PAPER

Analysis of Blood Cell Image Recognition Methods Based on Improved CNN and Vision Transformer

Pingping WANG[†*], **Xinyi ZHANG**[††*a)], **Yuyan ZHAO**[†††*], **Yueti LI**[††††*], **Kaisheng XU**[†††††*], *and* **Shuaiyin ZHAO**[††††††*], ***Nonmembers***

SUMMARY Leukemia is a common and highly dangerous blood disease that requires early detection and treatment. Currently, the diagnosis of leukemia types mainly relies on the pathologist's morphological examination of blood cell images, which is a tedious and time-consuming process, and the diagnosis results are highly subjective and prone to misdiagnosis and missed diagnosis. This research suggests a blood cell image recognition technique based on an enhanced Vision Transformer to address these problems. Firstly, this paper incorporate convolutions with token embedding to replace the positional encoding which represent coarse spatial information. Then based on the Transformer's self-attention mechanism, this paper proposes a sparse attention module that can select identifying regions in the image, further enhancing the model's fine-grained feature expression capability. Finally, this paper uses a contrastive loss function to further increase the intra-class consistency and inter-class difference of classification features. According to experimental results, The model in this study has an identification accuracy of 92.49% on the Munich single-cell morphological dataset, which is an improvement of 1.41% over the baseline. And comparing with sota Swin transformer, this method still get greater performance. So our method has the potential to provide reference for clinical diagnosis by physicians.

key words: *vision transformer, CNN, self attention mechanisms, blood cell recognition, leukemia*

1. Introduction

Blood cells are the cellular components of blood that play important roles in carrying oxygen and nutrients throughout the body, fighting off infections, and removing waste products. They are classified into three main types: red blood cells, white blood cells, and platelets. White blood cells, also known as leukocytes, play a vital role in protecting the human body from infections. They are crucial for the immune system to function and are responsible for fighting against harmful bacteria, viruses, and other pathogens. The five main types of white blood cells, including neutrophils, basophils, eosinophils, monocytes, and lymphocytes. Identifying and categorizing white blood cells has always been a crucial step in blood analysis as it helps physicians anticipate significant illnesses and monitor the progress of treatment by monitoring the alteration in the quantity and form of various white blood cell types [1].

The traditional method is to stain a blood smear and then identify it under a light microscope. This method is extremely tedious and time-consuming, and the results of counting and sorting are susceptible to human influence [2]. As a result of the quick growth of computing power and artificial intelligence, deep learning technology [3]–[8] has been widely used and has greatly advanced image processing [9], [10]. Blood cell classification using deep learning is a type of medical image analysis that involves using advanced machine learning algorithms to automatically identify and classify different types of blood cells present in digital images of blood samples. This method may enhance the precision and effectiveness of blood cell classification and assist medical professionals in the diagnosis and treatment of various blood-related diseases and disorders, such as leukemia.

Recent years, in computer vision tasks including picture classification, object recognition, and segmentation, Vision Transformer (ViT) has achieved astounding results. Unlike traditional convolutional neural networks (CNNs), ViT uses a self-attention mechanism to learn and extract relevant features from input images, allowing it to attend to the most important parts of the image for the task at hand. The self-attention mechanism allows ViT to learn the relationships between different parts of the input image and to attend to the most relevant parts of the image for the task at hand. This is achieved through a mechanism called the multi-head self-attention, where the input image is divided into multiple patches, and attention is calculated between each patch and all the other patches. The attention weights are then used to weight the importance of each patch for the final classification. Given the exceptional results demonstrated by ViT in generic visual recognition and the increasing demand for automated blood cell image classification, this paper were motivated to research the ViT for blood cell classification task.

While numerous studies have focused on blood cell recognition, the majority are predicated on general object

Manuscript received May 17, 2023.
Manuscript revised August 14, 2023.
Manuscript publicized September 15, 2023.

[†]The author is with School of Computer Science and Engineering, Jishou University, Zhangjiajie, 427000, China.
[††]The author is with School of Economics and Management, North China University of Technology, Beijing 100144, China.
[†††]The author is with School of Economics, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China.
[††††]The author is with School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China.
[†††††]The author is with Faculty of Control Systems and Robotics, ITMO University, St. Petersburg 197101, Russia.
[††††††]The author is with HDU-ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, 310018, China.
[*]All authors contributed equally to this study.
a) E-mail: xinyizhanng@outlook.com (Corresponding author)
DOI: 10.1587/transfun.2023EAP1056

detection and classification networks, without specific improvements tailored to the characteristics of blood cells. Moreover, many studies have concentrated on major blood cell categories, neglecting subcategories such as stages of granulocytes. The subtle differences between these blood cell subcategories make their automatic recognition more challenging. Recently, the Vision Transformer has demonstrated impressive performance in visual classification tasks, indicating that the self-attention mechanism of the Transformer can capture critical parts of image block sequences, endowing the model with stronger local and global feature representation capabilities. Although a visual problem has been transformed into a sequence to sequence problem by patch embedding in ViT, medical image classification tends to focus on discriminative regions' local information. ViT is hard to focus on information regarding the discriminative regions.

The rationale behind the choice of using an enhanced Vision Transformer and Sparse Attention Module in this study is that these methods can better identify and classify blood cells. The enhanced Vision Transformer, by integrating convolutions with token embedding, supersedes the positional encoding that represents coarse spatial information, thereby augmenting the model's feature representation capabilities. The Sparse Attention Module can select identifying regions in the image, further enhancing the model's fine-grained feature expression capability. Additionally, the choice of employing a contrastive loss function is predicated on its ability to further increase the intra-class consistency and inter-class difference of classification features, thereby enhancing the model's classification performance.

To sum up, our contributions are two-fold:

- Sparse attention module, which makes comprehensive use of the attention weight information of all coding layers to capture the discriminative region in the image. For the purpose of resolving the blood cell picture intra-class variance and inter-class similarity problems, this is extremely significant.

- Convolutional Token Emebedding keeps all the positive attributes of Transformers—dynamic attention, universal context integration, and more gigantic generalization—while utilizing all of CNN's advantages—Local correlation, stationarity, and spatially subsampling.

In the remainder of this paper, we delve into the details of our proposed method and its underlying principles. Section 2 provides a comprehensive review of related work, highlighting the advancements and limitations of existing methods in the field of blood cell recognition. In Sect. 3, we elucidate the materials and methods used in our study, including a detailed explanation of the enhanced Vision Transformer and Sparse Attention Module, and the rationale behind their implementation. Section 4 presents the results of our experiments, providing a comparative analysis of our method with existing techniques. Finally, Sect. 5 concludes the paper with a summary of our findings and potential directions for future research.

Author Contributions are here:

Pingping Wang: Writing-Original Draft, Methodology, Software, and Conceptualization. Xinyi Zhang: Writing-Original Draft, Supervision and Visualization. Yuyan Zhao: Data process, Investigation and Validation. Yueti Li: Investigation, Resources and Data Curation. Kaisheng Xu: Writing-Original Draft, Writing-Review & Editing, Formal analysis and Visualization. Shuaiyin Zhao: Project administration, Writing- Reviewing and Editing, and Funding acquisition. Above authors contributed equally to this article.

2. Related Work

Although previous studies [11]–[13] have made significant progress in blood cell recognition, most of them are based on general object detection and classification networks, without improvements specifically targeting blood cell characteristics. Furthermore, many studies have only focused on major blood cell categories, without paying attention to subcategories such as the stages of granulocytes, including primitive, early, intermediate, and late stages. The subtle differences between blood cell subcategories make their automatic recognition more challenging. Recently, the Vision Transformer [14] has shown good performance in visual classification tasks, indicating that the self-attention mechanism of Transformer [15] can capture important parts of image block sequences, allowing the model to have stronger local and global feature representation capabilities. Therefore, this paper combines blood cell characteristics to study fine-grained classification of blood cells and proposes a blood cell recognition method based on an improved Vision Transformer.

2.1 Medical Image Classification Based on Deep Learning Methods

Deep learning has made significant progress in recent years, mainly due to increasing computer hardware and the huge amount of data available, as well as deep learning technology [16].

A fundamental stage in medical image analysis, medical image classification tries to separate medical images based on a certain criterion, such as clinical pathologies or imaging modalities. A trustworthy technique for classifying medical images can help clinicians evaluate medical images quickly and correctly.

Deep learning techniques, particularly especially deep convolutional neural networks (DCNN), have significantly advanced medical image classification in recent years [17], [18].

Dhieb et al. used a Mask Region-Based Convolutional Neural Network (Mask R-CNN) to detect red and white blood cells. The model employed Resnet-101 as its backbone and utilized a Feature Pyramid Network (FPN) to extract multi-scale features for detecting cells of different sizes. The method achieved an accuracy of 92% for red blood cell recognition and 96% for white blood cell recognition [19].

Shakarami et al. [11] proposed a Fast and Efficient

YOLOv3 Detector (FED) based on the YOLOv3 (You Look Only Once v3) single-stage object detection network. This model used Efficientnet as its backbone and performed blood cell detection at three different scales. The method achieved average recognition accuracy of 90.25%, 80.41%, and 98.92% for platelets, red blood cells, and white blood cells, respectively, on the BCCD dataset.

In the field of blood cell image classification, researchers have also conducted extensive studies. Matek et al. [12] released an open-source blood cell dataset containing 15 classes and a total of 18,375 images. They then used the ResNext model for classification, achieving a recognition accuracy of 94% for common blood cells such as neutrophils, lymphocytes, and monocytes.

Huang et al. [20] first obtained individual blood cell slice images based on the RetinaNet detection network, then introduced an adaptive attention module into the convolutional neural network. This module enhanced the weight of regional features related to classification tasks, improving the model's feature representation capabilities. The model achieved an average classification accuracy of 95.3% for six types of white blood cells.

Mori et al. [21] divided blood cells into four categories based on the degree of cytoplasmic granule reduction, and then used the Resnet-152 network for classification. The average sensitivity and specificity were 85.2% and 98.9%, respectively.

Numerous studies demonstrate that the DCNN approach is more precise than hand-crafted feature-based alternatives [22], [23]. However, they have not been as successful [24] as ImageNet challenge [25] on medical image classification [26], [27]. Researchers have sought to apply Transformer in computer vision since its remarkable success in natural language processing, and they then suggested Vision Transformer, which has shown excellent results [15]. Therefore, researchers are attempting to apply variations of VIT to address the challenges of intra-class variance and inter-class similarity in the classification of medical images [28].

2.2 Research Process of Vision Transformer

CNNs have demonstrate excellent performances in image classification. CNNs are a type of deep learning technique that automatically extract features from image data. The CNNs process an input image through multiple stages to extract hierarchical and sophisticated feature representations [29]. Using CNN can easily construct an end-to-end model and there's no requirement for designing intricate, manually designed features [30]. Recently, a innovative transformer architecture has resulted in a significant advancement in Natural Language Processing tasks. The Transformer, specifically created for sequence modeling and transaction tasks, stands out for its implementation of attention mechanisms to capture long-range dependencies in the data. Due to its remarkable success in the language field, researchers are exploring its potential applications in computer vision.

Unlike traditional CNNs, ViT replaces convolutional layers with self-attention mechanism, which allows the network to directly process the full-resolution image. This design decision results in the ViT model being capable of processing images of variable sizes while maintaining high accuracy. Dosovitskiy et al. [14] designed the first example of a transformer-based. The ViT model replaces the fixed receptive field of CNNs with a self-attention mechanism that allows it to process full-resolution images. The ViT model is an encoder-only architecture and removes the transformer decoder for computer vision classification tasks [31]. The traditional approach to processing images in CNNs is to apply convolutional operations to reduce the spatial resolution of the image, which can lead to the loss of important details and features. In contrast, the ViT model processes the full-resolution image and retains all of its details and features. This enables the ViT model to exhibit an impressive balance between speed and accuracy in image classification relative to convolutional networks [32]. While VIT model relies on large dataset to perform better. Touvron et al. [33] presents multiple training techniques that enable the ViT model to perform effectively even when using the smaller ImageNet-1K dataset.

In the ViT model, each image is first divided into a sequence of non-overlapping patches, and each patch is then embedded into a vector representation. These vector representations are then processed through the Transformer encoder, which computes self-attention scores between the patches to capture the relationships between them. This self-attention mechanism allows the ViT model to learn a global representation of the image, regardless of its size or aspect ratio. The ViT model's architecture consists of a stack of multiple Transformer encoder blocks, each consisting of multi-head self-attention and position-wise feedforward layers. The final output of the model is then passed through a classifier layer to make a prediction for the image classification task.

2.3 Introducing Convolutions to Transformers

In Natural Language Processing (NLP), the Transformer block has been modified using convolutions. and speech recognition, either by substituting convolution layers for multi-head attentions [34] or by adding more convolution layers concurrently [35] or sequentially [36]. As well, some earlier research suggests propagating attention mappings to following layers using a residual link that is first convolutionally processed [37]. This paper propose, in contrast to these works, to add convolutions to two key components of the vision Transformer: initially, to substitute our convolutional projection for the current Position-wise linear projection. Secondly, to use our hierarchical multi-stage structure to enable variable resolution of two-dimensional reshaped token maps. Compared to earlier designs, ours offers notable performance and efficiency advantages.

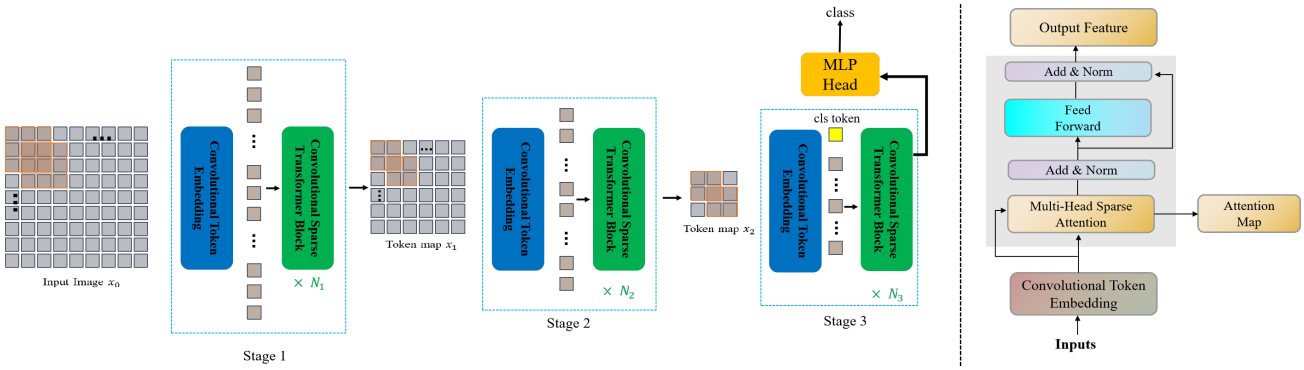


Fig. 1 This figure illustrates the pipeline for the proposed architecture. The overall architecture is shown in (a), which uses a convolutional token embedding layer to provide a hierarchical multi-stage structure. In (b), further information on the convolutional sparse transformer block.

3. Materials and Methods

This paper presents the Vision Transformer-based blood cell recognition network framework. This method add Convolutional Token Emebedding and Sparse Attention Module to Vision Transformer architecture. Meanwhile we use a multi-stage hierarchy design which is from CNNs [38], [39]. As shown in Fig. 1.

First put the image into the Convonlution Token Embedding layer, and it can be seen as Convolution of overlapping blocks of reconstructed Tokens into a 2D spatial grid as input. Then this paper add the another layer normalization to these tokens. Next, add a learnable classification vector. The token are subsequently fed into multiple stacked coding modules for feature extraction. Before the final layer of the coding module, a sparse attention module is used to find the distinguishing pixel blocks in the image and use their corresponding implicit features as input. Finally, The classification features output by the encoder are passed through the fully connected layer to obtain the class information of the blood cells.

3.1 Convolutional Token Embedding Layer

Traditional Vision Transformer [40] demonstrates the possibility of using pure Transformer structures in computer vision field. Nevertheless, it This layer uses a multi-stage hierarchical method, akin to CNNs, to represent local spatial contexts, ranging from basic edges to more complex semantic primitives. Essentially this paper assumed a two dimensions image or a reshaped output token map from the Earlier stages $x_{i-1} \in \mathbb{R}^{H_{i-1} \cdot W_{i-1} \cdot C_{i-1}}$ as for stage i's input. Which H_i represents height:

$$H_i = \frac{H_{i-1} + 2p - s}{s - o} + 1 \tag{1}$$

w_i represents the new token map's width:

$$W_i = \frac{W_{i-1} + 2p - s}{s - o} + 1 \tag{2}$$

This paper set a learnable function $f(\cdot)$ to map into tokens $f(x_{i-1})$ where $f(\cdot)$ is a 2-dimensional convolution operation with a kernel size of $s_i \times s_i$, stride $s - o$ and p padding. Padding is necessary to address boundary conditions. The new token map $f(x_{i-1}) \in \mathbb{R}^{H_{i-1} \cdot W_{i-1} \cdot C_{i-1}}$. Then $f(x_{i-1}) \in \mathbb{R}^{H_{i-1} \cdot W_{i-1} \cdot C_{i-1}}$ is flattened into size $H_i W_i \times C_i$ and normalized by layer normalization [41].

By changing the convolution operation's parameters, this layer enables us to modify the token feature dimension and the number of tokens at each step. So that, this method can gradually shorten the length of the token sequence while expanding the token dimension in each iteration. Similar to CNNs, It enable the tokens to represent sophisticated vision feature in the larger spatial grid [42].

3.2 Encoder

This paper has adopted a stacked encoder structure similar to the Vision Transformer. The structure of the module is shown in Fig. 2.

The module incorporates multi-head self-attention (MSA) and multi-layer perception (MLP). The MSA module consists of a single self-attention unit (SA) stitched together. For SA units, the input $I \in \mathbb{R}^{(N+1) \times D}$ is first transformed by the following formula Eqs. (3), (4) and (5) to obtain the query matrix Q, key matrix K, value matrix V. The formula is:

$$Q = I_p W^Q W^Q \in \mathbb{R}^D \times d_k \tag{3}$$

$$K = I_p W^K W^K \in \mathbb{R}^{D \times d_k} \tag{4}$$

$$V = I_p w^v W^v \in \mathbb{R}^{D \times d_k} \tag{5}$$

Where $d_k = \frac{D}{N_h}$, The attentional weighting matrix A and SA unit's output I' are calculated as follows:

$$\mathbf{A} = softmax\{\frac{QK^T}{\sqrt{d_k}}\}, \mathbb{A} \in \mathbb{R}^{(N+1) \times d_k} \tag{6}$$

$$I' = \mathbb{A} \cdot \mathbb{V}, I' \in \mathbb{R}^{(N+1) \times (N+1)} \tag{7}$$

Different SA unit learn relevant features in non-interfering and independent feature subspace. At last, MSA concat the

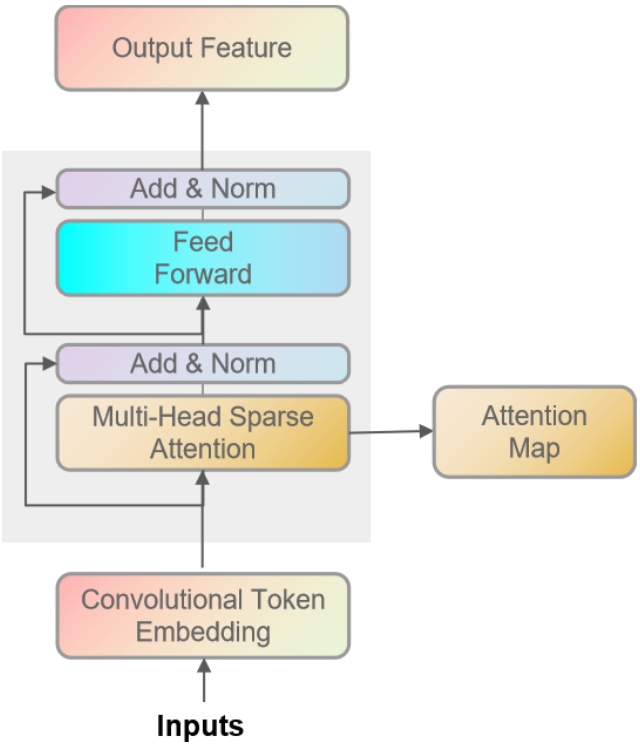


Fig. 2 Our model's encoder block architecture.

different SA units' output. After a linear transformation the output of the module is obtained. The output make residual connections with I_p, next go through layer normalization (LN) as the next MLP's input.

$$MSA(z_p) = concat(SA(z)^i)\mathbf{W}_{out} + \mathbf{b}_{out} \tag{8}$$

Where $\mathbf{W} \in \mathbb{R}^{(N+1)\times D}$ is the weight matrix, $\mathbf{b} \in \mathbb{R}^{(N+1)\times D}$ s the bias. MLP Module is composed of two Fully Connected Layer, the fist one used ReLu [43] as activation function and the second layer without activation function. The calculation formula is as follows:

$$MLP(\mathbf{X}) = ReLU(\mathbf{XW_1} + \mathbf{b_1}) \cdot \mathbf{W_2} + \mathbf{b_2} \tag{9}$$

The output of the module is shown in the Eqs. (10) and (11):

$$\mathbf{I'_p} = LN(MSA(\mathbf{I}_{p-1} + \mathbf{I}_{p-1})) \tag{10}$$

$$\mathbf{I}_P = LN(MSA(\mathbf{I}_{p-1} + \mathbf{I}_{p-1})) \tag{11}$$

were $\mathbf{I}_{p-1}$ is the input of pth coding module.

3.3 Sparse Attention Module

The capacity to precisely find the discriminative regions is the main challenge in blood cell classification. Take the granulocyte in Fig. 3 as an example. The nucleus of lobulated neutrophils is lobulated, and the cytoplasm is evenly distributed with many light red special particles (about 80% of the total particles) and a small amount of light purple azurophilic particles (about 20% of the total particles) The nucleus of eosinophils is mainly composed of two leaves of

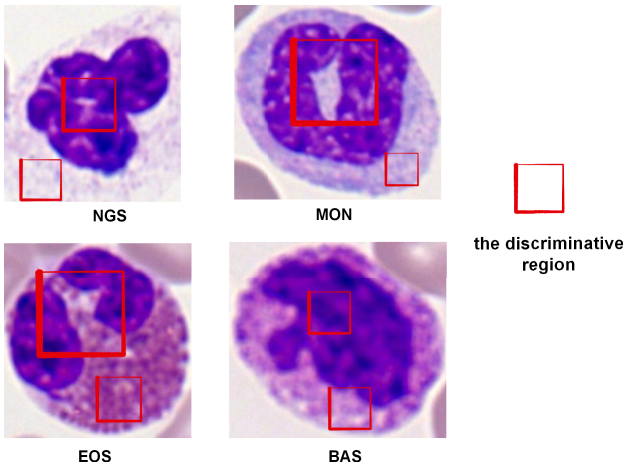


Fig. 3 The discriminative region.

cytoplasm filled with uniformly distributed and thick orange-red eosinophils. The nucleus of basophils is s-shaped, lobulated or irregular, and the color is light; The cytoplasm contains special basophilic particles with unequal size and uneven distribution. The nucleus of mononuclear cells is kidney shaped, horseshoe shaped or twisted and folded irregularly, the chromatin particles are fine and loose, and the color is light; More cytoplasm, weakly basophilic, gray-blue, containing lavender azurophilic particles.

In the Vision Transformer model, the multi-head self-attention mechanism can autonomously learn the weights of different image blocks. In this research, a sparse attention module is presented to fully utilize this weight information for the localization of discriminative regions.

It is assumed that the Vision Transformer network has n coding modules. The sparse attention module filters the hidden features input $I_{L-1} = [I^1_{L-1}; I^2_{L-1}; \cdots I^N_{L-1};]$ from the last coding layer using the weight learned from the first N-1 coding layers. The weights learned by the first N − 1 coding layers are shown in Equation Eqs. (12) and (13) :

$$\mathbf{A}_l = [\mathbf{A}^1_l, \mathbf{A}^2_l, \cdots \mathbf{A}^{N_h}_l] l \in 1, 2 \cdots, L-1 \tag{12}$$

$$\mathbf{A}^i_l = [\mathbf{a}^{class}_l; \mathbf{a}^1_l; \mathbf{a}^2_l; \cdots; \mathbf{a}^N_l]\mathbf{i} \in 1, 2 \cdots, n_h, \mathbf{a}^j_l \in R^{l\times(N+1)} \tag{13}$$

High layer feature attention maps do not accurately reflect the significance of the corresponding input picture blocks. Therefore, this paper combine the attentional map information from all previous encoding modules with the compressed excitation module to learn the weights of each attentional map autonomously. The sparse attention model is shown in Fig. 4.

After the global average pooling of attention graphs, the module uses two fully connected layers to model the correlation between attention graphs, and obtains the weight value a of each attention graph. Then the weight value is normalized and weighted with the attention diagram to get the final attention weight A_{attn}.

$$\mathbf{A}_{attn} = \sum_{i=1}^{L-1} \alpha_i \mathbf{A}_i \tag{14}$$

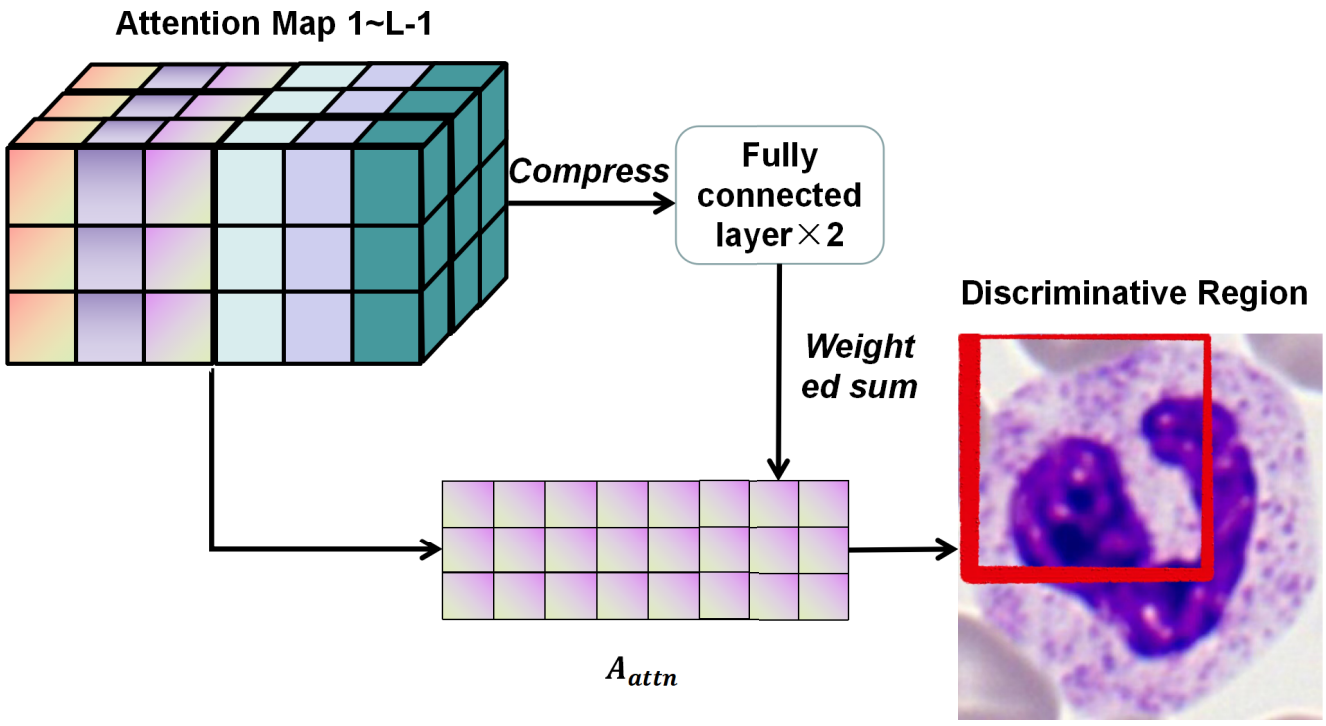


Fig. 4 Sparse attention module.

$\mathbf{A}_{attn}$ contains all the attention weight information of low-level features and high-level features $\mathbf{A}_{L-1}$, which is more suitable for screening identification regions than single-layer attention weight $\mathbf{A}_{L-1}$. We utilize the weights corresponding to the classification vectors $\mathbf{A}_{attn}^{class} = [\mathbf{a}_{final}^{1}, \mathbf{a}_{final}^{2} \cdots, \mathbf{a}_{final}^{N}]$ in $\mathbf{A}_{attn}$ to fliter out the implied features corresponding to the largest weights among the $n_h$self-attentive heads. These hidden features are finally combined with classification vectors as the input of the final layer of coding module.

$$\mathbf{I}_{\mathbf{L-1}}^{\mathbf{attn}} = [\mathbf{I}_{\mathbf{L-1}}^{\mathbf{class}}; \mathbf{I}_{\mathbf{L-1}}^{\mathbf{a_1}}; \mathbf{I}_{\mathbf{L-1}}^{\mathbf{a_2}}; \cdots \mathbf{I}_{\mathbf{L-1}}^{\mathbf{a_{N_h}}}] \tag{15}$$

The last coding module receives the result of the sparse attention module, which has replaced all sequence vectors with feature vectors corresponding to the identification region.

3.4 Loss Function

The network's The loss function consists of a cross-entropy loss L_{cross} and a contrast loss L_{con}. As shown following:

$$L = L_{cross}(\mathbf{y}, \mathbf{y}^{'}) + L_{con}(\mathbf{z}) \tag{16}$$

Cross-entropy loss is used to measure the similarity of the true labels y to the network predicted labels $y^{'}$. The definition is shown following.

This paper added contrast loss L_{con} to further increase the intra-class similarity and inter-class variability of network extracted features. Contrast loss minimizes the similarity of classification features corresponding to different labels and maximizes the similarity of classification features with the same label. To prevent losses from being dominated by different classes of features with little similarity, this paper introduce a threshold t_{con}. Only the similarity of features of different categories of samples is greater than t_{con}, it would be included in loss. $\mathbf{N}$ is the batch size of the input data. The contrast loss is defined as follows:

$$L_{con} = \frac{1}{N^2} \sum_{i=1}^{N} \left[\sum_{j:y_i=y_j} \left(1 - \frac{\mathbf{z_i} \mathbf{z_j}}{|\mathbf{z_i}||\mathbf{z_j}|} \right) + \sum_{j:y_i=y_j} max \left(\frac{\mathbf{z_i} \mathbf{z_j}}{|\mathbf{z_i}||\mathbf{z_j}|} \right) \right] \tag{17}$$

Table 1 Position embedding ablations experiments.

Model	Pos. Emb	ImageNet Top-1
Our model	Every stage	81.0
Our model	First stage	80.9
Our model	Last stage	80.9
Our model	N/A	81.1

Table 2 Contrastive model's position embedding ablations experiments.

Model	Pos. Emb	ImageNet Top-1
DeiT	Applicable	78.7
DeiT	N/A	77.1

4. Experiments

4.1 Dataset

This article uses the open source Munich AML Morphology Dataset (TMAMD) [44] on The Cancer Imaging Archive

Table 3 Ablation study on contrastive loss.

Model	Contrastive loss	Precision
Vision Transformer	Applicable	91.08
Vision Transformer	N/A	90.79
Our model(only sparse attention module)	Applicable	91.62
Our model(only sparse attention module)	N/A	91.28

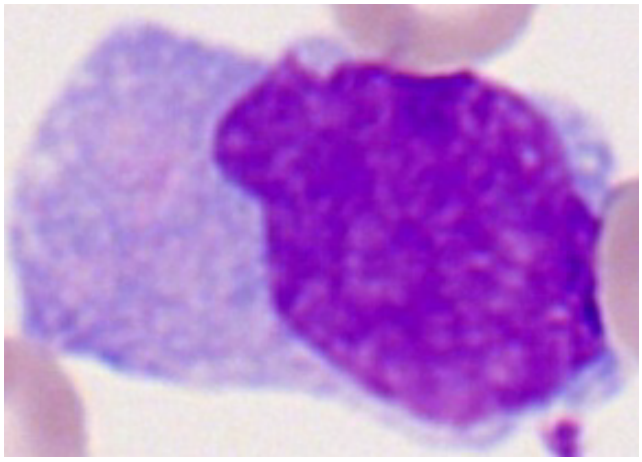


Fig. 5 This is a image of a single cell.

Table 4 The distribution of blood cell images which this paper have selected.

Blood cell type	Number of images	Data Augmentation
NGS	8484	1000
NGB	109	545
LYT	3937	1000
MON	1789	1000
EOS	424	848
BAS	79	395
MYO	3268	1000
PMO	70	350
MYB	42	210
EBO	78	390
Total	18280	6738

platform. The dataset contains 18,635 expert-labelled images of 15 categories of single-cell images. Some multicellular images and mature red blood cell images are present in the original dataset. Even though mature red blood cell are not the main concern in this paper. The above factors can lead to degradation of network classification performance. Therefore, this paper manually crop and screen the images. The image had processed is shown in the Fig. 5.

Considering the practicalities of clinical diagnosis and the small amount of data in some categories of the dataset. This paper selected 10 classes of red blood cells for the classification task.

4.2 Experimental Environment and Parameter Configuration

This paper adjusted the single-cell image size to 224×224. In the experiments, we adapted three stages. The convolutional token embedding layer's parameter is shown in Table 5. The

Table 5 Architectures for classification. Conv. Embed.: convolutional token embedding. Conv. Proj.: convolutional projection. $\mathbf{H}_i$ and $\mathbf{D}_i$ is the number of heads and embedding feature dimension in the i_{th} MHSA module. R_i is the feature dimension expansion ration in the i_{th} MLP layer.

	Output Size	Layer Name	
Stage 1	56×56	Conv. Embed.	7×7,64,stride 4
	56×56	Conv. Proj. MHSA MLP	3×3,64 H=1,D=64 ×1 R=4
Stage 2	28×28	Conv. Embed.	3×3,192,stide 2
	28×28	Conv. Proj. MHSA MLP	3×3,192 H=2,D=192 ×2 R=4
Stage 3	14×14	Conv. Embed.	3×3,384,stride 2
	14×14	Conv. Proj. MHSA MLP	3×3,384 H=6,D=384 ×10 R=4
Head	1×1	Linear	1000

threshold value t_{con} in Eq. (17) is 0.4, batch size is set to 32. We trained the model on an NVIDIA GeForce RTX 3090 graphics card in Windows 11. The version of deep learing frame is pytorch 1.10.1. We use ViT's technique for fine-tuning. The tuning is done with an SGD optimizor with a momentum of 0.9. The weight attenuation is set to $5e-4$ and the learning rate is initialized to 0.001, at the 40th, 70th, and 90th epoch, it becomes 1/10 of the original one. The whole training process stops at the 100th epoch.

This paper used 5-fold cross validation and evaluation indexes such as Precision, Recall and Accuracy to quantitatively estimate the performance of the classification algorithm. The definitions are shown in Eqs. (18), (19) and (20).

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

TP is the number of positive samples correctly predicted as positive class;

FP is the number of negative samples wrongly predicted to be positive;

TN is the number of negative samples correctly predicted as negative class;

FN is the number of positive samples of the wrong predicted negative class.

While our model was trained and tested using the Munich single-cell morphology dataset, the data settings used in this study are not exclusive to this dataset. The model's architecture and training parameters are designed to be adaptable to various types of blood cell recognition tasks. However, for different datasets or tasks, some adjustments may be necessary to achieve optimal performance.

4.3 Ablation Study

First, to study whether position embedding is still needed

Table 6 The performance of different methods.

Blood cell type	Precision(%)	Recall rate(%)	The number of testing images
NGS	94.32	94.29	200
NGB	92.71	92.98	110
LYT	95.78	96.75	200
MON	87.82	94.81	200
EOS	98.88	97.81	170
BAS	91.46	84.50	80
MYO	92.50	92.50	200
PMO	78.92	86.91	70
MYB	92.89	56.67	45
EBO	97.32	97.65	80
Total			1355

Table 7 The performance of different methods.

Method	Backone network	Precision(%)
VGG	VGG16	88.54
ResNet	ResNet50	88.71
ResNet	ResNet152	88.97
SENet	SE-ResNet50	89.56
SENet	SE-ResNet101	89.98
EfficientNet	EfficientNet-B0	90.27
Vision Transformer(2020)	vit-base-p16	91.08
Swin Transformer(2021)	vit-base-p16	92.09
TVT(2022)	t2t-vit-14	91.88
Our model(only Convolutional Token Embedding)	vit-base-p16	91.73
Our model(only sparse attention module)	vit-base-cell-p16	91.62
Our model	vit-base-cell-p16	92.49

for our model. This paper performed the ablation experiments on ImageNet [45] as shown in Table 1. Experimental results show that eliminating position embedding from different stages has little effect on the model effect. It can demonstrate that position embedding can be eliminated from the model with the use of convolutions.

Comparatively, eliminating DeiT's position embedding in Table 2 would result in a 1.6% reduction in ImageNet Top-1 accuracy. Because other than by adding the position embedding, it does not model the spatial connections between images. This demonstrates much more the efficiency of the convolutions we added.

Then, in order to prove the effectiveness of sparse attention module and contrast loss function, this paper conducted ablation experiments in Table 3. Experimental results show that by adding contrast loss, the recognition accuracy of Vision Transformer is improved by 0.29%, and the recognition accuracy of the model in this paper is improved by 0.34%

To sum up, we believe that our model can effectively expand the feature distance between similar subcategories and reduce the feature distance between the same categories, so as to improve the recognition performance of the model.

4.4 Network Performance Comparison Experiments

Different classes's accuracy rates and recall rates in the TMAMD data set are shown in Table 6. For the most prevalent types of blood cells, we found that the predicted results of the model were in good agreement with the doctor's annotation, and the accuracy rate and recall rate were both higher than 90%. Nevertheless, other categories' results are not quite ideal. It can be tolerated due to the small original sample size.

Additional, the baseline model is VGG [46], ResNet [38], EfficientNet [47], SENet [48] and Vision Transformer [14]. The results could be see in Table 7.

Meanwhile, comparing with the Tokens-to-Token ViT (TVT) [49] and Swin Transformer [32], the models' performance in this paper exceed 0.61% and 0.40%, respectively. The experiments strongly prove that the performance of the proposed model on the TMAMD dataset has been greatly improved.

5. Conclusions

Our study introduces an innovative method for blood cell recognition, which employs an enhanced Vision Transformer and Sparse Attention Module. This unique combination allows our model to effectively capture identifying regions in the image and enhance the model's fine-grained feature expression capability. Furthermore, our method utilizes a contrastive loss function to increase the intra-class consistency and inter-class difference of classification features, thereby enhancing the model's classification performance. Specifically, our model achieved an accuracy of 92.49% on the Munich single-cell morphology dataset, an improvement of 1.41% over the baseline. This approach represents a significant improvement over existing blood cell recognition methods, highlighting the novelty and superiority of our method.

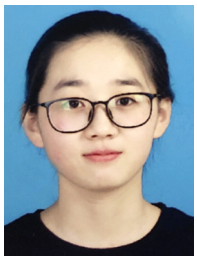
Despite the promising results, our study has certain limitations. Our model, while effective, may require larger

datasets for further performance improvement. The model's ability to handle certain types of blood cells may also be a potential area for improvement. Additionally, the computational resources required by our model may limit its applicability in resource-constrained environments. Future work will aim to address these limitations and further refine our model.

References

[1] X. Yao, K. Sun, X. Bu, C. Zhao, and Y. Jin, "Classification of white blood cells using weighted optimized deformable convolutional neural networks," Artificial Cells, Nanomedicine, and Biotechnology, vol.49, no.1, pp.147–155, 2021.
[2] Y. Duan, J. Wang, M. Hu, M. Zhou, Q. Li, L. Sun, S. Qiu, and Y. Wang, "Leukocyte classification based on spatial and spectral features of microscopic hyperspectral images," Optics & Laser Technology, vol.112, pp.530–538, 2019.
[3] M.I. Uddin, S.A. Ali Shah, M.A. Al-Khasawneh, A.A. Alarood, and E. Alsolami, "Optimal policy learning for COVID-19 prevention using reinforcement learning," J. Inform. Sci., vol.48, no.3, pp.336–348, 2022.
[4] F. Aziz, H. Gul, I. Uddin, and G.V. Gkoutos, "Path-based extensions of local link prediction methods for complex networks," Sci. Rep., vol.10, no.1, p.19848, 2020.
[5] I. Ullah, N.U. Amin, A. Almogren, M.A. Khan, M.I. Uddin, and Q. Hua, "A lightweight and secured certificate-based proxy signcryption (CB-PS) scheme for e-prescription systems," IEEE Access, vol.8, pp.199197–199212, 2020.
[6] Z. Ullah, A. Zeb, I. Ullah, K.M. Awan, Y. Saeed, M.I. Uddin, M.A. Al-Khasawneh, M. Mahmoud, and M. Zareei, "Certificateless proxy reencryption scheme (CPRES) based on hyperelliptic curve for access control in content-centric network (CCN)," Mobile Information Systems, vol.2020, pp.1–13, 2020.
[7] Z. Song and L. Wan, "Research of chinese relation extraction based on BERT," 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), pp.841–845, IEEE, 2023.
[8] L. Yang, Y. Li, S.X. Yang, Y. Lu, T. Guo, and K. Yu, "Generative adversarial learning for intelligent trust management in 6G wireless networks," IEEE Netw., vol.36, no.4, pp.134–140, 2022.
[9] J. He, S.L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," Nature Medicine, vol.25, no.1, pp.30–36, 2019.
[10] M.I. Jordan and T.M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol.349, no.6245, pp.255–260, 2015.
[11] A. Shakarami, M.B. Menhaj, A. Mahdavi-Hormat, and H. Tarrah, "A fast and yet efficient YOLOv3 for blood cell detection," Biomedical Signal Processing and Control, vol.66, p.102495, Feb. 2021.
[12] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks," Nat. Mach. Intell., vol.1, no.11, pp.538–544, Nov. 2019.
[13] X. Fu, M. Fu, Q. Li, X. Peng, J. Lu, F. Fang, and M. Chen, "Morphogo: An automatic bone marrow cell classification system on digital images analyzed by artificial intelligence," Acta Cytologica, vol.64, no.6, pp.588–596, July 2020.
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," Learning, 2020.
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.
[16] W. Wang, D. Liang, Q. Chen, Y. Iwamoto, X.H. Han, Q. Zhang, H. Hu, L. Lin, and Y.W. Chen, "Medical image classification using deep learning," Deep Learning in Healthcare: Paradigms and Applications, vol.171, pp.33–51, 2020.
[17] R. Li, T. Zeng, H. Peng, and S. Ji, "Deep learning segmentation of optical microscopy images improves 3-D neuron reconstruction," IEEE Trans. Med. Imag., vol.36, no.7, pp.1533–1541, March 2017.
[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, V. Vanhoucke, A. Rabinovich, and D. Erhan, "Going deeper with convolutions."
[19] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "An automated blood cells counting and classification framework using mask R-CNN deep learning model," 2019 31st International Conference on Microelectronics (ICM), March 2020.
[20] P. Huang, J. Wang, J. Zhang, Y. Shen, C. Liu, W. Song, S. Wu, Y. Zuo, Z. Lu, and D. Li, "Attention-aware residual network based manifold learning for white blood cells classification," IEEE J. Biomed. Health Inform., vol.25, no.4, pp.1206–1214, July 2020.
[21] J. Mori, S. Kaji, H. Kawai, S. Kida, M. Tsubokura, M. Fukatsu, K. Harada, H. Noji, T. Ikezoe, and T. Maeda, "Assessment of dysplasia in bone marrow smear with convolutional neural network," Sci. Rep., vol.10, no.1, pp.1–8, 2020.
[22] S. Koitka and C. Friedrich, "Traditional feature engineering and deep learning approaches at medical classification task of ImageCLEF 2016," 2016.
[23] S.H. Baloch and H. Krim, "Flexible skew-symmetric shape model for shape representation, classification, and sampling," IEEE Trans. Image Process., vol.16, no.2, pp.317–328, Feb. 2007.
[24] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Medical image classification using synergic deep learning," Medical Image Analysis, vol.54, pp.10–19, Feb. 2019.
[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," 2012.
[26] K. Sirinukunwattana, S.E.A. Raza, Y.W. Tsang, D.R.J. Snead, I.A. Cree, and N.M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," IEEE Trans. Med. Imag., vol.35, no.5, pp.1196–1206, Feb. 2016.
[27] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," IEEE Trans. Med. Imag., vol.36, no.3, pp.849–858, Dec 2016.
[28] Y. Song, W. Cai, H. Huang, Y. Zhou, D.D. Feng, Y. Wang, M.J. Fulham, and M. Chen, "Large margin local estimate with applications to medical image classification," IEEE Trans. Med. Imag., vol.34, no.6, pp.1362–1377, Jan. 2015.
[29] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "HEp-2 cell image classification with deep convolutional neural networks," IEEE J. Biomed. Health Inform., vol.21, no.2, pp.416–428, 2016.
[30] Q. Huang, W. Li, B. Zhang, Q. Li, R. Tao, and N.H. Lovell, "Blood cell classification based on hyperspectral imaging with modulated Gabor and CNN," IEEE J. Biomed. Health Inform., vol.24, no.1, pp.160–170, 2019.
[31] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," IEEE Trans. Geosci. Remote Sens., vol.60, pp.1–12, 2022.
[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," Proc. IEEE/CVF International Conference on Computer Vision, pp.10012–10022, 2021.
[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," International Conference on Machine Learning, pp.10347–10357, 2021.
[34] F. Wu, A. Fan, A. Baevski, Y.N. Dauphin, and M. Auli, "Pay less attention with lightweight and dynamic convolutions," International Conference on Learning Representations, 2019.
[35] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with

long-short range attention," arXiv: Computation and Language, arXiv:2004.11886, 2020.
[36] A. Gulati, J. Qin, C.C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," Proc. Interspeech 2020, pp.5036–5040, 2020.
[37] Y. Wang, Y. Yang, J. Bai, M. Zhang, J. Bai, J. Yu, C. Zhang, G. Huang, and Y. Tong, "Evolving attention with residual convolutions," arXiv: Learning, arXiv:2102.12895, 2021.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Computer Vision and Pattern Recognition, 2015.
[39] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," Shape, Contour and Grouping in Computer Vision, pp.319–345, 1999.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, arXiv:2010.11929, 2020.
[41] J. Ba, J.R. Kiros, and G.E. Hinton, "Layer normalization," arXiv: Machine Learning, arXiv:1607.06450, 2016.
[42] H. Wu, B. Xiao, N.C.F. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," arXiv: Computer Vision and Pattern Recognition, arXiv:2103.15808, 2021.
[43] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," Proc. Fourteenth International Conference on Artificial Intelligence and Statistics, pp.315–323, 2011.
[44] C. Matek, S. Schwarz, C. Marr, and K. Spiekermann, "A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls (AML-Cytomorphology_LMU)," The Cancer Imaging Archive (TCIA) [Internet], 2019.
[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," Int. J. Comput. Vis., vol.115, pp.211–252, 2015.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations, 2015.
[47] M. Tan and Q.V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019.
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7132–7141, 2018.
[49] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.



Pingping Wang received the B.S. degree in School of Computer Science and Engineering, Jishou University, zhangjiajie, China. She is the intern of Dream lab. Her research interests is medical image processing.

Xinyi Zhang received the B.S. degree in School of Economics and Management from North China University of Technology, Beijing China. She Will study in Tianjin University. Her research interests include data mining and quantitative trading.

Yuyan Zhao received the B.S. degree in School of Economics, Northeastern University at Qinhuangdao. The CEO of Qinhuangdao Chen Yan Energy Technology Co., LTD. Her research interest interests is quantitative trading.

Yueti Li received the B.S. degree in School of Control Engineering, Northeastern University at Qinhuangdao. Her research interest is deep learning.

Kaisheng Xu received the B.S. from the H DU-ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, China. He is now studying for a master's degree at Faculty of Control Systems and Robotics, ITMO University. His research interests are Robot control and computer vision.

Shuaiyin Zhao received the B.S. from the DU-ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, China. His research interest is computer vision.