applied sciences

MDPI

Article

Text Detection and Recognition for X-ray Weld Seam Images

Qihang Zheng and Yaping Zhang *

School of Information, Yunnan Normal University, Kunming 650500, China; yren0949@gmail.com
* Correspondence: zhangyp@ynnu.edu.cn

Abstract: X-ray weld seam images carry vital information about welds. Leveraging graphic–text recognition technology enables intelligent data collection in complex industrial environments, promising significant improvements in work efficiency. This study focuses on using deep learning methods to enhance the accuracy and efficiency of detecting weld seam information. We began by actively gathering a dataset of X-ray weld seam images for model training and evaluation. The study comprises two main components: text detection and text recognition. For text detection, we employed a model based on the DBNet algorithm and tailored post-processing techniques to the unique features of weld seam images. Through model training, we achieved efficient detection of the text regions, with 91% precision, 92.4% recall, and a 91.7% F1 score on the test dataset. In the text recognition phase, we introduced modules like CA, CBAM, and HFA to capture the character position information and global text features effectively. This optimization led to a remarkable text line recognition accuracy of 93.4%. In conclusion, our study provides an efficient deep learning solution for text detection and recognition in X-ray weld seam images, offering robust support for weld seam information collection in industrial manufacturing.

Keywords: weld seam image; text detection and recognition; attention mechanism

Citation: Zheng, Q.; Zhang, Y. Text Detection and Recognition for X-ray Weld Seam Images. *Appl. Sci.* **2024**, *14*, 2422. https://doi.org/10.3390/app14062422

Academic Editor: Douglas O'Shaughnessy

Received: 18 February 2024
Revised: 8 March 2024
Accepted: 11 March 2024
Published: 13 March 2024

1. Introduction

With the widespread application of modern welding technology in industrial manufacturing, there is an increasingly urgent demand for the automated analysis and processing of weld seam images. The text information in weld seam images contains important welding parameters, workpiece identification, and other key information such as dates, weld seam serial numbers, and positioning symbols. However, digitizing weld seam images often requires manually inputting specific numbering, which is inefficient and inconvenient for secondary queries. Accurately detecting and recognizing text in weld seam images can automate the collection of information from these images, greatly enhancing work efficiency. Additionally, timely access to relevant information from weld seam images facilitates subsequent work. Therefore, text detection and recognition in X-ray weld seam images hold significant theoretical and practical significance.

The recognition of text information in weld seam images can be divided into two key steps: text detection and text recognition. This division allows for the independent improvement of models for text detection and text recognition, thereby enhancing the overall accuracy of the recognition process. Text detection aims to extract connected text regions from weld seam images and form continuous text lines, while text recognition focuses on identifying semantically meaningful text characters within these text lines.

Traditional text detection and recognition methods (such as edge detection, sliding windows, template matching, etc.) have been widely applied to general text recognition tasks, performing well primarily on well-structured printed documents. However, their robustness is compromised when dealing with complex scenarios involving blurry images, low contrast, and other challenging conditions, making it difficult to achieve accurate detection and recognition.

With the rapid development of deep learning, the performance of text detection and recognition tasks has significantly improved. In the field of text detection, CTPN [1] is a text detection method that combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs). It utilizes a vertical anchor mechanism to predict scores for candidate regions, enabling the detection of blurred text in natural scene images. However, its effectiveness in detecting skewed text is limited. Mask R-CNN [2], widely used in image segmentation, also performs well in text detection applications, excelling in recognizing irregularly shaped text instances. However, it struggles with small-sized or sparse text instances. PSENe [3] is a progressive scale expansion network that generates a series of masks of different sizes from a base image for feature fusion and text line prediction. It effectively addresses the issue of overlapping detections in cases of small text line spacing but suffers from slow detection speeds. FCENet [4] proposes representing the bounding curve of text using the parameters of the Fourier transform. This method improves the detection accuracy of highly curved text instances in natural scene text detection. However, its performance in non-curved text detection is average. DBNet [5] addresses the time-consuming post-processing issue associated with threshold-based binarization in segmentation-based methods. It introduces a learnable threshold and designs a binarization function approximating a step function, allowing the segmentation network to learn the threshold during training. This not only improves the accuracy but also simplifies the post-processing. DBNet++ [6] builds upon DBNet by incorporating the ASF module to optimize the performance of multiscale segmentation, enhancing text detection. However, both DBNet and DBNet++ exhibit suboptimal performance when dealing with long text instances.

In the field of text recognition, ASTER [7] is a text recognition method with flexible correction capabilities. It employs a spatial transformer network to rectify input images with irregular text and utilizes a sequence recognition network based on attention mechanisms to achieve end-to-end text recognition. However, its recognition performance is constrained by the geometric features of the characters, and the model is susceptible to background noise. SAR [8] addresses irregular text scenes by proposing a weakly supervised approach that utilizes a two-dimensional attention mechanism module for character localization and individual recognition. It does not require character-level annotations, effectively enhancing the recognition accuracy of irregular text. However, the model training process is complex, leading to longer training times.

MASTER [9] adopts a multi-directional non-local self-attention module to address attention shift issues. During training, adjustments are made, and by creating a lower triangular mask matrix, the decoder can simultaneously output predictions for all the time steps, improving the training parallelism and significantly reducing the training time. However, the implementation cost is higher. ABINet [10] introduces a novel bidirectional cloze network (BCN) based on bidirectional feature representation as a language model. It iteratively corrects the features obtained from the visual model, effectively mitigating the impact of noisy inputs and achieving good performance on low-quality images. However, it faces challenges in achieving fully accurate recognition in long text lines with numerous characters. MATRN [11] enhances the text recognition performance by spatially encoding semantic features, connecting semantic features and visual features, and performing cross-modal feature enhancement on both feature types.

Furthermore, end-to-end natural scene text recognition methods based on deep learning integrate text detection and recognition tasks into a unified network model. STAR-Net [12] combines a spatial attention mechanism and residual connections to effectively capture and utilize the spatial information from input images for accurate text recognition. However, it struggles to accurately locate the text regions in blurry or low-resolution text images. TextNet [13] has developed a scale-aware attention mechanism, which learns from multiscale image features as the backbone network, sharing fully convolutional features and computations for localization and recognition. This approach is applicable to both regular and irregular text but exhibits mediocre performance in recognizing curved text.

Lyu et al. introduced Mask TextSpotter [14], an extension of Mask R-CNN, which adds an additional branch for single-character instance segmentation after RoI Align. Text recognition in this method depends on single-character classification from this branch. It is capable of detecting and recognizing text of arbitrary shapes but requires character-level annotations during training. Liao et al. proposed Mask TextSpotter v3 [15], which employs a Segmentation Proposal Network (SPN) instead of an RPN to predict salient maps for arbitrary-shaped text. Subsequently, a Hard RoI Masking operation is performed based on the mask of each text region, significantly improving the model's shape robustness for detecting arbitrary-shaped text. Yue et al. introduced RobustScanner [16] to address attention drift issues. This method proposes a novel position-enhanced branch and dynamically integrates its output with the decoder attention module output, offering improved robustness and practicality. While end-to-end methods simplify the overall model design and training process, debugging and optimization may consequently become more challenging.

In this study, we encountered challenges such as the complex backgrounds, high contrast, and brightness differences in the X-ray weld seam images, leading to significant visual differences between the text and background. Additionally, factors like irregular engineering operations may cause the text in weld seam images to appear blurred, distorted, and with varying lengths of text lines. To achieve good detection and recognition results under these interferences, we first constructed two datasets for the text detection and recognition tasks. Then, through a series of experimental analyses, we employed the DBNet model for text detection tasks and improved the post-processing methods to enhance the text detection accuracy. For text recognition tasks, we adopted the MATRN model and enhanced the model performance by introducing the CA [17], CBAM [18], and HFA [19] modules, thus improving the text recognition accuracy. Finally, the experimental results demonstrate that the proposed method performs excellently in adapting to the unique scenario of X-ray weld seam images.

The main contributions of this paper are as follows:

1. Through experimental analysis, we employed the DBNet model for weld seam image text detection and combined it with post-processing methods to improve the detection accuracy.
2. We improved the MATRN model, achieving a higher text recognition accuracy when targeting weld seam image datasets compared to other reference models. Our improved model also outperforms the original model in comparison with publicly available datasets.
3. We introduced a CA module into the backbone network to better extract visual features, thus enhancing the model performance. Additionally, we integrated the CBAM and HFA modules into the position attention module for more accurate spatial localization of characters, thereby improving the model's recognition accuracy.

2. X-ray Weld Seam Image Text Recognition Model Framework

The presented framework for X-ray weld seam image text detection and recognition, based on deep learning, comprises two primary components: one is the weld seam image text detection model based on DBNet, and the other is the text recognition model based on MATRN. As depicted in Figure 1, preprocessed weld seam images are put into the text detection model for the extraction of text lines. Subsequently, the extracted text lines are fed into the text recognition model, yielding the output of text information.

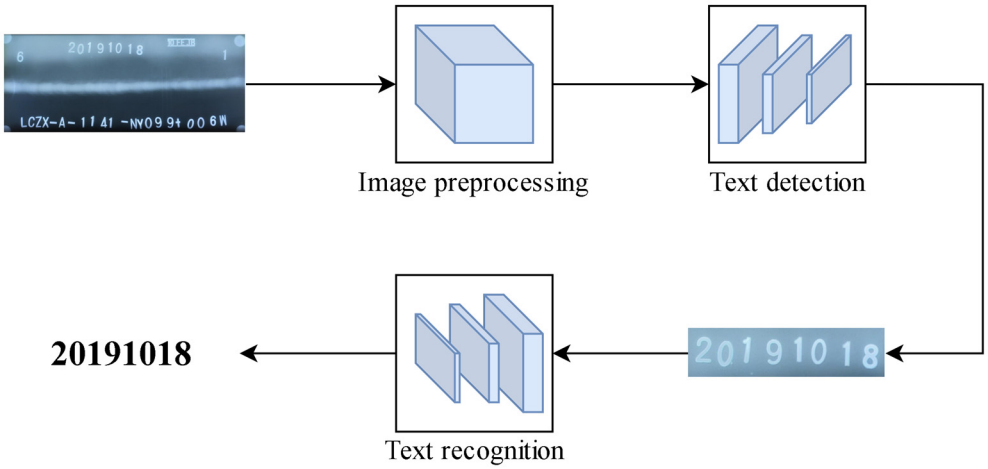


Figure 1. Weld seam image text recognition framework.

2.1. Text Detection Model Based on DBNet

DBNet is a classical text detection model based on segmentation algorithms. Its principle revolves around the introduction of a differentiable binarization method, addressing the challenge of threshold selection inherent in traditional segmentation algorithms. The network architecture of DBNet primarily consists of two modules: the Feature Pyramid Network (FPN) [20] and differentiable binarization (DB).

(1) The model initially utilizes a generic backbone network to extract the image features, with ResNet18 chosen to obtain feature maps at different scales. Feature maps corresponding to 1/2, 1/4, 1/8, 1/16, and 1/32 proportions of the original image are obtained. Subsequently, a top-down 2× downsampling is performed, and the resulting feature maps are merged with those obtained from bottom-up processing of the same size. Convolutional operations are then employed to eliminate the aliasing effects introduced during upsampling. After this FPN structure, four feature maps of sizes 1/4, 1/8, 1/16, and 1/32 of the original image are obtained. These feature maps are upsampled and unified into a 1/4-sized feature map, denoted as F.

(2) The feature map F undergoes a series of convolution and transposed convolution operations to obtain probability map P. A similar process is applied to obtain threshold map T.

(3) The probability map P and the threshold map T undergo the DB operation, resulting in approximate binary map B.

In traditional image segmentation algorithms, after obtaining the probability map, a standard binarization method is typically employed. This method involves processing the probability map using standard binarization, where pixels below a specified threshold are set to 0, and pixels above the threshold are set to 1. The formula is as follows:

$$B_{i,j} = \begin{cases} 1, if\ P_{i,j} \geq t \\ 0,\ otherwise \end{cases} \tag{1}$$

where t represents a pre-defined threshold, B is the binary map, P denotes the probability map, and (i,j) denotes the pixel coordinates. It is evident from this that standard binarization is non-differentiable, thereby precluding its integration into a network for optimization learning.

Differentiable binarization is the approximation of the step function in standard binarization, and the formula is as follows:

$$\mathcal{B}_{i,j} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \tag{2}$$

where k is the dilation factor, $\mathcal{B}$ is the approximate binary image, P is the probability map, T is the threshold map, and (i,j) are the coordinates of the pixel.

This form of approximated binarization function behaves similarly to a standard binarization function, but it is differentiable, allowing for optimization during the training of the segmentation networks. Differentiable binarization with adaptive thresholds not only distinguishes text regions from the background but also separates closely connected text instances. However, this approach may encounter issues of incomplete detection in long texts where the characters are closely connected. It might incorrectly detect a longer text as two or more shorter texts, thereby diminishing the performance of the detection model and adversely affecting the subsequent text recognition stages.

In accordance with the characteristics of weld seam image data, multiple short text lines are primarily distributed in the upper part of the weld seam images, while the lower part typically contains only one long text line. Consequently, the issue of incomplete detection primarily arises in the lower part of the images. To address this problem, we introduces a post-processing method based on the characteristics of the dataset. It involves merging multiple detected text line boxes into a single long text detection box, thereby improving the accuracy and performance of the detection model.

When multiple text detection boxes appear in the lower part of the image, we employ the Qhull algorithm [21] to merge them into a single detection box, aiming to encompass as much of the text area as possible. The Qhull algorithm is a commonly used 2D convex hull algorithm. It begins by constructing a 2D point set containing the coordinates of all the detection boxes. It then identifies the two points with the maximum distance in the set, connects them to form a line, and proceeds to find the point furthest from the line, creating a triangle. This process continues, rapidly expanding outward until all the points are within the convex hull. Finally, the algorithm filters the vertex set of the convex hull, determining the minimum and maximum x, y coordinates to construct a larger detection box. By employing this post-processing method, more complete text lines can be detected, effectively enhancing the accuracy of the text detection.

2.2. A Text Recognition Model Based on MATRN

We employ a text recognition model based on MATRN to recognize characters within the extracted text boxes containing various weld seam information. In order to enhance the accuracy of the text line recognition, we have introduced the CA, CBAM, and HFA modules to improve the network.

In the MATRN model, the detected text boxes undergo initial processing through the Feature Extraction Backbone, composed of ResNet and Transformer units [22], to generate visual features. Subsequently, the position attention module is employed to obtain the preliminary text recognition results. Following this, a language model is utilized for further optimization and correction of the obtained text recognition results, resulting in semantic features. To facilitate effective fusion of the visual and semantic features, a Multi-Modal Feature Enhancement module is employed to enhance both types of features. Finally, through the Output Fusion module, the enhanced visual features and semantic features are integrated to produce the ultimate recognition result. The formula is as follows:

$$V = F^{V.T}\left(F^{V.R}(X) + P^{V}\right) \tag{3}$$

where X represents the input image; $F^{V.R}$ corresponds to a ResNet unit; P^{V} denotes spatial position encoding, which is added to the input features through an addition operation; $F^{V.T}$ is a transformer unit; and V represents the visual features extracted through the backbone network.

$$A^{V-S} = softmax\left(\frac{P^{S}\mathcal{G}(V)^{T}}{\sqrt{D}}\right) \tag{4}$$

where P^S serves as the positional encoding for the character order, acting as the query vector. T represents the length of the character sequence, $\mathcal{G}(\cdot)$ is a small U-Net unit, $\mathcal{G}(V)$ functions as the key vector, D is the feature dimension, and A^{V-S} is the attention map calculated from the query vector and key vector.

$$Y_{(0)} = softmax\left(A^{V-S}\widetilde{V}W\right) \tag{5}$$

where $\widetilde{V}$ represents the flattened visual features, and through the attention map A^{V-S}, the visual features are abstracted into sequence features, denoted as $E^V = A^{V-S}\widetilde{V}$. W is a linear transformation matrix, and $Y_{(0)}$ is the character sequence feature obtained by applying a linear layer and softmax function.

$$S = F^{LM}\left(Y_{(0)}\right) \tag{6}$$

where F^{LM} represents the language model, comprising four transformer decoder units. S corresponds to the semantic features obtained by processing $Y_{(0)}$ through the language model.

$$P^{Align} = A^{V-S}\widetilde{P}^{V} \tag{7}$$

$$S^{Align} = S + P^{Align} \tag{8}$$

As A^{V-S} effectively provides visual features for the character estimation at each position, spatial position encoding can be introduced through flattened spatial position encoding $\widetilde{P}^{V}$ to determine the spatial positions of the semantic features, denoted as P^{Align}. Subsequently, the spatial position information is combined with the semantic features to obtain spatially aligned semantic features, S^{Align}.

$$G = \sigma\left(\left[E^{V^M}; S^M\right]W^{gated}\right) \tag{9}$$

$$F = G \bigodot E^{V^M} + (1 - G) \bigodot S^M \tag{10}$$

where W^{gated} is a weight matrix, [;] denotes concatenation, and $\odot$ represents element-wise multiplication. S^M and V^M, respectively, represent the multimodal semantic features and multimodal visual features obtained by enhancing the aligned semantic feature S^{Align} and visual feature V using a multimodal transformer unit [23]. E^{V^M} is the serialized representation of V^M. Subsequently, a gating mechanism is employed to combine the two sequence features, resulting in the feature F.

$$P = softmax(linear(F)) \tag{11}$$

where F is passed through a linear layer and softmax function to obtain the final predicted character sequence.

We introduce the CA module and integrate it into the backbone. The backbone is composed of ResNet units and Transformer units, with ResNet containing five layers and each layer consisting of a different number of BasicBlocks (Figure 2), with quantities of [3, 4, 6, 6, 3]. A CA module is added after each layer. The introduction of the CA module allows the model to flexibly focus on features at different levels, aiding in capturing local details and global features of different scales and dimensions more effectively. The Position Attention module aims to transform visual features into character probabilities, simultaneously determining the position information for each character and generating the corresponding attention map. We integrate the HFA module into the U-Net structure [24] within the Position Attention module. The U-Net structure can capture context information at different scales and essential features of the image, while the HFA module can more

accurately focus on the position information on the characters. Additionally, the CBAM module is introduced to enhance visual features, aggregating the spatial information and channel information of the features at a deeper level. The improved network structure is depicted in Figure 3.

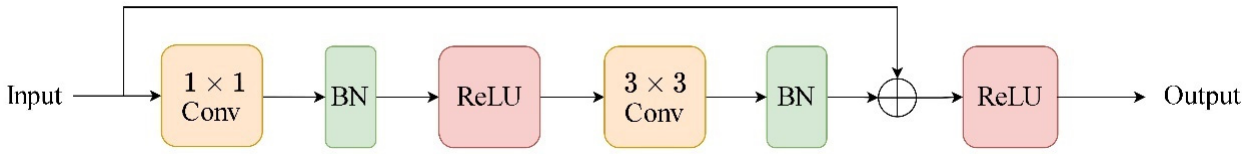


Figure 2. BasicBlock.

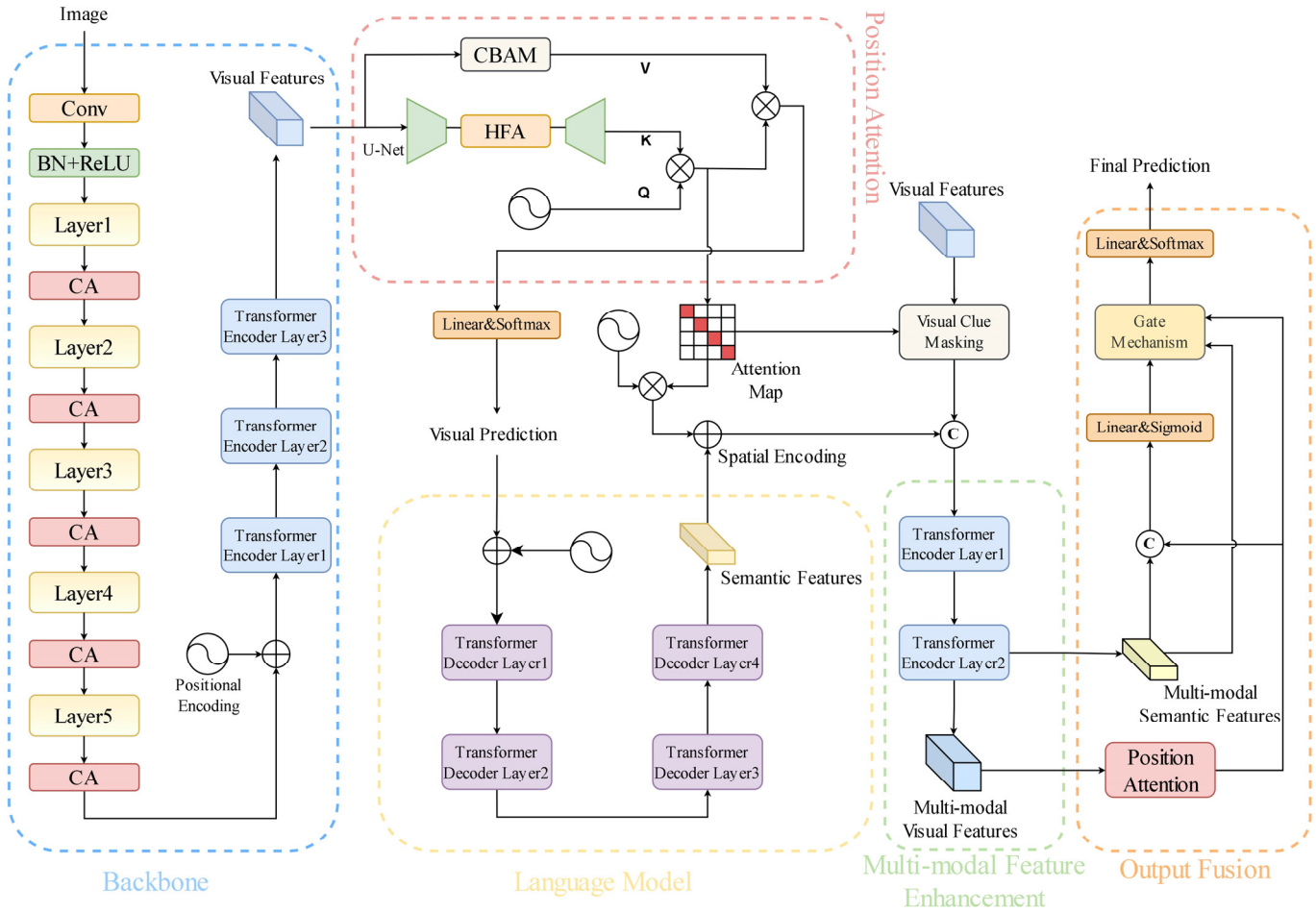


Figure 3. Overall network structure diagram.

2.2.1. The CA Module

The CA module (Figure 4) is a lightweight attention mechanism capable of capturing not only cross-channel information but also directional and position-sensitive information. This aids the text recognition model in more accurately locating and recognizing character regions.

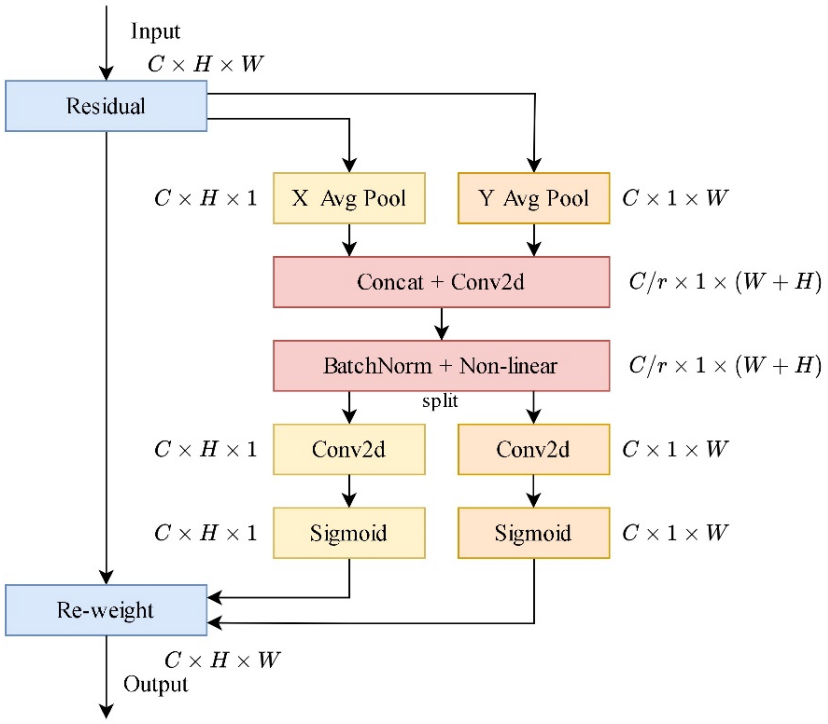


Figure 4. The structure of the CA module.

CA performs global average pooling operations in the horizontal and vertical directions on the input feature map, obtaining two aggregated features in the spatial directions. Subsequently, the obtained feature maps are individually encoded into a pair of direction-aware and position-sensitive attention maps. These attention maps can be complementarily applied to the input feature map to enhance the representation of the regions of interest.

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i). \tag{12}$$

where z_c^h represents the output of channel c at height h.

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w). \tag{13}$$

where z_c^w represents the output of channel c at width w.

The two aforementioned feature transformations enable the model to capture long-range dependencies in one spatial direction while preserving precise position information in another direction. This helps the network more accurately locate the position relationships of characters in long texts.

2.2.2. The CBAM Module

The CBAM module (Figure 5) is a feature-enhancing module that consists of two sub-modules: channel attention and spatial attention.

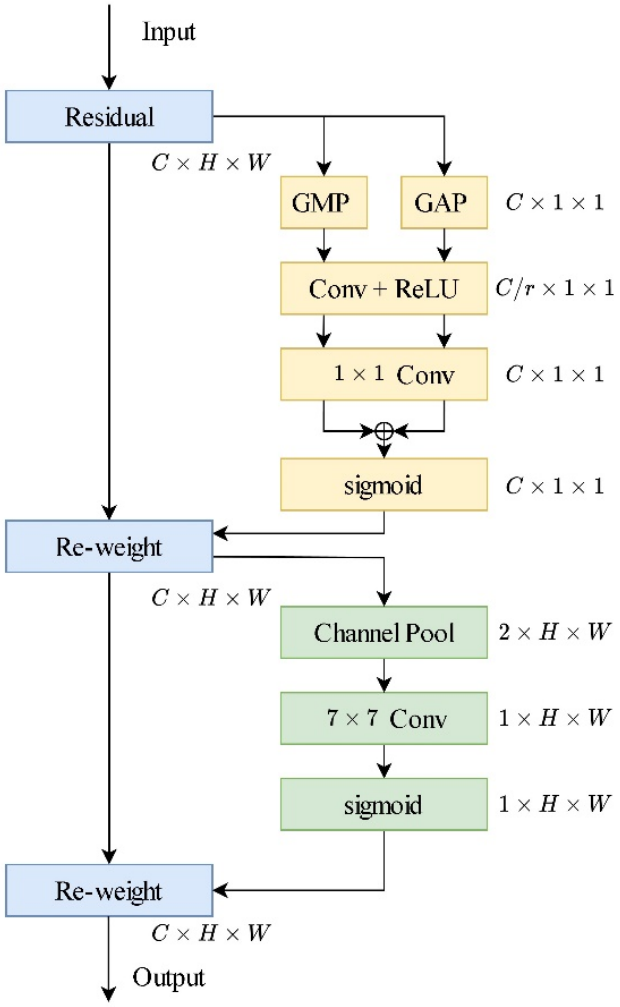


Figure 5. The structure of CBAM module.

(1) Channel attention: This part determines which channels are more important to the current task by learning the correlation between different channels in the feature map. Initially, the spatial information of the feature map is aggregated through adaptive average pooling and adaptive max pooling operations to generate average-pooled feature F^c_{avg} and max-pooled feature F^c_{max}. These two features are then individually passed through the same MLP structure, and the merged output feature vector produces channel attention map M_C.

$$\begin{aligned} M_C(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{max}))\right) \end{aligned} \tag{14}$$

σ represents the sigmoid function, and W_1 and W_0 are the weights of the MLP.

(2) Spatial attention: This part focuses on the correlation between different spatial positions in the feature map. It calculates the maximum and average values for each channel in the feature map, aggregating channel information to obtain two 2D feature maps F^s_{avg} and F^s_{max}. These two feature maps are then concatenated and passed through a standard convolutional layer with a kernel size of 7×7, resulting in spatial attention map M_S. This concentrates attention on the most important regions in the image, enhancing the representation of the regions of interest.

$$M_S(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) = \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \quad (15)$$

σ represents the sigmoid function, and $f^{7\times7}$ represents the convolution operation with a 7×7 kernel size.

After extracting the visual features, the CBAM module is applied to the high-level features, performing channel and spatial attention operations. This enables the network to focus more on important channels and spatial locations at higher levels of the feature representation, facilitating optimization of the feature representation. By enhancing the expressive power of the features, the network can more effectively handle text lines in welding images, leading to more accurate recognition and understanding of text lines in welding images. This enhances the accuracy and robustness of the text recognition.

2.2.3. The HFA Module

The U-Net structure in the Position Attention module performs multiple encoding and decoding operations on the visual features obtained through the backbone. Encoding operations progressively downsample the input features through multiple convolutional layers, reducing the size of the feature maps. This focuses on the higher-level features, providing a larger receptive field that helps the model capture global information more effectively. It aids in efficiently representing the overall structure and key features of the image. Decoding operations, achieved through upsampling, restore the encoded features to their original spatial dimensions, preserving more spatial location information. This contributes to accurate character localization.

However, our dataset contains numerous long text lines, for which the transformer architecture proves more effective. Therefore, we introduce the Hybrid Feature Aggregation (HFA) structure (Figure 6) into the U-Net architecture. The HFA structure comprises positional encoding and transformer units, embedded after the encoding layers into the U-Net structure. Initially, positional encoding is applied to the input features. Subsequently, the features with positional encoding are fed into the transformer encoding layer. Through self-attention mechanisms, this layer captures the correlations between different positions, allowing the features at one position to interact with the features at all other positions, not confined to just neighboring positions. This capability aids in capturing long-range relationships, particularly pertinent to the inter-character relationships within a text line and the contextual information between characters. It facilitates the extraction of more abstract features, thereby better expressing the structure of the text lines.

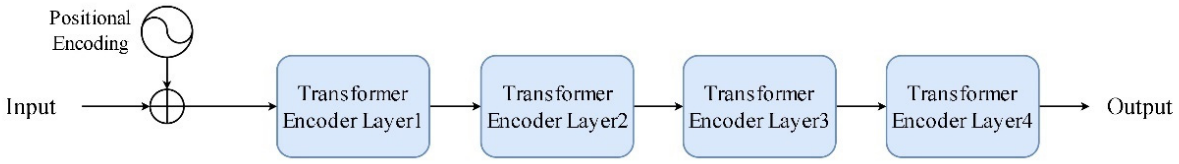


Figure 6. The structure of HFA module.

3. Experiments

3.1. Datasets

Due to the lack of publicly available X-ray welding seam image datasets, we established two welding seam image datasets specifically designed for text detection and text recognition experiments. The text detection dataset comprises complete welding seam images, each containing multiple text lines of varying lengths, as illustrated in Figure 7. This dataset consists of 2300 training images, 460 testing images, and 308 validation images. Meanwhile, the text recognition dataset consists of cropped single-line text images, as shown in Figure 8a,b. It includes 7565 training images, 1500 testing images, and 1458 validation images. By constructing these two datasets, we aimed to gain a deeper understanding

of and address the specific challenges associated with each task, ultimately providing more comprehensive and reliable solutions for integrated tasks in practical applications.

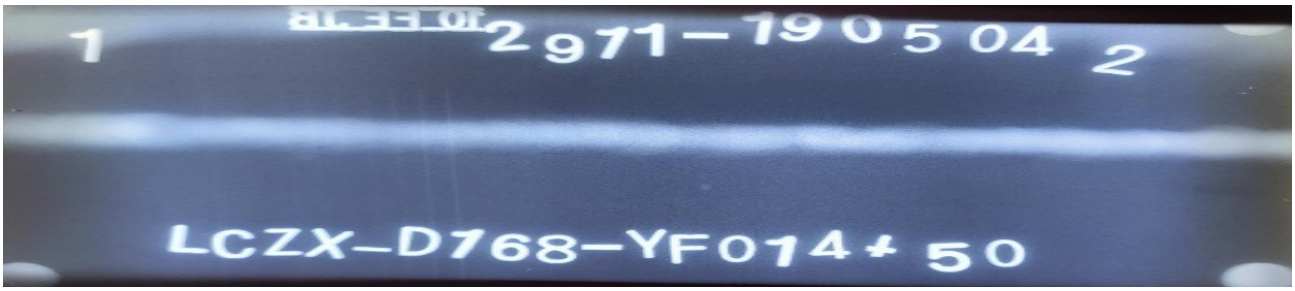


Figure 7. X-ray welding seam image.

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

Figure 8. Cropped image of the text line. (**a**) Serial number (**b**) Date.

3.2. Implementation Details

The experimental platform operates on Ubuntu 18.04, utilizing the PyTorch 1.10.2 deep learning framework. The hardware environment consists of an Intel(R) Core(R) i7-10700 CPU, 32 GB of RAM, and an NVIDIA(R) GTX (R) 3090 GPU.

In the text detection experiment, the input image size is set to 640 × 640, with a batch size of 32. The total number of epochs is set to 1200, and the SGD optimizer is employed with an initial learning rate of 7×10^{-3}. During training, the learning rate changes according to the formula $\left(1 - \frac{iter}{max\_iter}\right)^{power}$, where *iter* is the current iteration, *max_iter* is the maximum number of iterations, *power* is set to 0.9, and the weight decay is 10^{-4}.

In the text recognition experiment, the input image size is set to 32 × 320, with a batch size of 16. The total number of epochs is set to 600, and the Adam optimizer is used with an initial learning rate of 10^{-4}. The learning rate decays to 10^{-5} after 400 epochs.

3.3. The Evaluation Criteria

The text detection evaluation criteria include precision, recall, and the composite evaluation metric F1. The calculation methods are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{F1} = 2 \times \frac{P \times R}{P + R} \tag{18}$$

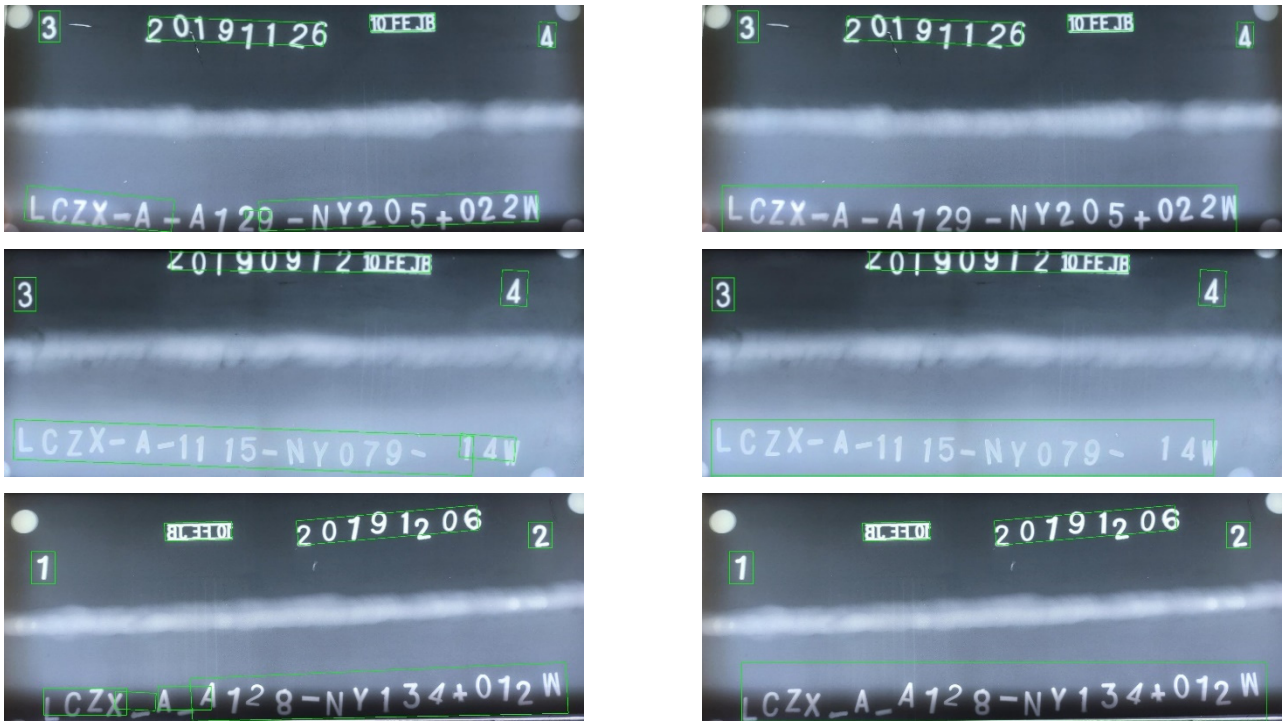
Predicted ground truth boxes with an Intersection over Union (IOU) ≥ 0.5 are considered positive samples, while samples with an IOU < 0.5 are considered negative samples. *TP* represents the number of instances where the model correctly predicts positive samples as positive, *FP* represents the number of instances where the model incorrectly predicts negative samples as positive, and *FN* represents the number of instances where the model incorrectly predicts positive samples as negative.

For text recognition, we employ text line recognition accuracy as the evaluation criterion. Specifically, when recognizing an image, any errors in detecting or omitting individual characters or multiple characters are considered recognition errors. Only when the entire text line is correctly identified in the recognition result is it considered a correct sample.

3.4. Evaluation

3.4.1. The Text Detection Experiment

In the text detection experiments, we compared the improved text detection method proposed in this paper with the DBNet method, and the experimental results are shown in Figure 9a,b. We observed that, especially when detecting longer text lines, the DBNet method has some issues, including incomplete detection of the text lines and the phenomenon of a long text line being recognized due to multiple overlapping bounding boxes. These problems significantly impact the subsequent text recognition, leading to a decrease in the recognition accuracy. However, by introducing a post-processing method, we successfully overcame these issues. The post-processing method ensures the complete and standardized detection of long text lines, providing high-quality data for downstream text recognition tasks.



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

Figure 9. Comparison chart of text detection results: (**a**) DBNet; (**b**) our.

Additionally, we compared our method with other state-of-the-art methods, and the experimental results are shown in Table 1. Although the precision of our method is slightly lower than the PSENet algorithm, both the recall and F1 value are superior to the other algorithms compared in this paper. Moreover, compared to DBNet, our method shows improvements of 1.9%, 0.3%, and 1.1% in precision, recall, and the F1 value, respectively.

Table 1. Comparison with the results of other text detection algorithms.

Model	Precision/%	Recall/%	F1/%	Parameters (M)
DBNet [5]	89.1	<u>92.1</u>	<u>90.6</u>	13.8
DBNet++ [6]	88.5	91.5	89.9	29.0
FCENet [4]	51.8	61.1	56.0	26.3
PSENet [3]	**93.5**	84.6	88.8	29.2
Mask R-CNN [2]	65.0	80.0	71.7	44.0
Our	<u>91.0</u>	**92.4**	**91.7**	13.8

Bold represents the best result, underline represents the second best result. This explanation applies to all tables below.

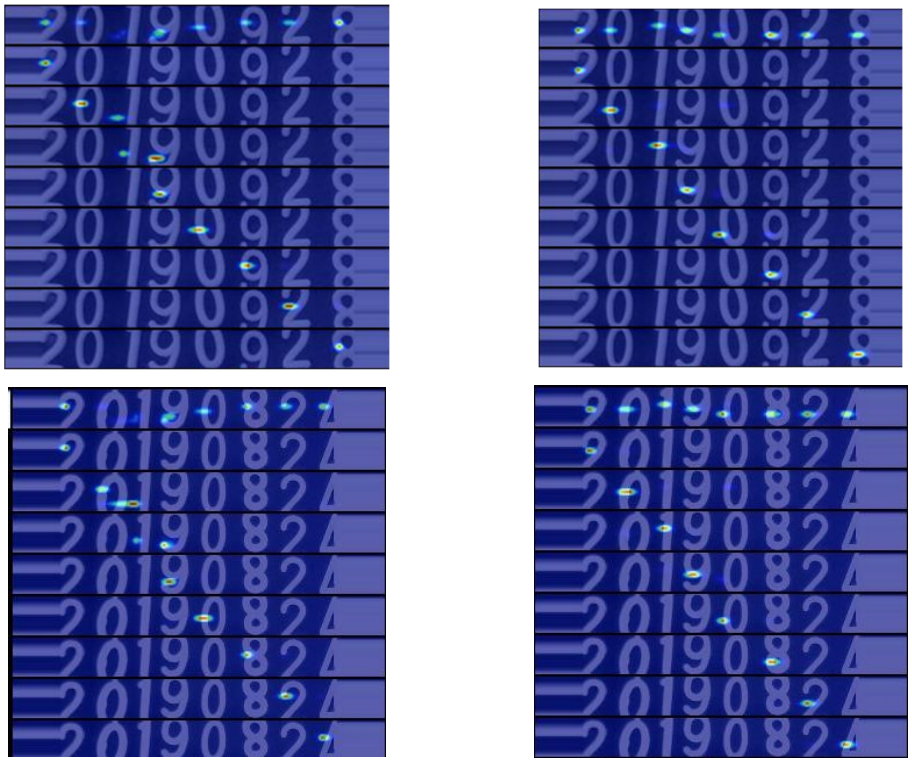
3.4.2. The Text Recognition Experiment

We conducted text recognition tests on the text line image dataset obtained in the text detection experiments using the recognition algorithm proposed in this paper and compared it with the current leading algorithms. The experimental results are shown in Table 2. Compared with the MATRN algorithm, our method achieved a 0.8% improvement in text line recognition accuracy and a 0.6% improvement compared to the currently best-performing PARSeq algorithm. This indicates the significant performance advantage of our method in the text recognition task with X-ray weld seam images.

Table 2. Comparison with the results of other text recognition algorithms.

Model	Precision/%	Parameters (M)
MATRN [11]	92.6	44.2
PARSeq [25]	92.8	36.9
SAR [8]	61.2	57.3
ABINet [10]	59.1	36.7
ASTER [7]	77.2	20.9
MASTER [9]	87.8	58.9
STAR-Net [12]	81.0	26.2
Our	**93.4**	46.5

Furthermore, during the text recognition testing phase, we generated attention maps for the MATRN algorithm and our method, as shown in Figure 10a,b. From these attention maps, we can clearly observe that our method is more accurate in capturing the character positional information within the text lines, contributing to better recognition results.



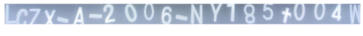
(**a**) (**b**)

Figure 10. Comparison of attention result images: (**a**) MATRN; (**b**) our.

Table 3 presents the recognition results of the MATRN algorithm and our method, where red characters indicate recognition errors, and "_" represents unrecognized charac-

ters. It can be observed from the table that our method achieves more accurate recognition for some incomplete characters.

Table 3. Comparison of recognition results.

X-ray Welding Seam Image	Ground Truth	MATRN	Our
	LCZX-A-2006-NY185+004W	LC7X-A-2006-NY185+004W	LCZX-A-2006-NY185+004W
	LCZX-A-1115-NY-132-008	LCZX-A-1115-NY-132-006	LCZX-A-1115-NY-132-008
	20190924	20190024	20190924
	20190823	20190723	20190823
	LCZX-B-273-SCXX5111-NY702-013W	LCZX-B-273-SCXX5111-NY702-013_	LCZX-B-273-SCXX5111-NY7022013W

Where red characters indicate recognition errors, and "_" represents unrecognized characters.

To validate the efficacy of the proposed methodology, experiments were conducted on a widely adopted dataset. Specifically, 1/10 of the MJSynth dataset [26] comprising 0.8 M images was chosen as the training set. Three datasets, namely ICDAR2003 (IC03) [27] with 860 images, ICDAR2013 [28] consisting of 857 images (IC13S), and 1015 images (IC13L), were employed as the testing datasets. The input image dimensions were set to 32×128, with a batch size of 48. The total number of epochs was set to 20, utilizing the Adam optimizer with an initial learning rate of 10^{-4}, which decayed to 10^{-5} after 12 epochs. The experimental results are presented in Table 4.

Table 4. Comparison of results on public datasets.

Model	Test Dataset		
	IC03	IC13S	IC13L
MATRN	0.919	0.914	0.893
Our	**0.923**	**0.919**	**0.894**

3.4.3. Ablation Study

To delve into the interrelationships between the introduced three modules and their impact on the system performance, we conducted relevant ablation experiments, and the results are presented in Table 5. The following conclusions can be drawn from the results: Introducing one or two modules does enhance the text recognition performance in certain cases, but it does not achieve optimal results. However, when a combination of the three modules CA, CBAM, and HFA is introduced simultaneously, significantly superior performance can be achieved.

Table 5. Comparison of ablative experiment results.

CA	CBAM	HFA	Precision/%	Parameters (M)
			92.6	44.2
✓			93.1	44.2
	✓		91.8	44.2
		✓	92.8	46.3
✓	✓		92.7	44.4
✓		✓	92.8	46.5
	✓	✓	92.4	46.5
✓	✓	✓	**93.4**	46.5

The checkmark indicates that the corresponding model adopts this module.

We also investigated the impact of the number of CA modules on the model performance, and the experimental results are shown in Table 6. It can be observed that adding

a CA module after each layer achieves the optimal performance of the model. This may be because adding a CA module after each layer better captures the global information at different levels, making it easier for the model to learn the features at each level. Additionally, due to the simple structure of the CA module, adding multiple CA modules does not significantly increase the model's parameter count, thereby not affecting the model performance.

Table 6. Ablation study on the number of CA modules.

Number	Precision/%	Parameters (M)
0	92.6	44.2
1	92.0	46.5
2	91.5	46.5
3	91.4	46.5
4	92.7	46.5
5	**93.4**	46.5

On the foundation of our proposed method, we further investigated the impact of different attention mechanisms on the model performance, as presented in Table 7. Specifically, A replaces the CA module with the CoT [29] module, B replaces the CA module with the EMA [30] module, and C substitutes the CBAM module with the CA module. However, the results indicate that these alternatives did not outperform our proposed method.

Table 7. Ablation study of other backbone networks and attention modules.

Model	Precision/%	Parameters (M)
A	91.6	49.5
B	92.0	46.5
C	92.7	46.5
D	92.3	75.4
E	91.7	54.7
F	92.1	80.5

Additionally, we explored the influence of different backbone architectures on the model performance, as illustrated in Table 7. Here, D corresponds to Res2net101 [31], E to UniNeXt [32], and F to FasterNet [33]. The experiments suggest that when replacing the backbone with a more complex and deeper network, the parameter count significantly increases, but the accuracy decreases. This phenomenon might be attributed to the relatively small size of the dataset, leading to overfitting issues during the training process.

4. Conclusions

This paper introduces a deep-learning-based algorithm for text recognition in X-ray weld seam images, successfully achieving automated recognition of the textual information in X-ray weld seam images. The algorithm comprises two key components: text detection and text recognition. Firstly, the text detection model accurately locates the positions of the text lines. Subsequently, these text lines are extracted from the weld seam images, and text recognition is performed, ultimately yielding precise text recognition results. By incorporating post-processing techniques into the text detection model, we achieve accurate detection of partially longer text lines, achieving an accuracy of 91.0%, a recall of 92.4%, and an F1 score of 91.7% on the test dataset.

In the text recognition phase, we introduce modules such as CA, CBAM, and HFA to further enhance the accuracy of the text recognition. Compared to other advanced algorithms, our text recognition algorithm demonstrates excellent performance on the test dataset, achieving a text line recognition accuracy of 93.4%. However, it is essential to note

that challenges exist in weld seam images, and our recognition algorithm may encounter issues in accurately recognizing characters with tightly connected text lines, as shown in Table 3. Therefore, further research on accurately recognizing densely arranged characters remains a crucial direction for future studies.

Author Contributions: Conceptualization, Y.Z. and Q.Z.; methodology, Y.Z. and Q.Z.; software, Q.Z.; validation, Q.Z.; investigation, Y.Z.; resources, Y.Z. and Q.Z.; data curation, Q.Z.; writing—original draft preparation, Y.Z. and Q.Z.; writing—review and editing, Y.Z. and Q.Z.; visualization, Y.Z. and Q.Z.; supervision, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Yunnan Provincial Agricultural Basic Research Joint Special Project (Grant No. 202101BD070001-042), and the Yunnan Ten-Thousand Talents Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public dataset is available at: https://www.dropbox.com/sh/i39abvnefllx2si/AAAbAYRvxzRp3cIE5HzqUw3ra?dl=0 (accessed on 17 February 2024). Weld seam images dataset is available at: https://zenodo.org/records/10618962 (accessed on 17 February 2024).

Conflicts of Interest: The authors declare no conflict 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: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part VIII 14*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 56–72.
2. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
3. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9336–9345.
4. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier contour embedding for arbitrary-shaped text detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3123–3131.
5. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11474–11481.
6. Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; Bai, X. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 919–931. [CrossRef] [PubMed]
7. Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2035–2048. [CrossRef] [PubMed]
8. Li, H.; Wang, P.; Shen, C.; Zhang, G. Show, attend and read: A simple and strong baseline for irregular text recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8610–8617.
9. Lu, N.; Yu, W.; Qi, X.; Chen, Y.; Gong, P.; Xiao, R.; Bai, X. Master: Multi-aspect non-local network for scene text recognition. *Pattern Recognit.* **2021**, *117*, 107980. [CrossRef]
10. Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; Zhang, Y. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7098–7107.
11. Na, B.; Kim, Y.; Park, S. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 446–463.
12. Liu, W.; Chen, C.; Wong, K.-Y.K.; Su, Z.; Han, J. Star-net: A spatial attention residue network for scene text recognition. In Proceedings of the BMVC, York, UK, 19–22 September 2016; Volume 2, p. 7.
13. Sun, Y.; Zhang, C.; Huang, Z.; Liu, J.; Han, J.; Ding, E. Textnet: Irregular text reading from images with an end-to-end trainable network. In *Proceedings of the Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018, Revised Selected Papers, Part III 14*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 83–99.
14. Lyu, P.; Liao, M.; Yao, C.; Wu, W.; Bai, X. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 67–83.

15. Liao, M.; Pang, G.; Huang, J.; Hassner, T.; Bai, X. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XI 16*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 706–722.
16. Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; Zhang, W. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 135–151.
17. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
18. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
19. Fang, S.; Mao, Z.; Xie, H.; Wang, Y.; Yan, C.; Zhang, Y. Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7123–7141. [CrossRef] [PubMed]
20. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
21. Barber, C.B.; Dobkin, D.P.; Huhdanpaa, H. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* **1996**, *22*, 469–483. [CrossRef]
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
23. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Conference. Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; NIH Public Access: Bethesda, MD, USA, 2019; Volume 2019, p. 6558.
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
25. Bautista, D.; Atienza, R. Scene text recognition with permuted autoregressive sequence models. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 178–196.
26. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv* **2014**, arXiv:1406.2227.
27. Lucas, S.M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; Young, R.; Ashida, K.; Nagai, H.; Okamoto, M.; Yamamoto, H.; et al. ICDAR 2003 robust reading competitions: Entries, results, and future directions. *Int. J. Doc. Anal. Recognit.* **2005**, *7*, 105–122. [CrossRef]
28. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L.G.I.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazàn, J.A.; de las Heras, L.P. ICDAR 2013 robust reading competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1484–1493.
29. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef] [PubMed]
30. 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; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
31. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P.H. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
32. Lin, F.; Yuan, J.; Wu, S.; Wang, F.; Wang, Z. UniNeXt: Exploring A Unified Architecture for Vision Recognition. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023.
33. Chen, J.; Kao, S.-H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Gary Chan, S.-H. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12021–12031.