

Artistic-style text detector and a new Movie-Poster dataset

Aoxiang Ning^a, Yiting Wei^b, Minglong Xue^{a,*} and Senming Zhong^c

^aCollege Of Computer Science And Engineering, Chongqing University of Technology, 400054, China

^bCollege of Agronomy and Biotechnology, Southwest University, 400715, China

^cIndustrial Training Center, Guangdong Polytechnic Normal University, Guangzhou, 510665, China

ARTICLE INFO

Keywords:

Text detection

Feature enhancement

Boundary proposal

Artistic-style text

ABSTRACT

Although current text detection algorithms demonstrate effectiveness in general scenarios, their performance declines when confronted with **artistic-style text featuring complex structures**. This paper proposes a method that utilizes **Criss-Cross Attention** and **residual dense block** to address the **incomplete and misdiagnosis of artistic-style text detection** by current algorithms. Specifically, our method mainly consists of a feature extraction backbone, a feature enhancement network, a multi-scale feature fusion module, and a boundary discrimination module. The feature enhancement network significantly enhances the model's perceptual capabilities in complex environments by **fusing horizontal and vertical contextual information**, allowing it to capture detailed features overlooked in artistic-style text. We incorporate residual dense block into the Feature Pyramid Network to suppress the effect of background noise during feature fusion. Aiming to omit the complex post-processing, we explore a **boundary discrimination module** that guides the correct generation of boundary proposals. Furthermore, given that movie poster titles often use stylized art fonts, we collected a **Movie-Poster** dataset to address the scarcity of artistic-style text data. Extensive experiments demonstrate that our proposed method performs superiorly on the Movie-Poster dataset and produces excellent results on multiple benchmark datasets. The code and the Movie-Poster dataset will be available at: <https://github.com/biedaxiaohua/Artistic-style-text-detection>

1. Introduction

Text detection, the process of locating and extracting textual information from various surfaces such as images, videos, or 3D environments, is a fundamental component in computer vision and artificial intelligence Long et al. (2021). In the digital age, where multimedia content is abundant, accurate text detection is crucial to enhance search capabilities, automate data entry, and enable intelligent image understanding. Following the groundbreaking success of convolutional networks exemplified by AlexNet Krizhevsky et al. (2017), the field of text detection has experienced swift advancement. Several regression-based methods (Tian et al. (2016); Liao et al. (2017); Zhou et al. (2017); Wang et al. (2020b)) inspired by general object detection algorithms (Girshick (2015); Fan et al. (2018, 2019)), employ rectangles or quadrilaterals for text instance localization. These methods have difficulties in detecting arbitrary shape text. To represent text instances flexibly, segmentation-based methods (Deng et al. (2018); Wang et al. (2019c,b); Liao et al. (2020); Zhu et al. (2021)) treat text detection as a semantic segmentation task. These methods effectively facilitate text detection in general scenes. However, specific text environments pose unique challenges that require further research. As shown in Fig. 1, titles in movie posters are often artistically rendered to have a variety of shapes. As shown in Fig. 2, current algorithms have difficulty accurately recognizing and segmenting text regions, hindering effective feature learning. On the other hand, there is very

little artistic-style text data on the market now. Therefore, detecting artistic-style text and supplementing artistic-style data is challenging and meaningful.

In existing datasets such as Total-Text Ch'ng & Chan (2017) and CTW1500 Yuliang et al. (2017), the majority of text instances have a more standardized shape, with only a few artistic-style text included. Therefore, we collected the **Movie-Poster** dataset to enrich the currently scarce artistic-style text data. Specifically, this dataset contains 1500 movie posters, of which 1100 are used for training and 400 for testing. In our dataset, poster titles are often treated artistically. As shown in Fig. 1(a), (c), and (d), these titles are highly personalized and stylized; they can appear at any angle, and there may be connections and overlaps between characters. As shown in Fig. 1(b) and (e), some instances have significant differences in aspect ratios, and image background elements may be mixed with the text, making the detection of text regions more difficult.

This paper introduces a novel feature enhancement network, primarily composed of the **Recycle Criss-Cross Attention module (RCCA)**. This module allows for effective information transfer and interaction between feature maps of different channels, capturing semantic information about different objects, textures, and structures in an image to enhance the model's perceptual capabilities. It effectively addresses the issue of incomplete detection of shape-changing artistic-style text, as shown in Fig. 2(a) and (c). TextPMS Zhang et al. (2022) and TextBPN++ Zhang et al. (2023) fuse the feature maps output by backbone through the Feature Pyramid Network (FPN), which effectively improves the representation of features. However, it still suffers from background noise interference for artistic-style text with complex backgrounds. Based on this, we design a new

*Corresponding author at: College Of Computer Science And Engineering, Chongqing University of Technology, 400054, China
ningax@stu.cqut.edu.cn (Aoxiang Ning);
wy18439532876@email.swu.edu.cn (Yiting Wei); xueml@cqut.edu.cn (Minglong Xue); itcsmzhong@gpnu.edu.cn (Senming Zhong)



Fig. 1: Some samples from the Movie-Poster dataset. Text in (a), (c), and (d) are highly artistic-style, text in (b) has widely varying aspect ratios, and in (e), the text area is mixed with background pixels.

multi-scale feature fusion approach **Residual Feature Pyramid Network (R-FPN)**, combining residual dense structure and FPN to suppress the effect of redundant information effectively. As shown in Fig. 2(b), we have effectively solved the problem of incorrectly recognizing non-text pixels as text pixels. As a mixture of strokes and background pixels, text regions are highly homogeneous textures that do not have natural and well-defined boundaries. TextBPN++ Zhang et al. (2023) systematically presents a unified coarse-to-fine framework via boundary learning for arbitrary shape text detection. However, this method can lead to boundary modelling distortions for artistic-style text areas with extreme aspect ratios. To this end, we explore a **Boundary Discriminant Module (BDM)** that guides the generation of boundary proposals by combining the priori information with the feature maps output from the feature enhancement network. Subsequently, the boundary proposals are fed into the boundary transformer module for refinement, thus omitting a series of complex post-processing procedures.

In summary, our main contributions are four-fold:

- We propose a new feature enhancement network and feature fusion method that greatly enhances the model's perceptual capabilities, effectively mitigates the problem of incomplete detection and suppresses the effect of background noise.
- We further explore a boundary discrimination module that guides the accurate generation of boundary proposals.
- We propose the Movie-Poster dataset to complement the existing shortage of artistic-style text data.
- Extensive experiments have demonstrated that our method achieves state-of-the-art performance on the Movie-Poster dataset and is competitive on the publicly available datasets.

The rest of the paper is organized as follows: Sect. 2 overviews the related work. Sect. 3 elaborates on our work. In Sect. 4, we demonstrate some experimental results and

analysis. Sect. 5 further discussed our proposed Movie-Poster dataset. Finally, we conclude our work in Sect. 6.

2. Related work

With the rapid development of deep learning and object detection technology (Kim et al. (2018); Kessentini et al. (2019); Jazayeri et al. (2019)), significant progress has been made in the field of text detection Liu et al. (2019). Text detection can be broadly categorized into regression-based, segmentation-based, and connected component-based methods.

2.1. Regression-based Methods

The regression-based methods (Wu & Natarajan (2017); Zhou et al. (2017); Liao et al. (2017); Ma et al. (2018); Wang et al. (2018); Xu et al. (2019); Zhu & Du (2021); Zhong et al. (2022)) rely on a regression-based object detection framework with word-level and line-level a priori knowledge. The difficulty of text detection is that text has irregular shapes with various aspect ratios, unlike standard images. RRPN Ma et al. (2018) and TextBoxes Liao et al. (2017) effectively address this problem by predicting anchor offsets to localize text boxes. Other methods (Zhou et al. (2017); He et al. (2021, 2017b)) are anchor-free methods that directly regress the offset from a boundary or vertex to the current point. EAST Zhou et al. (2017) directly predicts text or lines of text in images with arbitrary orientations and rectangular shapes, eliminating unnecessary intermediate steps. SAST Wang et al. (2019a) draws on the ideas of TextSnake Long et al. (2018) and EAST Zhou et al. (2017) and joins the prediction of some geometric features of the text (the text centerline region, text boundary bias, and text center bias, etc.) while directly regressing on the bounding box so that it can be applied to irregular text detection. MOST He et al. (2021) can dynamically adjust the receptive field of the localized prediction layer and adaptively merge the raw detections according to the predicted position, resulting in better detection of instances with large aspect ratios. TPLAANet Zhong et al. (2022) explores the prediction center mask to address the

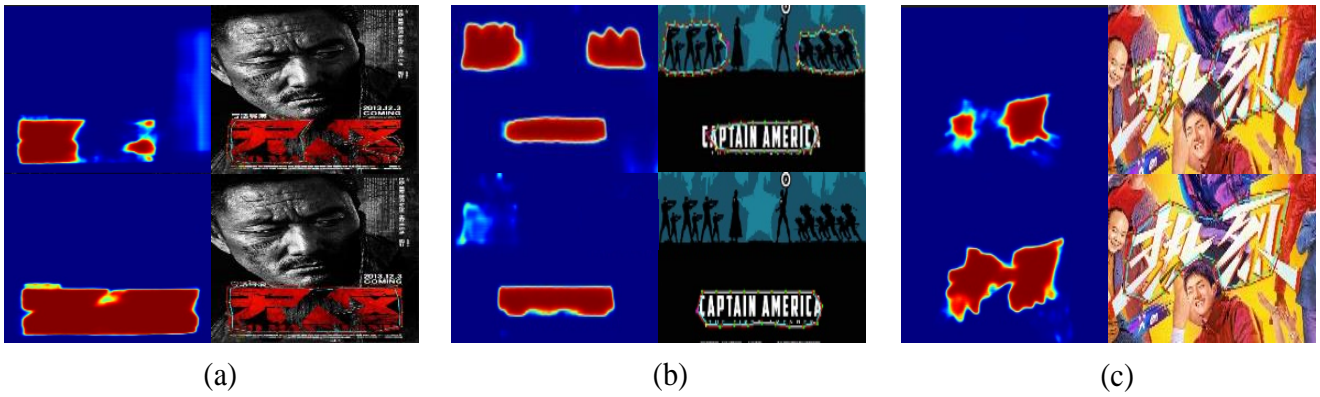


Fig. 2: Comparison of the effectiveness of our method with TextBPN++ on predicting mask maps and boundary proposals for the poster dataset. The visualization results of the TextBPN++ are on top of the image, and our method's results are on the bottom. (a) and (c) demonstrate that we have solved the problem of incomplete detection. (b) demonstrates that we have solved the problem of non-text pixels being mistaken for text pixels.

challenge of detecting text. Although the regression-based method is effective in quadrilateral text detection, it is not adapted to arbitrary shape text detection.

2.2. Segmentation-based Methods

Inspired by image segmentation methods (He et al. (2017a); Badrinarayanan et al. (2017); Schu & Scharcanski (2018); Yang et al. (2020); Farshi et al. (2020)), segmentation-based text detection methods (Wang et al. (2019b); Liao et al. (2020, 2022); Li et al. (2024a); Yu et al. (2023)) classify pixels at the pixel level to discriminate whether each pixel point belongs to a text target and its connectivity with the surrounding pixels. Then, they integrate the results of neighbouring pixels into a text box. These methods can be adapted to any shape and angle of the text. For example, PSENet Wang et al. (2019b) localizes text of any shape by pixel-level segmentation and uses a progressive scale expansion algorithm to identify neighbouring text instances. PAN Wang et al. (2019c) improves on PSENet Wang et al. (2019b) by using a learnable post-processing method, pixel aggregation, to guide text pixels to correct kernel parameters by predicted similarity vectors and to reconstruct complete text instances from the predicted kernel to reconstruct complete text instances. LSAE Tian et al. (2019) proposed pixel embedding, which groups pixels based on segmentation results to achieve more accurate text localization. DBNet Liao et al. (2020) embeds the threshold transformation process into the network for training by learning the threshold mapping and employing differentiable operations. In this way, the text detection model can adaptively learn the thresholds to capture the segmentation information of text more efficiently. DBNet++ Liao et al. (2022) adds the Adaptive Scale Fusion (ASF) module to DBNet Liao et al. (2020). Features at different scales are processed through the ASF module to obtain better-fused features. CBNet Zhao et al. (2024) uses knowledge distillation to detect arbitrary shape text efficiently. The most important thing for segmentation-based methods is to ensure the accuracy of segmentation.

2.3. Connected Component-based Methods

Connected Component-based Methods usually link or group the detected individual text parts or characters into final text instances through a post-processing process, for example (Yin et al. (2015); Shi et al. (2017); Baek et al. (2019); Feng et al. (2019); Tang et al. (2019); Zhang et al. (2020)). CRAFT Baek et al. (2019) innovatively breaks down the text detection task into two more recognizable components: fragments and links. In this framework, each text part is treated as separate fragments, while "links" connect two neighbouring fragments belonging to the same word, thus enabling complete detection of the entire word. SegLink Shi et al. (2017), inspired by TextSnake Long et al. (2018), simply groups detected local text regions by their geometric relationships. DRRG Zhang et al. (2020) uses Graph Convolutional Neural Networks Kipf & Welling (2016) (GCN) to learn and infer linking relationships of text components to group text components. Although linked component-based methods can work well for arbitrary shape text detection, they usually have complex post-processing procedures.

3. Methodology

3.1. Overall network architecture

Our proposed model architecture is shown in Fig. 3. It consists of five main components: feature extraction backbone, feature enhancement network, feature fusion module, boundary discrimination module, and boundary transformer. Specifically, we adopt ResNet-50 He et al. (2016) as the backbone to extract features. We input the feature maps output from the backbone to the feature enhancement module, RCCA, which can accurately capture the semantic information in complex images more acutely. Further, we propose a novel feature fusion strategy, R-FPN, to integrate the multi-scale feature maps processed by feature enhancement. After that, the coarse boundary proposals generated by BDM are utilized to sample the feature maps and extract key node features. Finally, these features are fed into the boundary transformer for refinement to achieve more accurate text bound-

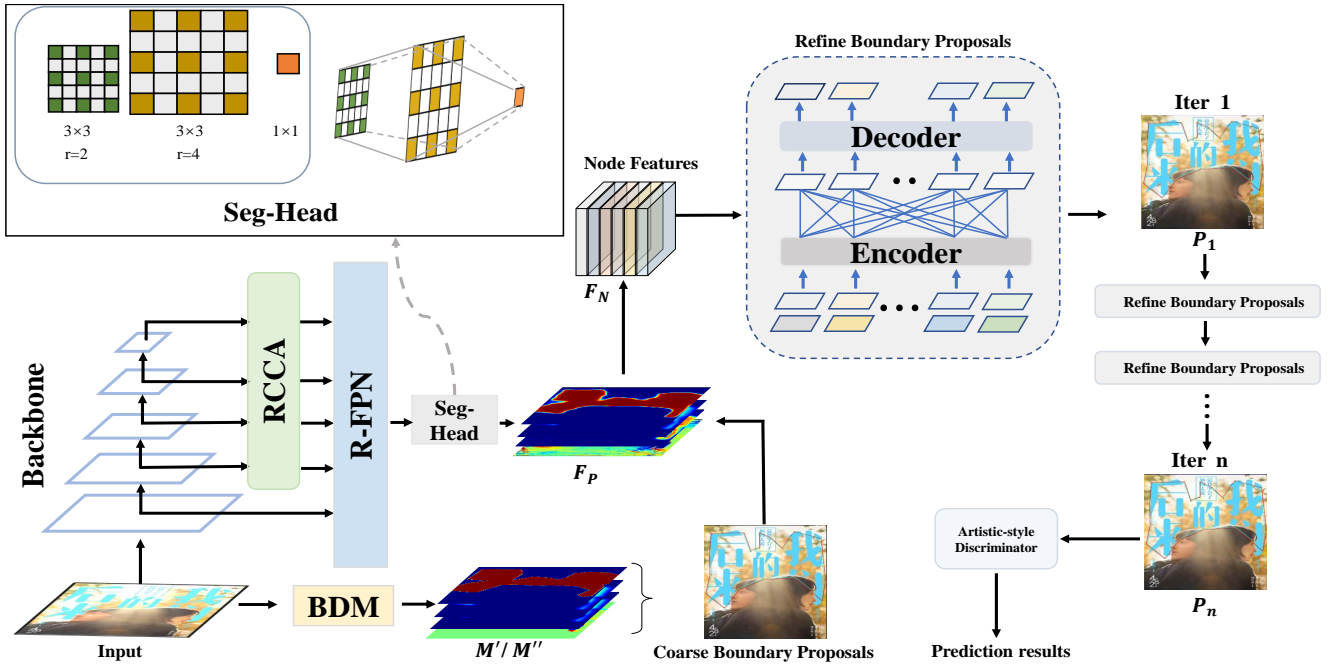


Fig. 3: Framework of our method. F_p denotes the feature maps output by Seg-Head, M'/M'' denotes the masks generated by BDM, F_N denotes the node features obtained by sampling on F_p , P_n are the proposals after iteratively. We use the Artistic-style Discriminator to determine whether the predicted detection proposals contain non-artistic-style text. The Prediction results are the final predicted boundary proposals.

ary localization.

3.2. Recycle Criss-Cross Attention Module

Current text detection algorithms often encounter issues with incomplete detection when dealing with artistic-style text. This is because they cannot effectively capture the complex structural information of artistic-style text. Therefore, we propose an RCCA module to enhance the model's perception of complex structures.

The structure is shown in Fig. 4(a). First, we pass the feature map output from the backbone through a convolutional layer with 3×3 convolutional kernels for feature dimensionality reduction, which can effectively reduce redundant features and model computation. Next, we feed the feature maps into the Criss-Cross Attention Module [Huang et al. \(2019\)](#), which aggregates contextual information in the horizontal and vertical directions for each pixel, enhancing the model's perceptual capabilities. Then, through a loop operation, each pixel can eventually capture the global dependencies of all pixels. Specifically, after a layer of 3×3 convolution, our feature maps are dimensionalized down to $1/4$ of the original ones, and then these feature maps are fed into the Criss-Cross Attention Module. The structure of the Criss-Cross Attention Module is shown at the top of Fig. 4(b). In the Criss-Cross Attention Module, the feature map $I \in R^{C \times W \times H}$ is firstly convolved by two 1×1 convolutions to generate Q and K , where $Q, K \in R^{C' \times W \times H}$, C' is the number of channel. After obtaining feature maps Q and K , we further generate attention maps $A \in R^{(H+W-1) \times W \times H}$ via Affinity operation. The Affinity operation is defined as

follows:

$$d_{i,u} = Q_u Y_{i,u} \quad (1)$$

where $d_{i,u}$ is the degree of correlation between feature Q_u and $Y_{i,u}$. $Q_u \in R^{C'}$ is the vector of Q at u , Y_u are the feature vectors of K which are in the same row or column with position u , $Y_{i,u}$ is the vector of Y_u at i . After getting all the $d_{i,u}$, a softmax operation is performed to the attention map A . Another convolutional layer with 1×1 convolutional kernels is applied on I to generate $V \in R^{C \times W \times H}$ for feature adaption. Similarly, at the same position u of V , we get the feature vectors $V_u \in R^C$, and finally, we get a set $\Psi_u \in R^{(H+W-1) \times C}$, which is the set extracted from the same rows and columns as u in V . Further, we get the semantic information of the duplicate rows and columns at position u through the aggregation operation, which is defined as follows:

$$I'_u = \sum_{i \in |\Psi_u|} A_{i,u} \Psi_{i,u} + I_u \quad (2)$$

where I'_u is a feature vector in output maps $I' \in R^{C \times W \times H}$. $A_{i,u}$ is a scalar value at channel i and position u in A . The contextual information is added to local feature I to enhance the local features and augment the pixel-wise representation.

As shown at the bottom of Fig. 4(b), since the remote context information is only captured horizontally and vertically in the Criss-Cross Attention Module, and the pixel-to-pixel connections around the pixel are still sparse, we designed a recycle module where the output feature map I'

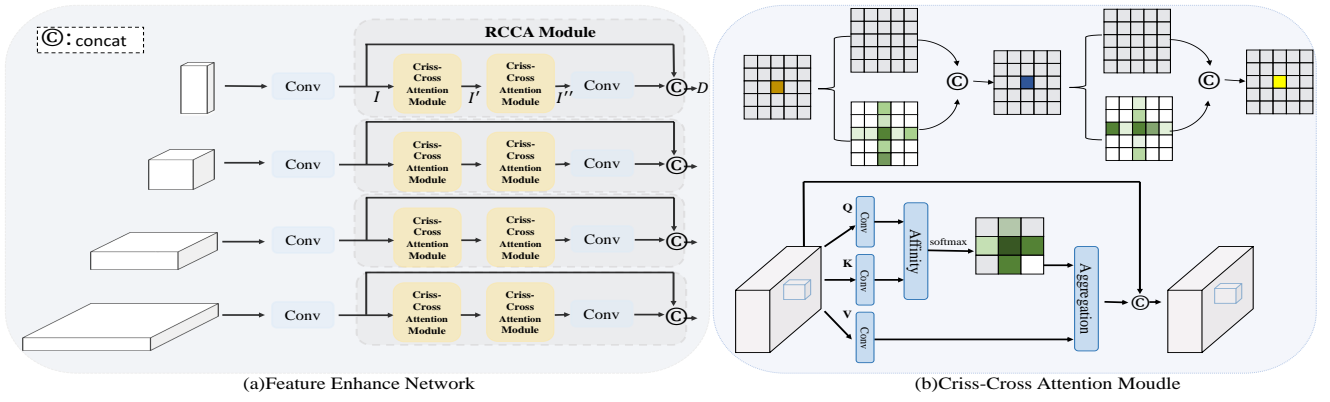


Fig. 4: (a) The overall framework of the feature enhancement network includes four layers of the RCCA module. I , I' , I'' , and D denote the feature maps after various operations. (b) The Criss-Cross Attention Module framework.

after passing through the first module is fed to the next Criss-Cross Attention Module so that our feature maps I'' contain global context information. The operation is defined as follows:

$$I'' = Att(I') \quad (3)$$

where Att denotes the operation of Criss-Cross Attention, $I'' \in R^{C \times W \times H}$ is the feature map output after two Att operations.

Then I'' again undergo a convolution operation and concat with I to get the final output feature maps $D \in R^{C \times W \times H}$.

3.3. Residual Feature Pyramid Network

Since artistic-style text is often mixed with background pixels, the current method usually mistakes the background for the text. Therefore, we propose an FPN-based feature fusion method, R-FPN, to suppress the effect of background noise. Its structure is shown in Fig. 5(a).

Specifically, we change the input at the third layer of the fusion stage and feed the feature map $D \in R^{C \times W \times H}$ output from the feature enhancement network into our proposed Redundant Feature Reduction Module (RFRM) to obtain a clean feature map $D' \in R^{C \times W \times H}$, which is then fed into the feature pyramid network for multiscale fusion. The structure of the RFRM is shown in Fig. 5(b). In RFRM, we first pass a feature map of size 512×512 through a convolutional layer with 3×3 convolutional kernels to better capture the global information. Then, we use the Residual Dense Block (RDB) with a four-layer structure to get the background noise features $O \in R^{C' \times W \times H}$. The RDB consists of residual structure and ReLU operation. It extracts rich local features through densely connected convolutional layers and introduces local residual learning to improve the information flow further. The residual operation is next used to fuse the features obtained from the RDB and ReLU combination, effectively preventing the long-term dependency problem. After the second RDB module and the ReLU operation, we perform an element-wise addition operation on the input raw features and the features from the first and second RDB block

to get the final redundant features $O' \in R^{C' \times W \times H}$. Finally, we subtract the redundant features obtained by RFRM from the original input features to obtain a clean feature map D' , and then input $D' \in R^{C \times W \times H}$ into FPN for feature fusion, which is defined by RFRM as follows:

$$D' = D - CRC(O + O') \quad (4)$$

where $D' \in R^{C \times W \times H}$ is the feature map after denoising, C denotes a function of Conv, R express a function of ReLU.

3.4. Boundary Discrimination Module

Due to irregular shapes and extreme aspect ratios, artistic-style text can suffer from boundary modelling distortion. In this regard, we propose the BDM to solve this problem.

Specifically, we map ground truth onto the input image to obtain a mask map $M \in R^{C \times W \times H}$. Based on M , we compute the Euclidean Distance transform from each pixel point in the image to the nearest nonzero pixel point to obtain a new mask map $M' \in R^{C \times W \times H}$. From the output mask map M' , we can compute the coarse boundary proposals P_0 . However, the initially generated coarse boundary proposal may be erroneous due to some images' highly irregular shapes of artistic-style text regions. This will introduce noise during training and affect the model's performance. Therefore, we design a discriminator to solve this problem. If the number of nonzero pixels in M' is much lower or higher than the number of nonzero pixels in M , it is determined that its modelling fails. If it fails, the Euclidean Distance transform between each pixel point in the ground truth mask and the nearest nonzero pixel point is computed to obtain the mask map M'' , and then the coarse boundary proposals are generated based on M'' .

$$P_0 = \begin{cases} f(M''), & \frac{N_{M'}}{N_M} < th_a \text{ or } \frac{N_{M'}}{N_M} > th_b \\ f(M'), & th_a \leq \frac{N_{M'}}{N_M} \leq th_b \end{cases} \quad (5)$$

where f stands for the function of calculates the coarse boundary proposals based on the mask map; N_M , $N_{M'}$ denotes the number of nonzero pixels in M and M' ; th_a and

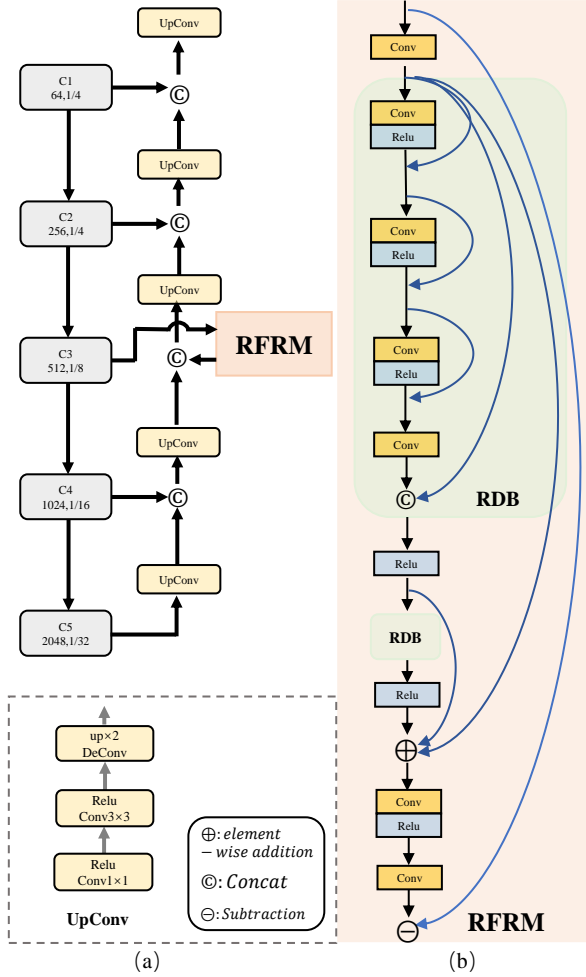


Fig. 5: (a) The overall structure of R-FPN. The 1/64, 1/256, 1/1024, and 1/2048 indicate the scale ratio compared to the input image. (b) Detailed architecture of RFRM in R-FPN.

th_b is the threshold that is experimentally proven to perform best when $th_a = 0.25$ and $th_b = 1.75$;

The Seg-Head consists of a multilayer expansion convolution, including two 3×3 convolution layers with different expansion rates and one 1×1 convolution layer. With the Seg-Head, we obtain the feature maps $F_P \in R^{C \times W \times H}$: the classification map, the distance field map, and the direction field map. We root P_0 to sample the specified location of the feature maps F_P to obtain the node features F_N . Then F_N are fed into the boundary transformer for n iterations of refinement to obtain the final boundary proposals P_n .

4. Experimental results

4.1. Datasets

Movie-Poster: It contains 1500 images, of which 1100 are in the training set and 400 in the testing set. The data consists of movie posters, most of which are artistic-style text of arbitrary shapes. These text regions can appear at any angle, and there may be joins and overlaps between characters. Fig. 6 illustrates some of the analysis of the Movie-Poster

dataset. This shows that our dataset contains text boxes of various shapes and sizes, which is highly challenging. In addition, the Movie-Poster dataset is multilingual, mainly in English and Chinese.

Total-Text (Ch'ng & Chan (2017)): It is a dataset containing the text of various shapes, including horizontal, multi-orientational, and curved. The dataset has a total of 1555 images and 11459 text lines. There are 1255 images in the training set and 300 images in the testing set.

CTW1500 (Yuliang et al. (2017)): It dataset contains 1,500 images, of which 1,000 are for training and 500 are for testing. Each image has at least one curved text, containing many horizontal and multi-orientational text. In addition, the dataset is multilingual, with mainly Chinese and English text.

MSRA-TD500 (Yao et al. (2012)): It contains 500 natural scene images, of which 300 are for training and 200 are for testing. It supports multiple languages, mainly Chinese and English. It contains a variety of text orientations and covers diverse scenes, such as billboards, signs, and walls.

ICDAR-Art (Chng et al. (2019)): It contains a total of 10,166 images, 5603 images in the training set and 4563 images in the testing set. It consists of three parts: Total-Text, SCUT-CTW1500, and Baidu Curved Scene Text (ICDAR2019-LSVT partially curved data), and contains horizontal, multi-orientational, and curved text of various shapes.

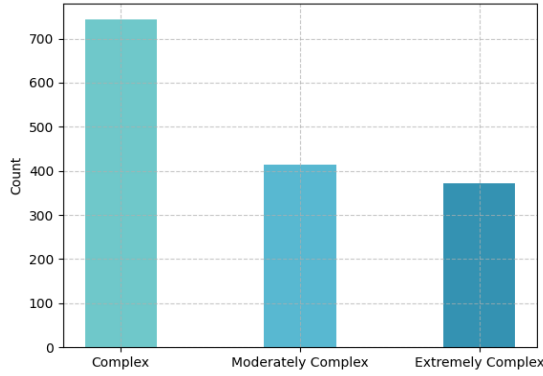
4.2. Implementation Details

The ResNet-50 He et al. (2016) is adopted as our backbone. In our experiments, we randomly crop the text region and resize it to 640×640 for training the model. To demonstrate the model's performance, we do not use pre-training data, and the initial learning rate is set to 0.0001 or 0.001 and decays by 0.9 after every 50 epochs. For all datasets, the training batch size is set to 4, and we use Adam Kingma & Ba (2014) as the optimizer. The data augmentation includes random rotation with an angle, random cropping, and random flipping. The other datasets were trained with 600 epochs except for MSRA-TD500, which was trained with 1200 epochs. In the inference, we keep the aspect ratio of the test images and then resize and fill them to the same size for the test. Training and testing are performed on a single GPU (NVIDIA GeForce RTX3090).

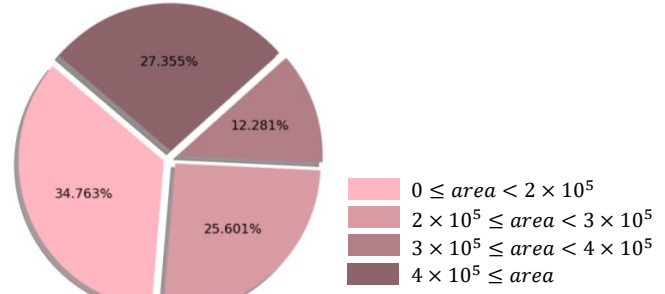
4.3. Ablation study

We perform a series of ablation experiments on the Movie-Poster dataset and Total-Text to demonstrate the effectiveness of our proposed modules. A detailed study of the performance of each module is given in Table 1.

Effectiveness of RCCA: Table 1 demonstrates the effectiveness of the RCCA module. On the Movie-Poster dataset, the use of the RCCA module alone results in a significant increase in F-measure by about 6.55%. This proves that RCCA can effectively utilize the global semantic information to better adapt to complex environments, significantly improving the model's effectiveness in detecting artistic-style text. In



(a) Complexity of the text region



(b) Area distribution of the text region

Fig. 6: Some analytical visualizations of the Movie-Poster dataset. (a) Complexity of the text region in Movie-Poster. We judge the complexity of a text region by **the number of coordinate points**, with coordinate points between 10 and 15 being Complex, coordinate points between 15 and 30 being Moderately Complex, and coordinate points than 30 being Extremely Complex. (b) Area distribution of all text regions in Movie-Poster, where the unit is square pixels.

Table 1

Ablation study of R-FPN, RCCA, BDM and Artistic-style Discriminator. '✓' means use the module, and '×' means don't use it. **Bold** indicates the best result, while underline indicates second.

Methods				Movie-Poster			Total-Text		
R-FPN	RDM	RCCA	Discriminator	Precision	Recall	F-measure	Precision	Recall	F-measure
×	×	×	×	71.92	86.83	78.67	<u>91.1</u>	83.4	87.08
✓	×	×	×	81.48	85.85	83.61	90.4	84.3	87.26
×	✓	×	×	84.67	84.87	84.77	91.12	83.4	87.09
×	×	✓	×	<u>86.84</u>	83.66	85.22	90.5	84.86	87.59
✓	×	✓	×	81.95	89.76	85.68	90.27	<u>85.53</u>	<u>87.84</u>
✓	✓	✓	×	83.76	<u>88.05</u>	<u>85.85</u>	90.17	85.67	87.86
✓	✓	✓	✓	88.89	85.85	87.34	-	-	-

Table 2

Ablation study of the number of Criss-Cross Attention Module cycles. 'P', 'R', and 'F' represent Precision, Recall, and F-measure. **Bold** indicates the best result, while underline indicates second.

Cycles num.	Movie-Poster		
	Precision	Recall	F-measure
0	81.5	85.89	83.64
1	87.28	83.66	85.43
2	88.89	<u>85.85</u>	87.43
3	<u>88.49</u>	84.39	<u>86.39</u>

addition, for the Total-Text dataset, our F-measure is also increased by 0.51%, proving that the module is useful for detecting curved text.

Effectiveness of R-FPN: The effectiveness of the R-FPN can be seen from the results in Table 1. Only using this module on the Movie-Poster and Total-Text datasets improves the F-measure by 4.94% and 0.18%, respectively. If both R-FPN and RCCA are used, the increment of F-measure reaches 0.78% on the totaltext and 7.01% on the Movie-

Poster.

Effectiveness of BDM: BDM is also effective on the Artistic Poster dataset, as shown in Table 1. If only BDM is used, the increment of F-measure reaches 6.1%. This shows that the correctness of boundary modelling is essential for detecting artistic-style text, and BDM effectively guarantees the correctness of boundary modelling. Since the text instances in the Total-Text have a more regular shape, BDM has relatively little impact on this dataset, with a 0.01% increase in F-measure. Our method uses R-FPN, RCCA, and BDM, increasing F-measure by 0.78% on Total-Text and 7.18% on Movie-Poster.

Effectiveness of the number of Criss-Cross Attention Module cycles: As shown in Table 2, the detection effect is optimized when the number of cycles is set to 2. The analysis yields that the Criss-Cross Attention Module collects semantic information in each cycle's horizontal and vertical directions. When the number of cycles is 1, the connection between pixels is still sparse. Each pixel incorporates global semantic information by increasing the number of cycles to 2. However, if the number of cycles is increased to 3 or more, too much noise may be introduced, which leads

Table 3

Detection results on the Movie-Poster dataset. **Bold** indicates the best result, while underline indicates second. [†] denotes the result reproduced using the original paper method.

Methods	Published	Movie-Poster/IOU 0.5			Movie-Poster/IOU 0.75		
		Precision	Recall	F-measure	Precision	Recall	F-measure
PSENet [†] Wang et al. (2019b)	CVPR'19	79.54	84.39	81.89	60.68	64.39	62.48
FAST [†] Chen et al. (2021)	Arxiv'21	74.65	78.29	76.42	46.38	48.53	47.43
PAN [†] Wang et al. (2019c)	ICCV'19	82.23	82.43	82.33	63.26	63.41	63.33
DBNet++ [†] Liao et al. (2022)	TPAMI'22	80.2	77.45	78.8	-	-	-
TextPMS [†] Zhang et al. (2022)	TPAMI'23	75.75	<u>86.09</u>	80.59	62.01	70.48	65.98
TextBPN++ [†] Zhang et al. (2023)	T-MM'23	71.92	86.83	78.67	44.84	54.14	49.06
CBNet [†] Zhao et al. (2024)	IJCV'24	<u>87.89</u>	81.46	<u>84.55</u>	<u>68.68</u>	63.65	<u>66.07</u>
Ours	-	88.89	85.85	87.34	70.7	<u>68.29</u>	69.47

to a decrease in the recall. Therefore, our method sets the number of cycles to 2 to balance the information's comprehensiveness with noise control.

When detecting the Movie-Poster dataset, we designed an Artistic-style Discriminator to determine whether the detected text region is an artistic-style title, effectively removing those redundant text boxes. The F-measure increases by 1.49%, reaching 87.34%, and the results are shown in Table 1.

4.4. Comparisons with previous methods

To comprehensively evaluate the performance of our proposed method, we conducted comparative experiments on five different datasets. These datasets include the Movie-Poster dataset and four recognized benchmark datasets: Total-Text, CTW1500, ICDAR-Art, and MSRA-TD500. For a comprehensive evaluation, we employed both qualitative and quantitative methods.

Movie-Poster: During training, we configured the input image dimensions to 640×640 and employed the Adam Kingma & Ba (2014) optimizer. We commenced the training process without leveraging any pre-trained data, setting the initial learning rate to 0.0001 and applying an exponential decay of 0.9 every 50 epochs. Given that the artistic-style text in the Movie-Poster dataset contains a variety of glyphs of larger sizes, we set intersection and concurrency (IOU) thresholds of 0.5 and 0.75 respectively when testing. The quantitative results are shown in Table 3. Our method achieved promising results in Precision (88.89%), Recall (85.85%) and F-measure (87.34%). The qualitative results are shown in Fig. 7. Compared with the state-of-the-art methods (Liao et al. (2022); Zhang et al. (2022, 2023); Zhao et al. (2024)), we have a clear advantage, and all indicators have achieved the most advanced results. When evaluating the F-measure, our method improves the performance by 8.67% compared to the baseline Zhang et al. (2023) when the intersection and concurrency ratio (IOU) threshold is set to 0.5 and improves the performance by 2.79% compared to the current state-of-the-art method CBNet Zhao et al. (2024). When the IOU threshold is increased to 0.75, our method

Table 4

Detection results on the Total-Text dataset. 'P', 'R', and 'F' represent Precision, Recall, and F-measure. **Bold** indicates the best result, while underline indicates second. [†] denotes the result reproduced using the original paper method.

Methods	Published	P	R	F
SPCNetXie et al. (2019)	AAAI'19	83	82.8	82.9
LOMOZhang et al. (2019)	CVPR'19	87.6	79.3	83.3
PSENet [†] Wang et al. (2019b)	CVPR'19	88.55	77.81	82.83
PAN [†] Wang et al. (2019c)	ICCV'19	83.54	77.45	80.38
ContourNetWang et al. (2020b)	CVPR'20	86.9	83.9	85.4
DRRGZhang et al. (2020)	CVPR'20	86.5	84.9	85.7
PAN++ [†] Wang et al. (2021)	TPAMI'21	76.21	69.05	72.46
FAST [†] Chen et al. (2021)	Arxiv'21	87.44	79.29	83.17
DBNet++ [†] Liao et al. (2022)	TPAMI'22	87.48	78.94	82.3
TextPMS [†] Zhang et al. (2022)	TPAMI'22	85.66	83.28	84.55
EMAZhao et al. (2022)	TIP'22	83.3	88.9	86
TextBPN++ [†] Zhang et al. (2023)	T-MM'23	91.1	83.4	<u>87.08</u>
CBNet [†] Zhao et al. (2024)	IJCV'24	87.53	80.21	83.71
Ours	-	<u>90.17</u>	<u>85.67</u>	87.86

Table 5

Detection results on the CTW1500 dataset. 'P', 'R', and 'F' represent Precision, Recall, and F-measure. **Bold** indicates the best result, while underline indicates second. [†] denotes the result reproduced using the original paper method.

Methods	Published	P	R	F
PSENet [†] Wang et al. (2019b)	CVPR'19	82.06	77.97	79.96
LOMOZhang et al. (2019)	CVPR'19	89.2	69.6	78.4
PAN [†] Wang et al. (2019c)	ICCV'19	78.18	78.85	78.51
TextRayWang et al. (2020a)	MM'20	77.9	83.5	80.6
PAN++ [†] Wang et al. (2021)	TPAMI'21	74.18	74.8	74.49
TextPMS [†] Zhang et al. (2022)	TPAMI'22	84.02	78.32	81.07
FAST [†] Chen et al. (2021)	Arxiv'21	83.51	76.27	79.73
TextBPN++ [†] Zhang et al. (2023)	T-MM'23	<u>87</u>	<u>79.63</u>	83.15
Ours	-	81.59	83.8	<u>83.12</u>

further improves the performance by 20.41% compared to the baseline Zhang et al. (2023), and 3.4% compared to the CBNet Zhao et al. (2024). This shows the effectiveness of our method for artistic-style text detection.



(a)GT

(b)CBNet

(c)DBNet++

(d)PSENet

(e)Ours

Fig. 7: Qualitative comparisons with CBNetZhao et al. (2024), DBNet++Liao et al. (2022), and PSENetWang et al. (2019b).

Total-Text: During training, we initialize the learning rate to 0.001, and other parameters are the same as training the Movie-Poster dataset. The quantitative results are shown in Table 4. Regarding the F-measure, our proposed method demonstrates a 0.78% enhancement in performance over the baseline Zhang et al. (2023) and a notable 4.15% improvement when compared with CBNet Zhao et al. (2024). The experimental outcomes provide compelling evidence of the efficacy of our methodology in detecting and processing curved text.

CTW1500: Similarly, during training, we configured the input image dimensions to 640×640 and employed the

Adam Kingma & Ba (2014) optimizer. We commenced the training process without leveraging any pre-trained data, setting the initial learning rate to 0.0001 and applying an exponential decay of 0.9 every 50 epochs. As shown in Table 5, the quantitative analysis results on the CTW1500 dataset indicate that our proposed method exhibits similar performance to TextBPN++ Zhang et al. (2023). Further comparing with method TextPMS Zhang et al. (2022), our F-measure achieves a 2.05% improvement.

MSRA-TD500: Considering the limited amount of data in the dataset MSRA-TD500, we performed 1200 epochs during the training process and set the initial learning rate



Fig. 8: Some visual results of failure cases. In (a) and (b), these text regions are very large in shape and more similar in color to the background. In (c), our method does not exclude the redundant text boxes in this poster. In (d), the detection fails because a character image splits the overall text region.

Table 6

Detection results on the MSRA-TD500 dataset. 'P', 'R', and 'F' represent Precision, Recall, and F-measure. **Bold** indicates the best result, while underline indicates second. [†] denotes the result reproduced using the original paper method.

Methods	Published	P	R	F
PAN [†] Wang et al. (2019c)	ICCV'19	58.7	68.03	63.02
TextPMS [†] Zhang et al. (2022)	TPAMI'22	69.79	74.23	71.94
FAST [†] Chen et al. (2021)	Chen et al.	75.42	68.59	72.04
TextBPN++ [†] Zhang et al. (2023)	T-MM'23	72.17	74.4	73.27
CBNet [†] Zhao et al. (2024)	IJCV'24	75.13	76.79	75.95
Ours	-	<u>75.38</u>	<u>76</u>	<u>75.8</u>

Table 7

Detection results on the ICDAR-Art dataset. 'P', 'R', and 'F' represent Precision, Recall, and F-measure. **Bold** indicates the best result, while underline indicates second. [†] denotes the result reproduced using the original paper method.

Methods	Published	P	R	F
PSENet [†] Wang et al. (2019b)	CVPR'19	81.1	57.5	67.3
ContourNetWang et al. (2020b)	CVPR'20	62.1	73.2	67.2
DBNetLiao et al. (2020)	AAAI'20	56	69.9	62.2
TextRayWang et al. (2020a)	MM'20	58.6	75.97	66.17
PCRDai et al. (2021)	CVPR'21	65	83.6	73.1
EMAZhao et al. (2022)	TIP'22	68.7	<u>80.8</u>	<u>74.3</u>
Wang et al. [†] Wang et al. (2022)	TIP'23	60.6	78.35	68.4
Ours	-	<u>70.5</u>	80.37	75.11

to 0.0001. In the testing stage, we adjusted the size of the input image to limit it to 640×960 pixels. As illustrated in the Table 6, our proposed method achieves a 2.6% improvement in the F-measure compared to the TextBPN++Zhang et al. (2023). In addition, we are also competitive with the state-of-the-art method CBNetZhao et al. (2024).

ICDAR-Art: To substantiate the generalizability of our proposed method, we conducted experiments on the ICDAR-Art dataset, which includes a large number of real-world instances of curved text. As shown in Table 7, our method outperforms current state-of-the-art models. Specifically, the F-measure of our method is 0.81% higher than EMA Zhao et al. (2022) and 6.71% higher than the newer

method Wang et al. (2022).

4.5. Challenges

While our methodology exhibits robust performance in detecting artistic-style text, it encounters challenges when dealing with more extreme textual variations. As shown in Fig. 8(a) and (b), our method still needs to be improved when dealing with text with extreme shapes and highly blended colors with the background. As shown in Fig. 8(c), there are few cases of error detection, categorizing non-artistic-style text errors as artistic style titles. Meanwhile, in Fig. 8(d), the severe occlusion of the character image resulted in incomplete text detection. We invite you to come up with innovative solutions to meet these challenges.

5. Discussion

Current text detection methods primarily focus on general scenarios, with few addressing the detection of artistic-style text. However, the importance of detecting artistic-style text in our lives has become increasingly apparent. We collected 1,500 movie posters featuring various artistic-style titles to address the current market's lack of artistic-style text data, and we conducted a comprehensive data analysis of this dataset. We evaluated the complexity of each text region based on the number of coordinate points. As shown in Fig. 6(a), the complexity is roughly divided into three categories: Complex, Moderately Complex, and Extremely Complex. The complexity distribution in this dataset is relatively balanced. Fig. 6(b) shows the area distribution of all annotated text boxes, revealing a wide range of text region sizes in our dataset, including various irregularly shaped artistic-style texts. This further proves the representativeness of our dataset. While our method represents a significant step forward in detecting artistic-style text, Sec. 4.5 discusses areas for improvement. We hope to achieve more efficient and faster detection of artistic-style text in the future.

6. Conclusion

This paper proposes a novel method that effectively enhances the model's perceptual capabilities to accurately de-

tect artistic-style text with complex structures, filling the gap in current text detection algorithms for detecting artistic-style text. In our method, the RCCA module, which consists of the **Criss-Cross Attention Module**, effectively utilizes global contextual information to enable the model to capture detailed features that are highly similar to the background. R-FPN based on FPN and residual dense block suppresses the effect of background noise and avoids treating non-text pixels as text pixels. BDM ensures the correctness of boundary modelling. We propose the Movie-Poster dataset to fill the market's gap in artistic-style text data. Extensive experiments demonstrate that our proposed method performs superiorly on the Movie-Poster dataset and produces excellent results on multiple benchmark datasets.

Acknowledgements

This work was supported in part by the Chongqing University of Technology high-quality development Action Plan for of graduate education gzlcx20243151, the Basic and Applied Basic Research Foundation of Guangdong Province under Project 2021A1515110298, and in part by the Science and Technology Program of Nansha under Project 2021ZD003

References

- Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39, 2481–2495.
- Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9365–9374).
- Chen, Z., Wang, J., Wang, W., Chen, G., Xie, E., Luo, P., & Lu, T. (2021). Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. *arXiv preprint arXiv:2111.02394*.
- Ch'ng, C. K., & Chan, C. S. (2017). Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)* (pp. 935–942). IEEE volume 1.
- Chng, C. K., Liu, Y., Sun, Y., Ng, C. C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E. et al. (2019). Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1571–1576). IEEE.
- Dai, P., Zhang, S., Zhang, H., & Cao, X. (2021). Progressive contour regression for arbitrary-shape scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7393–7402).
- Deng, D., Liu, H., Li, X., & Cai, D. (2018). Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*. volume 32.
- Dong, Y., Jiang, Z., Tao, F., & Fu, Z. (2023). Multiple spatial residual network for object detection. *Complex & Intelligent Systems*, 9, 1347–1362.
- Fan, D.-P., Cheng, M.-M., Liu, J.-J., Gao, S.-H., Hou, Q., & Borji, A. (2018). Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 186–202).
- Fan, D.-P., Wang, W., Cheng, M.-M., & Shen, J. (2019). Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8554–8564).
- Farshi, T. R., Drake, J. H., & Özcan, E. (2020). A multimodal particle swarm optimization-based approach for image segmentation. *Expert Systems with Applications*, 149, 113233.
- Feng, W., He, W., Yin, F., Zhang, X.-Y., & Liu, C.-L. (2019). Textdragon: An end-to-end framework for arbitrary shaped text spotting. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9076–9085).
- Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).
- Han, Q., Yin, Q., Zheng, X., & Chen, Z. (2021). Remote sensing image building detection method based on mask r-cnn. *Complex & Intelligent Systems*, (pp. 1–9).
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017a). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- He, M., Liao, M., Yang, Z., Zhong, H., Tang, J., Cheng, W., Yao, C., Wang, Y., & Bai, X. (2021). Most: A multi-oriented scene text detector with localization refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8813–8822).
- He, W., Zhang, X.-Y., Yin, F., & Liu, C.-L. (2017b). Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 745–753).
- Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 603–612).
- Jazayeri, S., Saghaei, A., Esmaeili, S., & Tsokos, C. P. (2019). Automatic object detection using dynamic time warping on ground penetrating radar signals. *Expert Systems with Applications*, 122, 102–107.
- Kessentini, Y., Besbes, M. D., Ammar, S., & Chabbouh, A. (2019). A two-stage deep neural network for multi-norm license plate detection and recognition. *Expert systems with applications*, 136, 159–170.
- Kim, J. H., Batchuluun, G., & Park, K. R. (2018). Pedestrian detection based on faster r-cnn in nighttime by fusing deep convolutional features of successive images. *Expert Systems with Applications*, 114, 15–33.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 84–90.
- Li, T., Bai, J., & Wang, Q. (2024a). Enhancing medical text detection with vision-language pre-training and efficient segmentation. *Complex & Intelligent Systems*, (pp. 1–13).
- Li, Z., Wei, X., Hassaballah, M., Li, Y., & Jiang, X. (2024b). A deep learning model for steel surface defect detection. *Complex & Intelligent Systems*, 10, 885–897.
- Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2017). Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence*. volume 31.
- Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020). Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11474–11481). volume 34.
- Liao, M., Zou, Z., Wan, Z., Yao, C., & Bai, X. (2022). Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45, 919–931.
- Liu, X., Meng, G., & Pan, C. (2019). Scene text detection and recognition with advances in deep learning: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 22, 143–162.
- Liu, Y., & Jin, L. (2017). Deep matching prior network: Toward tighter multi-oriented text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1962–1969).
- Long, S., He, X., & Yao, C. (2021). Scene text detection and recognition:

- The deep learning era. *International Journal of Computer Vision*, 129, 161–184.
- Long, S., Ruan, J., Zhang, W., He, X., Wu, W., & Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 20–36).
- Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., & Xue, X. (2018). Arbitrary-oriented scene text detection via rotation proposals. *IEEE transactions on multimedia*, 20, 3111–3122.
- Schu, G., & Scharcanski, J. (2018). A new method for building adaptive bayesian trees and its application in color image segmentation. *Expert Systems with Applications*, 98, 57–71.
- Shi, B., Bai, X., & Belongie, S. (2017). Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2550–2558).
- Tang, J., Yang, Z., Wang, Y., Zheng, Q., Xu, Y., & Bai, X. (2019). Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern recognition*, 96, 106954.
- Tian, Z., Huang, W., He, T., He, P., & Qiao, Y. (2016). Detecting text in natural image with connectionist text proposal network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14* (pp. 56–72). Springer.
- Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., & Jia, J. (2019). Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4234–4243).
- Tursun, O., Denman, S., Zeng, R., Sivapalan, S., Sridharan, S., & Fookes, C. (2020). Mtrnet++: One-stage mask-based scene text eraser. *Computer Vision and Image Understanding*, 201, 103066.
- Wang, F., Chen, Y., Wu, F., & Li, X. (2020a). Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 111–119).
- Wang, F., Xu, X., Chen, Y., & Li, X. (2022). Fuzzy semantics for arbitrary-shaped scene text detection. *IEEE Transactions on Image Processing*, 32, 1–12.
- Wang, F., Zhao, L., Li, X., Wang, X., & Tao, D. (2018). Geometry-aware scene text detection with instance transformation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1381–1389).
- Wang, P., Zhang, C., Qi, F., Huang, Z., En, M., Han, J., Liu, J., Ding, E., & Shi, G. (2019a). A single-shot arbitrarily-shaped text detector based on context attended multi-task learning. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1277–1285).
- Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., & Shao, S. (2019b). Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9336–9345).
- Wang, W., Xie, E., Li, X., Liu, X., Liang, D., Yang, Z., Lu, T., & Shen, C. (2021). Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 5349–5367.
- Wang, W., Xie, E., Song, X., Zang, Y., Wang, W., Lu, T., Yu, G., & Shen, C. (2019c). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8440–8449).
- Wang, Y., Xie, H., Zha, Z.-J., Xing, M., Fu, Z., & Zhang, Y. (2020b). Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11753–11762).
- Wu, Y., & Natarajan, P. (2017). Self-organized text detection with minimal post-processing via border learning. In *proceedings of the IEEE international conference on computer vision* (pp. 5000–5009).
- Xie, E., Zang, Y., Shao, S., Yu, G., Yao, C., & Li, G. (2019). Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 9038–9045). volume 33.
- Xu, Y., Wang, Y., Zhou, W., Wang, Y., Yang, Z., & Bai, X. (2019). Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 28, 5566–5579.
- Xue, C., Lu, S., & Zhang, W. (2019). Msr: multi-scale shape regression for scene text detection. *arXiv preprint arXiv:1901.02596*, .
- Yang, Y., Feng, C., & Wang, R. (2020). Automatic segmentation model combining u-net and level set method for medical images. *Expert Systems with Applications*, 153, 113419.
- Yao, C., Bai, X., Liu, W., Ma, Y., & Tu, Z. (2012). Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1083–1090). IEEE.
- Yin, X.-C., Pei, W.-Y., Zhang, J., & Hao, H.-W. (2015). Multi-orientation scene text detection with adaptive clustering. *IEEE transactions on pattern analysis and machine intelligence*, 37, 1930–1937.
- Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., & Bai, X. (2023). Turning a clip model into a scene text detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6978–6988).
- Yuliang, L., Lianwen, J., Shuaitao, Z., & Sheng, Z. (2017). Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, .
- Zhang, C., Liang, B., Huang, Z., En, M., Han, J., Ding, E., & Ding, X. (2019). Look more than once: An accurate detector for text of arbitrary shapes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10552–10561).
- Zhang, S.-X., Yang, C., Zhu, X., & Yin, X.-C. (2023). Arbitrary shape text detection via boundary transformer. *IEEE Transactions on Multimedia*, .
- Zhang, S.-X., Zhu, X., Chen, L., Hou, J.-B., & Yin, X.-C. (2022). Arbitrary shape text detection via segmentation with probability maps. *IEEE transactions on pattern analysis and machine intelligence*, 45, 2736–2750.
- Zhang, S.-X., Zhu, X., Hou, J.-B., Liu, C., Yang, C., Wang, H., & Yin, X.-C. (2020). Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9699–9708).
- Zhao, M., Feng, W., Yin, F., Zhang, X.-Y., & Liu, C.-L. (2022). Mixed-supervised scene text detection with expectation-maximization algorithm. *IEEE Transactions on Image Processing*, 31, 5513–5528.
- Zhao, X., Feng, W., Zhang, Z., Lv, J., Zhu, X., Lin, Z., Hu, J., & Shao, J. (2024). Cbnet: A plug-and-play network for segmentation-based scene text detection. *International Journal of Computer Vision*, (pp. 1–20).
- Zhong, D., Lyu, S., Shivakumara, P., Pal, U., & Lu, Y. (2022). Text proposals with location-awareness-attention network for arbitrarily shaped scene text detection and recognition. *Expert Systems with Applications*, 205, 117564.
- Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 5551–5560).
- Zhu, Y., Chen, J., Liang, L., Kuang, Z., Jin, L., & Zhang, W. (2021). Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3123–3131).
- Zhu, Y., & Du, J. (2021). Textmountain: Accurate scene text detection via instance segmentation. *Pattern Recognition*, 110, 107336.