Earf-YOLO:A Model for Recognizing Zhuang Minority Patterns Based on YOLOv3

XIN WANG

Guilin University of Electronic Technology

Jingke Yan (✉ 592499985@qq.com)

Guilin University of Technology <https://orcid.org/0000-0002-8007-4899>

QIN WANG

Guilin University of Electronic Technology

QIN QIN

Guilin University of Electronic Technology

JUN WANG

Guilin University of Electronic Technology

ZHUO SHI

Guilin University of Electronic Technology

HUANGHE LI

Guilin University of Electronic Technology

Research

Keywords: Object detection, YOLOv3 model, Earf-YOLO model, CBEAM, SRFB, Soft-NMS

Posted Date: July 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-713796/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Earf-YOLO:A Model for Recognizing Zhuang Minority Patterns Based on YOLOv3

XIN WANG^{a,b}, JINGKE YAN^{b,*}, QIN WANG^c, QIN QIN^b, JUN WANG^b, ZHUO SHI^a, HUANGHE LI^b

^aSchool of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China;
20172301005@mials.guet.edu.com(XIN WANG); 928311457@qq.com(ZHUO SHI)

^bSchool of Marine Engineering, Guilin University of Electronic Technology, Beihai 536000, China ;20172301005@mials.guet.edu.com(XIN WANG); 592499985@qq.com(JINGKE YAN);2694478971@qq.com(QIN QIN);766097024@qq.com(JUN WANG);Lihuanghe@mial.guet.edu.com(HUANGHE LI)

^cBasic Teaching Department, Guilin University of Electronic Technology, Beihai 536000, China;wangqin@mails.edu.guet.qq.com(QIN WANG)

Abstract

With reference to the limitations of YOLOv3 in recognizing symbols on Zhuang pattern, such as slow detection speed, unable to detect small object, and inaccurate positioning of bounding box, we propose a new model: Earf-YOLO (Efficient Attention Receptive Field You only look once) in this paper. In EarF-YOLO, we present an attention module: CBEAM (Convolution Block Efficient Attention Module) at first, which provides feature maps from channel and spatial dimensions. In CBEAM module, a local cross-channel interaction strategy without reducing dimensionality is used to improve the performance of the convolutional neural network. Besides, we put forward the SRFB (Strength Receptive Field Block) structure. During its training, more branch structures will be generated to enrich the feature space of the convolutional block. During its prediction, the multi-branched structures will be reparametrized and fused into one main branch to improve the performance of the model. Finally, we adopt some advanced training techniques to improve the detection performance. Experiments on the dataset of Zhuang patterns and the COCO dataset show that the Earf-YOLO model can effectively reduce the error of the prediction box and the ground-truth box, and decrease the calculation time. The mAP value of this model on the dataset of Zhuang patterns and on the COCO dataset reaches 82.1 (IoU=0.5) and 62.14 (IoU=0.5) respectively.

Keywords: Object detection, YOLOv3 model, Earf-YOLO model, CBEAM, SRFB, Soft-NMS

1. Introduction

Ethnic minorities usually integrate their culture into their costumes and architecture patterns, often with special and profound connotations [1]Recognizing symbols on minority patterns efficiently is significant for researchers to facilitate relative studies and to spread minority culture.

With the evolvement of deep learning, deep convolutional neural network has been utilized in object detection field. In specific object detections, there are two commonly-used object recognition models. One is the two-stage object recognition model, represented by Fast Region-based Convolutional Network(Fast R-CNN) [2]Faster Region Convolutional Neural Network(Faster R-CNN)[3]and Mask Region Convolutional Neural Network (Mask R-CNN) [4], which divides the recognizing process into two stages. In the first stage, use a region proposal networks to roughly determine the area of objects to be detected, thus saving time for subsequent classification and regression. In the second stage, categorize and fine-tune the generated detection proposals. Although this method obtains good result, two-stage object detection model increases the depth of network and computational cost of the model due to the use of region proposal networks, whose reasoning speed is slow. The other model is the one-stage object recognition model, represented by the You only look once (YOLO) series [5] and the Single Shot multi-box detector (SSD)[6]. This

*Corresponding author

Email address: 592499985@qq.com (JINGKE YAN)

model skips detecting generating detection proposals and adopts an end-to-end approach for training and recognition. Although the YOLOv3 and the SSD illustrate good real-time processing capability on multiple Graphic Processing Units (GPUs), they lose accuracy and cannot be run on a single GPU in real time. In the recognition stage, they need many GPUs to ensure its real-time performance. In the training stage, they also need many GPUs or expensive GPUs to increase the number of the batch size, so as to better fit the data of the model.

With regards to the above-mentioned problems, it is necessary to build a fast and efficient model for object detection. A sensible solution is not to increase the depth of the neural network but to increase the feature extraction ability of lightweight networks by introducing some new structures and modules into the models.

For recognizing symbols on the Zhuang patterns, we suggest the efficient attention receptive field Earf-YOLO model based on YOLOv3 in this paper. Earf-YOLO evolves YOLOv3 into a quick, accurate, and easily-trained model.

The main contributions of our work are summarized as follows:

(1) We propose a new attention module CBEAM, which can provide feature maps from channel and spatial dimensions. It contains a few parameters but can improve recognition performance.

(2) We suggest the SRFB structure based on the RFB structure [7] to replace the redundant convolutional layer in the feature pyramid of YOLOv3, thus making the convolutional neural network better extract deeper features and decreasing the computational overhead of the YOLOv3 model. The SRFB structure makes the model more suitable for single GPU training and recognition.

(3) We integrate some advanced techniques such as Soft-NMS, GIoU Loss, and Focal Loss into Earf-YOLO and then verify the effect in experiments. Results shows that this integration indeed improves the detection performance of the model.

To summarize, Earf-YOLO model can extract contextual information effectively, depict the complex features of images, and reduce the computational overhead.

2. Related works

There are two commonly-used object recognition models currently: one-stage object recognition model and two-stage object recognition model. Two-stage object recognition model, represented by the R-CNN [8] series such as R-FCN [9], Libra R-CNN [10], Reppoints [11], and etc., generates the suggesting regions and classifies the suggesting regions by the CNN model to achieve good results. However, the two-stage object recognition model requires large computational overhead to ensure its recognition effect. One-stage object recognition model accelerates the detection speed, discards the box of generating detection proposals, and predicts confidence and multiple objects according to the features of the whole image. Although its prediction time has been shortened, its accuracy is significantly lower than that of the two-stage object recognition model. Recently, some advanced one-stage object recognition models (such as Retinanet[12], EfficacientDet[13], YOLOv4[14], etc.) update their original network with some techniques, making their accuracy similar to that of the two-stage model. In this paper, instructed by the one-stage object recognition model, we optimize its network structure to achieve higher efficiency and accuracy.

2.1. Attention Module

In recent years, the attention module has been introduced into convolutional neural networks, which has significantly improved the feature extraction ability of convolutional neural networks with little computational overhead. Meanwhile, the attention module has shown great potential for further improvement. In some studies, for example, the Squeeze-and-Excitation(SE) module [15] utilizes compression and excitation to learn feature maps of channels through two-dimensional convolution and FC layer. SE module sets an FC layer, then divides the feature maps into multiple groups. Each group performs linear transformation independently, captures cross-channel interactive features, and establishes direct correspondence between channels and weights. The Spatial Attention Module (SAM) [16] uses FC layer to transform the information of the original image into another space to retain important information through spatial transformation, enhancing regions that interest the model and weakening irrelevant regions. The Convolutional Block Attention Module (CBAM) [17] combines the channel attention module with the extra spatial attention module. It achieves better results compared with modules focusing only on channels or on spatial information. SE, SAM, and CBAM all use the FC layer as it can cross channel non-linearly to make the model less complex. But

the FC layer will also capture some unimportant pseudo-attention feature maps, which will pose negative impacts on important channels worthy of attention. In addition, using a large number of two-dimensional convolutions, it will increase the cost of memory access, and will lose the dependencies of different groups. Considering the imperfection of current attention modules, we are committed to developing a more efficient one.

2.2. The receptive field

Some expanded network receptive field modules can improve the recognition accuracy, though they will increase a little bit computational cost. He et al. [18] proposed the Spatial Pyramid Pooling (SPP) structure and used max pooling with multiple parallel $k \times k$ convolution cores to increase the receptive field of the model and get feature information. Although the SPP structure can increase the receptive field of the model and obtain multi-scale information, it fails to consider the influence of the eccentricity of the receptive field. In its receptive field, the influence of every pixel is the same, and it does not emphasize the important information in the receptive field, so some important information may be omitted. Chen et al. [19] proposed the ASPP structure in DeepLabv3+ in order to capture contextual information at multiple scales. The main difference between the ASPP module and the SPP module is that the SPP extends the Max Pooling of the original ASPP with a step size of 1 and $k \times k$ convolution cores to the Max Pooling of the 3×3 convolution cores and the Dilated Convolution[20] with a void rate of k and a step size of 1. Although ASPP uses void rate to change the sampling of center distance, it treats the center distance of all feature images equally, which leads to the confusion between environment and object. Liu et al. [7] suggested the Receptive Field Block (RFB) structure based on the Inception network[21], which used several convolution kernels with different sizes to perform multi-branch pooling, and utilized the void rate of the convolution layer to control the eccentricity of the receptive field. However, during its training process, the RFB structure fails to expand the receptive field to maximum extent to give full play to its performance of multi-scale prediction. It takes longer reasoning time to carry out prediction. In the study of receptive field, we endeavor to re-parameterize the RFB structure, to develop a new module to replace redundant convolutions, and to improve the recognition performance of the model. We creatively combine the multi-branch, multi-scale and over-parameterization of RFB to enrich the feature space during training. SRFB can be equivalently converted into a single convolution when it is deducing.

2.3. Other studies

Usually, researchers adopt advanced training strategies, modules, and post-processing methods to make the model more accurate. The strategy that only increases the training cost not the deduction cost is called “Bag of freebies”. The insertion module and post-processing method, which only increase a small amount of prediction overhead but can significantly improve the prediction accuracy of the model are called “Bag of specials”. Some methods of “Bag of freebies” are to optimize the loss function to make the model better fit the data. Aiming at the problem that negative samples are more than positive samples, Lin et al. [12] proposed Focal Loss in the RetinaNet model to make the model focus on difficult-to-classify samples by reducing the weight of easy-to-classify samples during training. Their proposal solves the problem that the loss value of negative samples takes a large proportion in the loss value of object detection. In object detection tasks, IoU is the most commonly-used indicator, so IoU Loss can be directly used as regression Loss, but it cannot optimize the overlapping ground-truth box and prediction box. Rezatofighi et al.[22] put forward GIoU Loss on the basis of IoU Loss. GIoU Loss considers the distance between the centers of the two bounding boxes to solve the overlapping problem. Post-processing, a method of Bag of specials, can filter the prediction results of models. It uses the NMS algorithm to remove the wrong prediction box from the output results and finds the most suitable position of the prediction box. The Hard-NMS algorithm sorts the prediction boxes from high scores to low scores, selects the prediction box with the highest scores, sets a threshold, deletes the prediction box whose overlap rate with the highest-scored prediction box exceeds the threshold, and repeats the above-mentioned steps with the left prediction boxes until the last one. When the overlap rate of two objects in the image is larger than the fixed threshold, Hard-NMS will set the score of the prediction box as 0, then it will be deleted. This may lead to the low-scored objects not being detected and loss of accuracy. Soft-NMS algorithm considers that the object occlusion may reduce its confidence, so it does not delete prediction box directly. When prediction box is screened, confidence is taken into account. Therefore, Soft-NMS addresses the problem that Hard-NMS mistakenly deletes the prediction box when two objects overlap.

Figure 1: A Recognition Model for symbols of Zhuang Patterns Based on Earf-YOLO (a) Texture Features of Zhuang patterns Extracted by Darknet53 (b)Feature Pyramid Enhances Feature Information Extracted by SRFB and CBEAM(c) Earf-YOLO with Focal Loss, GIoU Loss, Soft-NMS

3. Methods

Accuracy and computational overhead are essential factors for choosing object recognition models. Although YOLOv3 is known as one of the classic models for object detection, it requires high computational overhead to ensure its accuracy. Recognizing symbols on Zhuang patterns calls for accuracy but with low computational overhead. Thus, we bring up the Earf-YOLO model based on YOLOv3 to recognize symbols on Zhuang patterns, as shown in Figure 1. Earf-YOLO is mainly composed of the Darknet 53 network and the feature pyramid. The model's structure is shown in Figure 1. (a) is Darknet 53, (b) is the feature pyramid and (c) is Focal Loss, GIoU Loss and Soft-NMS. The size of the input image is 416×416 . Darknet 53 is characterized by using residual network [23]. In Darknet 53, the residual network carries out a convolution operation of 3×3 convolution kernel to save the result of the convolution layer, and uses the saved result to carry out a convolution operation of 1×1 convolution kernel and a convolution operation of 3×3 convolution kernel, and adds this result to the input of residual network as the final output result. Residual network is used to increase the depth of Earf-YOLO and improve the ability of feature extraction. The feature pyramid operates on the feature maps extracted by Darknet 53 of Earf-YOLO from two aspects. One is up-sampling the feature map to the same size as those of the other feature layers and then SRFB and CBEAM fuse its features with those of other feature maps from other feature layers, enabling the network to extract more features. The other is using the convolution operation to predict the feature map directly and using the Soft-NMS algorithm to get the final prediction result. Earf-YOLO also uses Focal Loss and GIoU Loss to calculate the classified Loss value and regression Loss value respectively to make the gradient back propagate.

3.1. Convolution Block Efficient Attention Module

We develop an efficient attention module CBEAM, as shown in Figure 2. A spatial attention module is connected to the channel attention module. The information interactions in the channel and in the space are both considered simultaneously. The channel attention module utilizes a very lightweight channel attention to avoid the negative impact of reducing dimensions on the feature map of channel attention and to adapt to k neighbors. It makes the CBEAM attention module less complex. The spatial attention module can make the model focus on useful spatial feature information and suppress useless feature information.

The channel attention in CBEAM is shown in Figure 2(a). In the channel attention, a symmetric matrix is designed to realize local cross-channel communication. The matrix of is illustrated in Equation 1.

$$w_k = \begin{bmatrix} w^{1,1} & \dots & w^{1,k} & 0 & 0 & \dots & \dots & 0 \\ 0 & w^{2,2} & \dots & w^{2,k+1} & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & w^{C,C-k+1} & \dots & w^{C,C} \end{bmatrix}, \quad (1)$$

w_k has $k \times C$ parameters. Every group of vectors in will consider the interactive feature information of k neighbors, which avoids complete independence of feature information between groups. The channel features of channel attention in CBEAM only consider the information exchange between and k neighbor channel features, and they share learning parameters. The weight of channel attention can be calculated by formula 2.

$$\omega_i = \sigma \left(\sum_{j=1}^k w_i^j y_i^j \right), y_i^j \in \Omega_i^k, \quad (2)$$

138 The calculation strategy of ω channel attention can be achieved by 1D convolution with the kernel size as k . The
139 weight of the channel attention can be calculated by formula 3.

$$\omega = \sigma(C1D_k(y)), \quad (3)$$

In the formula, $C1$ represents 1D convolution. The kernel size k of 1D convolution determines the correlation range between feature information. If the channel attention of CBEAM is divided into fixed groups, the performance of CNN can be improved, but the same feature information will be communicated between the high-dimensional channels and between the low-dimensional channels for a long time. Therefore, in CBEAM channel attention, we define a nonlinear proportional relationship between the communication information range of features and channel dimension C , so as to solve the problem of long-time information communication between features. There's a mapping ϕ between k and C , as is shown in equation 4.

$$C = \phi(k) = \log 2^{\gamma*(k+2)+b}, \quad (4)$$

Given channel C in channel attention model, the convolution kernel size of k can automatically adapt to C . The channel attention module is expressed as formula (5):

$$\begin{aligned} \mathbf{M}_c(\mathbf{F}) &= \sigma(f_1^{3 \times 3}([AvgPool(\mathbf{F})])) \\ &= \sigma(f_1^{3 \times 3}([\mathbf{F}_{avg}^s])) \end{aligned}, \quad (5)$$

140 The spatial attention module in the CBEAM is shown in Figure 2(b). The spatial attention module in the CBEAM
141 firstly conducts global average pooling and global maximum pooling on the input feature maps of channel attention to
142 aggregate the spatial information of feature maps. Then, add the two results, and utilize a two-dimensional convolution
143 to change the channel number of the feature map to 1; finally, activate the feature map by Sigmoid, then generate the
144 final spatial attention map. The spatial attention module is expressed as formula (6):

$$\begin{aligned} \mathbf{M}_s(\mathbf{F}) &= \sigma(f_2^{7 \times 7}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})])) \\ &= \sigma(f_2^{7 \times 7}([\mathbf{F}_{avg}^s; \mathbf{F}_{max}^s])) \end{aligned}, \quad (6)$$

145 In the formula, $\mathbf{F}$ refers to feature map and σ indicates the Sigmoid operation. $f_1^{3 \times 3}$ represents the one-dimensional
146 convolution layer with a 3×3 convolution kernel, and $f_2^{7 \times 7}$ represents the two-dimensional convolution layer with a
147 7×7 convolution kernel.

148 3.2. Strength Receptive Field Block

We attempt to improve the detection performance of YOLOv3 with less computational overhead. In neural networks, the depth of the network plays an important role in improving the recognition rate of the neural network. However, under the influence of gradient divergence, only increasing the depth of the network may decrease the recognition rate of the network. Therefore, we do not increase the depth of the RFB network (the RFB structure is shown in Figure 3 (a)), but use some architecture-independent structures to enhance the performance of the RFB structure. SRFB structure based on the RFB structure, is presented to optimize the network structure of YOLOv3. Compared with the RFB structure, SRFB has more informative "microstructures", and increases the receptive field of feature extraction of the model, so that each feature extracted by convolution contains a large range of feature information. when SRFB conducts prediction, it can transform a variety of matrixes into a single convolution and reduce the loss of deduction. SRFB structure is shown in Figure 3 (b). We use parallel layers with kernel sizes of 3×3 , 1×3 , 1×1 , and 3×1 . Each layer is batch normalized. During training, SRFB uses the parallel 3×3 , 1×3 , 1×1

Figure 2: The CBEAM attention module consists of two sub-modules: (a) channel attention module and (b) spatial attention module. The convolution kernel of the channel attention module can adapt to the number of feature channels.

and 3×1 convolution kernel to increase the receptive field of SRFB structure, which enhances the aggregation ability of features, and deepens the expression of nonlinear layer of network. Batch normalization is used to reduce network overfitting and speed up training. Batch normalization is shown in formula (7).

$$\mathbf{O}_{:,j} = \left(\sum_{k=1}^C \mathbf{M}_{:,k} * \mathbf{F}_{:,k}^{(j)} - \mu_j \right) \frac{\gamma_j}{\sigma_j} + \beta_j, \quad (7)$$

In formula 7, $\mathbf{M}_{:,k}$ represents the input feature map of the k -th channel, $\mathbf{F}_{:,k}^{(j)}$ represents the convolution kernel of the k -th channel, and $\mathbf{O}_{:,j}$ represents the mapping channel of output features corresponding to the j -th convolution kernel. The additivity of convolution proves that two-dimensional convolution kernels with incompatible sizes operate at the same step to produce the same resolution. The outputs of two-dimensional convolution kernels with different sizes are added up. The additivity of convolution can be considered as adding the corresponding positions of the convolution kernels to produce an equivalent kernel with the same output, as shown in Equation (8). In the prediction, SRFB uses the additivity of convolution to convert 3×3 , 1×3 , 1×1 and 3×1 convolution kernel into a new 3×3 convolution kernel to enrich the convolution feature information, as shown in Figure 4.

$$\mathbf{I} * \mathbf{K}^{(1)} + \mathbf{I} * \mathbf{K}^{(2)} = \mathbf{I} * \left(\mathbf{K}^{(1)} \oplus \mathbf{K}^{(2)} \right), \quad (8)$$

Where $\mathbf{I}$ signifies the feature matrix, $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$ represent two-dimensional convolution kernels with compatible sizes, $\oplus$ represents the sum of the corresponding positions, $*$ represents the two-dimensional convolution operator, and compatibility represents that the smaller kernel can be patched to the larger kernel.

The homogeneity of convolution proves that batch normalization of the feature space of neural network can be equivalently integrated into convolution during the prediction. According to the homogeneity of the convolution, a new kernel $\frac{\gamma_j}{\sigma_j} \mathbf{F}^{(j)}$ plus bias $\frac{\mu_j \gamma_j}{\sigma_j} + \beta_j$ can be constructed for each branch, as shown in formula (9) and (10).

$$\mathbf{F}'^{(j)} = \frac{\gamma_j}{\sigma_j} \mathbf{F}^{(j)} \oplus \frac{\gamma_j}{\sigma_j} \bar{\mathbf{F}}^{(j)} \oplus \frac{\gamma_j}{\sigma_j} \hat{\mathbf{F}}^{(j)}, \quad (9)$$

Figure 3: (a) represents RFB; (b) represents SRFB during training; and (C) represents SRFB during prediction, 1×1 , 3×1 , 3×1 convolution (the convolution kernel consist of a large number of zeros) can be regarded as a special 3×3 convolution based on the convolution additivity of SRFB.

$$b_j = -\frac{\mu_j \gamma_j}{\sigma_j} - \frac{\bar{\mu}_j \bar{\gamma}_j}{\bar{\sigma}_j} - \frac{\hat{\mu}_j \hat{\gamma}_j}{\sigma_j} + \beta_j + \bar{\beta}_j + \hat{\beta}_j, \quad (10)$$

By adding the parallel convolution kernel to the asymmetric convolution kernel, three 3×3 , 1×1 , 1×3 and 3×1 convolution branch is normalized and merged into a standard convolution layer. Compared with RFB, SRFB can obtain rich feature information without increasing computational overhead after the merging. The result after the merging is shown in equation 11.

$$\mathbf{O}_{:,i,j} + \bar{\mathbf{O}}_{:,i,j} + \hat{\mathbf{O}}_{:,i,j} + \tilde{\mathbf{O}}_{:,i,j} = \sum_{k=1}^C \mathbf{M}_{:,k} * \mathbf{F}_{:,k}^{(j)} + b_j, \quad (11)$$

Where $\mathbf{O}_{:,i,j}$, $\bar{\mathbf{O}}_{:,i,j}$, $\hat{\mathbf{O}}_{:,i,j}$, $\tilde{\mathbf{O}}_{:,i,j}$ represents the output results of 3×3 , 1×3 , 3×1 and 1×1 convolutional layer, respectively. It is worth noting that the SRFB structure can be equivalently converted only when it conducts deduction, as shown in Figure 3 (C). Because the kernel weights are randomly initialized during training, they use different calculations to obtain gradients, so they cannot be converted equivalently during training.

3.3. Bag of freebies

In the field of object detection, there are thousands of objects in an image, but only a small part needs to be detected. Compared with the two-stage detector, the one-stage detector does not use region proposal network. This will result in imbalanced distribution of positive and negative samples during training, and the loss value of the object detection is susceptible to the loss value of the negative sample. In order to improve the training result, Lin et al.[12] proposed to modify the cross-entropy loss function to obtain Focal Loss. Focal Loss controls the total loss function by setting weights in the cross-entropy loss function, as shown in formula 12, which solves the unbalanced distribution of positive and negative samples and that of easy and complex samples. To address the imbalance of negative and positive samples, Focal Loss defines a weight factor $\alpha \in [0, 1]$, then take it to the cross-entropy loss function. When the number of positive samples is small, the value of will be large and the loss value of positive samples will increase. To solve the imbalance of easy and complex samples, Focal Loss suggests an adjustment factor to reduce the weight of easy samples and to make the model focus on training complex samples.

Figure 4: The sliding window shows the additivity of convolution. There are four convolution kernels with the size of 3×3 , 1×1 , 1×3 and 3×1 . They can share the sliding window based on the additivity of the convolution.

$$L(p, y) = \begin{cases} -\alpha(1-p)^{\gamma} \log_a p, & y = 1 \\ -(1-\alpha)p^{\gamma} \log_a (1-p), & y = 0 \end{cases} \quad (12)$$

Where $p \in [0, 1]$ represents category probability of predicted samples; is the category label. Set its value to 0 and 1.

At present, many object detection algorithms generally use L1 and L2 norm to calculate the loss value. L1 and L2 norm independently calculate the loss value of the four coordinate variables of the prediction box. The coordinate variables are independent, but there is some correlation among the coordinate variables in real situation. When detection performance of the model is evaluated, IoU is used to detect whether there is an object. If the norm regression of L1 and L2 is directly used to calculate the coordinate frame, the value of evaluation index will also be affected. Yu J et al.[24] proposed IoU as a regression loss function to calculate the coordinate frame, which solved the above problems. However, if IoU is directly used as the boundary loss, when the predicted box and the ground-truth box do not overlap, IoU will become 0, gradient will become 0 and the boundary loss cannot be optimized. Rezatofighi et al. [22] proposed GIoU Loss as a boundary loss. GIoU retains the scale invariant of IoU as loss function, and add the distance between two boxes to optimize the loss value, which solves the problem that prediction box and the ground-truth box do not overlap and gradient become 0. The calculation method of GIoU is shown in formula (13).

$$GIoU = \frac{|A \cap B|}{|A \cup B|} - \frac{|C/(A \cup B)|}{|C|}, \quad (13)$$

In the formula, A and B refer to the prediction box and the ground-truth box respectively, and C is the smallest closed box that contain both boxes.

When GIoU becomes larger, the GIoU Loss will become smaller and the network will be optimized to make the prediction box and the ground-truth box highly overlap. The boundary loss function of YOLOv3 optimized by GIoU is shown in formula (14).

$$\text{bbox_loss} = \sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{\text{obj}} (1 - GIoU) \times \left[2 - \left(\hat{w}_i^j \times \hat{h}_i^j \right) \right], \quad (14)$$

S is the grid size, S^2 is the $13 \times 13, 26 \times 26, 52 \times 52$ grid, B is the prediction box, I_{ij}^{obj} is the prediction box at i, j has a target, its value is 1, otherwise is 0, $\hat{w}_i^j$ and $\hat{h}_i^j$ are the width and height of the prediction box at i, j .

3.4. Bag of specials

In post-processing methods of object detection, Hard-NMS will directly delete the overlapping prediction box. To address the problem, Bold et al. [25] used the Soft-NMS algorithm to suppress the wrong prediction box from a new perspective. As formula 15 indicates, Soft-NMS does not delete low-scored prediction box directly. It will lower the scores further and then set a threshold to delete low-scored prediction boxes. Soft-NMS will also use the Gaussian weight function (as shown in formula 16): multiply the score of the current prediction box with a weight function. This function will attenuate the scores of adjacent prediction boxes that overlap the highest-scored prediction box M . The more overlapping the prediction box with the highest-scored one, the more serious the attenuation of the prediction box will be.

$$s_i = \begin{cases} s_i, iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)), iou(M, b_i) \geq N_t \end{cases}, \quad (15)$$

s_i is the score of the current prediction box; N_t is the threshold value; M is the prediction box with highest score. b_i is the prediction box with each score.

4. Results and discussion

In this section, we conduct experiments to compare Earf-YOLO with other detection methods. In evaluating the performance of detection methods, mAP (IOU=0.5), FPS and ParaWith reference to the limitations of YOLOv3 in recognizing symbols on Zhuang pattern, such as slow detection speed, unable to detect small object, and inaccurate positioning of bounding box, we propose a new model: Earf-YOLO (Efficient Attention Receptive Field You only look once) in this paper. In Earf-YOLO, we present an attention module: CBEAM (Convolution Block Efficient Attention Module) at first, which provides feature maps from channel and spatial dimensions. In CBEAM module, a local cross-channel interaction strategy without reducing dimensionality is used to improve the performance of the convolutional neural network. Besides, we put forward the SRFB (Strength Receptive Field Block) structure. During its training, more branch structures will be generated to enrich the feature space of the convolutional block. During its prediction, the multi-branched structures will be reparametrized and fused into one main branch to improve the performance of the model. Finally, we adopt some advanced training techniques to improve the detection performance. Experiments on the dataset of Zhuang patterns and the COCO dataset show that the Earf-YOLO model can effectively reduce the error of the prediction box and the ground-truth box, and decrease the calculation time. The mAP value of this model on the dataset of Zhuang patterns and on the COCO dataset reaches 82.1 (IoU=0.5) and 62.14 (IoU=0.5) respectively. are often used as evaluation indicators. Among them, the larger the mAP value is, the better the detection effect will be; the larger the FPS value is, the higher the detection efficiency will be; and the smaller the Param value is, the lower the network memory consumption will be. In this section, we first conduct experiments on the dataset of Zhuang patterns, which proves that Earf-YOLO can recognize symbols on Zhuang patterns successfully and efficiently. Then we compare the values of mAP and those of FPS of Earf-YOLO with other advanced networks on COCO dataset, which also illustrates Earf-YOLO network can also achieve good results on other datasets.

4.1. Zhuang pattern symbol dataset

The datasets used in the experiment are symbols on Zhuang patterns, as shown in Figure 5. The Zhuang people have incorporated their wisdom and culture into Zhuang patterns, usually reflecting their yearning for a better life. For example, the delicate and beautiful flowers on Zhuang patterns are believed to be representatives of natural beauty and colorful life; the birds on Zhuang patterns can arouse people's longing for a happy life as birds usually lead happy and free life in the forest. To date, there is no specific dataset composed by symbols on Zhuang patterns. The datasets used in this research are images taken by researchers in Zhuang tribes. There are about 19,199 images of Zhuang patterns in the dataset; and these images are classified into 20 categories. To ensure the justice of the model when it gets trained, we tried to maintain a balance in the number of images in each category. A total of 10,592 images are selected as training samples and 8,607 images as test samples. The samples distribution is shown in figure 6.

Figure 5: Symbols on Zhuang Patterns

Figure 6: Distribution of Training samples and test samples

Table 1: The comparison of Param, FPS and mAP of different attention modules in YOLOv3 on the dataset of Zhuang patterns.

SENet	ECANet	CBAM	CBEAM	Param(M)	FPS	mAP(%)
✓	✓	✓	✓	61.678	20	74.6
				61.77	21	75.4
				61.679	20	76.3
				61.771	21	76.1
				61.679	20	77.6

4.2. COCO 2014 data set

The COCO dataset is the most commonly-used database for object detection. There are on average 7.2 objects in every image and there are 80 categories in COCO dataset, making it more difficult to detect than dataset of Zhuang patterns because it contains much less categories. There are 82783 images in its training set and 604907 sampling frames. We selected 10,000 images as the validation set from the COCO 2014 dataset to verify our model.

4.3. Experiment Setting

The experimental data were trained on the environment of Python 3.6 and Keras 2.3.1 under the configuration of GTX 2070 8G and Windows 10, and the number of training iterations was 500. The image input size was fixed to 416 × 416, and the optimizer was Adam. The attenuation strategy of the learning rate was the Cosine annealing attenuation strategy. Set the initial learning rate of the Cosine annealing attenuation strategy to 0.001, the highest learning rate to 0.01, and the lowest learning rate to 0.0001. In the first 400 experiments, froze the first 170 convolution layers of the network, then trained the remaining convolution layers. In the last 100 experiments, opened and trained all convolution layers.

4.4. Ablation experiment on dataset of Zhuang patterns

To explore the influence of the CBEAM on the recognition results, YOLOv3 was used as the basic detection model. We compared the CBEAM with SENet[15], ECANet[26], and the CBAM[17]. As shown in Table 1, the mAP value of YOLOv3 with the CBEAM reached the highest 77.6%, 3% higher than that of single YOLOv3, 2.2% higher than that of YOLOv3 with SENet, 1.3% higher than that of YOLOv3 with ECANet, and 1.5% higher than that of YOLOv3 with CBAM. To summarize, the integrated attention module obviously improved the performance of the model for recognizing symbols on Zhuang patterns but among all of the above integration, YOLOv3 with CBEAM was the most accurate. To better study the effect of the SRFB in YOLOv3, we compared the SRFB structure with the SPP structure and the RFB structure in YOLOv3. As shown in Table 2, YOLOv3 with SRFB structure increased the mAP value to the highest 77.3%, which was 2.7% higher than that of single YOLOv3, 1.1% higher than that of YOLOv3 with SPP, and 1.4% higher than that of YOLOv3 with RFB. The Param of YOLOv3 with SRFB structure reduced by 6.14M and 13.78M compared with single YOLOv3 and YOLOv3 with SPP respectively. Besides, YOLOv3 with SRFB structure increased the FPS value to the highest 30, which was 10 higher than that of YOLOv3, 7 higher than that of YOLOv3 with SPP and 2 higher than that of YOLOv3 with RFB. Those data illustrated that YOLOv3 with SRFB has less complex design but achieves better predictive result only by simple manual design and re-parameterization than YOLOv3 with RFB structure. The above results proved that the SRFB structure gave YOLOv3 faster recognition speed and higher recognition rate. Finally, we studied the impact of different techniques on YOLOv3. We integrated Soft-NMS, Focal Loss, and GIoU Loss with YOLOv3 respectively. As shown in Table 3, The mAP value of YOLOv3 with Soft-NMS increased by 1.8% compared with YOLOv3; the mAP value of YOLOv3 with Focal Loss increased to 76.6%, the mAP value of YOLOv3 with GIoU Loss increased to 76.98%. The experimental results showed that on the dataset of Zhuang patterns, YOLOv3 with additional techniques had better detection results than single YOLOv3. Though it seemed that those techniques incorporated in YOLOv3 made no big difference in values of FPS and mAP respectively, when Soft-NMS, Focal Loss and GIoU Loss all incorporated in YOLOv3 simultaneously, the performance of model would be promoted.

Table 2: The comparison of Param, FPS and mAP of different expanding receiver field methods in YOLOv3 on the dataset of Zhuang patterns.

SPP	RFB	SRFB	Param(M)	FPS	mAP(%)
			61.678	20	74.6
✓			69.31	23	76.2
	✓		54.7	28	75.9
		✓	55.53	30	77.3

Table 3: The comparison of Param, FPS and mAP of different methods of Bag of Freebies methods and Bag of Specials in YOLOv3 on the dataset of Zhuang patterns.

Hard-NMS	Soft-NMS	Focal Loss	GIoU Loss	Param(M)	FPS	mAP(%)
✓				61.678	20	74.6
	✓			61.678	19	76.4
✓		✓		61.678	20	76.63
✓			✓	61.678	20	76.98

4.5. The influence of different prediction methods on the model results

We studied the impact of different prediction methods on its accuracy on the dataset of Zhuang patterns. In this paper, we used K-means to cluster the anchor box of the dataset of Zhuang patterns to obtain 9 anchor boxes with different but their best sizes. Earf-YOLO with and without the clustering scheme were compared, whose results were listed in Table 4. From Table 4, the mAP value of Earf-YOLO with clustering scheme was 2.8% higher than that of Earf-YOLO without clustering. Besides, we still explored the impact of image input size on the model’s ability to extract features. As shown in Figure 8, as the size of input image increases, its mAP value will continue to increase.

4.6. The influence of different prediction methods on the model results

We studied the impact of different prediction methods on its accuracy on the dataset of Zhuang patterns. In this paper, we used K-means to cluster the anchor box of the dataset of Zhuang patterns to obtain 9 anchor boxes with different but their best sizes. Earf-YOLO with and without the clustering scheme were compared, whose results were listed in Table 4. From Table 4, the mAP value of Earf-YOLO with clustering scheme was 2.8% higher than that of Earf-YOLO without clustering. Besides, we still explored the impact of image input size on the model’s ability to extract features. As shown in Figure 7, as the size of input image increases, its mAP value will continue to increase.

4.7. Comparison of different object detection models on the dataset of Zhuang pattern

We compared the Earf-YOLO model with other advanced models, including RFBNet[7], RetinaNet, YOLOv3-SPP[27], YOLOv3 [5]. From Table 5, it was obvious that the mAP value and FPS value of Earf-YOLO reached 82.1% and 26 respectively, and both of them reached the highest value. Earf-YOLO achieves a better recognition

Table 4: The comparison of mPA value of Anchor box using K-means clustering = {30,87,50,177,66,299,71,233,93,332,115,303,137,341,168,339,206,360} and Anchor box without clustering = {12,16,19,36,40,28,36,75,76,55,72,146,142,110,192,243,459,401}.

Method	Achor box	mAP(%)
Earf-YOLO with clustering scheme	9	82.1
Earf-YOLO without clustering scheme	9	79.3

Figure 7: The Comparison of mAP Value With Different Input Sizes

performance because Earf-YOLO with CBEAM attention module and SRFB structure can better capture information and detect objects in the picture. Moreover, it also incorporate Soft-NMS, Focal Loss, GIoU Loss algorithm to optimize YOLOv3 to make it more accurate.

4.8. Comparison of different models on COCO dataset

The COCO dataset is much more difficult to recognize than that of Zhuang patterns, and the evaluation criteria are stricter, so it can better evaluate the performance of different methods. We Compared Earf-YOLO with YOLOv4 and EfficientDet-D0. As shown in Table 6, the mAP value of our model was 9.82 % higher than that of EfficientDet-D0 and slightly lower than that of YOLOv4. The FPS value of our model was the highest of all, which increased by 10 compared with EfficientDet-D0 and by 9 compared with YOLOv4. All of these illustrated that Earf-YOLO could ensure the detection accuracy in a relatively fast speed.

Table 5: The comparison of the mAP and FPS of different object detection models on the dataset of Zhuang patterns

Method	Size	FPS	mAP(%)
YOLOv3	416	20	74.6
Earf-YOLO	416	26	82.1
RFBNet	512	18	73.9
RetinaNet	500	10	77.2
YOLOv3-SPP	416	23	76.2

Table 6: Comparison of FPS and mAP values of different models on COCO dataset

Method	Size	FPS	mAP%
Earf-YOLO	416	28	62.14
YOLOv4	416	19	64.23
EfficientDet-D0	512	18	52.32

Figure 8: Recognition results of Earf-YOLO for Zhuang pattern symbols

4.9. The recognition results of Earf-YOLO for Zhuang patterns

Figure 8 is the results of Earf-YOLO for recognizing symbols on Zhuang patterns, including the categories, the confidence, and the position of patterns. As shown in Figure 8. our model can accurately recognize the categories and positions of Zhuang patterns with high confidence, which is of great significance to further study the ethnic minority cultures.

5. Conclusion

In this paper, a fast and powerful Earf-YOLO detection model is proposed to detect and recognize the symbols on Zhuang patterns. First of all, our model incorporates CBEAM attention module which uses a local cross-channel interaction strategy without reducing dimensionality to improve the performance of the convolutional neural network. Besides, it integrates SRFB structure since this structure can sum the output of the convolution branch, which can enhance the feature extraction ability of the model without extra deduction time. Finally, it introduces some advanced techniques to increase the accuracy of the model. The mAP value of new-proposed model reaches 82.1% on the dataset of Zhuang patterns. Compared with the latest object detection methods, Earf-YOLO model is the most effective, but there is still room for further improvement in computational cost and recognition rate. We will continue to optimize this model to make it available on mobile phone, and to detect more decorative patterns of nationalities such as Yao brocade, Miao brocade, Dong brocade, Maonan brocade, and even the other 3 famous brocades: Yun brocade, Shu brocade and Song brocade.

Availability of data and materials

Not applicable

Abbreviations

Erf-YOLO:Efficient Attention Receptive Field You only look once
YOLO: You only look once
R-CNN:Region convolutional neural network
COCO:Common Objects in Context
GPU:Graphics processing unit
SRFB:Strength Receptive Field Block
CBEAM:Convolution Block Efficient Attention Module
SE:Squeeze-and-Excitation
SAM:Spatial Attention Module
Fast R-CNN:Fast Region-based Convolutional Network
Faster R-CNN:Faster Region Convolutional Neural Network
Mask R-CNN:Mask Region Convolutional Neural Network
SSD:Single Shot multi-box detector
CBAM:Convolutional Block Attention Module
SPP:Spatial Pyramid Pooling
RFB:Receptive Field Block

References

- [1] X. Qing, Semiotic logic in inheritance and application of patterns of southwest minorities, 36(07): 41-44 (2015).
- [2] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
- [3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv:1506.01497 (2015).
- [4] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- [5] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).
- [6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.
- [7] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.
- [8] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
- [9] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, arXiv preprint arXiv:1605.06409 (2016).
- [10] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra r-cnn: Towards balanced learning for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 821–830.
- [11] Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, Reppoints: Point set representation for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9657–9666.
- [12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
- [13] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.
- [14] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934 (2020).
- [15] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
- [16] X. Zhu, D. Cheng, Z. Zhang, S. Lin, J. Dai, An empirical study of spatial attention mechanisms in deep networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6688–6697.
- [17] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
- [18] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE transactions on pattern analysis and machine intelligence 37 (9) (2015) 1904–1916.
- [19] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters—improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4353–4361.

- [20] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122 (2015).
- [21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
- [22] N. He, L. Fang, A. Plaza, Hybrid first and second order attention unet for building segmentation in remote sensing images, Science China Information Sciences 63 (4) (2020) 1–12.
- [23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [24] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, Unitbox: An advanced object detection network, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 516–520.
- [25] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-nms—improving object detection with one line of code, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5561–5569.
- [26] W. Qilong, W. Banggu, Z. Pengfei, L. Peihua, Z. Wangmeng, H. Qinghua, Eca-net: Efficient channel attention for deep convolutional neural networks. (2020).
- [27] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147 (2016).

Acknowledgment

The author would like to thank Peng Xie from Southwest Jiaotong University and Jiaqi Xu from University of Science and Technology of China for their insights and feedback on the first draft of this article.

Funding

This work is supported by General Project of Guangxi Natural Science Foundation (2019GXNSFAA245053), Guangxi Science and Technology Major Project (AA19254016), the National Natural Science Foundation of China (61862018), Guangxi Natural Science Foundation Project (2018GXNSFAA138084).

Author information

Affiliations

School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China
XIN WANG, ZHUO SHI

School of Marine Engineering, Guilin University of Electronic Technology, Beihai, China

XIN WANG, JINGKE YAN, QIN QIN, JUN WANG, HUANGHE LI

Basic Teaching Department, Guilin University of Electronic Technology, Beihai, China

WANG QIN

Contributions

XIN WANG: writing—original draft, writing—review and editing, ZHUO SHI, JINGKE YAN: writing—original draft, conceptualization, data curation, validation. JINGKE YAN: conceptualization, data curation, validation. QIN QIN, JUN WANG, HUANGHE LI: conceptualization, formal analysis, writing—review and editing, supervision, and funding. WANG QIN: formal analysis, supervision, writing—review and editing. All authors read and approved the final manuscript.

Corresponding author Correspondence to JINGKE YAN.

Ethics declarations

Ethics approval and consent to participate

This research does not involve any human or animal participation.

Competing interests

The authors declare that they do not have any conflict of interests. All authors have checked and agreed the submission.