2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

BGE-YOLO: An Improved YOLOv8 for Chinese Handwritten Text Detection

Rui Xiong[1], Shiyu Li[2], Xiuwei Yang[1], and Zhiwu Liao[1] (✉)

[1] Sichuan Normal University School of Computer Science, Chengdu 610101, China
20060097@sicnu.edu.cn
[2] Sichuan Normal University School of Mathematical Sciences, Chengdu 610066, China

Abstract. Handwritten text detection is a crucial step in converting handwritten text images into editable text. However, in practical applications, text detection still faces numerous challenges, including the complexity of environmental backgrounds, diversity of target scales, and the contact between complex characters. To address these challenges, this paper proposes a BGE-YOLO model for handwritten text detection. Firstly, a new feature fusion module is designed to achieve bidirectional information flow through cross-scale connections and rapid planning, ensuring effective integration of features across multiple scales. On this basis, a Global Attention Mechanism (GAM) is incorporated, which reduces information loss and amplifies the interaction of global dimensional features, enabling the model to extract meaningful information in complex backgrounds. Additionally, the incorporated Multi-Scale Attention (EMA) module utilizes a novel cross-spatial learning approach, enhancing the interaction of local features and further improving feature fusion efficiency. Furthermore, a data augmentation strategy enriches the self-constructed handwritten text image dataset, further improving the model's generalization ability. Experimental results indicate that compared to the YOLOv8 model, the mAP50 and accuracy P of this model have increased by 2.8% and 3.9%, respectively. This validates the advantages of the BGE-YOLO model in handwritten text detection and facilitates more convenient information extraction from handwritten text.

Keywords: handwritten text, text detection, BGE-YOLO, BiFPN, Global Attention Module, EMA, self-constructed dataset.

1 Introduction

With the advent of the internet era and the widespread use of smart terminal devices, various aspects of people's lives are being recorded and preserved in the form of photographs. In these natural scene images, both printed and handwritten text play significant roles. Printed text is commonly found in road signs, license plates, and product labels. Handwritten text is prevalent in contexts such as handwritten notes, memos, and whiteboards in classrooms and meeting rooms. Thus, Chinese Handwritten Character Detection (CHCD) has become a topic of great interest to researchers [1]. CHCD has a wide range of application scenarios in various business systems, including document digitization, automated logistics, signature verification, and medical services. However, as shown in Fig. 1, several factors significantly increase the difficulty of character segmentation. These include the diversity of handwriting styles, complex character overlaps, variations in character shapes, and varying image backgrounds. These factors adversely affect text detection and recognition performance [2]. Therefore, proposing a method that can accurately and efficiently locate the positions of handwritten text has become an important research topic today.

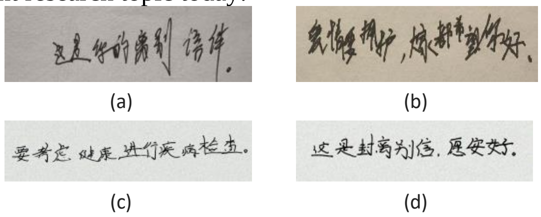

(a) (b)
(c) (d)

Fig. 1. Chinese handwritten text in various backgrounds

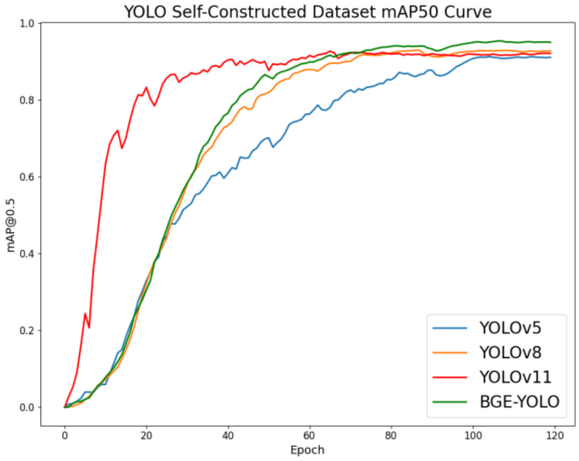


Fig. 2. The mAP@0.5 variation curves for YOLOv5, YOLOv8, YOLOv11, and the model proposed in this paper under the condition of no pre-trained weights.

To tackle these challenges, we propose an improved Chinese handwritten text detection algorithm based on YOLOv8, referred to as BGE-YOLO. This method introduces modules such as GAM, BiFPN, and EMA to enhance the backbone and neck layers of the original YOLOv8. These improvements not only facilitate the fusion of attention features across channel and spatial dimensions, enhancing the model's ability to extract features from handwritten Chinese text, but also achieve a better balance between network efficiency and performance. Moreover, they accelerate the convergence of model loss, enabling the model to produce more accurate localization results. Finally, a data augmentation strategy enriches the handwritten text image dataset, further improving the model's generalization ability. Our proposed model achieves outstanding results on self-constructed dataset, as shown in Fig. 2. The contributions of this paper are as follows:

- Considering the limited availability of line-level Chinese handwritten text datasets with bounding boxes and character category labels, we independently created and annotated a dataset containing 3,000 images. This dataset includes the handwriting styles of 20 individuals and covers 240 commonly used Chinese characters.
- To enhance the model's ability to extract and fuse effective features, we introduced the GAM module to reduce information loss in complex backgrounds, improved YOLOv8's feature fusion strategy to better integrate shallow and deep features for small-target detection, and incorporated EMA-S and EMA-M modules to enhance feature representation across different receptive fields while maintaining computational efficiency.
- While improving the accuracy of text detection, we explored an approach based on YOLO's detection and classification framework to efficiently detect and predict specific handwritten Chinese characters under the constraints of a small dataset and a limited set of character classes.

2 Related Work

2.1 Traditional Research Methods

Previous research on Chinese handwritten text detection can be broadly categorized into traditional approaches and computer vision-based methods. Traditional approaches primarily include two methods: template matching [3] and feature extraction. For example, early researchers compared handwritten characters with predefined character templates and selected the template with the highest similarity as the detection result [4]. Traditional feature extraction methods achieve character detection by segmenting the text into individual characters and extracting features such as stroke width and direction [5]. However, these methods rely on predefined templates based on manually classified handwritten samples, making it difficult to adapt to variations in handwriting styles and layouts. As a result, computer vision methods capable of accurately and efficiently locating text positions have gradually gained widespread application.

2.2 Traditional Machine Learning Methods

Research in computer vision can primarily be divided into traditional machine learning methods and deep learning methods. Traditional machine learning methods include Hidden Markov Models (HMM) [6], Support Vector Machines (SVM) [7], K-Nearest

Neighbors (K-NN) [8], among others. Xiao et al. [9] utilized an SVM-based video text detection method using color, edge, and HOG features, achieving high recall and accuracy rates in both single-frame and three-frame text detection. Xia et al. [10] adopted a discrete HMM to detect spam SMS by utilizing word sequence information, achieving a precision of 89.20% and a recall rate of 81.60%. However, traditional machine learning methods are computationally expensive and inefficient for large-scale data, limiting their practical applicability. In contrast, deep learning-based methods have been widely applied in Chinese handwritten text detection.

2.3 Two-Stage Deep Learning Methods

Deep learning methods can be primarily categorized into two-stage deep learning methods and single-stage deep learning methods. In handwritten text detection, two-stage deep learning methods typically involve two primary steps: firstly, detecting text regions (such as characters, words, or lines); secondly, performing regression on these detected regions. Text region detection methods primarily encompass the R-CNN series of algorithms, including the original R-CNN [11], Faster R-CNN [12], and Mask R-CNN [13], among others. Joseph Raj et al. [14] utilized Faster R-CNN to extract potential text regions and employed a simple yet effective classifier for prediction, achieving an F1 score of 0.70 on the MSRA-TD500 dataset. Qiu et al. [15] proposed a method that over-segments text line images into character segments to generate candidates, constructing a segmentation-recognition grid. Recognition scores and semantic context are then used to determine the optimal path for character segmentation and recognition. Although two-stage deep learning methods have achieved promising results in handwritten text detection, they still face certain challenges, including high computational resource demands [16], which limit their application on resource-constrained devices. Additionally, these methods exhibit high computational complexity and processing time [17], making it difficult to meet the requirements of real-time applications. Thus, single-stage deep learning methods have gradually gained widespread recognition and have become a crucial approach for handwritten text detection.

2.4 Single-Stage Deep Learning Methods

Single-stage deep learning methods primarily include the YOLO [18] series, SSD [19], RetinaNet [20], and others. Compared to Faster R-CNN, the YOLO model adopts a single-stage object detection architecture, which does not include a Region Proposal Network (RPN) [21]. Wang et al. [22] proposed a new text detection method called R-YOLO, which is a powerful real-time convolutional neural network (CNN) model capable of effectively detecting text in any orientation, achieving an F-measure score of 82.3%. Wang et al. [23] designed a new feature fusion structure based on an improved YOLOv3 algorithm, achieving a defect detection performance of 89.2%. Subsequently, YOLOv5 [24] has further optimized the YOLO series, showing significant improvements over YOLOv3 [25] in detection accuracy, network structure compactness, and the versatility of application scenarios. Thereafter, YOLOv7 [26] optimized convolutional operation speed and model size, achieving further performance enhancements.

Khan et al. [27] proposed an Arabic text detection model based on YOLOv7, achieving an accuracy of 91.30%, which is an improvement of 2.80% over the YOLOv5 model. This method achieves high precision while requiring low computational resources, making it a promising solution for text detection applications.

Subsequently, YOLOv8 [28] has made significant strides in detection accuracy and speed. However, it still overlooks issues related to the classification and localization loss of small targets. In summary, although previous researchers have made positive contributions in this field, their studies have not sufficiently addressed issues such as background complexity and target scale diversity. Therefore, to tackle these challenges, this paper proposes a novel BGE-YOLO model for handwritten text detection.

3 Proposed Method

The improved network framework BGE-YOLO is illustrated in Fig. 3. Specifically, in the feature extraction module, we introduced the GAM module, which enhances the interaction of global-dimensional features to reduce the loss of critical information. This method can enhance the model's ability to detect targets in complex backgrounds. In the feature fusion module, we replaced the Concat structure in YOLOv8 with BiFPN-Concat2 and BiFPN-Concat3. By employing a bidirectional cross-scale connection structure, the method enhances the transfer and fusion of semantic information between feature maps of different scales and network levels, enabling the network to better focus on regions containing small and medium-sized targets. To supervise the recognition of different targets with varying receptive fields, we integrated the EMA-S and EMA-M modules before the small-scale and medium-scale detection heads, respectively. This module combines channel and spatial information using a multi-scale parallel sub-network structure, focusing on preserving information from each channel while reducing computational overhead, thereby further enhancing fusion efficiency. These modules enhance the extraction and fusion of key information, making the model more accurate in detecting and recognizing small and medium-sized targets. This optimization improves the model's ability to handle handwritten text detection tasks, especially in effectively addressing varying text sizes and complex background scenarios.

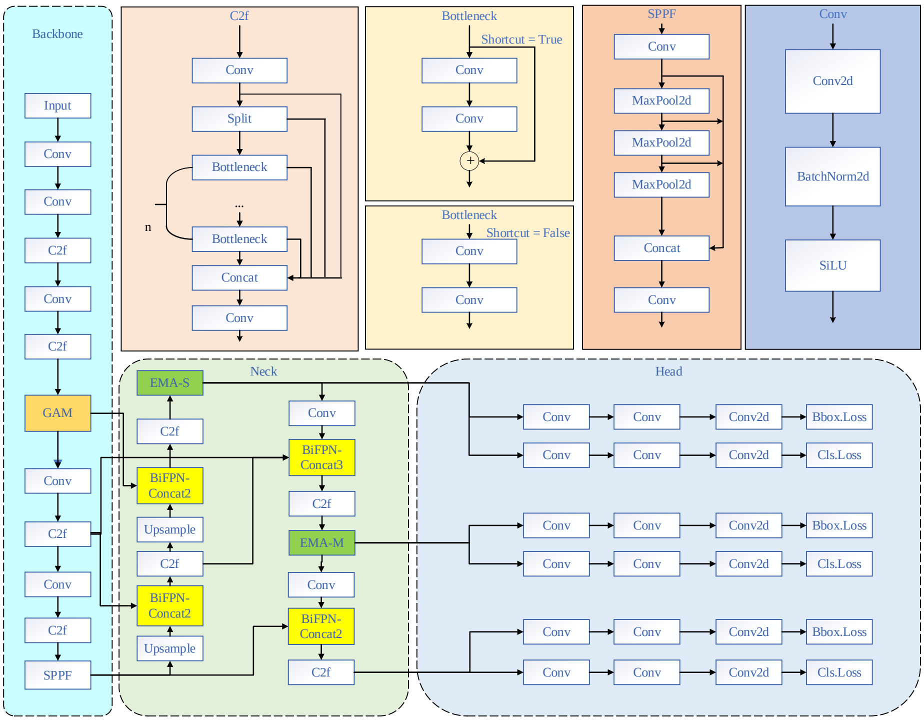


Fig. 3. BGE-YOLO Network Structure

3.1 Bidirectional Feature Pyramid Network

In the process of handwritten text detection, challenges such as the diversity of text sizes and shapes, as well as complex character overlaps, pose significant difficulties. YOLOv8 adopts a strategy that combines the Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) to integrate multi-scale feature information. This approach increases the model's parameter count and computational burden while failing to fully account for the weight differences between various input features. Therefore, this study introduces the Bidirectional Feature Pyramid Network (BiFPN) [29], as shown in Fig. 4(b). The network achieves dynamic feature fusion through learnable weights, rather than relying on simple concatenation or addition. This strategy allows the network to adaptively adjust the contribution of different features based on task-specific requirements, thereby significantly improving overall performance. Moreover, BiFPN introduces skip connections between the initial input and output nodes. This approach allows the fusion of more features without significantly increasing computational cost.

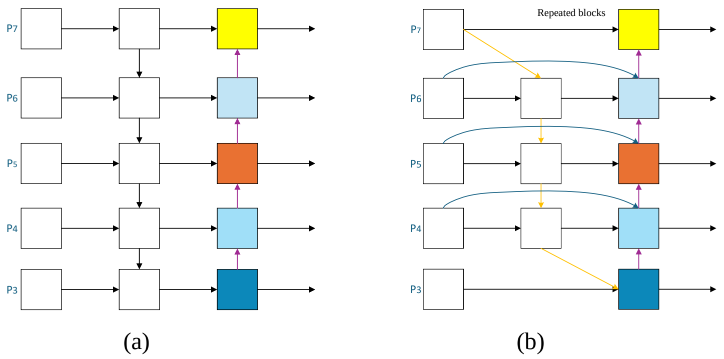


Fig. 4. (a)Structure of PANet (b)Structure of BiFPN

To address the varying contributions of input features at different resolutions to handwritten text detection, this study adopts a rapid normalized fusion method with additional weights. This method performs feature fusion by dividing each weight by the sum of all accumulated weights, enabling the network to recognize the importance of each feature layer. As a result, this strategy significantly improves the performance of handwritten text detection tasks. The specific calculation process is shown in the following equation:

$$O=\sum_i \frac{w_i}{\varepsilon+\sum_j w_j}\cdot I_i \tag{1}$$

In the formula, O represents the output feature, I_i denotes the input feature, and w_i indicates the node weight. It is important to note that the learning rate ε is set to 0.0001 to prevent the occurrence of unstable results.

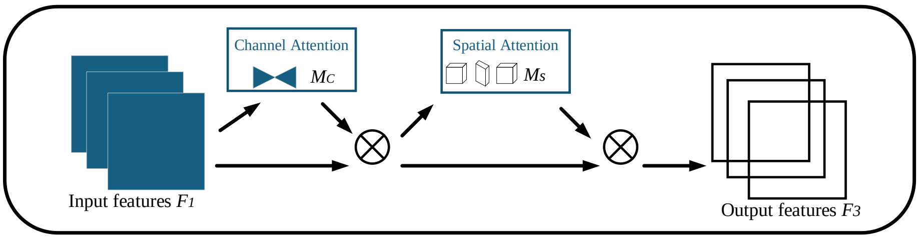


Fig. 5. Structure of the GAM module, which utilizes channel and spatial attention mechanisms to process the final output of the given feature map.

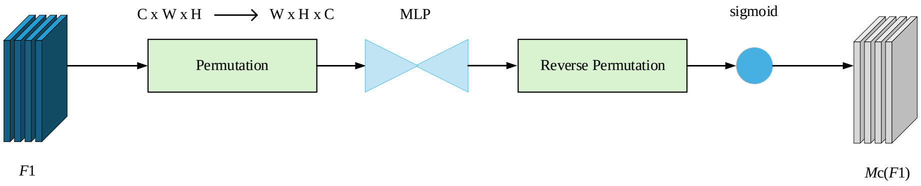


Fig. 6. Structure of the channel attention module

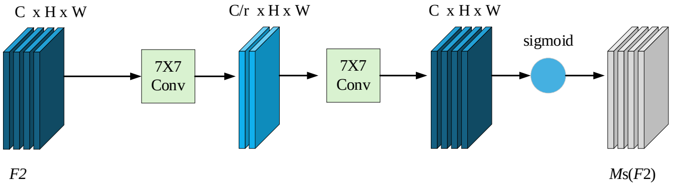


Fig. 7. Structure of the spatial attention module

3.2 Global Attention Mechanism

In practical handwritten text detection, various complex detection scenarios are often encountered. The complexity of the background in handwritten text images significantly increases the difficulty of detection. The backbone layer of YOLOv8 employs the C2f module as its basic building block, which features a lower parameter count and excellent feature extraction capabilities. However, while this design enhances feature extraction capabilities, it does not improve the detection of small targets in complex backgrounds. Additionally, the attention in the backbone layer is typically focused on two dimensions, potentially failing to fully utilize information across all three dimensions. To improve the detection accuracy of the model, a Global Attention Mechanism (GAM) [30] was introduced into the backbone network module. By effectively capturing critical information across three dimensions, the model's capacity for feature extraction in complex backgrounds has been enhanced. The detailed architecture of this module is shown in Fig. 5.

The GAM module enhances global-dimensional target interaction features, thereby reducing the loss of critical information. It adopts the sequential channel-spatial attention mechanism from CBAM while introducing new channel and spatial attention modules to replace the submodules of the original CBAM model. The structures of the new channel and spatial attention modules are shown in Fig. 6 and Fig. 7. The entire process is illustrated in Eqs. (2) to (3), while the following equations define the intermediate stages and outputs of the specific input feature maps:

$$F_2 = M_C(F_1) \otimes F_1 \tag{2}$$

$$F_3 = M_S(F_2) \otimes F_2 \tag{3}$$

In the formula, M_C and M_S denote the channel attention module and the spatial attention module, respectively; $\otimes$ signifies multiplication; F_1 , F_2 ,and F_3 represent the input features, intermediate features, and output features, respectively.

3.3 EMA Attention Mechanism

In handwritten text detection tasks, due to the small size of the text and significant background interference, missed detections frequently occur. To address this issue, the model needs to possess stronger contextual capture capabilities and enhanced local feature interaction abilities.

The Channel Attention (CA) mechanism learns the correlation between channels through global average pooling and fully connected layers, and it normalizes the attention weights using the softmax function to achieve the fusion of cross-channel and spatial information. However, this method incurs high computational costs, limiting its practical application. The EMA module builds upon this foundation by employing parallel subnetwork blocks to effectively capture cross-dimensional interactions and establish dependencies between different dimensions [31]. Specifically, the EMA integrates more contextual information into the intermediate feature map by concurrently utilizing 3×3 and 1×1 convolutions, thereby achieving multi-channel information fusion and optimizing feature grouping. By introducing this cross-dimensional logical reasoning capability, our model enhances the semantic expression ability between contexts while reducing computational overhead, thereby alleviating the issue of missed detections in handwritten text detection tasks. The structure is shown in Fig. 8.

Assume the input feature is represented as: $\mathrm{x} \in R^{C\times H\times W}$, where H and W denote the height and width of the feature map, and C represents the number of channels in the feature map. The EMA module divides the channels C evenly into G groups， with each group containing C/G channels , while maintaining the spatial dimensions as $H\times W$, i.e., $C//G\times H\times W$. Three parallel computational paths are applied to each group of features. Among these, two paths involve subnetworks with 1 × 1 convolution kernels, which perform 1D global average pooling along the H and W directions. The formula for the pooling operation is as follows:

$$\mathrm{z}_C^H(H) = \frac{1}{W}\sum\nolimits_{0\le i\le W} X_C(H,i) \tag{4}$$

$$\mathrm{z}_C^W(W) = \frac{1}{H}\sum\nolimits_{0\le j\le H} X_C(j,W) \tag{5}$$

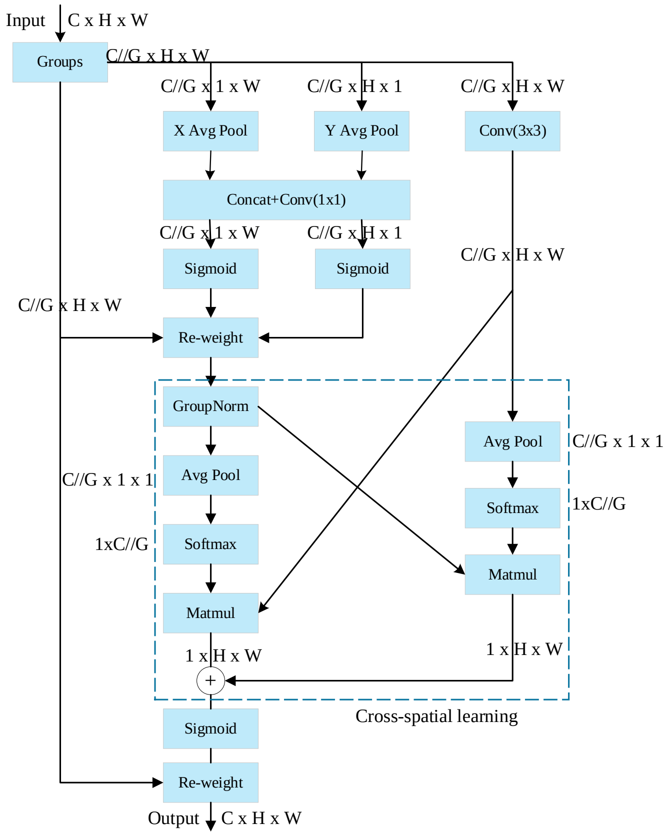


Fig. 8. Structure of the EMA attention mechanism

In Eqs. (4) to (5), X_C represents the input characteristics of channel C. $z_C^H(H)$ and $z_C^W(W)$ represent the pooling results along fixed width and height, respectively, used to capture information from different spatial locations. Afterward, concatenation and group normalization (GN) are applied to perform intra-group statistics, enhancing the ability to extract local features. Next, 2D pooling is conducted to further compress the features. This specific implementation can be represented by Eq. (6).

$$Z_c = \frac{1}{H \times W} \sum_j^H \sum_i^W X_C(i, j) \tag{6}$$

where Z_c represents the result of performing 2D pooling on the features. The features extracted from the subnetworks are aggregated, passed through a Sigmoid activation function, and then element-wise multiplied with the original features. Finally, the enhanced features are reshaped to align with the initial input size, generating the final output.

4 Experiments

4.1 Dataset

The dataset used in this study is a self-constructed dataset for line-level handwritten Chinese character detection, which includes bounding boxes and character category labels. The dataset consists of 3,000 images, encompassing different writing styles from 20 individuals and including 240 common Chinese characters. These characters were randomly selected from the first-level character list of the General Standard Chinese Character Table. Each character is treated as a separate category, and bounding boxes and category labels are annotated for each character. Data augmentation was then performed, including operations such as blurring, noise processing, and image scaling. Finally, the dataset was divided into training, validation, and test sets in a ratio of 7:1:2. As shown in Fig. 9, this dataset includes several challenging scenarios encountered in line-level handwritten text detection, such as complex character interactions, diversity in the size of handwritten characters, and variations in writing styles.

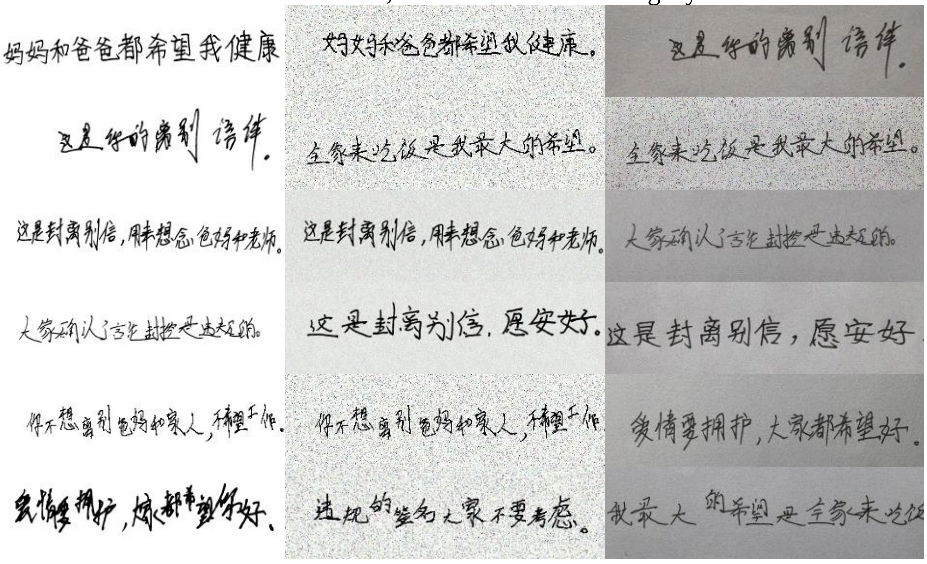


Fig. 9. A subset of the dataset illustrates the irregular handwriting styles of different individuals. Various data processing techniques were applied, including binarization, salt-and-pepper noise, and Gaussian blur.

4.2 Experimental Environment and Evaluation Metrics

Our experimental environment is based on a Linux operating system and utilizes a hardware platform equipped with an NVIDIA Tesla V100 32GB GPU, employing the

PyTorch deep learning framework. Our hardware configuration includes 64GB of DDR4 memory, providing sufficient memory resources for the experiments. During training, we uniformly employed the Adam optimizer, setting the number of epochs to 120 and the batch size to 32, while not utilizing any pre-trained weights.

To comprehensively evaluate the effectiveness of the proposed model in text detection, this study employed four fundamental evaluation metrics: precision (P), recall (R), mean average precision (mAP), and F1 score. The corresponding computational formulas are presented below (refer to Eqs. (7) to (10)):

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{9}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{10}$$

Precision quantifies the proportion of actual positive samples among the predicted positive categories made by the model, while Recall measures the proportion of accurately predicted actual positive samples. Mean Average Precision (mAP) is the average precision calculated at different levels of recall, serving as a comprehensive metric for model evaluation. mAP50 denotes the average precision at an IoU threshold of 0.5, and mAP50:95 denotes the average precision averaged over IoU thresholds from 0.5 to 0.95. F1 score is the harmonic mean of precision and recall, utilized for a comprehensive assessment of the model's accuracy and stability.

4.3 Results

To assess the performance of the proposed BGE-YOLO model, we performed a comparative evaluation against the baseline network YOLOv8 on the same test dataset after training. The specific results can be found in Table 1.

Table 1. Comparison results with the baseline model. The best results in each column are in bold.

Model	Precision(%)	Recall(%)	F1(%)	mAP50(%)	mAP50:95(%)
YOLOv8	84.0	85.5	81.2	92.8	81.7
Ours	**87.9**	**89.1**	**85.0**	**95.6**	**83.8**

The data presented in Table 1 indicates that the BGE-YOLO model exhibits significant improvements over YOLOv8n across several metrics, including accuracy, recall, F1 score, and mAP50, with increases of 3.9%, 3.6%, 3.8%, and 2.8%, respectively. The proposed model not only enhances detection performance but also improves the ability to perceive targets of varying scales.

4.4 Ablation Experiments

To verify the effectiveness of each module in the proposed BGE-YOLO model, we conducted ablation experiments comparing the performance of different modules on the same test dataset. Detailed results are presented in Table 2.

Table 2. Results of the ablation experiments.

Model	Precision(%)	Recall(%)	F1(%)	mAP50(%)	mAP50:95(%)
YOLOv8	84.0	85.5	81.2	92.8	81.7
+BiFPN	83.9	87.7	83.0	93.9	82.5
+BiFPN+GAM	86.1	86.9	84.2	94.7	83.2
Ours	**87.9**	**89.1**	**85.0**	**95.6**	**83.8**

Table 2 shows that the introduction of BiFPN resulted in a significant increase of 1.1% in the model's mAP50. This improvement can be attributed to the cross-scale connections and efficient planning methods employed by BiFPN, which facilitate bidirectional information flow and significantly enhance the model's ability to fuse features across different scales. Consequently, the model can identify text targets of varying sizes with greater precision.

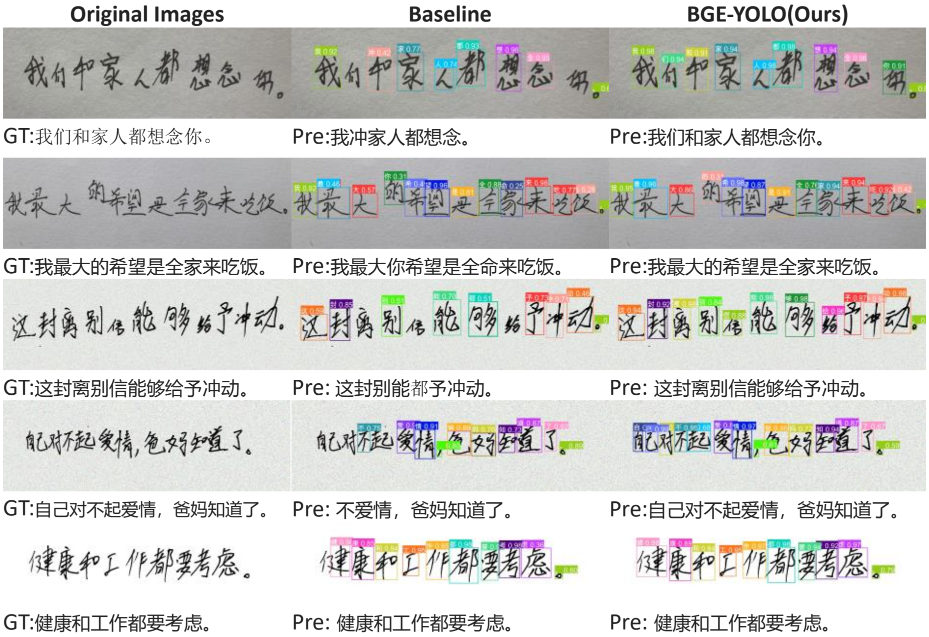
With the introduction of GAM, the model's mAP50 increased significantly by 1.9%. This enhancement can be primarily attributed to GAM's mechanism, which reduces information reduction and amplifies global dimensional interaction features, thereby improving the model's ability to extract effective information in complex backgrounds.

Subsequent to the integration of EMA, the model's mAP50 increased significantly by 2.8%. This enhancement is attributed to EMA's cross-space learning approach and multi-scale parallel subnetworks that establish dependencies. Consequently, the model's computational efficiency and local feature interaction capabilities improved, leading to better fusion efficiency.

As shown in Fig. 10, this study presents the detection results of the improved algorithm on line-level handwritten text images of specific Chinese characters. The model effectively classifies the detected text targets, identifies the text category with the highest predicted probability, and thereby obtains the predicted value for each target. The proposed model exhibits remarkable performance in detecting small text targets, achieving a significant reduction in the missed detection rate. These improvements provide higher precision and reliability for the detection and recognition of line-level handwritten text images, which is anticipated to yield better outcomes in practical applications.

4.5 Comparison Experiments

To validate the performance of the proposed BGE-YOLO model, we conducted comparative verification using the same test dataset against current mainstream models after training, as well as performing comparative experimental analysis between single-stage and two-stage object detection models. The specific results can be found in Table 3.



GT:我们和家人都想念你。 Pre:我冲家人都想念。 Pre:我们和家人都想念你。

GT:我最大的希望是全家来吃饭。 Pre:我最大你希望是全命来吃饭。 Pre:我最大的希望是全家来吃饭。

GT:这封离别信能够给予冲动。 Pre: 这封别能都予冲动。 Pre: 这封离别信能够给予冲动。

GT:自己对不起爱情，爸妈知道了。 Pre: 不爱情，爸妈知道了。 Pre:自己对不起爱情，爸妈知道了。

GT:健康和工作都要考虑。 Pre: 健康和工作都要考虑。 Pre: 健康和工作都要考虑。

Fig. 10. Test results of the improved model. GT represents the ground truth of the text in the image, while Pre denotes the text category with the highest predicted probability.

Table 3. Results of the comparison experiments. The model weight file size was introduced to evaluate model complexity and storage requirements.

Model	Precision (%)	Recall (%)	F1 (%)	mAP50 (%)	mAP50:95 (%)	Weight (MB)
Faster-RCNN[12]	80.10	82.00	77.30	87.60	71.20	172.12
SSD[19]	80.50	83.20	77.80	88.20	71.80	160.47
RetinaNet[20]	81.20	83.10	78.10	89.10	73.20	153.53
YOLOv3-tiny[25]	82.90	83.40	79.40	90.20	74.80	17.22
DBNet[32]	82.50	84.30	79.80	90.60	74.50	136.51
YOLOv5[24]	83.30	84.60	80.00	91.50	77.30	14.35
YOLOv7-tiny[26]	83.00	83.90	79.80	91.00	76.90	12.39
DBNet++[33]	84.70	87.10	83.40	93.40	81.10	118.62
YOLOv11[34]	84.80	85.30	81.30	92.70	80.40	5.36
Ours	**87.90**	**89.10**	**85.00**	**95.60**	**83.80**	**7.84**

Table 3 shows that the detection accuracy of the Faster R-CNN model is relatively low, primarily due to its reliance on anchor boxes generated from single-scale feature maps. Compared to other two-stage detection models, this characteristic reduces its efficiency in detecting multi-scale objects. As the network depth increases, the resolution of the feature maps gradually decreases, further weakening the SSD and RetinaNet models' ability to detect small-scale targets, such as text. In contrast, YOLOv3-Tiny maintains a lightweight model while utilizing three feature maps of different scales for bounding box predictions, thereby enhancing the localization accuracy for text detection. Although DBNet demonstrates stable performance in end-to-end text detection tasks, it exhibits certain limitations in feature fusion mechanisms, which constrain its ability to accurately detect small-scale text instances. As an improved version of YOLOv3, YOLOv5 optimizes the loss function, allowing the model to focus more on learning critical features during the training phase. Although YOLOv7-Tiny achieves better lightweight performance through a combination of ELAN and MaxPool2d structures, its information extraction capability diminishes accordingly. DBNet++ enhances detection accuracy by refining the feature fusion mechanism of DBNet and optimizing the post-processing steps. However, the relatively complex network architecture still presents challenges in terms of deployment efficiency. YOLOv11 builds upon YOLOv8 by introducing a more simplified network architecture and improved training strategies, further optimizing the model's lightweight characteristics, resulting in a reduction of the weight file size to 5.36MB. However, while achieving a lightweight model, YOLOv11 inevitably results in some loss of accuracy, particularly in the current task, where its performance is not as strong as that of YOLOv8.

In contrast, the proposed BGE-YOLO model significantly enhances detection performance by strengthening feature fusion and introducing a more effective small target detection module. This not only improves the ability to extract relevant information but also optimizes feature fusion across different scales. Although the size of the model's weight file has slightly increased, the improved model demonstrates superior performance in accuracy and small target detection capabilities. This outcome substantiates the efficacy of the optimization strategies employed.

5 Conclusion

Handwritten text detection is essential for converting handwritten text images into editable text. This study presents BGE-YOLO, an innovative deep learning approach designed to address challenges encountered in text detection, such as varying scales and complex backgrounds. The significant advantages of this model largely arise from improvements in the BiFPN architecture, which enhance its capability to integrate information across different scales. Additionally, the newly introduced GAM mechanism enables the model to focus more on extracting key features relevant to text detection, thereby minimizing the risk of extracting irrelevant information in complex backgrounds. Meanwhile, the new EMA module effectively strengthens the interaction of local features, improving fusion efficiency. Experimental results indicate that the BGE-YOLO model achieves improvements of 2.8% in mAP50 and 3.9% in precision compared to the YOLOv8 model.

However, there are still certain limitations in the current study, particularly regarding the underrepresentation of cluttered backgrounds and low-light conditions in the custom dataset. Future research will focus on enhancing the efficiency of text detection and recognition, with an emphasis on improving the ability to detect and recognize large amounts of text, as well as exploring its potential applications in end-user devices.

Acknowledgments. This research was funded by the Criminal Inspection Sichuan Key Laboratory of Colleges and Universities 2023 (grant number: 2023ZD01) and the Sichuan Science and Technology Program (grant number: 24GJHZ0388).

References

1. Chen, Z., Yin, F., Zhang, X.Y., Yang, Q., Liu, C.L.: MuLTReNets: multilingual text recognition networks for simultaneous script identification and handwriting recognition. Pattern Recognit 108, 107555 (2020)
2. Wang, Y., Yang, Y., Ding, W., Li, S.: A residual-attention offline handwritten Chinese text recognition based on fully convolutional neural networks. IEEE Access 9, 132301–132310 (2021)
3. Shen, L., Chen, B., Wei, J., Xu, H., Tang, S.-K., Mirri, S.: The challenges of recognizing offline handwritten Chinese: a technical review. Appl. Sci. 13(6), 3500 (2023)
4. Sun, Y., Mao, X., Hong, S., Xu, W., Gui, G.: Template matching-based method for intelligent invoice information identification. IEEE Access 7, 28392–28401 (2019)
5. Zoizou, A., Zarghili, A., Chaker, I.: A new hybrid method for Arabic multi-font text segmentation and a reference corpus construction. Journal of King Saud University - Computer and Information Sciences 32, 576–582 (2020)
6. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
7. Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. 13(4), 18–28 (2008)
8. Peterson, L.E.: K-nearest neighbor. Scholarpedia 4(2) (2009)
9. Xiao, B., Zhao, J., Zhao, C., Ma, J.: Video text detection based on multi-feature fusion. Journal of Intelligent & Fuzzy Systems 37(2), 2125–2136 (2019)
10. Xia, T., Chen, X.: A discrete hidden Markov model for SMS spam detection. Applied Sciences 10(14), 5011 (2020)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: 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. IEEE, New York (2014)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 28, pp. 91–99. Curran Associates, Red Hook (2015)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969. IEEE, New York (2017)

14. Joseph Raj, A.N., Junmin, C., Nersisson, R., Mahesh, V.G.V., Zhuang, Z.: Bilingual text detection from natural scene images using faster R-CNN and extended histogram of oriented gradients. Pattern Analysis and Applications 25(4), 1001–1013 (2022)
15. Wang, Q.-F., Yin, F., Liu, C.-L.: Handwritten Chinese text recognition by integrating multiple contexts. IEEE transactions on pattern analysis and machine intelligence 34(8), 1469–1481 (2012)
16. Gao, Y., Chen, Y., Wang, J., Lu, H.: Semi-supervised scene text recognition. IEEE Transactions on Image Processing 30, 3005–3016 (2021)
17. Redmon, J., Farhadi, A.: YOLO9000: Better faster stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7263–7271. IEEE, New York (2017)
18. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. IEEE, New York (2016)
19. Liu, W., et al.: SSD: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 21–37. Springer, Berlin (2016)
20. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988. IEEE, New York (2017)
21. Schönfelder, P., Stebel, F., Andreou, N., König, M.: Deep learning-based text detection and recognition on architectural floor plans. Automation in Construction 157, 105156 (2024)
22. Wang, X., Zheng, S., Zhang, C., Li, R., Gui, L.: R-YOLO: A real-time text detector for natural scenes with arbitrary rotation. Sensors 21(3), 888 (2021)
23. Wang, J., Su, S., Wang, W., Chu, C., Jiang, L., Ji, Y.: An Object Detection Model for Paint Surface Detection Based on Improved YOLOv3. Machines 10(4), 261 (2022)
24. Jocher, G., et al.: Ultralytics/YOLOv5: Initial release. Zenodo (2020)
25. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arxiv preprint arxiv:1804.02767 (2018)
26. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7475. IEEE, New York (2023)
27. Khan, N., et al.: Robust Arabic and Pashto Text Detection in Camera-Captured Documents Using Deep Learning Techniques. IEEE Access 11, 135788–135796 (2023)
28. Hussain, M.: YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines 11(7), 677 (2023)
29. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10781–10790. IEEE, New York (2020)
30. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561 (2021)
31. Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., et al.: Efficient multi-scale attention module with cross-spatial learning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, New York (2023)

32. Liao, M., Wan, Z., Yao, C., et al.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11474–11481. AAAI Press, Palo Alto (2020)
33. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE transactions on pattern analysis and machine intelligence *45*(1). 919-931 (2022)
34. Khanam, R., Hussain, M.: YOLOv11: An overview of the key architectural enhancements. arxiv preprint arxiv:2410.17725 (2024)