Attention-Based Fine-Grained Classification of Bone Marrow Cells

Weining Wang[1], Peirong Guo[1], Lemin Li[1], Yan Tan[2], Hongxia Shi[2], Yan Wei[2], and Xiangmin Xu[1,3](✉)

[1] South China University of Technology, Guangzhou, China
xmxu@scut.edu.cn
[2] Peking University People's Hospital, Beijing, China
[3] Institute of Modern Industrial Technology of SCUT in Zhongshan, Zhongshan, China

Abstract. Computer aided fine-grained classification of bone marrow cells is a significant task because manual morphological examination is time-consuming and highly dependent on the expert knowledge. Limited methods are proposed for the fine-grained classification of bone marrow cells. This can be partially attributed to challenges of insufficient data, high intra-class and low inter-class variances.
In this work, we design a novel framework Attention-based Suppression and Attention-based Enhancement Net (ASAE-Net) to better distinguish different classes. Concretely, inspired by recent advances of weakly supervised learning, we develop an Attention-based Suppression and Attention-based Enhancement (ASAE) layer to capture subtle differences between cells. In ASAE layer, two parallel modules with no training parameters improve the discrimination in two different ways. Furthermore, we propose a Gradient-boosting Maximum-Minimum Cross Entropy (GMMCE) loss to reduce the confusion between subclasses. In order to decrease the intra-class variance, we adjust the hue in a simple way. In addition, we adopt a balanced sampler aiming to alleviate the issue of the data imbalance.
Extensive experiments prove the effectiveness of our method. Our approach achieves favorable performance against other methods on our dataset.

Keywords: Fine-grained classification · Bone marrow cell · Attention · Weakly supervised learning

1 Introduction

Morphological examination of bone marrow cells is the most commonly used diagnostic method for acute leukemia. The manual classification method, which is time-consuming and repetitive, highly depends on the knowledge of experts and is prone to variances among observers. Automatical classification based on computers is conducive to diagnosing acute leukemia efficiently. The computer-aided classification can quickly obtain objective results [1, 2], effectively process

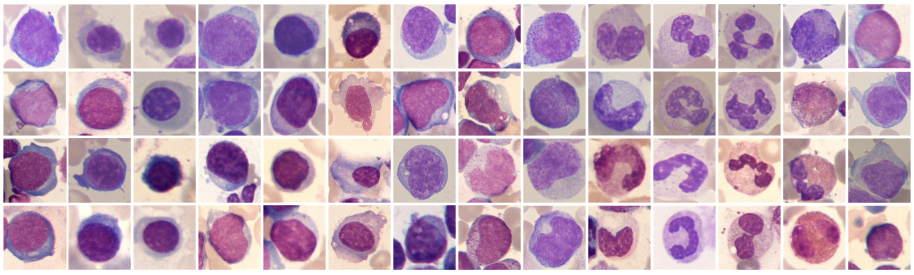

Fig. 1. Some instances in our bone marrow cell dateset. Bone marrow cells are classified into 14 categories for fine-grained classification in our dataset and one column represents one category. Four cell images in each category are randomly selected and shown in a single column.

a large amount of data [3] and greatly reduce the workload of doctors. It has attracted the attention of many scholars. We conduct a survey on the bone marrow cell classification in bone marrow smears and a similar task done on blood cells in peripheral blood smears.

In the past 40 years, most of the methods for these two similar tasks are based on traditional image processing methods like [4–10], consisting of three steps including image segmentation, feature extraction and image classification. With the development of deep learning, many researchers [11–19] consider to adopt it to these tasks.

To be specific, bone marrow cell classification can be divided into two categories including coarse-grained and fine-grained classification. Coarse-grained classification only considers meta-categories like lymphocytes and monocytes, which does not meet the requirements in the clinical practice. Fine-grained classification aims to distinguish similar sub-categories in different maturity stages.

Challenges come from three aspects: (1) High intra-class variances. Cells belonging to the same class may be variant in the saturation, hue and brightness due to the light, dyeing duration and dose. In addition, crowding degree and background complexity vary especially in bone marrow smears. (2) Low inter-class variances. Bone marrow cells in adjacent maturity stages can be easily confused and hard to be distinguished even by experts. (3) Limited and imbalanced datasets. It takes a large amount of time for experts to label them with fine-grained annotations. Currently, there are few studies on the fine-grained classification of bone marrow cells to tackle the aforementioned challenges. As shown in Fig. 1, we find the above challenges in our dataset.

Fine-grained classification has made great progress in some other areas, such as birds [20], flowers [21] and cars [22]. Common methods imitate the process of experts to locate discriminative regions and extract the features. The existing methods can be divided into two categories according to whether leveraging extra annotations or not. Methods relying on extra annotations like [20] meet the demands on the knowledge of experts so they do not suit for medical images analysis because of its high cost. Some other methods like [23–33] with only

image-level annotations locate key parts of an object based on the weakly supervised learning, which improves the availability and scalability. Motivated by it, we aim to design a flexible and efficient architecture for the task of classification on bone marrow cells with the help of weakly supervised learning.

Based on our observation, the existing weakly supervised methods have one or more limitations as follows: (1) Some models like [26] can locate object parts or mine discriminative regions with a limited quantity, which cannot be easily modified. (2) Some models can cover only a part of discriminative region because the loss focuses mainly on the most discriminative region and ignores the others. (3) Currently, some methods like Attention-based Dropout Layer [23] and Diversification Block [33] force the model to learn from other regions by discarding some discriminative regions while they have some disadvantages. ADL may discard excessive information. (4) Some models like [25, 29] have disadvantages of repeated training, extra classification branches and multiple forward computations. Our model overcomes the disadvantages mentioned above.

The contributions of this paper are as follows:

1. We propose a two-branch framework called Attention-based Suppression and Attention-based Enhancement Net (ASAE-Net) well designed for the fine-grained classification of bone marrow cells. ASAE-Net achieves the superiority without increasing training parameters compared with the backbone.
2. Our model can flexibly locate multiple discriminative regions with an unlimited quantity. Our Attention-based Suppression (AS) branch adopts a suppression approach with two restrictions leading to a good performance.
3. Our Gradient-boosting Maximum-Minimum Cross Entropy (GMMCE) loss alleviates the problem of the category confusion and improves the classification performance.
4. We adopt a balanced sampler and the hue adjustment aiming to alleviate the problem of data imbalance and the high intra-class variance.
5. Our method outperforms the existing fine-grained classification methods for bone marrow cells and some other methods based on attention mechanism.

The rest of the paper is organized as follows. We first review the related work of bone marrow cell classification and Gradient-boosting Cross Entropy loss in Section 2. Then we illustrate our proposed ASAE-Net, GMMCE loss and improved strategies on dataset in Section 3. In Section 4, adequate experiments and results are presented and analyzed. We conclude our work in Section 5.

2 Related Work

2.1 Bone Marrow Cell Classification

As we know, convolutional neural network (CNN) is a strong tool for image processing. In the last few years, CNNs are widely adopted in many medical applications [18,34–36]. In the coarse-grained classification on bone marrow cells and blood cells, many methods have gained great performance. However, there are few studies for the fine-grained classification on them.

In [11], Chandradevan et al. propose a two-stage method to detect and classify bone marrow cells in the fine-grained field. In the classification phase, they adopt ensembled models to improve the performance. Proposed by Qin et al., Cell3Net [17] is a residual model to classify cells into 40 categories. The model consists of seven convolutional layers, three pooling layers, three residual layers, two fully-connected layers and an output layer, gaining an accuracy of 76.84%. Similarly, Jiang et al. [14] design a 33-layers network called WBCNet referring to AlexNet [37], and improve its accuracy by modifying the residual layers and the activation function. Based on our observation, the model ensemble is effective but complicated. Cell3Net and WBCNet are simple but lack of pertinence for the fine-grained classification and the distribution of categories.

Compared with them, our model is well-designed, showing its effectiveness on the fine-grained classification task of bone marrow cells.

2.2 Gradient-Boosting Cross Entropy Loss

Cross entropy loss is one of the most widely used loss functions in the classification task. $\mathbf{s}$ and l are respectively the confidence scores and the true label. J is the set of all categories. The loss can be defined as:

$$CE(\mathbf{s}, l) = -log\frac{exp(s_l)}{\sum_{i \in J} exp(s_i)} \tag{1}$$

According to the definition, CE loss treats all negative categories equally. However, in our task, some negative classes are confusable and obtain relatively high confidence scores from the model. To ease this problem, Gradient-boosting Cross Entropy (GCE) loss presented in [33] focuses on the true category and some of the negative categories according to different confidence scores. Concretely, it only considers negative classes with top-k highest scores. $J^{'}_{>}$ is the set of negative classes with top-k highest scores. GCE can be defined as:

$$GCE(\mathbf{s}, l) = -log\frac{exp(s_l)}{exp(s_l) + \sum_{i \in J^{'}_{>}} exp(s_i)} \tag{2}$$

Our Gradient-boosting Maximum-Minimum Cross Entropy (GMMCE) loss enhances GCE loss by considering negative classes with bottom-k lowest scores.

3 Proposed Method

In this section, we present our attention-based framework ASAE-Net for the fine-grained classification of bone marrow cells, as shown in Fig. 2. This is a two-branch framework containing an attention-based suppression branch and an attention-based enhancement branch. Furthermore, we make a comparison between the proposed ASAE layer and other similar works. Then we introduce our proposed GMMCE loss, a preprocessing strategy and a sampling strategy for specific problems.

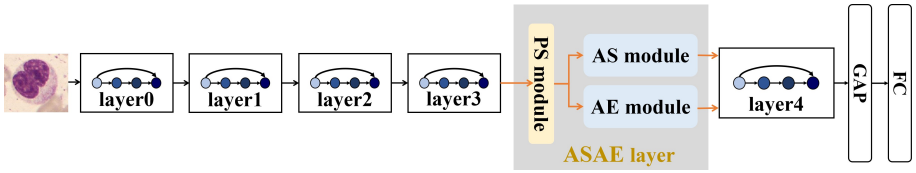

Fig. 2. The network structure of ASAE-Net. It contains the backbone (divided into five layers) and an ASAE layer. The ASAE layer consists of PS module, AS module and AE module.

3.1 Framework

ASAE-Net is designed aiming to capture more discriminative regions to improve the ability of distinguishing cells with subtle differences. Our ASAE-Net is formed by a common backbone and a novel ASAE layer, which is plug-and-play to capture more discriminative regions precisely without training parameters.

To build our ASAE-Net, we adopt ResNet-50 [38] as our backbone and divide it into five layers (layer0, layer1, layer2, layer3, layer4) based on the output feature size. As we know, shallow layers prefer to learn fundamental features like the texture and the shape. As pointed out by Choe et al. in [23], masks in the deep layers discard discriminative regions more precisely. Hence, we decide to insert our ASAE layer, which includes Attention-based Suppression branch and Attention-based Enhancement branch, in a deep layer (layer3 or layer4). Subsequently, we conduct experiments on the effect of the inserted position of our Attention-based Suppression branch.

We divide our framework into two parts by the inserted position. In the training phase, the front part takes an image as input, and feeds its own output, $F_{mid} \in R^{C \times H \times W}$, to two parameter-shared branches. Then, an attention-based suppression mask M_{sup} and an attention-based enhancement map M_{enh} are generated respectively. Next, they are applied to F_{mid} by spatial-wise multiplication. Two new features are sent to the latter part to gain two probability distributions. We average them to gain the final probability. While in the testing phase, the input image skips our ASAE layer.

Our framework has two characteristics: (1) It is an end-to-end model without extra training parameters than the backbone. (2) It consists of two branches sharing parameters and promoting each other.

3.2 ASAE Layer

In this part, we introduce our proposed ASAE layer shown in Fig. 3. It contains three key modules including: (1) Peak Stimulation (PS) module. (2) Attention-based Suppression (AS) module. (3) Attention-based Enhancement (AE) module.

Peak Stimulation Module. It is proved in [39] that the local peak value of the class response map is corresponding to strong visual cues. Actually, when it

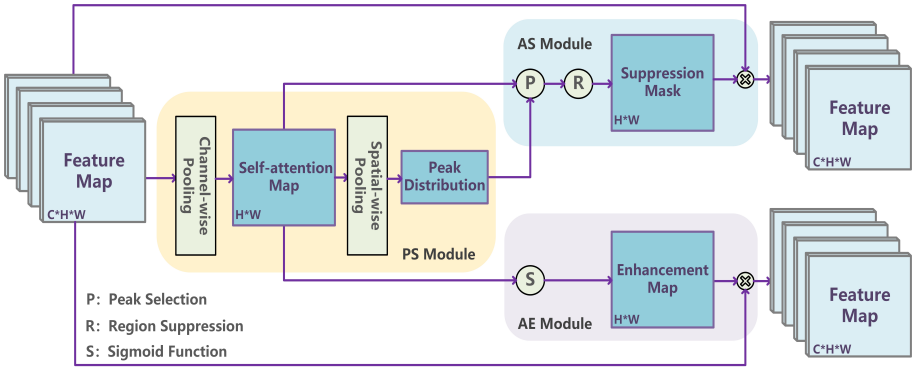

Fig. 3. Attention-based Suppression and Attention-based Enhancement (ASAE) layer. The self-attention map and peak distribution are generated by Peak Stimulation (PS) Module with two pooling layers. Attention-based Suppression (AS) Module generates a suppression mask by adopting two strategies. Attention-based Enhancement (AE) Module transfers the self-attention map into an enhancement map by using Sigmoid Function. Two new feature maps are generated by using spatial-wise multiplication.

comes to the feature maps in the hidden layer, we have the similar result. Hence, to obtain strong visual cues, we design a peak stimulation module to generate the peak distribution of feature maps.

Given the feature map $F \in R^{C\times H\times W}$, where C denotes the channel number and $H \times W$ denotes the shape of maps, the peak is defined as the maximum in a local region of $r \times r$. The positions of peaks are represented as $P = \{(i_1, j_1), (i_2, j_2), \cdots, (i_N, j_N)\}$, where N is the total number of local peaks, which may be different in different feature maps.

Our peak stimulation module consists of two pooling layers as shown in Fig. 3. The first one is an average pooling layer to squeeze the channel information to form a self-attention map $M_{att} \in R^{H\times W}$. The higher value in the map represents the more discriminative ability. The second one is a max pooling layer of size r to obtain the peak distribution including values and positions in the self-attention map. In this way, we can effectively approximate the spatial distribution of components.

The peak stimulation module can effectively capture multiple local peaks with an unlimited quantity but without extra training parameters. In this module, we generate a self-attention map for the following two modules and the peak distribution for the following AS module.

Attention-Based Suppression Module. Based on the self-attention map and the peak distribution from the peak stimulation module, our proposed AS module hides one discriminative region in the training phase aiming to force the model to locate and learn from other discriminative regions.

As illustrated above, local peak values are corresponding to strong visual cues within images. Hence, we define a peak and its local region as a discrimi-

native region. Higher values represent higher discrimination. It has advantages as follows: (1) No manual part annotations required. (2) No extra training parameters. (3) Flexibility on the number of discriminative regions. (4) Flexibility to be applied in different layers.

A keypoint of AS module is how to suppress the discriminative regions. To design a better method, we firstly explore some similar approaches. Dropout [40], Spatial Dropout [41], MaxDrop [42], Dropblock [43] are some popular methods of suppression. We find these methods unsuitable for our task. When discarding the max values, it encounters the high-correlation problem and when randomly discarding regions, it may not discard a discriminative one. Also, without restrictions on areas and values, the model may discard more than one discriminative region, which is bad for the latter learning.

To avoid the above disadvantages, we restrict the suppression size and the value threshold when discarding a region. Also, we ensure our model to discard only one discriminative region effectively. We present our strategies in the following.

The first one is *the peak selection strategy*. After the peak stimulation module, we obtain multiple local peaks in each image. Our strategy is to choose one from the top-k peaks to be the center of the discriminative region to be suppressed. Concretely, our strategy is as follows:

(1) Set the number of candidates k and arrange peaks in a descending order.

(2) Select a local peak randomly. The peak with the highest value is selected with probability p_{top1}, which means the most discriminative one. With probability $1 - p_{top1}$, a local peak is randomly selected from top-2 to top-k ($k \geq 2$).

In the fine-grained classification, different classes of cells often have subtle differences. With our strategy of peak selection, we choose randomly a discriminative region to suppress. In this way, we decrease the risk of mis-classification when cells suffer from lack of salient characteristics like cytoplasmic particles.

The second one is *the region suppression strategy* with two restrictions. We define the suppressed region as the one centered on the selected local peak. We leverage an area threshold and a value threshold to restrict the local region. The concrete operations are as follows:

(1) Set a suppression ratio γ and an area threshold β.

(2) Define the value threshold α as the product of peak value $Peak$ and γ:

$$\alpha = Peak * \gamma \tag{3}$$

(3) Define the original suppression region R_{sup} in M_{att} to be a square with a local peak at the center and the length of side $\sqrt{\beta}$.

(4) Define a suppression mask M_{sup} with the same shape of M_{att}.

(5) Set the value v_{ij}^{sup} in M_{sup} to be 0, if v_{ij}^{att} in M_{att} is higher than α and the pixel (i, j) is in R_{sup} at the same time. Otherwise set it to be 1.

$$v_{ij}^{sup} = \begin{cases} 0, & (i,j)\ in\ R_{sup} \cap v_{ij}^{att} \geq \alpha \\ 1, & otherwise \end{cases} \tag{4}$$

After obtaining M_{sup}, we apply it to the input feature map by spatial-wise multiplication to obtain F_{sup} with a discriminative region suppressed. The same region is suppressed in different channels.

The superiors of our AS module lie in: (1) It hides only one discriminative region in a single iteration. (2) It randomly discards different discriminative regions to encourage the model to learn from other regions. (3) It adopts useful strategies to avoid the model from discarding excessive information by two restrictions. (4) It has no training parameters.

Attention-Based Enhancement Module. The AE module aims to strengthen the most discriminative region to improve the performance. In this module, we adopt the same method as ADL [23] to generate an enhancement map $M_{enh} \in R^{H \times W}$. It is yielded by using sigmoid function on the self-attention map. In M_{enh}, the pixel is discriminative if the value is close to 1. Similar to the AS module, M_{enh} is applied to the feature map by spatial-wise multiplication to obtain F_{enh}. Similarly, the AE module has no training parameters.

3.3 Comparison with Similar Methods

Comparison with Attention-Based Dropout Layer (ADL) [23]. From the motivation, we both aim to hide the discriminative regions and encourage the model to learn from others.

From the architecture, we both generate a suppression mask and an enhancement map but we differ from each other when generating the suppression mask. ADL discards all pixels whose values are higher than the value threshold. Without restricting the area or the continuity, it may easily discard the most discriminative and several sub-discriminative regions at the same time. Thus the feature maps lose excessive information. However, our ASAE layer adopts the suppression ratio and the area threshold to restrict the discriminative region with a local peak in the center, in which way it avoids discarding too many continuous discriminative regions. At the same time, our AS module can sometimes discard a sub-discriminative region instead of the most discriminative one, which alleviates the problem of subtle differences.

From the combination of two branches, ADL randomly selects one branch in one time while ASAE-Net uses two branches at the same time by averaging the two probability distributions, which is better for the information learning.

Comparison with Diversification Block (DB) [33]. From the motivation, we both aim to erase discriminative regions to enhance the ability of feature learning.

From the input size, the channel of the input is limited to be the class number in DB while our ASAE layer does not restrict the size of the input or the channel number, which is more flexible to be inserted into the backbone.

From the suppression mask, DB discards different regions in different channels while our layer discards the same regions in all input channels, which is more close to the human behavior and increases the robustness.

3.4 GMMCE Loss

In multiple-instance learning, it is proved that the regions with max and min scores are both important [44] though they can bring in different kinds of information. Motivated by it and GCE loss introduced in Section 2.2, we design an enhanced loss called GMMCE loss to focus on both high and low confidence scores.

Following the setting and denotations of GCE loss, J is the set of all categories and $J^{'}$ is the set of negative categories. s_i is the confidence score of category i. $J^{'}_{>}$ and $J^{'}_{<}$ can be defined as:

$$J^{'}_{>} = \left\{ i : i \in J^{'} \cap s_i \geqslant t_{k^+} \right\} \tag{5}$$

$$J^{'}_{<} = \left\{ i : i \in J^{'} \cap s_i \leqslant t_{k^-} \right\} \tag{6}$$

where t_{k^+} and t_{k^-} denote the k^+th highest and k^-th lowest confidence scores in $J^{'}$. We denote l as the true label. It is easy to find the relationship:

$$J^{'}_{>} + \{l\} \subset J^{'}_{>} + J^{'}_{<} + \{l\} \subset J^{'} + \{l\} = J \tag{7}$$

Our GMMCE loss and its gradient can be defined as:

$$GMMCE(\mathbf{s}, l) = -log\frac{exp(s_l)}{exp(s_l) + \sum_{i \in J^{'}_{>}} exp(s_i) + \sum_{i \in J^{'}_{<}} exp(s_i)} \tag{8}$$

$$\frac{\partial GMMCE(\mathbf{s}, l)}{\partial s_c} = \begin{cases} \frac{exp(s_l)}{exp(s_l) + \sum_{i \in J^{'}_{<}} exp(s_i) + \sum_{i \in J^{'}_{>}} exp(s_i)}, & c \in \left\{ J^{'}_{>} + J^{'}_{<} \right\} \\ \frac{exp(s_l)}{exp(s_l) + \sum_{i \in J^{'}_{<}} exp(s_i) + \sum_{i \in J^{'}_{>}} exp(s_i)} - 1, & c = l \end{cases} \tag{9}$$

Compared with CE and GCE [33], we can obtain the relationship:

$$\frac{\partial GCE(\mathbf{s}, l)}{\partial s_c} > \frac{\partial GMMCE(\mathbf{s}, l)}{\partial s_c} > \frac{\partial CE(\mathbf{s}, l)}{\partial s_c} \tag{10}$$

For the ground truth label and the confusing negative classes, the gradient of GMMCE falls in between GCE and CE, which implies a moderate update rate. GMMCE focuses on negative classes with high scores and low scores. The negative classes with low scores can be regarded as a regularization method like label smoothing, which can avoid the model from over-fitting.

3.5 Preprocessing and Sampling Strategies

As illustrated above, bone marrow cell datasets generally suffer from severe imbalance and high intra-class variances. In this part, we introduce our approaches to adjust the hue and sample data during training.

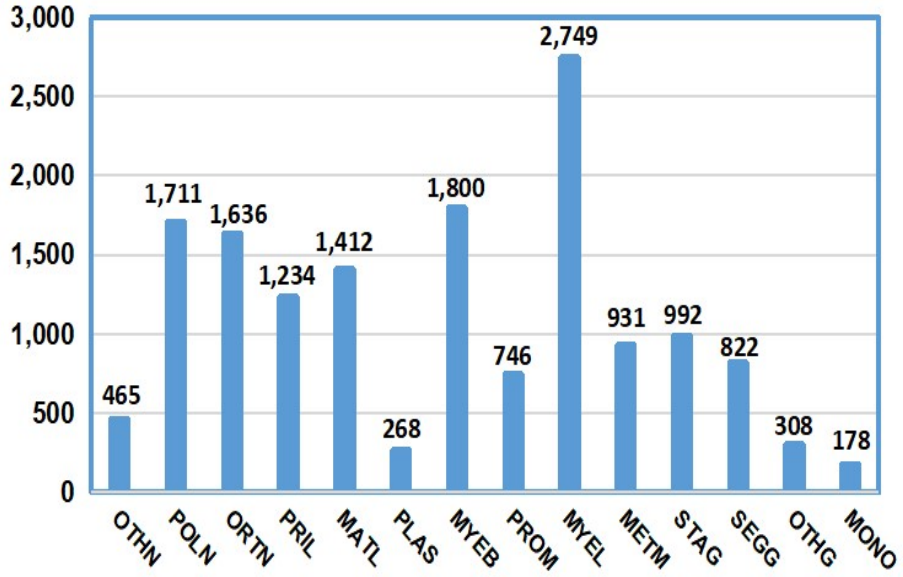

Fig. 4. The category distribution of our dataset.

Hue Adjustment. Generally, images of bone marrow cells have variant saturation, hue and brightness due to the dyeing duration, dyeing dose and light, even in the same class. The uniform hue improves the classification accuracy by decreasing the intra-class variance. Based on the consideration, we adopt a linear transformation in H channel of images to adjust the hue of images. Steps are as follows:

(1) Convert an RGB image into an HSV image.
(2) Conduct a linear transformation in the H channel of images as:

$$h_{out} = \frac{Max - Min}{max - min}(h_{in} - min) + Min \tag{11}$$

where h_{in} and h_{out} denote the values of H channel before and after the transformation. Max and Min denote the highest and lowest target values of the output while max and min denote the highest and lowest values of the input.

Balanced Sampler. Based on our observation, bone marrow cell datasets generally face the challenge of severe imbalance due to the distribution of cells in bone marrow smears. Classes with few training samples may have a low true positive rate without balancing strategies. Because of it, we adopt a balanced sampler to alleviate the imbalance without augmenting data offline.

Concretely, in a single iteration, we sample the same number of images in different classes from the train set. Hence, the balanced sampler can make the model learn from all classes in an equal probability.

4 Experiments

4.1 Dataset

Our dataset contains 15,252 images of 27 classes of bone marrow cells. Cells are cropped by 10 medical students and annotated by 3 medical experts. Some classes contain few images so we combine them into 14 classes according to the expert advice. The 14 classes are named respectively Other Normoblast (OTHN), Polychromatic Normoblast (POLN), Orthochromatic Normoblast (ORTN), Primitive Lymphocyte (PRIL), Mature Lymphocyte (MATL), Plasma Cell (PLAS), Myeloblast (MYEB), Promyelocyte (PROM), Myelocyte (MYEL), Metamyelocyte (METM), Neutrophilic Stab Granulocyte (STAG), Neutrophilic Segmented Granulocyte (SEGG), Other Granulocyte (OTHG), Monocyte (MONO). Some instances in our dataset are shown in Fig. 1. The 14 columns from the left to the right follow the order listed above. The distribution of classes is shown in Fig. 4.

4.2 Implementation Details

We adopt the stratified sampling to divide our dataset into the train set and the test set in a ratio of 8:2. Concretely, the train set includes 12,201 images and the test set includes 3,051 images. We conduct the same method of data augmentation as [38]. Especially, we random crop the input images into the size of 224×224. Besides, we add random vertical flip, rotation and color jitter to our train set, while an offline augmentation is excluded.

In this work, we choose ResNet-50 [38] as our backbone. We use SGD with a batch size of 84. The weight decay is set as 1e-4 and the momentum is set as 0.9. The initial learning rate is set to be 1e-3 and decreases to 85% every 2 epochs.

The kernel size in PS module is set as 5. We set p_{top1} as 0.7 and k as 5 for peak selection. γ and β in AS module are 0.8 and 36. k^+ and k^- for GMMCE loss are both 3. Our experiments are implemented with PyTorch.

4.3 Ablation Study

To fully evaluate our method, progressive experiments are conducted to verify the effectiveness of the hue adjustment and the balanced sampler, our ASAE layer and GMMCE loss. Further, we compare our method with other works on bone marrow cells.

Hue Adjustment and Balanced Sampler. As illustrated in Section 3.5, the variance of hue increases the intra-class variance. In our dataset, we adjust the hue of images to be close to the purple hue. Hence, we set Max as 155 and Min as 125. Some examples before and after the adjustment are seen in Fig. 5. As shown in Table 1, applying the hue adjustment to our dataset increases our accuracy from 72.61% to 73.02%. It indicates that the uniform hue makes sense to increase the accuracy of our model.

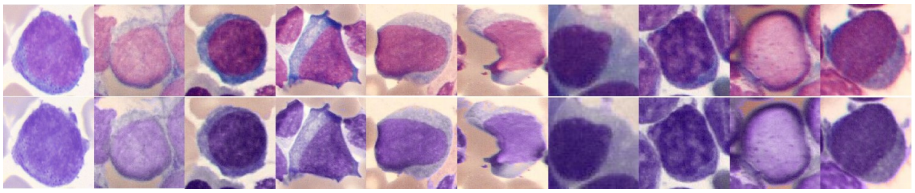

Fig. 5. Effect of the hue adjustment. The upper row shows some cells in our dataset before the hue adjustment. The below row shows the corresponding cells after the hue adjustment.

Table 1. Ablation experiments on the hue adjustment and the balanced sampler

Hue adjustment	Balanced sampler	Accuracy(%)
		72.61
✓		73.02
✓	✓	**73.53**

As shown in Fig. 4, the problem of the class imbalance is severe in our dataset. Without the balanced sampler, we find the number of monocytes to be 0 or 1 in many batches. We randomly sample 6 images in each class to form a single batch in an iteration and the balanced sampler effectively avoids the absence of some classes of cells in a batch. With the balanced sampler, the accuracy increases from 73.02% to 73.53%. It indicates the effectiveness of a balanced sampler.

ASAE Layer. We propose an AS branch to hide a discriminative region in order to force the model to learn better from the others. Also, we propose a novel ASAE-Net based on ResNet-50 with two branches sharing parameters. We conduct experiments to verify the effectiveness of our ASAE-Net and explore which layer to insert our AS branch. To better verify the effectiveness of our proposed ASAE layer, we also conduct experiments on different backbones including VGG-16 [45] and ResNet-18 [38].

Based on the results in Table 2, ResNet-50 benefits from our AS branch. With AS branch inserted after layer3, our model gains 1.26% improvement from 73.53% to 74.79% and with AS branch inserted after layer4, our model gains 0.50% improvement from 73.53% to 74.03%. It demonstrates AS branch is more effective when inserted after layer3.

Based on our analysis, with AS branch after layer3, suppressed feature maps are sent to layer4 and then modeled in a non-linear way. The non-linear model exploits the information of suppressed feature maps. However, when AS branch is inserted after layer4, suppressed feature maps are sent to the classifier directly, which cannot maximize its effect.

Also, when we insert our AE branch into ResNet-50, we gain 0.57% improvement. When we combine AS branch and AE branch together with sharing parameters, our model gains 0.32% improvement than using a single AS branch.

Table 2. Ablation experiments on ASAE layer

Network	Accuracy(%)
ResNet-50	73.53
ResNet-50 + AS branch (after layer4)	74.03
ResNet-50 + AS branch (after layer3)	74.79
ResNet-50 + ASAE layer (after layer3)	**75.11**

Table 3. Experiments on different backbones

Network	Accuracy(%)
VGG-16	71.21
VGG-16 + ASAE layer	**72.03**
ResNet-18	72.13
ResNet-18 + ASAE layer	**73.36**
ResNet-50	73.53
ResNet-50 + ASAE layer	**75.11**

It shows that the two branches promote each other. Also, it verifies that our ASAE-Net with ASAE layer after layer3 is effective without increasing training parameters.

Based on the results in Table 3, three different backbones gain different degrees of improvement when an ASAE layer is inserted after the penultimate layer. VGG-16 gains 0.82% improvement from 71.21% to 72.03% and ResNet-18 gains 1.23% from 72.13% to 73.36%. Also, ResNet-50 gains 1.58% improvement from 73.53% to 75.11%. The results amply prove the effectiveness of our ASAE layer.

GMMCE Loss. We propose a GMMCE loss to alleviate the problem of category confusion. To verify the effectiveness of it, we train the model with different losses including CE loss, GCE loss and GMMCE loss. Based on Table 4, training with our GMMCE loss outperforms the others. When trained with our proposed GMMCE loss, the model gains 2.52% and 1.21% improvement than CE and GCE loss respectively. The results prove that considering negative classes unequally is an effect strategy. Also, it verifies that taking negative classes with low confidence scores into consideration is effective when compared with GCE loss. As we illustrate, it can be regarded as a regularization method like label smoothing.

Table 4. Ablation experiments on the training loss

Training loss	Accuracy(%)
CE	75.11
GCE	76.42
GMMCE	**77.63**

4.4 Comparison with Other Methods

There are some related studies including WBCNet [14] and Cell3Net [17] in the fine-grained classification task of bone marrow cells. Also, there are several methods based on attention mechanism including Non-local Neural Network (NLNet) [46], Squeeze-and-Excitation Net (SENet) [47] and ADL [23]. We conduct experiments on different methods on our dataset. Among them, NLNet, SENet and ADL are based on the backbone ResNet-50 [38]. We discover our method outperforms all the others. It indicates our well-designed method for the fine-grained bone marrow cells classification is effective.

Table 5. Comparison with other methods

Method	Accuracy(%)
WBCNet [14]	69.63
Cell3Net [17]	70.21
ResNet-50 [38]	72.61
NLNet [46]	73.12
SENet [47]	73.29
ADL [23]	73.84
Proposed method	**77.63**

5 Conclusion

In the paper, we proposed a two-branch framework called ASAE-Net based on attention mechanism and weakly supervised learning for the fine-grained classification of bone marrow cells. It can locate several discriminative regions with an unlimited quantity to capture subtle differences between subclasses. The inserted layer is plug-and-play without training parameters. Besides, we put forward a novel suppression approach to encourage the model to learn better. In addition, a GMMCE loss is designed to alleviate the confusion. Due to the problems of data, we adopted a hue adjustment strategy and a balanced sampler. The experiments have shown the superiors of our method.

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China under Grant U1801262, the Guangzhou Key Laboratory of Body Data Science under Grant 201605030011, the Science and Technology Project of Zhongshan under Grant 2019AG024 and the Guangzhou City Science and Technology Plan Project under Grant 202002020019.

References

1. Labati, R.D., Piuri, V., Scotti, F.: All-idb: The acute lymphoblastic leukemia image database for image processing. In: International Conference on Image Processing (ICIP). (2011) 2045–2048
2. Pan, C., Park, D.S., Yang, Y., Yoo, H.M.: Leukocyte image segmentation by visual attention and extreme learning machine. Neural Comput. Appl. **21** (2012) 1217–1227
3. Razzak, M.I., Naz, S.: Microscopic blood smear segmentation and classification using deep contour aware cnn and extreme machine learning. In: Computer Vision and Pattern Recognition Workshops (CVPRW). (2017) 801–807
4. Hegde, R.B., Prasad, K., Hebbar, H., Singh, B.M.K.: Development of a robust algorithm for detection of nuclei and classification of white blood cells in peripheral blood smear images. J. Med. Syst. **42** (2018) 110
5. Madhloom, H.T., Kareem, S.A., Ariffin, H.: A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia. In: Advanced Computer Science Applications and Technologies (ACSAT). (2012) 330–335
6. Rajendran, S., Arof, H., Mokhtar, N., Mubin, M., Yegappan, S., Ibrahim, F.: Dual modality search and retrieval technique analysis for leukemic information system. Sci. Res. Essays **6** (2011)
7. Ramesh, N., Dangott, B., Salama, M.E., Tasdizen, T.: Isolation and two-step classification of normal white blood cells in peripheral blood smears. J. Pathol. Inform. **3** (2012) 13–13
8. Sinha, N., Ramakrishnan, A.G.: Automation of differential blood count. In: Conference on Convergent Technologies for Asia-Pacific Region. Volume 2. (2003) 547–551
9. Theera-Umpon, N., Dhompongsa, S.: Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification. IEEE Trans. Inf. Technol. Biomed. **11** (2007) 353–359
10. Vincent, I., Kwon, K., Lee, S., Moon, K.: Acute lymphoid leukemia classification using two-step neural network classifier. In: Frontiers of Computer Vision (FCV). (2015) 1–4
11. Chandradevan, R., Aljudi, A.A., Drumheller, B.R., Kunananthaseelan, N., Amgad, M., Gutman, D.A., Cooper, L.A., Jaye, D.L.: Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. Lab. Invest. **100** (2020) 98–109
12. Choi, J.W., Ku, Y., Yoo, B.W., Kim, J.A., Lee, D., Chai, Y.J., Kong, H., Kim, H.C.: White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. PLoS One **12** (2017)
13. Hegde, R.B., Prasad, K., Hebbar, H., Singh, B.M.K.: Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images. Biocybern. Biomed. Eng. **39** (2019) 382–392
14. Jiang, M., Cheng, L., Qin, F., Du, L., Zhang, M.: White blood cells classification with deep convolutional neural networks. Int. J. Pattern Recognit. Artif. Intell. **32** (2018)
15. Liang, G., Hong, H., Xie, W., Zheng, L.: Combining convolutional neural network with recursive neural network for blood cell image classification. IEEE Access **6** (2018) 36188–36197

16. Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. bioRxiv (2019)
17. Qin, F., Gao, N., Peng, Y., Wu, Z., Shen, S., Grudtsin, A.: Fine-grained leukocyte classification with deep residual learning for microscopic images. Comput. Meth. Programs Biomed. **162** (2018) 243–252
18. Shahin, A.I., Guo, Y., Amin, K.M., Sharawi, A.A.: White blood cells identification system based on convolutional deep neural learning networks. Comput. Meth. Programs Biomed. **168** (2017) 69–80
19. Tiwari, P., Qian, J., Li, Q., Wang, B., Gupta, D., Khanna, A., Rodrigues, J.J.P.C., De Albuquerque, V.H.C.: Detection of subtype blood cells using deep learning. Cogn. Syst. Res. **52** (2018) 1036–1044
20. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: European Conference on Computer Vision (ECCV). (2014) 834–849
21. Angelova, A., Zhu, S., Lin, Y.: Image segmentation for large-scale subcategory flower recognition. In: Workshop on Applications of Computer Vision (WACV). (2013) 39–45
22. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: European Conference on Computer Vision (ECCV). (2018) 438–454
23. Choe, J., Lee, S., Shim, H.: Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1–1
24. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Computer Vision and Pattern Recognition (CVPR). (2017) 4476–4484
25. Hu, T., Qi, H.: See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv:1901.09891 (2019)
26. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: European Conference on Computer Vision (ECCV). (2018) 834–850
27. Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Computer Vision and Pattern Recognition (CVPR). (2015) 842–850
28. Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: International Conference on Computer Vision (ICCV). (2015) 2399–2406
29. Wei, Y., Feng, J., Liang, X., Cheng, M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Computer Vision and Pattern Recognition (CVPR). (2017) 6488–6496
30. Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: Computer Vision and Pattern Recognition (CVPR). (2016) 1134–1142
31. Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S.: Diversified visual attention networks for fine-grained object classification. IEEE Trans. Multimedia **19** (2017) 1245–1256
32. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: International Conference on Computer Vision (ICCV). (2017) 5219–5227

33. Sun, G., Cholakkal, H., Khan, S., Khan, F.S., Shao, L.: Fine-grained recognition: Accounting for subtle differences between similar classes. arXiv:1912.06842 (2019)
34. Gao, X., Li, W., Loomes, M., Wang, L.: A fused deep learning architecture for viewpoint classification of echocardiography. Inf. Fusion **36** (2017) 103–113
35. Zhang, J., Zhong, Y., Wang, X., Ni, G., Du, X., Liu, J., Liu, L., Liu, Y.: Computerized detection of leukocytes in microscopic leukorrhea images. Med. Phys. **44** (2017) 4620–4629
36. Zhao, J., Zhang, M., Zhou, Z., Chu, J., Cao, F.: Automatic detection and classification of leukocytes using convolutional neural networks. Med. Biol. Eng. Comput. **55** (2017) 1287–1301
37. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS). (2012) 1097–1105
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR). (2016) 770–778
39. Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J.: Weakly supervised instance segmentation using class peak response. In: Computer Vision and Pattern Recognition (CVPR). (2018) 3791–3800
40. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15** (2014) 1929–1958
41. Tompson, J., Goroshin, R., Jain, A., Lecun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Computer Vision and Pattern Recognition (CVPR). (2015) 648–656
42. Park, S., Kwak, N.: Analysis on the dropout effect in convolutional neural networks. In: Asian Conference on Computer Vision (ACCV). (2016)
43. Ghiasi, G., Lin, T., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: Neural Information Processing Systems (NIPS). (2018) 10727–10737
44. Durand, T., Thome, N., Cord, M.: Mantra: Minimum maximum latent structural svm for image classification and ranking. In: International Conference on Computer Vision (ICCV). (2015) 2713–2721
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). (2015)
46. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Computer Vision and Pattern Recognition (CVPR). (2018)
47. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42** (2020) 2011–2023