


Research Article

Earf-YOLO: An Efficient Attention Receptive Field Model for Recognizing Symbols of Zhuang Minority Patterns

Xin Wang ^{1,2,3}, Jingke Yan ², Qin Qin ², Qin Wang ⁴, Jingye Cai,¹ Jianhua Deng,¹ Jun Wang,² Zhuo Shi,³ Yi Feng,² and Bingxu Chen²

¹School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China

²School of Marine Engineering, Guilin University of Electronic Technology, Beihai 536000, China

³School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China

⁴Basic Teaching Department, Guilin University of Electronic Technology, Beihai 536000, China

Correspondence should be addressed to Jingke Yan; 592499985@qq.com, Qin Qin; qinqin@guet.edu.cn, and Qin Wang; 283252764@qq.com

Received 25 September 2021; Accepted 16 February 2022; Published 21 March 2022

Academic Editor: Muhammad Haroon Yousaf

Copyright © 2022 Xin Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As for recognizing Zhuang minority pattern symbols, current recognition models often cause high computational overhead and low accuracy since Zhuang minority pattern symbols have large feature vectors and some complex features. In this paper, we present the efficient attention receptive field you only look once (Earf-YOLO), a new scheme to address those problems. Firstly, a global-local-transformer (GLocalT) structure is proposed, through which other control systems are introduced into the axial self-attention module, and global-local training strategies are also designed. The structure can use other control systems to compensate for the lost feature information along the height, width, and channel axes. The global-local training strategy can encode long-term dependencies between features and reduce local information loss, fully illustrating that the structure has high feature expression ability. Besides, strength receptive field block (SRFB) is suggested to use the dilated convolution to control the receptive field's eccentricity and enrich the feature information of the receptive field during its training. With more branches, it can better extract multiscale features, enrich the feature space of the convolution block, and reparametrize multibranch during prediction to fuse them into the main branch, all of which contribute to the improvement of the model performance. Finally, some advanced training techniques are adopted to enhance the detection effect further. In the end, comparative experiments are conducted on the datasets of Zhuang pattern symbols and PASCAL VOC, whose results indicate that the AP and FPS of the suggested model reach their highest values, manifesting its high efficiency.

1. Introduction

Ethnic minorities integrate their religious culture and totem culture into the pattern symbols of clothing and architectural decoration, usually with profound connotations. They are the basis for classifying national images [1]. As the minority patterns often feature their exquisite colors and structural and artistic styles, they are significant in retrieving the origin, distribution, and development of ethnic groups. With globalization and modernization, ethnic pattern culture is disappearing gradually. How to inherit, protect, spread, and utilize the traditional ethnic culture of pattern

symbols in China should be valued. Therefore, correctly and efficiently recognizing the symbols of minority patterns is vital in realizing digital protection and in inheriting ethnic culture.

Different from the modern symbols, the minority symbols have the following characteristics: (1) complex pattern structures; (2) bright colors; (3) rich accessories with different visual styles; and (4) rich connotations, especially texture details, often with rich ethnic characteristics. For example, the symbols of Zhuang patterns are bright in color, evident in color gradation, and different in components. Besides, different branches of the same nationality have

different pattern symbols. Let us take Zhuang patterns as an example: different pattern symbols in different branches reflect their unique aesthetics and styles.

Object detection, an essential branch of AI and pattern recognition, has been successfully applied to many areas, such as transportation [2, 3], rescue [4, 5], and the demand in these areas is still growing. Without object detection, symbol extraction will fail to obtain all features in images as they often contain many totems, patterns, and designs that serve as the basis for extracting and detecting features from minority patterns.

Recently, Huo et al. [6] classified ethnic costumes in natural settings into 11 representative ethnic pattern symbols, including Miao, Mongolian, and Korean, based on the component detection and the feature fusion of the costume pattern symbols. Sun et al. [7] classified the ethnic costumes by using Faster R-CNN to extract attribute features from symbols of costume patterns. However, the large feature vector of the pattern symbols extracted from ethnic costumes will increase data storage and computational overhead. Besides, the semantic gap between low-level features and high-level attributes presents the following difficulties: (1) the symbols of ethnic patterns have distinct colors, various styles, and distinctive texture patterns. How to divide the visual style of ethnic pattern symbols and bridge the semantic gap between high-level visual attributes and geometric features is critical to improving recognition accuracy. (2) Some small ethnic symbols with small coverage, low resolution, and inconspicuous features decrease the detection efficiency. (3) The current object detection models such as the YOLO series [8–11] often need high computational overhead.

The YOLO series [8–11] plays a vital role in object detection tasks in the single-stage detector. We propose an improved model, Earf-YOLO, based on YOLOv4 [11] to solve the above three problems. Earf-YOLO can extract global and local features and increase the model's receptive field, improving the detection accuracy at a relatively fast detection speed. The overview of Earf-YOLO is shown in Figure 1.

- (1) A new transformer architecture is designed to better describe the feature information of the pattern symbols. It adopts a gating self-attention mechanism to better converge features from height, width, and channel axes. It divides the feature map into patches and inputs them and the original feature map into the transformer to learn long-distance dependencies between features and reduce local information loss between features.
- (2) To increase the receptive field of pattern symbol extraction, enhance the ability of complex pattern symbols, and reduce the computational overhead of pattern symbol recognition model, the strength receptive field block (SRFB) structure is designed to replace the redundant convolution layer in the feature pyramid of the model. It not only improves the ability of the convolutional neural network to extract deeper features but also reduces the

computational overhead of the model, accelerating the model training and recognition speed.

- (3) Some advanced techniques, including the Soft-NMS [12], GIoU Loss [13], and Focal Loss [14], are integrated into Earf-YOLO, and their effects during training are verified. Experiment results demonstrate that these advanced techniques can improve detection performance.
- (4) The frames per second (FPS) and average precision (AP) of previous models and the proposed model are compared on the Zhuang pattern symbol datasets, as shown in Figure 2. The final result illustrates that Earf-YOLO achieves high performance in detecting pattern symbols.

2. Related Work

2.1. Traditional Object Detection Model. The detection task of Zhuang pattern symbols is to extract the style element features of Zhuang pattern symbols through the model to realize the positioning and classification of Zhuang pattern symbols. In recent years, many researchers have researched the object recognition models. Ribeiro et al. [15] proposed an end-to-end dual neural network architecture to recognize expiration dates in snack packaging. They used neural networks to fuse global and local features to recognize features. In recognizing Zhuang pattern symbols, we should focus on the classification and shape of multiple pattern symbols. Symbol classification and object positioning are different in detecting, so a new detection network is needed. The network of our model shows that the object classification focuses on judging local features, and the object positioning focuses on judging the global feature region. Nguyen et al. [16] demonstrated an object frame generation method based on a deep convolutional neural network (DCNN), which trained an object positioning detector to learn deep feature information from the bounding candidate frame detected in the image. When recognizing pattern symbols, there are color overlaps between geometric features and background features, making the model unable to explore the deep feature information of the relevant graph primitives and background. Erhan et al. [17] focused on processing similar instance objects in an image and proposed a display-inspired neural network to detect objects of an unknown category. Although the pattern symbol image of Zhuang nationality contains multiple similar objects and is also detected for multiple objects, the detection accuracy of the model is low due to the complex background of the Zhuang patterns.

2.2. Two-Stage Object Detection Model. Because of the low accuracy of traditional object detection algorithms, Girshick et al. [18] proposed a two-stage detection model based on R-CNN. Firstly, R-CNN uses a selective search algorithm to extract 2000 candidate frames from the images to be detected. Then, R-CNN scales 2000 candidate frames into 227×227 and uses a convolutional neural network to extract features from candidate frames to obtain feature vectors.

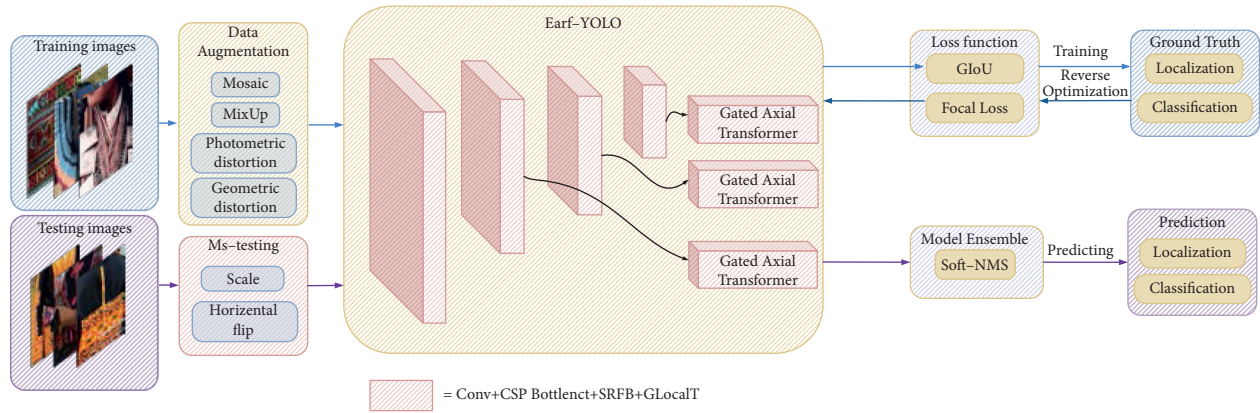


FIGURE 1: An overview of Earf-YOLO. Transformer and SRFB are used to optimize YOLOv4. Then, some advanced techniques such as data augmentation, Giou Loss, Focal Loss, and Soft-NMS were employed to improve the Earf-YOLO detection efficiency on Zhuang pattern symbol datasets. It is worth noting that the feature map in the figure can obtain three prediction results through the gating axial transformer.

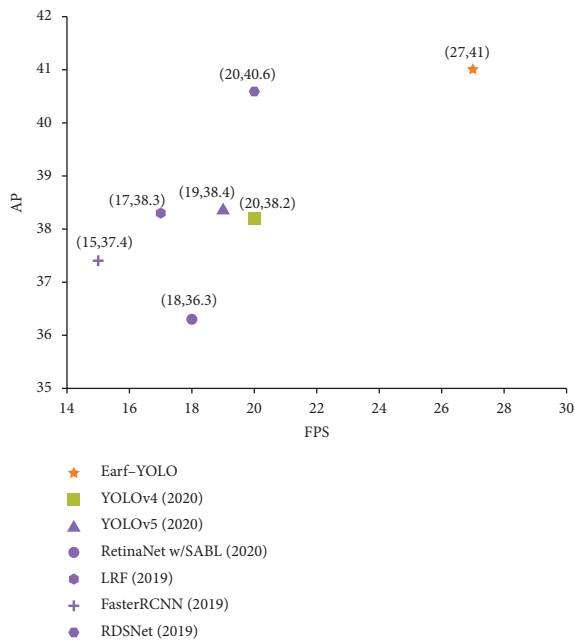


FIGURE 2: Comparison of FPS and AP of the presented model and other six benchmark models on the dataset of Zhuang pattern symbols.

Finally, the model inputs feature vectors into the support vector machine and the fully connected network. The support vector machine can classify feature vectors to get category information, and the fully connected network performs regression operations on feature vectors to obtain corresponding coordinates. Although R-CNN is cleverly designed, the model detection is divided into multiple stages, resulting in a significant decrease in detection efficiency. Therefore, Girshick [19] proposed a Fast RCNN. It does not need to input all the candidate frames into the deep learning model. Instead, it only needs to select all candidate frames, input the selected candidate frames into the network for feature mapping, and obtain the prediction category and the position of the prediction frame. The model performs a selective search to improve detection speed but spends a lot

of time selectively searching for candidate frames. To solve this problem, Faster RCNN [20] and Mask RCNN [21] added a region proposal network based on Fast RCNN. It extracts candidate frames by setting anchors of different scales and replaces the traditional candidate frame generation methods, such as the selective search method, which improves the computing speed of the network. With the development of deep learning, affected by the complexity of the primary network, the number of candidate frames, classification, and the complexity of the regression sub-network among other factors, the above techniques require high computation overhead, which seriously influences the model prediction and training performance.

2.3. *One-Stage Object Detection Model.* As for the low efficiency of the two-stage object detection algorithm, YOLOv1 [8] removes the candidate frame extraction branch of the two-stage algorithm and directly implements feature extraction, candidate frame classification, and regression in the same deep convolutional network, making a single network complete classification determination and locate regression. YOLOv1 abandons the candidate frame stage and speeds up the detection speed. However, it is not accurate enough in locating objects and has a low recall rate, resulting in low detection accuracy. Farhadi et al. [9] proposed the YOLOv2 model to address this problem, mainly using a multiscale classifier and multiscale object frame position detector to improve the model accuracy. Although the accuracy of YOLOv2 improves a lot, its accuracy is still not ideal in subsequent industrial applications. YOLOv3 [10] designs a Darknet53 residual network and feature pyramid network by learning the residual network and the RPN of Faster RCNN to improve network depth and network space representational ability. Therefore, a large number of scholars have made relevant studies on YOLOv3. Based on the YOLOv3 model, Li et al. [22] conducted rapid detection of cracks in the fuselage or engine blade of an aircraft structure by depthwise separable convolution and feature pyramid. Shi et al. [4] optimized the YOLOv3 model by reducing its parameters, improving its detection speed for underwater

objects, optimizing the residual network, and strengthening its feature extraction ability. Although the methods mentioned above based on YOLOv3 can identify large objects well, it is easy to neglect hard-detected and overlapping features. Bochkovskiy et al. [11] proposed the YOLOv4 model to solve those problems by applying advanced bag of freebies and bag of specials methods to achieve better detection results. However, the model is difficult to deploy on the platform with few resources because of its large number of network parameters and its large computation overhead. As for the massive overhead of neural networks that limited the model's detection and inference on mobile devices, Zhou et al. [23] proposed the RSANET model, which introduced lightweight convolution (LCNet) and attentional pyramid networks with residual as the prediction head. Their experiments proved that the model could reduce computational overhead effectively. John and Mita [24] proposed a residual semantic-guided attention feature pyramid network, including input and output branches. The model used the input branch to extract the features of a single sensor and then used the residual connection to integrate the extracted features into the output perception branch. Although both models can perform well on certain experimental datasets, they have low detection accuracy, high detection error, and a high neglected detection rate for detecting small specific objects in Zhuang patterns. Based on the previous research, we present an improved Earf-YOLO model in this paper to optimize the YOLOv4 to address the above problems.

3. Methods

This section details the Earf-YOLO, the proposed Zhuang pattern symbol recognition model, including the introduction of its structure and its contributions.

3.1. Network Structure. Accuracy and computational overhead are essential indexes to determine the performance of an object detection model. YOLOv4, one of the classical models for detecting an object, requires a high computational overhead to ensure accuracy. Therefore, we focus on detecting symbols of Zhuang patterns accurately with minimal computational overhead. Based on the previous studies, we propose the Earf-YOLO model based on the YOLOv4, as shown in Figure 3, mainly composed of the Backbone, neck, and transformer predict. First of all, Backbone mainly uses the CSPDarkNet53 featured by introducing the CSPNet structure [25] to reduce the computational overhead, eliminate the redundant gradient information when the network is optimized reversely, enhance the convolutional network's learning ability, and ensure accuracy while making the network lightweight. Secondly, Neck network adopts the structure of strength receptive field block (SRFB), global-local-transformer (GLocalT), and path aggregation network (PANet) [26]. The SRFB structure can effectively improve the receptive field of the network and extract important context features. GLocalT can extract the local and global features of Zhuang pattern symbols. PANet is the improved version of feature pyramid

network (FPN) [27], to which a bottom-up path augmentation structure is added to avoid losing shallow feature information during transmission, improving prediction accuracy. Finally, transformer predict is used for regression and classification. Unlike YOLOv4, the Earf-YOLO uses the global transformer to predict three feature maps of different sizes to detect small, medium, and large objects. The size of the prior frame is obtained by clustering the sample objects through the k-means algorithm, based on which the size and position of the prediction frame can be calculated by relative offset.

3.2. Global-Local-Transformer. With the wide application of transformer [28] in natural language, transformer [29] is also used for computer vision tasks. Recent studies show that transformer-based models can achieve good detection results only by training datasets with rich features. It is challenging for models to learn the position-coding of the image if traditional transformer architecture is used because there are various pattern symbols and small textures in datasets of Zhuang pattern symbols. Therefore, according to the gating attention mechanism of medical segmentation proposed by Valanarasu et al. [30], we designed a global-local-transformer (GLocalT) structure to detect symbols of Zhuang patterns. It inputs global and local features into the transformer for extracting and fusing features, respectively. In addition, it adopts a gating axial attention layer to serve as the basic structure block of the transformer. The architecture of GLocalT is shown in Figure 4(a).

Gating axial attention layer. In traditional transformers, $\text{softmax}(q^T k)$ is often used to calculate the global affinity, and the value matrix v is aggregated together, where $q = W_Q$, $k = W_K$, and $v = W_V$ represent query matrix, key-value matrix, and value matrix, respectively. They all calculate the mapping matrix by inputting x . The mapping matrix $W_Q, W_K, W_V \in R^{C_{in} \times C_{out}}$ can be obtained through the model's learning. This approach enables the model to capture nonlocal information from the global feature mapping. However, this calculation requires a large amount of computational overhead, and its computation overhead will increase when the feature map size increases. The self-attention layer of this method is not conducive to extracting any position feature information when calculating nonlocal context feature information. The positional feature information is crucial in recognizing symbols of Zhuang patterns and is usually used to locate objects. Researchers [30–32] decompose the self-attention module into two self-attention modules to make the computational affinity less complex. The first module performs self-attention on the height axis, and the second module performs self-attention on the width axis. Adding axial attention to the height and width axis can effectively simulate the original self-attention mechanism and better calculate efficiency. In addition, to make the self-attention mechanism more sensitive to position information when calculating affinity, they attached position bias items and gating mechanism to all q , k , and v , thus enabling the model to capture remote interactions with precise positional information. Therefore, based on the above discussions, for

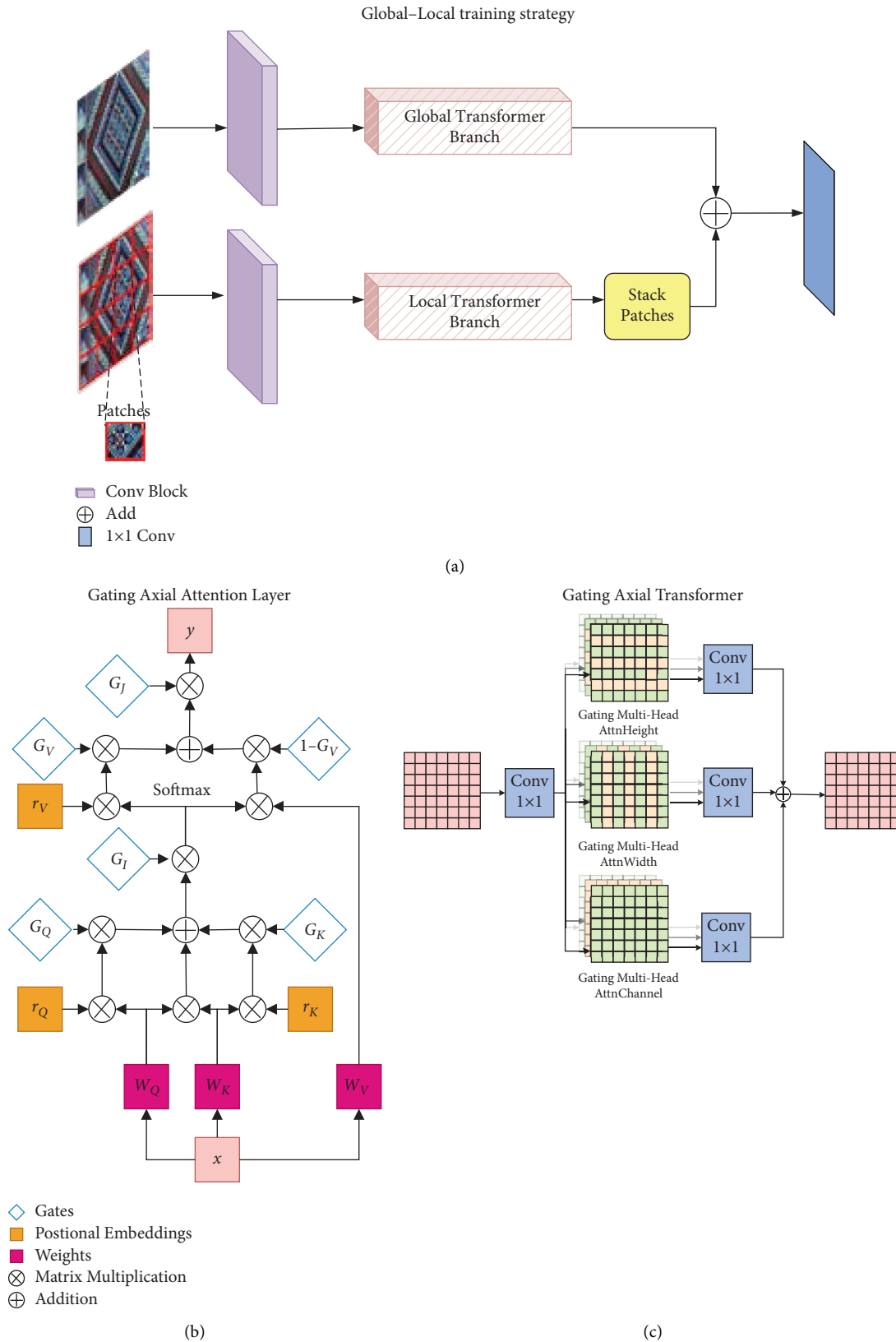


FIGURE 4: (a) represents the main architecture of GLocalT with the global-local training mechanism, (b) represents gating axial attention layer, a multiple attention block based on gating control to control the information amount of the key, query, and value provided by the position embedding, and (c) means that a gated axial transformer is used in GLocalT, and the gated axial transformer contains the gated axial attention layer encoded along the height, width, and channel.

proposed spatial pyramid pooling (SPP) block, which used max pooling of multiple parallel $k \times k$ convolution kernels to obtain receptive fields and extract feature information. The SPP structure can increase the receptive field of the model and get multiscale feature information of Zhuang pattern symbols. However, it fails to consider the effect of the eccentricity of the model's receptive field in the recognition process of Zhuang pattern symbols, which makes the effect of each pattern symbol image pixel in the perception field of the model is the same, and the vital information in the receptive field is not emphasized. It also makes the model increase inference time when the model conducts prediction. Based on the above discussion, we put forward the strength receptive field block (SRFB) structure, which not only adopts multiple convolutional kernels of different sizes to carry out multibranch pooling. In the branches, the SRFB structure uses the convolutional layer's void rate to control the receptive field's eccentricity and transforms various matrices into a single convolution during prediction to optimize the network structure of YOLOv4 [11]. Compared with the SPP structure, the SRFB structure has more "microstructures" with rich feature information, increasing the receptive field of the model's feature extraction. Each feature extracted by convolution contains extensive feature information that reduces the computational overhead during prediction. The SRFB structure, as shown in Figure 5(a), uses parallel layers with kernel sizes of 3×3 , 1×3 , 1×1 , and 3×1 , each of which will be batch normalized.

During training, the SRFB structure uses parallel convolution layers with kernels of 3×3 , 1×3 , 1×1 , and 3×1 to increase the receptive field of the structure, enhance the model's feature aggregation ability, and deepen the network's expression ability of the nonlinear layer. The SRFB involves batch normalization to reduce network overfitting and speed up training. The batch normalization formula is shown in (5).

$$\mathbf{O}_{:,j} = \left(\sum_{k=1}^C \mathbf{M}_{:,k} * \mathbf{F}_{:,k}^{(j)} - \mu_j \right) \frac{\gamma_j}{\sigma_j} + \beta_j, \quad (5)$$

where $\mathbf{M}_{:,k}$ is the input k -th channel feature map. $\mathbf{F}_{:,k}^{(j)}$ represents the input k -th channel convolution kernel, and $\mathbf{O}_{:,j}$ represents the mapping channel of output features corresponding to the j -th convolution kernel. $\gamma_j, \sigma_j, \beta_j$ denote learnable parameters, which can be obtained by the gradient descent algorithm.

The additivity of convolution proves that two-dimensional convolution kernels with different sizes operate at the same step to produce the same resolution, whose outputs can be added. The additivity of convolution can be considered the addition of the corresponding positions of the convolution kernels to produce an equivalent kernel with the same output, as shown in (6). During prediction, the 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 6.

$$\mathbf{I} * \mathbf{K}^{(1)} + \mathbf{I} * \mathbf{K}^{(2)} = \mathbf{I} * (\mathbf{K}^{(1)} \oplus \mathbf{K}^{(2)}), \quad (6)$$

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, and $*$ represents the two-dimensional convolution operator. Compatibility means that the smaller kernel can be patched to the larger kernel.

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

$$\mathbf{F}'^{(j)} = \frac{\gamma_j}{\sigma_j} \mathbf{F}^{(j)} \oplus \frac{\bar{\gamma}_j}{\bar{\sigma}_j} \bar{\mathbf{F}}^{(j)} \oplus \frac{\hat{\gamma}_j}{\hat{\sigma}_j} \hat{\mathbf{F}}^{(j)}, \quad (7)$$

$$b_j = \frac{\mu_j \gamma_j}{\sigma_j} + \frac{\bar{\mu}_j \bar{\gamma}_j}{\bar{\sigma}_j} + \frac{\hat{\mu}_j \hat{\gamma}_j}{\hat{\sigma}_j} + \beta_j + \bar{\beta}_j + \hat{\beta}_j. \quad (8)$$

By adding the parallel convolution kernel to the asymmetric convolution kernel, the three normalized 3×3 , 1×1 , 1×3 , and 3×1 convolutional branches are merged into the standard convolutional layer. This new structure can obtain rich feature information without the additional computational overhead. The result after the merging is

$$\mathbf{O}_{:,j} + \hat{\mathbf{O}}_{:,j} + \bar{\mathbf{O}}_{:,j} + \tilde{\mathbf{O}}_{:,j} = \sum_{k=1}^C \mathbf{M}_{:,k} * \mathbf{F}_{:,k}^{(j)} + b_j, \quad (9)$$

where $\mathbf{O}_{:,j}$, $\hat{\mathbf{O}}_{:,j}$, $\bar{\mathbf{O}}_{:,j}$, and $\tilde{\mathbf{O}}_{:,j}$ represents the output results of 3×3 , 1×3 , 3×1 , and 1×1 convolutional layer, respectively.

However, the kernel of the SRFB structure can be equivalently converted when it implements inference, as shown in Figure 5(b). The kernel uses different calculations to obtain the gradient because the SRFB structures are randomly initialized during training, so they cannot be converted equivalently.

3.4. Bag of Freebies. Generally, strategies that only increase training cost but does not increase inference loss are called "bag of freebies" in the object detection field. Bag of freebies mainly optimizes the loss function to make the model better fit the data. An image may have thousands of objects in the object detection field, but only a small part needs to be detected. Compared with the two-stage detector, the one-stage detector does not use a region proposal network, which will result in imbalanced distribution of positive and negative samples during training, and the loss value of the object detection susceptible to the loss value of the negative samples. Lin et al. [14] proposed that focal loss could be obtained by modifying the cross-entropy loss function to reduce the negative sample influence. In this paper, focal loss optimized the classification loss function of YOLOv4 to decrease the background influence when recognizing Zhuang pattern symbols. Focal loss decides the total loss

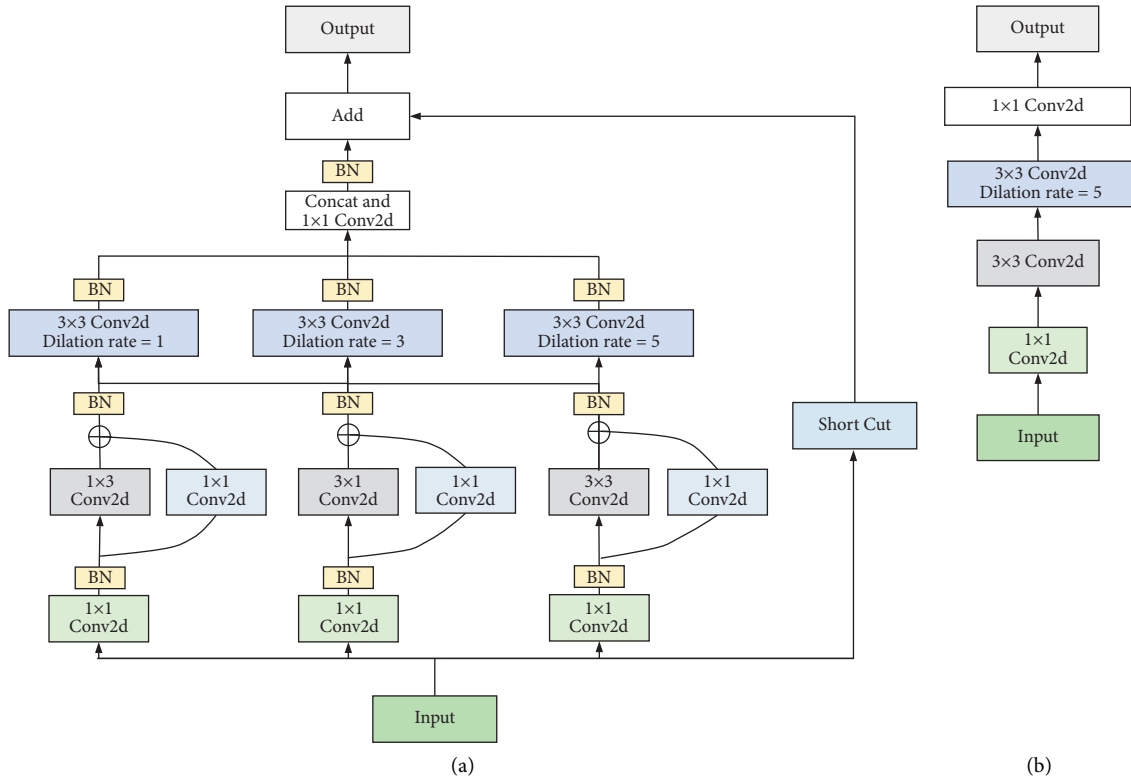


FIGURE 5: (a) is the SRFB during training and (b) the SRFB during prediction. According to the additivity of convolution, the SRFB can treat 1×1 , 3×1 , and 3×1 convolution (consisting of a large number of zeros in the convolution kernel) as a special 3×3 convolution.

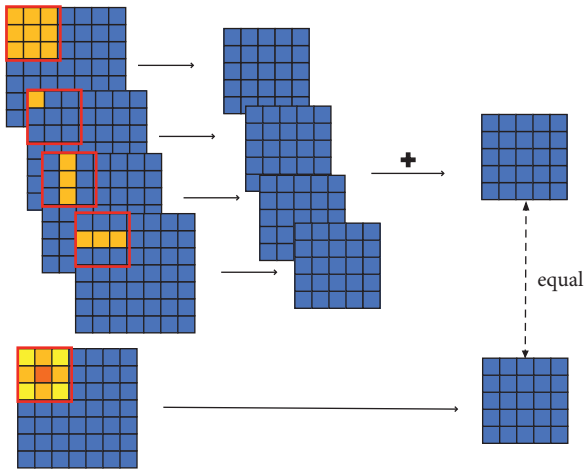


FIGURE 6: The sliding window shows the additivity of convolution. There are four convolutional kernels of 3×3 , 1×1 , 1×3 , and 3×1 , respectively, sharing the sliding windows based on the additivity of convolution.

function by setting weights to the cross-entropy loss function, as shown in (10), thus solving the unbalanced distribution of positive and negative samples and easy and complex samples. Focal loss defines a weight factor $\alpha \in [0, 1]$ and then takes it to the cross-entropy loss function to solve the imbalance of negative and positive samples. When the number of positive samples is small, the value of α will be large, and the loss value of positive samples will increase.

Focal loss suggests an adjustment factor γ to reduce the weight of easy samples and make the model focus on training complex samples to solve the imbalance between easy and complex samples.

$$\text{Focal_loss}(p, y) = \begin{cases} -\alpha(1-p)^{\gamma} \log_a p, & y = 1, \\ -(1-\alpha)p^{\gamma} \log_a(1-p), & y = 0, \end{cases} \quad (10)$$

where $p \in [0, 1]$ represents the classification probability of predicted samples and y indicates the label of positive and negative samples. If y is 0, it is negative sample, and y is 1, it is positive sample.

At present, many object detection models [36, 37] generally use L1 and L2 norms to calculate the loss value. The L1 and L2 norms independently calculate the loss value of the four coordinate variables of the prediction frame. The coordinate variables are irrelevant, but there is some correlation among the coordinate variables in real situations. When the model's performance 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 values of the evaluation indexes will also be affected. Yu et al. [38] proposed that IoU as a regression loss function could calculate the coordinate frame, which solved the above problems. However, if IoU is directly used as the boundary loss when the prediction frame and the ground-truth frame do not overlap, both IoU and the gradient will become 0, and the boundary loss cannot be optimized. Rezatofghi et al. [13] proposed the GIoU loss as a boundary loss. It retained the

scale invariance of IoU as a loss function and added the distance between two frames to optimize the loss value, which solved the problem that the gradient was 0 because the prediction frame and the ground-truth frame did not overlap. The calculation of GIoU is

$$\text{GIoU} = \frac{|A \cap B|}{|A \cup B|} - \frac{|C/(A \cup B)|}{|C|}, \quad (11)$$

where A and B are the prediction frame and the ground-truth frame, respectively, and C is the smallest closed frame containing both. When GIoU becomes larger, the GIoU loss will become smaller, and the network will be optimized to make the prediction frame and the ground-truth frame highly overlap. The boundary loss function of YOLOv4 optimized by GIoU is shown in the following formula:

$$\text{bbox_loss} = \sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{\text{obj}} (1 - \text{GIoU}) \times \left[2 - \left(\hat{w}_i^j \times \hat{h}_i^j \right) \right], \quad (12)$$

where \hat{w} and \hat{h} represent the width and height of the boundary frame, respectively. I^{obj} represents the probability of the object inside the current boundary frame.

In the Earf-YOLO model, the prediction results include the prediction category, confidence, and position of each prediction frame. Therefore, the model's loss function in this paper is

$$\begin{aligned} \text{Loss} = & \lambda_{\text{coord}} b \text{ box.loss} \\ & + \sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{\text{obj}} \left[\hat{C}_i^j \log_a(C_i^j) + (1 - \hat{C}_i^j) \log_a(1 - C_i^j) \right] \\ & - \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B I_{ij}^{\text{noobj}} \left[\hat{C}_i^j \log_a(C_i^j) + (1 - \hat{C}_i^j) \log_a(1 - C_i^j) \right] \\ & - \sum_{i=0}^{S^2} I_{ij}^{\text{obj}} \sum_{c \in \text{classes}} \left((1 - \alpha) P_i^j \log_a(1 - P_i^j) + \alpha (1 - P_i^j)^y \log_a P_i^j \right), \end{aligned} \quad (13)$$

where S^2 indicate $S \times S$ grids, B indicates that each grid has a B prediction frame. If the IoU of the j -th prediction frame and truth-ground frame in the i -th grid is greater than the threshold, then $I_{ij}^{\text{obj}} = 1$, otherwise $I_{ij}^{\text{obj}} = 0$. If the IoU of the j -th prediction frame and truth-ground frame in the i -th grid is less than the threshold, then $I_{ij}^{\text{noobj}} = 1$, otherwise $I_{ij}^{\text{noobj}} = 0$. If the IoU of the j -th prediction frame and truth-ground frame in the i -th grid is the greatest, then $\hat{C}_i^j = 1$, otherwise $\hat{C}_i^j = 0$. C_i^j is the confidence score of the existing object in the j -th prediction frame of the current i -th grid. λ_{coord} and λ_{noobj} represent the penalty weights of the loss function.

3.5. Bag of Specials. Postprocessing is a method to screen the prediction results of models, which belongs to the bag of specials. It can significantly improve the model's prediction accuracy only by adding a small prediction overhead. The postprocessing method uses the NMS algorithm on the output result to delete the wrong prediction frame and find the most appropriate

position for the prediction frame. The Hard-NMS algorithm sorts the prediction frames from high scores to low scores, selects the prediction frame with the highest scores, sets a threshold, deletes the prediction frames whose overlap rates with the highest-scored prediction frames exceed the threshold, and repeats the steps mentioned above with the left prediction frames until the last one. When the overlap rate of two objects in the image is larger than the fixed threshold, the Hard-NMS will set the score of the prediction frame as 0 and then delete it, which may lead to the low-scored objects not being detected and loss of accuracy.

The Soft-NMS [12] addresses the problem that the Hard-NMS mistakenly deletes the prediction frame when two objects overlap from a new perspective. As formula (14) indicates, the Soft-NMS does not delete low-scored prediction frames directly. It will lower their scores further and then set a threshold to delete low-scored prediction frames. The Soft-NMS will also use the Gaussian weight function (as shown in formula (15)) to multiply the scores of the current prediction frame with a weight function. This function will attenuate the scores of adjacent prediction frames that overlap the highest-scored prediction frame M . The more overlapping the prediction frame is with the highest-scored one, the more serious the attenuation of the prediction frame will be.

$$s_i = \begin{cases} s_i, & \text{IoU}(M, b_i) < N_t, \\ s_i(1 - iou(M, b_i)), & \text{IoU}(M, b_i) \geq N_t, \end{cases} \quad (14)$$

$$s_i = s_i e^{-\left(iou(M, b_i)^2 / \sigma\right)}, \quad (15)$$

where s_i is the score of the current prediction frame, N_t is the threshold, and M is the prediction frame with the highest scores. b_i is the score of each prediction frame.

4. Experimental Results and Analysis

In this section, we introduce the experimental datasets and the parameters of the experimental settings, and then verify the performance of Earf-YOLO in experiments.

4.1. Zhuang Pattern Symbol Datasets. The datasets used in the experiment are symbols of Zhuang patterns. 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 represent 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 lives in the forest.

So far, there are no specific dataset composed of symbols of Zhuang patterns. The datasets used in this research are images taken by researchers in the Zhuang tribes. There are about 19,199 images of Zhuang patterns in the datasets, divided into 20 classifications. To ensure the justice of the model when it gets trained, we try to balance the number of images in each classification. We selected 10,592 images as

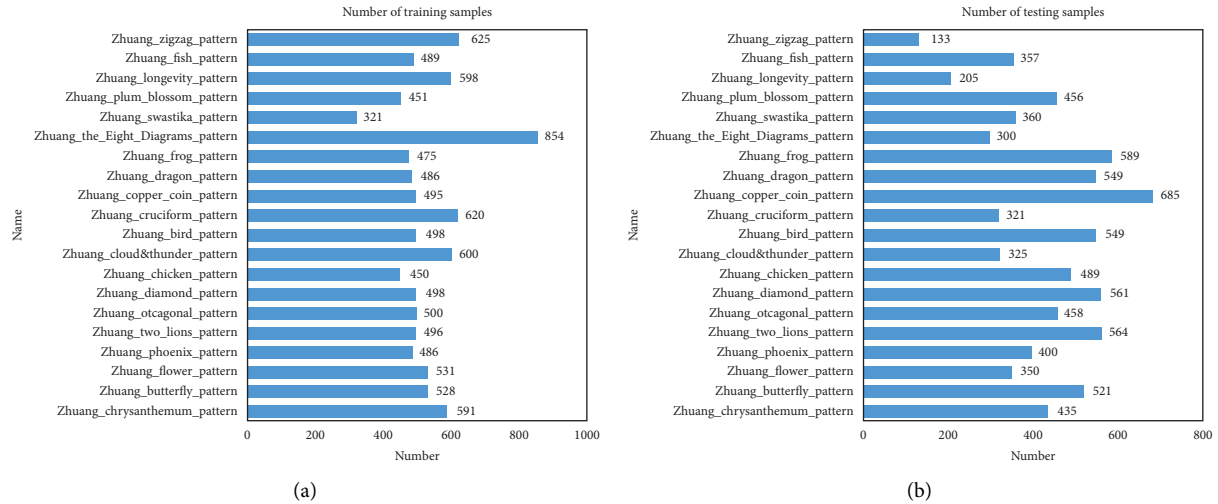


FIGURE 7: Datasets of Zhuang pattern symbols; (a) represents the training samples and (b) represents the testing samples.

training samples and 8,607 images as testing samples. The sample distribution is shown in Figure 7.

4.2. Experimental Settings. Experimental data are trained on Python 3.6, Keras 2.3.1, GTX 2070 8G, and Windows 10. The number of training iterations is 500. The image input size is fixed as 416×416 . The optimizer is Adam. The attenuation strategy of the learning rate is the Cosine annealing attenuation strategy, whose initial learning rate is set as 0.001, the highest learning rate as 0.01, and the lowest learning rate as 0.0001. In the first 400 experiments, the first 170 convolution layers of the network are frozen; then, the remaining convolution layers are trained. In the last 100 experiments, all convolution layers are opened and trained. Average precision (AP), frames per second (FPS), and Param are taken as evaluation indexes to evaluate the model's performance. AP stands for the average accuracy of IoU from 0.5 to 0.95, with the threshold increasing at intervals of 0.05. AP_{50} indicates the average accuracy when the IoU threshold is 0.5, and AP_{75} indicates the average accuracy when the IoU threshold is 0.75. AP_S , AP_M , and AP_L represent the average accuracy of small, medium, and large objects, respectively. The AP is proportional to the model detection effect. The larger the AP is, the better the detection effect will be. The larger the FPS is, the higher the detection efficiency will be. The smaller the Param is, the lower the network memory consumption will be.

4.3. Zhuang Nationality Pattern Symbols Contrast Experiment and Result Visualization. In this section, the suggested Earf-YOLO is evaluated on the datasets of Zhuang pattern symbols. In order to simplify the comparison results on the datasets of Zhuang pattern symbols, we compare the improved Earf-YOLO model with the latest one-stage and two-stage models on the network with ResNet101 and CSPDarkNet53 as backbones, respectively, as shown in Table 1. Table 1 indicates that the AP of Earf-YOLO on ResNet101 and CSPDarkNet53 reached 39.1% and 41.0%,

respectively. Figure 1 and Table 1 show that the Earf-YOLO model achieves the best results in both speed and accuracy compared to other models.

In addition, in the testing set of Zhuang pattern symbols, the experiment was conducted between Earf-YOLO (with CSPDarkNet53 as the backbone) and the original YOLOv4 model to compare classification accuracy, whose results are shown in Figure 8 demonstrate that the average classification accuracy of Earf-YOLO is higher than that of YOLOv4. Besides, for some small and complex pattern symbols such as the Zhuang two lions pattern, the Zhuang copper coin pattern, and the Zhuang bird pattern, the average classification accuracy of Earf-YOLO remains high.

Meanwhile, some images are randomly selected from the datasets of Zhuang pattern symbols for visualization. This paper selects four pairs of representative detection results for comparison. Figure 9(a) shows the visualized result of YOLOv4, and Figure 9(b) shows the visualized result of Earf-YOLO (with CSPDarknet as the backbone). The visualized results suggest that Earf-YOLO is more accurate in detecting complex and small pattern symbol frames.

4.4. Contrast Experiments on PASCAL VOC Dataset. In the previous section, the evaluation indexes of the Earf-YOLO model were acquired on the Zhuang pattern symbol dataset, which does not prove the general efficiency of the model. Therefore, we conduct experiments on PASCAL VOC2007 and VOC2012, the public dataset, to further verify the model's efficiency. The VOC dataset is composed of 20 categories. The dataset is annotated with the actual label position and the corresponding category information for each image. On PASCAL VOC2007 and VOC2012, we compared the proposed model with other advanced object detection models. The experimental results are shown in Table 2. Compared with YOLOv1, YOLOv2, YOLOv3, and YOLOv4, the AP of Earf-YOLO increases by 25.3%, 13.1%, 12.4%, and 2.8%, respectively. Compared with Faster RCNN, RefineDet512, and R-FCN-3000, the AP of Earf-YOLO

TABLE 1: Comparison of AP of Earf-YOLO with that of other latest models on the datasets of Zhuang pattern symbols.

Methods	The backbone network	AP(%)	AP ₅₀ (%)	AP ₇₅ (%)	AP _S (%)	AP _M (%)	AP _L (%)
YOLOv4 640(baseline) [11]	ResNet-101	36.8	56.2	39.2	20.4	39.8	46.4
YOLOv5s8	ResNet-101	37.7	56.8	40.4	21.1	40.8	47.6
Libra RetinaNet [39]	ResNet-101	36.5	55.7	39.1	21.0	40.6	46.3
RetinaNet w/AugFPN [40]	ResNet-101	37.2	55.4	40.2	20.2	40.3	47.1
Our model	ResNet-101	39.1	58.5	41.7	22.3	42.3	48.9
RetinaNet w/SABL [41]	CSPDarkNet53	36.3	59.2	39.1	18.3	39.1	48.3
LRF [42]	CSPDarkNet53	38.3	60.4	41.8	20.2	41.2	50.3
Faster RCNN [43]	CSPDarkNet53	37.4	58.6	39.8	19.8	41.9	50.2
YOLOv4 640 [11]	CSPDarkNet53	38.2	58.6	40.9	21.3	42.6	48.3
RDSNet [44]	CSPDarkNet53	40.6	59.4	41.9	21.1	41.7	50.4
YOLOX [45]	CSPDarkNet53	40.7	59.9	42.5	22.5	43.2	51.6
Our model	CSPDarkNet53	41.0	61.7	43.8	24.3	44.4	51.8

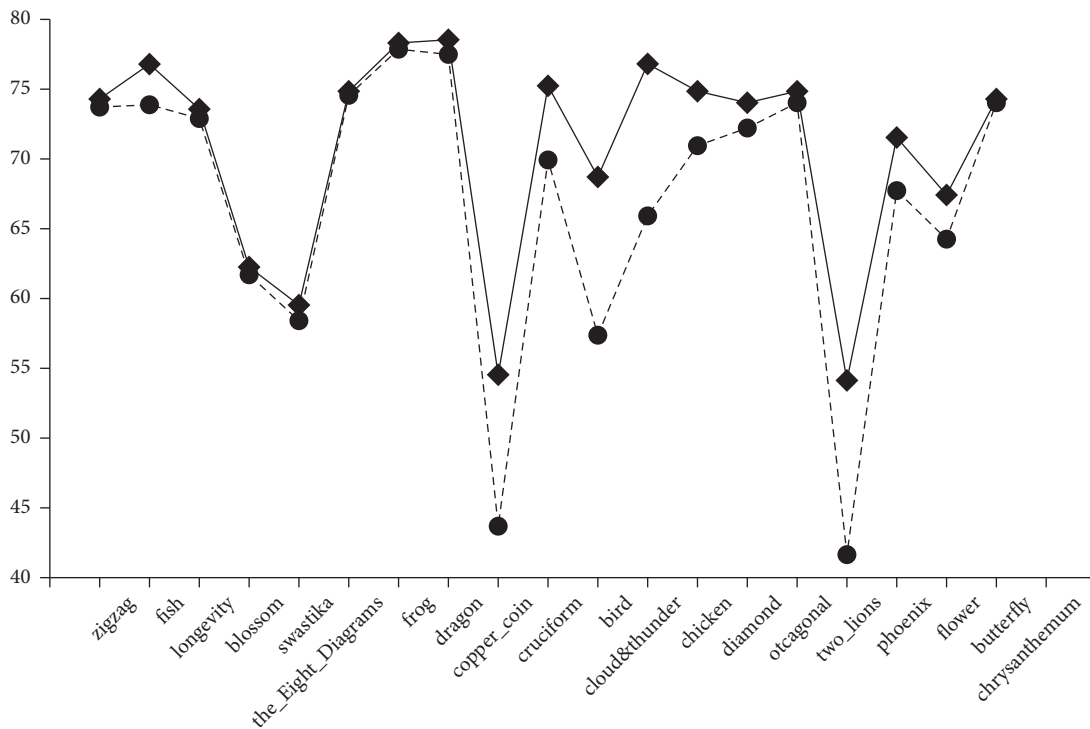


FIGURE 8: Comparison of the average classification accuracy of Earf-YOLO and the baseline model for 20 types of objects in the Zhuang pattern symbol datasets.

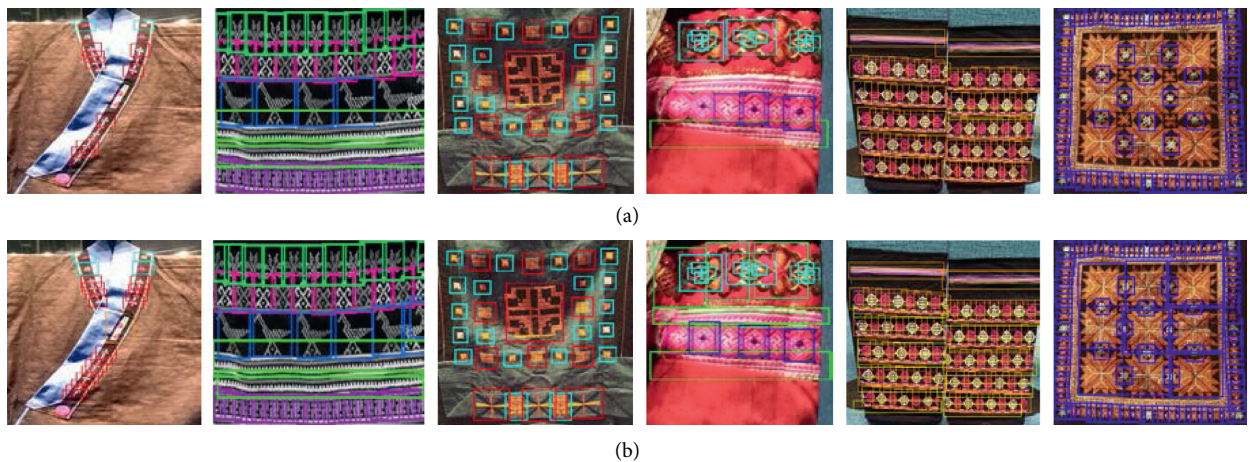


FIGURE 9: Visualized results of YOLOv4 and Earf-YOLO; (a) of YOLOv4 and (b) Earf-YOLO.

TABLE 2: Experimental results of different models on the VOC2007 and VOC2012 testing sets.

Methods	The backbone network	AP(%)
YOLOv1 448	VGG16	66.4
YOLOv2 544	Darknet19	78.6
YOLOv3 416	Darknet53	79.3
YOLOv4 640	CSPDarknet53	88.9
YOLOv5s	CSPDarknet53	85.7
Faster RCNN	ResNet-101	76.4
RefineDet512 [46]	VGfotG-16	80.1
R-FCN-3000 [47]	ResNet-101	80.5
DES512 [48]	VGG-16	80.3
DSSD [49]	ResNet-101	81.5
ASSD [50]	ResNet-101	83.0
Earf-YOLO	CSPDarknet53	91.7

increases by 15.3%, 11.6%, and 11.2%, respectively. Compared with DES512, DSSD, and ASSD, the AP of Earf-YOLO increases by 11.4%, 10.2%, and 8.7%, respectively. It can be seen from Table 2 that Earf-YOLO has the best performance, which illustrates the general efficiency of Earf-YOLO on other datasets.

4.5. Ablation Experiments. All ablation experiments in this section are first conducted on the Zhuang pattern symbol dataset. The experimental results are compared with the baseline, the YOLOv4 algorithm, with the backbone network of CSPDarkNet53. Finally, we integrate GLocalT and SRFB, the main contribution points of this article, into YOLOv3, YOLOv4, and YOLOv5 for comparative experiments on PASCAL VOC2007 and VOC2012.

The performance analysis of global-local-transformer (GLocalT) and strength receptive field block (SRFB) are discussed.

4.5.1. Baseline + GLocalT. As shown in line 2 of Table 3, compared with the baseline model, though FPS of YOLOv4 with GLocalT decreases by 2 and its Param increases by 2.093 M, its AP increases by 1.6%, demonstrating that YOLOv4 with GLocalT can better extract features with complex and small pattern symbols.

4.5.2. Baseline + SRFB. As illustrated in line 3 of Table 3, compared with the baseline model, the AP of YOLOv4 with SRFB increases 0.7%, its Param decreases by 6.148 M, and its FPS increases by 10, which proves that replacing redundant convolution with SRFB can improve the recognition accuracy, significantly reduce the computational overhead, and improve computational efficiency.

4.5.3. Baseline + SRFB + GLocalT. As illustrated in line 4 of Table 3, the AP, Param, and FPS of YOLOv4 with SRFB and GLocalT reach 40.1%, 57.623 M, and 27, respectively. The results demonstrate that YOLOv4 with SRFB and GLocalT reaches the highest performance, as the combination can address the problem that it is difficult to extract and fuse

TABLE 3: Comparative results of models with different additional structures.

Methods	AP(%)	Param(M)	FPS
Baseline	38.2	61.678	20
Baseline + GLocalT	39.8	63.771	18
Baseline + SRFB	38.9	55.53	30
Baseline + GLocalT + SRFB	40.1	57.623	27

TABLE 4: Comparison of the results of YOLOv4 with different techniques.

Hard-NMS	Soft-NMS	Focal loss	GIoU loss	AP(%)
				38.2
	√			38.4
√		√		38.8
√			√	38.7
	√	√	√	39.1

TABLE 5: In VOC dataset, the comparative results of YOLOv3, YOLOv4, and YOLOv5 with the main contributions of the proposed model are shown.

Methods	AP(%)
YOLOv3 416	79.3
YOLOv4 640	88.9
YOLOv5s	86.8
YOLOv3+GLocalT + SRFB	82.5
YOLOv4+GLocalT + SRFB	90.1
YOLOv5+GLocalT + SRFB	88.4

multilayer features thoroughly with less computational overhead.

Then the detection results of YOLOv4 integrated with some techniques are compared, shown in Table 4. When soft-NMS is involved, AP increases by 0.2% compared with YOLOv4 because when two object frames are close to each other, soft-NMS will not directly delete the frame with a large overlap area between the prediction frame and the ground-truth frame but decrease the score of the prediction frame. When focal loss is involved, AP increases by 0.6% compared with YOLOv4, which alleviates the imbalance of positive and negative samples and that of simple and complex samples. When GIoU loss as boundary loss is involved, AP reaches 38.7%. Finally, the model achieves the highest AP when the bag of freebies and the bag of specials are combined.

To prove the general efficiency of our main contributions, we incorporate GLocalT and SRFB into YOLOv3, YOLOv4, and YOLOv5 on the VOC dataset. Table 5 shows that with GLocalT and SRFB, the AP values of YOLOv3 and YOLOv5 increase. With GLocalT and SRFB, the AP value of YOLOv4 reaches the highest, indicating that the optimization based on YOLOv4 can detect more complex scenes.

5. Conclusions

Since present object detection models cannot fully extract features in different stages, and their computational overheads are too high in recognizing symbols of the Zhuang patterns, an object detection model, Earf-YOLO, is

suggested in this paper. To be specific, first we propose the global-local-transformer structure. This structure uses gating axial attention layer to make the model better collect features on height, width, and channel axes, and more sensitive to the position information. It also uses the global-local training strategy to help the model focus on the global dependence between features and reduce local information loss. Then we design the strength receptive field block (SRFB), which uses the dilated convolution of multiscale branches to enhance the model's feature extraction ability and fuses the convolution branches to reduce the inference time. Finally, we incorporate some advanced techniques to optimize the model. The structures and techniques mentioned above effectively address the problems YOLOv4 faces and improve its detection performance in recognizing Zhuang pattern symbols. However, Earf-YOLO has not been widely applied to the two-stage object detection models. Its detection effect on other datasets has not been discussed, either. Therefore, further improvement of the proposed structures and techniques will be the future focus, making them applicable to other two-stage object detection models and available to various datasets.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest with this study.

Acknowledgments

This work is supported by the General Project of Guangxi Natural Science Foundation (2019GXNSFAA245053), the Guangxi Science and Technology Major Project (AA19254016), the National Natural Science Foundation of China (61862018), the Guangxi Natural Science Foundation Project (2018GXNSFAA138084), the Beihai city science and technology planning project (202082033), the Beihai city science and technology planning project (202082023), the Translation and Introduction of Guangxi Marine Culture under the Strategy of Maritime Power (2021KY0184), and the Guang xi graduate student innovation project(-YCSW2021174). The author would like to thank Peng Xie from Southwest Jiaotong University and Jiaqi Xu from the University of Science and Technology of China for their insights and feedback on the first draft of this article.

References

- [1] Q. Kong, Z. Shi, Y. Feng et al., "Classification method of ethnic minority patterns based on faster R-CNN," *Journal of Physics: Conference Series*, vol. 1575, no. 1, Article ID 012137, 2020.
- [2] C. Sun, W. Zhan, J. She, and Z. Yangyang, "Object detection from the video taken by drone via convolutional neural networks," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4806359, 10 pages, 2020.
- [3] S. Zhang, Y. Wu, C. Men, R. Ning, and L. Xiaosong, "Channel compression optimization oriented bus passenger object detection," *Mathematical Problems in Engineering*, vol. 2020, Article ID 3278235, 11 pages, 2020.
- [4] T. Shi, M. Liu, Y. Niu et al., "Underwater targets detection and classification in complex scenes based on an improved YOLOv3 algorithm," *Journal of Electronic Imaging*, vol. 29, no. 4, p. 043013, 2020.
- [5] M. A. I. Mahmoud and H. Ren, "Forest fire detection using a rule-based image processing algorithm and temporal variation," *Mathematical Problems in Engineering*, vol. 2018, Article ID 7612487, 8 pages, 2018.
- [6] P. Huo, Y. Wang, and Q. Liu, "A Part-Based and Feature Fusion Method for Clothing classification," in *Proceedings of the Pacific Rim Conference on Multimedia*, pp. 231–241, Springer, Xian, China, September 2016.
- [7] Y. H. Sun and Q. J. Liu, "Attribute recognition from clothing using a faster R-CNN based multitask network," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 2, Article ID 1840009, 2018.
- [8] J. Glenn, S. Alex, L. C. JirkaBorovec, and P. Rai, *ultralytics/yolov5: v3.1 - bug fixes and performance improvements (version v3.1)*. Zenodo, 2020, <https://github.com/ultralytics/yolov5/tree/v3.1>.
- [9] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, HI, USA, July 2017.
- [10] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, <https://arxiv.org/abs/1804.02767>.
- [11] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, <https://arxiv.org/abs/2004.10934>.
- [12] N. Bodla, B. Singh, R. Chellappa, and S. D. Larry, "Soft-NMS--improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, pp. 5561–5569, Venice, Italy, October 2017.
- [13] H. Rezatofighi, N. Tsoi, J. Y. Gwak, S. Amir, R. Ian, and S. Silvio, "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, CA, USA, June 2019.
- [14] T. Y. Lin, P. Goyal, R. Girshick, H. Kaiming, and D. Piotr, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, Venice, Italy, October 2017.
- [15] F. D. S. Ribeiro, L. Gong, F. Calivá, G. Kjartan, S. Mark, and M. Yu, "An end-to-end deep neural architecture for optical character verification and recognition in retail food packaging," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2376–2380, IEEE, Athens, Greece, October 2018.
- [16] A. Nguyen, D. Kanoulas, D. G. Caldwell, and G. T. Nikos, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5908–5915, IEEE, Vancouver, British Columbia, Canada, September 2017.
- [17] D. Erhan, C. Szegedy, A. Toshev, and A. Dragomir, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147–2154, Columbus, OH, USA, June 2014.
- [18] R. Girshick, J. Donahue, T. Darrell, and M. Jitendra, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on*

- computer vision and pattern recognition, pp. 580–587, Columbus, OH, USA, June 2014.
- [19] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, December 2015.
- [20] S. Ren, K. He, and R. Girshick, “Faster r-cnn: towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.
- [21] K. He, G. Gkioxari, P. Dollár, and G. Ross, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, Italy, October 2017.
- [22] Y. Li, Z. Han, H. Xu, L. Liu, X. Li, and K. Zhang, “YOLOv3-lite: a lightweight crack detection network for aircraft structure based on depthwise separable convolutions,” *Applied Sciences*, vol. 9, no. 18, p. 3781, 2019.
- [23] Q. Zhou, J. Wang, J. Liu, S. Li, W. Ou, and X. Jin, “RSANet: towards real-time object detection with residual semantic-guided attention feature pyramid network,” *Mobile Networks and Applications*, vol. 26, no. 1, pp. 77–87, 2021.
- [24] V. John and S. Mita, “Deep feature-level sensor fusion using skip connections for real-time object detection in autonomous driving,” *Electronics*, vol. 10, no. 4, p. 424, 2021.
- [25] C. Y. Wang, H. Y. M. Liao, Y. H. Wu et al., “CSPNet: a new backbone that can enhance learning capability of CNN,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 390–391, Washington D. C., USA, June 2020.
- [26] K. Wang, J. H. Liew, Y. Zou, Z. Daquan, and F. Jiashi, “Panet: few-shot image semantic segmentation with prototype alignment,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9197–9206, Seoul, South Korea, October 2019.
- [27] T. Y. Lin, P. Dollár, R. Girshick, H. Kaiming, H. Bharath, and B. Serge, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, HI, USA, July 2017.
- [28] A. Vaswani, N. Shazeer, N. Parmar, J. Feng, and D. Zhou, “Attention is all you need,” in *Proceedings of the Advances in neural information processing systems*, pp. 5998–6008, CA, USA, December 2017.
- [29] Q. Qin, J. Yan, Q. Wang, X. Wang, M. Li, and Y. Wang, “ETDNet: an efficient transformer d model,” *IEEE Access*, vol. 9, Article ID 119893, 2021.
- [30] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and M. P. Vishal, “Medical transformer: Gating axial-attention for medical image segmentation,” 2021, <https://arxiv.org/abs/2102.10662>.
- [31] J. Ho, N. Kalchbrenner, D. Weissenborn, and S. Tim, “Axial attention in multidimensional transformers,” 2019, <https://arxiv.org/abs/1912.12180>.
- [32] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, “Axial-deeplab: stand-alone axial-attention for panoptic segmentation,” in *Proceedings of the European Conference on Computer Vision*, pp. 108–126, Springer, Glasgow, United Kingdom, August 2020.
- [33] Y. Ren, C. Zhu, and S. Xiao, “Object detection based on fast/faster RCNN employing fully convolutional architectures,” *Mathematical Problems in Engineering*, vol. 2018, Article ID 3598316, 7 pages, 2018.
- [34] B. Y. Chen, Y. K. Shen, and K. Sun, “Research on object detection algorithm based on multilayer information fusion,” *Mathematical Problems in Engineering*, vol. 2020, Article ID 9076857, 13 pages, 2020.
- [35] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [36] H. Huang, D. Sun, R. Wang, Z. Chun, and L. Bangquan, “Ship target detection based on improved YOLO network,” *Mathematical Problems in Engineering*, vol. 2020, Article ID 6402149, 10 pages, 2020.
- [37] Z. Zhao, J. Han, and L. Song, “YOLO-highway: an improved highway center marking detection model for unmanned aerial vehicle autonomous flight,” *Mathematical Problems in Engineering*, vol. 2021, Article ID 1205153, 14 pages, 2021.
- [38] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and H. Thomas, “Unitbox: an advanced object detection network,” in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, Amsterdam The Netherlands, October 2016.
- [39] J. Pang, K. Chen, J. Shi, H. Feng, D. Lin, and O. Wanli, “Libra r-cnn: towards balanced learning for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 821–830, CA, USA, June 2019.
- [40] C. Guo, B. Fan, Q. Zhang, C. Pan, and S. Xiang, “Augfpn: improving multi-scale feature learning for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Article ID 12604, Seattle, WA, USA, June 2020.
- [41] J. Wang, W. Zhang, Y. Cao et al., “Side-aware boundary localization for more precise object detection,” in *Proceedings of the European Conference on Computer Vision*, pp. 403–419, Springer, Glasgow, UK, August 2020.
- [42] T. Wang, R. M. Anwer, H. Cholakkal, S. K. Fahad, P. Yanwei, and S. Ling, “Learning rich features at high-speed for single-shot object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1971–1980, Seoul, South Korea, October 2019.
- [43] Y. Cao, K. Chen, C. C. Loy, and D. Lin, “Prime sample attention in object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Article ID 11591, Seattle, Washington, USA, June 2020.
- [44] S. Wang, Y. Gong, J. Xing, L. Huang, C. Huang, and W. Hu, “Rdnet: a new deep architecture for reciprocal object detection and instance segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12208–12215, NY, USA, February 2020.
- [45] Z. Ge, S. Liu, F. Wang, L. Zeming, and S. Jian, “Yolox: exceeding yolo series in 2021,” 2021, <https://arxiv.org/abs/2107.08430>.
- [46] S. Zhang, L. Wen, X. Bian, L. Zhen, and Z. L. Stan, “Single-shot refinement neural network for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, Salt Lake City, Utah, USA, June 2018.
- [47] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: object detection via region-based fully convolutional networks,” in *Proceedings of the Advances in neural information processing systems*, pp. 379–387, Barcelona, Spain, December 2016.
- [48] Z. Zhang, S. Qiao, C. Xie, S. Wei, W. Bo, and L. Y. Alan, “Single-shot object detection with enriched semantics,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5813–5821, Salt Lake City, Utah, USA, June 2018.
- [49] C. Y. Fu, W. Liu, A. Ranga, T. Ambrish, and C. B. Alexander, “Dssd: deconvolutional single shot detector,” 2017, <https://arxiv.org/abs/1701.06659>.
- [50] J. Yi, P. Wu, and D. N. Metaxas, “ASSD: attentive single shot multibox detector,” *Computer Vision and Image Understanding*, vol. 189, Article ID 102827, 2019.