Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-025-03332-y

ORIGINAL ARTICLE

OCCMNet: Occlusion-Aware Class Characteristic Mining Network for multi-class artifacts detection in endoscopy

Chenchu Xu[1] · Yu Chen[1] · Jie Liu[2] · Boyan Wang[3] · Yanping Zhang[1] · Jie Chen[1] · Shu Zhao[1]

Received: 14 May 2024 / Accepted: 11 January 2025

Abstract
Multi-class endoscope artifacts detection is crucial for eliminating interference caused by artifacts during clinical examinations and reducing the rate of misdiagnosis and missed diagnoses by physicians. However, this task remains challenging such as data imbalance, similarity, and occlusion among artifacts. To overcome these challenges, we propose an Occlusion-Aware Class Characteristic Mining Network (OCCMNet) to detect eight classes of artifacts in endoscope simultaneously. The OCCMNet comprises the following: (1) A Dual-Branch Class Rebalancing Module (DCRM) rebalances the impact of various classes by fully exploiting the benefits of two complementary data distributions, sampling and detecting from the majority and minority classes respectively. (2) A Class Discrimination Enhancement Module (CDEM) effectively enhances the discrepancy of inter-class by enhance important information and introduce nuance information nonlinearly. (3) A Global Occlusion-Aware Module (GOAM) infers the obscured part of the artifacts by capturing the global information to initially identify the obscured artifacts and combining local details to sense the overall structure of the artifacts. Our OCCMNet has been validated on a public dataset (EndoCV2020). Compared to the latest methods in both medical and computer vision detection, our approach demonstrated 3.5–6.5% improvement in mAP50. The results proved the superiority of our OCCMNet in multi-class endoscopic artifact detection and demonstrated its great potential in reducing clinical interference.

Keywords Endoscopic artifacts detection · Data imbalance · Class characteristic mining · Occlusion-aware

1 Introduction

The accurate detection of multi-class endoscopic artifacts is essential to aid endoscopists in mitigating artifact interference and enhancing the precision and efficiency of endoscopic examinations during clinical diagnosis, as shown in Fig. 1. These artifacts consist of eight common classes, including specularity, saturation, artifact, blur, contrast, bubbles, instrument, and blood, as shown in Fig. 2. Each class presents unique detection challenges due to its visual properties. Specifically, specularity is caused by light reflection on surfaces, saturation results from overexposure, and contrast arises from inadequate or excessive contrast, all of which can obscure subtle differences in tissue texture and color. Blur is caused by movement or focusing problems during image capture, leading to a loss of edge definition. Bubbles are thin films of liquid containing air that distort the appearance of underlying tissues. Instruments and blood can cause partial occlusion of the viewable area. Artifact is a comprehensive class that covers various artifacts that cannot be grouped into a specific category.

Accurately detecting these artifacts is crucial as it assists endoscopists in reducing the risk of misdiagnosis and missed diagnoses. Typically, artifacts severely damage the endoscopic video frames, making the presentation of initial lesions invisible or changed, which affects endoscopists' preliminary assessment of whether a lesion exists and its type [1]. Additionally, it enables to improve the efficiency of diagnostic and treatment. The inevitable presence of numerous diverse artifacts increases the endoscopists' information processing steps during endoscopic examinations, which, in turn, reduces the efficiency of the examination. Finally, it is crucial for the subsequent computer-assisted clinical analysis techniques [2]. High-quality endoscopic artifact detection is a necessary prerequisite for subsequent processing, laying a solid foundation for high-quality frame restoration and offering more possibilities for the development of other reliable intelligent endoscopic tools.

Extended author information available on the last page of the article

Published online: 05 March 2025

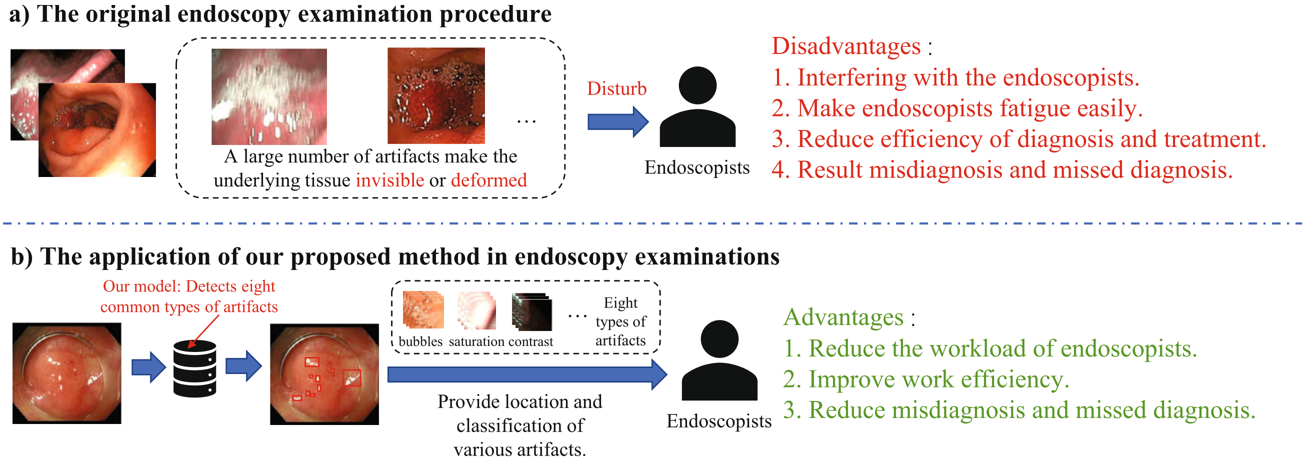

a) The original endoscopy examination procedure
A large number of artifacts make the underlying tissue invisible or deformed
…
Disturb
Endoscopists
Disadvantages :
1. Interfering with the endoscopists.
2. Make endoscopists fatigue easily.
3. Reduce efficiency of diagnosis and treatment.
4. Result misdiagnosis and missed diagnosis.
b) The application of our proposed method in endoscopy examinations
Our model: Detects eight common types of artifacts
bubbles
saturation
contrast
…
Eight types of artifacts
Provide location and classification of various artifacts.
Endoscopists
Advantages :
1. Reduce the workload of endoscopists.
2. Improve work efficiency.
3. Reduce misdiagnosis and missed diagnosis.

Fig. 1 The OCCMNet detects eight classes of endoscopic artifacts simultaneously, alleviates the effects of artifacts, and helps to improve the efficiency of endoscopy

However, the multi-class detection of endoscopic artifacts presents significant challenges, as shown in Fig. 3: (1) Artifacts of different classes are extremely imbalanced [3, 4]. Due to the significant differences in the frequency of occurrence of different artifact classes in actual data, where specularity has the highest proportion at 38.80%, while blood and instrument have much lower proportions at 1.71% and 1.99%, the model tends to be more inclined to predict the majority classes and ignore the minority classes, leading to poor performance on the minority classes. (2) The characteristics of inter-class are very similar while intra-class are varied in artifacts [5, 6]. Artifacts of different classes (e.g., specularity, bubbles, and artifact) may have similar characteristics (e.g., brightness and texture). Artifacts of the same classes (e.g., artifact class) may have very different characteristics (e.g., size, shape, and texture). It brings a huge challenge to the

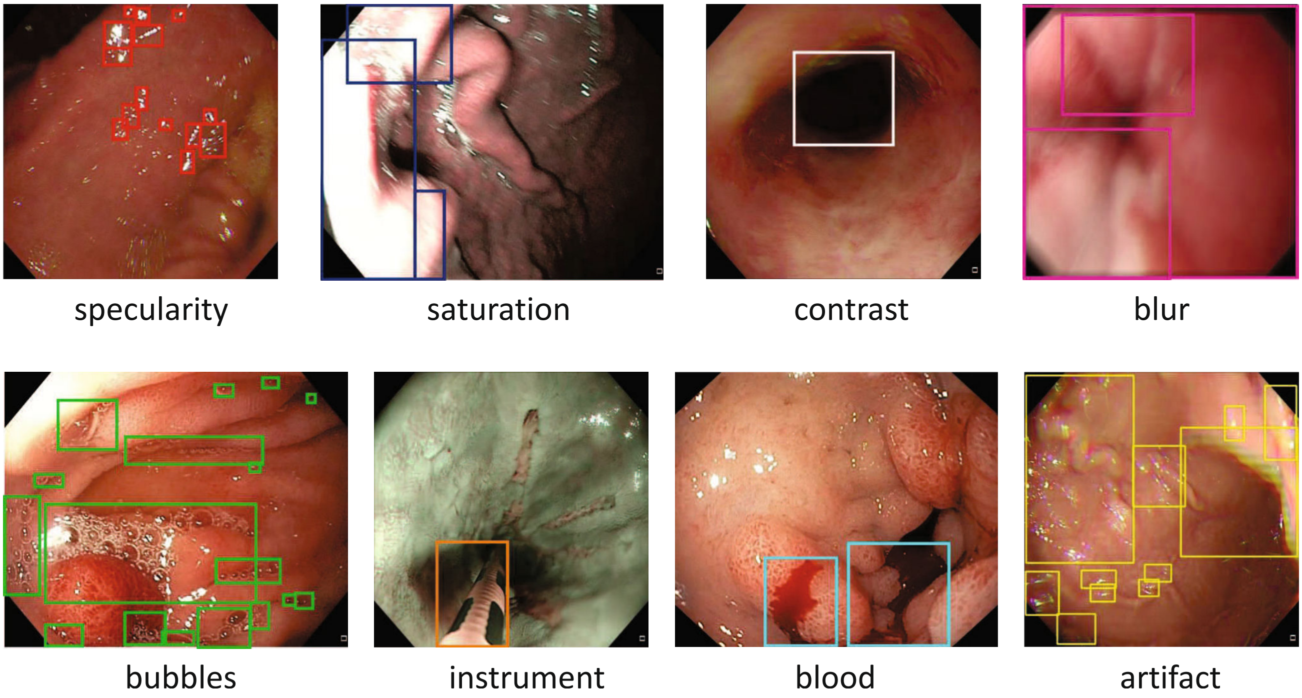


Fig. 2 Sample frames of the eight classes of artifacts in endoscopy. Specularity is caused by light reflection on surfaces. Saturation results from overexposure. Contrast arises from inadequate or excessive contrast. Blur is caused by movement or focusing problems during image capture. Bubbles are thin films of liquid containing air. Instrument means biopsy or any other instrument. Blood is an opaque red fluid. Artifact is a comprehensive class that covers various artifacts that cannot be grouped into a specific category mentioned above, such as chromatic aberration fragments

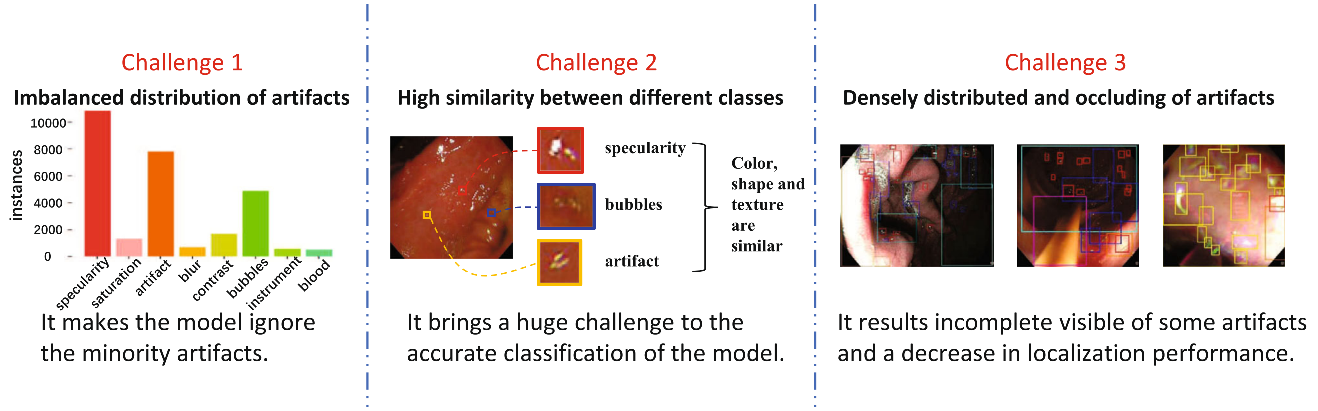


Fig. 3 Three challenges for multi-class endoscopic artifacts detection: (1) artifacts of different classes are imbalanced; (2) the characteristic of inter-class are similar while intra-class are varied in artifacts; (3) artifacts heavily overlap and occlude each other

accurate classification of the artifacts. (3) Artifacts heavily overlap and occlude each other [7]. The occlusion mainly refers to the whole or part of one artifact occluding another artifact, resulting in incomplete visual information of some artifacts. Occlusion frequently appears in endoscopic video frames, and they are numerous and densely distributed. This results in the model's susceptibility to both false positives and false negatives, and a decrease in localization performance.

To tackle these challenges, we propose an Occlusion-Aware Class Characteristic Mining Network (OCCMNet) to detect eight different classes of artifacts in endoscope simultaneously, mitigating artifact interference and enhancing the precision and efficiency of endoscopic examinations during clinical diagnosis. The OCCMNet includes three parts: (1) The Dual-Branch Class Rebalancing Module (DCRM) enables to correct the imbalance of sample amount in different artifact classes. It adopts two complementary sampling strategies, reserving more majority class and minority class artifacts respectively, which are then used to train two detection strategies for processing majority class and minority class respectively. This will increase the focus on minority class artifacts while keeping the majority class artifacts effect unaffected by the change of the overall data distribution. (2) The Class Discrimination Enhancement Module (CDEM) enables to enhance the discrepancy of inter-class. It groups features by calculating the information density of each channel, and the features in the group with high information density represent the key information of categorization, while the information in the group with low information density consists of some subtle differences in information, which then is introduced to further broaden inter-class distance. (3) The Global Occlusion-Aware Module (GOAM) enables to identify artifacts based on areas that are not obscured. It employs an attention-based multi-perspective fusion strategy, which captures the global context information to initially identify and locate the obscured artifacts and senses the overall structure of the artifacts by combining local details, thereby inferring the obscured part of the artifacts and making up for the information loss caused by occlusion.

In summary, the main contributions of this work are including the following:

- We propose a novel method for the precise detection of multi-class artifacts in endoscopy, aiming to reduce the disturbance caused by artifacts in clinical diagnosis and treatment.
- A novel data imbalance correction strategy is proposed to improve the effect of minority samples and maintain the effect of majority samples by utilizing two complementary data distributions.
- A novel class-wise discrimination enhancement approach is proposed to maximize inter-class difference, which effectively increases the complementarity and complexity of features.
- A novel occlusion-aware framework is proposed to mitigate the effect of occlusion by fusing multi-scale global and local features.

2 Related work

2.1 Existing methods for multi-class artifacts detection in endoscopy

Despite many specific networks [1, 7] have been designed for multi-class artifacts detection in endoscopes, they have not accounted for the negative impact of similarity and occlusion on artifacts detection. In previous studies, a series of models [8, 9] have been proposed for natural image detection. These models can be broadly classified into two categories: single-stage and two-stage. Although they achieved some success in the field of natural image detection, their main limitation comes from the inadequate consideration of the unique properties of endoscope images and artifacts. Consequently, certain models [10, 11] have integrated a feature pyramid network (FPN) [12] into their framework to better adapt to the diverse sizes of artifacts in endoscope images. Additionally, certain models [13, 14] integrated data augmentation and focal loss [15] to address the challenges arising from the class imbalance that exists in endoscope artifacts, improving the detection performance of minority-class artifacts and the model's generalization ability. Furthermore, certain models [16–19] leveraged ensemble learning and transfer learning, which combined the predictions of several basic models, enhancing the robustness and accuracy of the detection. Although the aforementioned methods have improved the performance of models to some extent in endoscope artifact detection tasks, they still fail to effectively address the challenges posed by similarity and occlusion in endoscope image detection.

2.2 Data imbalance in medical image analysis

The impact of class imbalance on multi-class detection is significant. However, most methods [20] are designed to achieve a relatively average data effect, which will inevitably affect the original effect of majority samples while improving the effect of the minority samples. Generally, there are three main solutions to the imbalance of image type data: resampling, data enhancement, and weight redistribution. Common resampling strategies [21] are divided into oversampling and undersampling. However, the oversampling for minority classes will easily lead to repeated sampling and overfitting of minority samples. Undersampling for majority classes is easy to cause the loss of discriminative samples. Data augmentation [22, 23] usually utilizes geometric changes, adding noise, color transformation, stitching, and other technologies to generate more data. However, this method mainly improves the effect by increasing the amount of data equally for all data and is not targeted. The redistribution of weights [24] mainly sets different weights for each class in the loss function, making the model pay more attention to the minority classes during training. Our method avoids the defect of a single resampling strategy, and the improvement of the detection effect of our method is not caused by the increase of data volume.

2.3 Context-aware in medical image detection

Context-aware utilizes surrounding information to improve the recognition and understanding capabilities of the targets. Although significant progress has been made in this work, these models are inclined to overlook certain details, potentially affecting the precision of the resulting analyses. The most basic is multi-scale fusion, such as [12, 25, 26], and they learn local features and context information in images through convolution and pooling and fuse feature maps of different scales to capture context information at different scales. The second type is based on global and local methods, such as [27]. They further consider the relationship between global and local context by establishing a connection between them, which helps the model to better understand the shape and structure of the object.

With the development of attention mechanism, more attention-mechanism-based methods began to appear, such as [27–30]. They generate weights through attention mechanisms or use auxiliary information to make the model focus on meaningful positions in the image, so as to better capture contextual information. Our method fully combines the advantages of these works to further explore the information relationship of feature fusion at different scales, while retaining more details and key features.

3 Methodology

The overview of OCCMNet is shown in Fig. 4. OCCMNet can accurately detect multiple classes of artifacts in endoscopic images through three different functional parts: (1) The Dual-Branch Class Rebalancing Module (The DCRM, Sect. 3.1) raises the awareness of minority and maintains the sensitivity towards majority by designing two distinct sampling and detection strategies for majority classes and minority classes respectively. (2) The Class Discrimination Enhancement Module (The CDEM, Sect. 3.2) enhances feature distinctiveness by assigning higher weights to critical feature and promoting the complementarity and complexity of features through a reconstruction mechanism. (3) The Global Occlusion-Aware Module (The GOAM, Sect. 3.3) is capable of mining both the spatial structural and the boundary information between artifacts by integrating features from multi-scales and embedding an attention-based global and local fusion strategy.

3.1 Dual-Branch Class Rebalancing Module

The Dual-Branch Class Rebalancing Module (DCRM) employs two complementary sampling and detection strategies to augment the effect of minority artifacts and maintain the effect of majority artifacts, as shown in Fig. 5. First, it utilizes two bias samplers to alter the input distribution for different classes, generating two distinct bias sampling strategies. These strategies tend to select more samples from the majority and minority classes respectively, thereby introducing different biases into the data distribution.

Then, it utilizes two bias detection heads, which are part of the CDEM behind GOAM, each trained to suit the characteristics of the two types of classes majority and minority. The network architecture of both Heads is the YOLOv5 detection

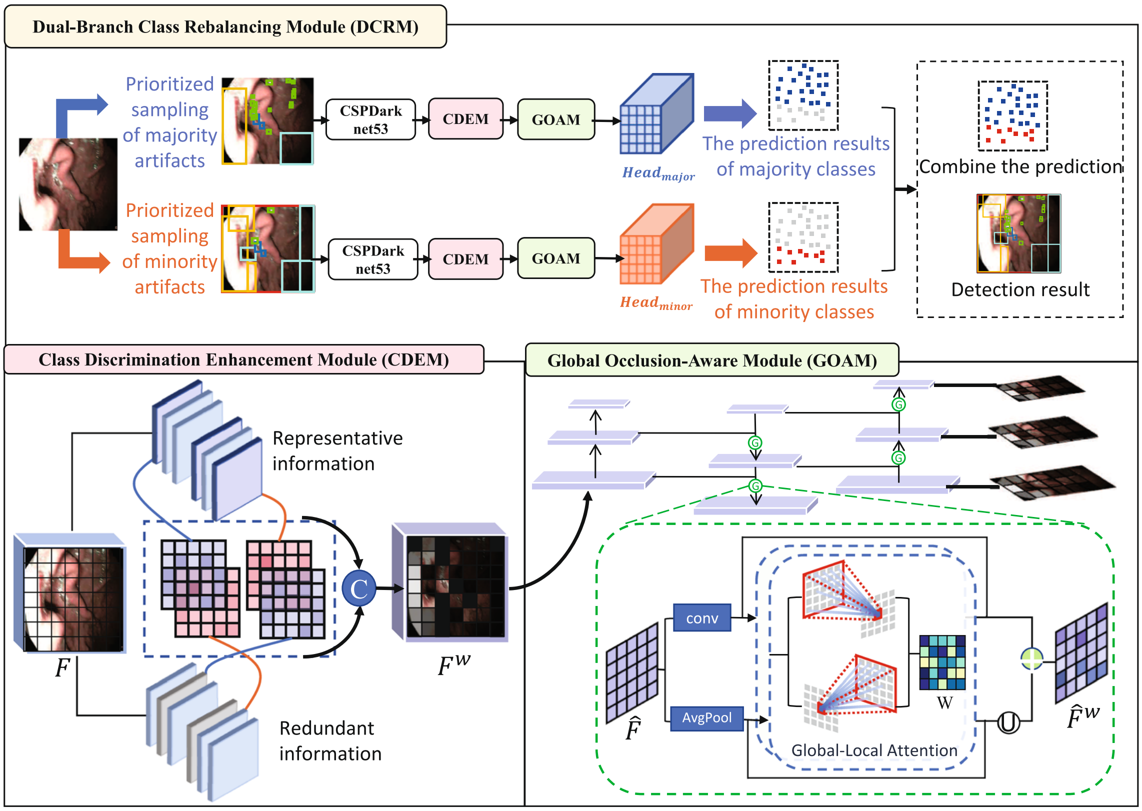

Dual-Branch Class Rebalancing Module (DCRM)
Prioritized sampling of majority artifacts
CSPDarknet53
CDEM
GOAM
$Head_{major}$
The prediction results of majority classes
Prioritized sampling of minority artifacts
CSPDarknet53
CDEM
GOAM
$Head_{minor}$
The prediction results of minority classes
Combine the prediction
Detection result
Class Discrimination Enhancement Module (CDEM)
Representative information
Redundant information
F
C
F^W
Global Occlusion-Aware Module (GOAM)
conv
AvgPool
$\hat{F}$
W
Global-Local Attention
$\hat{F}^W$

Fig. 4 Our proposed Occlusion-Aware Class Characteristic Mining Network (OCCMNet) consists of a Dual-Branch Class Rebalancing Module (DCRM), a Class Discrimination Enhancement Module (CDEM), and a Global Occlusion-Aware Module (GOAM)

head, ensuring a consistent framework for object classification and localization tasks. This enables the detection heads to learn features consistent with the altered input distribution. By integrating these two strategies, the module effectively mitigates the suppressive effects on minority class samples while preserving the original performance of the majority of samples.

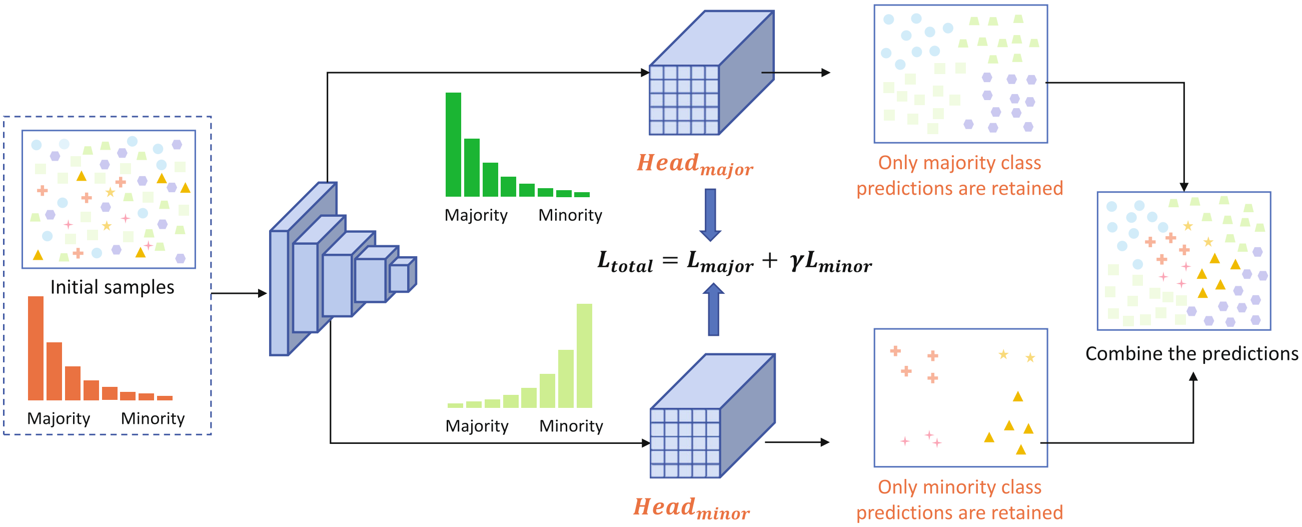

Initial samples
Majority
Minority
Majority
Minority
$Head_{major}$
Only majority class predictions are retained
$L_{total} = L_{major} + \gamma L_{minor}$
Majority
Minority
$Head_{minor}$
Only minority class predictions are retained
Combine the predictions

Fig. 5 The Dual-Branch Class Rebalancing Module (DCRM) eliminates the model's negligence of minority class bias by performing extensive sampling of the original majority and minority classes in front of their respective detection heads, and predicting the results separately using two types of bias detection heads

Algorithm 1 Class-biased samplers.

```
Input: Images and their corresponding labels
Output: Image sets of DCRM(major) and DCRM (minor)
▷ Count the occurrences of each class in each image;
Initialize [ImageID, C1_Num,...,C8_Num] = 0;
for int ImageID=0; ImageID<Image_Num; ImageID++ do
    Ci_Sum = 0;
    for int i=0; i<8; i++ do
        Ci_Sum += Ci_Num;
    end
end
▷ Define majority and minority based on the proportion
  threshold (e.g., 100/8);
Sum = 0;
for int i=0; i<8; i++ do
    Sum += Ci_Sum;
end
Initialize W = [w1,...,w8] = 0;
for int i=0; i<8; i++ do
    pi = Ci_Sum / Sum;
    if pi >(100/8) then
        wi = 1;
    end
end
▷ Calculate the proportion of majority and minority classes in
  per image;
▷ Image sets of DCRM(major) start sampling from the largest
  proportion of majority classes;
▷ Image sets of DCRM(minor) start sampling from the largest
  proportion of minority classes;
```

Specifically, two biased samplers resample at the image level. The first sampler DCRM (major) preferentially selects images with a large proportion of the majority classes, and the second sampler DCRM (minor) preferentially selects images with a large proportion of the minority classes. More details are shown in Alg. 1. The DCRM consists of two heads, Head (major) and Head (minor), which are used to process the majority-biased and minority-biased proposals, respectively. Besides, the DCRM computes loss for all classes during training. However, during inference, it predicted only the corresponding category (majority and minority), and the two prediction results were then fused to obtain the final result. Additionally, we can manually adjust the weight ratio between the losses of the two branches. The ultimate loss function for DCRM enables to be expressed as follows:

$$L_{total} = L_{major}(p_{major}, y) + \gamma L_{minor}(p_{minor}, y) \qquad (1)$$

Here, L_{major} and L_{minor} represent the focal loss functions of Head (major) and Head (minor), respectively. p_{major} and p_{minor} denote the predictions of Head (major) and Head (minor), encompassing box regression and class scores. y represents the labels for bounding boxes and classes, while γ is a balancing coefficient, which was set 2 in experiments.

3.2 Class Discrimination Enhancement Module

The Class Discrimination Enhancement Module (CDEM) employs a weighted grouping and cross-reconstructs strategy to capture the commonalities of intra-class and the distinctiveness of inter-class, as shown in Fig. 6. First, it utilizes a weighted grouping strategy to divide the extracted features into two distinct classes based on their informational density, one for representative features and the other for redundant features. Then, it cross-reconstructs the two sets of features

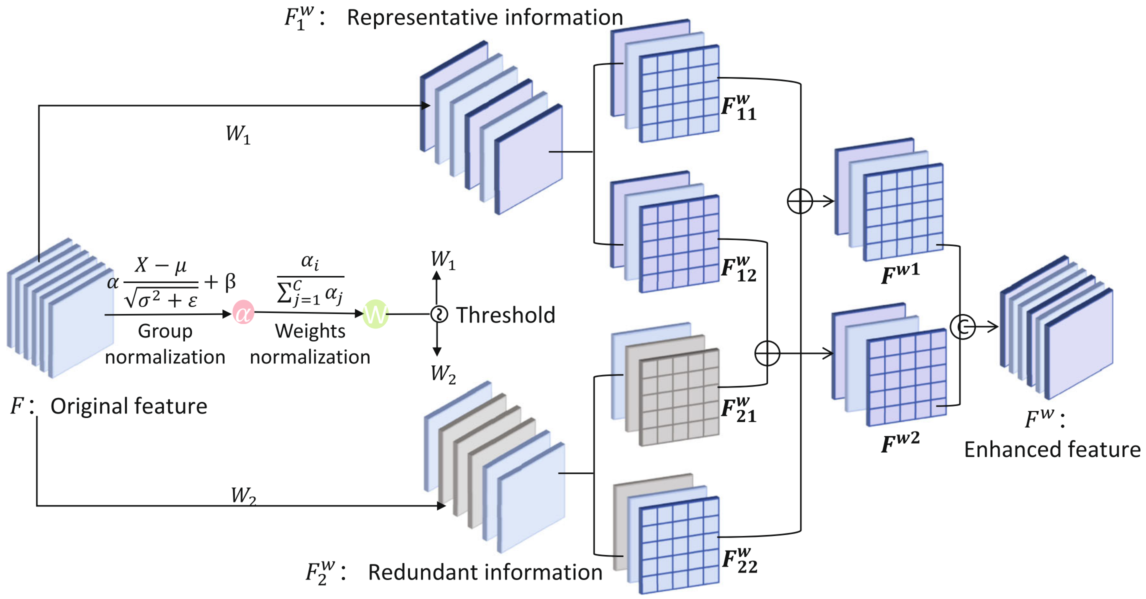


Fig. 6 The Class Discrimination Enhancement Module (CDEM) improves the discriminative and robust characteristics of fine-grained features by attention-induced weight-enhancing representative channels and dividing features into strong and weak groups for cross-reconstruction

to avoid the interference caused by too much redundant information while retaining more detailed information, thus increasing the complementarity and complexity of features.

Specifically, CDEM begins by performing group standardization on the input feature F to align the processed data with a standard normal distribution. F is obtained by feature extraction through the CSPDarknet53 network architecture after sampling. Subsequently, we assign a weight W_γ to the standardized features and employ a sigmoid function to map the re-weighted feature maps to the range (0,1). To apply gating to the weights, we introduce a threshold, setting weights above the threshold to 1 to obtain information weights W_1 and setting others to 0 to obtain non-information weights W_2 (with a threshold of 0.5 used in experiments). Finally, we multiply the input feature F separately by W_1 and W_2 to obtain two weighted features, F_1^w, which retains spatial content with information and expressiveness, and F_2^w, which contains little or no information and is considered redundant.

For the reconstruction part, instead of directly concatenating the two sets of features, we perform cross-reconstruction. F_1^w and F_2^w are divided into two groups sequentially along the channel dimension. Specifically, the channels of F_1^w are denoted as $[c_1, c_2, ..., c_k]$. These channels are then divided into two groups, resulting in F_{11}^w and F_{12}^w. Here, F_{11}^w contains the first half of the channels, $[c_1, c_2, ..., c_{\frac{k}{2}}]$, while F_{12}^w consists of the remaining channels, $[c_{\frac{k}{2}+1}, c_{\frac{k}{2}+2}, ..., c_k]$. Similarly, F_2^w is divided into F_{21}^w and F_{22}^w. Then, we add F_{11}^w to F_{22}^w and F_{12}^w to F_{21}^w. Finally, we concatenate the cross-reconstructed features F^{w1} and F^{w2} to obtain the spatially refined feature map F^w. The specific formulas are as follows:

$$\begin{cases} F_1^w = W_1 \otimes F, \\ F_2^w = W_2 \otimes F, \\ F_{11}^w \oplus F_{22}^w = F^{w1}, \\ F_{12}^w \oplus F_{21}^w = F^{w2}, \\ F^{w1} \cap F^{w2} = F^w. \end{cases} \tag{2}$$

$$W = Gate\left(Sigmoid\left(W_\gamma\left(\gamma\frac{F-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta\right)\right)\right) \tag{3}$$

Here, μ and σ represent the mean and standard deviation of F, respectively, and ε is a small positive constant added for division stability. The formula for calculating the normalization-related weight W_γ in the above equations is as follows:

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=i}^{C}\gamma_j}, i, j = 1, 2, ..., C \tag{4}$$

where γ is a trainable parameter with dimensions R^C, C is the number of channels. A larger value of γ indicates that the feature maps of the corresponding channel have richer information in terms of spatial pixel variance, reflecting greater variations across spatial pixels.

3.3 Global Occlusion-Aware Module

The Global Occlusion-aware Module (GOAM) employs an attention-based global–local fusion strategy to mine the spatial structural and boundary information between artifacts, as shown in Fig. 7.

First, it leverages the feature fusion strategy of PANet to integrate multi-scale features via both bottom-up and top-down pathways, thereby establishing an initial level of context awareness. Subsequently, at each stage of feature fusion, it employs a dual-branch architecture endowed with global–local self-awareness to autonomously perceive and establish the focus and connectivity of the features. One branch focuses on extracting local features through convolution, which is proficient in identifying and highlighting image details, including edges, textures, and other fine-grained characteristics. The other branch provides global contextual information through global average pooling, which is used to comprehend the comprehensive structural framework and spatial positioning of the occluded regions. Finally, the features of the two branches are fused, resulting in a composite representation that encapsulates both global contextual information and local detail relationships.

Specifically, the output $\hat{F} \in \mathbb{R}^{W\times H\times C}$ of each feature fusion gap is fed to two branches through a convolution (kernel size 3, stride 1, and padding 1) and an average pooling (stride 2) to obtain two sub-features $\hat{F}_a \in \mathbb{R}^{W\times H\times C}$ and $\hat{F}_b \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$ respectively. The sub-features are then input into a multi-scale-based global local attention to get two weights W_a and W_b. The attention uses the point-wise convolutions as the local channel context aggregator, which only exploits point-wise channel interactions for each spatial position. On the other hand, global average pooling is used as the global channel context aggregator. The attention-specific process can be described as follows:

$$\begin{cases} f_1(x) = BN(PwCon(ReLU(BN(PwCon(X))))), \\ f_2(x) = GAP(X_1), \\ f(x) = f_1(x) \oplus f_2(x). \end{cases} \tag{5}$$

$$\begin{cases} W_a = Sigmoid(f(\hat{F}^a)), \\ W_b = Sigmoid(f(\hat{F}^b)). \end{cases} \tag{6}$$

Here, X refers to the input feature of the attention mechanism. PwCon, BN, ReLU, and GAP refer to pointwise convolution, batch normalization, ReLU activation function, and global average pooling, respectively. $\oplus$ denotes the broadcasting addition.

After the attention mechanism, element-wise multiplication is adopted to fuse the weights (W_a and W_b) and the corresponding features ($\hat{F}_a$ and $\hat{F}_b$), and then we obtain

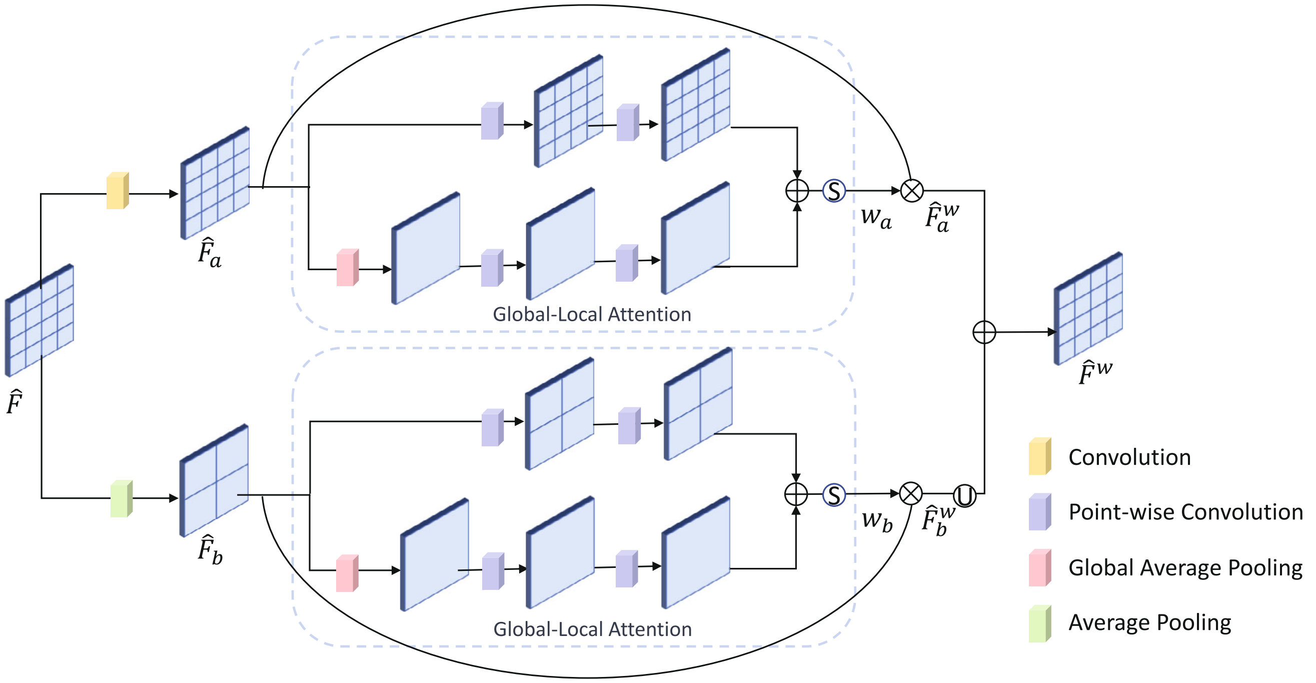


Fig. 7 The detailed display of the dotted green box in Fig. 4. The Global Occlusion-Aware Module (GOAM) integrates features from different hierarchical levels and utilizes a dual-branch fusion of ordinary convolutions and average pooling to enhance contextual information

$\hat{F}_a^W \in \mathbb{R}^{W\times H\times C}$ and $\hat{F}_b^W \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$. Finally, we directly fuse the features from the two branches using an addition operation to obtain $\hat{F}^W$. The above process can be described as follows:

$$\begin{cases} \hat{F}_a^W = \hat{F}_a \otimes W_a, \\ \hat{F}_b^W = \hat{F}_a \otimes W_b, \\ \hat{F}^w = \hat{F}_a^W \oplus Upsample(\hat{F}_b^W). \end{cases} \tag{7}$$

4 Experiment setup

4.1 Dataset

The EndoCV2020 [31] dataset is a comprehensive and extensive dataset designed for endoscopic artifacts detection. It encompasses eight common types of endoscopic artifacts, including specularity, saturation, artifact, blur, contrast, bubbles, instruments, and blood. The eight classes account for 38.80%, 4.50%, 27.34%, 2.44%, 5.98%, 17.25%, 1.99%, and 1.71%, respectively. The data are sourced from six different centers worldwide, namely the John Radcliffe Hospital in Oxford, UK, the ICL Cancer Research Institute in Nancy, France, the Ambroise Par'e Hospital in Boulogne-Billancourt, France, the Veneto Institute of Oncology in Padua, Italy, the Vaudois University Hospital in Lausanne, Switzerland, and the Botkin Clinical City Hospital in Moscow. This dataset covers multiple organs (gastroscopy, cystoscopy, gastroesophageal examination, and colonoscopy), multiple modes (white light, fluorescence, and narrow-band imaging), and diverse populations from the UK, France, Russia, and Switzerland.

4.2 Implementation details

This research dataset comprises a total of 2531 images, with 2278 images (90% of the total) allocated for the training set and 253 images (10%) for the test set. During the training phase, we employed data augmentation techniques to enhance the dataset, including random horizontal and vertical flips, rotations, scaling, and translations. These data augmentation techniques, along with their respective parameters, were adopted in accordance with YOLOv5 [32]. All images were cropped to a size of 640×640 as input. We set the batch size to 4, used the Adam optimizer with a momentum of 0.948, and applied a weight decay of 0.00005 for model optimization. Initial weights were set to 0.001. The training process consisted of 200 epochs, and our framework model was implemented using the PyTorch framework.

4.3 Experimental setup and evaluation metrics

To comprehensively demonstrate the advantages of OCCM-Net, we conducted experiments that first validated its performance in multi-class endoscopic artifacts detection, reaching the state-of-the-art performance. Therefore, the experiment compares our OCCMNet with seven recent state-of-the-art or typical target detection methods. These methods are divided into three types: (1) the single-stage object detection methods, including YOLOv5 [32] (2020) and YOLOv9 [33] (2024); (2) the typical two-stage object detection methods Faster RCNN [8] (2016) and Cascade RCNN [9] (2018); (3) the latest end-to-end transformer-based detector RT-DETR [34] (2023); (4) the latest multi-class endoscopic artifact detection methods, including Faster RFD3-CNN [11] (2022) and Sun [35] (2023). In addition, we conducted ablation studies to evaluate the performance of each primary novel component of OCCMNet, including the DCRM, the CDEM, and the GOAM. This allowed us to validate the individual superiority of each component. And we utilized five commonly recognized evaluation metrics for object detection in our experiments: mAP50, mAP50-95, precision, recall, and F1-score. These metrics provided us with a comprehensive assessment of our approach from various perspectives.

5 Experimental results

5.1 The accurate detection of multi-class artifacts in endoscopic images

The results from Table 1 demonstrated that our OCCMNet is able to accurately locate and classify multiple class artifacts from endoscopic images. The detection results (i.e., bounding boxes and classes of artifacts) of our OCCMNet are highly visually consistent with the label obtained by endoscopist. In addition, Table 1 presents the high performance of our OCCMNet in various detection metrics, and our OCCMNet achieves a mAP50 of 58.1%, a mAP50-90 of 33.3%, a precise of 60.9%, a recall of 56.4%, and a F1-score of 58.6%.

5.2 Outperformance of our OCCMNet than all comparative methods

Table 1 demonstrates that the evaluation of visualization and all five detection metrics demonstrate that our OCCMNet surpasses all seven comparative methods. Specifically, when compared to YOLOv5 [32], YOLOv9 [33], Faster RCNN [8], Cascade RCNN [9], RT-DETR [34], Faster RFD3-CNN [11], and Sun [35], all of which were trained using default network structures and parameters listed in their references, OCCMNet achieved the highest detection performance, as shown in Fig. 8. In Table 1, comparing seven comparative methods by various detection metrics, our OCCMNet improved the mAP50 by 3.3–6.5%, the mAP50-95 by 0.7–4.6%, the precise by 0.6–8.3%, the recall by 1.9–8.2%, and the F1-score by 2.3–6.5%. Table 2 demonstrates that OCCMNet achieves superior mAP50 performance across most artifact classes in comparison to other methods. Notably, OCCMNet shows significant improvements in detecting blood artifacts, with an mAP50 of 48.3%, which is a 5.6% increase compared to the second-best result.

5.3 Superiority of the dual-branch class rebalancing module

Table 3 and Fig. 9 depict that the usage of DCRM effectively contributes to change the distribution of data sampling during the training phase, thus solving the problem of the neglect of minority class artifacts caused by data imbalances. We created two versions of the model: (1) without DCRM version (wDCRM), which lacks the distinction treatment of majority and minority class samples, only uses ordinary random sampling to determine the input distribution and training indiscriminateness of the model; (2) the resampling version (wDCRM_R), which makes a distinction between the majority and minority classes, but only samples according to the reciprocal of the number of each class and does not discriminate majority and minority classes during the training phase. Table 3 shows that our DCRM performs well in all five aspects of detection performance. Specifically, DCRM improved in mAP50 by 5.5–7.4%, mAP50-90 by 5.8–6.2%, precise by 1.5–3.2%, recall by 3.0–4.4%, and F1-score by 2.4–3.9%. These results validate the effectiveness of DCRM, which changes the input distribution of different classes and trains two types of detection heads for the majority and

Table 1 The quantitative evaluation of multi-class artifacts detection

Method	mAP50 ↑	mAP50-95 ↑	Precision ↑	Recall ↑	F1-score ↑
YOLOv5 [32]	51.6%	28.7%	58.3%	54.5%	56.3%
YOLOv9 [33]	54.8%	32.0%	56.3%	55.6%	55.9%
Faster RCNN [8]	52.0%	29.4%	52.6%	53.2%	52.8%
Cascade RCNN [9]	53.3%	31.7%	56.7%	48.2%	52.1%
RT-DETR [34]	54.7%	32.6%	60.3%	50.4%	54.9%
Faster RFD3-CNN [11]	53.4%	31.0%	52.8%	52.4%	52.5%
Sun [35]	54.3%	29.6%	60.0%	52.1%	55.7%
OCCMNet (our)	**58.1%**	**33.3%**	**60.9%**	**56.4%**	**58.6%**

The results show that our proposed OCCMNet is superior to seven methods for multi-class artifacts detection
The bold entries indicate the best data values

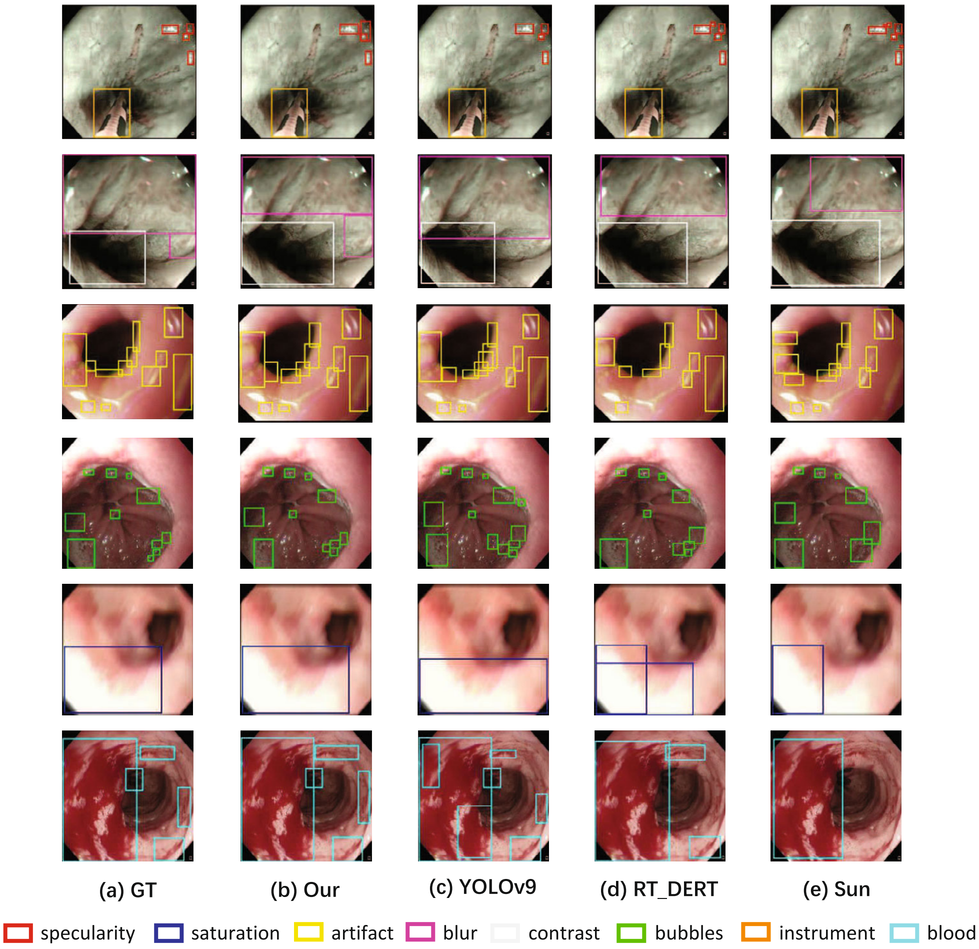


Fig. 8 Visual comparisons of the multi-class artifacts detection among different methods. **a** GT represents the ground truth; **b** represents the detection results of our method; **c**–**e** represent detection results from different methods. The first line shows the results of the specularity and instrument classes. The second line shows the results of the blur and contrast classes. The third line shows the results of the artifact class. The fourth line shows the results for the bubbles class. The fifth line shows the result of the saturation class. The sixth line shows the results of the blood class

minority class artifacts, respectively. As a result, it not only increases the attention of the minority class artifacts, but also keeps the decision of the majority artifacts unaffected to solve the problem of data imbalance and achieve accurate detection of multi-class artifacts in endoscopic images.

5.4 Superiority of the Class Discrimination Enhancement Module (CDEM)

Table 3 and Fig. 9 depict that the usage of CDEM effectively contributes to obtain more representative and robust

Table 2 The map50 of all classes in artifact detection

Method	Spec.	Sat.	Arti.	Blur	Cont.	Bubb.	Inst.	Blood
YOLOv5 [32]	38.8%	61.6%	46.9%	33.4%	70.7%	26.5%	92.4%	42.7%
YOLOv9 [33]	**42.1%**	69.8%	52.2%	41.6%	82.0%	28.4%	92.5%	30.1%
Faster RCNN [8]	38.7%	57.2%	51.4%	29.3%	77.1%	29.9%	90.7%	41.9%
Cascade RCNN [9]	36.9%	60.1%	50.4%	33.0%	78.9%	33.0%	**93.3%**	41.3%
RT-DETR [34]	41.6%	63.3%	51.3%	42.6%	79.2%	33.1%	90.2%	36.8%
Faster RFD3-CNN [11]	42.0%	61.9%	52.9%	32.8%	78.6%	30.8%	92.8%	35.1%
Sun [35]	40.8%	64.7%	51.5%	35.9%	78.2%	32.6%	91.6%	40.2%
OCCMNet (our)	42.0%	**70.4%**	**53.5%**	**42.7%**	**82.2%**	**33.2%**	92.6%	**48.3%**

The results show that our proposed OCCMNet achieves superior mAP50 performance overall compared to other methods. “Spec.,” “Sat.,” “Arti.,” “Cont.,” “Bubb.,” and “Inst.” are abbreviations for specularity, saturation, artifact, contrast, bubbles, and instrument, respectively

The bold entries indicate the best data values

Table 3 The quantitative evaluation of ablation studies

Method	mAP50 ↑	mAP50-95 ↑	Precision ↑	Recall ↑	F1-score ↑
wDCRM	52.6%	27.5%	57.7%	52.0%	54.7%
wDCRM_R	50.7%	27.1%	59.4%	53.4%	56.2%
wCDEM	53.1%	28.5%	60.5%	55.2%	57.7%
wCDEM_G	54.3%	28.5%	60.6%	53.1%	56.6%
wGOAM_P	53.0%	29.7%	58.7%	55.3%	56.9%
wGOAM_A	54.2%	27.6%	60.2%	56.0%	58.0%
OCCMNet	**58.1%**	**33.3%**	**60.9%**	**56.4%**	**58.6%**

These studies show that each module of the newly designed OCCMNet plays a role in multi-class artifacts detection

The bold entries indicate the best data values

fine-grained features, thus solving the problem of small distinction between different classes and big distinction within the same class. We created two versions of the model. We created two versions of the model:(1) without CDEM version (wCDEM), which lacks weighted grouping and cross-reconstruction two steps in the process of feature extraction, use the default feature extraction network CSPDarknet53 of YOLOv5 [32] directly; (2) the common attention module version (wCDEM_A). Here, we use GAMAttention [36], a model that combines channel and spatial attention mechanisms, which can also pay more attention to valuable information during feature extraction. Table 3 shows that our CDEM performs well in all five aspects of detection performance. Specifically, CDEM improved in mAP50 by 3.8–5.0%, mAP50-90 by 4.8–4.8%, precise by 0.3–0.4%, recall by 1.2–3.3%, and F1-score by 0.9–2.0%. These results validate the effectiveness of CDEM, which utilizes weighted grouping and cross-reconstruction to obtain more representative and robust fine-grained features, thereby increasing the gap between different classes and narrowing

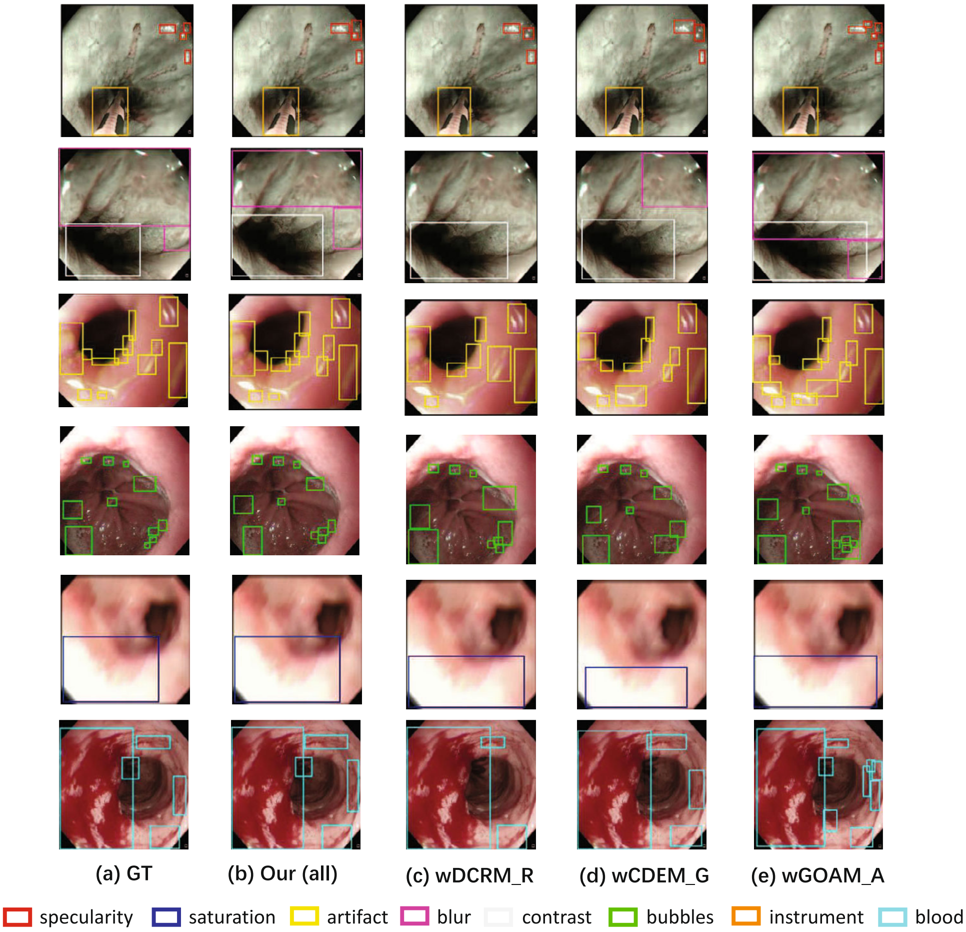


Fig. 9 Visualization results of the ablation experiment. **a** GT represents the ground truth; **b** represents the full model; c–**e** represent the effect of eliminating different components

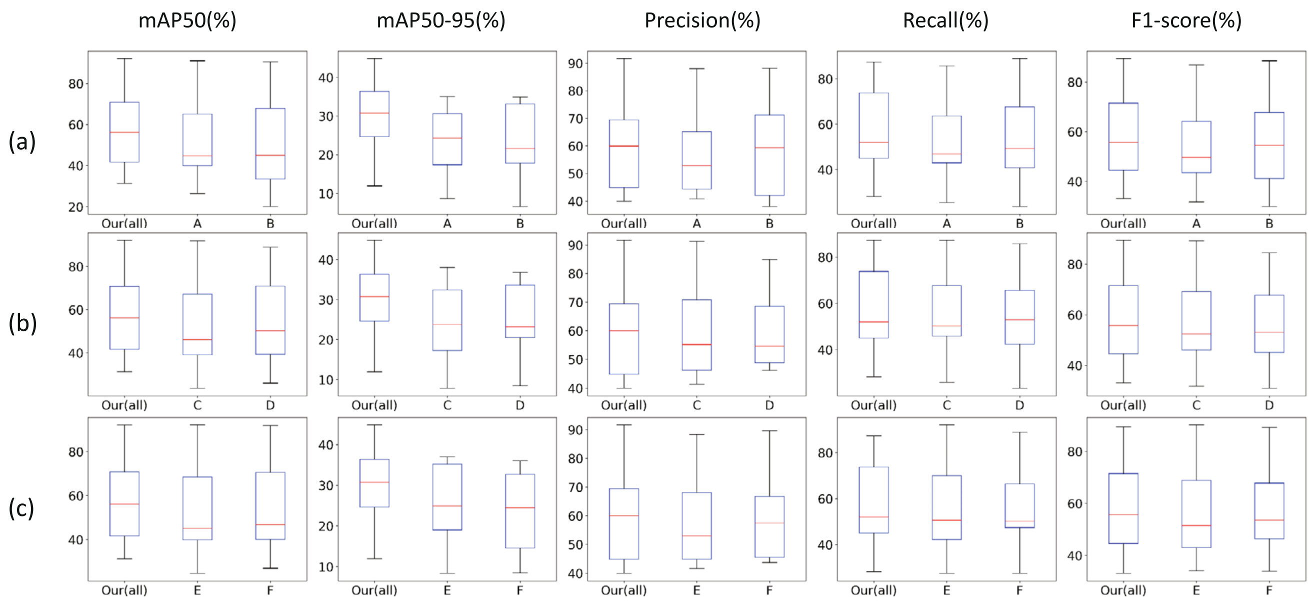


Fig. 10 **a**, **c**, and **c** respectively represent the ablation of the three innovative modules proposed by our OCCMNet, as demonstrated by metrics such as mAP50. A, B, C, D, E, and F represent "wDCRM," "wDCRM_R," "wCDEM," "wCDEM_G," "wGOAM_P," and "wGOAM_A," respectively

the gap within the same class. As a result, it not only helps distinguish the similar classes, but also enhances the representative feature to achieve accurate detection of multi-class artifacts in endoscopic images.

5.5 Superiority of the Global Occlusion-Aware Module

Table 3 and Fig. 9 depict that the usage of GOAM effectively contributes to enhance the semantic understanding of objects, thus solving the problem of occlusion between different artifacts. We created two versions of the model: (1) without GOAM version(wGOAM_P), which uses the typical path aggregation network (PANet), but only combines features of different scales; (2) the AFPN version(wGOAM_A), which replaces PANet with Asymptotic Feature Pyramid Network (AFPN), a newer multi-scale fusion network. Table 3 shows that our GOAM performs well in all five aspects of detection performance. Specifically, GOAM improved in mAP50 by 3.9–5.1%, mAP50-90 by 3.6–5.7%, precise by 0.7–2.2%, recall by 0.4–1.1%, and F1-score by 0.6–1.7%. These results validate the effectiveness of GOAM, which utilizes dual-branch context enhancement to understanding the connection between global and local features, thereby mining the

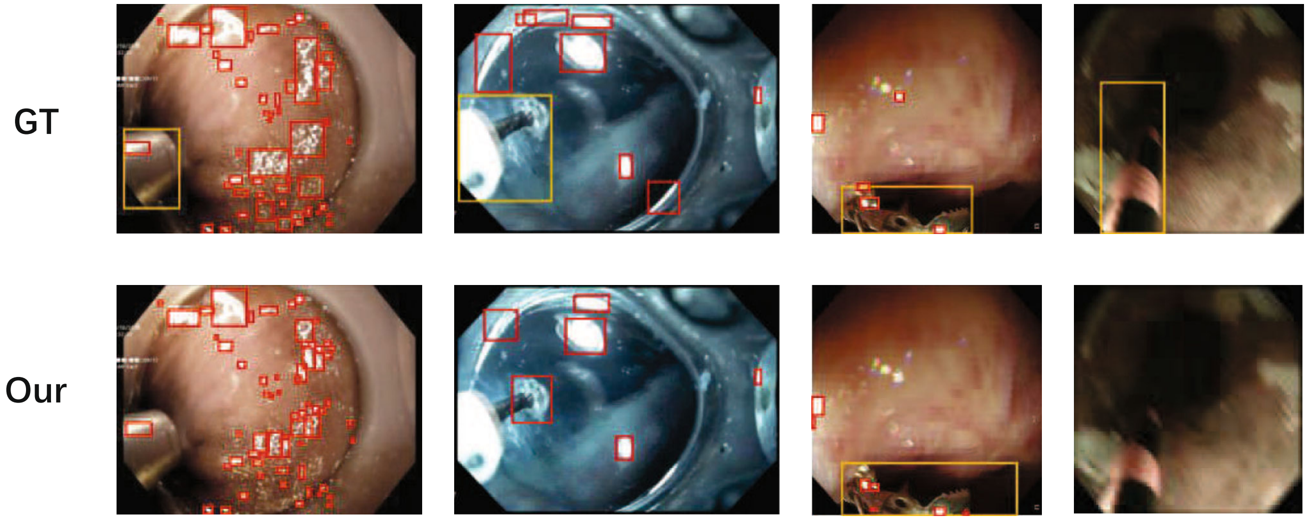


Fig. 11 The poorly performing images in specularity and instruments

Table 4 The map50 of all classes in artifacts detection

Method	Spec.	Sat.	Arti.	Blur	Cont.	Bubb.	Inst.	Blood
wDCRM	40.1%	61.9%	48.3%	34.2%	71.8%	29.3%	91.9%	43.1%
wCDEM	38.8%	64.5%	47.2%	36.7%	74.9%	26.7%	92.1%	44.2%
wGOAM_P	39.1%	63.7%	47.0%	36.8%	73.8%	27.9%	92.2%	43.6%
OCCMNet (our)	**42.0%**	**70.4%**	**53.5%**	**42.7%**	**82.2%**	**33.2%**	**92.6%**	**48.3%**

These studies show that each module of our OCCMNet plays a role in each class detection
The bold entries indicate the best data values

contextual relationship between occlusion and occluded artifacts to assist in their localization. As a result, it not only helps eliminate the effects of occlusion, but also enhances contextual information to achieve accurate detection of multi-class artifacts in endoscopic images.

6 Discussion

In this paper, we introduce the Occlusion-Aware Class Characteristic Mining Network (OCCMNet), a novel approach for detecting eight different classes of artifacts in endoscope images simultaneously. Compared with previous work, our approach further focuses on the influence of similarity and occlusion in the endoscopic artifact detection. To evaluate the performance of OCCMNet, we conducted experiments on a publicly available EndoCV2020 [31] dataset with seven state-of-the-art methods using five well-established metrics. OCCMNet outperformed the others, improving the mAP50 metric by 3.3–6.5%. Notably, our model achieved the best performance across six artifact classes, with a particularly significant improvement in the "blood" class, where the mAP50 increased by at least 5.6%. Additionally, the main modules of OCCMNet were analyzed in the experiment to quantify their contributions to the detection accuracy, as shown in Fig. 10. Horizontally, the overall model consistently outperforms the ablated versions, and vertically, it maintains robust performance across all five key metrics, particularly excelling in mAP50 and mAP50-95. The research results demonstrate that OCCMNet achieves precise detection of multi-class endoscopic artifacts by integrating these three modules.

Despite its strengths, OCCMNet can still be improved. The performance of OCCMNet on the two classes (specularity and instruments) did not achieve the best among all the compared methods. The poorly performing images are shown in Fig. 11. However, we conducted ablation experiments on each class to further elucidate the impact of our introduced modules, as shown in Table 4. The results indicate that all three modules have a positive influence across all classes. The suboptimal performance of the two classes among all compared methods can likely be attributed to the increased complexity of artifacts in dynamic environments. This complexity is exacerbated by variations in ambient lighting, as well as the low contrast of gastrointestinal tracts and other anatomical structures. Therefore, our future research may explore the integration of temporal information and the application of transfer learning techniques to enhance detection capabilities under a variety of conditions.

7 Conclusion

In this study, we proposed a novel approach to effectively detecting eight different classes of artifacts in endoscope simultaneously. First, it changes the input distribution of different classes and trains two types of detection heads suitable for the majority and minority classes respectively, thereby eliminating the negative effects of data imbalance. In addition, it extracts more representative and robust fine-grained features to recognize the commonness of intra-class artifacts and the difference of inter-class artifacts. Finally, it mines the contextual relationship between occlusion and occluded artifacts to eliminate the effects of occlusion. The experimental results on the EndoCV2020 [31] datasets demonstrate the superiority of the proposed method over existing state-of-the-art approaches. It indicates that our OCCMNet enables to be used as an effective and accurate clinical tool to detect multi-class endoscopic artifacts, so as to eliminate the interference caused by artifacts in clinical diagnosis and treatment and alleviate the fatigue of endoscopists.

Funding This work was supported in part by the National Natural Science Foundation of China (62476001), the University Synergy Innovation Program of Anhui Province under Grant (GXXT- 2021-007), and the Anhui Provincial Natural Science Foundation (2208085Y19), the Anhui Province Excellent Scientific Research and Innovation Team (2024AH010004).

Data availability A dataset of the first subtask, Endoscopy Artefact Detection and Segmentation (https://ead2020.grand-challenge.org/), used in the Endoscopic Computer Vision Challenge and Workshop (https://endocv.grand-challenge.org/).

Declarations

Conflict of interest The authors declare no competing interests.

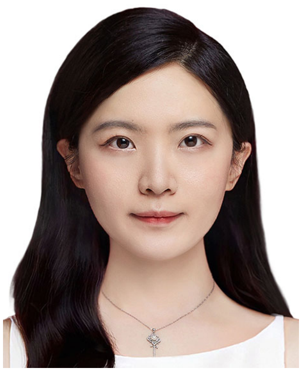
References

1. Ali S, Dmitrieva M, Ghatwary N, Bano S, Polat G, Temizel A, Krenzer A, Hekalo A, Guo YB, Matuszewski B et al (2021) Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med Image Anal 70:102002
2. Ali S, Zhou F, Daul C, Braden B, Bailey A, Realdon S, East J, Wagnières G, Loschenov V, Grisan E, Blondel W, Rittscher J (2019) Endoscopy artifact detection (EAD 2019) challenge dataset. ArXiv https://doi.org/10.17632/c7fjbxcgj9.1
3. Polat G, Sen D, Inci A, Temizel A (2020) Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination. In: EndoCV@ ISBI, pp 8–12
4. Chavarrias-Solano PE, Ali-Teevno M, Ochoa-Ruiz G, Ali S (2022) Improving artifact detection in endoscopic video frames using deep learning techniques. In: Mexican international conference on artificial intelligence, Springer, pp 327–338
5. Zhang C, Zhang N, Wang D, Cao Y, Liu B (2020) Artifact detection in endoscopic video with deep convolutional neural networks. In: 2020 2d international conference on transdisciplinary AI (TransAI), IEEE, pp. 1–8
6. Slavescu RR, Sporis IC, Gombos K, Slavescu KC (2022) Exploring two deep learning based solutions for improving endoscopy artifact detection. In: 7th international conference on advancements of medicine and health care through technology: Proceedings of MEDITECH-2020, 13-15 October 2020, Springer, pp 102–112
7. Ali S, Zhou F, Braden B, Bailey A, Yang S, Cheng G, Zhang P, Li X, Kayser M, Soberanis-Mukul RD et al (2020) An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci Rep 10(1):2748
8. Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
9. Cai Z, Vasconcelos N (2018) Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6154–6162
10. Zhang YY, Xie D (2019) Detection and segmentation of multi-class artifacts in endoscopy. J Zhejiang Univ-Sci B 20(12):1014–1020
11. Zhang H, Adjei P, Hu X, Wang X, Hu Y, Gan TRao N (2022) A study on endoscopy artefact detection based on deep learning
12. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
13. Kirthika N, Sargunam B (2021) YOLOV4 for multi-class artefact detection in endoscopic images. In: 2021 3rd international conference on signal processing and communication (ICPSC), IEEE, pp 73–77
14. Xu Z, Ali S, Gupta S, Celik N, Rittscher J (2021) Improved artifact detection in endoscopy imaging through profile pruning. In: Medical image understanding and analysis: 25th annual conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25, Springer, pp 87–97
15. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
16. Artunc F, Oksuz I (2021) An ensemble approach for automatic artefact detection on gastroendoscopy images. In: 2021 6th international conference on computer science and engineering (UBMK), IEEE, pp 741–746
17. Yin TK, Huang KL, Chiu SR, Yang YQ, Chang BR (2022) Endoscopy artefact detection by deep transfer learning of baseline models. J Digit Imaging 35(5):1101–1110
18. Natarajan K, Balusamy S (2022) Transfer learning based deep neural network for detecting artefacts in endoscopic images. Int J Electr Comput Eng Syst 13(8):633–641
19. Zhang D, Xu C, Li S (2023) Heuristic multi-modal integration framework for liver tumor detection from multi-modal non-enhanced MRIS. Expert Syst Appl 221
20. Wang YC, Cheng CH (2021) A multiple combined method for rebalancing medical data with class imbalances. Comput Biol Med 134
21. Bae SH, Yoon KJ (2015) Polyp detection via imbalanced learning and discriminative feature learning. IEEE Trans Med Imaging 34(11):2379–2393
22. Afzal S, Maqsood M, Nazir F, Khan U, Aadil F, Awan KM, Mehmood I, Song O-Y (2019) A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. IEEE access. 7:115528–115539
23. Escobar Díaz Guerrero R, Carvalho L, Bocklitz T, Popp J, Oliveira JL (2024) A data augmentation methodology to reduce the class imbalance in histopathology images. J Imaging Inf Med 1–16
24. Gessert N, Sentker T, Madesta F, Schmitz R, Kniep H, Baltruschat I, Werner R, Schlaefer A (2019) Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. IEEE Trans Biomed Eng 67(2):495–503
25. Tan M, Pang R, Le QV (2020) EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790
26. Gao Z, Xu C, Zhang H, Li S, Albuquerque VHC (2020) Trustful internet of surveillance things based on deeply represented visual co-saliency detection. IEEE Internet Things J 7(5):4092–4100
27. Song P, Yang Z, Li J, Fan H (2023) DPCTN: dual path context-aware transformer network for medical image segmentation. Eng Appl Artif Intell 124
28. Alam MS, Wang D, Liao Q, Sowmya A (2022) A multi-scale context aware attention model for medical image segmentation. IEEE J Biomed Health Inf
29. Zhang W, Lu F, Zhao W, Hu Y, Su H, Yuan M (2023) ACCPG-Net: a skin lesion segmentation network with adaptive channel-context-aware pyramid attention and global feature fusion. Comput Biol Med 154
30. Xu C, Gao Z, Zhang H, Li S, Albuquerque VHC (2021) Video salient object detection using dual-stream spatiotemporal attention. Appl Soft Comput 108:107433
31. Ali S, Dmitrieva M, Zhou F, Daul C, Braden B, Bailey A, East J, Realdon S, Georges W, Loshchenov M et al (2021) Endoscopy artefact detection (EAD) dataset (includes updated 2020 version). Mendeley Data 3
32. Jocher G (2021) ultralytics/yolov5: V6.0 - YOLOv5n 'Nano' Models, Roboflow Integration, TensorFlow Export, OpenCV DNN Support. https://doi.org/10.5281/zenodo.5563715
33. Wang CY, Yeh IH, Liao HYM (2024) YOLOV9: learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616
34. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J (2023) DETRS beat YOLOS on real-time object detection. arXiv preprint arXiv:2304.08069
35. Sun W, Li P, Liang Y, Feng Y, Zhao L (2023) Detection of image artifacts using improved cascade region-based CNN for quality assessment of endoscopic images. Bioengineering 10(11):1288
36. Liu Y, Shao Z, Hoffmann N (2021) Global attention mechanism: retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chenchu Xu received the MSc degree in computer science from the Hefei University of Technology, Hefei, China, in 2013, and the PhD degree in computer science from Anhui University, in 2017. He is currently an associate professor with the School of Computer Science and Technology, Anhui University, Hefei, China. His current research interests include machine learning and medical image processing.

Yu Chen is with the School of Computer Science and Technology, Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, China, and the Institute of Artificial Intelligence.



Jie Liu is a doctor in the Imaging Center of Anhui Provincial Hospital, and is pursuing PhD in Electrical Engineering at University of Science and Technology of China.

Boyan Wang received the M.S. and Ph.D. degrees in computer science from the Hefei University of Technology, Hefei, China. His current research interests include data mining and machine learning.

Yanping Zhang is a full professor with the School of Computer Science and Technology, Anhui University, China. Her main research interests include Computational intelligence and quotient space theory, machine learning methods and applications, artificial neural network and intelligent information processing.

Jie Chen is a full professor with the School of Computer Science and Technology, Anhui University, China. Her research interests include intelligent computing and knowledge engineering, text sentiment analysis and machine learning.

Shu Zhao is a full professor with the School of Computer Science and Technology, Anhui University, China. Her research interests include network representation learning, knowledge graph, and social network analysis.

Authors and Affiliations

Chenchu Xu[1] · Yu Chen[1] · Jie Liu[2] · Boyan Wang[3] · Yanping Zhang[1] · Jie Chen[1] · Shu Zhao[1]

✉ Jie Liu
528579064@qq.com

Chenchu Xu
cxu332@gmail.com

Yu Chen
1925252840@qq.com

Boyan Wang
boyanwang@nju.ed

Yanping Zhang
zhangyp2@gmail.com

Jie Chen
chenjie200398@163.com

Shu Zhao
zhaoshuzs2002@hotmail.com

[1] Department of Computer Science and Technology,
Anhui University, 111 Jiulong Road, Hefei 230601,
Anhui, China

[2] Anhui Provincial Hospital, 17 Lujiang Road, Hefei 230001,
Anhui, China

[3] School of Intelligence Science and Technology,
Nanjing University, 1520 Taihu Road, Huqiu District, Suzhou
215163, Jiangsu, China