RESEARCH ARTICLE

ASATrans: Adaptive spatial aggregation transformer for cervical nuclei segmentation on rough edges

Hualin Sun[1]*, **Shengyao Hu**[2]

1 ChangZhou Vocational Institute Of Mechatronic Technology, ChangZhou, China, **2** ChangZhou Institute Of Technology, ChangZhou, China

* sunhualin021@126.com

Citation: Sun H, Hu S (2024) ASATrans: Adaptive spatial aggregation transformer for cervical nuclei segmentation on rough edges. PLoS ONE 19(7): e0307206. https://doi.org/10.1371/journal.pone.0307206

Editor: Mohammed Abdelsamea, University of Exeter, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: November 19, 2023

Accepted: June 26, 2024

Published: July 12, 2024

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The main characteristic of cervical cytopathy is reflected in the edge shape of nuclei. Existing computer-aided diagnostic techniques can clearly segment individual nuclei, but cannot clearly segment the rough edges of adherent nucleus. Therefore, we propose an effective method (ASATrans) to accurately segment rough cervical nuclei edges by exploring adaptive spatial aggregation methods. ASATrans creates a Multi-Receptive Embedding Layer that samples patches using diverse-scale kernels. This approach provides cross-scale features to each embedding, preventing semantic corruption that might arise from mapping disparate patches to analogous underlying representations. Furthermore, we design Adaptive Pixel Adjustment Block by introducing a long-range dependency and adaptive spatial aggregation. This is achieved through the stratification of the spatial aggregation process into distinct groups. Each group is given an exclusive sampling volume and modulation scale, fostering a collaborative learning paradigm that combines local features and global dependencies. This collaborative approach to feature extraction achieves adaptability, mitigates interference from unnecessary pixels, and allows for better segmentation of edges in the nucleus. Extensive experiments on two cervical nuclei datasets (HRASPP Dataset, ISBI Dataset), demonstrating that our proposed ASATrans outperforms other state-of-the-art methods by a large margin.

Introduction

Cell nuclei, as a major factor in assessing cervical cytopathology [1, 2], are crucial due to their significant manifestation in cervical cancer lesions [3]. Traditional cervical cancer pathology diagnosis mainly relies on physicians observing the morphological features of nuclei through a microscope [4, 5]. However, this method suffers from problems such as limited field of view and visual fatigue, which can easily lead to misdiagnosis. Recognition of cell nuclei in real scenarios remains challenging for current computer-aided diagnostic techniques [6, 7].

Existing vision transformers have wider receptive fields [8, 9], can effectively model long-distance relationships, and show excellent performance under large-scale training data and sufficient model parameters [10–12]. However, transformers lack some of the inductive biases inherent in convolutional neural networks (CNNs), and they often require large amounts of data to accurately model relationships, making them generally less performant than CNN models. This is especially true in real-life scenarios, where cervical cells are numerous, distributed in clusters, and have a stacking phenomenon. There are almost no studies that have applied transformer to the field of cervical cell nuclei segmentation. This is because directly applying existing transformer models may lead to poor segmentation accuracy and blurred nuclei edges [11, 13, 14]. The purpose of this paper is to improve the transformer model on small-scale datasets by exploring adaptive spatial aggregation methods to effectively segment rough cervical cancer cell nuclear edges [15, 16].

The primary distinction between vision transformers and Convolutional Neural Networks (CNNs) lies in their approach: vision transformers partition the image into patches and present a sequence of linear embeddings derived from these patches as input to the transformer block. Nevertheless, due to the varying scales of objects in different images [17, 18], the use of fixed-size patches often encounters challenges in capturing comprehensive local structures associated with objects. The rigidity of fixed patches introduces the risk of compromising semantic information, consequently resulting in a decline in performance. Existing segmentation methods pay little attention to this. To address this, We introduce a novel module termed the Multi-Receptive Embedding Layer (MREL), positioned at the initiation of each stage. MREL accepts the output (or input image) from the preceding stage and employs diverse-scale kernels to sample patches. This methodology imparts cross-scale features to each embedding, mitigating the potential semantic corruption arising from the assignment of disparate patches to analogous underlying representations. Consequently, MREL possesses the capability to reconfigure otherwise isolated patches into overlapping patches with varied receptive field sizes. This capability compensates for the loss of image information at the edges of patches due to simplistic patch segmentation and averts semantic corruption resulting from the convergence of different patches into similar latent representations.

Existing multi-attention mechanisms treat all pixels equally, which may lead to optimization bias towards smoothing the inner regions while underestimating the boundary pixels. This can unbalance foreground and background information, resulting in rougher predicted mask boundaries that do not align well with the boundaries of real objects [19–25]. To address this problem, we design Adaptive Pixel Adjustment Block(APAB) by introducing a long-range dependency and adaptive spatial aggregation. This is achieved through the stratification of the spatial aggregation process into distinct groups. Each group is given an exclusive sampling volume and modulation scale, fostering a collaborative learning paradigm that combines local features and global dependencies. This collaborative feature extraction effort enables adaptivity, mitigating interference from unnecessary pixels and allowing for better segmentation of edges in nuclei.

In summary, the contributions of this paper are:

1. Aiming at the problem that existing methods cannot achieve good segmentation results on small-scale data sets, we propose a novel transformer model—ASATrans, which generates finer nuclei edge shapes by exploring adaptive spatial aggregation methods.
2. Specifically, Multi-Receptive Embedding Layer in ASATrans samples patches using diverse-scale kernels. This approach provides cross-scale features to each embedding, preventing semantic corruption that might arise from mapping disparate patches to analogous underlying representations.

3. In addition, we design Adaptive Pixel Adjustment Block by introducing a long-range dependency and adaptive spatial aggregation. This is achieved through the stratification of the spatial aggregation process into distinct groups. Each group is given an exclusive sampling volume and modulation scale, fostering a collaborative learning paradigm that combines local features and global dependencies.
4. Extensive experiments on two cervical nuclei datasets (HRASPP Dataset, ISBI Dataset), demonstrating that ASATrans outperforms other state-of-the-art methods by a large margin.

Literature review

In the past decade, Deep Convolutional Neural Networks (CNNs) have been extensively employed for medical image segmentation, consistently exhibiting satisfactory performance. The preference for CNN architectures in numerous medical tasks due to their rapidly converge on modest datasets, yielding commendable accuracy and robustness. Building upon the success of transformers developed for Natural Language Processing (NLP) [26–30], researchers have tailored specific vision transformers for visual tasks, leveraging their potent attention mechanisms. Notably, ViT [9] and DeiT [31] successfully adapted the original transformer to vision domains, yielding impressive outcomes. Subsequent innovations, such as PVT [17], Swin [10], and ViTAE [32], introduced the pyramid structure to vision transformers, substantially reducing the number of patches in underlying layers. Furthermore, these advancements extended the applicability of vision transformers to diverse visual tasks, including object detection and segmentation. In addition, these advances have successfully extended Transformer to various other visual tasks [33], including detection, classification, segmentation, etc. Task scenarios include liver tumor segmentation [34], cell segmentation [35], etc.

While vision transformers (ViTs) [36, 37] have demonstrated exceptional performance on large datasets, their efficacy tends to diminish when trained on smaller datasets, possibly attributable to the absence of localized inductive bias in their architecture. Recent investigations [15] have addressed this limitation by introducing locality to the architecture, thereby enabling ViTs to achieve performance comparable to CNNs in scenarios involving smaller datasets. To address the challenge of indistinct edges, Wang et al. [38] proposed the Boundary-Aware Transformer (BAT), incorporating boundary-aware gates in the transformer architecture to leverage prior knowledge about boundaries. BAT was effectively trained with assisted supervision to enhance performance. Additionally, Pu et al. introduced the Transformer-Based Edge Detector (EDTER) [39], employing two distinct phases to extract global context and local cues, which are subsequently fused by a feature fusion module for precise edge prediction. Although these methods exhibit success in diverse domains, the scarcity of foreground pixels in cervical cell nuclei segmentation poses challenges in rapidly establishing a local vision structure. Consequently, there is an urgent need for a transformer model tailored to excel on small-sized cervical cell nuclei segmentation datasets.

Methodology

We proposed ASATrans to solve the problem of blurred edge segmentation of transformer on small-scale datasets. The overall structure is shown in Fig 1. The input image first passes through the Multi Receptive Embedding Layer, and then passed through the Swin Transformer Block in stages 1 and 2 and the Adaptive Pixel Adjustment Block and applied in stages 3 and 4. In the decoding part, we use UperNet Head as the decoding head, which mainly includes FPN and PPM modules. Finally, we get the final prediction result through a classifier.

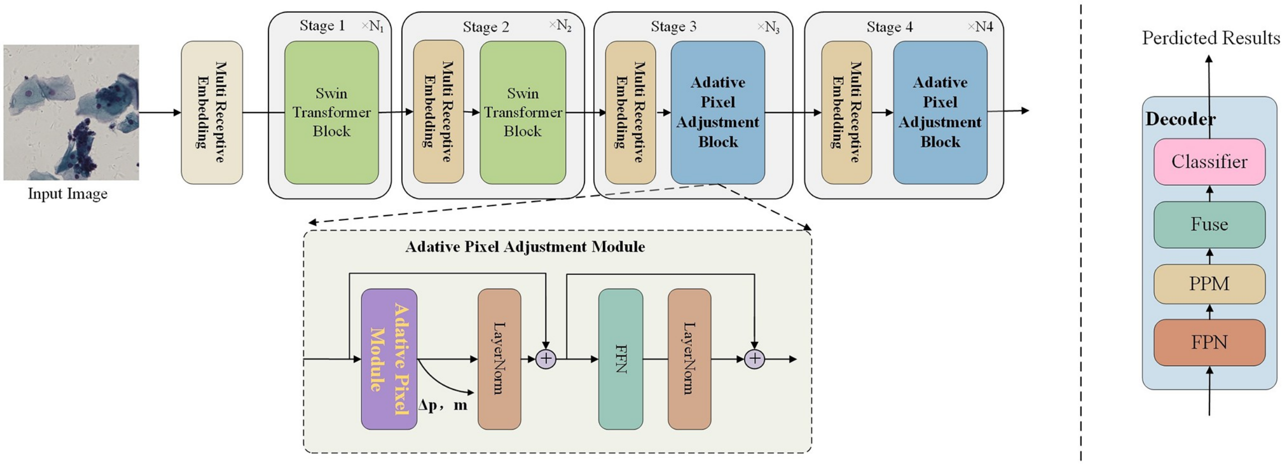

Input Image
Multi Receptive Embedding
Stage 1 ×N1
Swin Transformer Block
Stage 2 ×N2
Multi Receptive Embedding
Swin Transformer Block
Stage 3 ×N3
Multi Receptive Embedding
Adative Pixel Adjustment Block
Stage 4 ×N4
Multi Receptive Embedding
Adative Pixel Adjustment Block
Perdicted Results
Decoder
Classifier
Fuse
PPM
FPN
Adative Pixel Adjustment Module
Adative Pixel Module
Δp，m
LayerNorm
FFN
LayerNorm

Fig 1. An illustration of the model architecture of ASATrans. The overview of our proposed ASATrans architecture, including the global overview of backbone (above), details of the internal structure of our Adaptive Pixel Adjustment Block (below).

https://doi.org/10.1371/journal.pone.0307206.g001

Adaptive pixel adjustment block

Unlike CNN models, visual transformer has a larger receptive field and excels at modeling long-distance relationships, which shows excellent performance on large datasets. However, since transformer lacks some of the inductive biases that CNNs inherently have, causing it to often require a large amount of data to model relationships, it tends to perform less well than CNN models on undersized medical image datasets. In particular, in the field of cervical cancer cell nuclei segmentation, the small number of datasets and the disparity in the ratio of foreground pixels to background pixels make it difficult for existing transformer models to achieve satisfactory results.

So we designed the Adative Pixel Adjustment Block to replace the traditional transformer block and use it to make up for the shortcomings of convolution and multi-head self-attention. Compared with MHSA whose weights are dynamically adjusted by the input, the Adative Pixel Module is an operator with static weights and strong inductive bias, so that APAB can fully enjoy the advantages of both mechanisms. Due to their highly inductive nature, models composed of regular convolutions may converge faster than VITS and require less training data. The traditional multi-attention mechanism is shown below:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_{head}}}V\right) \tag{1}$$

A direct approach to reconcile the disparity between convolutional operations and Multi-Head Self-Attention (MHSA) is to imbue conventional convolution with the capacity for long-range dependencies and adaptive spatial aggregation. Analogous to the DCNv2, this represents a generalized form of traditional convolution. For a given input tensor $w_k \in \mathbb{R}^{C\times H\times W}$ and the present pixel p_0, the formulation of DCNv2 can be expressed as:

$$y(p_0) = \sum_{p_n \in \mathbb{R}} W(p_n) \cdot x(p_0 + p_n) \tag{2}$$

$$y(p_0) = \sum_{p_n \in \mathbb{R}} W(p_n) \cdot x(p_0 + p_n + \triangle p_n) \tag{3}$$

$$y(p_0) = \sum_{k=1}^{K} w_k m_k x(p_0 + p_k + \Delta p_k) \tag{4}$$

Here, K signifies the overall count of sampling points, with k serving as the index for individual sampling points. The notation $w_k \in \mathbb{R}^{C\times H\times W}$ designates the projection weights associated with the k-th sampling point, while $m_k \in \mathbb{R}$ denotes the modulation scalar corresponding to the k-th sampling point. This modulation scalar is subject to normalization through a sigmoid function. Additionally, p_k represents the k-th location within the predefined grid sampling, akin to conventional convolutional processes.

$$\mathbb{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\} \tag{5}$$

The symbol Δp_k denotes the displacement corresponding to the k-th grid sampling location. It is discernible from the mathematical expression that, in the context of long-range dependencies, the sampling offset Δp_k exhibits a degree of flexibility, enabling its interaction with features of both short and long-range characteristics. Moreover, for the purpose of adaptive spatial aggregation, both the sampling offset Δp_k and the modulation scalar m_k are endowed with learnable attributes and are conditioned by the input variable x. It is thereby evident that DCNv2 shares analogous advantageous characteristics with Multi-Head Self-Attention (MHSA), prompting our initiative to construct foundation models of large-scale Convolutional Neural Networks (CNNs) grounded upon this operator.

In order to augment the efficacy of the convolutional structure embedded within the transformer block, we have operationalized the ensuing strategies:

Weight sharing among convolutional neurons. In convolution, different convolution neurons have independent linear projection weights, so their parameters and complexity are linearly related to the total number of sampling points. We use the idea of separable convolution to share the weights between neurons, which effectively reduces the auxiliary degree of the model and makes it possible to apply it in large-scale models. We separate the weight w_k in normal convolution into a depth part and a point part. The original position-aware modulation scalar m_k is responsible for the depth part, and the shared projection weight w between sample points represents the point part.

Introducing the multi-group mechanism. Second, weight sharing just for convolutions is not enough. In MHSA, attention is often divided into multiple groups for calculation. Inspired by this, we also divide the spatial aggregation process of APAB into G groups. The basic idea is similar to the MHSA widely used in transformre, except that each group of ours has a separate sampling offset Δp_{gk} and modulation scale m_{gk}. Therefore, different spatial aggregations exist in different groups on a single convolutional layer, so they can better adapt to different downstream tasks and achieve better convergence speed and performance.

Normalization of the modulation scalar across sampling points. In order to control the sampling offset in each group not to exceed a reasonable range, we need to normalize the sampling points. Because the gradient in convolution is unstable when training with large-scale parameters or data. The offset of all sample points in APAB may be outside the normal range. In order to solve this problem, we use the improved sigmoid function to normalize the elements of the modulation parameter scalar, which can make the offset parameters more stable during the training process. The original formula is as follows:

$$\mathrm{softmax}(x_i) = \frac{e^{x_i - c}}{\sum_{j=1}^{d} e^{x_j - c}} = \frac{e^{x_i} e^{-c}}{e^{-c} \sum_{j=1}^{d} e^{x_j}} \tag{6}$$

We have made some changes to reduce the computational cost and increase the speed. The specific implementation method is to take the log of the value, as follows:

$$\mathrm{softmax}(x_i) = \log \frac{e^{x_i - c}}{\sum_{j=1}^{d} e^{x_j - c}} = x_i - c - \log \sum_{j=1}^{d} e^{x_j - c} \tag{7}$$

Compared with softmax, using Log-softmax has many advantages, including improved numerical performance and gradient optimization. These advantages are very important for implementation, especially when the computational cost of training the model is high, they can bring very objective benefits. Moreover, the use of log probability has better information theory interpretability.

In order to control the sampling offset in each group not to exceed a reasonable range, we need to normalize the sampling points. Because when training with large-scale parameters or data, the gradient in convolution is unstable. The offset of all sampling points in APAB may exceed the normal range. In order to solve this problem, we use the sigmoid function to normalize the elements of the modulated parameter scalars, so that the sum of all scalars is 1, which can make the offset parameters during the training process more stable.

Combined with the above modifications, the extended DCNv2, can be formulated as:

$$y(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} w_g m_{gk} x_g (p_0 + p_k + \Delta p_{gk}) \tag{8}$$

Let G denote the total number of aggregated groups. In the context of the g-th group, $w_g \in \mathbb{R}^{C \times C'}$ represents the position-independent projection weight, where $C' = C/G$ signifies the dimension of the group. The term $m_{gk} \in \mathbb{R}$ pertains to the offset associated with the k-th sampling point in the g-th group, which is normalized along the k-th dimension through the application of the softmax function. The variable $x_g \in \mathbb{R}^{C' \times H \times W}$ denotes the sliced input feature map within the g-th group. Furthermore, Δp_{gk} represents the offset corresponding to the grid sampling position p_k within the g-th group.

The predefined scalar γ used to modulate the offset amplitude is empirically set to 0.1, which we believe is unreasonable empirically. Because this limits the offset distance, if a large range of edge distortion is encountered, the degree of deformation will not be enough to cope with large changes. And if you encounter smaller deformations, it will be difficult to identify the degree of distortion. We set γ as a variable variable, and its formula is

$$\triangle p_{ij} = \gamma \triangle \hat{p_{ij}} \odot (w, h) \tag{9}$$

where (w, h) are the width and height of the ROI, by element-wise product with the width and height of the ROI. γ is limited between 0.01–0.5, with an initial value of 0.1, which is obtained through adaptive learning. The parameters are continuously adjusted during the training process through the back propagation algorithm to improve the accuracy and generalization ability of the model, and then transform the deformed The amplitude is controlled within a reasonable range, thereby better segmenting the edges of distorted cervical cell nuclei and helping doctors better judge the extent of cancer lesions.

In summary, the APAM operator serves to rectify the limitations of regular convolution with respect to long-distance dependencies and adaptive spatial aggregation.

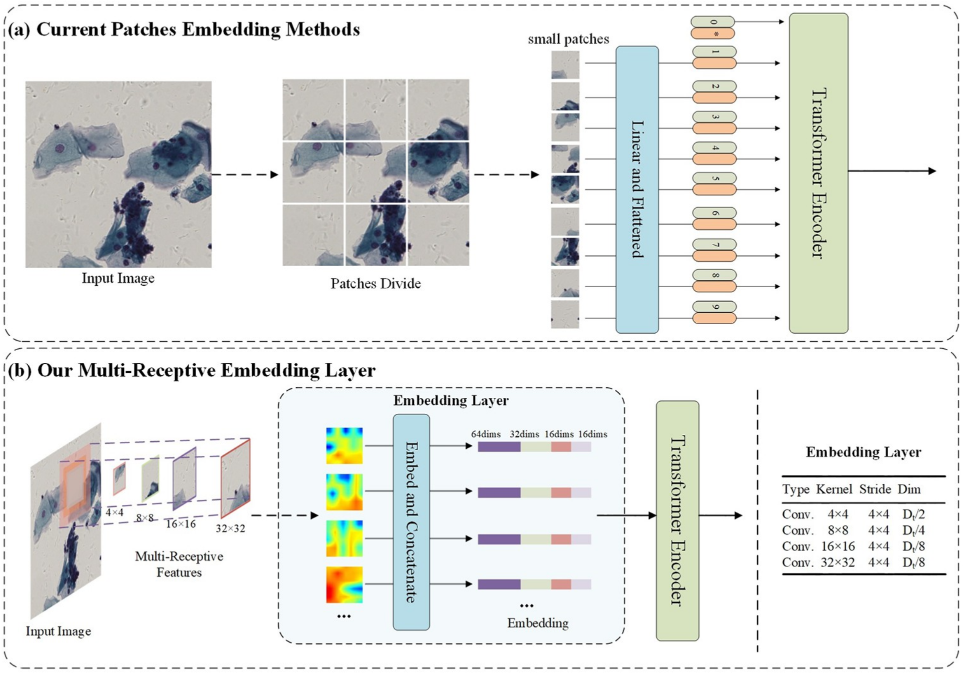
Multi receptive embedding layer

The main difference between visual transformers and CNN models is how images are processed. In visual transformers, images are divided into blocks, and these blocks are linearly embedded and passed through transformer blocks. However, this simple patch-based segmentation approach has two issues: (1) Loss of local structures: Regular patches (e.g., 16x16) struggle to capture complete local object structures as object scales vary in different images. (2) Semantic inconsistency: Objects in different images may have different geometric variations (scaling, rotation, etc.), and fixed patch segmentation may capture inconsistent object information, potentially degrading semantics and performance. As a result, some intrinsic inductive bias is lost during image segmentation, leading to inferior performance on small-scale datasets compared to CNN models.

In Fig 2(a), the current patch embedding method divides the image into small patches, linearizes them, and then flattens them before inputting them to the encoder. In contrast, Fig 2(b) shows our approach, which embeds and concatenates multi-sized image patches into a linear representation.

Multi-Resolution Embedded Layers (MRELs) are employed for the generation of input embeddings at each stage of the process. The initial MREL, depicted in Fig 2 and positioned prior to the first stage, accepts the image as its input. Subsequently, it samples patches using four kernels characterized by varying sizes. The step size of these kernels is appropriately adjusted to ensure uniform embedding counts. Notably, these four patches correspond to identical central regions but vary in scales. Ultimately, these patches undergo projection and consolidation into a unified embedding, a process typically executed through the utilization of four convolutional layers.

Dealing with cross-scale embeddings poses the challenge of selecting the right projection dimension for each scale. The computational cost of a convolutional layer scales with $K^2 \times D^2$,



Type	Kernel	Stride	Dim
Conv.	4×4	4×4	$D_t/2$
Conv.	8×8	4×4	$D_t/4$
Conv.	16×16	4×4	$D_t/8$
Conv.	32×32	4×4	$D_t/8$

Fig 2. An illustration of consecutive pixel patch moudle. (a) shows current patches embedding methods. (b) shows our multi-receptive embedding layer.

https://doi.org/10.1371/journal.pone.0307206.g002

where K is the kernel size, and D represents the input/output dimensions when they are equal. This means that larger kernels consume more computational resources compared to smaller ones for the same dimension. To efficiently manage the computational budget of the Multi-Resolution Embedding Layer (MREL), we assign lower dimensions to larger kernels and higher dimensions to smaller kernels. Specific allocation rules, along with a 128-dimensional example, are provided in sub-tables within Fig 2. Our approach significantly reduces computational overhead without significantly affecting the model's performance, compared to the conventional practice of evenly distributing dimensions. Similar processes are followed in the cross-scale embedding layers in other stages. As shown in Fig 1, MRELs in stages 2/3/4 utilize two different kernel sizes (2x2 and 4x4). Additionally, in stages 2/3/4, the MREL span is set to 2x2 to create a pyramid structure, effectively reducing the number of embeddings to one-fourth.

Experiments

Dataset and metric

This article presents experimental investigations conducted on both HRASPP datasets [3] and publicly available ISBI datasets [22]. Each dataset exhibits distinctive characteristics. The internally generated dataset is characterized by an abundance of cell clusters or stacks, posing challenges in the segmentation process. Notably, a mere 0.01% of pixels within each image correspond to nuclei, as shown in Fig 3. In contrast, the ISBI dataset involves cervical cells obtained through a Pap smear by a skilled medical professional, with subsequent presentation of slide images under a microscope. While the ISBI dataset also exhibits cell adhesion, it is observed to be of a milder degree. Furthermore, there is no substantial size difference in cell nuclei when compared to our proprietary dataset.

In the evaluation of the performance of our ASATrans model, three commonly utilized metrics were chosen: Intersection over Union (IoU), Dice coefficient, and Pixel Accuracy (PA). IoU and Dice coefficient are frequently employed in the assessment of medical image segmentation, as they provide direct quantification of pixel overlap between predicted outcomes and ground truth labels. Furthermore, Pixel Accuracy (PA) denotes the proportion of accurately classified pixels relative to the total pixel count.

These metrics are formally defined as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{11}$$

$$PA = \frac{TP + TN}{TP + TN + FT + FN} \tag{12}$$

Training details

The experiments were implemented using the PyTorch framework and executed on an NVIDIA GeForce RTX 3080. All methodologies underwent training with a Batch Size of 2 and employed SGD as the optimizer. The initial learning rate was established at 0.001, with a momentum of 0.9 and a weight decay of 0.0005.

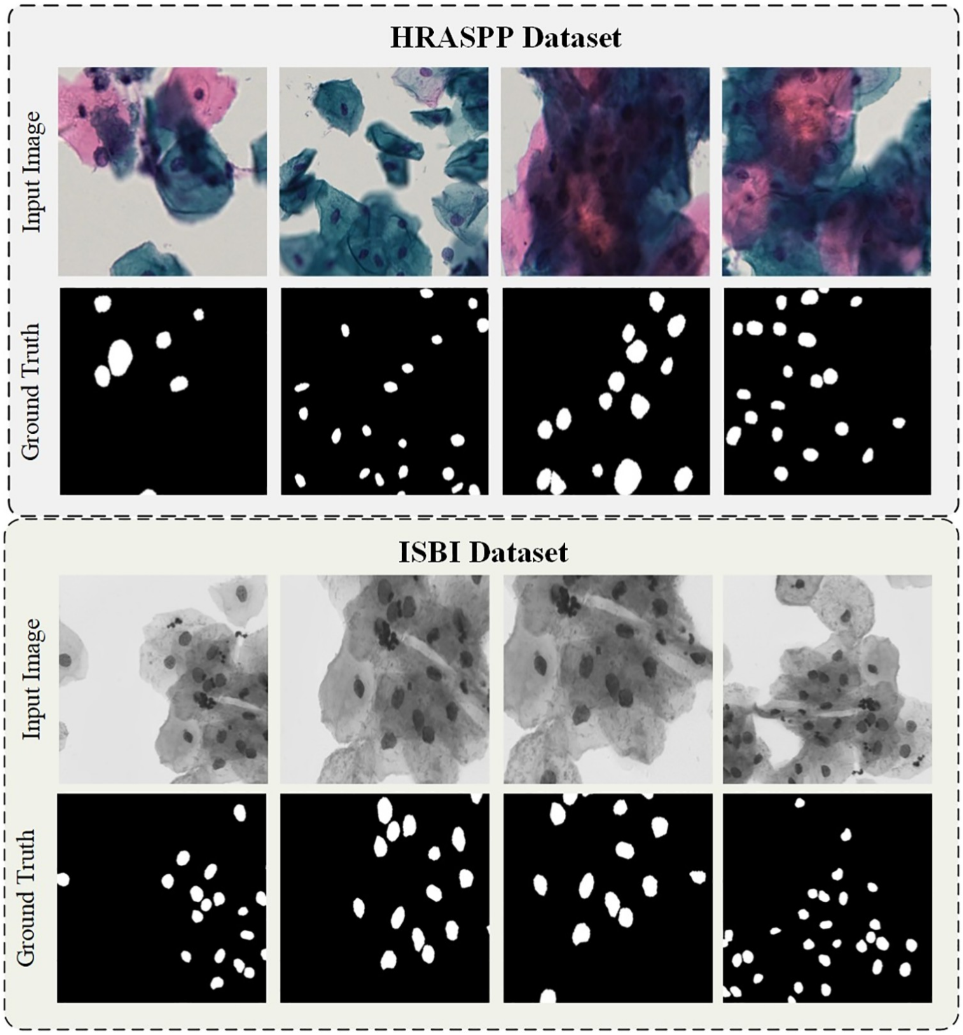


Fig 3. Display of dataset image.

https://doi.org/10.1371/journal.pone.0307206.g003

The training process incorporated **Multiple Losses**, specifically CrossEntropyLoss and DiceLoss, applied to the datasets. The ratio was set at 3. Data augmentation techniques, including flipping, rotating, and cropping operations, were applied to augment the three datasets. The input model operated on images with dimensions of 512 × 512, while the crop size for the transformer model was set to 224 × 224. Pre-training utilized the ImageNet-1k dataset, and model performance was evaluated on the official test set provided by the dataset itself. A total of 360,000 iterations were conducted during the training phase.

Comparison with other methods

To show the efficacy of ASATrans, a comparative analysis was conducted against several state-of-the-art methods using two distinct datasets. The selected methods encompassed Vision Transformer [9], Swin Transformer [10], Swin Unet [40], U2NET [41], UNET++ [42], and TransUNET [43]. To ensure equitable comparisons, all models were executed within the same configuration environment. The outcomes of these comparisons are presented in Tables 1 and 2, while visual contrasts are depicted in Figs 4 and 5.

Across two datasets with disparate characteristics (HRASPP dataset and the ISBI dataset), the results in Tables 1 and 2 consistently indicate the superior performance of our proposed method when compared to both CNN-based and Transformer-based models. This observation underscores the notable potential of our model for the segmentation of cervical cell nuclei edges, particularly in the context of limited dataset sizes.

Specifically, on HRASPP dataset, cervical cell nuclei often appear in cell clusters, leading to cell stacking, overlapping, and difficult segmentation. The CNN-based U2Net and Unet++ models converge quickly and fluctuate stably, and can quickly learn the morphological features of the cell nuclei, achieving notable results, with the best model achieving an IoU of 0.5034. In contrast, the Transformer-based model converges slower and fluctuates more, with a longer training cycle, and due to the lack of the CNN-based model's inherent of inductive biaes, their performance is often inferior to that of CNN-based models. For example, the ViT model only achieved an IoU of 0.4077, the worst performance on our dataset. However, our ASATrans model provides finer segmentation of cell nucleus edges with comparable model sizes by dynamic adaptive spatial aggregation, which allows the input patch embedding to contain more local detail information, and APAM to bias the transformer's attention more towards the foreground. Ultimately, our model achieves 0.89% higher IoU performance and 0.35% higher Dice performance than the next best CNN-based model, and 1.65% higher IoU

Table 1. Quantitative comparison of different excellent methods on HRASPP dataset.

Methods	Input Size	Crop Size	Params(M)	IoU	Dice	PA
ViT	512^2	224^2	142	0.4077	0.5792	0.5449
Swin-Trans	512^2	224^2	120	0.4629	0.6329	0.5709
Swin-UNet	512^2	224^2	79	0.4552	0.6256	0.5335
NucleiSegNet	512^2	256^2	93.54	0.4943	0.6866	0.5735
U2Net	512^2	256^2	44.63	0.5034	0.6697	0.5771
UNet++	512^2	256^2	35	0.4761	0.6451	0.5412
OCR	512^2	256^2	56.75	0.5058	0.6718	0.6124
HR-AS	512^2	256^2	71.23	0.5064	0.6723	0.5923
TransUNet	512^2	224^2	86	0.4958	0.6629	0.5854
Ours	512^2	224^2	63.71	**0.5123**	**0.6732**	**0.5991**

https://doi.org/10.1371/journal.pone.0307206.t001

Table 2. Quantitative comparison of different excellent methods on ISBI Dataset.

Methods	Input Size	Crop Size	Params(M)	IoU	Dice	PA
ViT	512^2	224^2	142	0.7698	0.8699	0.9627
Swin-Trans	512^2	224^2	120	0.8628	0.9264	0.9484
Swin-UNet	512^2	224^2	79	0.8216	0.9021	0.9532
NucleiSegNet	512^2	256^2	93.54	0.8456	0.9261	0.9475
U2Net	512^2	256^2	44.63	0.8321	0.9082	0.9480
UNet++	512^2	256^2	35	0.8474	0.9174	0.9493
OCR	512^2	256^2	56.75	0.8457	0.9163	0.9541
HR-AS	512^2	256^2	71.23	0.8498	0.9188	0.9776
TransUNet	512^2	224^2	86	0.8581	0.9236	0.9569
Ours	512^2	224^2	63.71	**0.8779**	**0.9362**	**0.9775**

https://doi.org/10.1371/journal.pone.0307206.t002

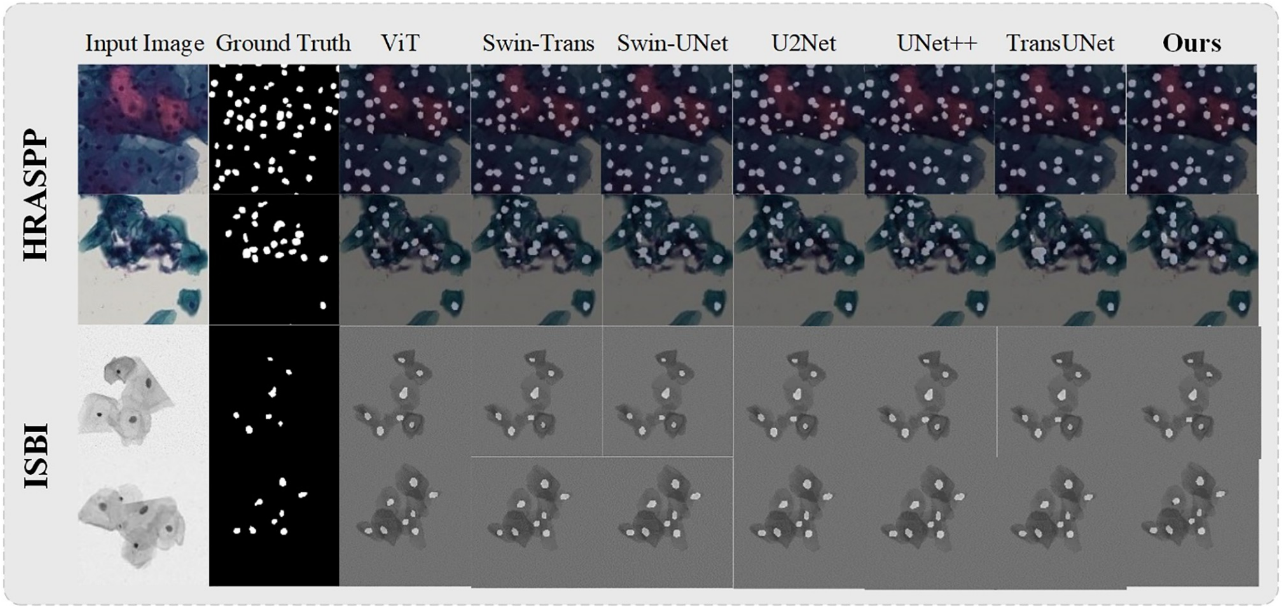


Fig 4. Display of transformer-based methods prediction results. Each dataset presents two nuclei images and ground truth images.

https://doi.org/10.1371/journal.pone.0307206.g004

performance and 1.03% higher Dice performance than the next best Transformer-based model. This also proves the effectiveness of our module design.

On the ISBI dataset, all models perform well with very clear nuclei and good contrast, despite the presence of cell stacking. In this case, the model structure based on Transformer

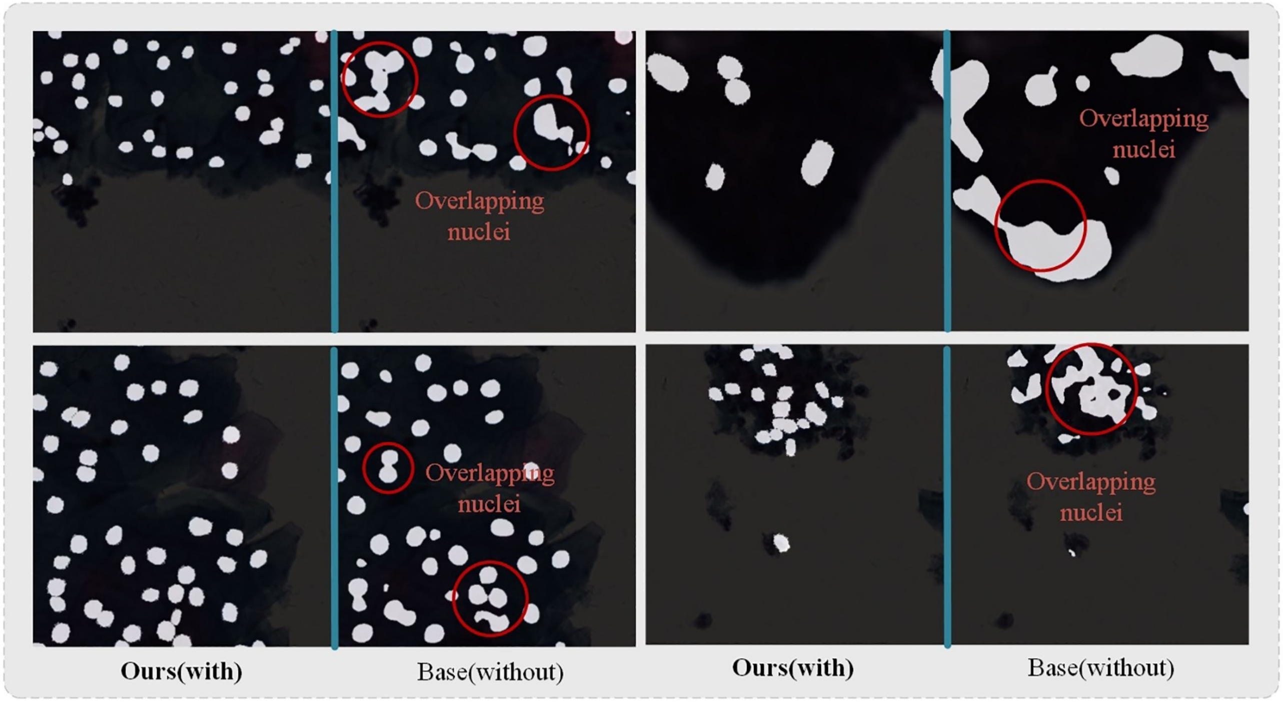


Fig 5. Refinement effect of ASATrans on edge. Left is the segmentation result with ASATrans, right is the segmentation result without ASATrans.

https://doi.org/10.1371/journal.pone.0307206.g005

can better establish long-distance dependency and fully utilize the performance of Transformer. As can be seen from Table 2, compared with the next best model, the model in this paper achieves 1.51% and 0.98% improvement in IoU and Dice metrics, respectively.

Combining the results of the two individual datasets, we find that the CNN-based model usually converges faster and works better than the Transformer-based model on HRASPP dataset, which has severe nucleus stacking, cell aggregation, and large picture differences. And on ISBI and Herlev, two small datasets with little difference, Transformer can fully utilize the performance. Our designed A and B modules perform adequately on all three datasets to better model cell nucleus images. ASATrans can improve the accuracy of spatial localization of small-size targets such as cell nuclei, act as a refinement of segmentation edges, and enhance the robustness of the model. These optimizations have a significant role in improving the performance in the task of cell nucleus image segmentation.

Ablation studies

Taking the ISBI dataset as an example, we have conducted experimental discussions on the selection of hyperparameters and ablation experiments under various settings to validate the effectiveness of the proposed individual modules for segmentation.

In order to demonstrate the effectiveness of the proposed modules, we conducted comparison experiments on the ISBI dataset using various combinations as shown in Table 3. Based on the results reported in Table 3, it is easy to see that the components in ASATrans are compatible with each other, while each component contributes to the improvement of segmentation rate.

Specifically, we compare the segmentation performance of the base model in the first row and the second row with the addition of the APAB module. Observe that the addition of APAB module improves IoU, Dice and PA by 0.7%, 0.38% and 0.23%, respectively. This indicates that the new embedding can maintain the local continuity of the pixels around the patch, and will not roughly break up the complete image into split chunks as in the case of plain block segmentation, thus avoiding the loss of image information at the edges of the block and maintaining the intrinsic scale invariance of the image. Compared to adding module APAB, comparing the segmentation performance of the base model in the first row and the third row, we find that the performance improvement is more with the addition of module MREL. The IoU and Dice are improved by 0.9% and 0.7%, respectively. This is that the APAM operator compensates for the shortcomings of regular convolution in terms of long distance dependence and adaptive spatial aggregation; Compared with common attention-based operators such as MHSA and closely related deformable attention, this operator inherits the inductive bias of convolution, which makes our model more efficient, with less training data and shorter training time; this operator is based on sparse sampling, which is more efficient than previous methods such as MHSA and heavily parameterized methods such as the large kernel with parameterization is more computationally and memory efficient. MREL supplements the

Table 3. Quantitative comparison of different excellent methods on ISBI dataset.

Components		Results		
APAB	MREL	IoU	Dice	PA
x	x	0.8628	0.9264	0.9484
√	x	0.8698	0.9302	0.9507
x	√	0.8718	0.9334	0.9594
√	√	**0.8779**	**0.9362**	**0.9775**

https://doi.org/10.1371/journal.pone.0307206.t003

image information lost at patch edges due to plain patch partitioning and prevents the semantic corruption caused by mapping different patches to similar latent representations. Comparing lines 2, 3, and 4, adding two modules results in greater performance improvement compared to adding one module. This demonstrates that the components in ASATrans are compatible with each other and that each component contributes to improving segmentation performance. Furthermore, it shows that our proposed ASATrans effectively enhances the segmentation performance of cervical cancer cell nucleus edges on small datasets.

Hyperparameter discussion

Multi-group (head) design first appeared in group convolution, which is widely used in MHSA of transformers and used with adaptive spatial aggregation to effectively learn richer information from different representation subspaces at different locations. Inspired by this, we divide the spatial aggregation process into G groups, each group has separate sampling offsets and modulation scales, so different groups on a single convolutional layer can have different spatial aggregation patterns, thus providing better performance for downstream tasks. Strong functionality.

In order to determine the optimal hyperparameter G, we set it to groups 2, 3, 4, 6, and 12 respectively. The experimental results are shown in Table 4.

Experimental results show that when G is set to 2, the performance improvement is minimal, only 0.12. However, setting the number of spatial aggregation groups to 4 results in the best performance improvement, up to 0.33. Therefore, we conclude that 4 is the optimal number of groups. This choice is similar to the spatial pyramid structure, where people usually aggregate 4 feature maps of different sizes to obtain comprehensive information. The reason we didn't choose 5 is that it is difficult to divide. Additionally, when the number of sets increases to 6 or 12, performance drops to about 0.32. We believe this is because aggregating a larger number of groups leads to information redundancy, thereby reducing performance. In addition, choosing a larger number of groups will increase the computational complexity and may lead to the problem of exploding gradients. Therefore, choice 4 considers both speed and accuracy.

Visualization

Segmentation results. Fig 4 shows the visualization results of the segmentation prediction, with two images selected for visualization for each dataset. Observing the visualization results of rows 2, 3, and 4 in Table 4, it can be seen that the visualization results of Transformer models with poorer performance (e.g., Vision Transformer) are also very rough, and the visualization results reflect the performance performance of the models very well. In contrast, as seen in the last column, our model has the best visualization results, with higher segmentation accuracy and finer segmentation of edges.

Table 4. Hyperparameter discussion.

Group Number	Improved(%)
2	0.13
3	0.24
4	**0.33**
6	0.323
12	0.322

https://doi.org/10.1371/journal.pone.0307206.t004

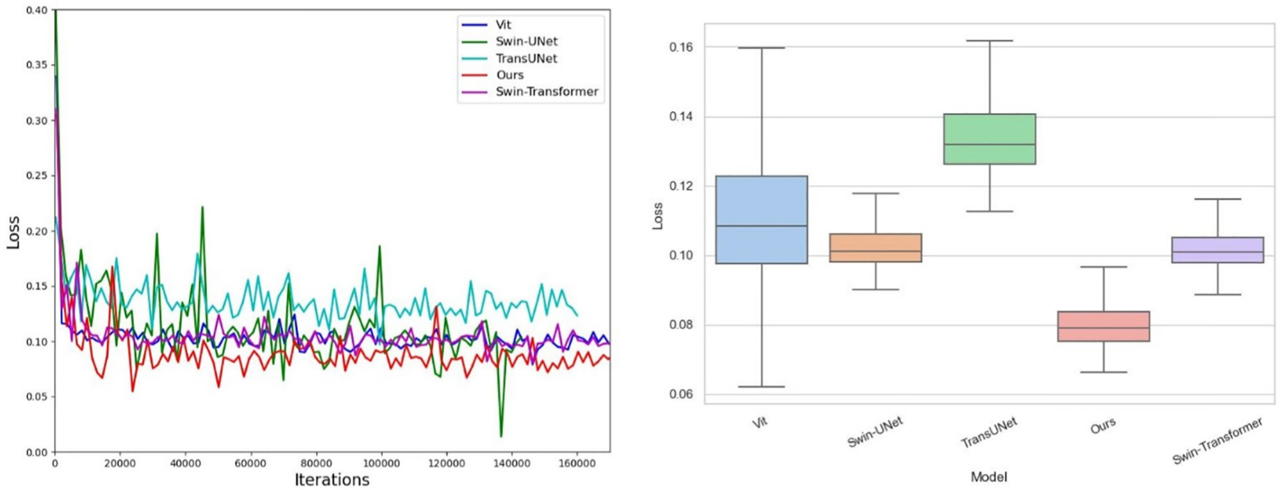


Fig 6. Effect on the convergence and stability.

https://doi.org/10.1371/journal.pone.0307206.g006

To further illustrate the improved effect of our proposed APAB on edge segmentation. We compare our model with the benchmark model and show the visualization results. It can be clearly observed that the originally adherent nuclei edges become clearer in the left-right comparison plots in Fig 5. The edges of Overlapping nuclei tend to be adherent and unclear, and we can observe in the region marked by the red box in Fig 5 that the overlapping boundaries segmented by ASATrans are more clearer than the edges segmented by the benchmark model as the training progresses.

Convergence and stability. The ASATrans we proposed can effectively improve the stability of the model and accelerate the convergence of the model. The proposed MREL module avoids the loss of local structure information through multi-scale feature extraction and accelerates the convergence of the model. At the same time, the inherent inductive bias is supplemented to enhance semantic consistency and make model training more stable. As shown in Fig 6, red represents our model. It is obvious that our model is more stable than other models.

Attention map. To further illustrate that the APAB does indeed bias attention that would otherwise be focused on the background more towards the foreground, we used GradCAM to visualize the model's attention. We visualized the 2D activations by weighting the 2D activations by the average gradient and selecting the maximum value channel. The first row of Fig 7 shows the distribution of attention before the addition of the APAB module, and it can be seen that the attention is scattered and much of it is focused on the background. Whereas after the addition of the APAB module, as shown in the second row of Fig 7 the attention is shifted from the background to the foreground, focusing more on the region of the nucleus clusters. Comparing columns 1, 2, and 3 in the figure, we can observe that as the training progresses, Transformer's attention becomes more refined, better segmenting the edges of the cell nuclei. This further proves the effectiveness of APAB.

Conclusions

In this paper, we delve into how the transformer model can be improved to finely segment blurred cell nuclei edges on small-scale datasets. First, we observe that the existing transformer model loses edge information when crudely dividing an image into small patches, making it difficult to quickly establish long-distance dependencies, which is detrimental to model

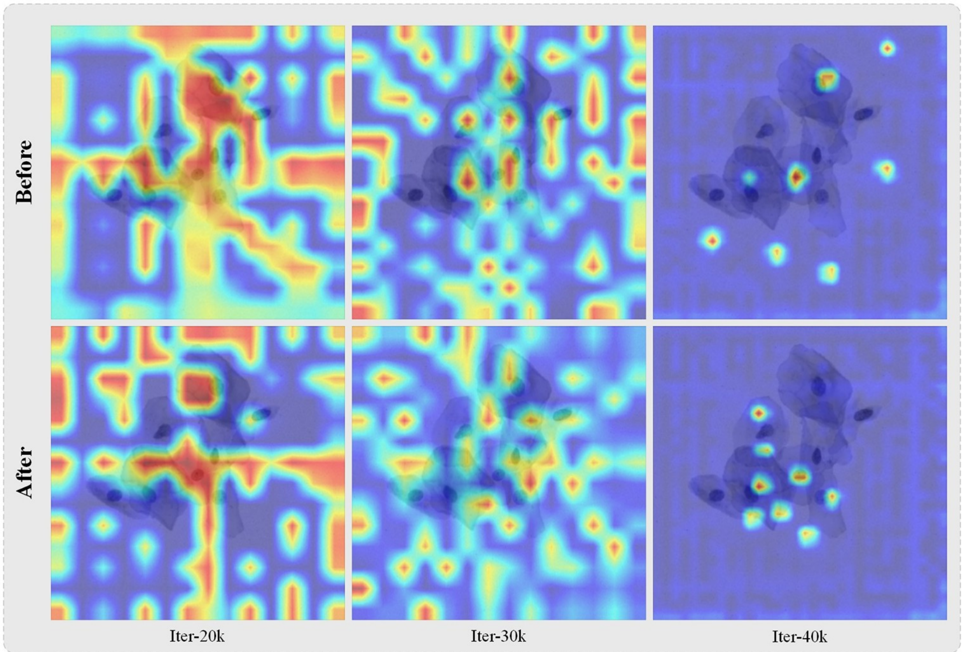


Fig 7. Effect of APAB on attention map. Above is the attention map without APAB, below is the attention map with APAB.

https://doi.org/10.1371/journal.pone.0307206.g007

convergence and edge segmentation. In addition, when extracting image features, the transformer's attention tends to focus too much on the background and ignores the important foreground information. To address these issues, we propose a simple yet effective transformer framework named ASATrans, which learns sparse attention in a data-dependent manner and models geometric transformations to bias the attention more from the background to the foreground. ASATrans effectively improves the accuracy of edge segmentation of cell nuclei. We applied ASATrans to the difficult task of cell nuclei segmentation with small datasets and obtained finer cell nuclei segmentation edges. Numerous experiments demonstrate the effectiveness of our ASATrans model, which performs better and achieves significant improvement compared to other baseline models.

Supporting information

S1 Fig. Original pathological images of Figs 1–5 and 7 in the main body.
(ZIP)

Author Contributions

Conceptualization: Hualin Sun, Shengyao Hu.

Data curation: Hualin Sun, Shengyao Hu.

Formal analysis: Hualin Sun, Shengyao Hu.

Funding acquisition: Hualin Sun, Shengyao Hu.

Investigation: Hualin Sun, Shengyao Hu.

Methodology: Hualin Sun, Shengyao Hu.

Project administration: Hualin Sun, Shengyao Hu.

Resources: Hualin Sun.

Software: Hualin Sun.

Supervision: Hualin Sun.

Validation: Hualin Sun.

Visualization: Hualin Sun.

Writing – original draft: Hualin Sun.

Writing – review & editing: Hualin Sun.

References

1. Yang J, Xu R, Wang C, Qiu J, Ren B, You L. Early screening and diagnosis strategies of pancreatic cancer: a comprehensive review. Cancer Communications. 2021; 41(12):1257–1274. https://doi.org/10.1002/cac2.12204 PMID: 34331845
2. Wardle J, Robb K, Vernon S, Waller J. Screening for prevention and early diagnosis of cancer. American psychologist. 2015; 70(2):119. https://doi.org/10.1037/a0037357 PMID: 25730719
3. Zhang J, Hu H, Yang T, Hu Q, Yu Y, Huang Q. HR-ASPP: An improved semantic segmentation model of cervical nucleus images with accurate spatial localization and better shape feature extraction based on Deeplabv3+. In: Proceedings of the 15th International Conference on Digital Image Processing; 2023. p. 1–8.
4. Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. The Lancet. 2019; 393(10167):169–182. https://doi.org/10.1016/S0140-6736(18)32470-X
5. Franco EL, Schlecht NF, Saslow D. The epidemiology of cervical cancer. The Cancer Journal. 2003; 9 (5):348–359. https://doi.org/10.1097/00130404-200309000-00004 PMID: 14690309
6. Guo P, Banerjee K, Stanley RJ, Long R, Antani S, Thoma G, et al. Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification. IEEE journal of biomedical and health informatics. 2015; 20(6):1595–1607. https://doi.org/10.1109/JBHI.2015.2483318 PMID: 26529792
7. Phoulady HA, Zhou M, Goldgof DB, Hall LO, Mouton PR. Automatic quantification and classification of cervical cancer via adaptive nucleus shape modeling. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE; 2016. p. 2658–2662.
8. Banik PP, Saha R, Kim KD. An automatic nucleus segmentation and CNN model based classification method of white blood cell. Expert Systems with Applications. 2020; 149:113211. https://doi.org/10.1016/j.eswa.2020.113211
9. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020;.
10. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 10012–10022.
11. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. ACM computing surveys (CSUR). 2022; 54(10s):1–41. https://doi.org/10.1145/3505244
12. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence. 2022; 45(1):87–110. https://doi.org/10.1109/TPAMI.2022.3152247 PMID: 35180075
13. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 6836–6846.
14. Zhou D, Kang B, Jin X, Yang L, Lian X, Jiang Z, et al. Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:210311886. 2021;.
15. Dong X, Bao J, Chen D, Zhang W, Yu N, Yuan L, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 12124–12134.

16. Chen CFR, Fan Q, Panda R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 357–366.

17. Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 568–578.

18. Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 12104–12113.

19. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, et al. Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 764–773.

20. Zhu X, Hu H, Lin S, Dai J. Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 9308–9316.

21. Wang W, Dai J, Chen Z, Huang Z, Li Z, Zhu X, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 14408–14419.

22. Lu Z, Carneiro G, Bradley AP. Automated nucleus and cytoplasm segmentation of overlapping cervical cells. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16. Springer; 2013. p. 452–460.

23. Hu H, Zhang J, Yang T, Hu Q, Yu Y, Huang Q. CNAC-Seg: Effective segmentation for cervical nuclei in adherent cells and clusters via exploring gaps of receptive fields. Biomedical Signal Processing and Control. 2024; 90:105833. https://doi.org/10.1016/j.bspc.2023.105833

24. Hu H, Zhang J, Yang T, Hu Q, Yu Y, Huang Q. PATrans: Pixel-Adaptive Transformer for edge segmentation of cervical nuclei on small-scale datasets. Computers in Biology and Medicine. 2024; 168:107823. https://doi.org/10.1016/j.compbiomed.2023.107823 PMID: 38061155

25. Khan SD, Alarabi L, Basalamah S. An encoder–decoder deep learning framework for building footprints extraction from aerial imagery. Arabian Journal for Science and Engineering. 2023; 48(2):1273–1284. https://doi.org/10.1007/s13369-022-06768-8

26. Gillioz A, Casas J, Mugellini E, Abou Khaled O. Overview of the Transformer-based Models for NLP Tasks. In: 2020 15th Conference on Computer Science and Information Systems (FedCSIS). IEEE; 2020. p. 179–183.

27. Tetko IV, Karpov P, Van Deursen R, Godin G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nature communications. 2020; 11(1):5575. https://doi.org/10.1038/s41467-020-19266-y PMID: 33149154

28. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations; 2020. p. 38–45.

29. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:191003771. 2019;.

30. Tunstall L, Von Werra L, Wolf T. Natural language processing with transformers. "O'Reilly Media, Inc."; 2022.

31. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. PMLR; 2021. p. 10347–10357.

32. Xu Y, Zhang Q, Zhang J, Tao D. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. Advances in neural information processing systems. 2021; 34:28522–28535.

33. Zheng W, Lu S, Yang Y, Yin Z, Yin L. Lightweight transformer image feature extraction network. PeerJ Computer Science. 2024; 10:e1755. https://doi.org/10.7717/peerj-cs.1755

34. Hu C, Xia T, Cui Y, Zou Q, Wang Y, Xiao W, et al. Trustworthy multi-phase liver tumor segmentation via evidence-based uncertainty. Engineering Applications of Artificial Intelligence. 2024; 133:108289. https://doi.org/10.1016/j.engappai.2024.108289

35. Zhan G, Wang W, Sun H, Hou Y, Feng L. Auto-csc: a transfer learning based automatic cell segmentation and count framework. Cyborg and Bionic Systems. 2022;. https://doi.org/10.34133/2022/9842349 PMID: 36285314

36. Wu K, Peng H, Chen M, Fu J, Chao H. Rethinking and improving relative position encoding for vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 10033–10041.

37. Zhang P, Dai X, Yang J, Xiao B, Yuan L, Zhang L, et al. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 2998–3008.

38. Wang J, Wei L, Wang L, Zhou Q, Zhu L, Qin J. Boundary-aware transformers for skin lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer; 2021. p. 206–216.

39. Pu M, Huang Y, Liu Y, Guan Q, Ling H. Edter: Edge detection with transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 1402–1412.

40. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. Springer; 2022. p. 205–218.

41. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern recognition. 2020; 106:107404. https://doi.org/10.1016/j.patcog.2020.107404

42. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer; 2018. p. 3–11.

43. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:210204306. 2021;.