Ship Number Recognition Method Based on An improved CRNN Model

Wenqi Xu[1], Yuesheng Liu[2], Ziyang Zhong[2], Yang Chen[3], Jinfeng Xia[3], Yunjie Chen[1,4,5*]
[1] School of Mathematics and Statistics, Nanjing University of Information Science, Nanjing, 210044, China
[2] ShenZhen Maritime Safety Administration of China, ShenZhen, 518032, China
[3] CSIC Pengli (Nanjing) Atmospheric Ocean Information System Co., Ltd, Nanjing, 211106, China
[4] Center for Applied Mathematics of Jiangsu Province, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5] Jiangsu International Joint Laboratory on System Modeling and Data Analysis, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[E-mail: priestcyj@nuist.edu.cn]
[*] Corresponding author: Yunjie Chen

Received November 22, 2022; revised February 15, 2022; accepted March 4, 2023; published March 31, 2023

Abstract

Text recognition in natural scene images is a challenging problem in computer vision. The accurate identification of ship number characters can effectively improve the level of ship traffic management. However, due to the blurring caused by motion and text occlusion, the accuracy of ship number recognition is difficult to meet the actual requirements. To solve these problems, this paper proposes a dual-branch network based on the CRNN identification network. The network couples image restoration and character recognition. The CycleGAN module is used for blur restoration branch, and the Pix2pix module is used for character occlusion branch. The two are coupled to reduce the impact of image blur and occlusion. Input the recovered image into the text recognition branch to improve the recognition accuracy. After a lot of experiments, the model is robust and easy to train. Experiments on CTW datasets and real ship maps illustrate that our method can get more accurate results.

Keywords: CRNN, CycleGAN, Pix2pix, Text Occlusion, Text Motion

http://doi.org/10.3837/tiis.2023.03.004 **ISSN : 1976-7277**

1. Introduction

Waterway transportation is an indispensable part of the human comprehensive transportation network and has irreplaceable advantages in the process of large-scale cargo transportation. With the increase in the number of ships, the frequent occurrence of waterway traffic accidents, as well as the illegal overloading of ships to transport goods, have brought huge security risks. Therefore, waterway traffic management is extremely urgent. At present, waterway traffic management mainly depends on the AIS system to determine the identity of ships. However, for some illegal ships, the AIS system cannot obtain ship information. Using computer vision technology to recognize traffic images can effectively improve the level of traffic management.

The recognition of text information in natural scene images has already attracted more and more attention from computer vision. Character recognition is an important part of Optical Character Recognition (OCR) [1]. OCR mainly consists of text area detection and text recognition. The main difficulties of text recognition in natural scene images come from the following factors. First of all, the scene text is very different in font or color. Secondly, most scene images will experience intensity inhomogeneity, motion blur, low contrast, low resolution, and occlusion. In addition, wild text may also have erratic shapes, including curve shapes[2]. Deep learning's advancement in recent years has greatly advanced text recognition technology. Convolutional neural network (CNN) was utilized by Jaderberg et al.[3] to categorize English words. Sequence decoding has frequently made use of connectionist temporal classification (CTC)[4] or attention mechanism[5]. To enable vocabulary-free recognition and achieve acceptable performance, scene text recognition was modeled as a sequence learning problem by them. Although the performance of regular text recognition is good, recognizing irregular text is more difficult (random shape or low quality).

Text that is irregular compared to conventional text frequently exhibits irregular features, such as an irregular shape (perspective deformation or curved shape) or inferior quality (such as motion blur, low contrast, intensity inhomogeneity, and occlusion). Some text images with poor quality are shown in **Fig. 1**. The current crop of irregular text recognizers largely concentrates on the identification of random shapes, and gives little thought to the issue of poor quality. This research suggests an irregular text image recognition network that contains three steps of text detection, text recovery, and text recognition to address the issue of low-quality text recognition.

Fig. 1. Low-quality text images. The first to right columns show the images with motion blur, low contrast, intensity inhomogeneity, and occlusion, respectively.

2. Related Work

Text region detection is an important prerequisite for text recognition. High-precision area detection is conducive to improving the ability of subsequent character recognition. At present, text region detection can be divided into two categories: traditional methods and deep learning methods. Matas et al.[6] proposed a maximally stable extreme region (MSER), which is a classical traditional method. Generally speaking, the gray level change in the text area is relatively small, while the gray level contrast between the text and the background is relatively large, which is consistent with the characteristics of the maximum extreme stable area. Therefore, this feature can be used to extract some connected regions that cannot be obtained by color clustering. Based on this assumption, MSER requires that the internal gray level of the extracted area is almost unchanged, and it is difficult to obtain ideal results when there are low contrast, occlusion, text blur , and other factors in the target area. Therefore, in recent years, most text region detection algorithms are based on deep learning methods.

Text region recognition algorithms based on depth learning methods can be divided into two categories: text detection based on segmentation and text detection based on regression[1]. CTPN[7] and EAST[8] are two classic regression-based methods. CTPN combines CNN and LSTM [9] to detect horizontally distributed texts in complex scenes. In this algorithm, a unique anchor is proposed to locate the text, and then LSTM is used to judge the continuous anchors, and finally, the target text area is obtained. The algorithm has high precision in detecting horizontal characters and small area characters. However, the recognition accuracy of the algorithm for oblique or irregular characters is low, and the detection speed is slow because the algorithm completes in two steps [10]. EAST is a pixel-based scene text recognition algorithm. It eliminates many complex post-processing operations, uses FCN to directly segment character regions, and regresses the distance between character pixels and character bounding boxes, so the algorithm is simple and efficient. However, this model has the following shortcomings: 1) It can only deal with text regions with rotation and quadrilateral transformation; 2) When the receptive field size leads to the regression of the distance between the character pixel and the surrounding boundary, it is easily affected by the length of the text area. Segmentation-based text detection method mainly depends on the distribution information of image pixels. PSENet[11] and CRAFT[12] are classical segmentation-based detection methods. PSENet gradually expands the detection area from small core to large and the whole instance map through multiple semantic segmentation, so it is easy to separate text instances that are very close to or even partially crossed. CRAFT uses the idea of small receptive field expansion to predict large text and long text, so it only needs to pay attention to the content of the character level rather than the whole text instance to get better results. However, this algorithm is not good at detecting conglutination characters and requires highly labeled data and complex training[13]. In recent years, the YOLO algorithm [14] has been widely used in text area detection because of its simple algorithm and high accuracy, but its detection accuracy for small target areas is not high[15]. In view of this, the YOLOv5 algorithm uses multi-scale information to improve the extraction accuracy of small target regions, so this paper uses this algorithm to extract text regions.

At present, there are two mainstream algorithms for deep learning character recognition, namely, the algorithm based on Attention and the algorithm based on CTC. The difference between the two methods is mainly in the decoding stage. The former is to access the sequence to the cyclic neural network module for cyclic decoding, while the latter is to access the sequence generated by coding to CTC for decoding. RAREnet is a specially designed deep neural network based on Attention, which consists of a Spatial Transformer Network (STN) [16]and a Sequence Recognition Network (SRN) [17]. This model combines the advantages

of the attention model and the STN model to improve the accuracy of identifying deformed texts. However, the algorithm uses two networks, which leads to high computational complexity.

CRNN[18] is a classic text recognition algorithm based on CTC. It combines convolutional neural network, recurrent neural network, and CTC loss function to improve the accuracy of scene text recognition. The convolution layer, recursive layer, and transcription layer are the three basic components of CRNN. In the convolution layer, convolutional neural network is used to transform the original image into a featured image. In the recursive layer, the character sequence features are extracted using the depth bidirectional LSTM network based on convolution features. The transcription layer converts features into character output. CTC loss is used to solve the one-to-one correspondence problem between the input sequence and the output sequence.

3. METHODS

The model proposed in this paper is mainly composed of three parts: text area detection, text correction , and text recognition. The whole model framework is shown in **Fig. 2**. Text area detection part uses the YOLOv5 method, and the text recognition part uses the CRNN method based on CTC. In the part of text restoration, a two-branch coupling model is used to restore low-quality images to improve the model recognition accuracy.

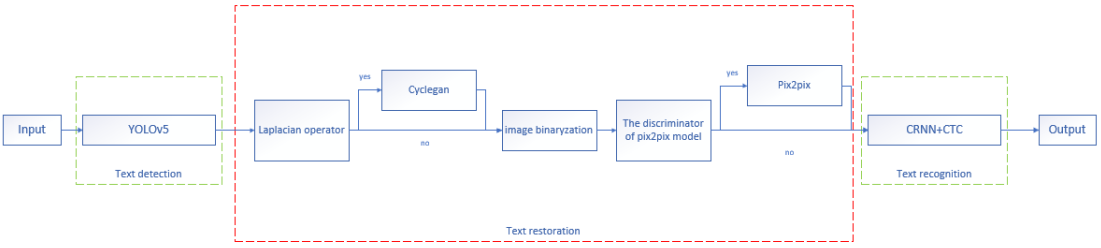

Fig. 2. The framework of text recognition

The current character recognition models mainly focus on the recognition of irregular text, but pay less attention to the recognition of low-quality text. Therefore, the character recovery part is added before the character recognition stage to improve the quality of the text to be detected and the recognition accuracy. For the image with motion blur, this paper selects the CycleGAN model [19] to restore the image. However, when faced with occluded images, the image restoration effect of CycleGAN will be significantly reduced. In this case, we use the Pix2pix model.

3.1 CycleGAN Model

With the development of deep learning, image-to-image style transformation or translation between images has attracted more and more attention. CycleGAN first introduced a neural network into the experiment of painting style migration. The neural network takes two inputs: one picture provides style, the other provides content, and then calculates the loss between the content picture and the style picture.

A two-player minimax game is used to iteratively train the generator $G: X \rightarrow Y$ and discriminator neural networks D_Y that make up a Generative Adversarial Network (GAN) [20]. The definition of adversarial loss $\mathcal{L}(G_{X\rightarrow Y}, D_Y)$ is :

$$\mathcal{L}(G_{X\to Y}, D_Y) = \min_{\Theta_1}\max_{\Theta_2}\left\{\mathbb{E}_y[logD_Y(y)] + \mathbb{E}_x\left[\log\left(1 - D_Y\left(G_{X\to Y}(x)\right)\right)\right]\right\} \quad (1)$$

where Θ_1 and Θ_2 are the parameters of the generator $G_{X\to Y}$ and discriminator D_Y, respectively. $x \in X$ and $y \in Y$ represent the input training data in source and target domain respectively. Meanwhile, $\mathcal{L}(G_{Y\to X}, D_X)$ is similarly characterized.

CycleGAN simultaneously learns the translation between $X \to Y$ and $Y \to X$, which are two distinct image representations. CycleGAN's training data is unpaired, hence. As a result, they implement Cycle Consistency, which may be thought of as fictitious pairings of training data, to guarantee forward-backward consistency. **Fig. 3** contains the CycleGAN framework. The CycleGAN loss function is shown as follows:

$$\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \mathcal{L}(G_{X\to Y}, D_Y) + \mathcal{L}(G_{Y\to X}, D_X) + \lambda\mathcal{L}_c(G_{X\to Y}, G_{Y\to X},) \quad (2)$$

where

$$\mathcal{L}_c(G_{X\to Y}, G_{Y\to X}) = ||G_{Y\to X}\left(G_{X\to Y}(x)\right) - x||_1 + ||G_{X\to Y}\left(G_{Y\to X}(y)\right) - y||_1 \quad (3)$$

is the Cycle Consistency Loss.

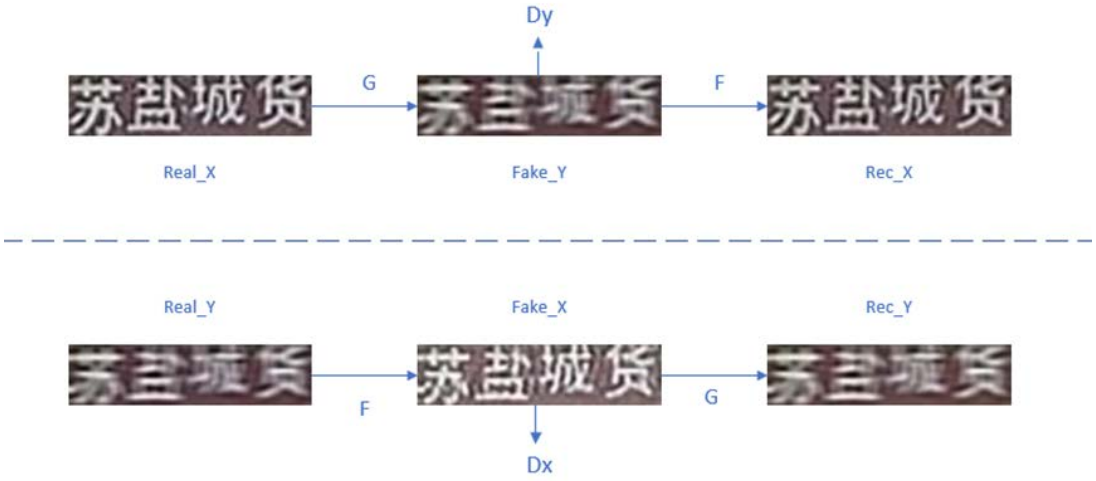

Fig. 3. The model contains two generators, $G: X \to Y$, $F: Y \to X$,which correspond to two different discriminators Dy and Dx. Dy encourages G to make the generated Fake_Y indistinct from the input Real_Y and vice versa. In order to achieve cycle consistency, the final output should be Real_X → Fake_Y → Rec_X ≈ Real_X, Real_Y → Fake_X → Rec_Y ≈ Real_Y.

The complementary roles of edge matching and cycle consistency in CycleGAN's model are one of its strengths. In each domain, marginal matching promotes the creation of realistic samples. Cycle consistency promotes close connections among domains. It can also assist in preventing numerous things from one domain from mapping to a single item from another domain at the same time. Another advantage is that the trained CycleGAN model also works very well for style transfer between mismatched data.

However, one of the basic weaknesses of the CycleGAN model is that it learns deterministic mapping. This causes CycleGAN to learn arbitrary one-to-one mapping, making it difficult to complete geometric changes. The focus of model learning is the transformation of image style, and the image content will not change too much, so the effect of occluded text restoration is not satisfactory. In order to solve this shortcoming, the Pix2pix network is used to recover the occluded words.

3.2 Pix2pix Model

As shown in **Fig. 4**, Pix2pix model [21] is a cGAN [22] based image reduction network. Compared to other GAN models, conditional GANs have the ability to generate a large number of high-quality images for various image transformation tasks. In terms of the loss function, the loss function of the Pix2pix model is also borrowed from the loss function of cGAN:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[logD(x,y)] + \mathbb{E}_{x,z}\left[\log\left(1 - D\left(x, G(x,z)\right)\right)\right] \quad (4)$$

This whole formula is made up of two terms. x represents the source domain image, y represents the real image,z represents the noise input to the generate network, and the generate network based on the source domain image and random noise generates the target domain image $G(x,z)$. $D(x,y)$ is the probability that D determines whether a real picture is real or not. $D(x,G(x,z))$ is the possibility that D determines whether the image generated by G is real or not. The difference between cGAN and GAN is that in addition to generating an image that can fool D, the G role of cGAN also needs to be as close as possible to the image y of the target domain.

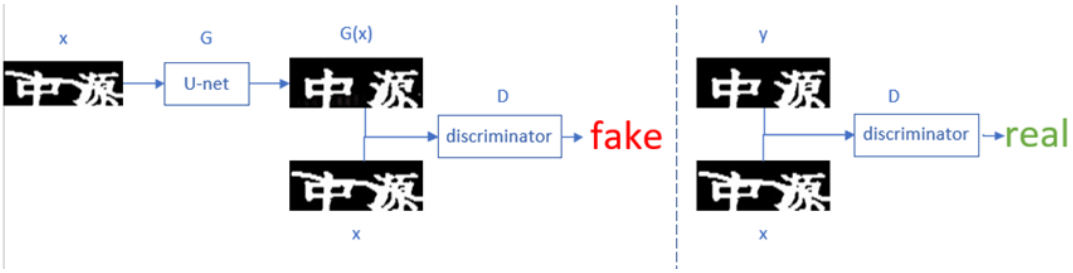

Fig. 4. The x in the figure represents the occluded map of the input, and y represents the occluded map of the input. The discriminator D classifies between fake images (generated images $G(x)$) and real images (images x with occlusion and y without occlusion). The generator G is learning to try and fool the discriminator. Different from the traditional GAN, the cGAN used in this model inputs the occluded image x in both generator and discriminator.

In terms of the generator, the Pix2pix network uses U-Net[23], which has a very obvious effect on promoting details. Mapping a high-resolution input grid to a high-resolution output grid is a defining feature of the image-to-image translation problem. At the same time, we would like to transmit a large amount of shared low-level information existing between inputs and outputs directly over the network.

In terms of the discriminator, PatchGAN is used, which has the property of solving low-frequency components with image reconstruction and solving high-frequency components with GAN. Pix2pix cuts an image into different N x N patches, the discriminator discriminates the authenticity of each patch, and conducts the average of all patches as the final output of the discriminator in an image. After all, the loss function of Pix2pix is defined as:

$$\mathcal{L} = argmin_G max_D L_{cGAN}(G,D) + \lambda \mathbb{E}_{x,y,z}[||y - G(x,z)||_1] \quad (5)$$

where λ is a hyperparameter.

The Pix2pix model intelligently uses GAN to provide a general framework for image translation, and compared with CycleGAN, it is much better for image details and image content filling, it can restore and reconstruct the occluded part of the image according to the mapping provided by the training set. At the same time, the image details are improved through U-net, and the high-frequency part of the image is processed by PatchGAN.

3.3 Proposed Model

Combining the advantages and disadvantages of the above two models, this paper proposes a dual-branch coupling model of CYC-PIX, which uses two networks to restore blurred and occluded images respectively, and finally performs unified recognition.

Firstly, we use Laplacian operator [24] to classify all the detected images as blurry and sharp images. The Laplacian is the detection operator for edge points that is independent of the orientation of one edge. The response of isolated pixels is much stronger than that of edges or lines. After processing, the gray contrast of the image is enhanced, so that the blurred image becomes clearer. The Laplacian operator is defined as:

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{6}$$

where ∇ represents the image gradient computation, which represents the total of the second-order differentiations in x-direction and y-direction. Its discrete form is as follows:

$$\nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4 \tag{7}$$

After many experiments, the images with Laplacian response variance value less than 1000 can be identified as blurred images, and the CycleGAN network is used to restore and deblur them.

Next, after binarization of all deblurred and clear images, the images are input into the discriminator based on cGAN trained by the Pix2pix network. The reason why the discriminator of Pix2pix can be used for classification is that as the generator learns to synthesize better non-occluded images, it will try to deceive the discriminator, while at the same time, the discriminator learns to distinguish between real non-occluded images and synthetic non-occluded images. Our hypothesis is that the features learned by the discriminator which discriminate between real and synthetic unoccluded images can also be used in discriminating between images with or without occlusion[25]. The cGAN network is different from the traditional GAN network in that two images need to be input for each discrimination. We choose a fixed binarized image with occlusion, and the other image is the binarized image output from the previous step[26]. Each image processed by the discriminator will output a one-dimensional vector, and the classification result of the discriminator can be obtained by processing it with the MSELoss [27] below:

$$MSELoss = \frac{1}{N}\sum(x_i - y_i)^2 \tag{8}$$

where N stands for the number of elements in each vector, x_i represents the input image to be discriminated, and y_i represents the fixed occluded image.

After discrimination, the Pix2pix model is used to restore the occluded images for text recognition, and the remaining images are directly input into the recognition network for recognition.

In the part of text recognition, this paper chooses the CRNN model based on CTC. Firstly, the binarized text image after reduction was scaled to the same scale image as the input. Then it is input into the convolutional layer, and the feature sequence of text image is extracted and output by a convolutional neural network. Then, it was input to a Bidirectional long short-term memory network (BiLSTM), and a combination of context information generates the target feature sequence. Finally, the feature sequence of each word is converted into a label sequence by the CTC model of the conversion layer to get the result of text recognition.

4. Experiments

The experimental datasets adopted in this paper include the real ship diagram dataset and the CTW standard dataset[28]. The CTW dataset contains more than 32000 high-resolution images, of which 75% is used as the training dataset, 5% as the validation dataset, 10% as the classification dataset, and 10% as the test dataset. Chinese Text in the Wild (CTW) dataset contains 32,285 images and 1,018,402 Chinese characters. The images sourced from Tencent Street View are captured from dozens of different cities in China, with no preference for any particular purpose. It contains flat text, protruding text, city street view text, partial display text, etc. For each image, all Chinese characters are annotated in the dataset. For each Chinese character, the dataset is annotated with its real character, bounding box, and 6 attributes to indicate whether it is occluded, has a complex background, distorted, artistic, handwritten, etc. The real ship image data set is a non-public data set composed of camera screenshots of the Yangtze River waterway, which contains 2625 pictures, mainly including cargo ships, fishing boats, coast guard ships, etc. The images were taken between 9 and 17 during the day. The text on the ship includes flat text, fuzzy text, and occluded text. The dataset of each image is annotated in Chinese.

Fig. 5. An enumeration of the pictures in the two-training data used in this paper, the left is the real ship map dataset, and the right is the CTW dataset.

In these two datasets, we compare the images recognition accuracy processed by the network in this paper with that of the unprocessed images and perform various occlusion and blur restoration experiments on the two data sets to illustrate the performance of the model. In the real ship dataset, we compare the text recognition accuracy that we have restored with the text accuracy that has not been restored. Meanwhile, we compare the text recognition accuracy of different occlusion degrees with each other. In terms of the text detection model, we use the YOLOv5 model trained by myself. The initial learning rate is set to 0.0003, the number of training epochs is set to 300, the size of the training input picture is 1920*1024, and the size of the batch processing of pictures in the training model is 640*640. The size of batch normalization used in our training is 4. The ADAM optimizer was used in the training. For the

training parameter Settings of CycleGAN and Pix2pix model, the number of epochs is 200, the batch size is set to 64, the optimizer is ADAM optimizer, and the model learning rate is set in the learning rate decay mode. The initial learning rate is set to 0.0002, and after 100 epochs of training, the learning rate starts to gradually decay. The size of the batch processing of pictures in the training model is 256*256. The least squares generative adversarial network(lsGAN) was chosen for the GAN network. First of all, **Table 1** shows the performance gap between the YOLOv5 detection model and traditional CTPN in detection. YOLOv5 is superior to CTPN in both detection accuracy and detection time. In terms of accuracy, it is 5.1% higher on average, and the detection time is 1.2s shorter on average, which is a huge improvement compared with CTPN.

Table 1. The detection accuracy and detection speed of two datasets by different text detection methods.

Model	Shipdata	Detect time	Shipdata	Detect time
YOLOv5	95.21%	0.035s	88.32%	0.04s
CTPN[7]	90.01%	1.2s	83.32%	1.35s

For motion-blurred images, due to the insufficient number of realistic ship images, we applied Gaussian blur processing of different degrees to the training set to test the accuracy and robustness of the model for restoring text with different degrees of ambiguity, which can be seen in **Table 2**.

Table 2. Different datasets, various levels of ambiguity, accuracy of original text recognition and accuracy of restored text recognition.

Degree of motion blur	Original images	Our model	Datasets
4	75.86%	78.91%	Shipdata
8	62.07%	79.31%	Shipdata
10	31.03%	68.96%	Shipdata
8	51.70%	57.14%	CTW
10	49.90%	56.23%	CTW

Fig. 6. The result of CycleGAN. The left column contains blurred images, the middle column contains restored images, and the right column contains real images.

CycleGAN has a good effect on image restoration under different degree of blur. In the real ship map data set, the average accuracy is improved by 19.4%. In the CTW dataset, the average accuracy is improved by 10.89%. It is worth noting that the accuracy of the restored picture has been improved, but the color of the picture has changed. However, it does not affect the subsequent text recognition accuracy, so it can be ignored. **Fig. 6** shows the resorted images and we can find that CycleGAN can obtain ideal results.

For shade really pictures, we also simulate the likely scenario in reality, then obscured text after binarization, the shade or missing rendering generally the same as the background and the target text color is black or white same shade. In this way, the color complexity in the picture is simplified and the restoration efficiency is improved. We simulate the dataset with different thicknesses, different shapes, and different color shades in order to verify the performance of the restored model. The experimental results are shown in **Table 3**.

Fig. 7 shows the resorted images and we can find that Pix2pix can obtain satisfactory results. In the real ship map data set, the recognition accuracy of black occluded text restoration is improved by 20%, and the accuracy of white occluded text restoration is improved by 14.09%. In the CTW dataset, both black and white occlusion was significantly improved, with an average recognition accuracy of 10.94%.

Fig. 7. The result of Pix2pix. On the left column are restored images, in the middle column are blocked images, and on the right column are real images.

Table 3. Different datasets, different occlusion types, original text recognition accuracy and restored text recognition accuracy.

Block type(mm)	Original images	Our model	Datasets
Blackline(8mm)	36.99%	56.99%	Shipdata
Whiteline(8mm)	41.90%	55.99%	Shipdata
Blackline(5mm)	79.50%	87.56%	CTW
Blackline(8mm)	71.99%	88.06%	CTW
Whiteline(8mm)	75.49%	89.55%	CTW
Whiteline(5mm)	81.99%	87.56%	CTW

For the selection of the Laplacian variance value used in motion-blurred image classification, this paper also conducts corresponding experiments on the real ship dataset. In the experiment, 200 blurred images and clear images are selected and the corresponding Laplacian variance values are calculated. According to the obtained results, the dataset is classified based on different experimental parameters (A), and the classification accuracy obtained is shown in **Table 4**.

Table 4. Different experimental parameter A, corresponding classification accuracy in the real ship dataset.

Determine parameter(A)	Classification accuracy
200	50%
400	55%
600	61%
800	70%
1000	75%
1200	76%
1400	76%

It can be seen from the data in above that when the parameter is less than 1000, the classification accuracy has great improvement with the increase of the parameter. Meanwhile, the classification accuracy has a small improvement with the increase of the parameter when the parameter is larger than 1000. Therefore, we choose 1000 as the final parameter that used in our model. In addition, during the experiment, we found that the clarity of some clear images will not be greatly changed after being restored by CycleGAN. Therefore, when selecting the classification parameters, we can consider setting a larger value as far as possible to ensure that the blurred images can be selected for restoration as much as possible.

5. Conclusion

In this paper, we propose a text recognition model based on CRNN for two-branch coupled image restoration, which can restore the text that will make errors in text recognition, and improve the accuracy of text recognition. Our experimental results also show the effectiveness of our proposed model in real-ship datasets as well as standard datasets.

Possible directions for future research include: 1) The model of the text detection part can be optimized to detect irregular characters and automatically correct them. 2) fuse the abnormal state text classification algorithm to simplify the complexity of the model 3) the model architecture of CRNN can be further improved.

Acknowledgment

This work was supported in part by the Six Talent Peaks Project in Jiangsu Province SWYY-034, the Natural Science Foundation of Jiangsu Province of China BK20191394 and the National Nature Science Foundation of China 61672291.

References

[1] Long S, He X, Yao C, “Scene text detection and recognition: The deep learning era,” *International Journal of Computer Vision*, 129(1), 161-184, 2021. Article(CrossRef Link).

[2] Wang C, Liu C L, "Multi-branch guided attention network for irregular text recognition," *Neurocomputing*, 425, 278-289, 2021. Article(CrossRef Link).
[3] Jaderberg M, Simonyan K, Vedaldi A, et al., "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014. Article(CrossRef Link).
[4] Alex Graves, Santiago Fern andez, Faustino Gomez, et al., "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML*, pp. 369–376, 2006. Article(CrossRef Link).
[5] Wang F, Jiang M, Qian C, et al., "Residual attention network for image classification," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 6450-6458, 2017. Article(CrossRef Link).
[6] Donoser M, Bischof H, "Efficient maximally stable extremal region (MSER) tracking," in *Proc. of 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Ieee, vol. 1, 553-560, 2006. Article(CrossRef Link).
[7] Tian Z, Huang W, He T, et al., "Detecting text in natural image with connectionist text proposal network," in *Proc. of European conference on computer vision*, Springer, Cham, 56-72, 2016. Article(CrossRef Link).
[8] Zhou X, Yao C, Wen H, et al., "East: an efficient and accurate scene text detector," in *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, 5551-5560, 2017. Article(CrossRef Link).
[9] Graves A, Fernández S, Schmidhuber J, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. of International conference on artificial neural networks*, Springer, Berlin, Heidelberg, 799-804, 2005. Article(CrossRef Link).
[10] L. Cao, H. Li, R. Xie and J. Zhu, "A Text Detection Algorithm for Image of Student Exercises Based on CTPN and Enhanced YOLOv3," *IEEE Access*, vol. 8, pp. 176924-176934, 2020. Article(CrossRef Link).
[11] Wang W, Xie E, Li X, et al., "Shape robust text detection with progressive scale expansion network," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9336-9345, 2019. Article(CrossRef Link).
[12] Baek Y, Lee B, Han D, et al., "Character region awareness for text detection," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9365-9374, 2019. Article(CrossRef Link).
[13] P. Dai, Y. Li, H. Zhang, J. Li and X. Cao, "Accurate Scene Text Detection Via Scale-Aware Data Augmentation and Shape Similarity Constraint," *IEEE Transactions on Multimedia*, vol. 24, pp. 1883-1895, 2021. Article(CrossRef Link).
[14] Redmon J, Divvala S, Girshick R, et al., "You only look once: Unified, real-time object detection," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 779-788, 2016. Article(CrossRef Link).
[15] Diwan T, Anirudh G, Tembhurne J V, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, 82, 9243-9275, 2023. Article(CrossRef Link).
[16] Jaderberg M, Simonyan K, Zisserman A., "Spatial transformer networks," *Advances in neural information processing systems*, 28, 2015. Article(CrossRef Link).
[17] Ke W, Chen J, Jiao J, et al., "SRN: Side-output residual network for object symmetry detection in the wild," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 1068-1076, 2017. Article(CrossRef Link).
[18] Tian Z, Huang W, He T, et al., "Detecting text in natural image with connectionist text proposal network," in *Proc. of European conference on computer vision*, Springer, Cham, 56-72, 2016. Article(CrossRef Link).
[19] Harms J, Lei Y, Wang T, et al., "Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography," *Medical physics*, 46(9), 3998-4009, 2019. Article(CrossRef Link).
[20] Goodfellow I, Pouget-Abadie J, Mirza M, et al., "Generative adversarial networks," *Communications of the ACM*, 63(11), 139-144, 2020. Article(CrossRef Link).

[21] Isola P, Zhu J Y, Zhou T, et al., "Image-to-image translation with conditional adversarial networks," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 1125-1134, 2017. Article(CrossRef Link).
[22] Mirza M, Osindero S, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. Article(CrossRef Link).
[23] Ronneberger O, Fischer P, Brox T, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical image computing and computer-assisted intervention*, Springer, Cham, 234-241, 2015. Article(CrossRef Link).
[24] Wang X, "Laplacian operator-based edge detectors," *IEEE transactions on pattern analysis and machine intelligence*, 29(5), 886-890, 2007. Article(CrossRef Link).
[25] Salimans T, Goodfellow I, Zaremba W, et al., "Improved techniques for training gans," *Advances in neural information processing systems*, 29, 2016. Article(CrossRef Link).
[26] Engelsma J J, Jain A K, "Generalizing fingerprint spoof detector: Learning a one-class classifier," in *Proc. of International Conference on Biometrics (ICB), IEEE*, 1-8, 2019. Article(CrossRef Link).
[27] Das K, Jiang J, Rao J N K, "Mean squared error of empirical predictor," *The Annals of Statistics*, 32(2), 818-840, 2004. Article(CrossRef Link).
[28] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, Shi-Min Hu, "A Large Chinese Text Dataset in the Wild," *Journal of Computer Science and Technology*, 34(3), 509-521, 2019. Article(CrossRef Link).

Wenqi Xu received a bachelor's degree in information and computing science from Nanjing University of Information &Science, Nanjing, China, in June 2021.He is currently pursuing the M.S. degree with the School of Mathematics and Statistics, Nanjing University of Information &Science, Nanjing, China. His research interests include Digital image processing, text recognition and deep learning.

Yuesheng Liu received the Master degree in electronic engineering from Shanghai Maritime University, Shanghai, China in 1999. He is the Deputy Director of Scientific and Technical Information Division of Shenzhen Maritime Safety Administration. His research interest is mainly focused on pattern recognition, Ship dynamic monitoring and ship identification.

Ziyang Zhong received the Master degree in Maritime Affairs- Maritime Safety and Environmental Management from World Maritime University, Malmo, Sweden in 2016. He is a chief staff member of Scientific and Technical Information Division of Shenzhen Maritime Safety Administration. His research interest is mainly focused on telecommunication and Intelligent ship monitoring.

Yang Chen received the Ph.D. degree in mechanical engineering from Nanjing University of Science & Technology, Nanjing, China, in 2010. His research interest is mainly focused on the overall technology research of shore based water surface surveillance radar, the overall technology research of shipborne navigation radar, and the overall technology research of ship shore cooperation.

Jinfeng Xia recesiced the Master degree in traffic information engineering and control from Dalian Maritime University in 2005. At present, he is a professional in VTS product development of CSIC Prade (Nanjing) Atmosphere and Information System Co., Ltd. His research interests mainly focus on tracking algorithms and image processing.

Yunjie Chen received the Ph.D. degree in Pattern recognition and intelligent system from Nanjing University of Science and Technology, Nanjing, China, in 2008. He is currently a professor with the School of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing. His research interest is mainly focused on pattern recognition, image segmentation, and image processing.