Feature Fusion Pyramid Network for End-to-end Scene Text Detection

YIRUI WU, College of Computer and Information, Hohai University, China and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
LILAI ZHANG, HAO LI, and YUNFEI ZHANG, College of Computer and Information, Hohai University, China
SHAOHUA WAN*, Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, China and Key Laboratory of AI and Information Processing, Hechi University, Guangxi, China

How to properly involve text characteristics like multi-scale, arbitrary direction, length aspect ratio, into detection network design has become a hot topic in computer vision. Feature Pyramid Network (FPN) is a typical method to achieve robust text detection, where its low-level and high-level feature map retains spatial structure and global semantic information, respectively. However, its strict hierarchical structure fails to fuse low-level and high-level information to improve distinguish ability of feature map. To address this problem, we propose a novel feature fusion pyramid network for end-to-end scene text detection by fusing multi-modal information. By diving pyramid structure into high-level and low-level layers, channel and spatial attention modules are adopted to enhance high-level and low-level feature representation by encoding channel and spatial -wise context information, respectively. In order to reduce information loss by layer transmission, a special residual network is designed to achieve short-cut between high-level and low-level features, so as to realize multi-modal feature fusion. Experiments show the precision and recall of the propose method on ICDAR2015, ICDAR2017-MLT and MSRA-TD500 datasets reach 88.7%/82.1%, 77.0%/60.3% and 85.3%/74.8%, respectively.

Additional Key Words and Phrases: Feature Pyramid Networks, Text Detection, Receptive Fields, Attention Module

1 INTRODUCTION

Text detection has received significant attention in applications such as iTown, Rosetta and many other smart city developments [18, 34]. These applications generally require accurate and robust text detectors to perform tasks of natural scene semantics understanding or visual content analysis. Due to the large variations in text rotations or illumination embedded in complex backgrounds with buildings, trees, etc, many techniques have been proposed to improve the accuracy and robustness of text detection in natural scene images.

Owing to the inevitable challenges and complexities, traditional text detection methods [7, 20, 31, 35] tend to apply multiple processing stages to perform text detection task, including steps of text candidates detection, candidates filtering, classifying text or not, grouping into textlines. With the development of deep learning structures, many works build on Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN) to complete separated steps of text detection like text candidates detection or classifying text. Inspired by the

Authors' addresses: Yirui Wu, wuyirui@hhu.edu.cn, College of Computer and Information, Hohai University, Fochengxi Road 8, Nanjing, Jiangsu, China, 210093 and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Qianjin Road 2699, Changchun , Jilin, China, 130012; Lilai Zhang, zhanglilai1999@gmail.com; Hao Li, lihao1998h@163.com; Yunfei Zhang, gogo@hhu.edu.cn, College of Computer and Information, Hohai University, Fochengxi Road 8, Nanjing, Jiangsu, China, 210093; Shaohua Wan*, Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Guanguang Road, Shenzhen, China, shaohua.wan@ieee.org and Key Laboratory of AI and Information Processing, Hechi University, Guangxi, Yizhou , Guangxi, China, 546300.

2375-4699/2023/1-ART $15.00
https://doi.org/10.1145/3582003

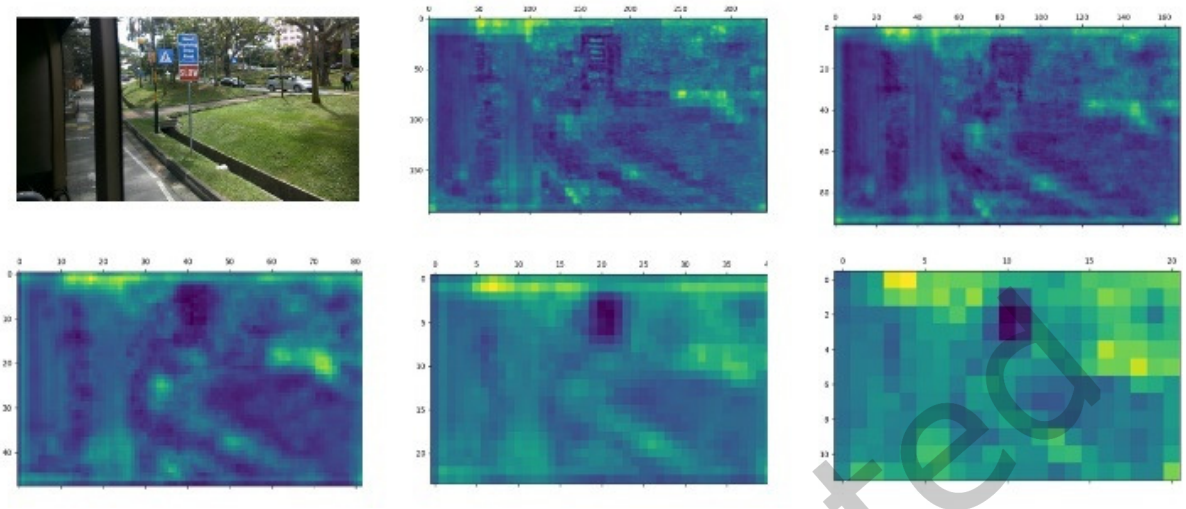

Fig. 1. The heat map of different levels of FFPM. Its brightness represents the feature distribution in the feature map, so the neural network will pay more attention to the areas with brighter colors. The first one is the original image, and the following are the feature maps of the 1st, 2nd, 3rd, 4th and 5th floors respectively. It's noted that the feature of low-levels are obvious to show details, meanwhile the high-level features can help distinguish the text and non text areas.

architecture of Faster-RCNN [21], Ma *et al.* [16] generate rotated anchors as text candidates to detect arbitrary-oriented scene text by their proposed Rotation Region Proposal Networks (RRPN), which has achieved significant improvements on detection accuracy. However, these methods suffer from slow optimization and detection speed, since each individual component must be trained and parameter tuning separately. Moreover, if there exist errors in the middle of pipelines, it will lead to chain reaction for afterward steps, greatly affecting detection accuracy and efficiency. All these drawbacks prevent their further usage to run on embedded systems with low computation resource.

Recently, researchers adopt feature pyramid network to realize multi-scale text detection, which successfully retains the spatial structure information in the lower layer of the network and constructs the global semantic information in the upper layer. However, the strict information stratification of the network has the defect of insufficient information sharing. How to improve the hierarchical characteristics of text features through the transmission and integration of low-level and high-level information has become a concern of researchers.

To solve the problem of insufficient feature information transmission between layers in feature pyramid network, a feature fusion pyramid module, named as FFPM is proposed to improve feature capability for description of text characteristics with different feature levels. First, the upper two layers and the lower two layers of the feature pyramid are fused by upsampling, and divided into high layers and low layers. Then, use feature channel attention to process high-level feature information to obtain contextual information suitable for text detection, and use spatial attention to process low-level features to obtain spatial information suitable for text positioning. Finally, in order to reduce the information loss of low-level space and high-level semantic information in the layer-by-layer transmission of the original pyramid network, this paper designs a residual network to short-circuit high-level and low-level features to achieve text feature fusion and hierarchical feature improvement. The effect of FFPM is visualized in Fig.1.

In summary, the main contributions are as follows.

Table 1. Current Methods on Scene Text Detection

Category	Methods	Year
Regression Based Methods	CTPN[24]	2016
	Textbox[9]	2017
	EAST[38]	2017
	RRPN[17]	2017
	SegLink[23]	2017
	Textbox++[8]	2018
	TextRay[25]	2020
	Xu et al.[32]	2021
	FCE[40]	2021
Segmentation Based Methods	Textsnake[14]	2018
	SPCNet[30]	2019
	TextFuseNet[33]	2020
	Zhang et al.[36]	2021

- A novel feature fusion pyramid network is proposed to address the issue of low-level and high-level feature fusion, which significantly improves performance of end-to-end scene detection task.
- Two attention modules and a residual network are specially designed on the basis of FPN, which achieve feature enhancement by encoding context information and reduce information loss via transmission among layers respectively, thus obtaining fused and distinguish feature map for text detection.
- Quantity of ablation and comparative experiments have proved the effectiveness of the proposed network design specially designed for text detection task.

2 RELATED WORK

The existing methods related to our work can be categorized into the following two types: text detection, and feature pyramid network.

2.1 Text Detection

Owing to the significant discriminative power of deep neural networks, text detection has achieved obvious progresses recently, especially with the rapid development of general object detection [1, 5], semantic segmentation [2, 37], and deep learning works [27–29]. Based on general object detection and semantic segmentation models, several well-designed modifications have been made to improve text detection for higher accuracy. Following such trend, we category recent methods on text detection as regression based and segmentation based text detection, listed in Tab.1 In addition, some of the available quardrilateral or arbitrary shape text detection datasets are listed in the following Tab.2.

2.1.1 Regression-based Methods. Regression based text detection usually use the existing object detection frameworks. Due to the characteristics of text multidirectionality and big aspect ratio, the detection area generated by the existing detection framework can not fit the text very well. Therefore. So it is necessary to further add operations such as rotation angle to adapt to the characteristics of text multidirectionality and length aspect ratio[26].

Early, Tian et al. [24] propose CTPN to detect text area, which uses CNN to get spatial information of pictures, Bidirectional LSTM to learn serialization information, and information of different modal to realize text detection.

Table 2. Current Available Publicly Datasets on Text Detection

Category	Dataset	Number of Samples	Language
Quadrilateral	ICDAR2013	462	English
	ICDAR2015	1500	English
	ICDAR2017-MLT	18000	9 Languages
	MSRA-TD500	500	En & Ch
Arbitrary shape	CTW1500	1500	En & Ch
	Total Text	1555	English

Later, Zhou et al. [38] propose EAST algorithm to detect text. Firstly, the feature information is obtained through the Full Convolution Network, and then the candidate frames are filtered through Non-Maximum Suppression (NMS). However, CTPN and EAST are not specially designed for multi-directional text scenes, resulting in poor robustness for natural scenes.

To solve multi-directional problem, Ma et al. [17] propose RRPN (Rotation Region Proposal Network) and RROI (Rotation Region of Interest) based on RPN and ROI pooling based on Fast R-CNN [22]. Later, Shi et al. [23] propose SegLink, which adapts to multi-direction text detection by using rotation. Afterwards, Xu et al. [32] propose a novel multi-directional detection model based on fast R-CNN, which defines candidate box as four rotation labels and rotation factors with center point, width and height of the rectangle.

Most recently, Liao et al. [9] propose Textbox to focus on text detection with different length aspect ratios, where default boxes of six length aspect ratios of textbox are settled to obey characteristics of text. On the basis of Textbox, Liao et al. [8] further propose Textbox++, which is an end-to-end text detection framework by useing a 3×5 convolution kernel to obtain feature information of text. Wang et al.[25] propose TextRay, where contour based geometric modeling can be conducted from top to bottom through a single shot anchor free architecture to generate text contours. However, this method uses contour point sequence, which has limited ability to express highly curved text.

Considering that Fourier coefficient expression can fit any closed curve theoretically and text outline is more concentrated on the low-frequency component, Zhu et al.[40] propose FCE, which solves the above problems by characterizing the text examples of irregular scenes in the fourier domain, thus owing the characteristics of simple, compact expression ability for complex contours.

2.1.2 Segmentation-based Methods. Most modern segmentation based text detection methods use Fully Convolutional Networks (FCN) to label pixel level objects and background. The segmentation based method is to obtain the segmentation image first, and then obtain the final boundary box according to the segmentation results. By dividing the graph, the network can detect text in any direction or shape without adding additional labels.

Early, Long et al. [14] propose Textsnake algorithm according to the characteristics of the text itself. In order to fit with multi-directional property of texts, Textsnake regards texts as disks, and sets multiple text disks in different directions, which makes it to detect text of different sizes. Later, Xie et al. [30] propose SPCNet (supervised Pyramid Context network), which is enhanced by Text Context Module (TCM), Pyramid Attention Module (PAM) and Pyramid Fusion Module (PFM). Afterwards, Ye et al. [33] propose TextFuseNet, which fuses character feature information, word feature information and global feature information, so as to obtain rich text feature information.

Essentially, there exist two problems in the segmentation based method. One is that adjacent text instances cannot be separated effectively, so complex post-processing is required. Another problem is that they depend on the accuracy of contour detection with quantity of defects and noise. To solve these problems, Zhang et al.[36]

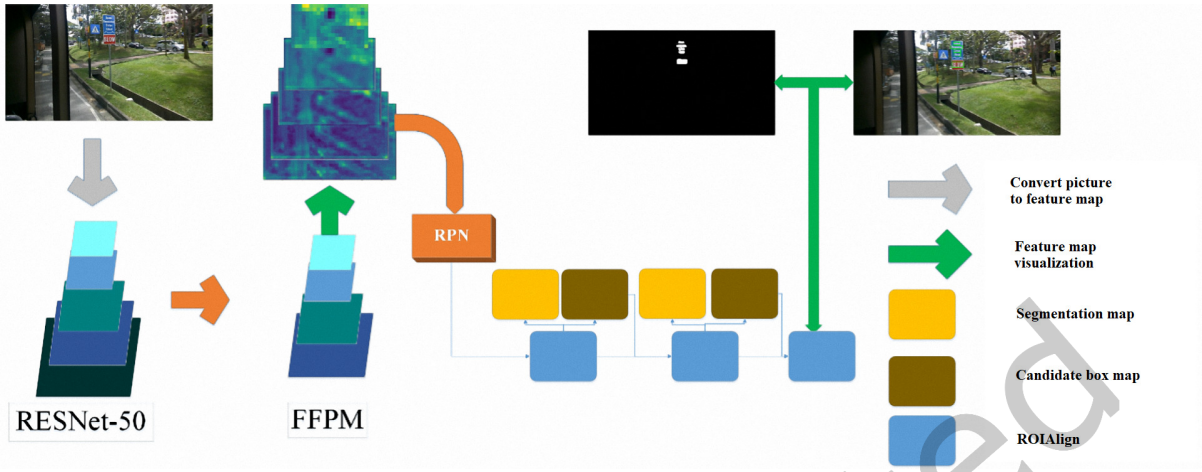

Fig. 2. Overall Framework of the proposed model.

propose to obtain thick border of text instances. Meanwhile, they design the adaptive border adjustment network to iterative refine the adjustment thick border for ground-truth border results.

2.2 Feature Pyramid Network

To better detect objects of different sizes and make use of feature information of different levels, Lin et al. [10] propose FPN, in which the top-level feature map is integrated with the low-level feature map by the up-sampling operation to form low-level feature map, thus obtaining semantic feature information of the high-level. Due to fewer times of convolution and pooling, low-level feature maps have smaller receptive fields and is capable to detect small objects more easily, while high-level feature maps have larger receptive fields through multiple convolution and pooling operations, so they are more sensitive to large objects.

Although FPN propose the idea of hierarchical detection, it's not accurate enough in the distribution of information for feature distribution. Therefore, Liu et al. [11] propose PANet (Path Aggregation Network), where only feature information of several layers receives the semantic information of high-level. Specially, PANet transfers the information of low-level features by adding a bottom-up path to the upper layers, thereby leading the upper layers to be more sensitive about appearance details and location information.

Since there are inevitably some defects in modifying the network of feature pyramid by artificial methods, GhiAsi et al. [4] propose NAS-FPN, which uses neural architecture search to obtain the optimal feature pyramid, and searches the optimal feature pyramid network through combination of pooling layer, activation layer and feature layer of different levels. Recently, most FPN architectures are often paired with RPN to detect objects of different sizes hierarchically, i.e., two-stage detection algorithms, which are often time-consuming. Therefore, Zhu et al. [39] propose FSAF (Feature Selective Anchor-Free) for single-stage detection, which adaptively divides instances of different scales into its appropriate feature layers.

3 THE PROPOSED METHOD

3.1 Overall Framework

The network framework of this paper is shown in Fig. 1, which is composed of backbone network ResNet-50, Feature Pyramid Network FPN, Region Proposal Network RPN and Cascade Modules.Firstly, resnet-50 is used to

obtain basic feature information, and its formula is as follows:

$$F = ResNet_{50}(I), F = \{F_i | i = 1, 2, 3, 4, 5\} \tag{1}$$

Where i represents different levels of the feature map, $i = 1$ represents the lowest level of the feature map, I represents the basic feature map transformed from the picture, $ResNet_{50}(i)$ represents the five levels of ResNet-50.

Then the FPN module is used to fuse the high-level and low-level feature information to generate a feature layer suitable for detecting objects of different sizes. Before FPN, the model did not make full use of the feature information at different levels of the feature layer. For example, the ResNet-50 only used the last layer of the feature layer for detection and recognition. Although the last layer of ResNet-50 feature layer has rich semantic information, it is not sensitive to location and detail information, and the low-level features of ResNet-50 are very precise in the positioning of target objects. The network model can share feature information through FFPM, and can detect objects with different scales hierarchically. The formula is as follows:

$$P = f_{FFPM}\left(F_i\right), i = \{2, 3, 4, 5\} \tag{2}$$

Where f_{FFPM} represents the FFPN module, which also uses four-layer network to refine feature information. Then, the candidate box is generated through RPN. Its formula is as follows:

$$b_0 = f_{RPN}(P) \tag{3}$$

Where B_0 is the initial candidate box, F_{RPN} stands for the generating process.

Because there is a cascade module, this paper does not directly obtain the final result of the candidate frame generated by RPN, but further screening through the cascade module. The cascade module is divided into three stages. The thresholds of the first, second and third stages are set to 0.5, 0.6 and 0.7 respectively. The cascade module removes the false positive results by setting the thresholds of several Intersection of Unions (IoU), which further improves the quality of candidate boxes.

$$\begin{aligned} m_k &= f_{M,k}\left(R_a\left(b_{k-1}, P\right)\right) \quad k = \{1, 2, 3\} \\ b_k &= f_{B,k}\left(R_a\left(b_{k-1}, P\right)\right) \quad k = \{1, 2, 3\} \end{aligned} \tag{4}$$

Where R_a represents the ROIAalign operation, k represents the sequence number of the stage, b represents the candidate box or final boundary box, and m represents the segmentation graph. $F_{B,k}$ is a module for generating classification results and regression results, $F_{M,k}$ is a module for generating segmentation results, which includes fully connected layer, activation function and loss function. The function of multi-stage cascade is to remove false positive results to improve the accuracy of the whole framework, and can also increase or reduce the stages of cascade. The experimental results show that reducing the cascade stage will increase the recall rate and reduce the accuracy rate.

The loss function of the network consists of four parts: RPN, classification, regression and segmentation. The formula is as follows:

$$\begin{aligned} Loss = L_{RPN} + L_{mask,3} + L_{cls,k} + L_{reg,k} \\ k = \{1, 2, 3\} \end{aligned} \tag{5}$$

Where k represents each stage of the cascade network, L_{RPN} is the loss function of RPN, $L_{mask,3}$ is the loss function of the third stage of cascade network segmentation, $L_{cls,k}$ is the loss function of the three-stage classification module, $L_{reg,K}$ is the loss function of the three-stage regression module, and the loss weights of RPN, classification, regression and segmentation are all 1. It is worth noting that the first and second stages of the mask do not affect the final result, so the loss function calculation is not included, and only the third stage of the segmentation is taken as the loss part.

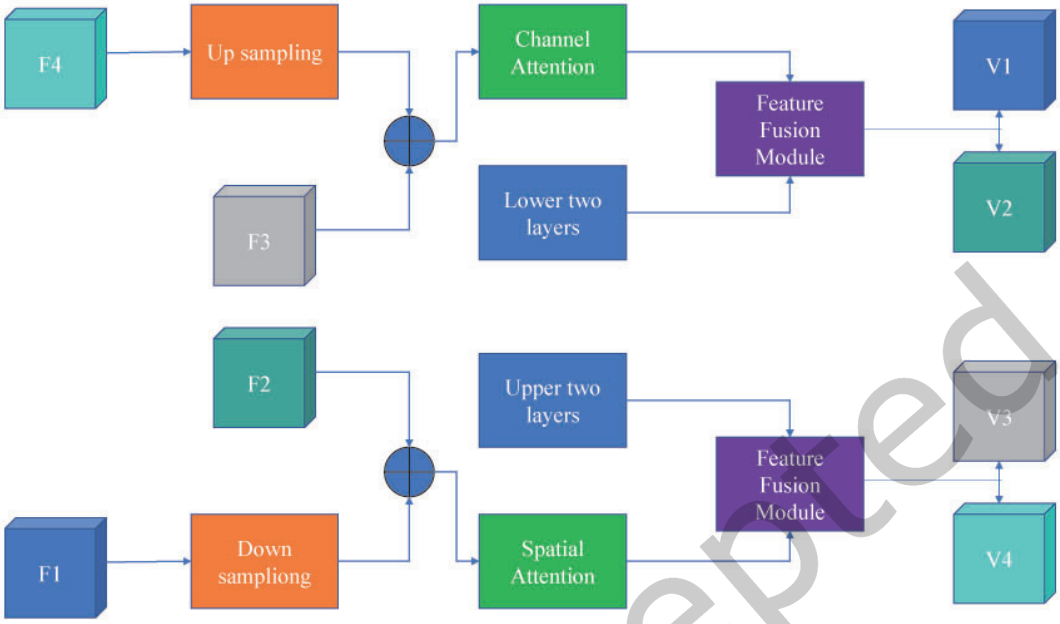

Fig. 3. Feature Fusion Pyramid Module

3.2 Design of Feature Fusion Pyramid Module

Based on FPN, this paper proposes FFPM, which directly transmits the high-level semantic feature information obtained by FPN to the low-level through residual connection, so as to avoid the loss of semantic feature information in the process of layer transmission. At the same time, in order to make the high-level more sensitive to the location information, this paper fuses the filtered low-level spatial feature information with the high-level feature information. After the above processing, it enriches the semantic feature information at different levels and improves the hierarchical characteristics of text features.

As shown in Fig. 3 , F1, F2, F3 and F4 respectively represent the feature information generated by the four levels of FPN. F1 and F2 represent the feature information of the lower two layers of the feature pyramid network, and F3 and F4 represent the feature information of the upper two layers of the feature pyramid network. F1 and F2 contain more spatial information and have smaller receptive fields, and have strong detection ability for small-size text object. F3 and F4 contain rich context information and have larger receptive fields, and have stronger detection ability for large-size text. In the figure, V1, V2, V3 and V4 respectively represent the feature layers of F1, F2, F3 and F4 after feature fusion. The highest layer is still V4 and the lowest layer is still V1.

Semantic information of natural scene text. In order to reduce the loss of semantic information in the transmission process, this model directly fuses the semantic information with the lower two layers of feature information. However, the dimensions of the high and low layers do not match, so the fused feature information needs to be fused with the second layer feature map through the 2-fold up sampling operation, and with the first layer feature through the 4-fold up sampling operation. The above operation is the feature fusion module. The fusion operation formula of the upper two layers of the feature pyramid network is as follows:

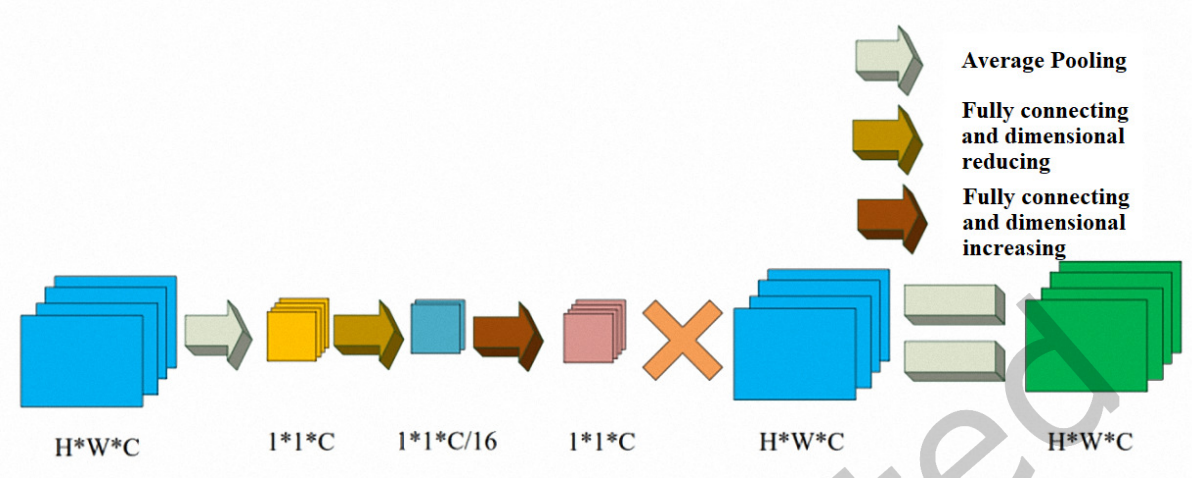

Fig. 4. The structure of Channel Attention

$$\widetilde{U_{34}} = \widetilde{U_4{}^{2up}} + \widetilde{U_3} \tag{6}$$

Where $\widetilde{U_3}$ represents the feature information of the third layer of the feature pyramid, $\widehat{U_4}$ represents the feature information of the fourth layer of the feature pyramid, $widehatU_4{}^{2up}$ represents the feature information obtained by the upper sampling layer, and the feature of the third layer is fused with the feature of the fourth layer after the upper sampling operation $\widetilde{U_{34}}$.

In order to obtain semantic information more suitable for scene text detection, this paper uses channel attention to screen the fused semantic information. The structure of channel attention is shown in Fig. 4 The attention structure makes the input feature map have different weights through the fully connected layer, and then multiplies it with the original feature information to obtain a new feature layer. The above process formula is as follows:

$$Q_{34} = FC_c\left(FC_{c/16}\left(FC_c\left(F_m\left(\widetilde{U_{34}}\right)\right)\right)\right) * \widetilde{U_{34}} \tag{7}$$

Where FC represents the operation of the full connection layer, and C and $C/16$ respectively represent the input and output dimensions of the full connection layer. Using $C/16$ can effectively reduce the amount of parameters of the fully connected layer, and then reduce the amount of parameters of the whole attention structure. $\widetilde{U_{34}}$ represents the input feature maps, F_m represents the mean pooling layer, and the channel attention uses the mean pooling layer to reduce the feature maps $\widetilde{U_{34}}$ and remove redundant information. The maximum pooling operation retains more texture information. Q_{34} represents a new feature map obtained after the feature map $\widetilde{U_{34}}$passes through the channel attention, and its dimension remains unchanged.

The feature map Q_{34} filtered by channel attention is directly fused with low-level features to reduce the loss of semantic feature information layer by layer. The formula is as follows:

$$\begin{aligned} Q_1 &= Q_{34}{}^{4up} + \widetilde{U_1} \\ Q_2 &= Q_{34}{}^{2up} + \widetilde{U_2} \end{aligned} \tag{8}$$

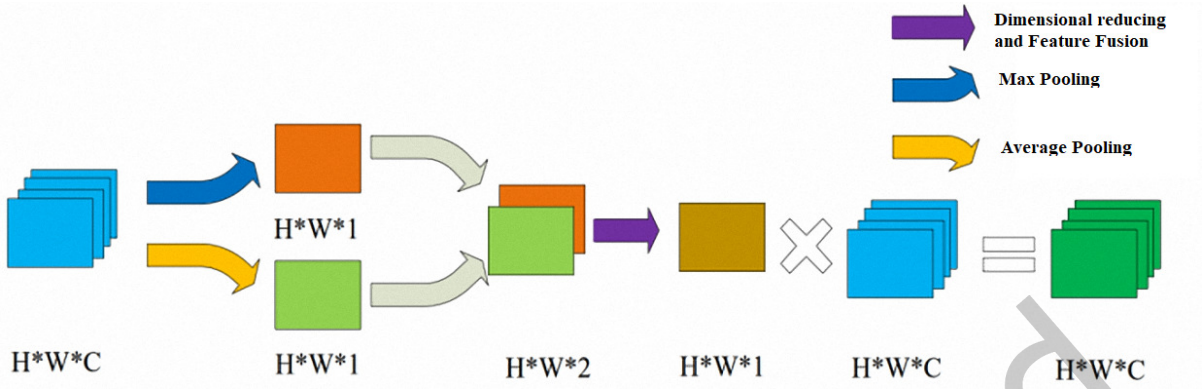

Fig. 5. The structure of Spatial Attention

The feature layer ${Q_{34}}^{4up}$ indicates that the feature dimension of the first layer of the feature pyramid network is consistent with that of the second layer of the feature pyramid network by using the 4-fold up sampling operation. ${Q_{34}}^{2up}$ indicates that the feature dimension of the second layer of the feature pyramid network is consistent with that of the second layer of the feature pyramid network by using the 2-fold up sampling operation, and Q_1 and Q_2 respectively represent the low-level feature information fused with the high-level semantic features.

F1 and F2 are located in the lower layer of the pyramid, with smaller receptive fields and more sensitive to location information. Spatial attention is used to screen the feature information after the fusion of F1 and F2, so as to obtain a feature map that is more sensitive to text location information. Then it is directly fused with the feature information of F3 and F4, and the generated high-level feature map is more sensitive to the location information. In order to solve the problem of dimension mismatch, the fused feature information is fused with the third layer feature information through 2-fold down sampling operation, and with the fourth layer feature information through 4-fold down sampling operation. The formula is as follows:

$$\widetilde{U_{12}} = \widetilde{U_1^{2down}} + \widetilde{U_2} \tag{9}$$

Where U_1 represents the feature information of the first layer of the feature pyramid, U_2 represents the feature information of the second layer of the feature pyramid, $\widetilde{U_1^{2down}}$ represents the feature information obtained by U_1 through the down sampling operation, and $\widetilde{U_{12}}$ represents the feature information after the fusion of the second layer of feature information and the first layer of feature information after the down sampling operation. In order to obtain effective spatial information, the fused position information is filtered by using the spatial attention structure, as shown in Fig. 5.

The fused feature map $\widetilde{U_{12}}$ uses the max pooling and the average pooling to reduce the size of feature maps and remove the redundant information. Then, the feature information is aggregated through the convolution layer and multiplied with the original feature map, and finally the feature map sensitive to spatial information is obtained. The above process formula is as follows:

$$Q_{12} = G\left(F_{\text{mean}}\left(\widetilde{U_{12}}\right) + F_{\max}\left(\widetilde{U_{12}}\right)\right) * \widetilde{U_{12}} \tag{10}$$

Where F_{mean} represents the average pooling, F_{max} represents the max pooling layer, G represents the convolution operation with convolution kernel of 3×3, and Q_{12} represents the feature information of feature information

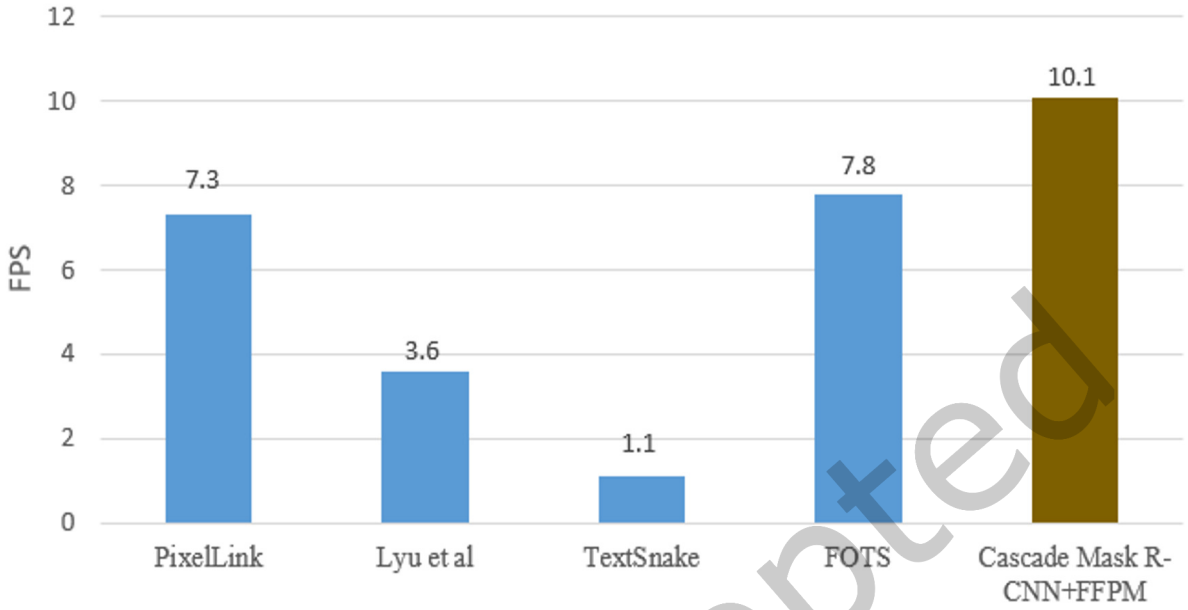

Fig. 6. Comparisons of FPS on ICDAR2015 dataset

Table 3. Ablation experiments of FFPM on ICDAR2015 dataset

Methods	Precision	Recall	F1
Cascade Mask R-CNN	86.70%	79.70%	83.10%
Cascade Mask R-CN+FFPM	88.70%	82.10%	85.30%

$\widetilde{U_{12}}$ after spatial feature information transformation. The size of feature map remains unchanged after spatial attention.

The feature map Q_{12} obtained by spatial attention is directly fused with the low-level feature information to reduce the loss of feature information layer by layer in the transmission process. The formula is as follows:

$$\begin{aligned} Q_4 &= Q_{12}^{4down} + \widetilde{U}_4 \\ Q_3 &= Q_{12}^{2down} + \widetilde{U}_3 \end{aligned} \tag{11}$$

The feature layer Q_{12}^{4down} indicates that a 4-fold down sampling operation is used to keep consistent with the feature dimension of the fourth layer of the feature pyramid network, and Q_{12}^{2down} indicates that a 2-fold down sampling operation is used to keep consistent with the feature dimension of the third layer of the feature pyramid network. Then Q_3 and Q_4 indicate the high-level feature information fused with the low-level detail feature information. Then the obtained feature information is transmitted to RPN, and candidate boxes are generated. Finally, the final boundary box and segmentation map are generated by cascade module.

Table 4. Ablation experiments of FFPM on ICDAR2017-MLT dataset

Methods	Precision	Recall	F1
Cascade Mask R-CNN	76.80%	59.60%	67.10%
Cascade Mask R-CNN +FFPM	77.00%	60.30%	67.70%

Table 5. Ablation experiments of FFPM on MSRA-TD500 dataset

Methods	Precision	Recall	F1
Cascade Mask R-CNN	84.20%	74.60%	79.10%
Cascade Mask R-CNN +FFPM	85.30%	74.80%	79.70%

Table 6. Comparison experiments of Cascade Mask R-CNN + FFPM and other models on ICDAR2015 dataset

Methods	Precision	Recall	F1
He et al.[6]	85.00%	59.60%	82.00%
TextBox++[8]	87.80%	61.20%	82.90%
EAST[38]	83.20%	61.10%	80.70%
PixleLink[3]	85.50%	82.00%	83.70%
SegLink[23]	73.10%	76.80%	75.00%
DMPNet[13]	68.20%	73.20%	70.60%
FOTS[12]	91.00%	**85.20%**	**88.00%**
Liu et al.[15]	**94.10%**	70.70%	80.70%
FCE.[40]	85.10%	84.20%	84.6%
Boundrary.[36]	88.10%	82.20%	85.0%
Cascade Mask R-CNN +FFPM	88.70%	82.10%	85.30%

Table 7. Comparison experiments of Cascade Mask R-CNN + FFPM and other models on ICDAR2017-MLT dataset

Methods	Precision	Recall	F1
He et al.[6]	85.00%	59.60%	82.00%
Liu et al.[15]	**94.10%**	70.70%	80.70%
TDN SJTU2017[19]	64.20%	47.10%	54.30%
SARI FDU RRPN[17]	71.20%	55.50%	62.40%
FOTS[12]	91.00%	**85.20%**	**88.00%**
Cascade Mask R-CNN +FFPM	88.70%	82.10%	85.30%

4 PERFORMANCE EVALUATION EXPERIMENT

4.1 Datasets

This paper mainly detects multi-directional and multi-language text in natural scenes. Since our method is not currently designed for curve text, we only use quadrilateral datasets ICDAR2015, ICDAR2017-MLT and MSRA-TD500 as the model's effect evaluation dataset. The ICDAR2015 dataset contains only English text and includes a total of 1500 pictures, including 1000 training pictures and 500 test pictures. ICDAR2017-MLT is

Fig. 7. The candidate box of the feature information of Cascade Mask R-CNN+FFPM generated in RPN

a multilingual text detection dataset consisting of text images from nine countries, including 7200 training images, 1800 validation images and 9000 testing images. MSRA-TD500 is a small-capacity Chinese and English multi-directional text detection dataset, including a total of 300 training pictures and 200 test pictures.

4.2 Ablation experiments

Ablation experiments are designed to verify the effectiveness of the FFPM module. It can be seen from Table 3, Table 4 and Table 5 that after modifying FPN to FFPM, the detection results of the detection framework on the ICDAR2015 and ICDAR2017-MLT datasets are improved. This is mainly because FFPM increases the multi-scale detection capability of the network. It can better grasp the text information of different sizes and distances. Due to the variability of the scenes of the ICDAR2017-MLT dataset and the diversity of language types, the learning difficulty of the ICDAR2017-MLT dataset is greater than that of the ICDAR2015 dataset.

4.3 Comparison experiments

As shown in tables 4 and 5, the regression based method is worse than the segmentation based method. By adding the fusion feature pyramid module, the model proposed in this paper has been better than most text detection algorithms.

On the ICDAR2015 dataset, the performance results of FOTS are very excellent, with an accuracy rate of 91.0%. The accuracy rate of Cascade Mask R-CNN + FFPM is 88.7%, which is 2.3% higher than that of Cascade Mask RCNN + RACAM. In terms of recall rate, FOTS was 85.2% and RACAM was 82.1%, higher than Cascade Mask R-CNN + RACAM by 3.1%. This is because FOTS uses ICDAR2013, ICDAR2015, ICDAR2017-MLT and synth800k datasets as additional training sets, and uses OHME to train difficult samples. The model proposed in this paper only uses 1000 images in ICDAR2017-MLT data set as additional dataset. Therefore, on ICDAR2015 dataset, the experimental results of Cascade Mask R-CNN + RACAM are not as good as FOTS.

Fig. 8. Detection results of Cascade Mask R-CNN+FFPM on MSRA-TD500 dataset

The method of Liu et al has the best precision among these methods. But there is also a phenomenon that its recall rate is far lower than precision rate, which indicates that it is more conservative in the detection tasks. When it is uncertain whether an object is text, it tends to ignore it rather than blindly predict it. In contrast, although our network is not SOTA in all indicators, it maintains a good balance between precision and recall.

4.4 Complexity and FPS on ICDAR2015 dataset

Inheriting the high efficiency of FPN, our method has 1107.4k parameters. It certain advantages in complexity and computational speed. We provides a speed comparison between FFPM and other classic text detection models, as shown in the Fig. 6. Although FFPM uses a fully connected layer in the attention structure, a smaller dimension c/16 is used in the fully connected layer to reduce the amount of parameters of the entire module, and FFPM uses more pooling and sampling operations and fewer convolution operations, so that only a small number of parameters are added.

4.5 Visualization of experimental results

In this paper, the network model uses a two-stage detection method to detect text. Fig. 7 shows the generation of candidate boxes of different sizes on the feature map of cascade mask r-cnn + ffpm. It can be seen from the figure that a large number of candidate boxes will be generated at the text in the picture through the learning of neural network. The two stages help detection by identifying some candidate boxes in advance, but it also increases the cost of calculation.

Fig. 8 shows the test results on MSRA-TD500 dataset. Different from ICDAR2015 and ICDAR2017-MLT datasets, MSRA-TD500 dataset contains two languages, and there are few text training sets. Due to different test standards, the effect of additional training on MSRA-TD500 using ICDAR2015 and ICDAR2017-MLT datasets is not obvious.

Fig. 9. Detection results of Cascade Mask R-CNN+FFPM on ICDAR2017-MLT dataset

It can be seen from the figure that there is only 300 picturesś text area of MSRA-TD500 training set can be accurately detected after adding FFPM.

4.6 Implementation Details

All these experiments are performed on a single Titan 1080Ti and measured on a 2.1GHz E5-2620 PC with 10G Memory. In the training process, parameters were optimized via the SGD optimizer with an initial learning rate of 0.1 and a mini-batch size of 128.

5 CONCLUSION

This paper aims at the problems of feature information transmission loss and insufficient sharing of different feature information in FPN, and studies and proposes a Feature Fusion Pyramid Network FFPM, which realizes text feature fusion and hierarchical characteristic improvement. In terms of further work, this paper uses FFPM

for hierarchical detection without further adaptive processing of RPN module. Therefore, RPN can generate candidate boxes of different scales according to the hierarchy, but can not generate candidate boxes more in line with arbitrary text shapes, which is the direction of subsequent research.

ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China under Grant No. 2021YFB3900601, National Natural Science Foundation of China under Grant No. 62172438, 61702160, the Fundamental Research Funds for the Central Universities under Grant No. B220202074, the Fundamental Research Funds for the Central Universities, JLU, the Joint Fundation of the Ministry of Education under Grant No.8091B022123, and Key Laboratory of AI and Information Processing (Hechi University), Education Department of Guangxi Zhuang Autonomous Region under Grant 2022GXZDSY014.

REFERENCES

[1] Hamid Bazargani, Olexa Bilaniuk, and Robert Laganiere. 2018. A fast and robust homography scheme for real-time planar target detection. *Journal of Real-Time Image Processing* 15, 4 (2018), 739–758.
[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2018), 834–848.
[3] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. 2018. PixelLink: Detecting Scene Text via Instance Segmentation. In *Proceedings of the AAAI*. 6773–6780.
[4] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. 2019. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In *Proceedings of the IEEE CVPR*. 7036–7045.
[5] Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, Weihua Ou, and Zhang Yi. 2022. Multi-level attention-based sample correlations for knowledge distillation. *IEEE Transactions on Industrial Informatics* (2022).
[6] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. 2018. Multi-Oriented and Multi-Lingual Scene Text Detection With Direct Regression. *IEEE Trans. Image Process.* 27, 11 (2018), 5406–5419.
[7] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision* 116, 1 (2016), 1–20.
[8] Minghui Liao, Baoguang Shi, and Xiang Bai. 2018. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Process.* 27, 8 (2018), 3676–3690.
[9] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. 2017. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. In *Proceedings of the AAAI*. 4161–4167.
[10] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE CVPR*. 936–944.
[11] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE CVPR*. 8759–8768.
[12] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. 2018. FOTS: Fast Oriented Text Spotting With a Unified Network. In *Proceedings of the IEEE CVPR*. 5676–5685.
[13] Yuliang Liu and Lianwen Jin. 2017. Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection. In *Proceedings of the IEEE CVPR*. 3454–3461.
[14] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. In *Proceedings of the ECCV*, Vol. 11206. 19–35.
[15] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. 2018. Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation. In *Proceedings of the IEEE CVPR*. 7553–7563.
[16] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multimedia* 20, 11 (2018), 3111–3122.
[17] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multim.* 20, 11 (2018), 3111–3122.
[18] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar J. Leite, and Jorge Stolfi. 2014. SnooperText: A text detection system for automatic indexing of urban scenes. *Computer Vision and Image Understanding* 122 (2014), 92–104.
[19] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, Wafa Khlif, Muhammad Muzzamil Luqman, Jean-Christophe Burie, Cheng-Lin Liu, and Jean-Marc Ogier. 2017. ICDAR2017

Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT. In *Proceedings of the IAPR ICDAR*. 1454–1459.
[20] Lukas Neumann and Jiri Matas. 2016. Real-Time Lexicon-Free Scene Text Localization and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 9 (2016), 1872–1885.
[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.
[22] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.
[23] Baoguang Shi, Xiang Bai, and Serge J. Belongie. 2017. Detecting Oriented Text in Natural Images by Linking Segments. In *Proceedings of the IEEE CVPR*. 3482–3490.
[24] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting text in natural image with connectionist text proposal network. In *Proceedings of the ECCV*. 56–72.
[25] Fangfang Wang, Yifeng Chen, Fei Wu, and Xi Li. 2020. TextRay: Contour-based Geometric Modeling for Arbitrary-shaped Scene Text Detection. In *Proceedings of ACM International Conference on Multimedia*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). 111–119.
[26] Yirui Wu, Hao Cao, Guoqiang Yang, Tong Lu, and Shaohua Wan. 2022. Digital Twin of Intelligent Small Surface Defect Detection with Cyber-Manufacturing Systems. *ACM Trans. Internet Technol.* (2022). https://doi.org/10.1145/3571734
[27] Yirui Wu, Yuntao Ma, and Shaohua Wan. 2021. Multi-scale relation reasoning for multi-modal Visual Question Answering. *Signal Process. Image Commun.* 96 (2021), 116319.
[28] Yirui Wu, Lilai Zhang, Stefano Berretti, and Shaohua Wan. 2023. Medical Image Encryption by Content-Aware DNA Computing for Secure Healthcare. *IEEE Trans. Ind. Informatics* 19, 2 (2023), 2089–2098.
[29] Yu Xia, Shiru Qu, and Shaohua Wan. 2018. Scene guided colorization using neural networks. *Neural Computing and Applications* (2018), 1–14.
[30] Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li. 2019. Scene Text Detection with Supervised Pyramid Context Network. In *Proceedings of the AAAI*. 9038–9045.
[31] Zhenzhen Xie, Yan Huang, Dongxiao Yu, Reza M Parizi, Yanwei Zheng, and Junjie Pang. 2022. FedEE: A Federated Graph Learning Solution for Extended Enterprise Collaboration. *IEEE Transactions on Industrial Informatics* (2022).
[32] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. 2021. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 4 (2021), 1452–1459.
[33] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. 2020. TextFuseNet: Scene Text Detection with Richer Fused Features. In *Proceedings of the IJCAI*. 516–522.
[34] Qixiang Ye and David S. Doermann. 2015. Text Detection and Recognition in Imagery: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 7 (2015), 1480–1500.
[35] Yuan Yuan, Feng Li, Dongxiao Yu, Jichao Zhao, Jiguo Yu, and Xiuzhen Cheng. 2020. Distributed social learning with imperfect information. *IEEE Transactions on Network Science and Engineering* 8, 2 (2020), 841–852.
[36] Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. 2021. Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection. In *Proceedings of International Conference on Computer Vision*. 1285–1294.
[37] Yue Zhang, Fanghui Zhang, Yi Jin, Yigang Cen, Viacheslav Voronin, and Shaohua Wan. 2022. Local Correlation Ensemble with GCN based on Attention Features for Cross-domain Person Re-ID. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).
[38] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: An Efficient and Accurate Scene Text Detector. In *Proceedings of the IEEE CVPR*. 2642–2651.
[39] Chenchen Zhu, Yihui He, and Marios Savvides. 2019. Feature Selective Anchor-Free Module for Single-Shot Object Detection. In *Proceedings of the IEEE CVPR*. 840–849.
[40] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. 2021. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 3123–3131.