (σ = 10) in localization performance. Secondly, the introduction of additional branches further enhances both the localization and counting performance of the model. Lastly, in comparison to the simple stacking approach employed in HRNet, the SAF module effectively boosts the model's localization performance.

Table 6
The ablation experiments of the MHFAN model on BCData val dataset.

Backbone	Branches 4/3/2/1	SAF	Counting MAE/RMSE↓	Localization F1(5)/F1(10)↑
✓			22.2/27.8	77.0/86.5
✓	✓/✗/✗/✗		19.6/25.9	78.9/88.0
✓	✗/✓/✗/✗		20.6/26.3	79.3/87.9
✓	✗/✗/✓/✗		20.4/26.0	79.2/87.6
✓	✗/✗/✗/✓		20.7/27.0	79.3/87.7
✓	✓/✓/✗/✗		19.0/24.9	79.0/87.7
✓	✓/✗/✓/✗		21.0/26.6	79.2/87.7
✓	✓/✗/✗/✓		22.8/29.6	78.8/87.4
✓	✗/✓/✗/✓		20.1/25.2	78.5/87.8
✓	✗/✗/✓/✓		19.8/25.9	78.7/87.5
✓	✗/✓/✓/✓		19.3/25.5	79.3/87.9
✓	✓/✗/✓/✓		18.7/24.3	79.2/87.6
✓	✓/✓/✗/✓		18.4/23.1	79.0/**88.4**
✓	✓/✓/✓/✗		18.5/23.3	79.1/88.2
✓	✓/✓/✓/✓		18.2/23.1	79.2/88.2
✓	✓/✓/✓/✓	✓	**18.1/22.3**	**80.0**/88.3

5. Discussion on multi-scale hypergraph attention

This subsection aims to provide further analysis to better comprehend the contribution of the multi-scale hypergraph attention module. Firstly, we conduct experimental validation to assess the impact of the multi-metric approach and multi-scale design. Secondly, we compare the MHA module with well-known attention mechanisms such as Squeeze-and-Excitation (SE) [54], Convolutional Block Attention Module (CBAM) [55], and Global Attention Mechanism (GAM) [56]. It is worth noting that the MHA module in this paper serves as an attention weight module for feature optimization, similar to existing attention mechanisms. Lastly, we also include a comparison with deformable convolution, considering that most existing works [7,57] related to feature alignment rely on deformable convolution extensions.

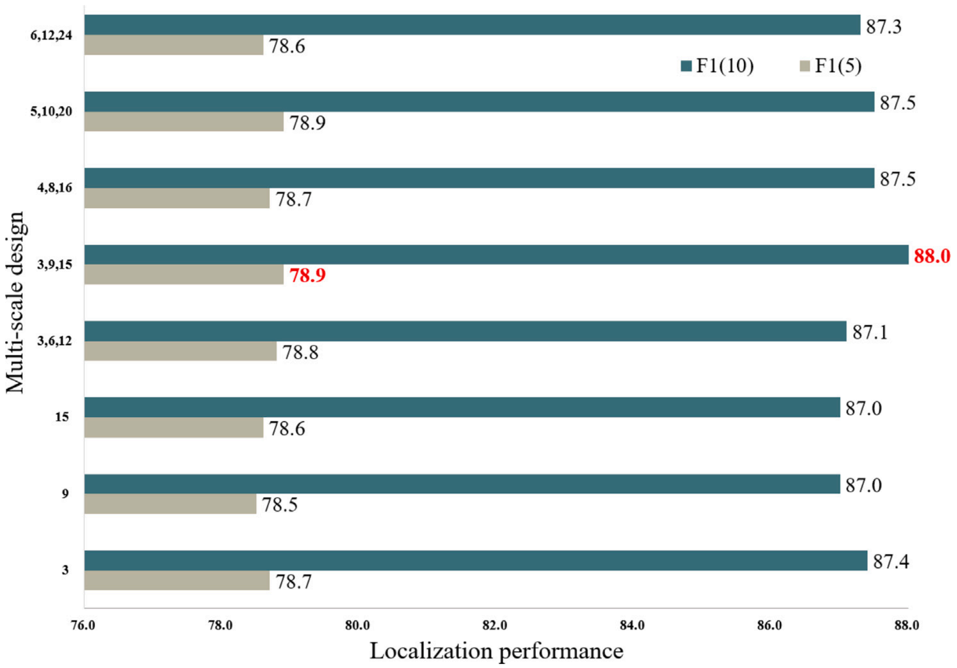

Fig. 6. Quantitative comparison of localization performance under different scale designs on the BCData val dataset. It can be observed that the model achieves its optimal localization performance when the multi-scale design is set to [3, 9, 15].

Table 7
Quantitative comparison of localization and counting performance of different metric methods on the BCData val dataset.

Methods	Counting MAE/RMSE↓	Localization F1(5)/F1(10)↑
Baseline	22.2/27.8	77.0/86.5
Dis($\alpha = 1, \beta = 0$)	22.0/28.3	77.7/86.6
Sim($\alpha = 0, \beta = 1$)	21.5/**27.5**	77.6/86.7
Dis+Sim × 0.5($\alpha = 1, \beta = 0.5$)	21.4/28.0	**78.7**/**87.4**
Dis × 0.5+Sim($\alpha = 0.5, \beta = 1$)	21.5/**27.5**	78.6/87.0
Dis+Sim($\alpha = 1, \beta = 1$)	**21.3**/27.7	78.1/86.9

Table 8
Quantitative comparison of the localization and counting performance of our MHA module, attention mechanism (GAM, CBAM, and SE), and deformable convolution (DC) on the BCData test dataset.

Methods	Counting		Localization	
	MAE↓	RMSE↓	F1(5)↑	F1(10)↑
Baseline	18.5	23.3	79.2	88.2
+ GAM	17.8	23.3	81.0	87.9
+ CBAM	17.7	23.1	80.7	87.6
+ SE	16.6	21.6	80.8	87.8
+ DC	17.2	22.8	81.5	88.4
+ MHA (Our)	**16.2**	**21.2**	**81.8**	**88.5**

5.1. Analysis of multi-metric and multi-scale

Multi-metric analysis: In this subsection, we delve into the multi-metric approach of the MHA module. Specifically, we conduct experimental analysis on the hyperparameters α and β in Eq. (2), and present the results in Table 7. The hyperparameters α and β determine the relative contribution of Euclidean distance and learnable cosine similarity to the correlation matrix. In order to investigate this, we experimented with five commonly used sets of ratios. The results demonstrate that the multi-metric approach effectively captures more correlations between features compared to the single metric approach, consequently enhancing both model counting and localization performance. Notably, when $\alpha = 1$ and $\beta = 0$, the classical hypergraph neural network is represented. Moreover, the performance is significantly influenced by the metric matrices with different weight ratios. Notably, the optimal performance is achieved when $\alpha = 1$ and $\beta = 0.5$, further affirming the efficacy of the multi-metric approach in improving performance.

Multi-scale analysis: This subsection presents an experimental analysis of the scale design, encompassing both single-scale and multi-scale approaches for different neighbor nodes, as illustrated in Fig. 6. Given the scale range of cells in pathological images, we set the K values ranging from 3 to 24. Each K value corresponds to a distinct correlation matrix H. The specific combinations of K values were determined based on some previous works [33,34] and sufficient experiments. By aggregating features from neighboring nodes at different scales, the model effectively captures information across various scales. The experimental results demonstrate that node aggregation at a single scale underperforms compared to node aggregation at multiple scales. Notably, the best performance is achieved when the scales are set to 3, 9, and 15, respectively.

5.2. Comparison with attention mechanism and deformable convolution

Referring to Section 3.2, we address the oversmoothing issue by utilizing the hypergraph output as attention weights to augment the original features. This strategy shares similarities with attention mechanisms frequently employed in visual tasks. Consequently, we compare attention modules with the proposed MHA module to further demonstrate the superiority of our approach. As discussed in Section 2.2, Deformable Convolution (DC) [7] is more commonly utilized in feature alignment tasks. Hence, we extend our comparison to include the DC module. The results are presented in Table 8.

Specifically, we compare our method with popular attention mechanisms, including SE [54], CBAM [55], and GAM [56]. Among these, SE focuses on channel attention, while CBAM and GAM incorporate both channel and spatial attention. Additionally, we include a comparison with DC [7], which are widely used in current feature alignment studies. By keeping all other parameters constant and solely replacing the MHA module with the aforementioned module, we evaluate and present the performance comparison in Table 8. It is evident that our MHA module surpasses attention mechanisms and DC in terms of cell localization and counting. Convolution-based attention mechanisms and DC struggle to capture the intrinsic connections among features and adaptively aggregate features around nodes for effective feature alignment.

5.3. Limitations of hypergraph

We discuss the limitations of our work from two perspectives. Firstly, the smoothed features after hypergraph optimization pose a challenge in direct utilization. In this paper, we address this issue by optimizing the original features using attention mechanisms, partly due to the inherent difficulty in avoiding the over-smoothing problem. However, treating features as attention to enhance the original ones inevitably leads to some information loss. Therefore, the over-smoothing issue imposes certain constraints on our module. Secondly, there is a computational cost concern. The current model design is relatively intricate and computationally intensive due to the substantial number of hypergraph nodes involved, resulting in high computational complexity.

6. Conclusion and outlook

This paper addresses the challenge posed by large variations in cell size, shape, and color in the localization task by reframing it as a feature misalignment problem between cell images and location maps. To tackle this issue, we propose a feature alignment network that harnesses the adaptive feature aggregation capability of hypergraphs. Our approach utilizes multi-scale hypergraphs to capture features with diverse correlations in proximity to nodes and leverages hypergraph neural networks to continuously aggregate and optimize node features for effective feature alignment. Additionally, we introduce a stepwise adaptive fusion module to extract useful feature information at different levels more efficiently and adaptively. Through extensive experiments, we demonstrate that our proposed multi-scale hypergraph module significantly mitigates the feature misalignment, leading to state-of-the-art performance in cell localization and counting tasks.

In future work, we plan to expand our research in two main directions: (1) Task expansion of existing approaches. As evident from Fig. 4, our approach of achieving feature alignment using hypergraphs has shown significant improvements in cell localization task. Therefore, our future plans involve extending this approach to a broader range of object localization task, such as crowd localization, few-shot object localization, and general object localization. (2) Technical refinement of the multi-scale hypergraph attention module. As demonstrated in Table 8, our proposed multi-scale hypergraph attention module outperforms various attention mechanisms and deformable convolution modules, indicating its substantial potential. In subsequent work, we aim to further optimize this module to make it a more versatile foundational component. These future directions align with our goal of advancing both the applicability and technical sophistication of our research.

CRediT authorship contribution statement

Bo Li: Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yong Zhang:** Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Chengyang Zhang:** Data curation, Formal analysis, Resources, Visualization, Writing – original draft. **Xinglin Piao:** Funding acquisition, Resources, Software, Supervision, Writing – original draft, Writing – review & editing. **Yongli Hu:** Funding acquisition, Resources, Supervision, Writing – review & editing. **Baocai Yin:** Funding acquisition, Project administration, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The experimental code and processed datasets will be publicly available at GitHub (https://github.com/Boli-trainee/MHFAN).

Acknowledgments

The research project is partially supported by the National Natural Science Foundation of China (No. 62072015, U21B2038, U19B2039, 61902053).

References

[1] V. Petukhov, R.J. Xu, R.A. Soldatov, P. Cadinu, K. Khodosevich, J.R. Moffitt, P.V. Kharchenko, Cell segmentation in imaging-based spatial transcriptomics, Nature Biotechnol. 40 (3) (2022) 345–354.

[2] N.F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C.C. Fullaway, B.J. McIntosh, K.X. Leow, M.S. Schwartz, et al., Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning, Nature Biotechnol. 40 (4) (2022) 555–565.

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

[4] B. Li, J. Chen, H. Yi, M. Feng, Y. Yang, H. Bu, Exponential distance transform maps for cell localization, 2023, http://dx.doi.org/10.36227/techrxiv.22275958.

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

[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.

[7] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9308–9316.

[8] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: The Advances in Neural Information Processing Systems, vol. 29, 2016.

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

[10] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 3558–3565.

[11] Z. Ma, Z. Jiang, H. Zhang, Hyperspectral image classification using feature fusion hypergraph convolution neural network, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–14.

[12] B. Li, Y. Zhang, C. Zhang, X. Piao, B. Yin, Hypergraph association weakly supervised crowd counting, ACM Trans. Multimed. Comput. Commun. Appl. (2023).

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

[14] Y. Guo, O. Krupa, J. Stein, G. Wu, A. Krishnamurthy, SAU-net: A unified network for cell counting in 2D and 3D microscopy images, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (4) (2021) 1920–1932.

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

[16] B. Li, Y. Zhang, Y. Ren, C. Zhang, B. Yin, Lite-unet: a lightweight and efficient network for cell localization, Engineering Applications of Artificial Intelligence 129 (2024) 107634.

[17] S. Huang, Z. Lu, R. Cheng, C. He, FaPN: Feature-aligned pyramid network for dense image prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 864–873.

[18] J. Xie, R. Zhu, Z. Wu, J. Ouyang, FFUNet: A novel feature fusion makes strong decoder for medical image segmentation, IET Signal Process. 16 (5) (2022) 501–514.

[19] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, A.K. Nandi, DefED-Net: Deformable encoder-decoder network for liver and liver tumor segmentation, IEEE Trans. Radiat. Plasma Med. Sci. 6 (1) (2021) 68–78.

[20] X. Li, Z. Xu, X. Shen, Y. Zhou, B. Xiao, T.-Q. Li, Detection of cervical cancer cells in whole slide images using deformable and global context aware faster RCNN-FPN, Curr. Oncol. 28 (5) (2021) 3585–3601.

[21] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: The Advances in Neural Information Processing Systems, vol. 28, 2015.

[22] J. Atwood, D. Towsley, Diffusion-convolutional neural networks, in: The Advances in Neural Information Processing Systems, vol. 29, 2016.

[23] M. Li, Y. Zhang, X. Li, Y. Zhang, B. Yin, Hypergraph transformer neural networks, ACM Transactions on Knowledge Discovery from Data 17 (5) (2023) 1–22.

[24] S. Bai, F. Zhang, P.H. Torr, Hypergraph convolution and hypergraph attention, Pattern Recognit. 110 (2021) 107637.

[25] H. Chen, L. Li, F. Hu, F. Lyu, L. Zhao, K. Huang, W. Feng, Z. Xia, Multi-semantic hypergraph neural network for effective few-shot learning, Pattern Recognit. (2023) 109677.

[26] M. Li, Y. Zhang, W. Zhang, Y. Chu, Y. Hu, B. Yin, Self-supervised nodes-hyperedges embedding for heterogeneous information network learning, IEEE Transactions on Big Data (2023).

[27] X. Hu, D. Wei, Z. Wang, J. Shen, H. Ren, Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints, Pattern Recognit. 111 (2021) 107688.

[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

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

[30] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.

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

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

[33] W. Ma, Y. Wu, F. Cen, G. Wang, Mdfn: Multi-scale deep feature learning network for object detection, Pattern Recognit. 100 (2020) 107149.

[34] V. Chalavadi, P. Jeripothula, R. Datla, S.B. Ch, et al., mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions, Pattern Recognit. 126 (2022) 108548.

[35] J. Huang, J. Yang, UniGNN: A unified framework for graph and hypergraph neural networks, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2021.

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

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

[38] S. Graham, M. Jahanifar, Q.D. Vu, G. Hadjigeorghiou, T. Leech, D. Snead, S.E.A. Raza, F. Minhas, N. Rajpoot, Conic: Colon nuclei identification and counting challenge 2022, 2021, arXiv preprint arXiv:2111.14485.

[39] S. Graham, M. Jahanifar, A. Azam, M. Nimir, Y.-W. Tsang, K. Dodd, E. Hero, H. Sahota, A. Tank, K. Benes, et al., Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 684–693.

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

[41] J.M.J. Valanarasu, V.M. Patel, Unext: Mlp-based rapid medical image segmentation network, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 23–33.

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

[43] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, in: Proceedings of the Medical Imaging with Deep Learning.

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

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

[46] S. Graham, Q.D. Vu, S.E.A. Raza, A. Azam, Y.W. Tsang, J.T. Kwak, N. Rajpoot, Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Med. Image Anal. 58 (2019) 101563.

[47] F. Tang, L. Wang, C. Ning, M. Xian, J. Ding, CMU-net: A strong ConvMixer-based medical ultrasound image segmentation network, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–5.

[48] F. Tang, J. Ding, L. Wang, C. Ning, S.K. Zhou, CMUNeXt: An efficient medical image segmentation network based on large kernel and skip fusion, 2023, arXiv preprint arXiv:2308.01239.

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

[50] R. Azad, R. Arimond, E.K. Aghdam, A. Kazerouni, D. Merhof, DAE-former: Dual attention-guided efficient transformer for medical image segmentation, 2022, arXiv preprint arXiv:2212.13504.

[51] J. Gao, M. Gong, X. Li, Congested crowd instance localization with dilated convolutional swin transformer, Neurocomputing 513 (2022) 94–103.

[52] J. Gao, T. Han, Q. Wang, Y. Yuan, X. Li, Learning independent instance maps for crowd localization, 2020, arXiv preprint arXiv:2012.04164.

[53] T. Han, J. Gao, Y. Yuan, X. Li, et al., LDC-Net: A unified framework for localization, detection and counting in dense crowds, 2021, arXiv preprint arXiv:2110.04727.

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

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

[56] Y. Liu, Z. Shao, N. Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions, 2021, arXiv preprint arXiv:2112.05561.

[57] C. Zhang, J. Chen, B. Li, M. Feng, Y. Yang, Q. Zhu, H.B. Bu, Difference-deformable convolution with pseudo scale instance map for cell localization, IEEE Journal of Biomedical and Health Informatics 28 (1) (2024) 355–366, http://dx.doi.org/10.1109/JBHI.2023.3329542.

Li Bo, received his bachelor degrees from Beijing Information Science and Technology University. He is currently a postgraduate student at the Department of Informatics, Beijing University of Technology. His research interests include computer vision and multimedia.

Zhang Yong, received the Ph.D. degree s in computer science from the BJUT, in 2010. He is currently an Associate Professor in computer science in BJUT. His research interests include intelligent transportation system, big data analysis and visualization, computer graphics.

Chengyang Zhang, received his bachelor degrees from Beijing Information Science and Technology University. He is currently a postgraduate student at the Department of Informatics, Beijing University of Technology. His research interests include computer vision and multimedia.

Xinglin Piao, received the Ph.D. degree from the Beijing University of Technology, Beijing, China, in 2017. He is currently a lecturer in the Faculty of Information Technology at Beijing University of Technology. His research interests include intelligent traffic, pattern recognition, and multimedia technology.

Yongli Hu, received the Ph.D. degree from the Beijing University of Technology in 2005. He is a Professor with the Faculty of Information Technology, Beijing University of Technology. He is a Researcher with the Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology. His research interests include computer graphics, pattern recognition, and multimedia technology.

Yin Baocai, received his Ph.D. degrees from Dalian University of Technology in 1993. He is a professor in the Faculty of Information Technology at Beijing University of Technology. He is a researcher at the Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology. He is a member of China Computer Federation. His research interests cover multimedia, multifunctional perception, virtual reality and computer graphics.