Computational and Structural Biotechnology Journal 21 (2023) 2985–3001

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

An improved Yolov5s based on transformer backbone network for detection and classification of bronchoalveolar lavage cells

Puzhen Wu [a,b], Han Weng [b], Wenting Luo [c], Yi Zhan [b], Lixia Xiong [c], Hongyan Zhang [d,*], Hai Yan [a,*]

[a] The Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, Beijing 100124, China
[b] Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China
[c] Department of Pathophysiology, Medical College, Nanchang University, 461 Bayi Road, Nanchang 330006, China
[d] Department of Burn, The First Affiliated Hospital, Nanchang University, 17 Yongwaizheng Road, Nanschang 330066, China

ARTICLE INFO

Article history:
Received 28 October 2022
Received in revised form 4 May 2023
Accepted 5 May 2023
Available online 6 May 2023

Keywords:
Deep learning
Convolutional neural network
Cell detection
Bronchoalveolar lavage cells
Transformer

ABSTRACT

Biological tissue information of the lung, such as cells and proteins, can be obtained from bronchoalveolar lavage fluid (BALF), through which it can be used as a complement to lung biopsy pathology. BALF cells can be confused with each other due to the similarity of their characteristics and differences in the way sections are handled or viewed. This poses a great challenge for cell detection. In this paper, An Improved Yolov5s Based on Transformer Backbone Network for Detection and Classification of BALF Cells is proposed, focusing on the detection of four types of cells in BALF: macrophages, lymphocytes, neutrophils and eosinophils. The network is mainly based on the Yolov5s network and uses Swin Transformer V2 technology in the backbone network to improve cell detection accuracy by obtaining global information; the C3Ghost module (a variant of the Convolutional Neural Network architecture) is used in the neck network to reduce the number of parameters during feature channel fusion and to improve feature expression performance. In addition, embedding intersection over union Loss (EIoU_Loss) was used as a bounding box regression loss function to speed up the bounding box regression rate, resulting in higher accuracy of the algorithm. The experiments showed that our model could achieve mAP of 81.29% and Recall of 80.47%. Compared to the original Yolov5s, the mAP has improved by 3.3% and Recall by 3.67%. We also compared it with Yolov7 and the newly launched Yolov8s. mAP improved by 0.02% and 2.36% over Yolov7 and Yolov8s respectively, while the FPS of our model was higher than both of them, achieving a balance of efficiency and accuracy, further demonstrating the superiority of our model.

1. Introduction

Bronchoalveolar lavage fluid (BALF) is the alveolar surface lining fluid obtained by repeated lavage with sterile saline at the level of the lung segments and sub-pulmonary segments below the bronchi by fibreoptic bronchoscopy [1]. Cytomorphological examination of BALF is of great importance in the diagnosis, observation of outcome and prognosis of respiratory diseases, especially interstitial lung disease, pulmonary infiltrates and some infectious diseases [2,3]. BALF is still regarded as the gold standard for the identification of pulmonary misfiring in chronic interstitial lung disease [4]. When clinical history, physical examination, common laboratory testing, pulmonary function tests, and radiography are insufficient to provide a conclusive diagnosis, it typically offers important diagnostic information [5]. Clarification of inflammatory cell alterations (e.g., lymphocytes, eosinophils, neutrophils and macrophages) in BALF can provide a better understanding of the disease and can be used as an objective parameter to assess the efficacy of treatment [6]. According to research, lymphocyte-dominated BALF substantially supports the diagnosis of pulmonary nodular disease or allergic pneumonia [7]. Increased percentages of neutrophils and/or eosinophils have now been found to be associated with reduced carbon monoxide lung diffusion [8].

Although cellular differential findings on BALF often lack specificity, they may still be useful in identifying certain diseases and thus narrowing the diagnosis [9]. For example, studies have shown that cell patterns in BALF can be used to differentiate specific diseases in

* Corresponding authors.
E-mail addresses: ndyfy00672@ncu.edu.cn (H. Zhang), yhai@bjut.edu.cn (H. Yan).

https://doi.org/10.1016/j.csbj.2023.05.008

patients with acute respiratory failure [10]. According to the 2020 Chinese Expert Consensus on BALF cytomorphology testing [11], the counting and classification of different cell types in BALF is still currently done by trained pathologists by flow cytometry or manually after filtration or cell centrifugation techniques. This process of manually counting and classifying cells is not only complex, tedious and time-consuming, but also has significant variation in image reporting between pathologists with different working experience [12]. The interpretation of BALF often calls for the use of a bio-oil lens with a 1000x magnification and the counting of at least 400 cells [13]. There is little question that cytological examination of bronchoalveolar lavage fluid is a major job for Chinese pathologists who must analyze hundreds of slides each day, and a shortage of skilled operators may impair the accuracy of results [11,13]. Furthermore, for patients with acute and severe diseases, such as refractory Mycoplasma pneumoniae pneumonia, obtaining accurate data in the quickest feasible time is vital for following clinical decisions to be made, which influences the patient's prognosis [14,15]. While bronchoalveolar lavage fluid is crucial for the detection of lung-related disorders, its use in clinical screening or fast field evaluation may be constrained for a variety of reasons, including examination time [16]. Therefore, it would be beneficial to develop an intelligent deep learning model for BALF cell detection and classification in order to improve the accuracy and efficiency of diagnosis and reduce the workload of specialist physicians.

In recent years, deep learning techniques have demonstrated excellent image processing and object recognition capabilities in a number of large-scale visual recognition challenge tasks, and many researchers and groups have explored the possibility of applying deep learning to medical image interpretation, and have made very significant breakthroughs. In the medical field, image processing tasks are mainly classification, detection and segmentation, with cell recognition and counting being an important part of these three tasks for medical examination. In the field of cell detection, Banik et al. proposed a convolutional neural network (CNN) model-based method for identifying different kinds of leukocytes by segmenting white blood cell (WBC) cell nuclei based on color space transformation and k-means mean algorithm [17]. They achieved an overall accuracy of 96% on BCCD (a small-scale dataset for blood cells detection) test database by using nine classification metrics. Tayebi et al. proposed an end-to-end deep learning-based system for detecting bone marrow cells [18]. Following retrospective analysis of digital whole slide images (WSI) from 1247 patients, the model achieved accuracy, precision, specificity, and NPV of over 90%. Delgado-Ortet et al. proposed a deep learning method for erythrocyte image segmentation and malaria detection [19]. According to the test set from Core Laboratory at the Hospital Clinic of Barcelona and the online repository Malaria Dataset, this model had a worldwide accuracy of 93.72% and a specificity of 87.04% for detecting malaria in red blood cells (RBCs). This work highlights the application of deep learning to the field of digital pathology. Bibi et al. proposed an Internet of Medical Things (IoMT-) based framework for identifying leukaemia-related cells [20]. Their proposed deep CNNs (DenseNet-121 and ResNet-34) both yielded accuracies above 99% on ALL-IDB and ASH image bank. In the current study on the detection and counting of different cells in BALF, Yi Tao et al. used an automated biomicroscopy platform to acquire visual images and proposed a convolutional neural network-based algorithm to automatically interpret BALF cytology [21]. The sample collected a total of 12,900 images from 139 subjects. The study successfully detected the majority of cells in BALF specimens. However, the sample images in that study had sparse cell density and the trained model may not work well in the dense situation of actual detection. Peng et al. integrated an enhanced master curve technique with an evolving neural network to create a unique hybrid method for rectal ultrasound image segmentation, resulting in increased image segmentation accuracy and resilience [22]. Sai Kit Lam et al. created an image-omics-based algorithm for determining whether patients with nasopharyngeal carcinoma can undergo radiation treatment [23]. Additionally, they suggested an automated A-LugSeg approach based on the Mask-RCNN deep learning and refinement subnetwork to precisely segment lung X-ray images to aid in the diagnosis of lung diseases [24].

In order to improve the detection accuracy and efficiency of BALF cells in dense situations we decided to use the you only look once (Yolo) algorithm [25]. The core idea of Yolo is to transform target detection into a regression problem by applying a single convolutional neural network to the whole image, using the whole image as input, to obtain the position of the bounding box and the probability of the class to which it belongs. Yolov5 is one of the models in the Yolo series. This network model has a high detection accuracy as well as a fast inference speed. Based on this, we propose an improved Transformer backbone network based on Yolov5s for the detection and classification of bronchoalveolar lavage fluid cells [26]. In summary, there were six main contributions in our study:

(1) The original data volume is expanded using data augmentation to prevent overfitting and improve the robustness of the model.

(2) The introduction of Swin Transformer V2 as the backbone network, which allows feature extraction of global information and thus improves the accuracy of cell detection.

(3) Presenting the Ghost module in the neck network of Yolov5s, which reduces the parameters to achieve a lightweight model and facilitates the capture and extraction of detailed features.

(4) We adapt the bounding box regression loss function to embedding intersection over union Loss (EIoU_Loss) and introduce the positive and negative sample allocation method of Yolov7 in order to speed up the bounding box regression and increase the precision of the algorithm for cell recognition.

(5) Using weighted non-maximum suppression (Weighted-NMS), we do a weighting and averaging of the prediction bounding box information, which can overlap the information of each bounding box and ensure the accuracy of localization and classification.

(6) We have deployed our Improved Yolov5s Based on Transformer Backbone Network for Detection and Classification of Bronchoalveolar Lavage Fluid Cells to a web client. People around the world can upload BALF cell images for cell detection. It is currently available at http://www.balfcell.cn/. Fig. 1 and Table 1.

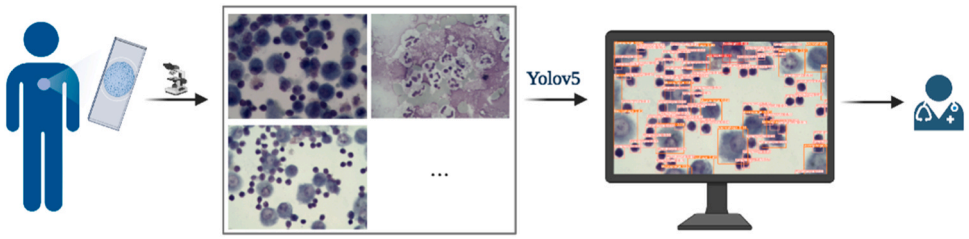


Fig. 1. Our model can detect and classify four types of cells in bronchoalveolar lavage fluid, allowing doctors to make a diagnosis.

Table 1
The application of deep learning techniques in cell detection.

References	Methods	Results	Applications
Banik 2020[17]	CNN, color space transformation, k-means mean algorithm	accuracy 96% on BCCD test database	segment WBC cell nuclei, identify different kinds of leukocytes
Tayebi 2022[18]	end-to-end deep learning	accuracy, precision, specificity, and NPV over 90%	detect bone marrow cells
Delgado-Ortet 2020[19]	deep learning	accuracy 93.72%; specificity 87.04%	segment erythrocyte image, detect malaria
Bibi 2020[20]	IoMT-based framework	accuracy above 99% on ALL-IDB and ASH image bank	identify leukaemia-related cells
Yi Tao 2022[21]	CNN	sensitivity, precision, and F1 score over 0.9	detect most cells in BALF
Tao Peng 2022[22]	evolutionary neural network, improved principal curve	DSC 96.8%; Ω95.7%; ACC 96.4%	segment ultrasound image
Sai Kit Lam 2022[23]	Ridge and MKL	an average AUC of 0.942 and 0.918 in training and hold-out test set	predict radiation therapy eligibility in nasopharyngeal carcinoma patients
Tao Peng 2022[24]	Mask-RCNN, refinement subnetwork	DSC 0.973; Ω 0.958; ACC 0.972 for JSRT	automatic lung segmentation in CXRs

CNN: convolution neural network; IoMT: Internet of Medical Things; DSC: Dice similarity coefficient; Ω: Jaccard similarity coefficient; ACC: accuracy; MKL: Multi-Kernel Learning; RCNN: region convolution neural network; JSRT: Japanese Society of Radiological Technology dataset

2. Methods and Materials

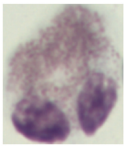
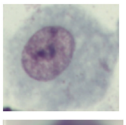
2.1. Dataset

The BALF dataset used in this study was derived from the publicly available dataset "Bronchoalveolar Lavage Fluid Cell Sorting Count Challenge" [27], the dataset that classifies cells in BALF into four categories in general: macrophages, lymphocytes, neutrophils and eosinophils. Normally, the number of nucleated cells is (90–260) x 106/L, with 85%– 96% alveolar macrophages, 6%– 15% lymphocytes, ≤ 3% neutrophils and < 1% eosinophils. This bronchoalveolar lavage fluid cell target assay dataset was annotated using the Computer Vision Annotation Tool (CVAT) [28] platform, with image files of the same filename and their label files providing information on the location and category of the data images and boxes. Table 2.

2.2. Image preprocessing

We divided the existing 180 images into two sets of 4:1, 144 as the training set and 36 as the test set. The generation of a deep learning based detection model requires training on a large amount of image data, which needs to be augmented in order to improve the robustness of the model and prevent overfitting. Image enhancement methods include image brightness enhancement and reduction, vertical flip, multi-angle rotation, etc. On the basis of the original picture, we first rotate the original picture by 2° and 5°, and then add Gaussian filtering and vertical flip to the rotated picture. Then perform 10° rotation and Gaussian filtering on the original image, and finally rotate 350° and add random brightness. The detailed process of the image enhancement method is shown in Fig. 2, with no overlap between the training and test sets. Fig. 3.

Table 2
Examples and numbers of the cells annotated.

Cell Types	Number of Cells Annotated	Number of trainset cells annotated	Number of validation sets cells annotated	Example
Eosinophil	308	226	82	
Lymphocyte	3547	2753	794	
Macrophage	2070	1655	415	
Neutrophil	649	427	222	
Total	6574	5061	1513	/

2.3. Yolov5s network architecture

Yolov5 is divided into Yolov5n, Yolov5s, Yolov5m, Yolov5l and Yolov5x, which have the same overall network structure, except that each sub-module uses a different depth and width, corresponding to the depth_multiple and width_multiple parameters in the sub-module yaml file. We have chosen Yolov5s as the base model, with some refinements to ensure accurate detection and a lightweight design of the parameters.

Yolov5 and Yolov4 use the same Mosaic data augmentation on the input side. It is stitched together using 4 images, randomly scaled, randomly cropped and randomly arranged. This method has more details and textures in the background area of the image, which enhances the network's understanding of the background and detection accuracy. In Yolov5, for different datasets, there are anchors with initial set length and width. In the network training, the network outputs predicted bounding boxes based on the initial anchors, and then compares them with the ground truth bounding box, calculates the difference between them, and then iterates the network parameters by backpropagation.

As shown in Fig. 4, Yolov5s backbone network consists of the Conv, C3 and spatial pyramid pooling-fast (SPPF) modules, and in this new version Yolov5 has replaced the Focus module, which was the first layer of the network, with a 6 × 6 sized convolutional layer. The Conv module is Yolov5s basic convolution unit, which performs 2D convolution, regularization and Swish activation function (SILU) activation function operations on the input. The Bottleneck module first reduces and then expands the number of channels (to half by default) by first halving the number of channels with 1 × 1 convolution, then doubling the number of channels with 3 × 3 convolution, and acquiring the features (using two standard convolution modules in total), without changing the number of channels in the input and output. In the new version of Yolov5, the authors have transformed the Bottleneck block with a CSP connection (BottleneckCSP) module into a C3 module. The C3 module is the main module for learning residual features and is structured in two

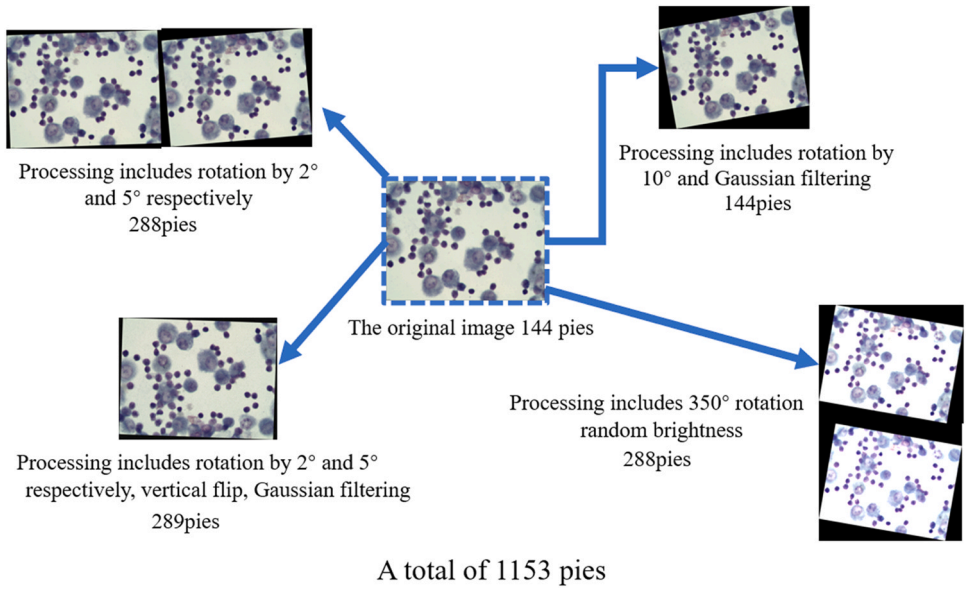


Fig. 2. Specific operations for image enhancement.

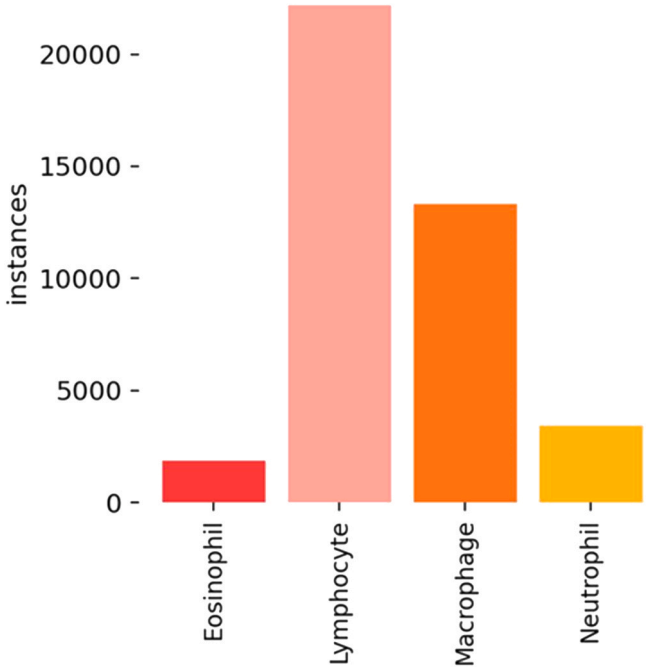


Fig. 3. Total number of individual labels in the processed image.

branches, one using multiple Bottleneck modules and one Conv module, the other passing through only one Conv module and finally concatting the two branches and passing through one more Conv module. The SPPF module is a spatial pyramidal pooling module that performs maximum pooling, with different kernel sizes and fuses features by connecting them to solve the target multiscale problem to some extent.

Yolov5's Neck is similar to Yolov4 in that it uses an Feature Pyramid Networks (FPN) + Path Aggregation Network (PAN) structure; the FPN is top-down, passing the high-level feature information through upsampling to fuse it and obtain a feature map for prediction, while the feature pyramid conveys strongly localized features from the bottom-up, aggregating parameters from various backbone layers to various detection layers.

Head is mainly used for the final detection part of the model, where the anchor boxes are filtered using Non Maximum Suppression (NMS). There are three detection layers with different sized feature maps, each outputting a corresponding vector, and finally generating and labelling predicted bounding boxes and classes of objects in the original image. Figs. 5 and 6.

2.4. Improvement of Yolov5s network architecture

2.4.1. Improvement of the backbone network

It has been shown that convolutional operations only extract features from local neighborhoods, while ignoring global feature information [29]. Self-attention can extract contextual information about the images and learn more semantic features [30]. For a single-head self-attention, the output of each pixel can be calculated using the following equation:

$$y_{ij} = \sum_{a,b \in N_k(i,j)} softmax_{ab}(q_{ij}^T K_{ab}) v_{ab} \quad (1)$$

where $q_{ij} = W_Q x_{ij}$, $k_{ab} = W_K x_{ab}$, $v_{ab} = W_V x_{ab}$ implies linear pixel points and changes in surrounding pixel points, and W_Q, W_K, $W_V \in R^{d_{out} \times d_{in}}$ are the parameters that the network needs to learn. A schematic representation of the multi-head self-attention is given in Fig. 7. During self-attention, the input embeddings are transformed into query, key, and value vectors. The model then computes a weighted sum of the values, where the weights are determined by the dot product of the query and key vectors. The resulting output is a weighted combination of the input embeddings, where the weights represent the importance of each input embedding for the current context.

As the Transformer architecture is a more geometric attention mechanism that allows different attention to be given to different features, the interference of slice priming, impurities, in BALF cells on recognition results is somewhat suppressed. Specifically, it allows information to flow freely at different locations in the cell image, establishing remote dependencies and providing better robustness in the recognition of obscured cells. For the target detection task, in order to introduce a self-attention, multiple convolutional layers need to be superimposed to build a larger scale model that aggregates all the local features extracted by the convolution. Although the stacking of multiple convolutional layers can effectively improve the network's ability to extract target features, it also leads to an

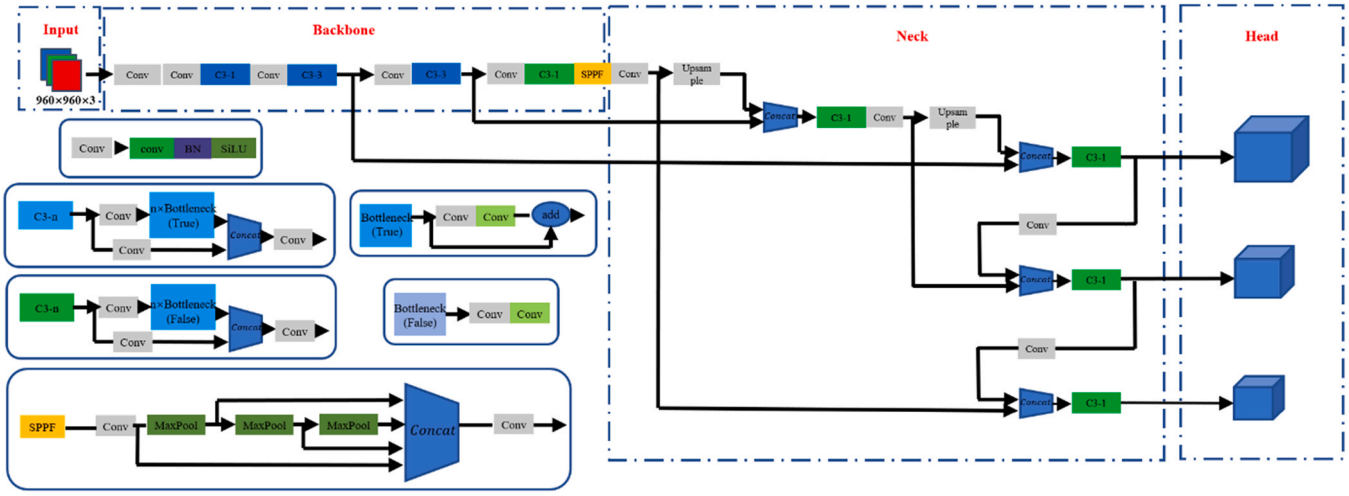


Fig. 4. Network structure of Yolov5s and its modules.

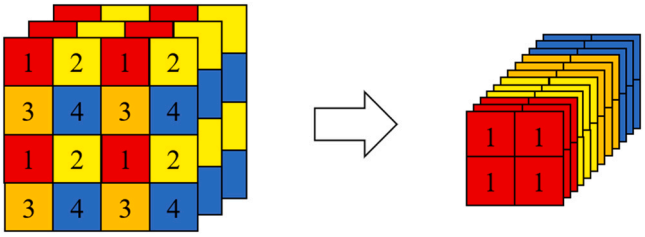


Fig. 5. Focus Module.

increase in the depth and computational effort of the network layers, so in order to introduce the self-attention while minimizing the computational effort, we initially considered using the Swin Transformer [31], which solves the problem of computational complexity in the Transformer in a very reasonable way. The Swin Transformer has a linear computational complexity with respect to image size, which means that the amount of computation required to process an image increases linearly with the number of pixels in the image. This is achieved through the use of a hierarchical architecture and a shifted windowing scheme that limits self-attention computation to non-overlapping local windows while allowing for cross-window connection.

As shown in Fig. 8(a), the Swin Transformer constructs a hierarchical representation by starting with small patches (grey contours) and gradually merging adjacent patches in deeper Transformer layers. It uses hierarchical feature mapping to facilitate dense prediction. Linear computational complexity is computed locally within non-overlapping windows of image partitions for self-attention (red contours), rather than over all patches of the entire9 image. The number of patches in each window is fixed, so the complexity is linearly related to the image size.

Swin Tranformer reduces the amount of computation required, but still leads to instability in training when deepening the number of model layers. Swin Tranformer V2 uses the normalization operation that comes with cosine to calculate the Attention and then normalizes the output to stabilize the output [32]. Swin Tranformer V2 also uses a two-layer multi-layer perceptron (MLP) to adaptively compute additional pixel values for relative position coding, which is more flexible than the usual binomial cubic difference, improving the model's adaptability to high resolution, and redefines the relative position coding to use logarithmically spaced coordinates instead of the original linearly spaced coordinates, as in Eq. (2) :

$$\hat{\Delta x} = sign(x) \cdot \log(1+|\Delta x|)$$
$$\hat{\Delta y} = sign(y) \cdot \log(1+|\Delta y|) \quad (2)$$

where Δx, Δy, $\hat{\Delta x}$, $\hat{\Delta y}$ are the linear scale coordinates and logarithmic space coordinates respectively. This operation ensures the accuracy of the relative position encoding and keeps the extrapolation within an acceptable range. Fig. 9.

We finally chose to introduce Swin Transformer V2, which improves on both of the above, as one of the backbone layers into the Yolov5 network structure, replacing the original C3 layer to improve the model's ability to process cell image information. The details of this addition are shown in Fig. 10.

In summary, for cell detection tasks, BALF cell images inherently suffer from low contrast and small differences in the number and shape of various types of cells. By introducing Swin Transformer V2, we have introduced a self-attentive mechanism with less impact on computational effort to enhance detection by effectively capturing global features, helping the skeleton network to focus on useful cellular objects and reducing the interference of confusing information. In addition, it additionally stabilizes the output values of the model and improves its adaptability to downstream tasks compared to Swin Transformer.

2.4.2. Improvement of the neck network

GhostModule is a plug-and-play innovative module proposed in GhostNet, which can use fewer parameters and computations to obtain more feature maps, making the network structure lighter [33]. The core idea is to divide the original convolution operation into two stages, where the first stage is a small number of convolution calculations in the segment, while the second stage is to perform linear convolution again on top of the feature maps obtained in the first stage, generating new feature maps and combining them together at the end to obtain a large number of feature maps. In this process, the new feature maps are called Ghost maps of the previous feature maps. the Ghost module is mainly used to eliminate redundant features and obtain a lighter model. For target detection tasks such as cells, the Ghost module helps the model to better capture detailed information and compensate for the incomplete extraction of Yolov5 detailed features, while also effectively reducing parameters. The working principle is shown in Fig. 11, which illustrates the specific operation of the convolutional layer and the implementation of the Ghost module.

Specifically, assume that the output feature map has height H, width W, and channel C. The size of the conventional convolution

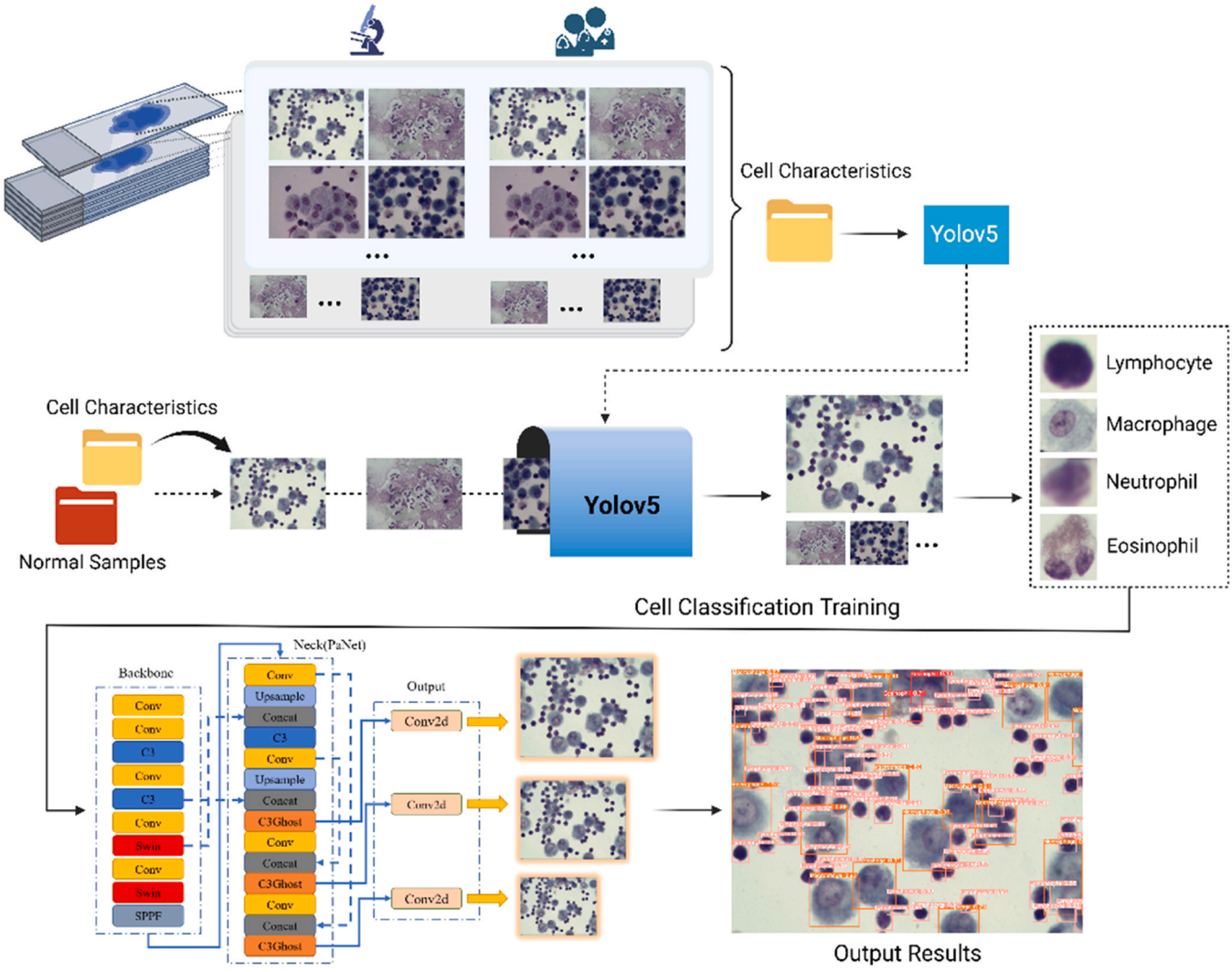


Fig. 6. The two-stage process of our framework (a) Training of the model: Our optimized Yolov5s model, which focuses on the detection of cells in a given image, is trained with annotated bronchoalveolar lavage fluid cells. (b) Use of the model: By using our optimized Yolov5s model, the cells detected in BALF and their corresponding probabilities can be obtained.

kernel is $k \times k$. The Ghost module has a linear operation of Identity and $(O-1)$ cheap transform, where and $O \ll c$. The convolution kernel for the linear operation is $a \times a$. Thus, comparing the computation consumed by the conventional convolution with that consumed by the Ghost convolution module, the Ghost convolution module has a speedup ratio R of:

$$R = \frac{N \times H \times W \times C \times k \times k}{\frac{N}{O} \times H \times W \times C \times k \times k + \frac{N}{O}(O-1) \times H \times W \times a \times a}$$

$$= \frac{C \times k \times k \times O}{C \times k \times k + (O-1) \times a \times a} \approx O\# \tag{3}$$

The structure of the GhostBottleneck and C3Ghost is shown in Fig. 12. The GhostBottleneck consists of two stacked GhostModules. The first GhostModule is utilized as an extension layer to increase the number of channels. The second GhostModule reduces the number of channels matching the shortcut path and is connected in the middle with a depthwise convolution (DWConv) with a step size equal to 2. The inputs and outputs of the two GhostModules are then connected using shortcut. The second GhostModule applies the Batch Normalization and rectified linear unit (ReLU) activation functions after each layer. The two Ghost modules are connected by deep convolution in steps of 2 to continuously deepen the refinement and capture of image features. Compared to the C3 module, this module not only reduces the model size, but also facilitates the capture and extraction of detailed features.

This module is reusable, so in this paper we chose C3Ghost, constructed from GhostBottleneck, as the main network for the feature extraction part of the YoloV5 network, replacing the three C3 modules that were originally located in the Neck network. We also tried to use the Ghost module as the backbone of Yolov5, but after experiments (As shown in Table 4), we showed that it was not as effective as replacing it on the Neck part.

2.4.3. Loss function improvement

2.4.3.1. Bounding box regression score improvement. The Yolov5 loss function consists of a bounding box regression score, an objectness score and a class probability score. In the bounding box regression score, the complete intersection over union loss (CIOU_Loss) is used to achieve the prediction [34], as shown in Eq. (4).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + av\# \tag{4}$$

$$a = \frac{v}{(1 - IoU) + v}\# \tag{5}$$

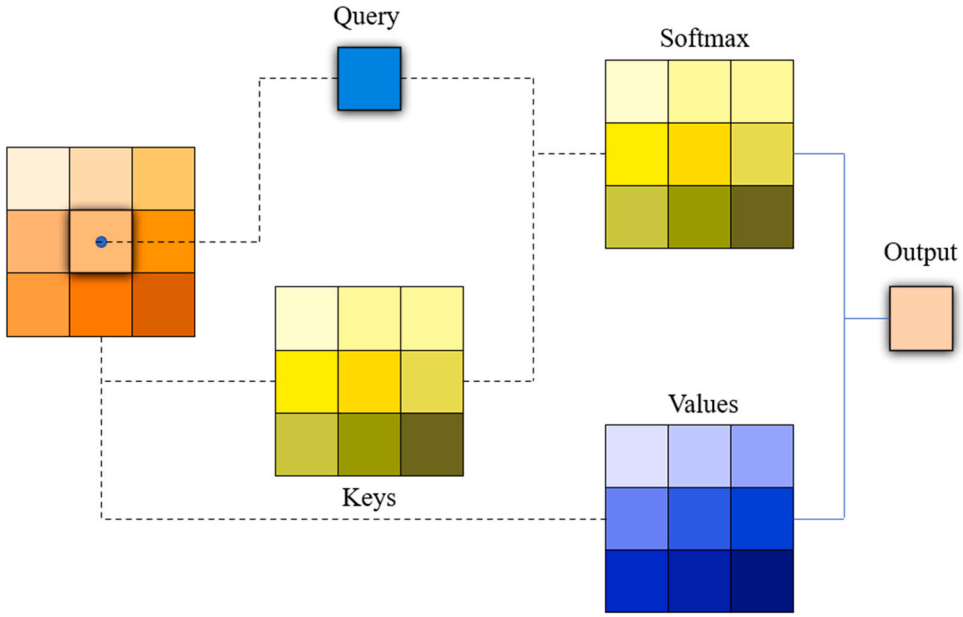


Fig. 7. Diagram of the multi-head self-attention.

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \# \tag{6}$$

$$IoU = \frac{A \cap B}{A \cup B} \# \tag{7}$$

where b, b^{gt}represent the centroids of the prediction bounding box and the ground truth bounding box respectively, ρ represents the Euclidean distance calculated between the two centroids, c represents the diagonal distance of the smallest closed region that can contain both the prediction bounding box and the ground truth bounding box, w, w^{gt}represent the width of the prediction bounding box and the ground truth bounding box respectively. h, h^{gt}represent the width of the prediction bounding box and the ground truth bounding box respectively, and IoU is the ratio of the intersection and the concatenation between the prediction bounding box and ground truth bounding box.

CIoU_Loss considers the overlap area, centroid distance and aspect ratio of the bounding box regression, but ignores the real difference between the width and height respectively and their confidence levels, which hinders the effectiveness of model optimization. To address this problem, Zhang et al. proposed an efficient intersection over union loss (EIoU_Loss) by splitting the aspect ratio on the basis of CIoU_Loss [35]. EIoU_Loss is calculated by splitting the aspect ratio influence factor to calculate the length and width of the ground truth bounding box and anchors. An overlap loss, a center distance loss, and a width-height loss are all included in the loss function, with the width-height loss directly minimizing the difference between the ground truth bounding box and the anchor. EIoU_Loss is shown in Eq. (8).

$$L_{EIoU} = L_{IoU} + L_{dis} + L_{asp} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \# \tag{8}$$

where $c^w$and $c^h$are the width and height of the minimum external frame covering the prediction bounding box and the ground truth bounding box. Compared to the CIoU_Loss in the original network, the width and height loss in the EIoU_Loss accelerates and improves convergence in the border regression loss, so the better

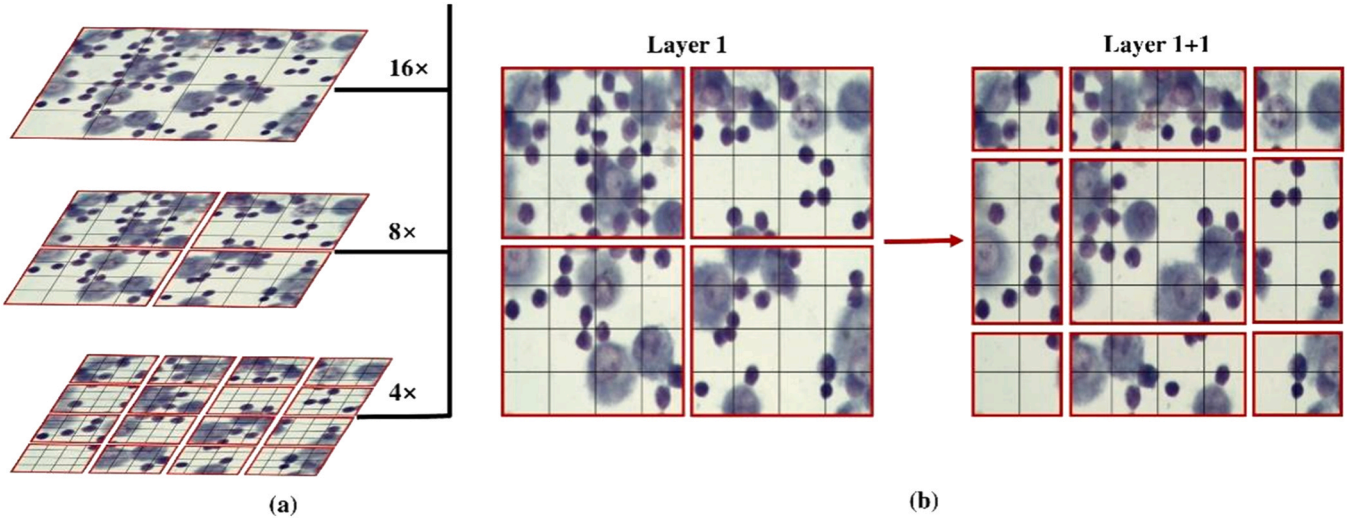


Fig. 8. Layered representation of the Swin Transformer (a) and sliding windows (b).

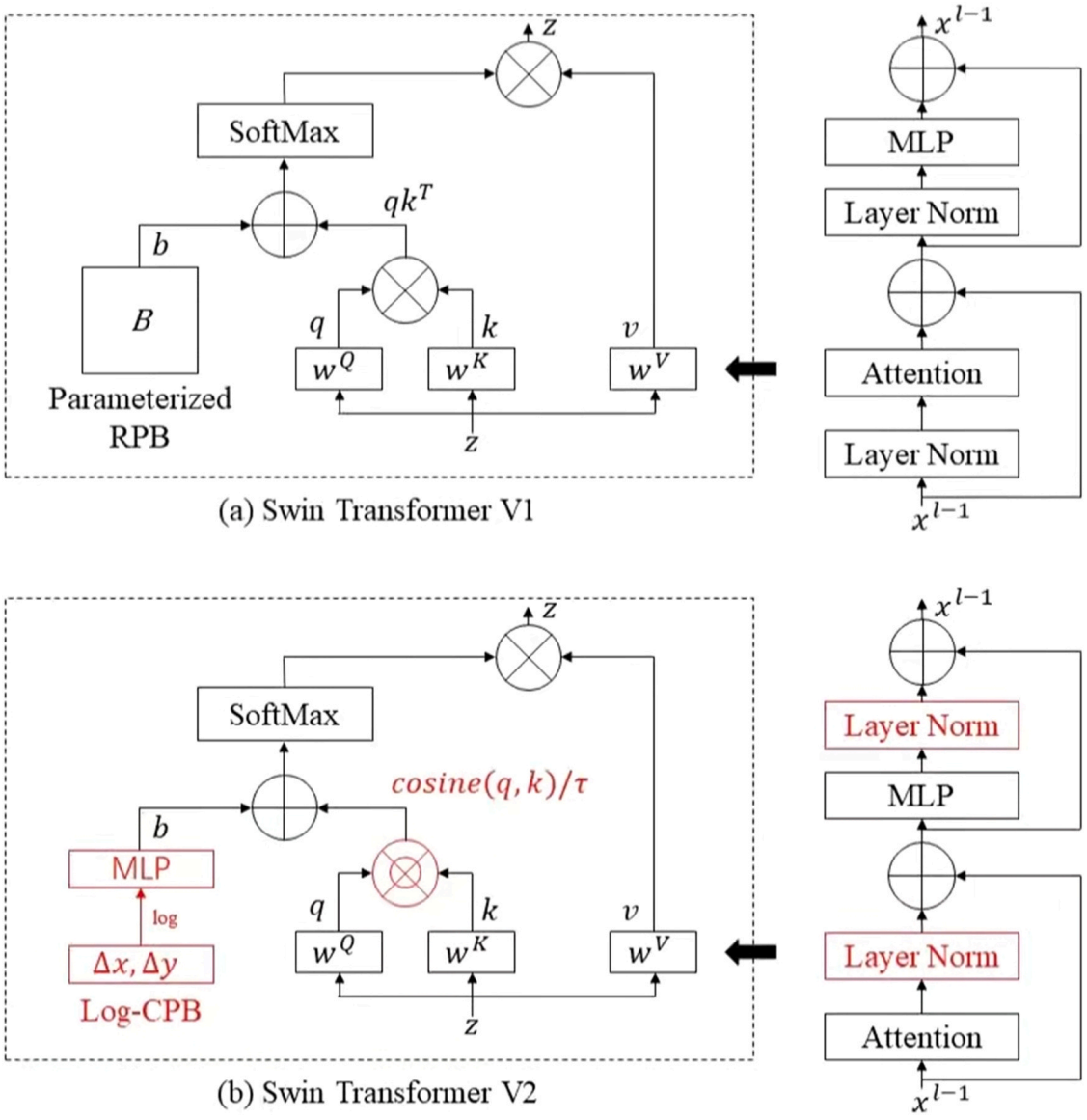


Fig. 9. Comparison of the structure of the Swin Transformer and Swin Transformer V2.

performing EIoU_Loss border regression loss function is used in this paper.

2.4.3.2. Positive and negative sample allocation strategies. Among the four cell types of BALF, there are differences in the number and distribution of cells, resulting in an uneven distribution of the number of samples in the training data set. Due to the density of the sliced cells, it is not possible to maintain an even distribution of samples even with a cropping operation on the images, for which we introduce a new sample distribution strategy.

Yolov5's positive and negative sample delineation strategy is based on an anchor-based strategy. Before starting training, 9 anchors are obtained a priori based on the ground truth (GT) in the training set by the k-means clustering algorithm, arranged from smallest to largest. Afterwards, each GT is traversed and matched with the 9 anchors, and the aspect ratio between the GT and the 9

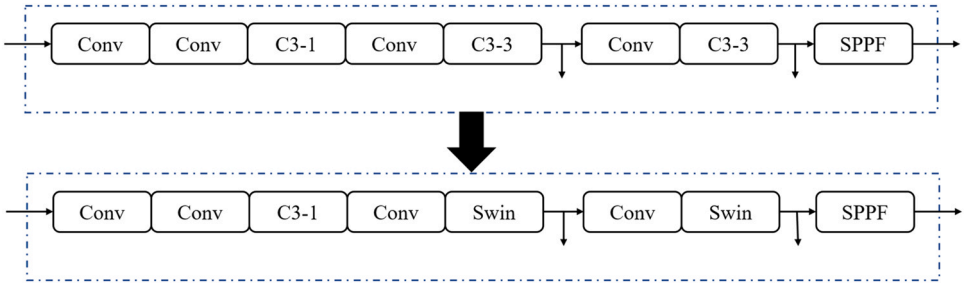


Fig. 10. Improved backbone network.

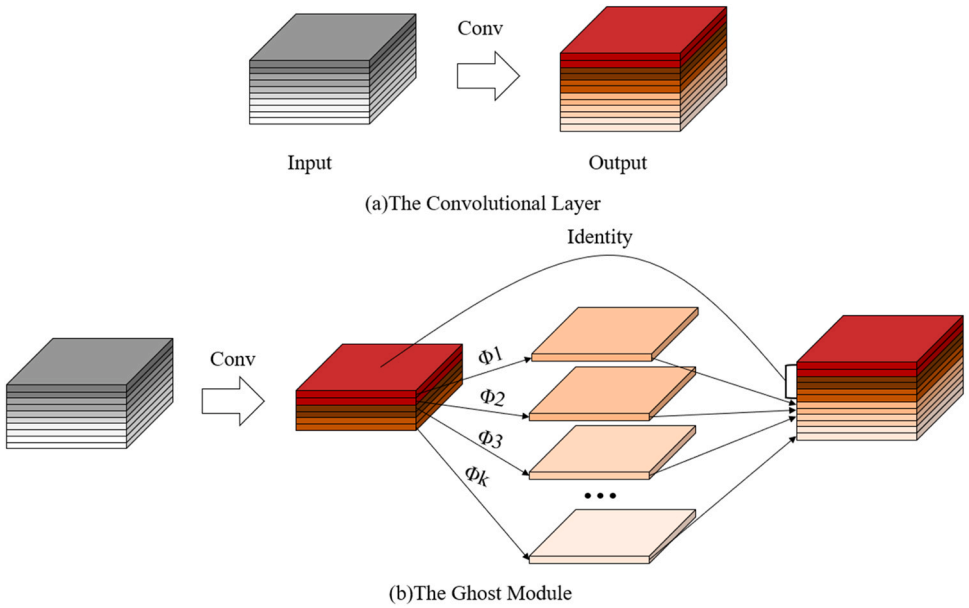


Fig. 11. The Convolutional Layer and Ghost Module.

anchors is calculated. If the aspect ratio is less than a set threshold, it means that the GT is a positive sample of the current feature map. Since negative samples are not involved in training in Yolov5, the number of positive samples should be increased. GT is matched with the anchors to get the corresponding grid of the anchor, and see which grid the GT centroid falls on. Not only the anchor in the grid where the GT centroid falls and the GT matches are taken as positive samples, but also the anchors in the two adjacent grids are taken as positive samples. Figs. 13 and 14.

Simulation-based online training algorithm (SimOTA) was proposed by Ge, Zheng et al. [36], which serves to select different numbers of positive samples for different targets. SimOTA can be understood as a method of matching strategies on how to achieve the lowest cost of assigning these anchors to GT. The cost function is described in Eq. (9).

$$c_{ij} = L_{ij}^{cls} + \lambda L_{ij}^{reg} \# \tag{9}$$

where λ is the balance coefficient, and L_{ij}^{cls} and L_{ij}^{reg} denote the bounding box regression score and class probability score between the ground truth bounding box and the prediction bounding box, respectively. That is, for all anchors on a feature map, the policy cost of the entire matching is the sum of the bounding box regression score and class probability score generated by all feature points and each GT. Yolov7 proposes a new sample allocation strategy, compute loss online training algorithm (ComputelossOTA) [37], which is also based on anchor based, drawing on ideas related to Yolov5 and Yolox. ComputelossOTA uses the strategy of Yolov5 for screening positive samples in the first step, followed by the further screening strategy of SimOTA. The detailed steps are as follows.

(1) First match each GT with 9 anchors, and also take the two adjacent anchors in the grid as positive samples.

(2) Calculate the IoU of the first screening positive samples and GT, and sort the IoU from the largest to the smallest, and take the sum of the top ten and round it up to b.

(3) Calculate the cost function of the initial sieve of positive samples, and rank the cost function from smallest to largest, and take the top b samples as positive samples. Also consider the case where the same grid prediction bounding box is associated with two GT, and take the smaller value of the cost function, the prediction bounding box is the positive sample of the corresponding GT.

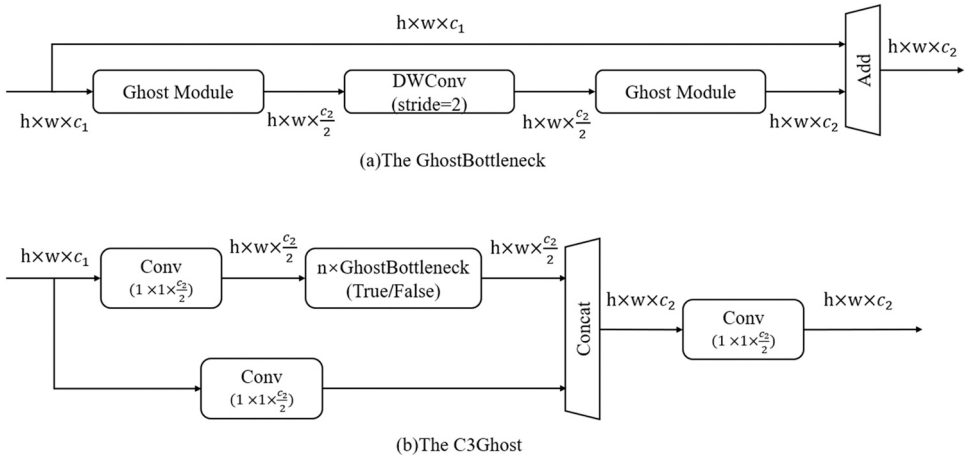


Fig. 12. The GhostBottleneck and The C3Ghost.

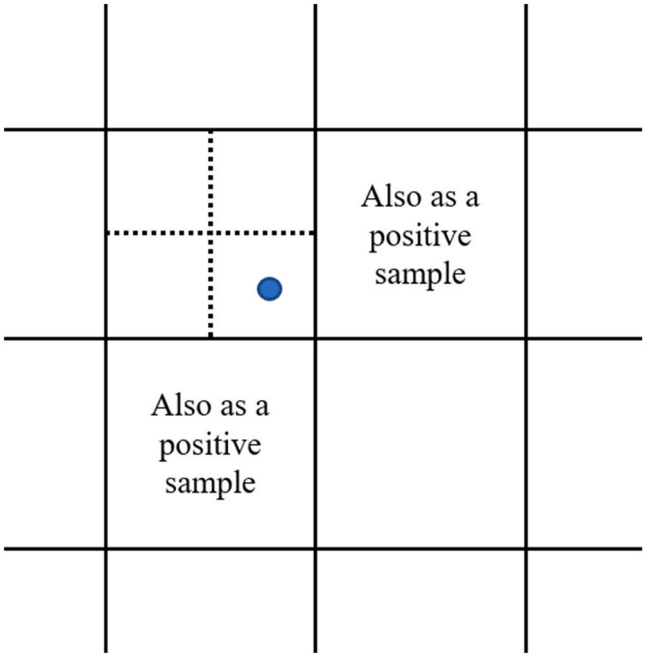


Fig. 13. Positive and negative sample allocation strategies for Yolov5.

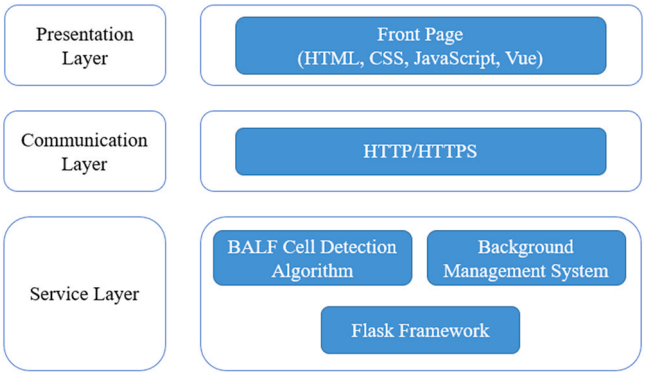


Fig. 14. System Technical Architecture Diagram.

The Yolov7′s positive and negative sample allocation strategy incorporates a cost function compared to the Yolov5′s strategy, which enables a further fine screening using the performance of the current model. The Yolov7′s strategy, on the other hand, is able to provide more accurate prior knowledge compared to using only SimOTA. We introduced the Yolov7′s positive and negative sample strategy into our model, increasing the number of positive samples while maintaining the quality of the anchor, and this change has been shown to improve the accuracy of BALF cell detection and classification significantly.

2.4.4. Weighted non-maximum suppression

The Non-Maximum Suppression (NMS) algorithm is used to suppress elements other than extreme values [38]. In the Yolo family of algorithms other candidate bounding box with low accuracy can be removed, leaving the best candidate bounding box with the highest accuracy. Taking target detection as an example, the target detection inference process generates many detection bounding boxes, many of which detect the same target, but ultimately only one bounding box is needed for each target. The traditional NMS algorithm is to sort the candidate bounding box by confidence, and the one with the highest confidence is used as the benchmark bounding box, calculate the IoU with other bounding box, delete the ones with values greater than the threshold and keep the ones smaller, and so on, and finally get the best candidate bounding box with the highest accuracy. The traditional NMS algorithm is defined as follows:

$$s_i = \begin{cases} 0, IoU(M, B_i) \geq thresh \\ s_i, IoU(M, B_i) < thresh \end{cases} \# \tag{10}$$

where s_i is the current confidence size; M is the benchmark bounding box with the highest confidence level; B_i is the other candidate bounding boxes with which the benchmark bounding box calculates the IoU.

However, the best candidate bounding box obtained from each traversal of the traditional NMS algorithm is not necessarily the precise one, and there is a possibility that the better candidate bounding box is mistakenly deleted. The weighted Non-Maximum suppression (Weighted-NMS) algorithm used in this paper obtains more accurate candidate bounding boxes by performing a weighted average of the benchmark bounding box M and the candidate bounding boxes larger than a threshold, rather than a deletion operation [39]. The weighting formula is as follows: Eq. (12) is the weight added by the weighting.

$$M = \frac{\sum_i w_i B_i}{\sum_i w_i}, B_i \in \{B \mid IoU(M, B) \geq thresh\} \cup \{M\} \# \tag{11}$$

$$w_i = s_i IoU(M, B_i) \# \tag{12}$$

Specifically, Weighted-NMS assigns a weight value to each bounding box based on the confidence level for a batch of bounding boxes with a high repetition rate, with the higher confidence level giving a higher weight because the higher confidence level gives a reason to value the bounding box more. So, for all the prediction bounding boxes a confidence weight is multiplied, and a weighting and averaging of the prediction bounding box information is done. The use of a weighting algorithm allows for a more accurate placement of the bounding boxes and improves the accuracy and recall of the overall algorithm. Therefore, the Weighted-NMS algorithm is used in this paper to replace the traditional NMS algorithm.

2.5. Web development of BALF cell detection

The experiments show that the bronchoalveolar lavage cell detection and classification algorithm proposed in this paper has achieved significant improvement in detection accuracy and light weight, and has a good application prospect. Based on the algorithm, we developed a web application to display the detection results clearly and intuitively. In addition to basic cell image detection, this support detects cell images enhanced with HSV, which expands the application scenarios.

The technical architecture of the application is mainly composed of three parts: presentation layer, communication layer and service layer. The presentation layer uses the Vue technology stack to build pages through HTML, JavaScript, and CSS, while the communication layer uses the Hypertext Transfer Protocol (HTTP) to establish information transmission connections. As the core layer of the system, the service layer completes the BALF cell detection task according to the proposed optimization algorithm, and returns the detection results to the front-end page for users to view.

Fig. 15 shows the system page layout. Users upload local images, and the original images will be displayed in the left area, which will be sent to the cell detection algorithm for target recognition, and the recognized and labeled detection images will be displayed in the right area. Click on the test result graph to display the marked image, which is convenient for observing and analyzing the test results. The detected target category, target size, confidence and other data are displayed in a table at the bottom of the page, which is convenient for objectively checking the performance of the algorithm.

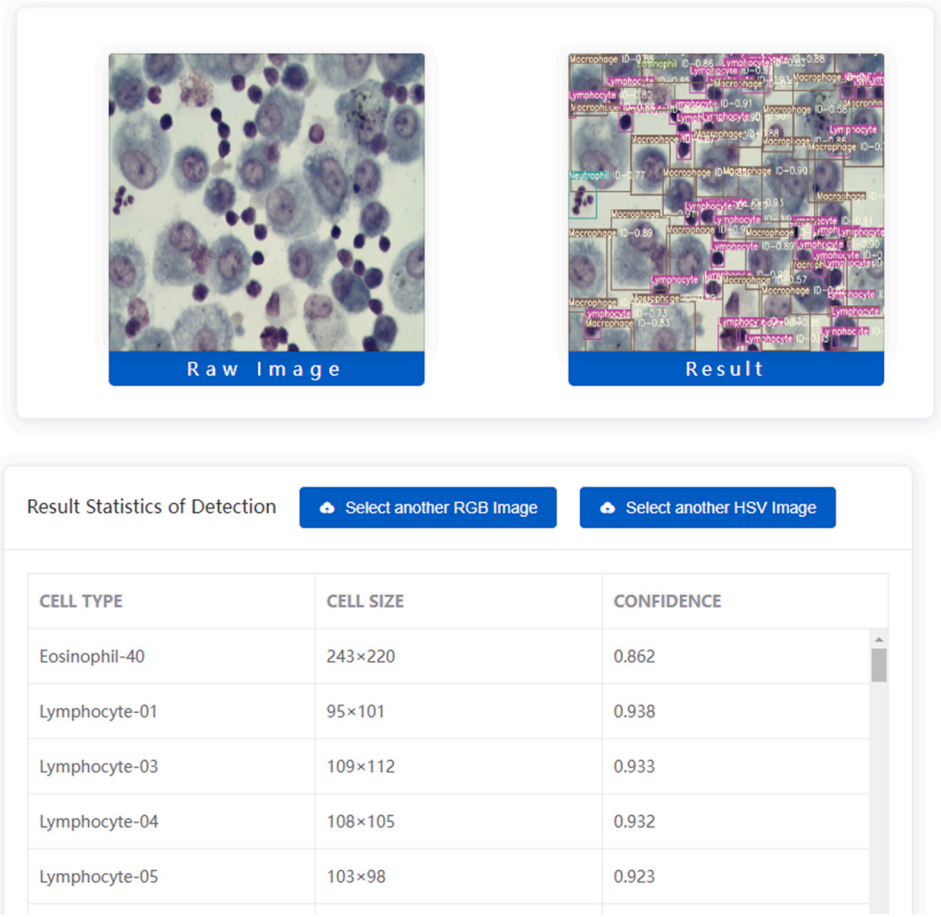


Fig. 15. Web Page Layout.

2.6. Overall network model structure

To achieve the accuracy of the BALF cell detection model, we proposed Improved Yolov5 Based on Transformer Backbone Network for Detection and Classification of Bronchoalveolar Lavage Fluid. Fig. 16 shows the network structure of our model. As shown, the Swin Transformer Layer (Red) replaces the last two C3s in the Yolov5s backbone network. The C3Ghost module (Orange) is introduced into the Yolov5s neck network. The purpose of the Swin Transformer Layer is to obtain global information in BALF cell detection to enhance the feature extraction capability of the network, thus improving the accuracy of the model for BALF cells in multiple

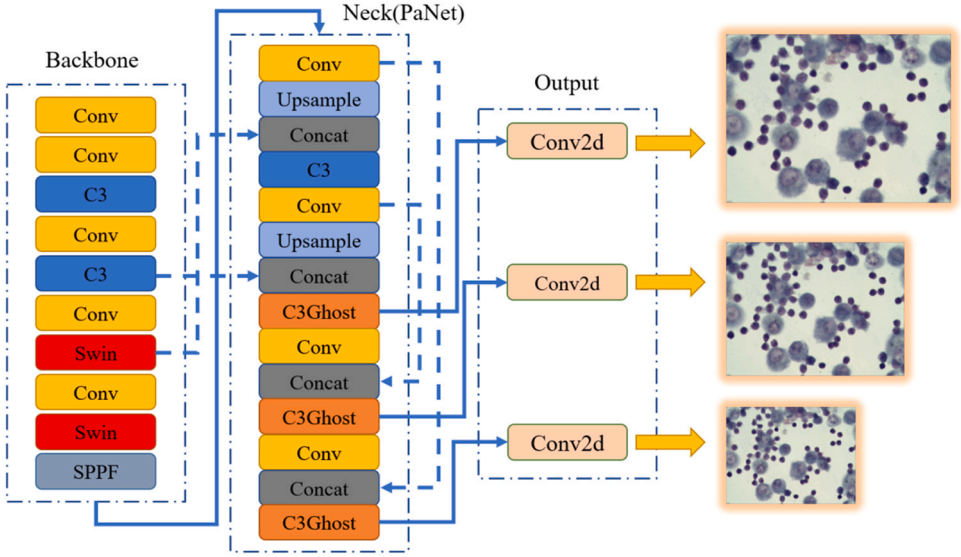


Fig. 16. The network structure of our model.

scenarios. The C3Ghost module is designed to further reduce the parameters of the model and facilitate the capture and extraction of detailed features.

3. Experimental results

3.1. Experimental setting

We used Windows and the Pytorch deep learning framework to train the model. The software environment was CUDA 11.3, CUDNN 8.2, and Python 3.9. The CPU used for training the dataset was a 12th Gen Intel(R) Core (TM) i7–12700 H 2.30 GHz 32 G and the GPU was a GeForce RTX 3070Ti Laptop GPU 16 G.

Yolov5's anchor is based on the dataset and calculated using the k-means clustering algorithm, which may also differ in datasets with different object scales and aspect ratios. When the best recall is greater than or equal to 0.98, there is no need to update the anchor. The default anchor in this dataset has a best recall of 1.000, so there is no need to update the anchor box. Note: The default anchor is [10,13, 16,30, 33,23], [30,61, 62,45, 59,119], [116,90, 156,198, 373,326].

3.2. Evaluation indicators

Intersection over Union (IoU) is a commonly used evaluation metric in object detection and image segmentation tasks. IoU is a measure of the overlap between two sets, where the sets in question are usually the predicted bounding boxes and the ground truth bounding boxes. IoU is calculated as the ratio of the area of the intersection between the predicted and ground truth bounding boxes, to the area of their union. IoU can be expressed as a value between 0 and 1, where a value of 0 indicates no overlap between the predict and ground truth bounding boxes, and a value of 1 indicates perfect overlap.

Precision refers to the proportion of true positives among all samples predicted as positive by the detector.

Objective evaluation metrics such as *recall rate*(*Recall*) and mean average precision (*mAP*) were used in the study to evaluate the performance of the trained bronchoalveolar lavage fluid cell recognition model. The calculation equations are as follows:

$$Recall = \frac{Ttue\quad positive\quad (TP)}{Ttue\quad positive\quad (TP) + False\quad negative\quad (FN)}\# \tag{13}$$

$$mAP = \frac{1}{C}\sum_{K=i}^{N} P(k)\Delta R(k)\# \tag{14}$$

where *TP* is the number of correctly classified positive cases, *FP* is the number of misclassified negative cases, *FN* is the number of misclassified positive cases, and *TN* is the number of correctly classified negative cases. *C* represents the number of cell target categories; *N* represents the number of *IoU* thresholds, *K* is the *IoU* threshold, $P(k)$ is precision, and $R(k)$ is *Recall*. *mAP*@0.5 refers to the average *AP* of all classes when *IoU* is set to 0.5, and *mAP*@0.5: 0.95 refers to the average *mAP* under different *IoU* thresholds. The *IoU* ranges from 0.5 to 0.95 in steps of 0.05. The size of the model saved after final training is referred to as its model size. FPS stands for "Frams Per second".

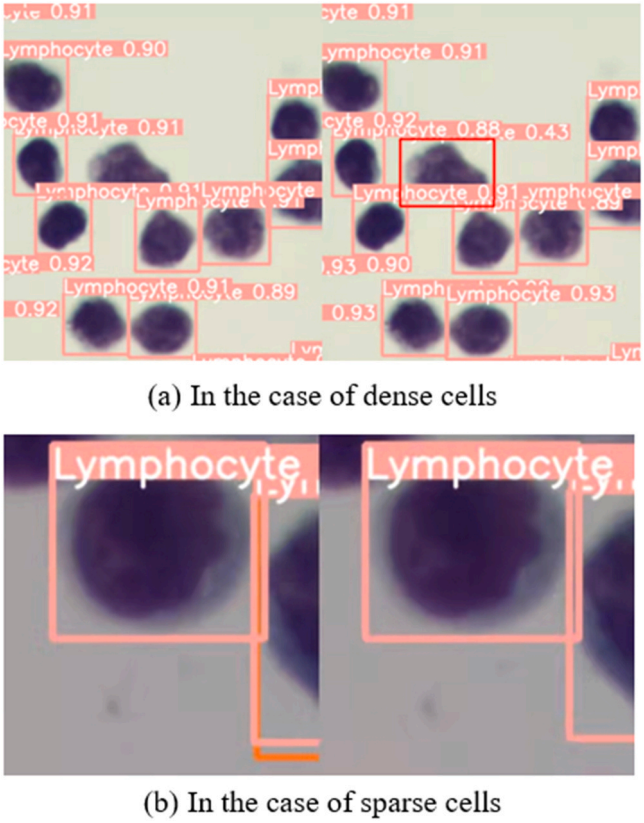
3.3. Results

We conducted four sets of experiments using the dataset provided by the Xunfei platform "Bronchoalveolar Lavage Cell Sorting Challenge" [23]. Firstly, we show the accuracy of the optimized model compared to the original and visualize the results to demonstrate the superiority of our model. Secondly, we conducted an ablation experiment with the optimized Yolov5 model, where we added six improvements to the Yolov5 model, each of which improved the accuracy and validity of the original model. In addition, to further investigate the performance of our method, we compare it with other classical backbone models of Yolov5s, and our model performs best in terms of accuracy. Finally, we compare the number of parameters, FPS and accuracy of our model with other classical object detection algorithms. The mAP of our model is 0.02% and 2.36% higher than the recently launched Yolov7 and Yolov8s, respectively, and the FPS value is the largest. Experimental results show that our model outperforms other models in both accuracy and efficiency.

3.3.1. The comparison results with Yolov5s and results visualization

Fig. 17 (a), (b) shows the performance of Yolov5s compared with the method proposed in this paper in both dense and sparse cell scenarios. The left panel shows the detection results of the original Yolov5s version, while the right panel shows the detection results of the model in this paper. As shown in Fig. 17 (a), in the case of dense cells, the original Yolov5s (left) version may miss detection in some images, while the method (right) in this paper has a better detection effect in the case of dense cells due to the introduction of Transformer to obtain global information, which effectively reduces the cases of missing detection. In Fig. 17 (b), the original Yolov5s (left) version is more likely to have redundant and overlapping bounding boxes in the case of sparse cells. Our model (right) uses Weighted-NMS to weight the prediction bounding boxes, which can reduce the appearance of redundant bounding boxes to a certain extent. By removing redundant frames, the analysis can focus on unique frames, resulting in better accuracy in detecting the number and location of cells. Removing redundant frames can also speed up the processing time for cell detection, as fewer frames need to be analyzed. This can be especially important when analyzing large datasets or real-time experiments where rapid detection is necessary.

The outcomes of a single class comparison on the test set between the benchmark model and our model are shown in Fig. 18. The



(a) In the case of dense cells

(b) In the case of sparse cells

Fig. 17. Comparison of test results.

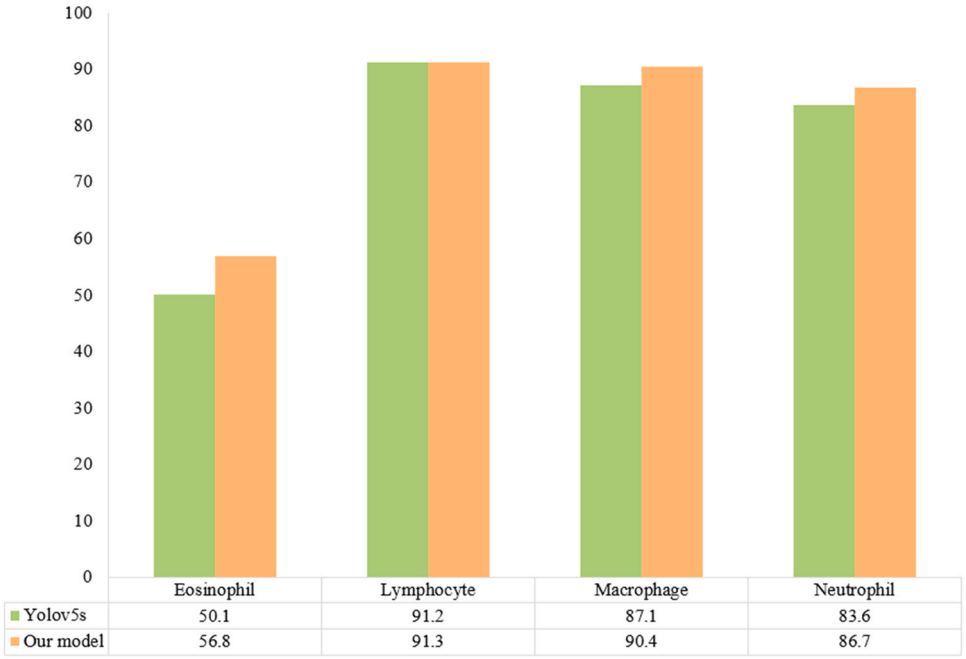


Fig. 18. Four types of cell test results.

findings demonstrate that, for all four cells, the single-class AP values of our model are greater than the benchmark model values, with the original version of Eosinophil, which had the lowest accuracy, showing the most improvement and detecting better than Yolov5s. The strategy still outperforms Yolov5s in multi-class target detection tasks, demonstrating the applicability of the suggested approach.

Furthermore, based on the confusion matrix provided in Fig. 19, our proposed method has a recall of more than 0.8 for three of the four BALF cell types and an improvement of 0.9% for the Eosinophil cell category relative to the original version, with the accuracy of the model detection meeting the needs of realistic applications.

3.3.2. Ablation experiment

In our ablation experiments, we verified the effectiveness of our improvement points, as shown in Table 3.

- We processed 180 images from the original dataset into 1189 images using three methods in data enhancement: image brightness enhancement and reduction, vertical flip, and fixed angle rotation. The training of a deep learning-based target detection model relies on a large amount of image data as the training set in order to improve the generalization ability of the target detection model and prevent overfitting of the model. After we expanded the data volume using data augmentation, the Recall of the Yolov5s model for the BALF cell detection model increased by 2.45%, demonstrating that our data pre-processing can significantly improve the detection coverage of the network.

- Using Swin Transformer V2, we introduced a self-attention mechanism to extract richer geometric features around pixel points, while maintaining linear computational complexity. Also, in terms of migration learning of the model, it is better adapted to high-resolution downstream tasks after training on low-resolution datasets, with stable model output and generalization capability. With the addition of Swin Transformer V2, the overall improvement in mAP(@0.5) was significant at 1.21%, although Recall decreased.

- The introduction of the Ghost module into the Neck network resulted in a decrease in the number of parameters and an increase in the ability to obtain feature maps from the neck network, with a steady increase in both mAP(@0.5) and Recall. We have also tried to introduce the Ghost module into the backbone network (see Table 3), but experiments have shown that it works better with the Neck network, which is mainly used to eliminate redundant features and obtain a lighter model. For target detection tasks such as cells, the Ghost module helps the model to better capture and extract detailed information, compensating for the incomplete extraction of detailed features from YoloV5, and also effectively reduces parameters.

- We then modified the loss part of Yolov5s to introduce the positive and negative sample assignment strategy of Yolov7, which adds a cost function to the model compared to the Yolov5-only strategy, allowing for a further fine screening using the current model performance. It also provides more accurate a priori knowledge, increases the number of positive samples and guarantees the quality of the anchor. The experimental improvement in the detection and classification of BALF cells is significant.

- We modify the bounding box loss function to EIoU_Loss, and the width-height loss in the border regression loss in EIOU_Loss makes convergence faster and more accurate. In particular, for the detection of BALF cells, which have high cell density and high similarity, this change has a significant improvement in accuracy, with mAP(@0.5) improved by 0.59%.

- Finally, we used Weighted-NMS to filter the prediction bounding box information. Compared with the traditional NMS algorithm, the Weighted-NMS algorithm is a weighted average of the benchmark bounding box M and the candidate bounding box larger than the threshold, rather than a deletion operation, which can obtain more accurate candidate bounding boxes. The final results obtained are the best among all the experiments. Compared to the original Yolov5s, mAP(@0.5) improved by 3.3%, mAP(@0.5:0.95) improved by 0.93%, and Recall improved by 3.67%.

3.3.3. The comparison results with related methods

To further investigate the performance of our method, we compared it with other classical backbone models of Yolov5s, namely Yolov5s_all_Ghost, Yolov5s_ShuffleNetV2, Yolov5s_CoT. We chose mAP(@0.5), mAP(@0.5:0.95), and Recall as the evaluation metrics. Recall as the evaluation metric, and from the experimental results in

Fig. 19. Confusion matrix for 4 cell types.

Table 4, it is clear that our model performs best when compared to Yolov5s replaced by other backbone networks.

GhostModule is a plug-and-play innovative module that can use fewer parameters and computations to obtain more feature maps, making the network structure lighter. In this experiment, Yolov5s_all_Ghost differs from our model in that it replaces the convolutional layers in the backbone network of Yolov5s with the Ghost module, rather than the Neck network. However, experiments have shown that replacing the backbone network of Yolov5s with the Ghost module instead affects the performance of the model, with a reduction of more than 5% in all metrics compared to the original version of Yolov5s.

Table 4
The results for different backbone network.

Model	mAP (@0.5)	mAP (@0.5:0.95)	Recall
Yolov5s	77.99%	53.31%	76.80%
Yolov5s_all_Ghost	71.41%	45.75%	71.16%
Yolov5s_ShuffleNetV2	75.09%	48.92%	72.99%
Yolov5s_CoT	78.32%	52.65%	75.11%
Our model	81.29%	54.24%	80.47%

ShuffleNetV2 was developed by Ningning Ma et al. through extensive experiments to propose four lightweight network design guidelines, which provide a detailed analysis of the effects of input

Table3
The results for different improvement points.

Model	mAP (@0.5)	mAP (@0.5:0.95)	Recall
Yolov5s	77.99%	53.31%	76.80%
Yolov5s+Data Augmentation	78.77%(↑0.78%)	53.39%	79.25%
Previous+Swin Transformer	79.98%(↑1.21%)	53.84%	77.64%
Previous+GhostNet	80.30%(↑0.78%)	53.50%	79.87%
Previous+ComputelossOTA	80.54%(↑0.32%)	53.08%	79.94%
Previous+EIoU	81.13%(↑0.59%)	53.84%	80.39%
Previous+Merge-NMS	81.29%(↑0.16%)	54.24%	80.47%

and output channels, the number of grouped convolutional groups [40], the degree of network fragmentation, and element-by-element operations on the speed and memory access cost (MAC) on different hardware. The optimal trade-off between speed and accuracy is achieved by replacing group convolution with Channel Split, which satisfies the four design criteria. After experiments, it can be seen that the detection accuracy of ShuffleNetV2 is slightly degraded compared to the original version.

Yehao Li et al. designed a novel Transformer-style module, the Contextual Transformer (CoT) block [41], for visual recognition. The design makes full use of the contextual information between input keys to guide the learning of the dynamic attention matrix, thus enhancing the visual representation. As our model backbone network also uses Transformer-related techniques, an experimental comparison with other Transformer backbone networks is warranted. The results show a small improvement in CoT compared to the original mAP(@0.5), but a decrease in both mAP(@0.5:0.95) and Recall compared to YoloV5s.

In summary, our model has a clear advantage in detection accuracy over other Yolov5s that replace the backbone network, and achieves the best scores in all three objective metrics of mAP(@0.5), mAP(@0.5:0.95), and Recall, making it quite competitive.

3.3.4. The comparison results with other object detection algorithms

To demonstrate the competitiveness of our model, it is necessary to compare the optimized model with other classical target detection algorithms, and the experimental results are presented in Table 5.

Faster-RCNN is an improved model based on R-CNN by Girshick et al. [42], which further improves the region based CNN baseline, increases the training and testing speed, and also improves the accuracy of target detection and solves the problem of multi-scale and small target detection [43]. SSD is a single-stage feedforward neural network based on Liu et al. [44], which proposes a target detection algorithm that eliminates the suggestion of bounding boxes and resamples subsequent pixels or features and differentiates the prediction by aspect ratio. Yolov5l is a larger but more accurate model compared to Yolov5s. Yolov7 proposes a new positive and negative sample allocation strategy based on Yolov5. Compared with existing target detection models for general-purpose GPUs and mobile GPUs, Yolov7 outperforms other target detection models in terms of both FPS and AP.

By generating multiple resamples of the original data and calculating the metric of interest for each resample. We re-validated each model separately using 5-fold cross validation, and the statistics of the results are presented in Table 5.

As can be seen from Table 5, our model mAP(@0.5) improves by 8.42%, 8.56%, and 1.78% compared to Faster-Rcnn, SSD, and Yolov5l, respectively, while the number of parameters and model size are much smaller than these three algorithms. We also compared it with Yolov7 and the newly launched Yolov8s. We have improved mAP by 0.02% compared to Yolov7, and at the same time, the amount of parameters, model size and FPS are all better than Yolov7. Compared with Yolov8, the mAP of our model has increased by 2.36%, and the FPS is also higher. further demonstrating the superiority of our model.

Table 5
Result of a 5-fold cross-validated object detection model.

Model	mAP (@0.5)	Params	FPS	Weight
Faster-RCNN	72.87%(± 0.15%)	136,750,479	0.1	522 Mb
SSD	72.73%(± 0.23%)	24,146,894	4.2	92.1 Mb
Yolov5l	79.51%(± 0.13%)	46,154,449	13.0	88.7 Mb
Yolov7	81.27%(± 0.11%)	37,622,682	22.4	71.4 Mb
Yolov8s	78.93%(± 0.08%)	11,137,148	21.8	21.4 Mb
Our model	**81.29%(± 0.05%)**	**12,511,853**	**23.6**	**24.0 Mb**

4. Discussion

In this study, we propose a Yolov5-based cellular assay technique to detect macrophages, lymphocytes, neutrophils and eosinophils in BALF and thus provide accurate counts of these four cell types. We subjected the optimized model to ablation experiments, where six improvements were added to improve accuracy and effectiveness. Comparisons with other classical backbone network models from Yolov5s show that our proposed model performs best in terms of accuracy. Comparing FPS and accuracy with other classical object detection algorithms, the proposed model outperforms others in terms of accuracy and efficiency. At the same time, the results show that the method has a recall rate of more than 0.8 for three of the four BALF cell types, and the accuracy of the model detection meets the practical needs. After the model training, we developed a WEB-side application, so that people all over the world can visit the website and quickly detect BALF cells.

BALF cytomorphological tests have important clinical implications in the diagnosis of lung inflammation, tuberculosis, tumors and parasitic infections [45]. Studies have shown that when clinical history, physical examination, routine laboratory tests, pulmonary function tests and radiography are not sufficient to reach a definitive diagnosis, cytological examination of BALF usually provides valuable diagnostic information [46]. Immunophenotyping of lymphocytes in BALF is particularly important in the differential diagnosis of interstitial lung disease. The standard method for lymphocyte phenotyping is the peroxidase-antiperoxidase technique, however, it is both time-consuming and experience-dependent [47]. Furthermore, with the large number of samples currently being processed in the clinic, BALF cell sorting and counting still needs to be performed manually by the testers, a tedious and time-consuming process, with variability between test results from different testers, even for trained experts, and different methods of preparing cytological tests for BALF significantly affect the results of cell quantification [48].

BALF cell counting and classification not only helps clinicians in diagnosis, but in recent years it has also shown great potential in the prognosis of diseases [49]. Recently, Bouros et al. [50] reported that higher levels of eosinophils in BALF were associated with increased mortality in patients with systemic sclerosis. Researchers have also found that high levels of BALF eosinophils are associated with poor prognosis in idiopathic pulmonary fibrosis (IPF) and that increased BALF neutrophils are strongly associated with early mortality in IPF [51]. Several studies have demonstrated the significant prognostic value of BALF lymphocytes for 1-year survival in patients with acute respiratory failure. Patients with BALF lymphocytes ≥ 20% had significantly higher 1-year survival compared to patients with BALF lymphocytes < 20% [52]. In studies related to acute exacerbations of chronic progressive interstitial pneumonia, mortality was significantly lower in patients with BALF lymphocytes ≥ 15% compared to those with BALF lymphocytes < 15% [53]. In addition, a BALF total leukocyte count ≥ 510/μL was considered an independent predictor of bacterial pneumonia [54]. increased BALF neutrophil percentage was considered an independent predictor of early mortality in patients with IPF [55].

Our method has numerous advantages in terms of clinical application. First, BALF cell morphology is a tedious and time-consuming task, and there is a shortage of skilled professionals involved in body fluid cytomorphology, and BALF cell classification and counting rely on the extensive clinical skills and experience of pathologists. Secondly, rapid and accurate BALF cell sorting and counting can free pathologists from the burden of image reading, allowing them to focus more on making diagnostic and therapeutic decisions for their patients. In addition, the diagnostic results for the same image may vary between pathologists, with consequent differences in diagnostic accuracy. The algorithm we proposed can be

used as a third-party cell counting tool to assist pathologists, which can help reduce the misdiagnosis rate of pathologists.

5. Conclusion

In this paper, an improved Yolov 5 s Based on transformer backbone network is proposed for the detection and classification of bronchoalveolar lavage fluid cells. The model has excellent performance and solves the challenges in the application of BALF cell detection algorithms. The main findings are as follows:

(1) The introduction of Swin Transformer V2 as the backbone network allows feature extraction of global information and thus achieves an improvement in cell detection accuracy. Although the computational cost is increased, the accuracy is significantly improved. The superiority of introducing SWinV2 was demonstrated by comparing it with other classical backbone networks of Yolov5s (Yolov5s_all_Ghost, Yolov5s_CoT, Yolov5s_ShuffleNetV2) in ablation experiments.

(2) By introducing the Ghost module at the Neck of Yolov5s, the model parameters are compressed to a large extent, facilitating the capture and extraction of detailed features and maintaining the detection accuracy and speed.

(3) The positive and negative sample allocation strategy of Yolov7 is introduced to increase the number of positive samples and optimize the quality of anchor, which improves significantly. Using EIoU_Loss as the bounding box regression loss makes the bounding box converge faster and with higher accuracy. Use Weighted-NMS to weight and average the prediction frame information to make full use of the information of each bounding box and improve the accuracy of localization and classification.

(4) Compared with related target detection methods, in terms of accuracy our model is much higher than Faster-Rcnn and SSD, higher than Yolov5l, and slightly higher than with Yolov7, the proposed method has better prediction performance.

(5) The application of our cloud to the web side facilitates online BALF cell detection for physicians around the world.

Although our proposed method is effective in classifying and counting the four types of cells in BALF, it has some limitations. Firstly, the data used in this study were obtained from the Bronchoalveolar Lavage Cell Sorting and Counting Challenge on the Xunfei platform, which is a single source of data, and more high-quality image data from multiple centers should be collected to further evaluate and validate our method. Secondly, in our training tests for BALF cell classification, we found that Eosinophil cells and Macrophage cells, due to the similarity of their characteristics and differences in the way they are handled or observed in the sections, change their color rendering in different section views, leading to confusion between the two, which will be explored in depth in our future studies, trying to reduce the parameters, add this will be explored further in our future studies, trying to reduce the parameters, add new attention mechanisms or replace the backbone network to reduce the occurrence of cell confusion.

Funding

This work was funded by the National Natural Science Foundation of China, grant number No.32160169; the National Natural Science Foundation of Jiangxi Province, grant number No. 20181BAB205041; the National College Student Innovation and Entrepreneurship Training.

CRediT authorship contribution statement

Puzhen Wu performed data augmentation on the original data and improved the model. Han Weng jointly performed ablation experiments and compared with other algorithms. Yi Zhan discussed the analysis of the results. Wenting Luo contributed significantly to the model structure image and visualization of the data. Puzhen Wu, Wenting Luo and Yi Zhan co-authored the manuscript. Lixia Xiong and Hongyan Zhang critically revised the manuscript and put forward valuable suggestions for the manuscript. Hai Yan is involved in the study design and the critical revision of the manuscript. All authors reviewed and approved the manuscript.

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.

References

[1] Couetil LL, Thompson CA. Airway diagnostics: bronchoalveolar lavage, tracheal wash, and pleural fluid. The Veterinary clinics of North America. Equine Pract 2020;36(1):87–103. https://doi.org/10.1016/j.cveq.2019.12.006
[2] Meyer KC, Raghu G. Bronchoalveolar lavage for the evaluation of interstitial lung disease: Is it clinically useful? Eur Respir J 2011;38(4):761–9. https://doi.org/10.1183/09031936.00069509
[3] Welker L, Jorres RA, Costabel U, Magnussen H. Predictive value of BAL cell differentials in the diagnosis of interstitial lung diseases. Eur Respir J 2004;24(6):1000–6. https://doi.org/10.1183/09031936.04.00101303
[4] Midulla F, Nenna R. Bronchoalveolar lavage: indications and applications. Paediatric Bronchoscopy. Karger Publishers; 2010. p. 30–41.
[5] Liu Z, Yan J, Tong L, Liu S, Zhang Y. The role of exosomes from BALF in lung disease. J Cell Physiol 2022;237(1):161–8. https://doi.org/10.1002/jcp.30553
[6] van Hoecke L, Job ER, Saelens X, Roose K. Bronchoalveolar lavage of murine lungs to analyze inflammatory cell infiltration. J Vis Exp 2017;2017(123). https://doi.org/10.3791/55398
[7] Drent M, van Velzen-Blad H, Diamant M, Wagenaar S, Hoogsteden H, van den Bosch J. Bronchoalveolar lavage in extrinsic allergic alveolitis: effect of time elapsed since antigen exposure. Eur Respir J 1993;6(9):1276–81.
[8] Silver RM, Clements PJ. Interstitial lung disease in systemic sclerosis: optimizing evaluation and management. Scleroderma Care Res 2003;1:3–11.
[9] Costabel U, Guzman J, Bonella F, Oshimo S. Bronchoalveolar lavage in other interstitial lung diseases (October). Seminars in Respiratory and Critical Care Medicine Vol. 28. © Thieme Medical Publishers; 2007. p. 514–24. (October).
[10] Kyo M, Hosokawa K, Ohshimo S, Kida Y, Tanabe Y, Shime N. Prognosis of pathogen-proven acute respiratory distress syndrome diagnosed from a protocol that includes bronchoalveolar lavage: A retrospective observational study. 54-54 J Intensive Care 2020;8(1). https://doi.org/10.1186/s40560-020-00469-w
[11] Zhou Daoyin, Wu Mao, Xu Shaoqiang, Zhang Shimin, Fan Ailin, Huang Daolian, et al. Chinese expert consensus on cytomorphological testing of bronchoalveolar lavage fluid (2020). J Mod Lab Med 2020.
[12] Hodge SJ, Hodge GL, Holmes M, Reynolds PN. Flow cytometric characterization of cell populations in bronchoalveolar lavage and bronchial brushings from patients with chronic obstructive pulmonary disease. Cytom Part B, Clin Cytom 2004;61B(1):27–34. https://doi.org/10.1002/cyto.b.20020
[13] Smith KP, Kang AD, Kirby JE. Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. J Clin Microbiol 2018;56(3):e01521. https://doi.org/10.1128/JCM.01521-17
[14] Zhang Y, Chen Y, Chen Z, Zhou Y, Sheng Y, Xu D, et al. Effects of bronchoalveolar lavage on refractory Mycoplasma pneumoniae pneumonia. Respir Care 2014;59(9):1433–9. https://doi.org/10.4187/respcare.03032
[15] Yu Y, Liu C, Zhang Z, Shen H, Li Y, Lu L, et al. Bronchoalveolar lavage fluid dilution in ICU patients: what we should know and what we should do. Crit Care 2019;23(1):23. https://doi.org/10.1186/s13054-018-2300-x
[16] Trisolini R, Cancellieri A, Tinelli C, Paioli D, Scudeller L, Casadei GP, et al. Rapid on-site evaluation of transbronchial aspirates in the diagnosis of hilar and mediastinal adenopathy: a randomized trial. Chest 2011;139(2):395–401. https://doi.org/10.1378/chest.10-1521
[17] Banik PP, Saha R, Kim K. An automatic nucleus segmentation and CNN model based classification method of white blood cell. Expert Syst Appl 2020;149:113211. https://doi.org/10.1016/j.eswa.2020.113211
[18] Tayebi RM, Mu Y, Dehkharghanian T, Ross C, Sur M, Foley R, et al. Automated bone marrow cytology using deep learning to generate a histogram of cell types. 45-45 Commun Med 2022;2(1). https://doi.org/10.1038/s43856-022-00107-6
[19] Delgado-Ortet M, Molina A, Alférez S, Rodellar J, Merino A. A deep learning approach for segmentation of red blood cell images and malaria detection. Entropy 2020;22(6):1–16. https://doi.org/10.3390/e22060657
[20] Bibi N, Sikandar M, Ud Din I, Almogren A, Ali S. IoMT-based automated detection and classification of leukemia using deep learning. 6648574-12 J Healthc Eng 2020;2020. https://doi.org/10.1155/2020/6648574
[21] Tao Y, Cai Y, Fu H, Song L, Xie L, Wang K. Automated interpretation and analysis of bronchoalveolar lavage fluid. 104638-104638 Int J Med Inform 2022;157. https://doi.org/10.1016/j.ijmedinf.2021.104638
[22] Peng Tao, Zhao Jing, Gu Yidong, Wang Caishan, Wu Yiyun, Cheng Xiuxiu, et al. H-ProMed: ultrasound image segmentation based on the evolutionary neural

network and an improved principal curve. ISSN 0031-3203 Pattern Recognit 2022;Volume 131:108890. https://doi.org/10.1016/j.patcog.2022.108890
[23] Lam S, Zhang Y, Zhang J, Li B, Sun J, Liu CY, et al. Multi-organ omics-based prediction for adaptive radiation therapy eligibility in nasopharyngeal carcinoma patients undergoing concurrent chemoradiotherapy. Front Oncol 2022;11. https://doi.org/10.3389/fonc.2021.792024
[24] Peng Tao, Gu Yidong, Ye Zhenyu, Cheng Xiuxiu, Wang Jing. A-LugSeg: automatic and explainability-guided multi-site lung detection in chest X-ray images. ISSN 0957-4174 Expert Syst Appl 2022;Volume 198:116873. https://doi.org/10.1016/j.eswa.2022.116873
[25] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., & IEEE. (2016). You only look once: Unified, real-time object detection. Paper presented at the, 2016- 779–788. https://doi.org/10.1109/CVPR.2016.91.
[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.,. & et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[27] ⟨https://challenge.xfyun.cn/topic/info?type=bronchoalveolar&option=phb⟩.
[28] CVAT.ai Corporation. (2023). Computer Vision Annotation Tool (CVAT) (v2.4.3). Zenodo. https://doi.org/10.5281/zenodo.7863887.
[29] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.,. & et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., , & et al. (2017). Attention is all you need.
[31] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proc IEEE/CVF Int Conf Comput Vis 2021:10012–22.
[32] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., & et al. (2022). Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12009–12019).
[33] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1580–1589).
[34] Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans Cybern 2021.
[35] Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing 2022;506:146–57.
[36] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding Yolo series in 2021. arXiv preprint arXiv:2107.08430.
[37] Wang, C.Y., Bochkovskiy, A., & Liao, H.Y.M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.
[38] Neubeck, A., & Van Gool, L. (2006, August). Efficient non-maximum suppression. In 18th International Conference on Pattern Recognition (ICPR'06) (Vol. 3, pp. 850–855). IEEE.
[39] Ning, C., Zhou, H., Song, Y., & Tang, J. (2017). Inception single shot MultiBox detector for object detection. Paper presented at the 549–554. https://doi.org/10.1109/ICMEW.2017.8026312.
[40] Ma, N., Zhang, X., Zheng, H., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss (Eds.), Computer vision - eccv 2018, pt xiv (pp. 122–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-01264-9_8.
[41] Li Y, Yao T, Pan Y, Mei T. Contextual transformer networks for visual recognition. 1-1 IEEE Trans Pattern Anal Mach Intell2022. https://doi.org/10.1109/TPAMI.2022.3164083
[42] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031
[43] Eggert, C., Brehm, S., Winschel, A., Zecha, D., Lienhart, R., & IEEE. (2017). A closer look: Small object detection in faster R-CNN. Paper presented at the 421–426. https://doi.org/10.1109/ICME.2017.8019550.
[44] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., & et al. (2016;2015;). SSD: Single shot MultiBox detector. In B. Leibe, J. Matas, N. Sebe & M. Welling (Eds.), Computer vision - eccv 2016, pt i (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978–3-319–46448-0_2.
[45] Fang X, Mei Q, Fan X, Zhu C, Yang T, Zhang L, et al. Diagnostic value of metagenomic next-generation sequencing for the detection of pathogens in bronchoalveolar lavage fluid in ventilator-associated pneumonia patients. 599756-599756 Front Microbiol 2020;11. https://doi.org/10.3389/fmicb.2020.599756
[46] Li L, Ye Z, Yang S, Yang H, Jin J, Zhu Y, et al. Diagnosis of pulmonary nodules by DNA methylation analysis in bronchoalveolar lavage fluids. 185-185 Clin Epigenet 2021;13(1). https://doi.org/10.1186/s13148-021-01163-w
[47] Midulla F, Nenna R. Bronchoalveolar lavage: indications and applications. Paediatric Bronchoscopy Vol. 38. Karger Publishers; 2010. p. 30–41.
[48] Ma W, Cui W, Lin Q. Improved immnunophenotyping of lymphocytes in bronchoalveolar lavage fluid (BALF) by flow cytometry. Clin Chim Acta 2001;313(1):133–8. https://doi.org/10.1016/S0009-8981(01)00664-7
[49] Kalidhindi RSR, Ambhore NS, Bhallamudi S, Loganathan J, Sathish V. Role of estrogen receptors alpha and beta in a murine model of asthma: Exacerbated airway hyperresponsiveness and remodeling in ER beta knockout mice. 2019; Front Pharmacol 2020;10. https://doi.org/10.3389/fphar.2019.01499
[50] Davidson KR, Ha DM, Schwarz MI, Chan ED. Bronchoalveolar lavage as a diagnostic procedure: A review of known cellular and molecular findings in various lung diseases. J Thorac Dis 2020;12(9):4991–5019. https://doi.org/10.21037/jtd-20-651
[51] Bouros D, Wells AU, Nicholson AG, Colby TV, Polychronopoulos V, Pantelidis P, et al. Histopathologic subsets of fibrosing alveolitis in patients with systemic sclerosis and their relationship to outcome. Am J Respir Crit Care Med 2002;165(12):1581–6. https://doi.org/10.1164/rccm.2106012
[52] Kinder BW, Brown KK, Schwarz MI, Ix JH, Kervitsky A, King Jr TE. Baseline BAL neutrophilia predicts early mortality in idiopathic pulmonary fibrosis. Chest 2008;133(1):226–32.
[53] Hirasawa Y, Nakada T, Shimazui T, Abe M, Isaka Y, Sakayori M, et al. Prognostic value of lymphocyte counts in bronchoalveolar lavage fluid in patients with acute respiratory failure: a retrospective cohort study. 21-21 J Intensive Care 2021;9(1). https://doi.org/10.1186/s40560-021-00536-w
[54] Takei R, Arita M, Kumagai S, Ito Y, Noyama M, Tokioka F, et al. Impact of lymphocyte differential count > 15% in BALF on the mortality of patients with acute exacerbation of chronic fibrosing idiopathic interstitial pneumonia. 67-67 BMC Pulm Med 2017;17(1). https://doi.org/10.1186/s12890-017-0412-8
[55] Choi S, Hong S, Hong H, Kim S, Huh JW, Sung H, et al. Usefulness of cellular analysis of bronchoalveolar lavage fluid for predicting the etiology of pneumonia in critically ill patients. e97346-e97346 PLoS One 2014;9(5). https://doi.org/10.1371/journal.pone.0097346