Biomedical Signal Processing and Control 82 (2023) 104518

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Deep learning detection network for peripheral blood leukocytes based on improved detection transformer

Bing Leng [a,b], Chunqing Wang [c], Min Leng [d], Mingfeng Ge [a,b,*], Wenfei Dong [a,b,*]

[a] School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, China
[b] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, China
[c] Department of Clinical Laboratory Medicine, The First Affliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, China
[d] Liaocheng Cardiac Hospital, China

ARTICLE INFO

Keywords:
improved DETR
Deep learning
Leukocyte detection
Convolutional neural network
Transformer

ABSTRACT

Leukocytes serve as an important barrier to healthy immunity in the body and play an important role in fighting diseases. Manual morphological examination of leukocytes is the gold standard for the diagnosis of certain diseases but is undoubtedly labour-intensive and requires a high level of expertise. Therefore, conducting research on computer-aided diagnostics is important. With the development of deep learning techniques in computer vision, an increasing number of deep learning-based methods are now being applied in the field of medical imaging. Recently, the detection transformer (DETR) model, which is based on the transformer architecture, has exhibited outstanding performances in object detection tasks and has attracted considerable attention. Our study aims to propose a pure transformer-based end-to-end object detection network based on DETR and apply it to a practical medical scenario of leukocyte detection. First, we introduce the pyramid vision transformer and deformable attention module into the DETR model to improve the model performance and convergence speed. Second, we train the improved model on the challenging Common Objects in Context dataset to obtain the pretrained weights. Third, we perform transfer learning on the modified Raabin leukocyte dataset to obtain the optimal model. The improved DETR shows a mean average precision detection performance of up to 0.961 and is therefore superior to the original DETR and convolutional neural network. The study findings are expected to be useful for the development of a transformer structural model for leukocyte detection.

1. Introduction

Leukocytes account for only 1 % of the peripheral blood content but are a key component of the immune system [1]. Leukocytes can actively deform and move. They travel across blood vessels to participate in immune response and protect the organism from bacteria, viruses, and immune diseases.

As spherical blood cells, leukocytes are usually in a colourless nucleated state. The morphology of leukocytes can be observed by creating stained and dispersed thin blood smears [2], which can be divided into two categories based on morphological differences: granular and nongranular. After staining, granular leukocytes can be distinguished as neutrophils, eosinophils, and basophils, whilst nongranular leukocytes can be divided into monocytes and lymphocytes. Fig. 1 shows the morphologies of the five types of leukocytes, when observed under a microscope. Normally, the percentages of each type of cell in leukocytes are as follows: 50–70 % for neutrophils, 20–40 % for lymphocytes, 3–8 % for monocytes, 0.5–5 % for eosinophils, and 0.5–1 % for basophils [3].

The occurrence of diseases in the human body can cause the number and morphology of leukocytes to change. Leukocyte testing plays a crucial role in the diagnosis or treatment of infectious, leukaemia, lymphoma, anaemia, dysplasia, and other diseases [4–5]. In a clinical diagnosis, physicians initially provide a rough judgment by the cell count in the patient's peripheral blood test, and then analyse the condition by the morphological characteristics of the nucleus and cytoplasm of the cells.

The haematocrit analyser—the most widely used automated leukocyte testing equipment in hospitals—is used to collect blood in a test tube for leukocyte counting. It improves the efficiency and reduce the stress of the laboratory staff. However, this automated equipment, which is based on the principles of light scattering or electrical

* Corresponding authors.
E-mail addresses: gemf@sibet.ac.cn (M. Ge), wenfeidong@sibet.ac.cn (W. Dong).

https://doi.org/10.1016/j.bspc.2022.104518
Received 23 July 2022; Received in revised form 12 October 2022; Accepted 18 December 2022
Available online 27 December 2022

impedance, does not enable the analysis of cell morphology. Further manual morphological confirmation of abnormal cells via the creation of peripheral blood smears is often required, as shown in Fig. 2(a). As the gold standard, the morphologic microscopy of leukocytes aids in diagnosing diseases [6].

However, manual morphological microscopy requires using a microscope for a labour-intensive visual inspection of blood smears. This technique may cause damage to the vision, shoulder, and neck of the examiner. Additionally, the professionalism of the examiners is associated with the accuracy of the results obtained using this equipment, as examiners are susceptible to bias [7]. In ordinary hospitals, this gold standard is not sufficiently appreciated because of the lack of testing manpower and associated low cost.

In recent years, computer-aided diagnostic (CAD) techniques have been rapidly developed in the field of digital haematology for leukocyte detection. By connecting a microscope, camera, and computer together to obtain microscopic images of leukocytes can be used to realize a low-cost, simple, and efficient approach. Moreover, the acquired images can be processed remotely, eliminating the need for blood sample storage and transportation [8]. The digital storage of blood image information can provide a traceable historical information for the clinics and improve the efficiency of diagnosis [9]. Therefore, developing an automated leukocyte digital image recognition device for peripheral blood based on cell morphology tests is important to achieve high accuracy and specificity [10].

The CAD methods are divided into two categories: traditional image processing methods and deep-learning (DL) methods. The former methods generally involves complex steps: image pre-processing, cell segmentation, feature extraction, feature selection, and cell classification. Putzu et al. converted microscopic images into the cyan, magenta, yellow, and key (CMYK) mode and implemented threshold-based segmentation using the Zack algorithm. They then imported the shape, colour, and texture features of the cells in the sub-images into a support vector machine (SVM) for classification [11]. In addition, industrial automated cell morphology analysis systems have been applied in clinical settings, such as the CellaVision DM96 (CellaVision Inc., https://www.cellavision.com). The CellaVision DM96 was developed based on traditional image processing techniques. The accuracy of the classification needs to be improved [12].

With the wide application of DL in downstream computer vision (CV) tasks—e.g., classification, detection, segmentation—an increasing number of DL methods are achieving good results in the task of leukocyte recognition in microscopic images. Based on the model structure, object detection models can be classified into two types based on convolutional neural network (CNN) and a transformer. For information extraction and processing, the CNN uses convolutional operations, whereas a Transformer uses a self-attentive mechanism.

Object detection networks can be divided into one-stage and two-stage networks, both of which have both advantages and disadvantages. One-stage network is the you only look once (YOLO) v3 [13], which directly classifies candidate boxes and performs regression on the positions of the boxes, can perform quick detection but with a low accuracy. Two-stage networks such as the faster R-CNN (region-based CNN) classifies and regresses with and without objects on the proposed frame using the region proposal network and then performs object multiclassification and frame regression [14]. A high detection accuracy can be achieved using this method, but its speed is low.

Kutlu et al. used Faster R-CNN for the detection of five classes of leukocytes and achieved good detection results on the public dataset LISC, with a mean average precision (mAP) = 0.74 [15]. Reena et al. used the DeepLabv3 + model to segment the leukocytes, and then imported them into AlexNet for classification [16]. Fan et al. proposed a Mask R-CNN-like model for leukocyte detection using a modified FPN network to extract feature maps and RPN to obtain region proposals [17]. Hung et al. proposed a Faster R-CNN-based Python package, Keras R-CNN, and validated the recognition capability of the package on cell nucleus detection and malaria classification tasks [18]. Inspired by R-CNN, Di et al. proposed a region proposal method for edge boxes that introduced knowledge-based constraints into the leukocyte detection process combined with nonmaximal suppression [19]. Wang et al. applied SSD and YOLO v3 networks to leukocyte detection using transfer learning and training the model on a dataset. In their study, SSD achieved a better detection result (mAP = 0.93) [10]. Khandekar et al. used YOLO v4 for the detection of both blast and healthy cells in blood smear microscopic images for the diagnosis of acute lymphoid leukaemia with mAP up to 0.96 [20].

Transformer-based models have been used widely in natural language processing. The vision transformer (ViT) model [21] proposed by Google showed the application prospect of transformers in the field of CV and set off extensive research on transformers. Some transformer-based network backbones, such as a Swin transformer [22], have been proposed and combined with the classical detection CNN to achieve promising detection effects. Facebook proposed a transformer-based end-to-end object detection network, DETR [23], which surpassed the performance of the classical CNN detection network faster R-CNN. The model combines bipartite matching and a transformer and views the task of object detection as direct set prediction, thereby eliminating the nonmaximal suppression and anchor generation of the manual design process.

Several improvements based on DETR have been successively proposed. The deformable DETR proposed by SenseTime can fuse multi-scale features to enhance small-object detection. The training process can converge quickly [24]. Dai et al. proposed UP-DETR, which can pretrain DETR models in an unsupervised manner using a pretext task to improve the detection performance of DETR [25]. In terms of applied research on DETR, Prangemeier et al. proposed Cell-DETR for instance segmentation of cells in microscopic images, which is comparable to the advanced Mask R-CNN [26].

In summary, traditional methods often have the disadvantages of complex steps and low accuracy. Despite DL-based leukocyte detection methods having achieved good results, the following problems still exist: (a) The CNN-based one-stage or two-stage detection methods are inseparable from the manual design steps, with the performance affected by factors such as postprocessing, design of proposal, and ground truth assignment. (b) The DETR model, as a newly proposed Transformer-based end-to-end set prediction model, has the shortcomings of high

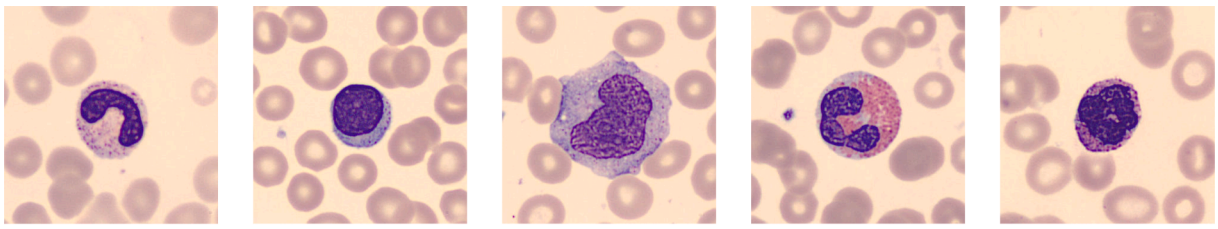

Fig. 1. Morphologies of the five types of leukocytes: (a) neutrophil; (b) lymphocyte; (c) monocyte; (d) eosinophil; (e) basophil.

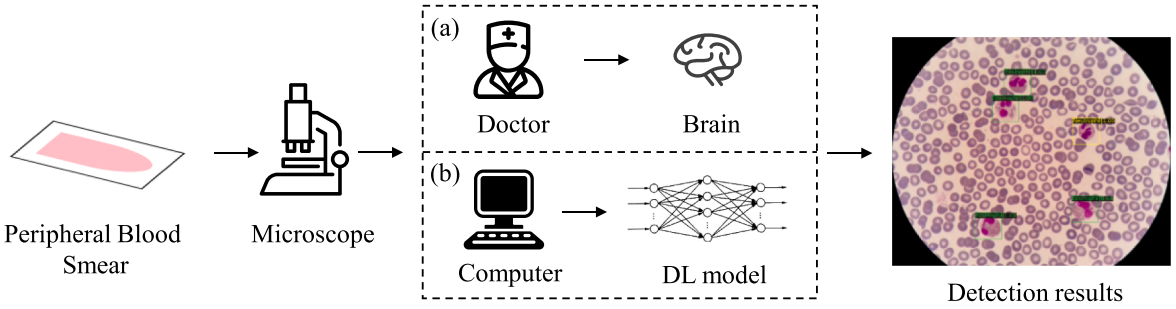

Fig. 2. Leukocyte detection with large-field-of-view microscopic images in a blood smear examination.

computational complexity and slow training convergence. Its application studies are fewer and its performance needs to be improved.

In addition to the detection models, the datasets used in some studies are mostly single-cell images similar to that presented in Fig. 1. Unlike the large-field-of-view images in actual detection scenarios, these images are more suitable for classification tasks. Some studies used classification accuracy as an evaluation metric for leukocyte detection, which lacks rationality.

In this paper, we propose an improved end-to-end leukocyte detection model based on DETR to detect leukocytes in a large-field-of-view image of a blood smear, thereby replacing the visual inspection process by the examiner, as shown in Fig. 2(b) earlier. This manuscript provides the following contributions:

1. We introduce an improved-DETR model for leukocyte detection for the first time and propose an end-to-end detection model utilizing a pure Transformer architecture that outperforms classical CNN models.
2. We used a Deformable Attention Module (DAM) to reduce the computational complexity and improve the convergence speed of the model.
3. We improved the detection performance of the model using the Pyramid Vision Transformer (PVT) model as the DETR's backbone to extract multiscale feature maps and combined it with the attention mechanism of the multiscale DAM.

The remaining parts of the paper are organised as follows. Section 2 describes the relevant theories and proposed method. Section 3 reports the datasets, experimental setup, analysis of the results, and discussion. Section 4 summarises the study and provides suggestions for future research.

2. Methods

In this section, we briefly review the theories involved in the proposed model. Subsequently, we present the proposed model.

2.1. Transformer

Google proposed the Transformer to solve the sequence-to-sequence task in natural language processing and address the long-range dependencies problem [27]. A Transformer uses an attention mechanism to weigh the association of each node of the input data. It can process the sentence as a whole and parallelise the inputs. It does not rely on past hidden states to capture dependencies on previous words, and losing past information is not a risk. Benefiting from the advantages of parallelised processing and no loss of information, a Transformer works better than traditional recurrent neural network (RNN) models.

As a Transformer's core module, the attention mechanism can be described as mapping a query and a set of key-value pairs to an output, where the query, key, value, and output are vectors. The result is a weighted sum of values, with the weight assigned to each value determined by the query's compatibility function with the corresponding key. The output matrix is computed as [27]:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

in which $1/\sqrt{d_k}$ is the scaling factor, Q, K, V are the matrix of query, key and value respectively.

The structure of a Transformer contains two parts: an encoder and a decoder. The role of the encoder is to encode the representation vector matrix—that is, feature extraction. The function of the decoder is to obtain the predicted output based on the input matrix of encoded information. By combining the two parts, the input sequence is transformed into an output sequence. A Transformer functions as follows: the representation vector of each word in the input sequence is obtained, the representation vector matrix is then input to the encoder to obtain the encoding information matrix, and finally the encoding information matrix is input to the decoder for word prediction.

2.2. DETR

The DETR algorithm [23] is used for end-to-end object detection and is based on a Transformer. DETR eliminates the need to manually design components in previous object detection models and simplifies the detection process. DETR combines bipartite matching loss with a Transformer for parallel decoding. Given a fixed set of small learnable object queries, the relationship between the objects and global image context is inferred. The final set of predictions is directly output in parallel.

Fig. 3 shows the detailed structure of the encoder and decoder parts of the DETR. First, the extracted image features are passed through the encoder along with the spatial positional encoding. Second, the decoder receives queries, output positional encoding and encoder memory. Third, multiple multi-headed self-attention and cross-attention are used to generate the predicted set. Finally, the class labels and bounding boxes are generated through the feed forward network (FFN). [23].

The bipartite matching is designed to minimize the matching loss between the ground truth set and prediction set [23]:

$$\widehat{\sigma} = \underset{\sigma\in G_N}{\text{argmin}} \sum_{i=1}^{N} L_{\text{match}}(y_i, \widehat{y}_{\sigma(i)}) \tag{2}$$

In Eq. (2), $L_{\text{match}}(y_i, \widehat{y}_{\sigma(i)})$ is the matching loss obtained using the Hungarian algorithm [28]. After the elements in the set are matched one by one, the model is trained by defining a Hungarian loss as the loss function. The loss function includes the classification loss and bounding box loss:

$$L_{\text{Hungarian}}(y,\widehat{y}) = \sum_{i=1}^{N}[-\log\widehat{p}_{\widehat{\sigma}(i)}(c_i) + 1_{\{c_i\neq\varnothing\}} L_{\text{box}}(b_i, \widehat{b}_{\widehat{\sigma}}(i))] \tag{3}$$

In Eq. (3), $\widehat{p}$ is the probability that the object is predicted to be c_i class. $L_{\text{box}}(b_i, \widehat{b}_{\widehat{\sigma}}(i))$ is the bounding box loss, b_i is the ground truth, and $\widehat{b}_{\widehat{\sigma}}(i)$ is the predicted bounding box. To alleviate the problem of loss

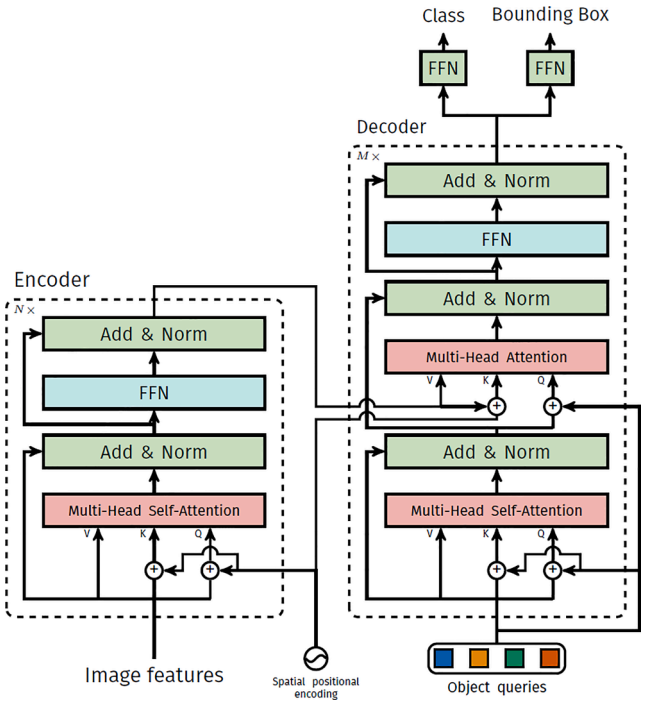

Fig. 3. Structure of the encoder and decoder in DETR [23].

scaling for different bounding box sizes, it contains *IoU* loss and L_1 loss:

$$L_{box}(b_i, \hat{b}_{\hat{\sigma}}(i)) = \lambda_{iou} L_{iou}(b_i, \hat{b}_{\hat{\sigma}}(i)) + \lambda_{L1} ||b_i - \hat{b}_{\hat{\sigma}}(i)||_1 \quad (4)$$

In Eq. (4), λ_{iou} and λ_{L1} are the corresponding hyperparameters.

2.3. PVT

Unlike the original CNN backbone of the DETR model, we used PVT [29] as a backbone to improve the performance of the DETR model and therefore build a pure Transformer object detection network. It can control the scale of the feature map by patching the embedding layer and construct a multiscale feature pyramid structure using a progressive shrinkage strategy. The PVT also includes a spatially reduced attention layer to replace the multi-head attention layer in the encoder, which effectively reduces the computational and memory overheads. The PVT can be designed following the rules of ResNet.

The entire model is divided into four phases. Each stage is divided into a patch embedding layer and transformer encoder layer, which can output feature maps at different scales. Fig. 4 depicts the structure of PVT. The version of PVT used in this study is PVT-small, which is benchmarked against the ResNet-50 model.

2.4. DAM

Inspired by [24], we introduce the DAM into the improved-DETR. DAM focuses only on a few key points around the reference point, regardless of the spatial size of the feature mapping. This module can effectively reduce the complexity of feature mapping and use a high-resolution feature mapping to better detect small objects. It can also accelerate the convergence of the model. We use DAM instead of self-attention in the encoder and cross attention in the decoder of DETR. The query in DAM only samples the keys at some of the global locations, interpolates the values based on the sampling of these locations, and finally applies the sparse attention weights to the corresponding values.

$x \in R^{C\times H\times W}$ is the input feature map, z_q and p_q are the content features and reference points corresponding to the qth query element, respectively. The deformable attention features are calculated as [24]:

$$DeformAttn(z_q, p_q, x) = \sum_{m=1}^{M} W_m [\sum_{k=1}^{K} A_{mqk} \cdot W'_m x(p_q + \Delta p_{mqk})] \quad (5)$$

where M is the total number of attention heads, K is the total number of sampled keys. Δp_{mqk} and A_{mqk} denote the sampling offset and attention weight of the kth sampled point in the mth attention head, respectively.

On the other hand, training with multiscale feature maps improves the performance of the object detection network. DAM can also naturally aggregate the multiscale feature maps generated by PVT through an attention mechanism without the need for an additional feature pyramid network. The input multiscale feature map is $\{x^l\}_{l=1}^{L}$, where $x^l \in R^{C\times H_l\times W_l}$. The normalized coordinates of the reference point of each query element are $\hat{p}_q \in [0,1]^2$ The multi-scale deformable attentional features are computed as [24]:

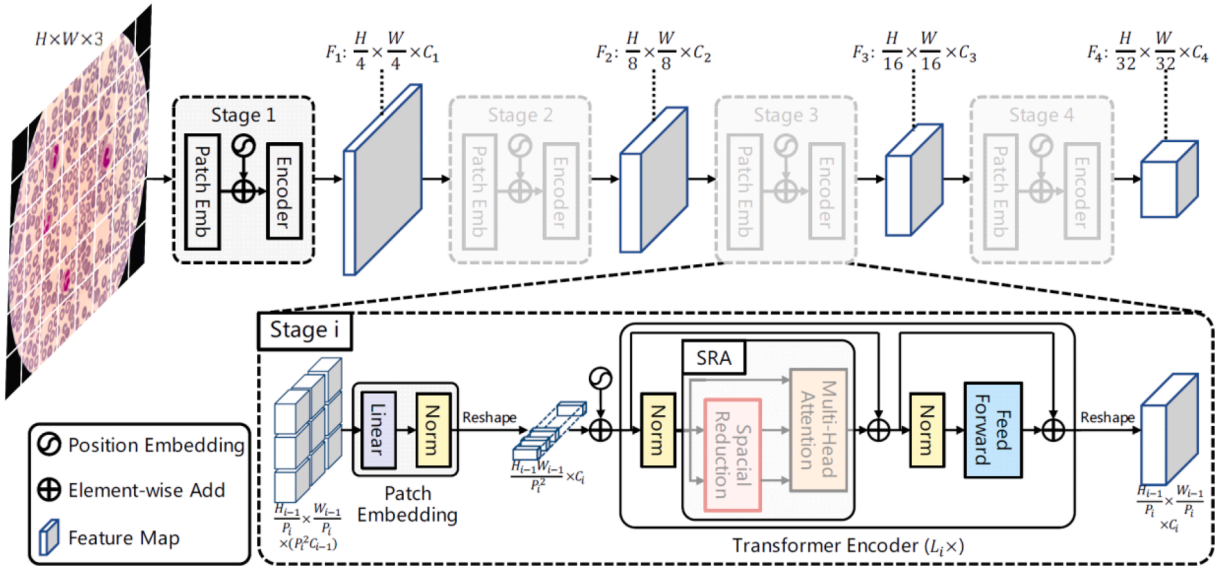

Fig. 4. Structure of PVT [29].

$$MSDeformAttn(z_q, \hat{p}_q, \{x^l\}_{l=1}^{L}) = \sum_{m=1}^{M} W_m \left[\sum_{l=1}^{L} \sum_{k=1}^{K} A_{mlqk} \cdot W'_m x^l(\phi_l(\hat{p}_q) + \Delta p_{mlqk}) \right] \quad (6)$$

Where Δp_{mlqk} and A_{mlqk} denote the sampling offset and attention weight of the kth sampling point in the mth attention head in the lth feature layer, respectively.

2.5. The proposed model

Because of the high computational complexity of the attention module in DETR, the original model uses only low-resolution single-scale feature maps, which leads to a low detection performance of small targets. On the other hand, DETR was initialised with the same weight for each query for all positions, which lead to a slow convergence during model training. Therefore, the proposed model improves on both aspects. Fig. 5 shows the whole framework roughly divided into three steps. We generated the multiscale hierarchical feature maps by using the Transformer structure of the PVT model, and then imported the multiscale feature maps into the encoder and decoder parts with DAM. Finally, we directly outputted the set predictions in parallel by combining the bipartite matching loss. The proposed improved-DETR improved the convergence speed and detection performance.

3. Experimental methodology

As presented in this section, we carried out numerous experiments to confirm the superiority of the proposed leukocyte detection model. We present the datasets, performance metrics, training strategy, analysis of experimental results, and discussion in the following subsections.

3.1. Datasets

Numerous studies on leukocyte identification achieved good results by using the LISC dataset [30] as a public dataset. However, the field of view and resolution of the images in this dataset were small (720 × 576), which lead to a large gap with the actual examination process, with a small number of images. The Raabin public dataset [31] used in this study had a resolution of 5312 × 2988 for images from real medical scenarios. We annotated the dataset by two senior haematologists, with some unidentified and controversial annotation results. However, the bounding box of cell annotation was large and included many background areas. Therefore, we screened the noncontroversial cell annotations and reannotated the extent of the bounding box into the format of the Common Objects in Context (COCO) dataset. The reannotated cells were three types of leukocytes: eosinophil, monocyte, and neutrophil. We used a total number of used images at 10,323, which we divided into training and validation sets in a ratio of 8:2.

This dataset does not contain patient information. The original authors did not indicate that the data source had compliance or ethical risks. The source dataset is currently in public use. The literature that using these datasets is publicly available. This study used this publicly published dataset for cell detection algorithm experiments and did not provide a new dataset. Table 1 lists the information of the used dataset.

Also, with the development and application of DL, transfer learning (TL) training in medical imaging has become a routine practice. We trained the designed models from scratch (FS) on natural image datasets—e.g., the COCO dataset—to obtain the pretrained weights. Then we fine-tuned the models on medical imaging datasets. Specifically, for training large DL models on small medical datasets, TL can achieve better results than FS [32]. For pretraining of the improved-DETR model in this study, we use the COCO dataset [33], which contains 118,000 training images and 5,000 validation images.

3.2. Performance evaluation

We used the precision and recall to evaluate the detection performance of the model. Further, we used average precision (AP) for a comprehensive evaluation. We calculated the AP values from the area under the precision and recall curves, which can be used to consider both accuracy and recall of the model. For multicategory object detection, we used the mean average precision (mAP) to evaluate the performance of detection. The calculation equations for each metric are:

Table 1
Details of the Raabin dataset.

Dataset	Number	Capturing Device	Magnification	Clinical Contribution
Raabin [45]	10,323	CX18 microscope with Samsung Galaxy S5 camera	100x	Leukocyte detection

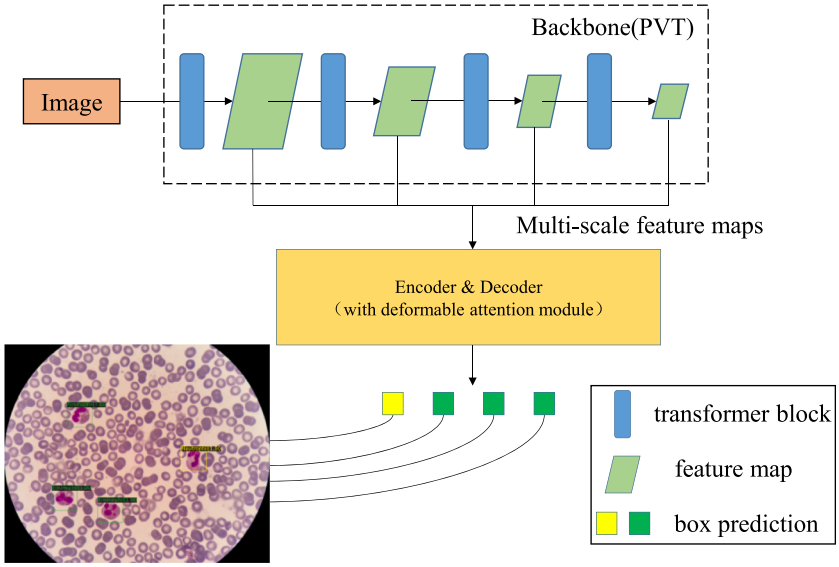

Fig. 5. Proposed overall model framework.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \sum_{i=1}^{N} p_i \Delta r_i \tag{9}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{10}$$

where *TP* and *FP* denote the number of true positives and false positives, *TN* and *FN* denote the number of true negatives and false negatives, respectively. N is the number of cells of the same class, p_i is a function of accuracy, and Δr_i is a function of recall. n is the total number of object classes and AP_i is the AP value of the ith class.

The metric mAP@[0.5:0.95] used in this study is the mean value of mAP corresponding to the ten values when the Intersection over Union (IOU) was in the 0.5–0.95 range, with an interval of 0.05. This metric can be used to evaluate the accuracy of model detection and localisation more comprehensively (abbreviated as mAP in this paper).

3.3. Training strategies

We built our model on the PyTorch 1.8.0 framework using a workstation with four NVIDIA RTX 3090 graphics processing units (GPUs), Ubuntu 20.104.1 LTS operating system, and two Intel® Xeon® Silver 4210R central processing units (CPUs). Because of the high GPU computing power, we were able to train the large model based on the COCO dataset.

We train the following object detection models: DETR, Improved-DETR, Faster R-CNN, and YOLO v3. For each model, we perform distributed training on four RTX 3090 GPUs.

(1) For the FS training of improved-DETR on the COCO dataset, the parameters are set as follows: optimiser = AdamW, learning_rate = 2E-4, weight_decay = 1E-4, epoch_number = 70, batch_size = 4. Owing to the large size of the COCO dataset, the training time is approximately 213 h.

(2) For the TL training of the DETR model on the leukocyte dataset (DETR-TL), the parameters are set as follows: optimiser = AdamW, learning rate = 1E-4, weight_decay = 1E-4, epoch_number = 100, batch_size = 8. The training time is approximately 10.6 h. The DETR model is trained FS on the leukocyte dataset with the same parameters; the training time is similar.

(3) For the TL training of improved-DETR on the leukocyte dataset (Improved-DETR-TL), the parameters are set as follows: Optimizer = AdamW, learning_rate = 2E-4, weight_decay = 1E-4, epoch_number = 100, batch_size = 4. Training time: about 24.3 h. The FS training of improved-DETR on leukocyte dataset (Improved-DETR-FS) uses the same parameters and the training time is close.

(4) For the TL training of Faster R-CNN on the leukocyte dataset (Faster R-CNN-TL), the parameters are set as follows: Optimizer = SGD, learning_rate = 2E-2, momentum = 0.9, weight_decay = 1E-4, epoch_number = 100, batch_ size = 8. Training time: about 11 h.

(5) For the TL training of YOLO v3 on the leukocyte dataset (YOLO v3-TL), the parameters are set as follows: Optimizer = SGD, learning_rate = 1E-2, momentum = 0.9, weight_decay = 5E-4, epoch_number = 100, batch_size = 8. Training time: about 9 h.

3.4. Results analysis

The experiments were divided into three main parts:

In the first part of the experiments, we trained the improved-DETR on the COCO dataset FS because, to implement the TL training of the improved-DETR on the leukocyte dataset, we needed to obtain the pretrained weights of the model on the COCO dataset. The optimal mAP obtained during training was 0.44. Fig. 6 shows the loss and mAP during training.

In the second part of the experiments, we performed a series of trainings and comparisons on the leukocyte dataset using the improved-DETR and original DETR. The improved-DETR demonstrated excellent performance. We used the pretrained weights obtained in the first part to train the improved-DETR on the leukocyte dataset via TL. The maximum mAP was 0.961. Fig. 7 shows the loss and mAP changes during the training process.

We then used the weights of the original DETR pretrained on the COCO dataset to train the DETR via TL. As shown in Fig. 7, the improved-DETR model converged at the 40th epoch. The mAP values throughout the training process were higher than those of the original DETR model, with the maximum mAP being 0.902. This indicated that the improved-DETR exhibits a better performance than that of the original DETR.

Finally, to verify the effectiveness of using TL, we trained the improved-DETR and original DETR on the leukocyte dataset FS. Notably, DETR failed to converge and the mAP was always 0. In contrast, the proposed improved-DETR model still showed a better performance with a maximum mAP of 0.948. This result surpasses that obtained for DETR with TL but is still lower than that obtained for the improved-DETR model where TL was applied. Thus, the superiority of the improved model and necessity of TL are confirmed.

The third part of the experiments compared the performance of the proposed improved-DETR model to those of the detection models based on CNN architectures, such as Faster R-CNN and YOLO v3. We trained all three models on the leukocyte dataset via TL. Fig. 8 shows the mAP variations of the three models. The detection performance of YOLO v3 was significantly lower than those of the other models, because YOLO v3 has high speed but low detection precision. The performance of the improved-DETR model was lower than that of the Faster R-CNN at the beginning and exhibited large fluctuations. This may be related to the convergence speed of the transformer. Although the improved model showed an accelerated convergence, it was still slower than the CNN model. In the later phase of the training, the improved-DETR model outperformed the Faster R-CNN and achieved an optimal detection performance.

We compared the results of the experiments. First, Table 2 lists the evaluation metrics of each model to compare the detection performances of the models. Because the detection was performed on three types of leukocytes, the AP for each type of leukocyte is also evaluated. The comparison of the precision, recall, and mAP values shows that the proposed improved DETR achieves an optimal overall detection performance. The AP values of the improved DETR model for each cell type

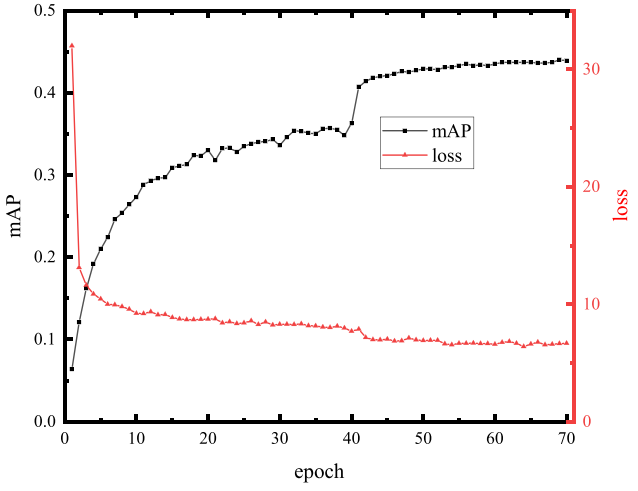

Fig. 6. Training loss and mAP of the improved-DETR on the COCO dataset.

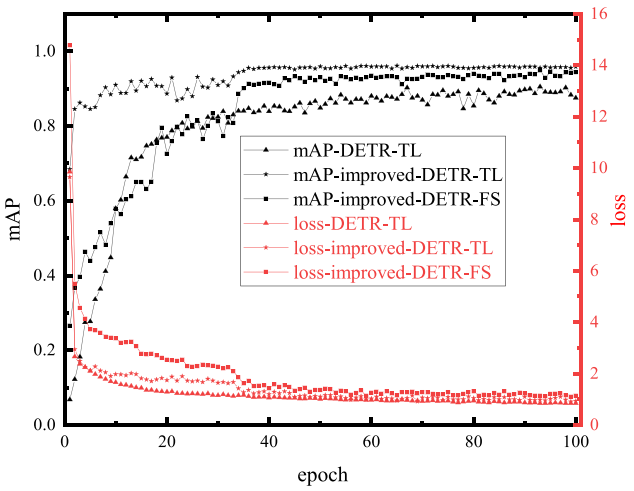

Fig. 7. Training loss and mAP of the models on the leukocyte dataset.

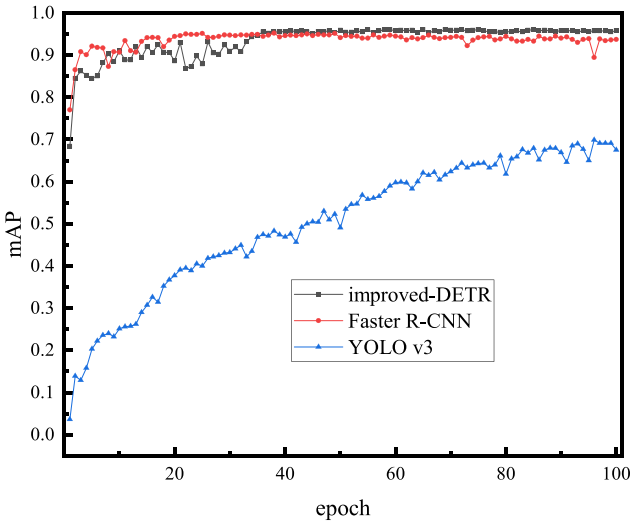

Fig. 8. Maps of the models on the leukocyte dataset.

are almost at their optimal values. Amongst the three types of cells, neutrophils have the highest AP, which may be related to their simpler morphology.

Second, to compare the actual detection effect of each model, we selected challenging images for test each model. The proposed models achieved the optimal results. In the test image shown in Fig. 9, two cells are relatively close to each other. Both DETR-TL and YOLO v3-TL show a missed detection, whilst Improved-DETR-FS and Faster R-CNN-TL detect cells with a low confidence. In the test image shown in Fig. 10, the number of leukocytes is high and affected by impurity. DETR-TL, Faster R-CNN-TL, and YOLO v3-TL show missed detections, whilst Improved-DETR-FS detects cells with an, overall, lower confidence than that of the Improved-DETR-TL model.

Additionally, we evaluated the computation and number of parameters of each model (with the input image size of 1200 × 800 as an example). Table 3 compares the inference speed of each model for leukocyte detection, calculated with a single GPU and 2000 images in the validation set used for testing. The comparison shows that the YOLO v3 model has the highest inference speed, whereas our proposed model has a lower inference speed. The reason for this can be explained as follows. The use of multiscale feature maps increases the computational volume. On the other hand, the optimisation of the Transformer is not as good as that of the CNN because of the existing GPU hardware computations; this is also the primary reason why the inference speed of our proposed model was limited [23].

3.5. Discussion

In this study, we constructed a pure Transformer-based end-to-end detection network by improving the DETR model. The optimal training results were obtained via concise experimental steps. To the best of our knowledge, this type of model has not been applied in the field of object detection using medical images; as a result, this study is novel. The experimental results demonstrate the superiority of the proposed method in terms of both its detection performance and practical detection results, respectively.

Therefore, for leukocyte detection, improving the DETR model using PVT and DAM can yield performances better than those of classical CNNs—that is, Faster R-CNN. However, this study has two limitations that need to be addressed by future studies. First, we tested the performance of our proposed model on a relatively large public dataset of leukocytes. However, we did not perform any clinical validation experiments. In the future, we will collect additional clinical data for model training and validation. Second, although our proposed improved-DETR model achieves optimal performance, its inference speed is lower than that of the CNN. Recent studies have shown that the DETR model without the backbone or encoder can still achieve promising performances [34–35]. In the future, we will further optimise the structure of our DETR model to improve its inference speed.

4. Conclusion

We propose an improved-DETR model for microscopic image leukocyte detection. The primary aim of our study was to achieve high-precision object detection of leukocytes from large-field-of-view microscopic images of real medical scenes. Our improved-DETR model is based on PVT and DAM, pretrained on the COCO dataset, and finally trained via TL on a modified leukocyte dataset.

Our proposed model achieved mAP of up to 0.961 on the validation set, thereby exceeding the performance of the classical CNN detection algorithm. Therefore, this study is a useful exploration of end-to-end detection models based on the pure Transformer structure. The proposed model with the Transformer structure, which has recently exhibited good performance in the CV field, can applied in the field of medical imaging to achieve superior detection accuracy.

Table 2
Detection performance of each model.

Model	Precision (%)	Recall (%)	mAP	Standard deviation	AP Eosinophil	Monocyte	Neutrophil
DETR-TL	72.8	88.0	0.902	0.051	0.843	0.896	0.968
Improved-DETR-TL	**75.1**	**94.3**	**0.961**	**0.015**	**0.954**	**0.947**	0.982
Improved-DETR-FS	74.5	94.1	0.948	0.023	0.925	0.938	0.980
Faster R-CNN-TL	72.5	92.5	0.952	0.023	0.946	0.927	**0.984**
YOLO v3-TL	66.6	80.9	0.699	0.083	0.607	0.681	0.808

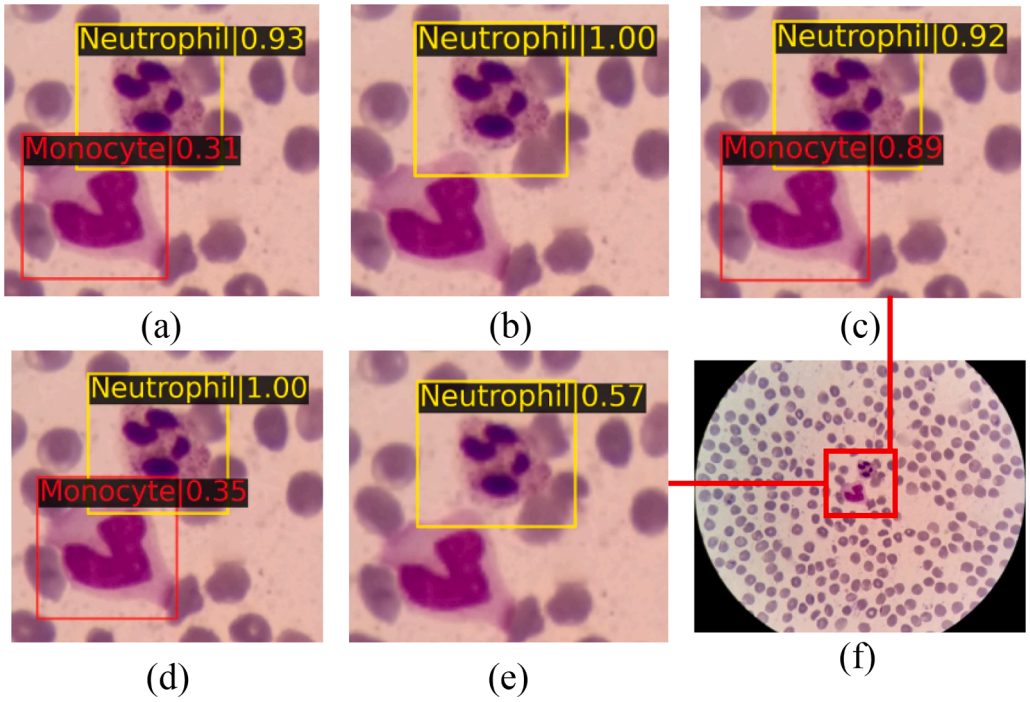

Fig. 9. Comparison of the detection results on the image containing adjacent leukocytes: (a) DETR-TL; (b) Improved-DETR-TL; (c) Improved-DETR-FS; (d) Faster R-CNN-TL; (e) YOLO v3-TL; (f) the original image.

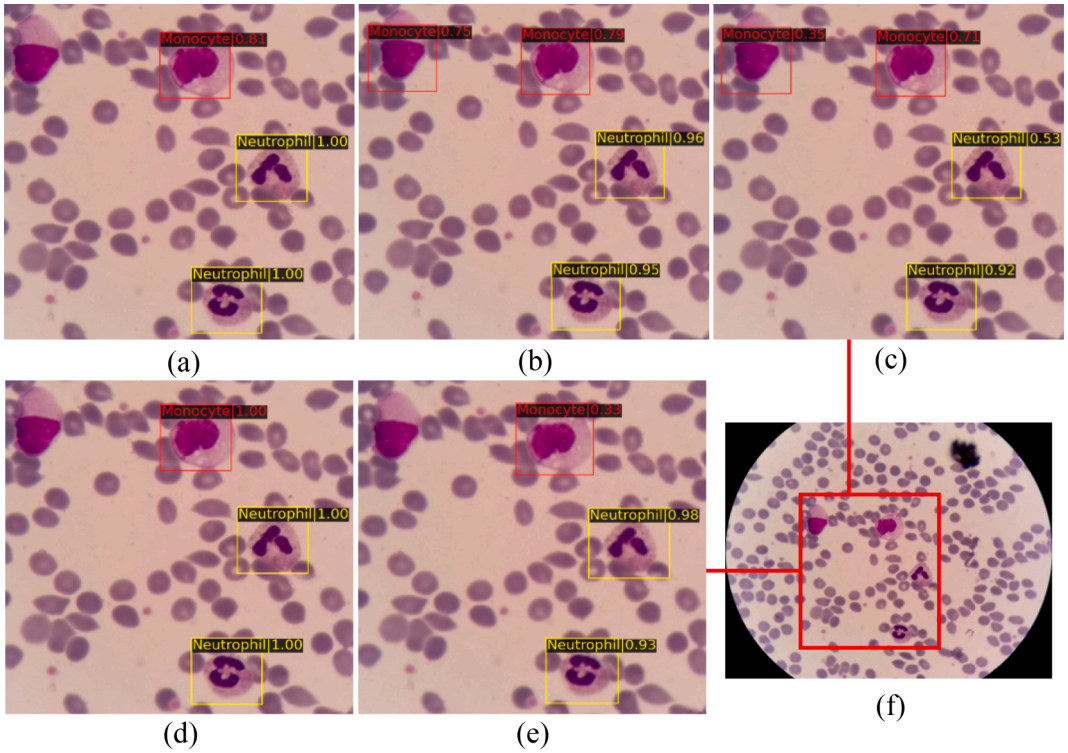

Fig. 10. Comparison of the detection results on the image containing multiple objects and impurity: (a) DETR-TL; (b) Improved-DETR-TL; (c) Improved-DETR-FS; (d) Faster R-CNN-TL; (e) YOLO v3-TL; (f) the original image.

CRediT authorship contribution statement

Bing Leng: Methodology, Software, Writing – original draft. **Chunqing Wang:** Formal analysis. **Min Leng:** Investigation. **Mingfeng Ge:** Writing – review & editing. **Wenfei Dong:** Conceptualization.

Declaration of Competing Interest

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

Table 3
Comparison of operations, parameters and inference speed of each model.

Row#	Model	Operations (GFLOPs)	Parameters (M)	Inference (fps)
1	DETR	91	41	22.4
2	Improved-DETR	169	36	11.4
3	Faster R-CNN	206	41	23.5
4	YOLO v3	193	61	57.3

Data availability

Data will be made available on request.

Acknowledgements

This research was funded by the National Key R&D Program of China (No. 2021YFB3602200), the Primary R&D Plan of Jiangsu Province (BE2019683), the Science Foundation of the Chinese Academy of Sciences (No. 2020SYHZ0041), and the Instrument Developing Project of Chinese Academy of Science (YJKYYQ20200038).

References

[1] Y.Y. Baydilli, Ü. Atila, Classification of white blood cells using capsule networks, Comput. Med. Imaging Graph. 80 (2020), https://doi.org/10.1016/j.compmedimag.2020.101699.
[2] R. Al-qudah, C.Y. Suen, A Survey on Peripheral Blood Smear Analysis Using Deep Learning, in: Pattern Recognit. Artif. Intell. ICPRAI, 2020: pp. 725–738. https://doi.org/10.1007/978-3-030-59830-3_63.
[3] B. Leng, M. Leng, M. Ge, W. Dong, Knowledge distillation-based deep learning classification network for peripheral blood leukocytes, Biomed. Signal Process. Control. 75 (2022), 103590, https://doi.org/10.1016/j.bspc.2022.103590.
[4] S. Zhang, Q. Ni, B. Li, S. Jiang, W. Cai, H. Chen, L. Luo, Corruption-Robust Enhancement of Deep Neural Networks for Classification of Peripheral Blood Smear Images, in: Med. Image Comput. Comput. Assist. Interv. – MICCAI, 2020: pp. 372–381. https://doi.org/10.1007/978-3-030-59722-1_36.
[5] Y. Lu, X. Qin, H. Fan, T. Lai, Z. Li, WBC-Net: A white blood cell segmentation network based on UNet++ and ResNet, Appl. Soft Comput. 101 (2021), 107006, https://doi.org/10.1016/j.asoc.2020.107006.
[6] R. Al-qudah, C.Y. Suen, Improving blood cells classification in peripheral blood smears using enhanced incremental training, Comput. Biol. Med. 131 (2021), 104265, https://doi.org/10.1016/j.compbiomed.2021.104265.
[7] C. Matek, S. Krappe, C. Münzenmayer, T. Haferlach, C. Marr, Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set, Blood. 138 (2021) 1917–1927, https://doi.org/10.1182/blood.2020010568.
[8] K.K. Anilkumar, V.J. Manoj, T.M. Sagi, A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of Leukemia, Biocybern, Biomed. Eng. 40 (2020) 1406–1420, https://doi.org/10.1016/j.bbe.2020.08.010.
[9] J. Banks, K. Nugyen, A. Al-sabaawi, I. Tomeo-reyes, V. Chandran, Segmentation of White Blood Cell, Nucleus and Cytoplasm in Digital Haematology Microscope Images: A Review-Challenges, Current and Future Potential Techniques, IEEE Rev. Biomed. Eng. 3333 (2020) 1–16, https://doi.org/10.1109/RBME.2020.3004639.
[10] Q. Wang, S. Bi, M. Sun, Y. Wang, D. Wang, S. Yang, Deep learning approach to peripheral leukocyte recognition, PLoS One. 14 (2019) e0218808.
[11] L. Putzu, G. Caocci, C. Di Ruberto, Leucocyte classification for leukaemia detection using image processing techniques, Artif. Intell. Med. 62 (2014) 179–191, https://doi.org/10.1016/j.artmed.2014.09.002.
[12] P.P. Banik, R. Saha, K.D. Kim, An Automatic Nucleus Segmentation and CNN Model based Classification Method of White Blood Cell, Expert Syst. Appl. 149 (2020), 113211, https://doi.org/10.1016/j.eswa.2020.113211.
[13] J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement, (2018). http://arxiv.org/abs/1804.02767.
[14] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
[15] H. Kutlu, E. Avci, F. Özyurt, White blood cells detection and classification based on regional convolutional neural networks, Med. Hypotheses. 135 (2020), 109472, https://doi.org/10.1016/j.mehy.2019.109472.
[16] M.R. Reena, P.M. Ameer, Localization and recognition of leukocytes in peripheral blood: A deep learning approach, Comput. Biol. Med. 126 (2020), 104034, https://doi.org/10.1016/j.compbiomed.2020.104034.
[17] H. Fan, F. Zhang, L. Xi, Z. Li, G. Liu, Y. Xu, LeukocyteMask: An automated localization and segmentation method for leukocyte in blood smear images using deep neural networks, J. Biophotonics. 12 (2019) 1–17, https://doi.org/10.1002/jbio.201800488.
[18] J. Hung, A. Goodman, D. Ravel, S.C.P. Lopes, G.W. Rangel, O.A. Nery, B. Malleret, F. Nosten, M.V.G. Lacerda, M.U. Ferreira, L. Rénia, M.T. Duraisingh, F.T.M. Costa, M. Marti, A.E. Carpenter, Keras R-CNN: Library for cell detection in biological images using deep neural networks, BMC Bioinformatics. 21 (2020) 1–7, https://doi.org/10.1186/s12859-020-03635-x.
[19] C. Di, A. Loddo, L. Putzu, Detection of red and white blood cells from microscopic blood images using a region proposal approach, Comput. Biol. Med. 116 (2020), 103530, https://doi.org/10.1016/j.compbiomed.2019.103530.
[20] R. Khandekar, P. Shastry, S. Jaishankar, O. Faust, N. Sampathila, Automated blast cell detection for Acute Lymphoblastic Leukemia diagnosis, Biomed. Signal Process. Control. 68 (2021), 102690, https://doi.org/10.1016/j.bspc.2021.102690.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR. (2021). http://arxiv.org/abs/2010.11929.
[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, ICCV. (2021). http://arxiv.org/abs/2103.14030.
[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End Object Detection with Transformers, ECCV. (2020), https://doi.org/10.1007/978-3-030-58452-8_13.
[24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable Transformers for End-to-End Object Detection, ICLR. (2021) 1–14. https://doi.org/arXiv:2010.04159.
[25] Z. Dai, B. Cai, Y. Lin, J. Chen, UP-DETR: Unsupervised Pre-training for Object Detection with Transformers, CVPR. (2021) 1601–1610. https://doi.org/10.1109/CVPR46437.2021.00165.
[26] T. Prangemeier, C. Reich, H. Koeppl, Attention-Based Transformers for Instance Segmentation of Cells in Microstructures, Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020. (2020) 700–707. https://doi.org/10.1109/BIBM49941.2020.9313305.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process, Syst, 2017 https://doi.org/10.48550/arXiv.1706.03762.
[28] H.W. Kuhn, The Hungarian method for the assignment problem, Nav. Res. Logist. 52 (2005) 7–21, https://doi.org/10.1002/nav.20053.
[29] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions, in: 2021 IEEE/CVF Int. Conf. Comput. Vis., IEEE, 2021: pp. 548–558. https://doi.org/10.1109/ICCV48922.2021.00061.
[30] S.H. Rezatofighi, H. Soltanian-Zadeh, Automatic recognition of five types of white blood cells in peripheral blood, Comput. Med. Imaging Graph. 35 (2011) 333–343, https://doi.org/10.1016/j.compmedimag.2011.01.003.
[31] Z.M. Kouzehkanan, S. Saghari, S. Tavakoli, P. Rostami, M. Abaszadeh, F. Mirzadeh, E.S. Satlsar, M. Gheidishahran, F. Gorgi, S. Mohammadi, R. Hosseini, A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm, Sci. Rep. 12 (2022) 1–14, https://doi.org/10.1038/s41598-021-04426-x.
[32] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, Transfusion: Understanding transfer learning for medical imaging, Adv. Neural Inf. Process. Syst (NeurIPS). 32 (2019). https://doi.org/10.48550/arXiv.1902.07208.
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common Objects in Context, in: ECCV, 2014: pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
[34] F. Liu, H. Wei, W. Zhao, G. Li, J. Peng, Z. Li, WB-DETR: Transformer-Based Detector without Backbone, in: 2021 IEEE/CVF Int. Conf. Comput. Vis., IEEE, 2021: pp. 2959–2967. https://doi.org/10.1109/ICCV48922.2021.00297.
[35] J. Lin, Y. Chen, D^{2}ETR: Decoder-Only DETR with Computationally Efficient Cross-Scale Attention, https://arxiv.org/abs/2203.00860.