electronics

MDPI

Article

RA-YOLOv8: An Improved YOLOv8 Seal Text Detection Method

Han Sun [1], Chaohong Tan [2], Si Pang [1], Hancheng Wang [2] and Baohua Huang [1,2,*]

[1] School of Computer and Electronic Information, Guangxi University, Nanning 530004, China; 2213393040@st.gxu.edu.cn (H.S.); 2213393033@st.gxu.edu.cn (S.P.)
[2] Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center, Nanning 530000, China
* Correspondence: bhhuang66@gxu.edu.cn

Abstract: To detect text from electronic seals that have significant background interference, blurring, text overlapping, and curving, an improved YOLOv8 model named RA-YOLOv8 was developed. The model is primarily based on YOLOv8, with optimized structures in its backbone and neck. The receptive-field attention and efficient multi-scale attention (RFEMA) module is introduced in the backbone. The model's ability to extract and integrate local and global features is enhanced by combining the attention on the receptive-field spatial feature of the receptive-field attention and coordinate attention (RFCA) module and the cross-spatial learning of the efficient multi-scale attention (EMA) module. The Alterable Kernel Convolution (AKConv) module is incorporated in the neck, enhancing the model's detection accuracy of curved text by dynamically adjusting the sampling position. Furthermore, to boost the model's detection performance, the original loss function is replaced with the bounding box regression loss function of Minimum Point Distance Intersection over Union (MPDIoU). Experimental results demonstrate that RA-YOLOv8 surpasses YOLOv8 in terms of precision, recall, and F1 value, with improvements of 0.4%, 1.6%, and 1.03%, respectively, validating its effectiveness and utility in seal text detection.

Keywords: YOLOv8; seal text detection; RFEMA; AKConv

Citation: Sun, H.; Tan, C.; Pang, S.; Wang, H.; Huang, B. RA-YOLOv8: An Improved YOLOv8 Seal Text Detection Method. *Electronics* **2024**, *13*, 3001. https://doi.org/10.3390/electronics13153001

Academic Editor: José Carlos Castillo

Received: 1 July 2024
Revised: 18 July 2024
Accepted: 26 July 2024
Published: 30 July 2024

1. Introduction

Text serves as a crucial carrier for information transfer and preservation in the era of information technology. The proliferation of smartphones has led to the accumulation of a large amount of visual data. Efficiently and accurately extracting text information from this massive amount of visual data has become an urgent problem. As a prerequisite for text recognition, text detection plays a crucial role. A superior text detection model can accurately locate text areas, avoid background interference, and provide a solid foundation for subsequent text recognition. As an important means of anti-counterfeiting and anti-tampering, seals are usually stamped on documents or bills. However, since most seals are red or blue, stamping them on black text can make it difficult to extract the seal's text information due to color coverage. Additionally, uneven seal color or mutilations in the seal's outline also pose challenges to text detection. Therefore, achieving an accurate seal text detection model with strong robustness is particularly important.

Numerous text detection techniques have been developed due to the rapid advancement of computer vision technology, including CTPN [1], SegLink [2], EAST [3], and others. CTPN uses a vertical anchor box to position text. CTPN can handle long text by adopting a vertical anchor box, but it is difficult to detect non-horizontal text. SegLink forms a complete line of text by connecting many parts of a line. SegLink introduces a rotation Angle compared to CTPN. This allows SegLink to handle not only long text, but also text with different orientations. But SegLink has trouble detecting crooked text. EAST has a simple structure. This method can predict the text box directly and reduce the complexity of post-processing, but it cannot deal with long text. However, since most seals

contain curved text, these methods often suffer from low accuracy and practicality for such cases. To address the issue of text detection in arbitrary directions, many methods such as TextSnake [4], Inceptext [5], and R2CNN [6] have emerged. TextSnake transforms the text detection problem into a series of circular parameter prediction problems based on the geometric properties of the text area. TextSnake is good at handling curved text, but the algorithm has time-consuming, post-processing operations. Inceptext enhances the detection performance of the model by extracting multi-scale features, but the algorithm has high computational complexity. R2CNN introduces a rotating bounding box to handle text in any direction, but this algorithm requires higher hardware. Although these methods improve the detection performance to some extent, they often struggle to achieve an ideal balance between accuracy and computational efficiency.

To address these problems, an improved YOLOv8 method for seal text detection, called RA-YOLOv8, is proposed. This method introduces modules such as AKConv [7], RFAConv [8], and EMA [9] based on YOLOv8, enhancing the backbone and neck of the original YOLOv8. These improvements not only promote the attentional feature fusion in channel and spatial dimensions, improve the model's ability to extract seal features, but also reduce the computational load and achieve a better balance between network overhead and performance. Additionally, the loss function in YOLOv8 is replaced with the MPDIoU loss function [10], a loss function based on MPDIoU for bounding box regression. By considering the shape differences of the bounding boxes, this loss function minimizes the distance between the corresponding top-left and bottom-right corners of the predicted box and the ground truth box when they have the same aspect ratio but different widths and heights. This not only simplifies the computational process but also makes the bounding box regression more accurate and improves the convergence speed of the model loss, enabling the model to produce better localization results.

The contributions of this paper are as follows:

1. Considering the large number of Chinese character categories and the limited number of existing Chinese seal datasets, a Chinese seal dataset consisting of 7004 images was created, including 2002 blurred real seals and 5002 electronic seals.
2. An improved seal text detection method RA-YOLOv8 is proposed. YOLOv8 backbone and neck are enhanced, with the RFEMA method replacing the original Conv layer in the backbone. The RFEMA method integrates RFAConv receptive field attention convolution operation and the EMA cross-latitude interaction operation, significantly improving the model's accuracy and performance while adding a negligible number of parameters. In the neck, the improved AKConv method replaces the original Conv layer, further enhancing the model's feature extraction efficiency and precision.
3. Replace the loss function with the MPDIoU loss function. The MPDIoU loss function considers the geometric features of the bounding box. This allows the model to better adapt to different scenarios, handle targets, and improve regression performance.

The rest of the paper is organized as follows: Section 2 provides an overview of related work, Section 3 presents detailed information about the proposed method, Section 4 describes the self-constructed dataset and the experiments testing the performance of the proposed method, and Section 5 concludes the paper and discusses plans for future work.

2. Related Work

The seal text detection problem can essentially be regarded as natural scene text detection. Current natural scene text detection methods can be mainly divided into two categories: one applies traditional text detection methods to seal text detection, typically focusing on utilizing the unique features of seals, and the other uses deep learning techniques for text detection.

The traditional text detection method is to extract the outline features of the seal. Gao et al. [11] proposed a seal discrimination method based on stroke edge matching, which identifies the tested seal by comparing the similarity between the edge images of the template seal and those of the tested seal. However, when the outline of the seal is

damaged, the stroke edges may lose key information, leading to a decrease in accuracy in specific cases. Chen et al. [12] proposed an identification approach for valid seal imprints based on the center-rays model, which takes advantage of the geometric properties of the seal. First, the connected component of the seal is segmented using image segmentation methods. Then, the region growth method is used to locate the candidate region. Finally, the topological relationship between the seal frame and the region within the frame is explored using eight rays extending from the center of the model in different directions to extract the seal. However, when the seal outline is mutilated, key information may be lost, resulting in decreased identification accuracy. Cai et al. [13] proposed a method by selecting the contour shape skeleton as a lantern ring, which takes advantage of the fact that most seals are red or blue. The method normalizes the color of colored seals and then extracts the color of the seal to simplify the image. It judges by calculating the Euclidean distance between any two black point pixels on the contour shape skeleton map. Yao et al. [14] took full advantage of the fact that most seals are red and proposed using the red component for seal detection and localization via the HSI color model. However, when the seal is gray and similar to the background color, this method becomes ineffective. Zhang et al. [15] used a diffuse water filling algorithm to process the grayscale image and achieve seal detection through binarization. But the robustness of this algorithm is not high. Kang et al. [16] first extracted the seal ontology using the SN color space model, and then used the adaptive Canny operator for edge detection on the morphologically processed image, further localizing the seal text. When the edge information is destroyed, the detection performance of this method is low.

Based on deep learning, text detection methods can be mainly divided into two types: regression-based text detection algorithms and segmentation-based text detection algorithms. Each type showcases its unique advantages and applicable scenarios in the text detection task. The text detection method based on Regression mainly predicts the coordinates of the bounding box. Methods with a preset anchor box are called indirect regression, while those without a preset anchor box are called direct regression. For detecting horizontal text, Zhong et al. [17] proposed the Inception and RPN (Inception RPN) method. The method uses convolution and max pooling of different sizes to extract text features. The method improves the accuracy of model detection of horizontal text by adjusting the anchor box. Zhong et al. [18] also proposed the Anchor-Free Region Proposal Network (AF-RPN) method, which utilizes the Feature Pyramid Network (FPN) to detect text of different sizes such as large, medium, and small, thereby producing high-quality text region proposals. Then, the sliding window detector is used for classification and regression. This method can detect scene text regions at low resolution, but it cannot deal with extremely small text instances.

In order to detect multi-directional text, Liao et al. [19] proposed a trainable text detection model named TextBoxes++. By using inclined bounding boxes to detect text accurately, the model increases the acceptance area of long text area, and enhances the ability of the model to detect long text. When the character spacing is large or the text is curved, the detection performance of this method will be degraded. Xu et al. [20] created a new Geometry Normalization Module (GNM). The module can normalize the text instances to a geometry range through a single branch. This improvement enables the model to better adapt to changes in text size and orientation, thus improving the accuracy and robustness of detection. This method performs well in conventional text detection. However, the detection performance of this method can be improved greatly under special and complex conditions. He et al. [21] proposed a scene text detection method named MOST. The method first compares the initial detection results with the image features. Then, according to the obtained comparison results, the receptive field of image features is dynamically adjusted, and the final detection result is obtained by refining continuously. This method solves the problem of inaccurate long text detection. MOST is optimized for three problems that exist in EAST. This method may perform well in natural scenes, but may degrade in special cases such as text missing. Wang et al. [22] created a scene text detection method. This

method creates an end-to-end network that reduces the complexity of text detection. This method improves the performance of the model by ranking the similarity of the detected text instances. This method is mainly aimed at the problem of text detection in natural scenes. When the text is fuzzy and cannot be displayed completely, the detection effect of this method may be poor.

To detect curved or arbitrarily directional text, Liu et al. [23] proposed a curved text detector (CTD). The model optimizes the regression module by adding curved locating points. The model first performs offset prediction for width and height, respectively. The bounding box is then corrected by predicting the offsets of 14 points on the proposal bounding box. The coordinates of these 14 points represent the curved shape of the text. Wang et al. [24] proposed a robust method for scene text detection. The method utilizes Long Short-Term Memory (LSTM) to iteratively regress the coordinates of points on the proposal bounding box. This enables the detection of arbitrarily shaped text. In order to improve the detection performance of curved text, Liu et al. [25] proposed a Conditional Spatial Expansion (CSE) method. The model models local features in vertical and horizontal directions, respectively. Then, the model extracts the contour points and performs processes such as Non-Maximum Suppression (NMS). This effectively suppresses similar features and reduces false discriminations. Zhang et al. [26] created a text detection model named LOMO. LOMO model first generates the initial quadrilateral text. Then, the model extracts feature blocks from the obtained initial text and extracts the long text through continuous refinement. Finally, the model comprehensively considers the geometric properties of the text instances. This makes the detection results more accurate. Liu et al. [27] proposed a model that can use parameterized Bezier curves to adaptively fit arbitrarily shaped text. The model can also simplify the detection of scene text and reduce the computational overhead, but cannot detect Chinese text. Zhang et al. [28] proposed an adaptive boundary proposal network. The model first generates the prior information and an initial bounding box through multi-layer convolution. Then, an adaptive boundary deformation model is used to iteratively change the shape of the bounding box. This makes the bounding box constantly fit the text region. Dai et al. [29] created a Progressive Contour Regression (PCR) model. The model first generates an initial horizontal text bounding box by estimating the center and size of the text. It then predicts the corner points of the bounding box. Then, it generates a rotated text box based on the position and semantic information of these points. Finally, the model iterates over the text bounding box. This makes the bounding box fit the shape of the text. However, the final detection result of this method is greatly affected by the number of selected points. Zhu et al. [30] proposed a scene text detection method called TextMountain. This method can be used to locate the text centers by predicting the Text Center-Border Probability (TCBP) and Text Center-Direction (TCD) using the border-center information. This method has no advantage when the word is short.

Segmentation-based text detection algorithms use neural networks to extract features. It then determines whether the pixel belongs to the text region by classifying each pixel in the image. This enables the segmentation of text and background. Deng et al. [31] proposed a scene text detection algorithm called PixelLink. The algorithm segments the text by connecting pixels of the same text instance and then extracts the text bounding box from the segmentation result. However, this method is not accurate in the detection of large objects. Since this method only looks at the relationship between the pixel and its neighbors, it ignores the context information and may cause some false detections in the model. Baek et al. [32] introduced a character-level-based scene text detection method designed to detect long lines of text efficiently. This method uses a convolutional neural network to predict the affinities between characters. But it has limitations in stroke sticking and curved text. Tian et al. [33] developed a model called LSAE. By using Shape-Aware Embedding, the model can distinguish between different text instances and make pixels belonging to the same instance closer. Additionally, a Shape-Aware Loss and a new post-processing operation are introduced to generate more accurate bounding box predictions. Wang et al. [34] proposed a Progressive Scale Expansion Network (PSENet).

The network generates different proportions of kernels for each text instance. This can separate two closely text instances, so that the model can get more accurate detection results. The disadvantage of this method is that the selection of hyperparameters is particularly important for different data sets. Selecting a hyperparameter that is not suitable for the model may affect the detection effect. Xu et al. [35] proposed a text detection method called TextField. This algorithm segments the text region from the background by encoding a binary text mask and direction information through a direction field. However, its detection effectiveness is compromised when the text region is occluded in the image. Liao et al. [36] proposed a Differentiable Binarization (DB) module and a network named DBNet. The network can adaptively set the binarization threshold. This simplifies the post-processing operation and improves the text detection performance. This model does not solve the case of circular text with text inside. Zhu et al. [37] proposed a new Fourier Contour Embedding (FCE) method and created the FCENet network. The network uses Inverse Fourier Transformation (IFT) and Non-Maximum Suppression (NMS) to generate a more accurate detection bounding box for text instances in any direction. This method may not perform as well as expected when dealing with low resolution or blurred images. Because text detail may be insufficient in low-resolution images, this can affect feature extraction. Cai et al. [38] proposed a new arbitrary shape text detection method named DText. This method can dynamically generate convolution kernels for different text instances according to features. This approach overcomes the limitations of fixed convolution kernels, which cannot adapt to all resolutions and prevent information loss across multi-scale instances. However, this method is difficult to deal with sharpened text instances. Zhong et al. [39] created a new Progressive Region Prediction Network (PRPN) with directional pooling. The network first predicts the probability distribution of the text region. Then, the network converts this distribution into a bounding box using a watershed-based, post-processing algorithm. This can achieve the purpose of text detection. The high computational complexity of the directional pooling module in this method leads to a decrease in speed. Yu et al. [40] proposed a text detection method called TCM. The method uses the CLIP model for unsupervised perception of text images. This model applies CLIP model to scene text detection through adaptive learning. The model also incorporates features between different levels of the CLIP model to obtain more accurate text detection results. Shi et al. [41] proposed a scene text detection algorithm based on result fusion. This method synthesizes the results of different text detection algorithms and improves the performance of text detection by using the advantages of these text detection algorithms. However, if the fused algorithm has the wrong detection result, the algorithm will also have the wrong result. Naveen et al. [42] proposed a new text detection method. The method improves accuracy by combining Generative Adversarial Network (GAN) and Network Variational Autoencoder (VAE). The method first generates diverse text regions, then continuously optimizes these text regions, and finally detects these text regions. However, in some cases, the complexity of the model can affect its effectiveness. Zheng et al. [43] proposed a text detection method based on boundary points dynamic optimization (BPDO). The method first extracts the image features. Then, text region and text awareness features are obtained according to the extracted features. Finally, based on the text perception features, the boundary points are iteratively optimized to obtain a complete boundary box.

Many existing text detection methods are optimized for some problems in natural scenes. The text in the natural scene is clear and complete, but the seal text detection will face problems such as text missing or text blurring. These detection methods do not take into account the particularity of seal text. This results in these models often working less well in environments with complex backgrounds and multiple fonts. Especially in the seal text detection, many models do not perform as well as they do in ordinary scene text. When dealing with seal text, these models often have difficulty in distinguishing between text and background. In addition, due to the lack of seal text, it is difficult for the model to extract image features, which makes the model detection accuracy decline. Therefore, it is

crucial to develop a detection model specialized for seal text scenes. To solve the above problems, we proposed RA-YOLOv8. This model can extract the detailed features of the seal text and effectively distinguish the text from the complex background.

3. Method

3.1. YOLOv8 Model Introduction

YOLOv8 is an improvement on YOLOv5. YOLOv8 consists of three main parts, namely the backbone, neck, and head. The backbone mainly extracts features from the input image. It incorporates YOLOv7 ELAN design concept and replaces the original C3 module with the C2F module, which captures more key detail information in complex backgrounds and noise environments. The neck removes 1×1 convolutional to reduce channel layer and replaces the C3 module with the C2F module. The head mainly transforms the fused feature maps into final detection results. The head structure is changed to decoupled head structure, and the Anchor-Based is changed to Anchor-Free. The original structure of YOLOv8 is illustrated in Figure 1.

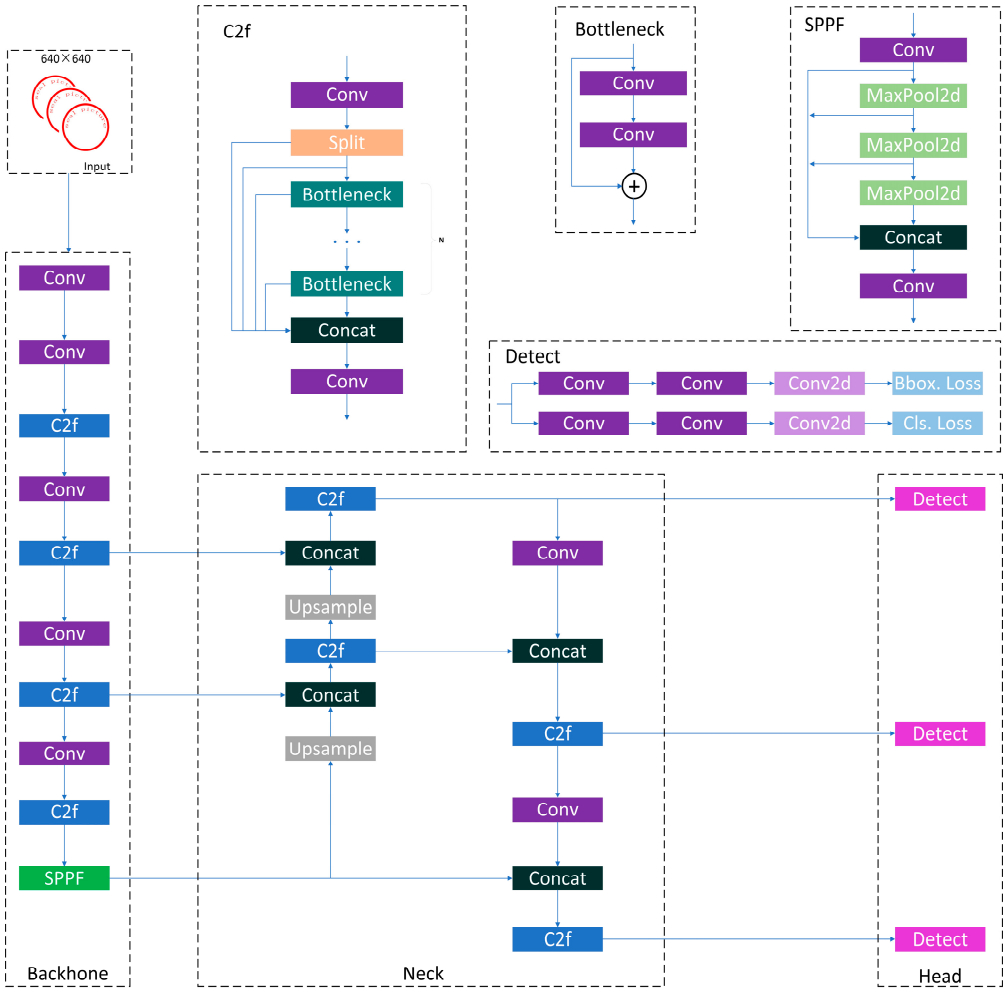


Figure 1. Structure of YOLOv8.

Seals are usually stamped on a variety of documents. However, these documents have various colors of text or graphics, which can obscure the text content on the seal and increase the difficulty of text detection. Additionally, uneven pressure during stamping may lead to blurred seal impressions. In such complex backgrounds, YOLOv8 may struggle to distinguish between the target and the background. Furthermore, when the seal contains densely packed text, leading to reduced character spacing, YOLOv8 may experience detection omissions, resulting in decreased accuracy. YOLOv8 structure is also complex, requiring substantial computational resources and longer training times. To address these issues, specific optimizations have been made to YOLOv8, focusing mainly on improving the backbone, the neck, and the loss function. The Conv layers in the backbone and the neck have been replaced with RFEMA module and AKConv module, respectively, enhancing the model's feature extraction capabilities and detection accuracy in complex backgrounds. The improved YOLOv8 is showed in Figure 2.

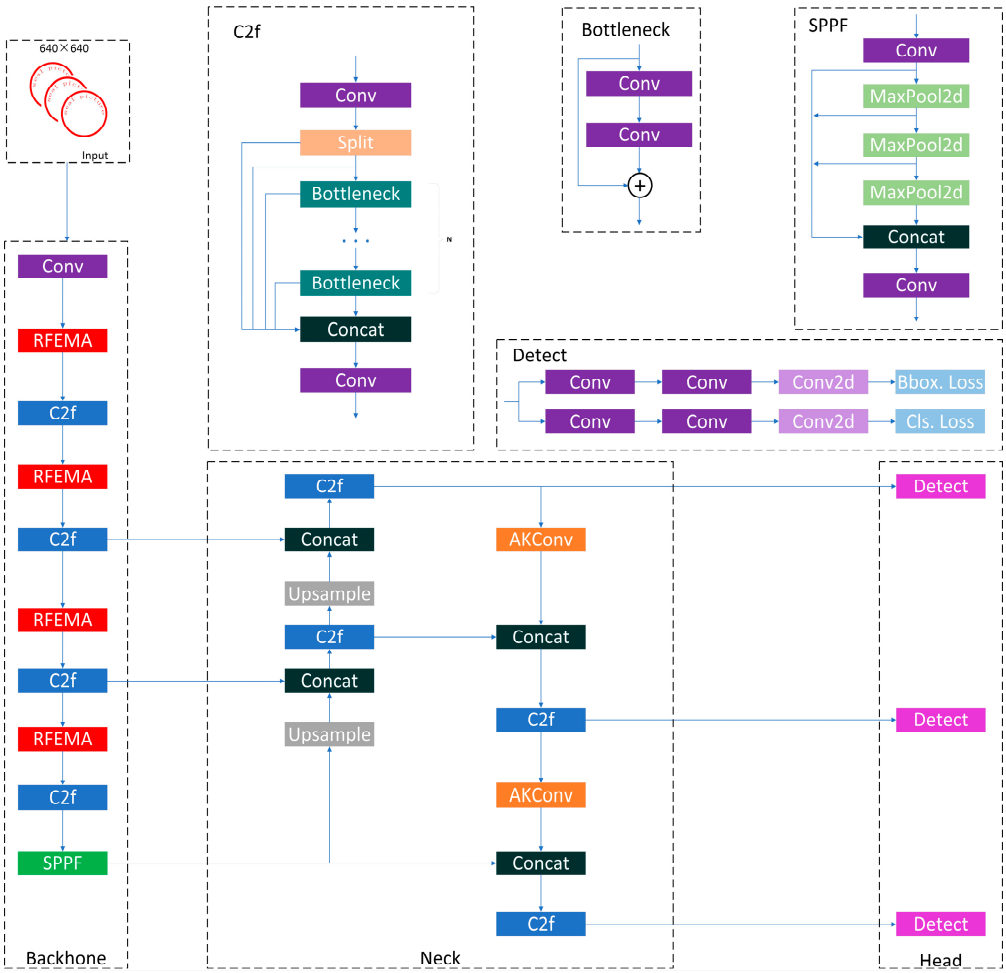


Figure 2. Improved structure of YOLOv8.

3.2. RFEMA Module

RFEMA module is a modification of RFCA module and EMA module. RFEMA module incorporates the feature of both modules and combines them in a new structure by connecting the two modules in series. The structure of RFEMA module is showed in Figure 3.

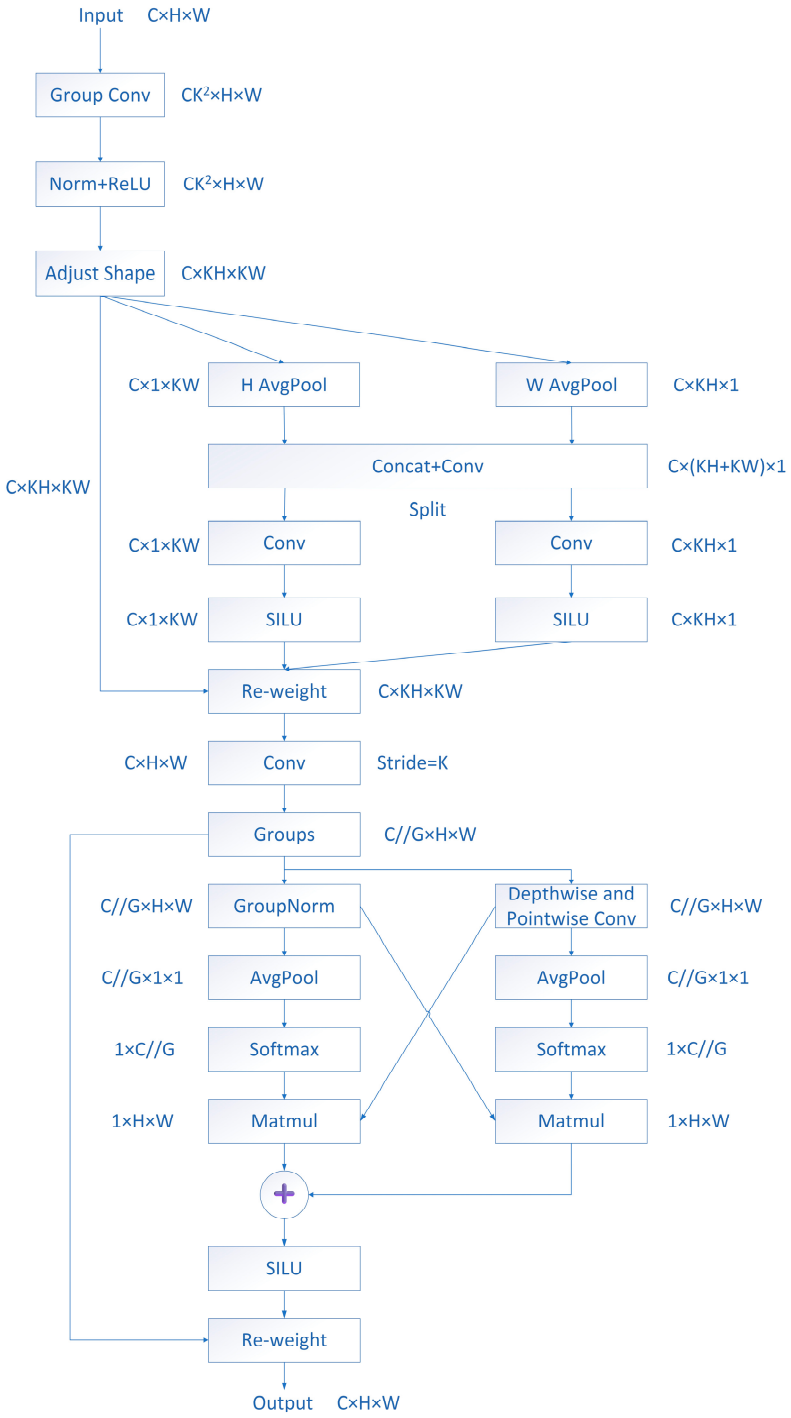


Figure 3. Structure of RFEMA module.

RFCA module enhances existing spatial attention mechanisms by combining Receptive Field Attention (RFA) with Coordinated Attention (CA), resolving the issue that CA only focuses on spatial features without enabling the sharing of convolution kernel parameters. The original structure of RFCA module is depicted in Figure 4. RFCA module adds several operations to CA module, including group convolution operations, batch normalization, activation functions, and adjusting the shape of the feature map. Additionally, it employs a $K \times K$ convolution operation to output the final feature information after the output from CA module. RFCA module first makes a feature extraction operation through group convolution to generate multiple feature maps of different sizes. Then, the generated feature map is reshaped to isolate the local receptive field features of each location, which

is conducive to the subsequent feature rearrangement and integration. Compared to CA module, RFCA module places greater emphasis on the spatial features of the receptive fields, allowing the model to better handle local regions within the image. For stamp text detection, RFCA module makes the model more focused on emphasizing smaller text or edges and also effectively suppresses background noise. When the content of the seal is blurred, RFCA module enables the model to better concentrate on relevant areas, thus improving the distinction between seal text and background and enhancing detection accuracy. RFCA module combines spatial attention with convolution through the integration of the attention mechanism, enabling flexible adjustment of convolution kernel parameters, solving the problem of convolution parameter sharing. The combination of RFA and CA directs the attention of existing spatial attention mechanisms toward the receptive field features. This allows the model to solve the problem of remote information parameter sharing and requires fewer parameters than self-attention. Through this focused attention on receptive field spatial features, the model better adapts to the deformation of seals, enhancing its ability to detect a diverse range of seals.

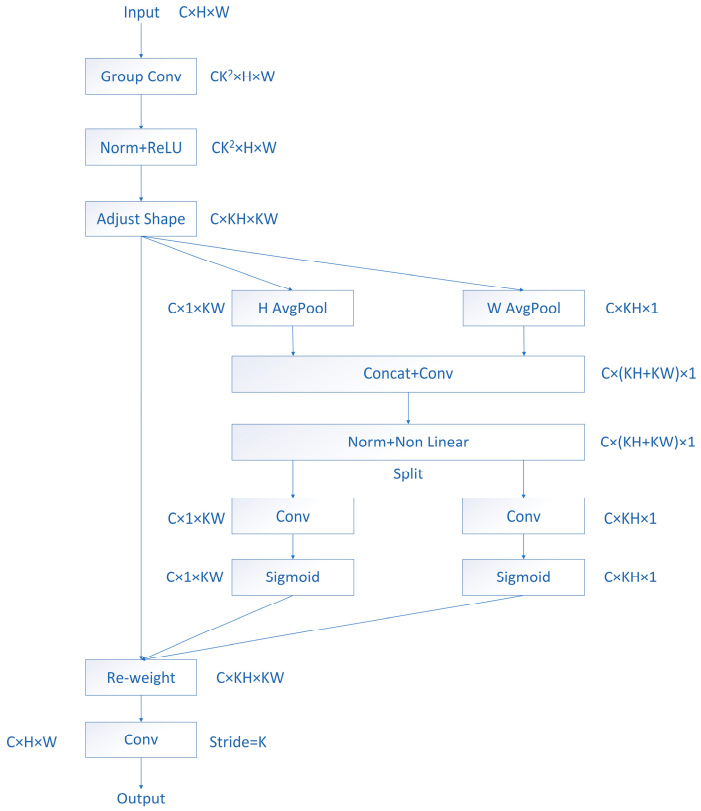


Figure 4. Structure of RFCA module.

RFEMA module is a simplified and optimized version of RFCA module. First, it splices the features of row and column directions, and then splits the features of row and column directions directly after 1×1 convolution, which simplifies the process of feature splicing and segmentation. Following this, SiLU activation function is used in both directions to calculate the respective attention weights. By removing the original batch normalization operation, RFEMA module reduces computational complexity and simplifies the model training process. These improvements make RFEMA module more efficient and concise while preserving the advantages of RFCA module.

The original structure of EMA module is shown in Figure 5. EMA module preserves the information of each channel and reduces computational overhead by reshaping part of the channels into batch dimensions. Initially, EMA module selects a portion of the 1×1 convolution in CA module as its 1×1 branch and sets a 3×3 convolution in parallel

as the 3 × 3 branch. Compared to CA module, EMA module introduces a cross-spatial information aggregation method to achieve richer feature aggregation. The input for the cross-spatial learning method consists of two parts: the outputs of the 1 × 1 branch and the 3 × 3 branch. EMA module encodes global spatial information in the output of the 1 × 1 branch, followed by Softmax activation function to generate attention weights for each channel. Simultaneously, the output of the 3 × 3 branch is reshaped to the corresponding shape, and the outputs of these two branches are matrix-multiplied to obtain the weighted feature aggregation results, generating the first spatial attention map. Similarly, the output of the 3 × 3 branch is subjected to a 2D global average pooling operation and Softmax activation function is used to generate attention weights. The 1 × 1 branch is then reshaped into the corresponding dimensions and the outputs of the two branches are matrix multiplied to obtain another weighted result to generate the second spatial attention map. Finally, EMA module fuses the output features of the two branches and adds the corresponding attention weight values in each spatial attention map to obtain a new set of spatial attention weights. Cross-space learning through the output features of the two parallel branches enables the model to obtain richer contextual information.

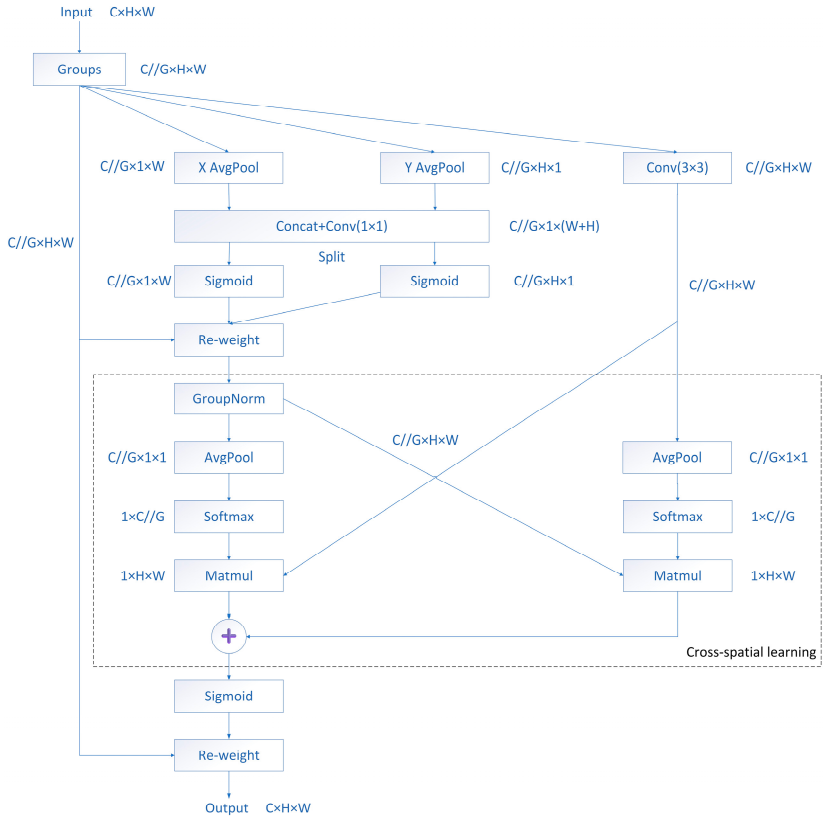


Figure 5. Structure of EMA module.

In RFEMA module, the output of RFCA module is used as the input for the 1 × 1 branch. The 3 × 3 convolution is replaced by depthwise separable convolution, Sigmoid activation function of the original EMA module is replaced by SiLU activation function, and a residual structure is added. Depthwise separable convolution decomposes the traditional convolution operation into two steps: depthwise convolution and pointwise convolution. This convolution first applies depthwise convolution for independent spatial feature extraction on the input channels and then performs cross-channel feature combination using pointwise convolution. This achieves more efficient feature extraction and computational optimization. By extracting spatial and channel features separately, the parameters are significantly reduced, making the model more lightweight and efficient.

In seal text detection, capturing complex features in a deep network can be challenging if the gradient vanishes, making training difficult. SiLU activation function has good gradient flow properties, which can alleviate the problem of gradient vanishing. It allows for negative outputs and combines linear and nonlinear properties, enhancing the feature representation of the model. To minimize information loss and retain more original feature information, the residual connection is introduced in RFEMA module. This design not only accelerates the training process of the model but also promotes its rapid convergence.

RFEMA module not only retains the advantages of RFCA module's focus on receptive-field features but also incorporates the benefits of EMA's cross-spatial learning. RFCA component of RFEMA module enhances the capture of local features by emphasizing the fusion of receptive field features and convolution with spatial attention. Meanwhile, EMA module can obtain global features of different scales through adaptive average pooling and multi-scale convolution operations. The use of depthwise separable convolution enhances the ability to integrate global information and detailed information. The combination of these two methods makes the model to be more comprehensive in extracting local features and global features. This also allows the model to handle noise more efficiently, effectively distinguish between background and text, and reduce false and missed detections. RFCA module uses group convolution to improve the feature extraction ability of the model while reducing the parameters. EMA module reduces the dimensions by reshaping the dimensions, in this way avoiding the dimension reduction of the convolution method. The module also combines depthwise convolution and pointwise convolution to improve the feature extraction ability while making the model more computationally efficient. This combination alleviates the high demand for computational resources of YOLOv8 without sacrificing the feature extraction ability, and achieves a good balance between feature extraction effectiveness and computational efficiency. Since RFCA module can accurately extract the local features, it improves the robustness of the model in dealing with the detailed features. EMA module has a strong ability to integrate global functions, which enhances the generalization ability of the model in various cases, and makes the model perform well in various complex environments.

3.3. Formatting of Mathematical Components

AKConv module uses a new coordinate generation algorithm to define the initial positions of the convolutional kernels. This balances the relationship between fixed shape convolutional kernels and network performance. It adjusts the shape of the convolutional kernels by introducing offsets to suit different application scenarios. AKConv module is illustrated in Figure 6.

AKConv module first obtains the corresponding offsets through depthwise separable 2D convolution operations. Depthwise separable convolution significantly reduces computational complexity and improves training efficiency by decomposing standard convolution into two independent steps: depthwise convolution and pointwise convolution. Specifically, it performs depthwise convolution independently for each input channel and then combines these results through pointwise convolution. The modified coordinates are obtained by adding the initial coordinates and offsets, where the initial coordinates are generated by the initial coordinate generation algorithm. Finally, the features at the corresponding locations are extracted by interpolation and resampling. AKConv module borrowed the idea of RFCA module performing separate convolutions in the row and column directions. This can solve the problem that irregular convolution kernels are difficult to extract image features. AKConv module convolves features with convolution kernels appropriate size in the column direction, and then uses row convolution to complete the extraction of irregular convolution features. In order to speed up the training process, AKConv module performs batch normalization of the input feature maps. In addition, AKConv module employs Mish activation function, which allows the model to learn more complex features compared to SiLU activation function. To alleviate the problem of gradient vanishing in deep neural networks, residual connection is added to AKConv module

to improve the stability of model training. This residual connection is usually a mapping that simply transfers the input directly to the output without any changes. But there is a different situation in this model. When the number of input channels are different from the number of output channels or the stride is not 1, a sequence container is used to adjust the number of input channels and the spatial dimensions. The sequence container is composed of a 1×1 convolutional layer and a batch normalization layer. This ensures that the inputs and outputs have the same shape, thus maintaining the stability of the model training.

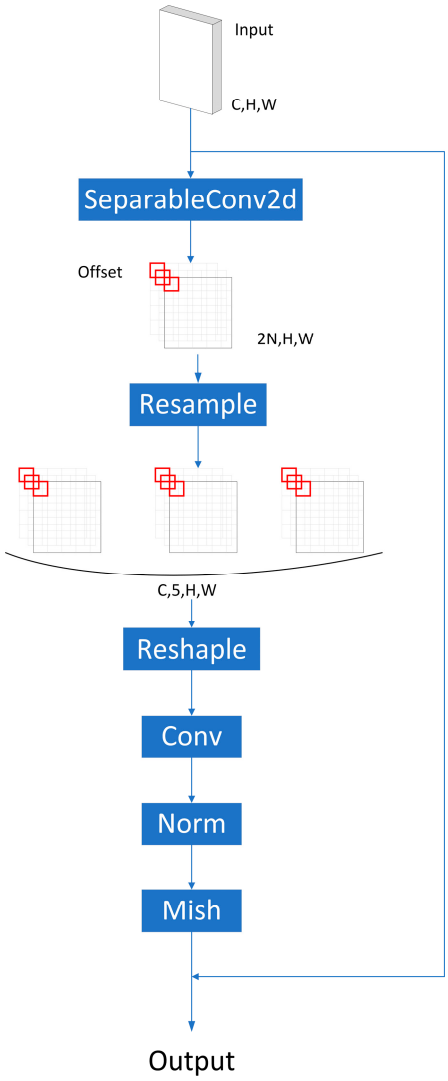


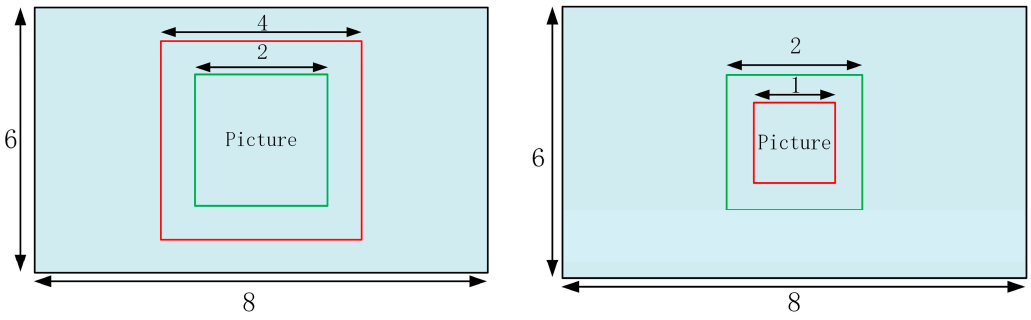
Figure 6. Structure of AKConv module. The red frame in the figure represents the sample shape.

AKConv module has an adaptive convolution kernel, which can dynamically adjust the sampling position based on the feature map. This enhances the model's ability to extract local features and enables the model to better handle character connections in seal text detection. compared to fixed sampling position, AKConv module can efficiently handle rotated images, allowing the model to better handle the seals of font changes. In addition, AKConv module also includes an adaptive learning rate adjustment mechanism, which makes the model to dynamically adjust the gradient during the backpropagation process.

3.4. Loss Function

The loss function of YOLOv8 mainly consists of two parts: classification loss and regression loss. BCE Loss is used for classification loss, while DFL Loss and CIOU Loss are used for regression loss. These three losses are weighted using specific weight proportions to form the complete loss function of YOLOv8.

In this study, a bounding box regression loss function based on MPDIoU is introduced. In some cases, as shown in Figure 7, when there are two images where the predicted box and the ground truth box have the same aspect ratio but are visually inconsistent—one predicted box is inside the ground truth box and the other is outside—the calculation results of CIOU and GIOU may be the same. This leads to the ineffectiveness of these loss functions in handling such cases, which can limit the model's convergence speed. MPDIoU loss function, however, can compute the difference between these two boxes, thus addressing this issue.



(**a**) (**b**)

Figure 7. (**a**) The predicted bounding box is outside the ground truth bounding box, at this time L_{CIoU} = 0.75, L_{GIoU} = 0.75, L_{MPDIoU} = 0.79; (**b**) The predicted bounding box is inside the ground truth bounding box, at this time L_{CIoU} = 0.75, L_{GIoU} = 0.75, L_{MPDIoU} = 0.76.

Inspired by the geometric properties of bounding boxes, which determine a rectangular shape through the coordinates of the points in the top-left and bottom-right corners, MPDIoU loss function directly minimizes the distances between these two sets of corresponding points between the predicted box and the ground truth box. This approach simplifies the process of similarity comparison and improves the accuracy of bounding box detection. This makes the model more suitable for overlapping and non-overlapping bounding box regression. In addition, MPDIoU loss function considers the central point distance and the deviations in width and height, improve the efficiency of bounding box regression. The formula of MPDIoU loss function is as follows:

$$L_{MPDIou} = 1 - MPDIou, \tag{1}$$

$$MPDIou = \frac{B_{gt} \cap B_{prd}}{B_{prd} \cup B_{gt}} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2}, \tag{2}$$

$$d_1^2 = \left(x_1^{prd} - x_1^{gt}\right)^2 + \left(y_1^{prd} - y_1^{gt}\right)^2, \tag{3}$$

$$d_2^2 = \left(x_2^{prd} - x_2^{gt}\right)^2 + \left(y_2^{prd} - y_2^{gt}\right)^2, \tag{4}$$

where w is the width of the input image, h is the height of the input image, B_{prd} is the predicted box, and B_{gt} is the ground truth box. The coordinates of the predicted box is $\left(x_1^{prd}, y_1^{prd}\right)$, and the coordinates of the groundtruth box is $\left(x_1^{gt}, y_1^{gt}\right)$, d_1 is the distance between the top-left corner of the predicted box and the top-left corner of the ground truth

box, and d_2 is the distance between the bottom-right corner of the predicted box and the bottom-right corner of the ground truth box.

In the seal text detection, MPDIoU loss function can more accurately recognize the differences between bounding boxes. This loss function can more efficiently handles seal text with complex shapes and rich details. It can improve the localization accuracy of the model and enhance the accuracy and robustness of detection.

4. Experiments

4.1. Datasets

The dataset used in this study is a self- constructed dataset for Chinese seal text detection. This dataset contains various difficult situations that are encountered when detecting seal text. These situations include partial absence of seal text due to document overwriting, or blurring of seal text caused by uneven stamping force, etc. The dataset consists of 7004 images, including 2002 real seals and 5002 electronic seals. The real seals were collected from the public information of the people's government of Guangxi Zhuang Autonomous Region, mainly collecting seals on documents or pictures. The electronic seals were created by ourselves. During the creation process, we enhanced data diversity of the data by adding anti-counterfeiting marks, rotating text angles, modifying word spacing, and increasing simulation effects. The modifications allow the model to learn from more realistic scenarios, thus improving its performance in practical applications. This also prevents the model from overfitting the training and ensures better performance on the test. The dataset is divided into a training dataset and a test dataset in a ratio of 8:2, resulting in 5603 training images and 1401 test images. The training set contains 1614 real seals and 3989 electronic seals. The test set contains 388 real seals and 1013 electronic seals. The real seals in the training set contained 1302 blurred seals, 263 missing text seals, and 49 distorted seals, with a ratio of 263:53:10. The real seals in the test set contained 349 blurred seals, 35 missing text seals, and 4 distorted seals, with a ratio of about 70:7:1. The electronic seals in the training set and the test set are different angles, different texts and blurred seals. All the images have a resolution of 640 × 640 pixels and saved in PNG format.

As shown in Figure 8, the data set contains different types of seals, such as seals with missing text, seals with blurred text, seals with distorted text, seals with complex backgrounds, and so on. Compared with text in natural scenes, these seals are difficult to detect and challenging. The dataset contains seals of different sizes, shapes, and angles, which allows the model to learn the diversity of the dataset and better adapt to different scales of data. The distribution of data samples is shown in Figure 9.

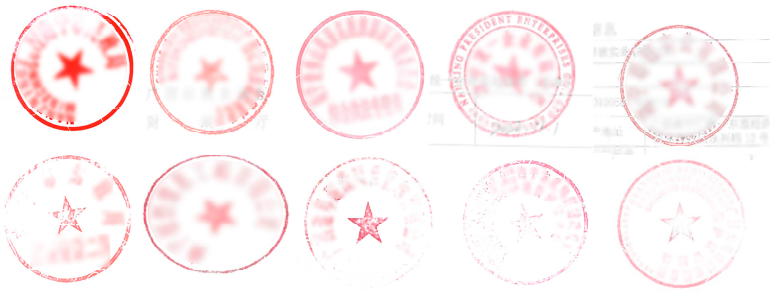


Figure 8. Different types of seals in the data set.

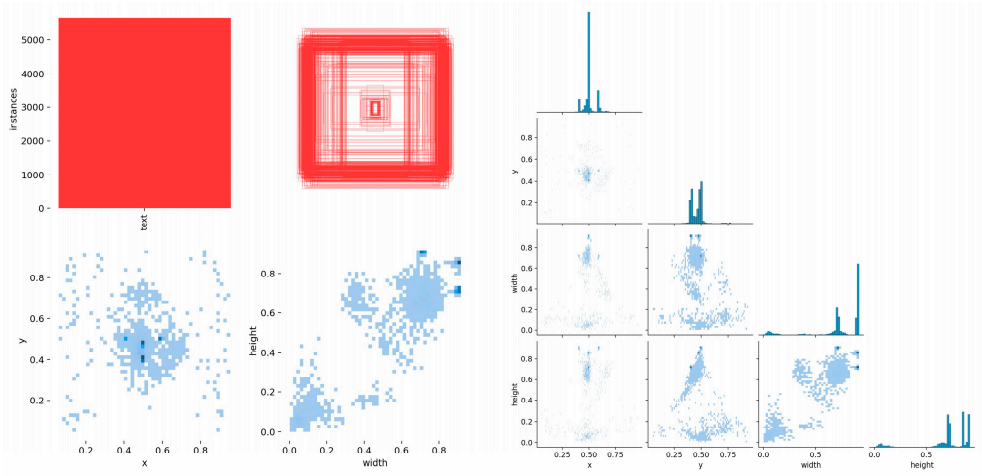


Figure 9. Dataset label distribution map.

4.2. Evaluation Metrics

In this study, Precision, Recall, F1 score, and mean Average Precision (mAP) were used as the primary metrics to evaluate the model.

$$P = \frac{TP}{TP + FP}, \tag{5}$$

$$R = \frac{TP}{TP + FN}, \tag{6}$$

$$AP = \int_0^1 p(R)dR, \tag{7}$$

$$mAP = \frac{\sum_{i=0}^{n} AP(i)}{n} \tag{8}$$

where, TP denotes the number of samples that are actually positive and correctly detected as positive by the model. FP denotes the number of samples that are actually negative but incorrectly detected as positive by the model. FN denotes the number of samples that are actually positive but incorrectly detected as negative by the model. Positive samples refer to labeled textual objects, while negative samples represent background regions that are not related to the text.

4.3. Experimental Platform and Parameters

The experiments were conducted on a hardware platform featuring an AMD EPYC 7T83 CPU and an RTX 4090 graphics card, using the PyTorch deep learning framework.

During the training process, the number of epochs was set to 80, with a batch size of 32. The AdamW optimizer was used, with an initial learning rate (lr0) set to 0.002, a final learning rate (lrf) set to 0.01, and momentum set to 0.937.

4.4. Experimental Results

To evaluate the detection performance of the model, a self-constructed seal text dataset was used. Under the same experimental conditions, the model was compared with several up to date seal text detection algorithms. The values of Precision (P), Recall (R), and F1 score for each algorithm were compared, and the results are shown in Table 1.

Table 1. Comparative experiments of different detection models.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)
PSENet	86.13	90.62	88.32	-
DB	76.30	71.16	73.64	-
FCENet	91.05	82.70	86.67	-
DB++	93.69	71.16	80.88	-
TextBPN++	83.79	80.90	-	68.39
DPText-DETR	79.78	67.03	72.85	-
YOLOv8	95.00	92.00	93.48	94.90
RA-YOLOv8	95.40	93.60	94.51	96.50

As shown in Table 1, DB model does not perform well in seal text detection. DB++ model and FCENet model have relatively high precision but relatively low recall and F1 scores. DB++ model is an improvement over DB model, mainly enhancing the detection of text regions by introducing a binarization mechanism. However, it is not suitable for handling complex backgrounds and detail-rich seal text content, as the binarization operation may ignore some edge information, resulting in some text regions not being detected correctly. FCENet model primarily detects text regions through Fourier transform, but it struggles when dealing with small or blurred text, as some detailed information may not be captured by the model, leading to lower detection performance in these regions. The overall performance of PSENet and DPText-DETR models is not as effective as that of RA-YOLOv8. PSENet model's overall performance may be lower due to its weak ability to detect small text. DPText-DETR model may be interfered with by background noise in complex backgrounds, causing the dynamically updated control points to deviate slightly from the actual text regions. In high-density text scenes, control points in neighboring text regions may overlap, making it difficult for the model to distinguish each individual text region, leading to detection omissions. RA-YOLOv8 improves precision by 0.4%, recall by 1.6%, and F1 score by 1.03% compared to the original YOLOv8. The experimental results demonstrate that RA-YOLOv8 can avoid background noise interference, accurately distinguish between background and target, and achieve precise localization of stamped text under complex backgrounds, making the model robust.

As shown in Figure 10, the comparative experiment was mainly conducted on the S-scale and M-scale of YOLOv5, the N-scale, S-scale and M-scale of YOLOv6, the N-scale, S-scale and M-scale of YOLOv8, and the N-scale and S-scale of RA-YOLOv8. As shown in Figure 10, with the increase of parameters, the precision of RA-YOLOv8 is greater than other models. We also compared the performance of different models in the yolo series, and the results are shown in Figure 11. It can be seen from Figure 11 that RA-YOLOv8 model has the best performance.

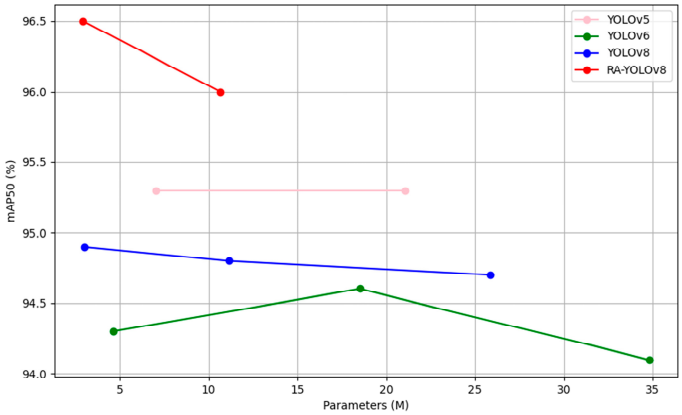


Figure 10. Comparative experiments of yolo series.

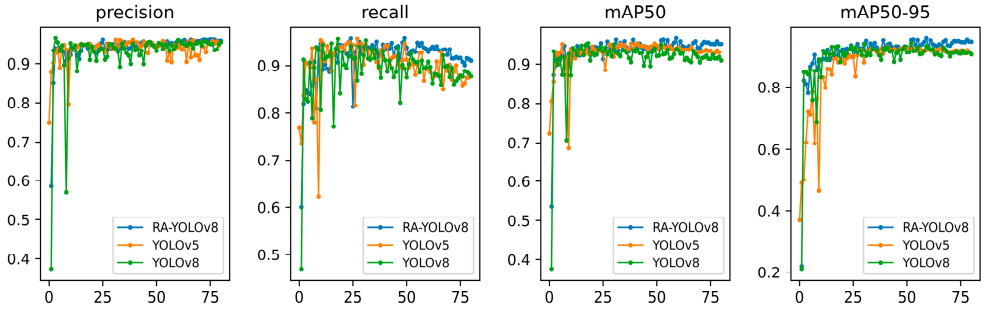


Figure 11. Performance comparison experiment of yolo series models.

We selected 6 challenging seal pictures from the test dataset for testing. As shown in Figure 12, for seals with uneven colors, compared with RA-YOLOv8, the detection results of YOLOv8 and YOLOv5 are lower, while the detection results of YOLOv6 are very low. For seals containing two texts, compared with RA-YOLOv8, the positioning results of YOLOv8 and YOLOv5 are biased, and the positioning results of YOLOv6 are inaccurate. For complex background or fuzzy seals, RA-YOLOv8 can accurate detection, and other models are underperforming. For seals with missing text, the localization results of YOLOv8, YOLOv5 and YOLOv6 are inaccurate, and these models ignore the missing text part. It can be seen that the detection performance of RA-YOLOv8 is better than that of the other models.

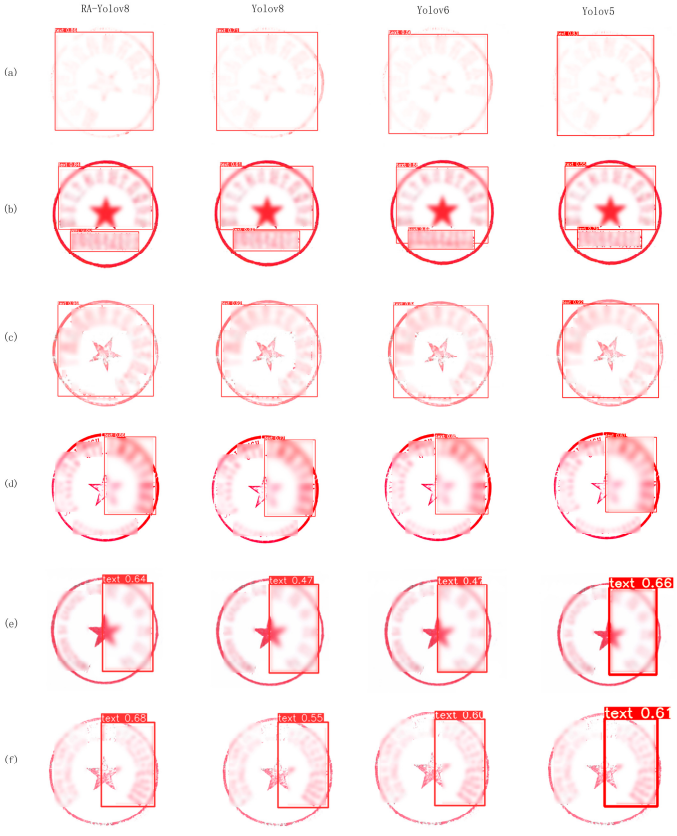


Figure 12. Visualization results of different models.

As you can see from Figure 13, RA-YOLOv8 can detect a seal that is missing text. However, the model is not very accurate at detecting seals that have a large amount of text content missing. As shown in the left image of Figure 13, the strokes of this seal are incoherent, with many horizontal or vertical strokes missing, leaving only a few dots. As shown in the right picture of Figure 13, this seal has a large amount of blank space due

to the missing upper part of many characters. For this kind of seal, although the model can extract the detailed features of the image, it is difficult for the model to obtain enough context information due to the excessive white space, so they cannot be correctly judged as a complete text instance.

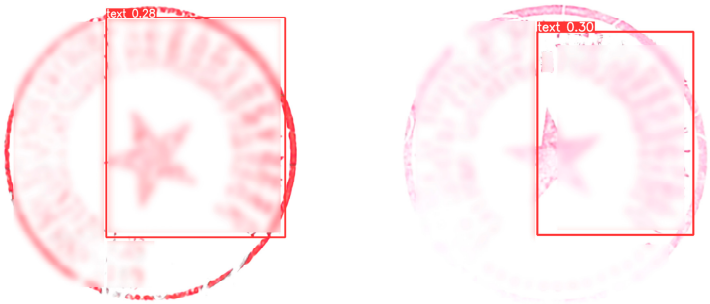


Figure 13. Visualized results of failed cases.

4.5. Ablation Experiments

To assess the validity of this study, ablation experiments will be conducted on the two improvement points, using the values of Precision (P), Recall (R), F1 score, and mean Average Precision (mAP) as indicators for evaluating the model. The results are shown in Table 2.

Table 2. Ablation experiments of different modules.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)
YOLOv8	95.00	92.00	93.48	94.90
YOLOv8 +RFEMA	95.40	92.80	94.07	96.40
YOLOv8 +AKConv	95.20	92.50	93.81	94.80
RA- YOLOv8	95.40	93.60	94.51	96.50

As shown in Table 2, the addition of RFEMA module and AKConv module to the backbone and neck of YOLOv8, respectively, improves the detection performance of YOLOv8. Specifically, the precision increased by 0.4% and 0.2%, the recall increased by 0.8% and 0.5%, and the F1 score increased by 0.59% and 0.33%, respectively. The experimental results show that RFEMA module and AKConv module significantly enhance model's ability to detect seal text in complex backgrounds and improve model's feature extraction capability.

As can be seen from Figure 14, RFEMA module can enhance the model's ability to extract the detailed features of seal text, so that the model can accurately extract seals with missing text or complex background. AKConv module enables the model to accurately extract seals with high text density.

Figure 15 shows the enhancement effect of each module. It can be seen that both RFEMA module and AKConv module can enhance the performance of the model. In particular, when the RFEMA module is added to YOLOv8, recall, map50 and MAP50-95 all have certain improvements.

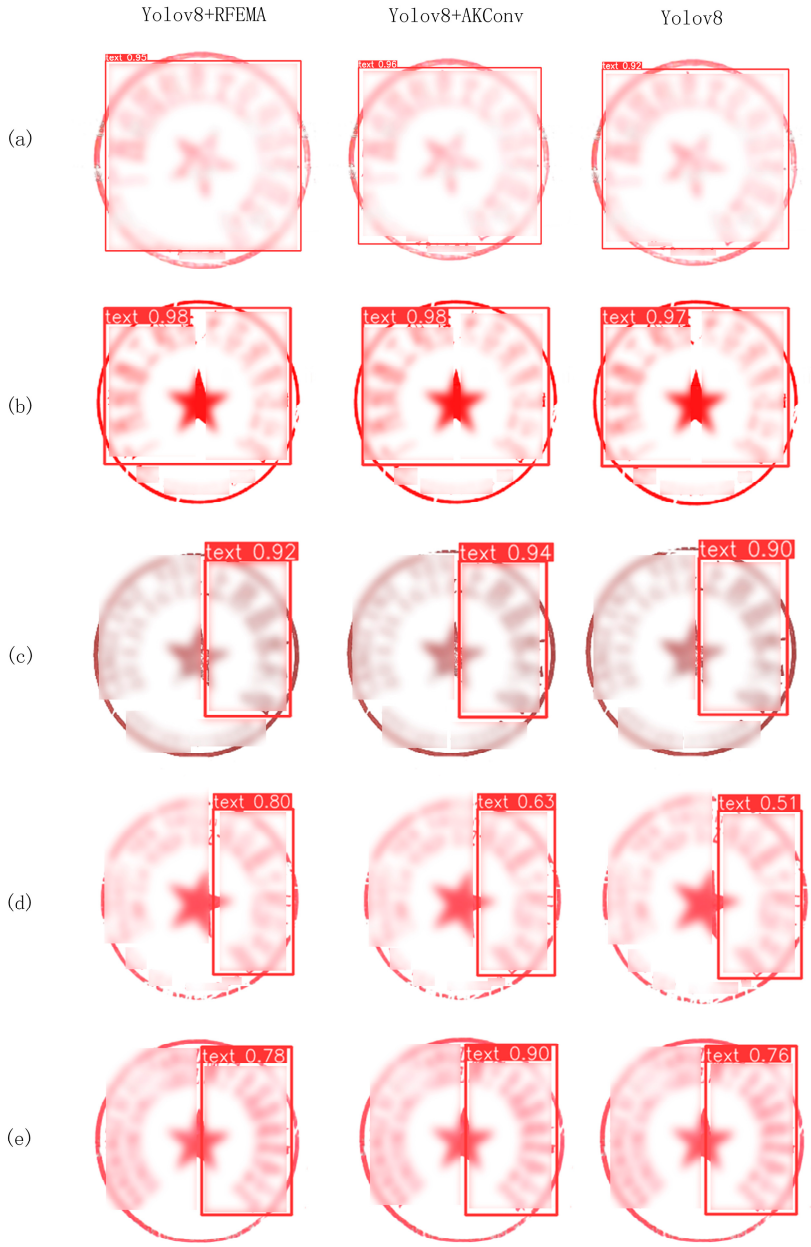


Figure 14. Visualization results of different modules.

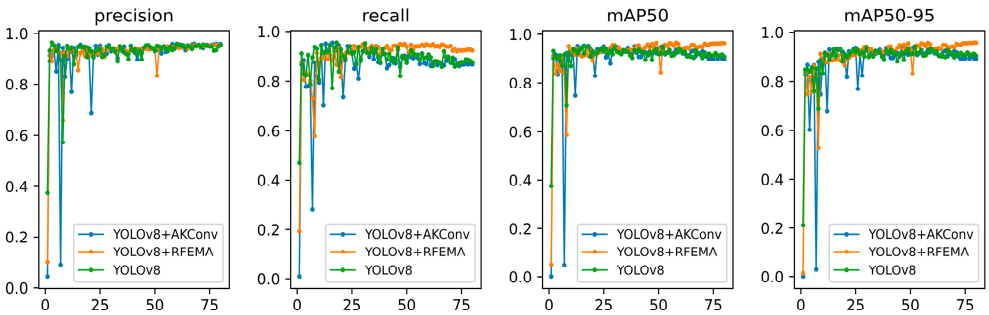


Figure 15. Performance comparison experiment of different modules.

5. Conclusions

In this paper, an improved YOLOv8 model called RA-YOLOv8 is proposed. This model can solve the difficult problem of seal text detection when detecting complex backgrounds. RA-YOLOv8 enhances the backbone and neck of YOLOv8 by adding RFEMA module and AKConv module. These upgrades significantly improve the model's ability to extract local features and global features, enabling it to efficiently handle curved seals and improve detection performance of the model. Experimental results show that compared with YOLOv8, RA-YOLOv8 improves the precision by 0.4%, the recall by 1.6%, and the F1 score by 1.03%.

In future work, the seal text dataset can be extended further, allowing the dataset to cover a wider variety of situations. In addition, a more lightweight seal text detection model can be created to reduce the model size while maintaining high detection performance. This would improve the efficiency and practicality of the detection process.

Author Contributions: Conceptualization, H.S. and C.T.; methodology, H.S.; software, H.S.; validation, H.S., C.T. and S.P.; formal analysis, H.S. and C.T.; investigation, H.S. and H.W.; resources, H.S.; data curation, H.S.; writing—original draft preparation, H.S.; writing—review and editing, B.H.; visualization, H.S.; supervision, B.H.; project administration, B.H.; funding acquisition, B.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Open Project Program of Guangxi Key Laboratory of Digital Infrastructure, grant number GXDINBC202406 and National Natural Science Foundation of China, grant number 61962005.

Data Availability Statement: The data presented in this study are available on request from the corresponding author, as the real seals were obtained from the Guangxi Zhuang Autonomous Region Government Information disclosure website. For legal and ethical reasons, we believe that the real seal dataset should be limited to: Under article 280 of the Criminal Law, anyone who forges, falsifies, trades in or steals, snatches or destroys official documents, papers or seals of State organs shall be sentenced to fixed-term imprisonment of not more than three years, detention, control or deprivation of political rights, and shall also be punished by a fine; if the circumstances are serious, he shall be sentenced to fixed-term imprisonment of not less than three and not more than 10 years, and shall also be punished by a fine. According to article 52 of the Law of the People's Republic of China on Punishments for Public Security Administration, anyone who forges, alters, or trades in the official documents, papers, certificates or seals of State organs, people's organizations, enterprises, institutions, or other organizations shall be sentenced to detention of not less than 10 days and not more than 15 days, and may be sentenced to a fine of not more than CNY 1000. The leaked seals may be misused to sign false contracts, resulting in damage to the legitimate rights and interests of enterprises. It can also be illegally withdrawn or transferred by criminals, causing financial losses. Making data sets public could cause problems for government agencies. Relevant researchers may contact the corresponding author (bhhuang66@gxu.edu.cn) to obtain the dataset by providing information verifying identity, organization, and purpose.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tian, Z.; Huang, W.; He, T.; He, P.; Qiao, Y. Detecting Text in Natural Image with Connectionist Text Proposal Network. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016.
2. Shi, B.; Bai, X.; Belongie, S. Detecting Oriented Text in Natural Images by Linking Segments. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
3. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
4. Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018.
5. Yang, Q.; Cheng, M.; Zhou, W.; Chen, Y.; Qiu, M.; Lin, W. Inceptext: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.

6. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2 CNN: Rotational Region CNN for Arbitrarily-Oriented Scene Text Detection. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
7. Zhang, X.; Song, Y.; Song, T.; Yang, D.; Ye, Y.; Zhou, J.; Zhang, L. AKConv: Convolutional Kernel with Arbitrary Sampled Shapes and Arbitrary Number of Parameters. *arXiv* **2023**, arXiv:2311.11587.
8. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. RFAConv: Innovating Spatial Attention and Standard Convolutional Operation. *arXiv* **2023**, arXiv:2304.03198.
9. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023.
10. Ma, S.; Xu, Y. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. *arXiv* **2023**, arXiv:2307.07662.
11. Gao, W.; Dong, S.; Zhou, S. Stroke Edge Matching Based Automatic Seal Imprint Verification. *Patt. Recog. Artif. Intell.* **1994**, *7*, 338–342.
12. Chen, L.; Liu, T.; Chen, J.; Ma, S. ldentification of Seal Imprint Based on Center-Rays Model and its Application. *Opt. Technol.* **2006**, *32*, 511–513.
13. Cai, L.; Mei, L. Wedge-Ring Based Method for Color Seal Registration. *J. Zhejiang Univ. (Eng. Sci.)* **2006**, *40*, 1696–1700.
14. Yao, M.; Mou, X.; Chen, P.; Zhao, M.; Li, Z. Research on Detection, Location and Recognition of Seals in Images. *Inf. Technol. Inf.* **2018**, *12*, 148–150.
15. Zhang, X.; Qin, Y.; Dong, Z.; Huang, Q.; Li, J. Chinese Seal Recognition Method Based on Flood Filling Algorithm. *Appl. Eletron. Tech.* **2022**, *48*, 1–6.
16. Kang, Y.; Sun, P.; Lang, Y.; Wang, Y. Adaptive Canny Detection of Obsolete Seals in Reconstructed Color Space. *Comput. Simul.* **2023**, *40*, 230–234.
17. Zhong, Z.; Jin, L.; Huang, S. DeepText: A New Approach for Text Proposal Generation and Text Detection in Natural Images. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
18. Zhong, Z.; Sun, L.; Huo, Q. An Anchor-Free Region Proposal Network for Faster R-CNN-Based Text Detection Approaches. *IJDAR* **2019**, *22*, 315–327. [CrossRef]
19. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *TIP* **2018**, *27*, 3676–3690. [CrossRef]
20. Duan, J.; Xu, Y.; Kuang, Z.; Yue, X.; Sun, H.; Guan, Y.; Zhang, W. Geometry Normalization Networks for Accurate Scene Text Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
21. He, M.; Liao, M.; Yang, Z.; Zhong, H.; Tang, J.; Cheng, W.; Yao, C.; Wang, Y.; Bai, X. MOST: A Multi-Oriented Scene Text Detector with Localization Refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
22. Wang, H.; Bai, X.; Yang, M.; Zhu, S.; Wang, J.; Liu, W. Scene Text Retrieval via Joint Text Detection and Similarity Learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
23. Liu, Y.; Jin, L.; Zhang, S.; Luo, C.; Zhang, S. Curved Scene Text Detection via Transverse and Longitudinal Sequence Connection. *Pattern. Recognit.* **2019**, *90*, 337–345. [CrossRef]
24. Wang, X.; Jiang, Y.; Luo, Z.; Liu, C.-L.; Choi, H.; Kim, S. Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
25. Liu, Z.; Lin, G.; Yang, S.; Liu, F.; Lin, W.; Goh, W.L. Towards Robust Curve Text Detection with Conditional Spatial Expansion. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
26. Zhang, C.; Liang, B.; Huang, Z.; En, M.; Han, J.; Ding, E.; Ding, X. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
27. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. ABCNet: Real-Time Scene Text Spotting with Adaptive Bezier-Curve Network. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
28. Zhang, S.-X.; Zhu, X.; Yang, C.; Wang, H.; Yin, X.-C. Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
29. Dai, P.; Zhang, S.; Zhang, H.; Cao, X. Progressive Contour Regression for Arbitrary-Shape Scene Text Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
30. Zhu, Y.; Du, J. TextMountain: Accurate Scene Text Detection via Instance Segmentation. *Pattern. Recognit.* **2021**, *110*, 107336. [CrossRef]
31. Deng, D.; Liu, H.; Li, X.; Cai, D. PixelLink: Detecting Scene Text via Instance Segmentation. *AAAI* **2018**, *32*, 6773–6780. [CrossRef]

32. Baek, Y.; Lee, B.; Han, D.; Yun, S.; Lee, H. Character Region Awareness for Text Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
33. Tian, Z.; Shu, M.; Lyu, P.; Li, R.; Zhou, C.; Shen, X.; Jia, J. Learning Shape-Aware Embedding for Scene Text Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
34. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape Robust Text Detection with Progressive Scale Expansion Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
35. Xu, Y.; Wang, Y.; Zhou, W.; Wang, Y.; Yang, Z.; Bai, X. TextField: Learning a Deep Direction Field for Irregular Scene Text Detection. *TIP* **2019**, *28*, 5566–5579. [CrossRef] [PubMed]
36. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-Time Scene Text Detection with Differentiable Binarization. *AAAI* **2020**, *34*, 11474–11481. [CrossRef]
37. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
38. Cai, Y.; Liu, Y.; Shen, C.; Jin, L.; Li, Y.; Ergu, D. Arbitrarily Shaped Scene Text Detection with Dynamic Convolution. *Pattern. Recognit.* **2022**, *127*, 108608. [CrossRef]
39. Zhong, Y.; Cheng, X.; Chen, T.; Zhang, J.; Zhou, Z.; Huang, G. PRPN: Progressive Region Prediction Network for Natural Scene Text Detection. *KBS* **2021**, *236*, 107767. [CrossRef]
40. Yu, W.; Liu, Y.; Hua, W.; Jiang, D.; Ren, B.; Bai, X. Turning a CLIP Model Into a Scene Text Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
41. Shi, X.; Peng, G.; Shen, X.; Zhang, C. TextFuse: Fusing Deep Scene Text Detection Models for Enhanced Performance. *Multimed. Tools Appl.* **2024**, *2*, 22433–22454. [CrossRef]
42. Naveen, P.; Hassaballah, M. Scene Text Detection Using Structured Information and an End-to-End Trainable Generative Adversarial Networks. *Pattern. Anal. Appl.* **2024**, *27*, 33. [CrossRef]
43. Zheng, J.; Zhang, L.; Wu, Y.; Zhao, C. BPDO: Boundary Points Dynamic Optimization for Arbitrary Shape Scene Text De-tection. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024.