Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20 (10): 18301–18317.
DOI: 10.3934/mbe.2023813
Received: 07 July 2023
Revised: 05 September 2023
Accepted: 17 September 2023
Published: 22 September 2023

Research article

Boosting microscopic object detection via feature activation map guided poisson blending

Haixu Yang[1,2,†], Yunqi Zhu[1,†], Jiahui Yu[2], Luhong Jin[1], Zengxi Guo[3], Cheng Zheng[3], Junfen Fu[4] and Yingke Xu[1,2,4,*]

[1] Department of Biomedical Engineering, MOE Key Laboratory of Biomedical Engineering, State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang Provincial Key Laboratory of Cardio-Cerebral Vascular Detection Technology and Medicinal Effectiveness Appraisal, Zhejiang Provincial Key Laboratory of Traditional Chinese Medicine for Clinical Evaluation and Translational Research, Zhejiang University, Hangzhou, 310027, China.

[2] Binjiang Institute of Zhejiang University, Hangzhou, 310053, China.

[3] Zhejiang Institute for Food and Drug Control, NMPA Key Laboratory of Quality Evaluation of Traditional Chinese Medicine (Traditional Chinese Patent Medicine), Hangzhou 310052, China

[4] Department of Endocrinology, Children's Hospital of Zhejiang University School of Medicine, National Clinical Research Center for Children's Health, Hangzhou, 310051 China

* **Correspondence:** Email: yingkexu@zju.edu.cn.

†These authors contributed to the work equally and should be regarded as co-first authors.

Abstract: Microscopic examination of visible components based on micrographs is the gold standard for testing in biomedical research and clinical diagnosis. The application of object detection technology in bioimages not only improves the efficiency of the analyst but also provides decision support to ensure the objectivity and consistency of diagnosis. However, the lack of large annotated datasets is a significant impediment in rapidly deploying object detection models for microscopic formed elements detection. Standard augmentation methods used in object detection are not appropriate because they are prone to destroy the original micro-morphological information to produce counterintuitive micrographs, which is not conducive to build the trust of analysts in the intelligent system. Here, we propose a feature activation map-guided boosting mechanism dedicated to microscopic object detection to improve data efficiency. Our results show that the boosting mechanism provides solid gains in the object detection model deployed for microscopic formed elements detection. After image augmentation, the mean Average Precision (mAP) of baseline and strong baseline of the Chinese herbal medicine micrograph dataset are increased by 16.3% and 5.8% respectively. Similarly, on the urine sediment dataset, the boosting mechanism resulted in an improvement of 8.0% and 2.6% in mAP of the baseline and strong baseline maps respectively. Moreover, the method shows strong generalizability and can be easily integrated into any main-stream object detection model. The performance enhancement is interpretable, making it more suitable for microscopic biomedical applications.

Keywords: microscopic examination; deep learning; object detection; synthetic images

1. Introduction

Medical micrographs contain a wealth of formed elements information and are important reference for biomedical research and clinical medical diagnosis. Microscopic examination of visible components is the gold standard for testing in many biomedical fields. Microscopic morphological examination is very tedious, time-consuming and experience-dependent for analysts, due to the complexity of biological entities. Computerized methods are a very important part of modern biomedical diagnosis. The application of computerized image processing technology to the auxiliary analysis of microscopic formed elements not only improves the efficiency of the analyst but also provides decision support to ensure the objectivity and consistency of diagnosis [1]. In particular, in formed elements detection, deep learning-based methods [2] have made great strides in overcoming the limitations of traditional approaches and have been successfully used in micrograph analysis [3]. This has promoted research in computer-aided formed elements detection [4,5].

Object detection models based on state-of-the-art convolutional networks [6–8] are often data-hungry. Their performance increases logarithmically based on the amount of training data, with larger datasets achieving better models [9]. However, collecting and annotation the data is a particularly common limiting step in the application of deep learning for the analysis of biomedical micrographs [10]. Unlike the process of constructing non-biomedical datasets, the rarity of diseases and privacy constraints make it particularly difficult to obtain large-scale datasets. Moreover, even if enough images are acquired, expert knowledge is commonly required to identify and segment structures of interest in micrographs, especially for the comparison of multiple biological conditions [11]. This is often an expensive and time-consuming task. The lack of large annotated datasets is a major impediment to the rapid deployment of object detection models for microscopic formed elements detection, which resulting in even the most advanced detection models not being directly deployable. We realize that it is of modest improvement in real performance to build and optimize end-to-end frameworks for microscopic image analysis, comparing with developing new methods to improve the data efficiency for state-of-the-art object detection models when adapting them to specific tasks.

Leveraging data augmentations to boost data efficiency is a promising direction toward addressing this challenge. Recent studies have taken advantage of the unreasonable effectiveness [12] of images in deep learning employed mix-based [13,14] or erasure-based [15,16] augmentation methods in object detection tasks to elevate performance . However, these methods have some obvious drawbacks in micrograph analysis applications. First, the augmented data are inexplainable from a human perspective. The performance enhancement from counterintuitive images produced by mixing is very difficult to understand or explain. In biomedical image processing or microscopic morphological analysis, such images are not a useful transformation for human. It is important to establish the trust of the analyst in the intelligent system[17], in other words, the interpretability of methods is important. An augmentation that is more object-aware is more likely to be useful for microscopic morphological examination. Second, in the case of extremely limited data, such approaches are prone to destroy rare or even unique morphological information, causing further deepen the overfitting of the model or inefficiency in training.

In this study, we propose a feature activation map-guided boosting mechanism dedicated to microscopic object detection, which can perform interpretable augmentation and provide solid gains in detection baselines to be deployed for microscopic morphological examination. The boosting mechanism exploits the spatial attention itself to extract biologically meaningful features. It reuses these information to increase the number of microscopic formed elements in the training data. The results show that this novel method significantly improves the data efficiency on the Chinese herbal medicine micrograph datasets we collected from Zhejiang Institute for Food and Drug Control. Using YOLOV5 as the object detection backbone, we achieve 16.3% and 5.8% mean Average Precision (mAP) improvements under baseline and strong baseline (with general data augmentation methods). And a similar improvement was observed in the urine sediment dataset, which resulting in 8.0% and 2.6% increase in mAP under baseline and strong baseline. Moreover, generalizability is demonstrated in the mainstream object detection models. In summary, our contributions are as follows:

(1) We propose a feature activation map-guided boosting mechanism for microscopy object detection. Compared to generic methods, our solution maintains fidelity to the augmented images without producing counter-intuitive data or inducing artifacts. The performance improvements from the proposed method are interpretable, which facilitate to build analyst trust.

(2) The boosting mechanism greatly improves data efficiency for microscopic morphological examination. It is additive to other data augmentation methods, causing further improves in model performance. In addition, it has strong generalizability and can be easily integrated into other mainstream object detection models.

(3) We validated the performance of our proposed method in typical micrograph datasets, which demonstrates the substantial improvement in object detection tasks.

2. Related work

In this section, we briefly review the object detection models and data augmentation methods commonly used in object detection.

Object detection. Object detection [2] is one of the most important and challenging branches of computer vision, to provide location and category information about a target in a video or image. Object detectors are evolved into two main categories: two-stage and one-stage. The two-stage detector has a separate module to generate region recommendations. These models generate region proposals in an image during the first stage and then classify and localize them in the second stage. Representative two-stage models include Regions with CNN features (R-CNN) [18], Spatial Pyramid Pooling in Deep Convolutional Networks (SPP-net) [19], Feature Pyramid Network (FPN) [20], and Fast Region-based Convolutional Network(Fast R-CNN) [21]. The one-stage detector uses pre-defined bounding boxes of various scales and aspect ratios to localize objects and then directly predicts the category and location of the targets of interest. The one-stage model represented by YOLO (You Only Look Once) [22] is ahead of the two-stage models in terms of real-time performance and has a simpler structure. Most of the state-of-the-art object detection models are based on single-segment architecture such as YOLOv6[6], YOLOv7 [7], and EfficientDet [23].

Data Augmentations. Data augmentation is the key to the success of most neural networks in the machine vision domain. Common forms of data augmentation include random crops [24], color jittering , scale jittering , and random flipping, which mainly exploit the invariance of data transformations to alleviate the overfitting of the model for improving the robustness of the model. Linear and Non-Linear mixing augmentations, such as Mixup [13], Cutmix [14], and Cutout [16], mix the information contained in different images together, takes advantage of the unreasonable effectiveness of images in deep learning to achieve performance improvement . The counterintuitive images produced by this type of augmentation are very effective in training models that can significantly improve performance across a variety of tasks and domains even after encode invariances augmentation are considered, but little is known about why such methods work. The interpretability of methods is important in microscopic morphological analysis, such mixing images are not a useful transformation for a human observer. In contrast to these studies, with the help of feature reuse and natural blending, the augmentation images produced by our method are more realistic and natural in the context of microscopy. Moreover, the performance improvement of this method is interpretable, thus the method would be more feasible to be employed in bioimage analysis.

3. Methods

3.1. Approach overview

Our goal is to perform interpretable augmentation to improve the data efficiency of state-of-the-art object detection models and adapt them to microscopic formed elements detection tasks. The overview of our proposed method is given in Figure 1, we exploit a gradient-weighted class activation map to gain spatial attention from the final convolutional layer of the object detection model. The resultant attention maps were leveraged in microscopic images to extract the formed element coarse segmentation masks. Finally, reconstruct images from the gradient of the mask bootstrap region by the Poisson equation, leading to a combinatorial number of new training data to ultimately improve the data utilization efficiency of the detection model.

3.2. Approach details and analysis

3.2.1 Microscopic object detection baseline

Unlike images in nature, microscopic images contain a large number of tiny, fine objects, and the density of objects in different batches of samples varies significantly. This makes the detection of microscopic formed elements more challenging when only a small number of annotated samples are available.

We choose YOLOv5 [25] as the microscopic detection framework because it is reliable and stable enough as a baseline. In particular, the addition of shallow bypass branches in YOLOv5-6.0 improves the focus of the model on fine-grained features, which is beneficial for microscopic morphological

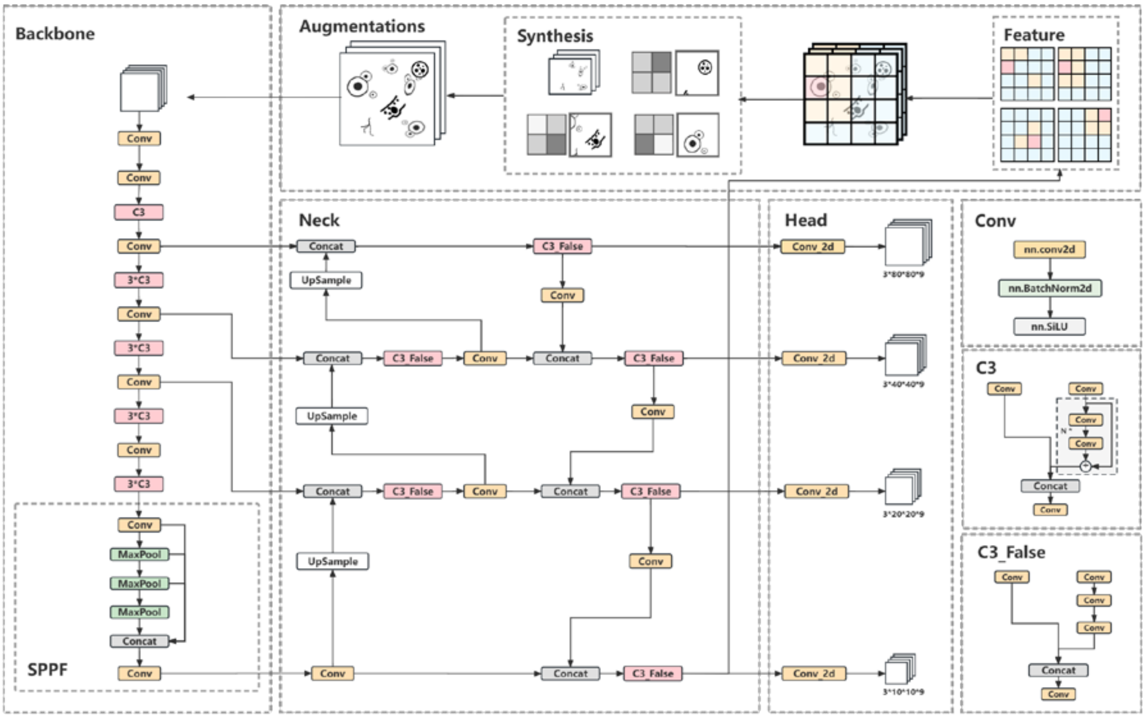

Figure 1. Overview of the feature activation map guided object detection model boosting framework. The boosting mechanism is integrated in the object detection baseline. After the Neck aggregates the features of different layers in Backbone, spatial attention is captured from the final convolutional layer of Neck and is used to guide the acquisition of biomedically meaningful regions and coarse segmentation masks for feature reuse. The obtained formed element instances are randomly blending into the micrograph background to obtain the combined number of new training data. SPPF represent Spatial Pyramid Pooling- Fast. Conv2d and BatchNorm2d denote 2d convolution and Batch Normalization, respectively. SiLU is the activation function Sigmoid Linear Unit.

feature extraction. The structure of YOLOv5 consists of three main parts: backbone, neck, and head. The Backbone is the core structure of YOLOv5 for initial feature extraction, consisting of Conv, C3 (CSPNet Bottleneck with three convolutions), and Spatial Pyramid Pooling - Fast (SPPF) modules. The role of the Neck is to aggregate the features of different layers in the Backbone to improve the recall and positioning accuracy of objects at different scales. YOLOv5 combines FPN [20] and PANet [17] as Neck for feature bi-directional aggregation, where FPN transfers the semantics of high-level features from top to bottom, in contrast to PAN, which transfers low-level localization features upwards. The resulting four feature maps are used to detect tiny, small, medium, and large objects in the images. Finally, the Head performs confidence calculations and bounding box regressions with pre-defined priori bounding boxes in the four feature maps to obtain object information, which include category, confidence and bounding box coordinates.

Most of the experiments in this work will be performed on this baseline, but we also explore the generalizability of the proposed approach to other object detection frameworks.

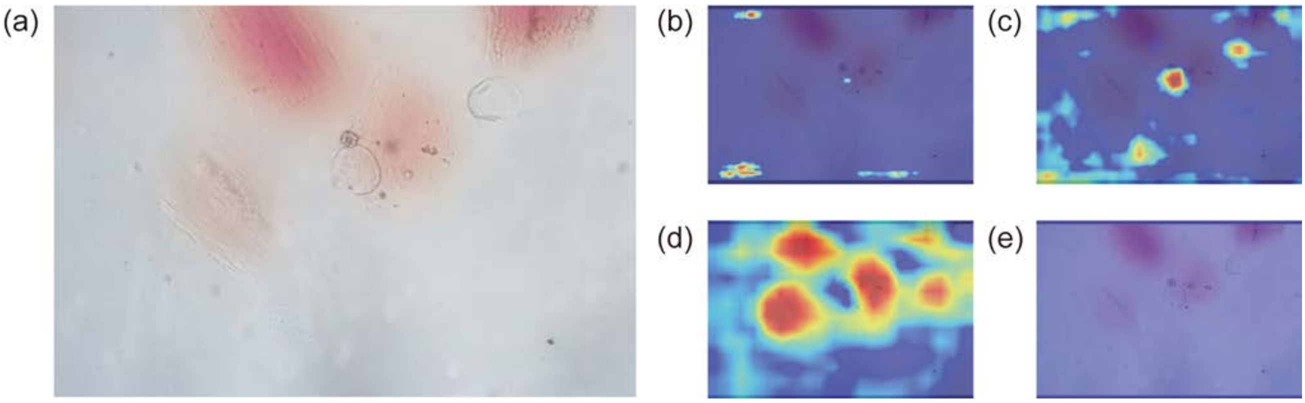

Figure 2. Spatial attention gained by Grad-CAM. (a) is the micrograph input of model. (b)–(e) are spatial attention maps captured from the 23rd, 26th, 29th and 32nd convolution layers (the final convolution layer of Neck), respectively. The attention maps of objects with different scales appear on different convolution layers. Spatial attention at layer 26 captures the features of pollen, and spatial attention at layer 29 captures the features of epidermalcells.

3.2.2 Acquisition of the biological regions using gradient-weighted class activation map (Grad-CAM)

Visualizing the spatial attention of a model is an effective way to assess whether the decisions of the model are based on biologically meaningful information [26]. The acquisition of biomedically meaningful regions based on model attention is key to information reuse when few annotated instances are available, as it provides unannotated biomedical instances. Gradient-weighted class activation map (Grad-CAM) [27] is a class-discriminative localization technique that can produce "visual interpretation" for decisions of convolutional neural network (CNN)-based models. We exploit Grad-CAM to gain spatial attention from the final convolutional layer of the object detection model as show in Figure 2. The resultant attention maps were leveraged in microscopic images for weakly supervised localization to extract the formed element coarse segmentation masks.

To obtain the formed elements discriminative localization map in micrographs, the input image is inferred through detection model by forward propagation and a score y^c for the target category is obtained. Grad-CAM then calculates the gradient for class c in the feature map A^k of the four convolutional layers before the Head of the detector, i.e. $\frac{\partial y^c}{\partial A_{i,j}^k}$, and performs a global average pooling of the gradient to obtain the neuron importance weights α_k^c. The aforementioned calculation process can be represented as:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \tag{1}$$

Where i, j represent the horizontal and vertical coordinates of the pixels in the feature map A^k, Z represents the sum of pixel points in the feature map. The weights α_k^c represents the 'importance' of each pixel on the feature map k for a target class c. Finally, the class activation map is obtained after the linear combination of weights and feature maps by Rectified Linear Unit (ReLU).

$$L_c = ReLU\left(\sum_k \alpha_k^c A^k\right) \tag{2}$$

Following the above calculations, we can obtain biological instances and their coarse segmentation masks guided by activation mapping, which provides good boundary conditions for Poisson blending in the next step.

3.2.3 Synthetic images

Directly blending the target into the microscope background images by nonlinear methods in the spatial domain (e.g., Cutmix) can result in pixel artifacts, as well as photometric inconsistencies between the instance and the background. Such artifacts are not only unkindly to human observers, but also may affect the model judgment of local semantic information and deteriorate the accuracy of the detector [28]. The dissimilarity between the images is defined by the distance between their derivatives , so we can blend images in the gradient domain instead of the spatial domain to reduce pixel artifacts and photometric differences.

Poisson blending [29] is a common gradient-domain image blending method. In this paper, Poisson blending is used to combine instance images captured by the class activation map with microscope background images. The basic principle of the Poisson blending is to exploit the two conditions that the gradient vector fields of the mixed region and the background image are equal, the pixel values of the boundary of the mixed region and the boundary of the target image are equal, to construct the corresponding Poisson equation by the Laplace operator, and to interpolate each position of the blending region by using the solution of the equation to naturally and smoothly reconstruct the synthetic image. The blending process as shown in Figure 3. Poisson blending first calculates the vector field v for the source mixed region g. Then the Poisson equation is formulated to establish the relationship between the vector field of the source mixed region and the gradient of the target background image S, ensuring the consistency of color and texture. By iteratively solving the Poisson equation (as shown in formula 3), we obtain the reconstructed result for the source mixed region. Subsequently, the reconstructed portion of the source region is seamlessly blended with the remaining parts of the target background, resulting in the creation of the final synthetic image.

$$\min_f \iint_\Omega |\nabla f - v|^2 \; with \;\; f|_{\partial\Omega} = \; f^*|_{\partial\Omega} \tag{3}$$

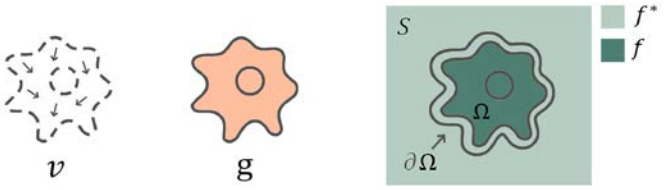

Figure 3. Poisson image blending. Here, g is the source mixed region, v is the vector field of g, and S is the background.

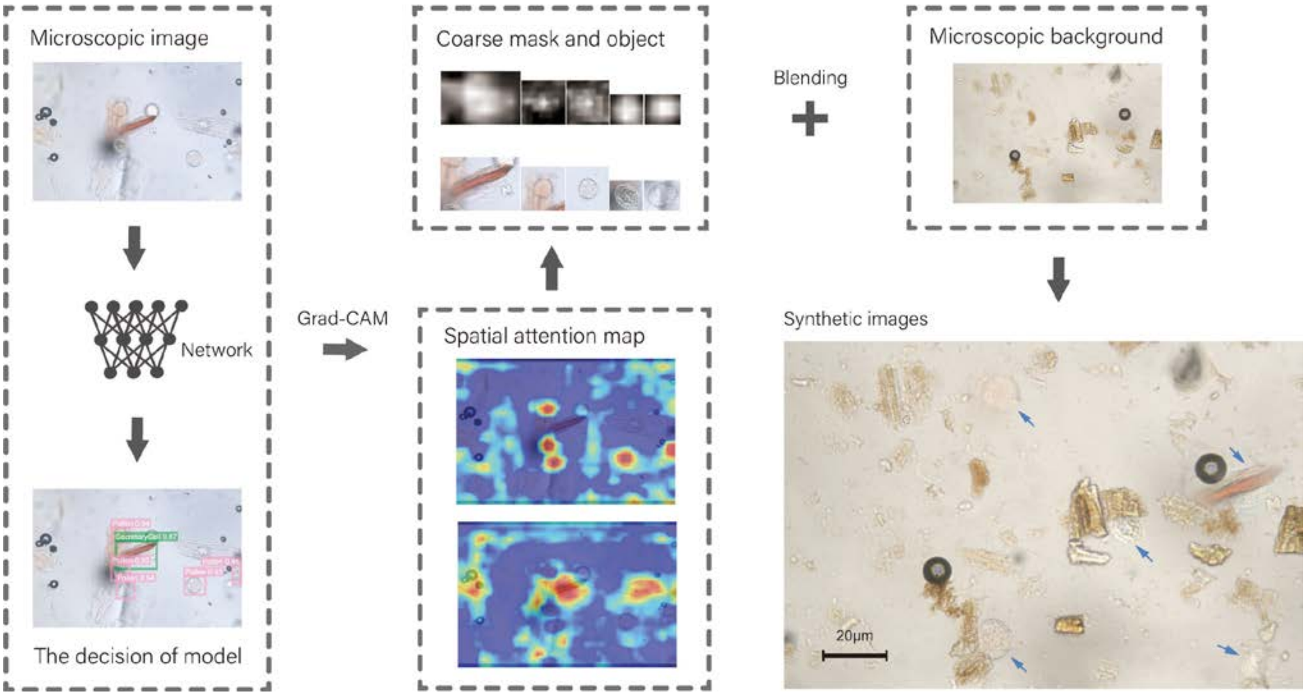

Figure 4. The process of feature activation map guided Poisson blending. The spatial attention map accurately reflects the region of interest of the network, and the formed elements within the region are randomly blended into the microscopic background image by Poisson blending. The blue arrows indicate the objects that our method added for feature reuse.

where Ω is the area covered by the target formed elements in the synthetic image, and $\partial\Omega$ is the boundary of Ω. The pixel representation function of the synthetic image inside Ω is f, and the pixel value outside Ω represents f^*.

The above process requires the user to provide the boundary of the desired region. In this paper, coarse segmentation masks obtained by class activation map are used as instances boundary to guide Poisson blending, where regions with formed elements are randomly blended into the microscopic background images during the model training for feature reuse and to improve the data utilization efficiency of the detector. The process of blending is shown in Figure 4. With the help of spatial attention and Poisson blending, the augmentation mechanism can effectively maintain the original biomedical information and visual features in microscopic images. The performance enhancement of this mechanism is interpretable, that is more applicable to bioimage applications. The resulting synthetic images are shown in Section 4.4.2 and are quantitatively compared with other methods.

4. Experimentation and results

In this part, we perform a series of experiments to demonstrate the effectiveness and generalizability of the feature map-guided boosting method. The comparison results with other augmentation methods commonly used in object detection are illustrated to validate the advantage of the proposed method. These experimental results are described separately in the following sections in details.

4.1. Data preparation

The microscopic datasets (Chinese herbal medicine datasets) for our method assessment were collected from Zhejiang Institute for Food and Drug Control. The collection of microscopic characteristic images of Chinese herbals was carried out in accordance with the relevant technical provisions in the general rules of Chinese Pharmacopoeia [30]. After grinding, sieving, permeabilization, and placement, images were collected with an Olympus CKX53 microscope with a Sony E3ISPM20000KPA camera at a magnification of 200 X and 400 X for small targets. These images were saved in JPEG format with pixel size of 5440x3648 and annotated by experienced laboratory experts using the graphical image annotation tool LabelImg. The datasets we collected consist of 774 images, including 1277 feature samples from 11 categories, which divided 7:3 between the training and test sets, as shown in details in Table 1.

Furthermore, to demonstrate the generality of the result, we also performed a methodological evaluation on the urine sediment dataset. Urine sediment data were collected from the clinical laboratory at university affilicated hostpital and annotated by experienced laboratory experts. The data

Table 1. Chinese herbal medicine sample quantity statistics.

Category	Training set	Test set
Pollen	271	147
Raphides	148	75
vug	201	69
hypha	80	30
SecretoryCell	25	22
epidermalcells	20	20
ParenchymatousCell	29	11
PericarpEpidermalCell	30	8
BrownParenchymatousCell	27	13
StoneCell	18	9
NonglandularHair	19	5
Total	868	409

Table 2. Urine sediment sample quantity statistics.

Category	Training set	Test set
erythrocyte	7583	3250
leukocyte	2753	1180
crystalline	1960	840
bacteria	4031	1728
epithelial cell	645	276
tubular	156	67
fungus	153	65
sperm	157	68
Total	17438	7474

were composed of 772 images with 5440x3648 pixels, including 24912 labeled instances across 8 distinct categories. The category and quantity information are shown in Table 2.

4.2. Experimental platform

The experiments were executed on five computers as following:

I. Computer with Windows 10 operating system, Intel (R) Core (TM) i9-11900F CPU, GeForce RTX 3080 graphic processing cards (with 12 GB memory).

II. Local workstation with Ubuntu 18.04 operating system, Intel® Xeon® E5-2699C v4 CPU, GeForce RTX 3080 graphic processing cards (with 10 GB memory).

III. Local workstation with the same specifications as II.

IV. Computer with Ubuntu 20.04 operating system, Inter® Core™ i9-11900 CPU, GeForce RTX 3080 graphic processing cards (with 10 GB memory).

V. Computer with Ubuntu 20.04 operating system, Intel® Xeon® W-2245 CPU, GeForce RTX 3090 graphic processing cards (with 24 GB memory).

All of the code were programmed in Python. All the models are run over GPU using the PyTorch deep learning framework. Each set of comparison tests was done separately on the same computer to avoid fluctuations in results caused by equipment differences.

4.3. Implementation details

The inputs of the detection models involved in the work are micrographs resized to 3 × 640 × 640. All models are initialized with ImageNet pre-trained parameters to speed up training, and trained by Adam for 300 epochs using a batch size of 16 and weight decay of 0.0005. For other hyperparameters and general data augmentation methods, including HSV (hue, saturation, and value) augmentation, translating, scaling, flipping (horizontally and vertically), and mosaic, we use the default configuration published unless otherwise stated.

The proposed boosting mechanism requires the user to provide the source mixed targets, the boundary of the source mixed targets and microscopic background image. During the model training, the model detects formed element instance and treats them as fusion targets. The spatial attention weights from the last convolutional layer are filtered to generate a non-binary coarse segmentation mask as the boundary of the mixed targets. This is done by selecting parts of the weight matrix that have a weight value greater than or equal to 50. The selected parts indicate the regions of the weight matrix that the model considers to be the most important for detection and classification. Additionally, microscopic background images are randomly selected samples from the database. Finally, the aforementioned information is utilized to perform Poisson fusion. The boosting mechanism randomly blends with an average of five different formed element instances into the microscopic background images with a given probability of 20%. The mechanism intervenes in training after 100 epochs to ensure the accuracy of the class activation region.

4.4. Assessment of the proposed method

4.4.1 Evaluation indicators

Average precision (AP) is applied to assess the performance of object detection methods. Specifically, the correctness of the object detection model depends on the value of the Intersection over Union (IOU). The objects detected by the model are sorted by confidence, and we can calculate the Precision and Recall to plot the P-R curve (precision-recall curve). The area under the P-R curve is AP, and the mAP is calculated by averaging the AP over different classes. Precision and Recall are defined in formulas 4–5:

$$Precision\ = \frac{TP}{TP+FP} \times 100\% \tag{4}$$

$$Recall\ = \frac{TP}{TP+FN} \times 100\% \tag{5}$$

Where TP means the number of true positive samples, TN means the number of true negative samples, FP means the number of false positive samples, and FN means the number of false negative samples.

In this study, we assert that the IOU between the inference box and the ground truth box over 50% is the true positive sample.

4.4.2 Performance evaluation

Firstly, we used YOLOV5 as a baseline to illustrate feature activation map-guided boosting mechanism elevates data efficiency. During the model training, the boosting mechanism randomly blends with an average of five different formed element instances into the microscopic background images with a given probability of 20%. The mechanism intervenes in training after 100 epochs to ensure the accuracy of the class activation region. All experiments were performed three times and the median value was measured.

Table 3. The detection results of the baseline and after augmentation.

Dataset	Chinese herbal medicine			Urine sediment		
Indicator	Precision (%)	Recall (%)	mAP (%)	Precision (%)	Recall (%)	mAP (%)
Baseline	51.4	44.1	41.3	39.2	21.5	22.2
Ours	**54.3**	**63.3**	**57.6**	**40.4**	**33.9**	**30.2**

With the same amount of training data, the higher the mAP, the better the data efficiency. We compared the detection results of the model before and after data augmentation on two datasets, Chinese herbal medicine and Urine sediment. The comparative results are summarized in Table 3.

In the experiment conducted on the Chinese herbal medicine dataset, our method demonstrated remarkable enhancements in the accuracy of the target detection model, recall rate, and mean Average Precision (mAP). Specifically, we observed an increase of 2.9% in accuracy, 19.2% in recall rate, and 16.3% in mAP after applying our method, showcasing substantial improvements compared to the baseline. These improvements can be primarily attributed to the feature activation map-guided boosting mechanism, which incorporates the detection of formed element instances, identified by the model, into other batches of microscopic images. This strategic integration facilitates efficient information reuse and greatly enhances the data efficiency of the detection model. Moreover, by fusing data from different batches, our method effectively alleviates the influence of data batch variations on the model, thereby improving its overall robustness. Similarly, we observed consistent performance improvements when applying our method to urine sediment datasets. In this case, we achieved a 1.2% increase in accuracy, a 12.4% increase in recall, and an 8.0% increase in mAP. These results demonstrate the universal applicability of our method. In both datasets, we observed a considerable improvement in the recall of the model. This improvement can be attributed to the integration of data from different batches, which mitigated the impact of batch variations and enhanced the overall robustness of the model. This characteristic is particularly valuable for solving real-world clinical problems, as recall levels directly affect the accuracy and reliability of test results.

Table 4. The performance comparison of the detection model trained with different augmentation methods. (Baseline-aug means strong baseline).

Dataset	Chinese herbal medicine			Urine sediment		
Indicator	Precision (%)	Recall (%)	mAP (%)	Precision (%)	Recall (%)	mAP (%)
Baseline-aug	87.7	72.9	81.6	58.0	57.4	55.9
Mixup	77.9	79.4	82.5	61.3	**59.1**	57.4
Cutout	83.8	79.1	85.2	52.4	58.7	54.9
Cutmix	**90.5**	72.4	85.5	57.7	57.2	55.7
Ours	87.1	**82.8**	**87.4**	**64.6**	57.0	**58.5**

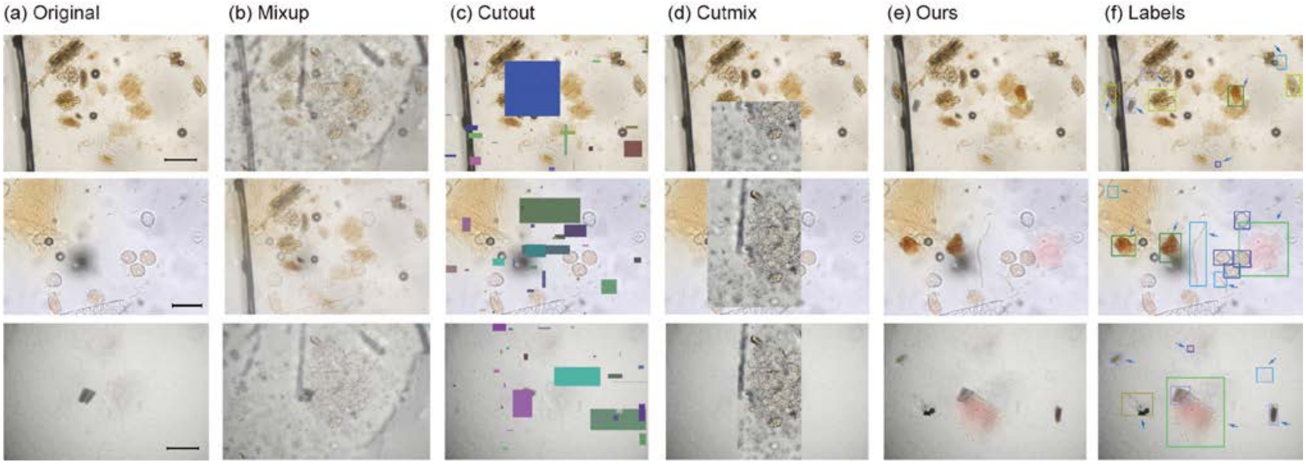

Figure 5. The visualized results of different augmentation methods used in experiment**s.** From (a) to (e) are Original, Mixup, Cutout, Cutmix and our proposed method. (f) The annotated object by using bounding box, where the blue arrows indicate the objects that were added by our method for feature reuse. Scale bars for the three figures in (a): 25 μm for the middle figure and 50 μm for the other figures.

Furthermore, the boosting mechanism can also significantly improve performance even after other forms of data augmentation are considered. The results are shown in Table 4. In our experiments with the Chinese herbal medicine dataset, we further increased the mAP by 5.8% on a strong baseline trained with general data augmentation. A similar improvement was observed in the urine sediment dataset, with a 2.6% increase in mAP. Compared with other mix-based methods commonly used in object detection, our method not only achieves the best results on both datasets, but also generates synthetic images that are more realistic and more faithful to the microscopic background image. As shown in Figure 5, although all methods have improvements in strong baseline, Mixup superimposes different images through linear interpolation, resulting in morphological aliasing, whereas Cutmix and Cutout destroy the local original features in micrographs, or even completely cover original object. Particularly, the Cutout, due to the higher target density in urine sediment data, causes destruction of critical data characteristics through large-scale random cutouts, resulting in a decrease in mAP (from 55.9% to 54.9%). Overall, compared to the above methods, our method not only delivers remarkable performance improvements, but also preserves the biomedical characteristics of the objects and is an appropriate method in biomedical applications.

4.4.3 Generalizability evaluation

In this part, we explore the generalizability of the proposed method on the Chinese herbal medicine dataset.

Among the modern object detection algorithms, the YOLO framework and its variations stand out for their remarkable balance of speed and accuracy. We test the generality of the proposed method on yolov3 [31], v6, and v7, as well as excellent detection models in the non-YOLO families, such as DetectoRS [32], Deformable DETR [33], ATSS [34], DINO [35], DDOD [36], AutoAssign [37] and DCNv2 [38]. The results of the models trained with the default configuration are used as the baseline and compared with the results after augmentation. All the mentioned non-YOLO methods were carried out with the help of the object detection toolbox MMDetection [39].

As shown in Table 5, all of these models, except for yolov6, show better mAP when boosting with our method. Especially on the DetectoRS, the performance gain can reach 8.2%, and on other models, mAP can also increase from 1.8% to 6.1%. These results demonstrate that our proposed method has strong generalizability and can be easily integrated with any mainstream object detection models to be deployed in microscopic formed elements detection to improve model performance.

5. Conclusions

In this paper, we proposed a feature activation map-guided boosting mechanism dedicated to microscopic formed elements detection. The method exploits a gradient-weighted class activation map to gain spatial attention from the object detection model. Under the guidance of spatial attention, the feature reuse of biomedical regions is performed using Poisson blending. The experimental results show that boosting mechanism effectively improves the data efficiency of the detection model. Specifically, on the Chinese herbal medicine micrograph dataset, the mAP of baseline and strong baseline are increased by 16.3% and 5.8%, respectively. Similarly, on the urine sediment dataset, the boosting mechanism resulted in an improvement of 8.0% and 2.6% in mAP of the baseline and strong baseline maps, respectively. At the same time, the method has strong generalizability and can be easily integrated with any other mainstream object detection models deployed in microscopic formed elements detection to further improve their performance. More importantly, compared to other methods, the performance enhancement achieved by this method is interpretable, which would greatly facilitate the building of trust for the analysts in the intelligent system. We hope this boosting mechanism will act as a baseline to assist excellent object detection models be more easily deployed to annotation lacking microscopic formed elements detection tasks to further advance performance in

Table 5. The mAP comparison of the mainstream detection models before and after augmentation.

Model	Baseline	Ours
YOLOv3	81.2	**85.3**
YOLOv6	**85.9**	85.6
YOLOv7	83.3	**85.1**
DetectoRS	77.5	**85.7**
Deformable DETR	76.6	**80.8**
ATSS	74.8	**80.1**
DINO	79.2	**85.3**
DDOD	78.3	**80.4**
AutoAssign	66.7	**71.2**
DCNv2	73.0	**76.4**

this domain.

Use of AI tools declaration

The authors declare that we have not used artificial intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by Zhejiang Provincial Natural Science Foundation (LZ23H180002), National Key Research and Development Program of China (2021YFF0700305), Zhejiang University K.P.Chao's High Technology Development Foundation (2022RC009), and the Fundamental Research Funds for the Central Universities (226-2023-00091).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. J. Hipp, T. Flotte, J. Monaco, J. Cheng, A. Madabhushi, Y. Yagi, et al., Computer aided diagnostic tools aim to empower rather than replace pathologists: Lessons learned from computational chess, *J. Pathol. Inform.*, **2** (2011), 25. https://doi.org/10.4103/2153-3539.82050
2. Z. Q. Zhao, P. Zheng, S. T. Xu, X. Wu, Object detection with deep learning: A review, *IEEE Transact. Neural Networks Learn. Syst.*, **30** (2019), 3212–3232. https://doi.org/10.1109/icABCD49160.2020.9183866
3. Z. Liu, L. Jin, J. Chen, Q. Fang, S. Ablameyko, Z. Yin, et al., A survey on applications of deep learning in microscopy image analysis, *Comput. Biol. Med.*, **134** (2021), 104523. https://doi.org/10.1109/TNNLS.2017.2766168

4. C. Matek, S. Schwarz, K. Spiekermann, C. Marr, Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks, *Nat. Machine Intell.*, **1** (2019), 538–544. https://doi.org/10.1038/s42256-019-0101-9
5. B. Midtvedt, J. Pineda, F. Skärberg, E. Olsén, H. Bachimanchi, E. Wesén, et al., Single-shot self-supervised object detection in microscopy, *Nat. Commun.*, **13** (2022), 7492. https://doi.org/10.1038/s41467-022-35004-y
6. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, et al., YOLOv6: A single-stage object detection framework for industrial applications, ArXiv:2209.02976, 2022. https://doi.org/10.48550/arXiv.2209.02976
7. C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in *CVF Conference on Computer Vision and Pattern Recognition*, 2023, 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
8. Z. Liu, H. Zhang, L. Jin, J. Chen, A. Nedzved, S. Ablameyko, et al., U-Net-based deep learning for tracking and quantitative analysis of intracellular vesicles in time-lapse microscopy images, *J. Innov. Opt. Health Sci.*, **15** (2022), 2250031. https://doi.org/10.1142/S1793545822500316
9. C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 843–852. https://doi.org/10.1109/ICCV.2017.97
10. V. Cheplygina, M. de Bruijne, J. P. W. Pluim, Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, *Med. Image Anal.*, **54** (2019), 280–296. https://doi.org/10.1016/j.media.2019.03.009
11. A. Bilodeau, C. V. L. Delmas, M. Parent, P. De Koninck, A. Durand, F. Lavoie-Cardinal, Microscopy analysis neural network to solve detection, enumeration and segmentation from image-level annotations, *Nat. Mach. Intell.*, **4** (2022), 455–466. https://doi.org/10.1038/s42256-022-00472-w
12. A. Halevy, P. Norvig, F. Pereira, The unreasonable effectiveness of data, *IEEE Intell. Syst.*, **24** (2009), 8–12. https://doi.org/10.1109/MIS.2009.36
13. H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in *International Conference on Learning Representations (ICLR)*, 2018.
14. S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, J. Choe, CutMix: Regularization strategy to train strong classifiers with localizable features, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6022–6031. https://doi.org/10.1109/ICCV.2019.00612
15. T. Devries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, ArXiv:1708.04552, 2017. https://doi.org/10.48550/arXiv.1708.04552
16. Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. https://doi.org/10.1609/aaai.v34i07.7000
17. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8759–8768. https://doi.org/10.1109/CVPR.2018.00913
18. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. https://doi.org//10.1109/CVPR.2014.81

19. K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, **1** (2005), pp.1458–1465. https://doi.org/10.1109/ICCV.2005.239
20. T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. https://doi.org/10.1109/CVPR.2017.106
21. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169
22. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91
23. M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and efficient object detection, arXiv:1911.09070, 2019. https://doi.org/10.1109/CVPR42600.2020.01079
24. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2012), 84–90. https://doi.org/10.1145/3065386
25. G. Jocher, A. Stoken, A. Chaurasia, J. Borovec, Y. Kwon, K. Michael, et al., ultralytics/yolov5: v6. 0—YOLOv5n ‘Nano’models, Roboflow integration, TensorFlow export, OpenCV DNN support, *Zenodo Tech. Rep.*, (2021).
26. W. Ouyang, C. F. Winsnes, M. Hjelmare, A. J. Cesnik, L. Åkesson, H. Xu, et al., Analysis of the Human Protein Atlas Image Classification competition, *Nat. Methods*, **16** (2019), 1254–1261. https://doi.org/10.1038/s41592-019-0658-6
27. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74
28. N. Dvornik, J. Mairal, C. Schmid, Modeling visual context is key to augmenting object detection datasets, in *European Conference on Computer Vision (ECCV) 2018*, Springer International Publishing, Cham, 2018, pp. 375–391. https://doi.org/10.1007/978-3-030-01258-8_23
29. P. Pérez, M. Gangnet, A. Blake, Poisson image editing, *ACM Trans. Graph.*, **22** (2003), 313–318. https://doi.org/10.1145/1201775.882269
30. C.C. Pharmacopoeia, Pharmacopoeia of the People's Republic of China, 2010.
31. J. Redmon, A. J. A. P. A. Farhadi, Yolov3: An incremental improvement, arXiv:1804.02767. 2018. https://doi.org/10.48550/arXiv.1804.02767
32. S. Qiao, L. C. Chen, A. Yuille, DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10208–10219. https://doi.org/10.1109/CVPR46437.2021.01008
33. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, in *International Conference on Learning Representations*, 2021.
34. S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://doi.org/10.1109/cvpr42600.2020.00978
35. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, et al., DINO: DETR with improved denoising anchor boxes for end-to-end object detection, arXiv:2203.03605, 2022. https://doi.org/10.48550/arXiv.2203.03605

36. Z. Chen, C. Yang, J. Chang, F. Zhao, Z. J. Zha, F. Wu, DDOD: Dive deeper into the disentanglement of object detector, *IEEE Transact. Mult.*, (2023), 1–15. https://doi.org/10.1109/TMM.2023.3264008
37. B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, et al., AutoAssign: Differentiable label assignment for dense object detection, arXiv:2007.03496, 2020. https://doi.org/10.48550/arXiv.2007.03496
38. X. Zhu, H. Hu, S. Lin, J. Dai, Deformable ConvNets V2: More deformable, better results, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9300–9308. https://doi.org/10.1109/CVPR.2019.00953
39. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, et al., MMDetection: Open MMLab detection toolbox and benchmark, arXiv:1906.07155, 2019. https://doi.org/10.48550/arXiv.1906.07155