Efficient Medical Images Text Detection with Vision-Language Pre-training Approach

Tianyang Li TIANYANGLI@NEEPU.EDU.CN
Computer Science, Northeast Electric Power University, Jilin, China
Jiangxi New Energy Technology Institute, Jiangxi, China

Jinxu Bai 1092945306@QQ.COM
Computer Science, Northeast Electric Power University, Jilin, China

Qingzhu Wang 150681573@QQ.COM
Computer Science, Northeast Electric Power University, Jilin, China

Hanwen Xu 1429969970@QQ.COM
Computer Science, Northeast Electric Power University, Jilin, China

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Text detection in medical images is a critical task, essential for automating the extraction of valuable information from diverse healthcare documents. Conventional text detection methods, predominantly based on segmentation, encounter substantial challenges when confronted with text-rich images, extreme aspect ratios, and multi-oriented text. In response to these complexities, this paper introduces an innovative text detection system aimed at enhancing its efficacy. Our proposed system comprises two fundamental components: the Efficient Feature Enhancement Module (EFEM) and the Multi-Scale Feature Fusion Module (MSFM), both serving as integral elements of the segmentation head. The EFEM incorporates a spatial attention mechanism to improve segmentation performance by introducing multi-level information. The MSFM merges features from the EFEM at different depths and scales to generate final segmentation features. In conjunction with our segmentation methodology, our post-processing module employs a differentiable binarization technique, facilitating adaptive threshold adjustment to enhance text detection precision. To further bolster accuracy and robustness, we introduce the integration of a vision-language pre-training model. Through extensive pretraining on large-scale visual language understanding tasks, this model amasses a wealth of rich visual and semantic representations. When seamlessly integrated with the segmentation module, the pretraining model effectively leverages its potent representation capabilities. Our proposed model undergoes rigorous evaluation on medical text image datasets, consistently demonstrating exceptional performance. Benchmark experiments reaffirm its efficacy.

Keywords: vision-language pre-training, medical text detection, feature enhancement, differentiable binarization

1. Introduction

Text detection is a longstanding research area dedicated to accurately locating the bounding boxes or polygons of text instances within images. This pursuit is driven by its extensive practical applications across diverse domains, including text recognition, intelligent healthcare, real-time translation, and autonomous driving. Medical images can unveil crucial information associated with diagnostic outcomes. Therefore, the swift and precise extraction of such information from text bears practical significance.

Recent advancements have highlighted the significance of pretrained vision and language knowledge, notably exemplified by the large-scale Contrastive Language-Image Pretraining (CLIP) model . CLIP has demonstrated its efficacy in diverse downstream tasks, including image classification, object detection Fan et al. (2019, 2018a), and semantic segmentation Zhao et al. (2017); Fan et al. (2018b). In contrast to general object detection, scene text within natural images often presents a blend of visual and textual information, making it inherently relevant to the CLIP model. Consequently, recent research endeavors have increasingly focused on harnessing cross-modal information encompassing visual, semantic, and textual knowledge to enhance the performance of text detection models.

In contrast to general object and scene text detection, medical image text detection faces specific challenges that require attention:(1)During the scanning or capture of original medical images, text content may become skewed due to non-flat surfaces. (2)Medical images often feature a substantial volume of text, characterized by a prevalent dense text distribution. (3)Medical text can encompass Chinese characters, alphabets, identifiers, as well as images and tables, which can complicate text boundary detection, potentially resulting in incomplete results.

To address these challenges, the integration of vision and language pre-training models has emerged as a promising approach. Vision-language pre-training models, trained on large-scale visual language understanding tasks, acquire rich visual and semantic representations. Integrating these models with text detection systems enhances their accuracy and robustness.

While conducting this study, it is pertinent to note that our previous work titled 'Multi-level Feature Enhancement Method For Medical Text Detection' has been accepted for publication but is pending formal publication. The work presented in this paper serves as an extension and further exploration of the methodologies introduced in the aforementioned forthcoming paper. In this study, we extend the research presented in our previous work from two distinct perspectives. First, We introduce a vision-language pre-training model that has undergone extensive pre-training on large-scale visual language understanding tasks. Second, we conducted a comprehensive experimental analysis of the proposed text detection network.

In this paper, we propose an efficient and accurate text detection system specifically designed for medical text image datasets. Our system incorporates an optimized segmentation module, a learnable post-processing method, and leverages the power of visual language pre-training models. Through rigorous experimentation and benchmarking, we demonstrate the outstanding performance of our proposed method in handling the complexities of medical text image detection.

In summary, our contributions are three-fold:

- This paper introduces an efficient and accurate text detection system specifically designed for medical text images, addressing the challenges of text-dense images, extreme aspect ratios, and multi-oriented text.

- The proposed system leverages the power of vision-language pre-training models, acquired through large-scale visual language understanding tasks, to enhance the representation capabilities, improving the accuracy and robustness of text detection.
- Using the proposed method, we achieve competitive results in terms of efficiency, accuracy, F-score and robustness on five public text detection datasets.

The rest paper is organized as follows: Sec. 2. reviews the relevant text detection methods. Our proposed approach is described in Sec. 3. The experiments are discussed and analyzed in Sec. 4. The conclusions are summarized in Sec. 5.

2. Related Works

2.1. Text Detection and Spotting

In recent years, significant advancements in the field of scene text detection have been driven predominantly by deep learning approaches. These approaches can be broadly classified into two categories: regression-based methods and segmentation-based methods, depending on the level of granularity in predicting text instances.

Regression-based methods directly estimate the bounding boxes of text instances. For instance, EAST Zhou et al. (2017) and Deep-Reg He et al. (2017b) are anchor-free techniques that employ pixel-level regression to detect multi-oriented text instances. DeRPN Xie et al. (2019b), on the other hand, introduces a dimension-decomposition region proposal network to tackle scale variation challenges in scene text detection. Despite benefiting from straightforward post-processing algorithms, accurately representing irregular shapes, such as curved text, with precise bounding boxes remains a formidable challenge.

On the other hand, segmentation-based methods typically combine pixel-level predictions with post-processing algorithms to derive bounding boxes. PSENet Wang et al. (2019b) introduces progressive scale expansion by segmenting text instances using different scale kernels. Tian et al. Tian et al. (2019) propose pixel embedding to group pixels based on segmentation results. DBNet Liao et al. (2020) focuses on enhancing the segmentation results by incorporating the binarization process during training without sacrificing inference speed.

2.2. Vision-Language Pre-training

Moreover, the integration of vision-language pre-training models Xue et al. (2022) has recently gained attention in the field of text detection. Vision-language pre-training involves training models on large-scale visual language understanding tasks, enabling them to acquire rich visual and semantic representations. These pre-training models have shown promising results in various computer vision tasks, including text detection. By leveraging the power of vision-language pre-training, the text detection systems can benefit from enhanced representation capabilities, leading to improved accuracy and robustness.

3. Methodology

3.1. Overall Network Architecture

Our proposed model's architecture is illustrated in Fig. 1. It incorporates a vision-language pre-training model along with four main components: a feature extraction backbone, a feature enhancement backbone, a Multi-scale feature fusion model, and a post-processing procedure. In the initial stage, the input image undergoes processing by the vision-language pre-training backbone, which has been pre-trained on large-scale visual language understanding tasks. This backbone learns rich visual and semantic representations, providing a robust foundation for subsequent stages(as depicted in Fig. 1(a)(b)). Simultaneously, the input image is passed through an FPN structure to obtain multi-level features (refer to Fig. 1(c)(d)). Subsequently, the pyramid features are up-sampled to a uniform scale and then fed into the Efficient Feature Enhancement Module (EFEM). The EFEM module is designed to be cascaded, offering the advantage of low computational cost. It can be seamlessly integrated behind the backbone network to enrich and enhance the expressive power of features at different scales(see Fig. 1(e)(f)). Following that, we introduce the Multi-scale Feature Fusion Module (MSFM) to effectively amalgamate the features generated by the EFEMs at various depths, resulting in a comprehensive final feature representation for segmentation(as illustrated in Fig. 1(g)). Subsequent to feature extraction, feature F is employed for the prediction of both the probability map (P) and the threshold map (T). Subsequently, we compute the approximate binary map ($\hat{B}$) using the probability map and feature F (refer to Fig. 1(h)(i)(j)(k)). Throughout the training phase, supervision is applied to the probability map, threshold map, and approximate binary map. Notably, the probability map and approximate binary map share the same supervision. During the inference phase, bounding boxes can be effortlessly generated from either the approximate binary map or the probability map using a dedicated box formulation module.

3.2. Efficient Feature Enhancement Module

The EFEM, depicted as a U-shaped module (see Fig. 2), plays a crucial role in enhancing the features of different scales in our proposed text detection system. It consists of three distinct phases that work together harmoniously to achieve optimal performance. In the up-scale enhancement phase, the feature maps undergo iterative enhancement using strides of 32, 16, 8, and 4 pixels. This iterative process ensures that the features at each scale are progressively refined and enriched with more detailed information. One of the key components of the EFEM is the Pyramid Squeeze Attention (PSA) module Zhang et al. (2022), which operates as an efficient attention block. By effectively extracting multi-scale spatial information, the PSA module captures the intricate relationships between different regions of the feature maps. Furthermore, it establishes long-range channel dependencies, allowing the network to capture global context information while preserving local details. This attention mechanism greatly enhances the discriminative power of the features and improves the overall accuracy of text detection.

In the down-scale phase, the EFEM takes advantage of the feature pyramid generated by the up-scale enhancement. Starting from a 4-stride, the enhancement process continues iteratively until reaching a 32-stride. This down-scale enhancement further refines the

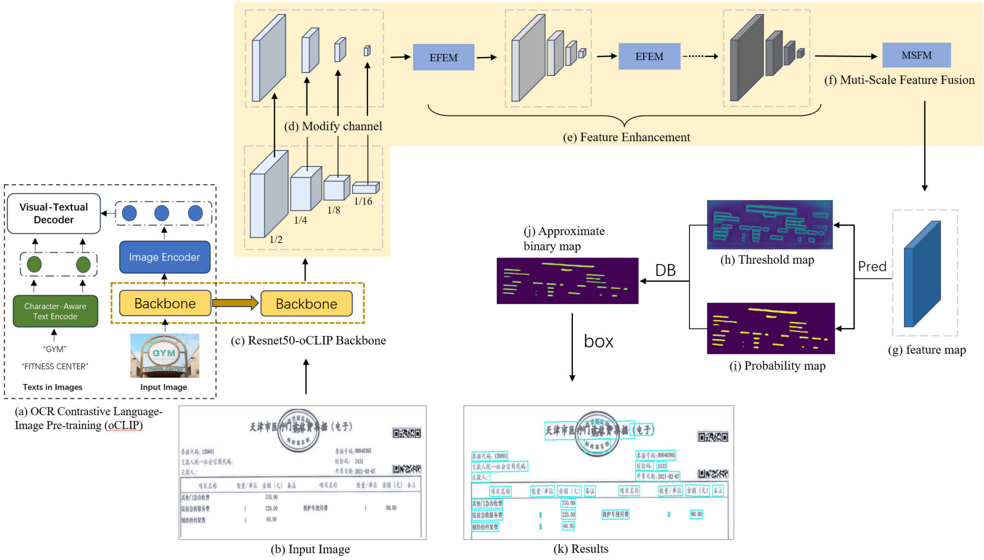

Figure 1: Overall Architecture of our method. The 1/4, 1/8, 1/16, 1/32 indicate the scale ratio compared to the input image.

features by incorporating multi-level information from both lower and higher scales. By integrating low-level details with high-level semantic information, the EFEM ensures that the final feature representation is rich, informative, and capable of capturing the diverse characteristics of text instances.

Compared to existing methods, the EFEM offers two advantages. Firstly, it can be cascaded multiple times to enhance feature fusion and expand receptive fields. This improves the model's ability to handle text instances of various sizes and aspect ratios. Secondly, the EFEM is computationally efficient, making it suitable for real-world applications with limited resources.

3.3. Multi-Scale Feature Fusion Module

In our proposed methodology, we introduce the Multi-scale Feature Fusion Module, which enables the fusion of features from different depths. This module is designed to address the significance of both low-level and high-level semantic information in semantic segmentation. To merge the feature pyramids, we adopt a straightforward and efficient approach that involves upsampling and concatenation. Specifically, we first perform element-wise addition to merge the feature maps of corresponding scales. This step allows for the combination of local details and global context, capturing both fine-grained information and overall scene understanding. Next, the resulting feature maps are upsampled and concatenated, creating a comprehensive final feature map. This fusion method enhances the model's ability to perceive objects and scenes by effectively integrating multi-scale features.

Our proposed fusion method, depicted in Fig. 3, improves the overall performance of semantic segmentation by leveraging the complementary information provided by different

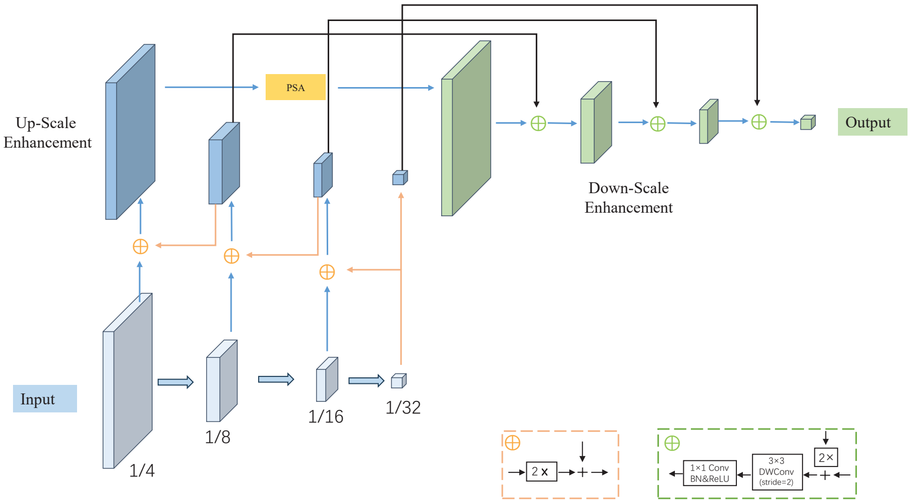

Figure 2: The details of EFEM. “+”, “2×”, “DWConv”, “Conv” and “BN” represent element-wise addition, 2× linear upsampling, depthwise convolution, regular convolution and Batch Normalization respectively.

scales. By incorporating multi-scale features, our approach enables the model to capture fine details, spatial relationships, and contextual cues, leading to more accurate and comprehensive segmentation results.

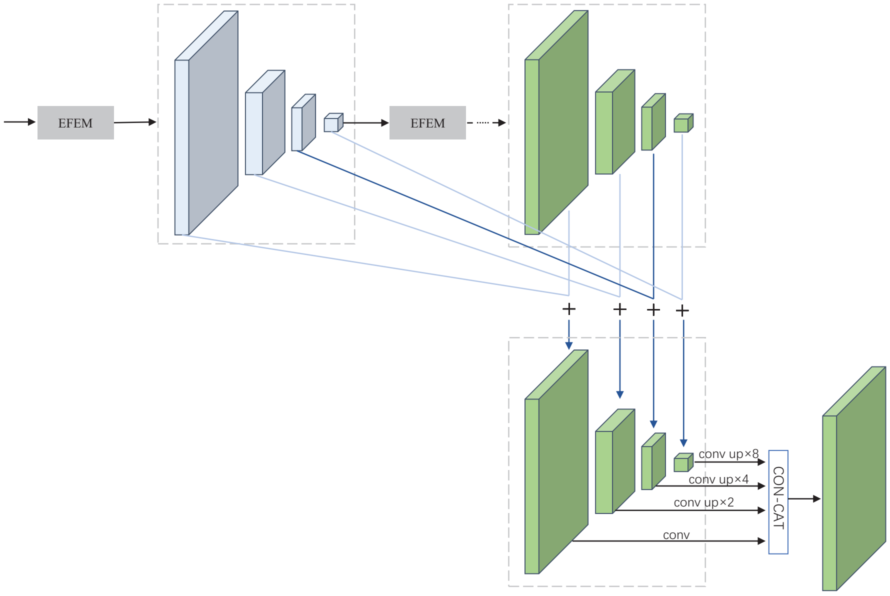

Figure 3: Illustration of the MSFM module. “+” is element-wise addition.

3.4. Deformable Convolution and Label Generation

Deformable convolution Zhu et al. (2019a) can offer a versatile receptive field for the model, which is particularly advantageous for text instances with extreme aspect ratios. In line with Zhu et al. (2019b), modulated deformable convolutions are employed in all the 3×3 convolutional layers within the conv3, conv4, and conv5 stages of the ResNet-50 backbone.

The label generation for the probability map is influenced by PSENet Wang et al. (2019b). Typically, post-processing algorithms display the segmentation results using a collection of vertices that define a polygon:

$$G = \{S_{\mathrm{k}}\}_{k=1}^{n} . \tag{1}$$

n represents the number of vertices, which typically varies depending on the labeling rules in different datasets, and S denotes the segmentation results for each image. Mathematically, the offset D can be calculated as follows:

$$D = \frac{\mathrm{Area}(P) \times (1-\mathrm{r}^2)}{\mathrm{Perimeter}}. \tag{2}$$

In this context, Area(·) refers to the calculation of the polygon's area, and Perimeter(·) denotes the calculation of the polygon's perimeter. The shrink ratio, denoted as "r," is empirically set to 0.4. By employing graphics-related operations, the shrunken polygons can be derived from the original ground truth, serving as the fundamental building block for each text region.During the inference phase, we have the option to utilize either the probability map or the approximate binary map to generate text bounding boxes, which yield nearly identical results.

4. Experimental Results

4.1. Datasets

In addition to the medical text dataset, our experiments involve several widely used arbitrary-shaped public scene text detection datasets, such as SynthText, Total-Text, CTW1500, ICDAR2015, and MSRA-TD500. These datasets are utilized to assess the robustness and generalization capabilities of our proposed model. By evaluating our model on diverse datasets, we can validate its performance across different text detection scenarios and demonstrate its effectiveness in handling various text types and shapes beyond medical text.

MEBI-2000 dataset is derived from the public dataset of Ali Tianchi competition 31. It consists of 2000 scanned images and photos taken with mobile phones, specifically created for detecting and recognizing medical ticket text. MEBI-2000 provides a practical and representative dataset for medical insurance image analysis.

SynthText Gupta et al. (2016) is a synthetic dataset consisting of more than 800k synthetic images. It is only used to pre-train our model.

Total-Text Ch'ng and Chan (2017) is a curved text dataset including 1255 training and 300 testing images. It contains horizontal, multi-oriented, and curve text instances labeled at the word level.

CTW1500 Yuliang et al. (2017) is another curved text dataset, including 1000 training images and 500 testing images. It contains both English and Chinese texts annotated at the text-line level with polygons.

ICDAR 2015 dataset Karatzas et al. (2015) is a commonly used dataset for text detection. It consists of 1000 training images and 500 testing images, which are captured by Google Glass.

MSRA-TD500 dataset Yao et al. (2012) includes 300 training images and 200 test images with text line level annotations. It is a dataset with multi-lingual, arbitrary-oriented and long text lines.

4.2. Implementation Details

We pre-train our network on SynthText Gupta et al. (2016) and then finetune it on the real datasets (MEBI, Total-Text, CTW1500, ICDAR2015, MSRA-TD500). Following the pre-training phase, we proceed with fine-tuning the models for 1200 epochs on the corresponding real-world datasets. During training, our primary data augmentation techniques encompass random rotation, random cropping, random horizontal and vertical flipping. Additionally, we resize all images to 640 × 640 to enhance training efficiency. For all datasets, the training batch size is set to 16, and we adhere to a ”poly” learning rate policy to facilitate gradual decay of the learning rate. Initially, the learning rate is set to 0.007, accompanied by an attenuation coefficient of 0.9. Our framework employs stochastic gradient descent (SGD) as the optimization algorithm, with weight decay and momentum values set to 0.0001 and 0.9, respectively.

4.3. Ablation Study

To demonstrate the effectiveness of key modules, namely deformable convolution, Efficient Feature Enhancement Module, and Multi-Scale Feature Fusion Module, we perform an ablation study on the MEBI dataset and the ICDAR2015 dataset. The experimental results, presented in Table 1, provide detailed insights into the performance of each module. By evaluating the performance of the individual modules, we can assess their contributions to the overall text detection system and validate their effectiveness in improving accuracy and robustness.

The effectiveness of Deformable Convolution As shown in Table 1, for the ICDAR2015 dataset, the deformable convolution increase the F-measure by 2% . For the MEBI dataset, 3.3% improvements are achieved by the deformable convolution. Clearly, deformable convolution can provides a flexible receptive field for the backbone, with small extra time costs.

The effectiveness of EFEM As shown in Table 1, for the ICDAR2015 dataset, The effectiveness of EFEM increase the F-measure by 8%, 10.5% improvements are achieved by the EFEM+DConv. For the MEBI dataset, The effectiveness of MSFM increase the F-measure by 7.7%, 10.4% improvements are achieved by the EFEM+DConv. The inference speed decreases slightly.

The effectiveness of MSFM As shown in Table 1, for the ICDAR2015 dataset, The effectiveness of MSFM increase the F-measure by 5.6%, 7.5% improvements are achieved by the MSFM+DConv. For the MEBI dataset, The effectiveness of MSFM increase the F-measure by 6.6%, 9.3% improvements are achieved by the MSFM+DConv. The inference speed decreases slightly.

Table 1: Detection results with different settings of Deformable Convolution, EFEM and MSFM.

Backbone	DConv2	EFEM	MSFM	ICDAR 2015				MEBI			
				P	*R*	*F*	*FPS*	*P*	*R*	*F*	*FPS*
oclip-ResNet50	×	×	×	83.3	71.2	76.8	**30**	82.1	71	76.1	**38.7**
oclip-ResNet50	✓	×	×	85.4	73.1	78.8	28.5	85.2	74.4	79.4	34.5
oclip-ResNet50	×	×	✓	88.7	76.9	82.4	26	87.6	78.3	82.7	33
oclip-ResNet50	✓	×	✓	90.2	79.1	84.3	22.7	87.8	83.2	85.4	27.2
oclip-ResNet50	×	✓	×	89.9	80.2	84.8	25.1	86.2	81.5	83.8	32.4
oclip-ResNet50	✓	✓	×	91.3	83.7	87.3	19.4	87.4	85.6	86.5	25.5
oclip-ResNet50	✓	✓	✓	**92.4**	**86.1**	**89.1**	16	**89.5**	**87.1**	**88.3**	23.9

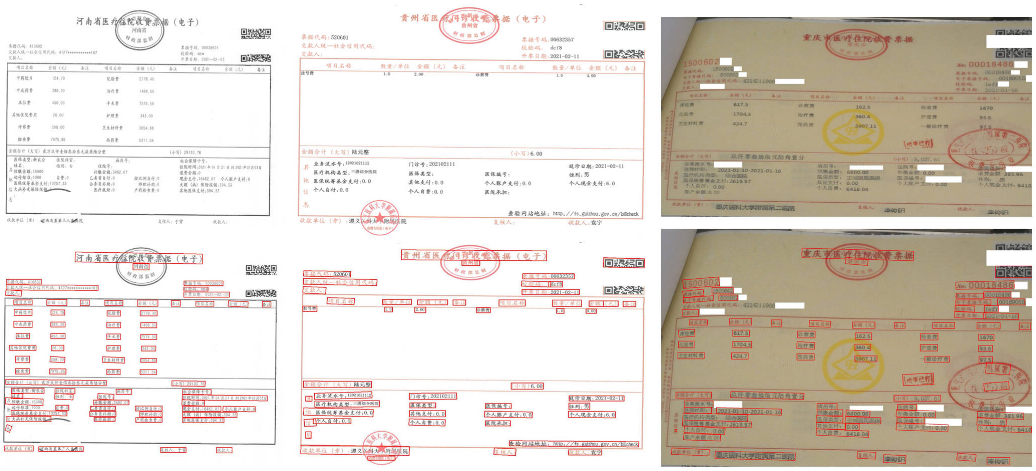

Figure 4: Some visualization results on text instances of MEBI.

Table 2: Detection results on the MEBI dataset.

Method	P	R	F	FPS
EAST Zhou et al. (2017)	78.7	70.4	74.3	15
PSE-Net Wang et al. (2019b)	84.9	85.7	85.3	4
SAST Wang et al. (2019a)	85.1	80.4	82.7	-
PAN Wang et al. (2019c)	87.6	82.1	84.8	-
DBNet Liao et al. (2020)	82.3	74.2	78.1	20
DBNet++ Liao et al. (2022)	83.3	80.1	81.7	15
Ours(ResNet-50)	88.5	86.3	87.4	**25**
Ours(oCLIP-ResNet-50)	**89.5**	**87.1**	**88.3**	23.9

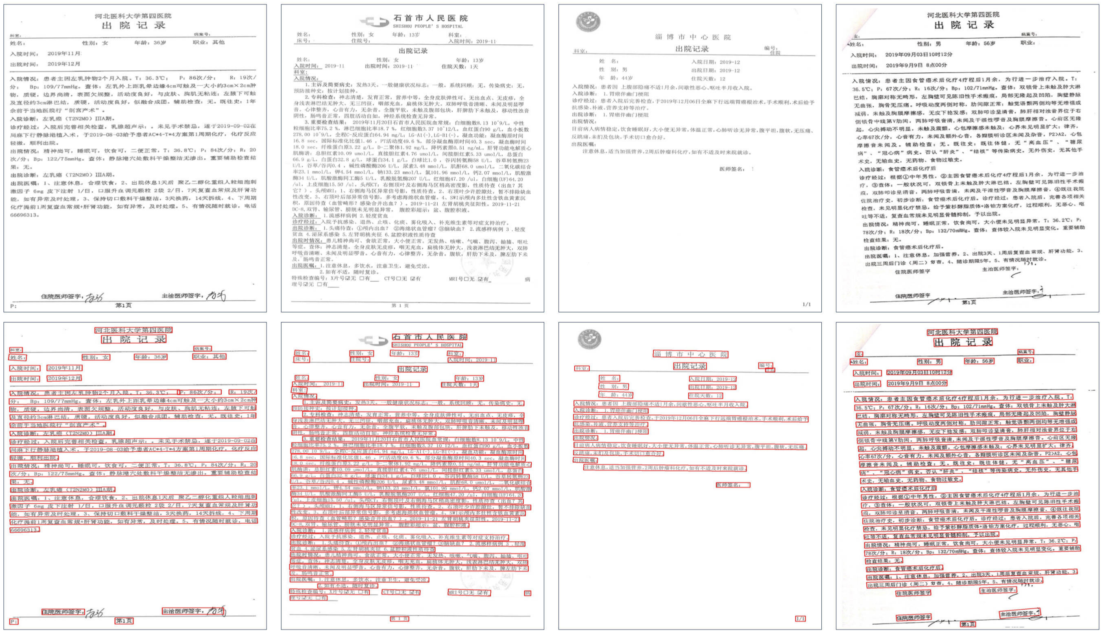

Figure 5: Some visualization results on text instances of MEBI.

4.4. Comparisons with previous methods

We evaluate the performance of our proposed method by comparing it with previous methods on five well-established benchmarks. These benchmarks cover a range of text scenarios, including medical text, curved text, multi-oriented text, and long text lines in multiple languages. To provide a comprehensive assessment, we conduct both quantitative and qualitative evaluations. The quantitative results are presented in terms of benchmark scores, while the qualitative results are visualized in Fig. 6. By analyzing the results, we can demonstrate the superiority of our proposed method and its ability to handle diverse text scenarios effectively.

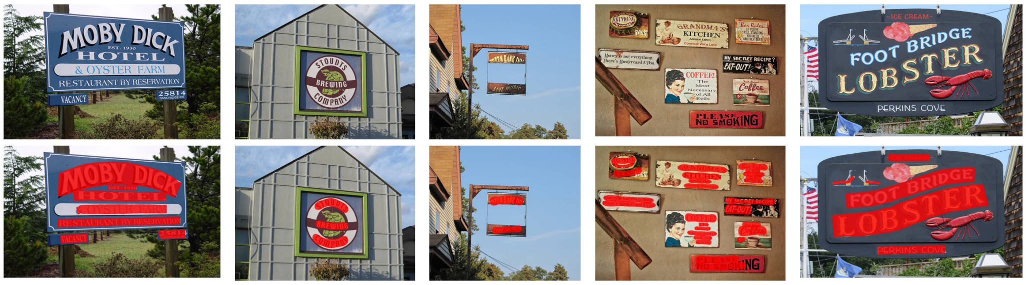

Figure 6: Some visualization results on text instances of various shapes, including curved text, multi-oriented text, vertical text.

Table 3: Detection results on the Total-Text dataset.

Method	P	R	F	FPS
TextSnake Long et al. (2018)	82.7	74.5	78.4	-
PSE-Net Wang et al. (2019b)	84.0	78.0	80.9	3.9
SPCNet Xie et al. (2019a)	83.0	82.8	82.9	-
LOMO Zhang et al. (2019)	87.6	79.3	83.3	-
PAN Wang et al. (2019c)	89.3	81.0	85.0	**39.6**
DBNet Liao et al. (2020)	87.1	82.5	84.7	32
ContourNet Wang et al. (2020)	86.9	83.9	85.4	-
DRRG Zhang et al. (2020)	86.5	**84.9**	85.7	-
DBNet++ Liao et al. (2022)	88.9	83.2	86.0	28
Ours	**89.6**	83.7	**86.5**	38

Medical text detection The MEBI dataset is specifically designed for medical text, encompassing a wide range of textual instances with diverse scales, irregular shapes, and extreme aspect ratios. In Table 2, we compare our model with previous methods, and our model achieves state-of-the-art performance in terms of accuracy, f-measure, recall, and fps. Furthermore, we provide visualizations of medical text instances from the MEBI dataset in Fig. 4 and Fig. 5.

Table 4: Detection results on the CTW1500 dataset.

Method	P	R	F	FPS
EAST Zhou et al. (2017)	78.7	49.1	60.4	21.2
TextSnake Long et al. (2018)	67.9	85.3	75.6	-
PSE-Net Wang et al. (2019b)	84.8	79.7	82.2	3.9
LOMO Zhang et al. (2019)	**89.2**	69.6	78.4	-
PAN Wang et al. (2019c)	86.4	81.2	83.7	-
DBNet Liao et al. (2020)	86.9	80.2	83.4	22
ContourNet Wang et al. (2020)	83.7	84.1	83.9	-
DRRG Zhang et al. (2020)	85.9	83.0	84.5	-
DBNet++ Liao et al. (2022)	87.9	82.8	85.3	26
Ours	88.2	**83.1**	**85.6**	**32**

Curved text detection Table 3 and Table 4 present the performance of scene text detection on two curved text datasets, namely Total-Text and CTW1500. Our approach demonstrates outstanding performance in terms of F-measure on both CTW1500 and Total-Text datasets. Specifically, compared to the DBNet++ method, which is also a segmentation-based approach and performs well on these datasets, our method achieves similar performance with a significantly reduced inference time, approximately 75% of the time required by DBNet++. This highlights the efficiency and effectiveness of our proposed approach in handling curved text instances.

Table 5: Detection results on the ICDAR 2015 dataset.

Method	P	R	F	FPS
CTPN Tian et al. (2016)	74.2	51.6	60.9	7.1
EAST Zhou et al. (2017)	83.6	73.5	78.2	13.2
SSTD He et al. (2017a)	80.2	73.9	76.9	7.7
WordSup Hu et al. (2017)	79.3	77	78.2	-
TB++ Liao et al. (2018a)	87.2	76.7	81.7	11.6
PSE-Net Wang et al. (2019b)	86.9	84.5	85.7	1.6
SPCNet Xie et al. (2019a)	88.7	85.8	87.2	-
LOMO Zhang et al. (2019)	91.3	83.5	87.2	-
PAN Wang et al. (2019c)	84.0	81.9	82.9	**26.1**
SAE Tian et al. (2019)	85.1	84.5	84.8	3
DBNet Liao et al. (2020)	91.8	83.2	87.3	12
DBNet++ Liao et al. (2022)	90.9	83.9	87.3	10
Ours(ResNet-50)	**92.5**	83.1	87.5	15.4
Ours(oCLIP-ResNet-50)	92.4	**86.1**	**89.1**	13

Table 6: Detection results on the MSRA-TD500 dataset.

Method	P	R	F	FPS
DeepReg He et al. (2017b)	77	70	74	1.1
RRPN Ma et al. (2018)	82	68	74	-
RRD Liao et al. (2018b)	87	73	79	10
MCN Liu et al. (2018)	88	79	83	-
PixelLink Deng et al. (2018)	83	73.2	77.8	3
CRAFT Baek et al. (2019)	88.2	78.2	82.9	8.6
SAE Tian et al. (2019)	84.2	81.7	82.9	-
PAN Wang et al. (2019c)	84.4	83.8	84.1	30.2
DBNet Liao et al. (2020)	91.5	79.2	84.9	32
DRRG Zhang et al. (2020)	88.1	82.3	85.1	-
MOST He et al. (2021)	90.4	82.7	86.4	-
TextBPN Zhang et al. (2021)	86.6	**84.5**	85.6	-
DBNet++ Liao et al. (2022)	91.5	83.3	87.2	29
Ours(ResNet-50)	92.2	81.4	86.5	**36**
Ours(oCLIP-ResNet-50)	**92.3**	82.9	**87.3**	34.2

Multi-oriented and Multi-oriented text detection In Table 5, our method surpasses DBNet++, which is recognized as one of the most influential and efficient segmentation-based approaches, achieving a 1.8% improvement in F-measure. In Table 6, our method exhibits comparable or even superior performance in F-measure compared to DBNet++, while also enjoying a 1.24 times faster inference speed on MSRA-TD500. These results demonstrate that our method strikes a favorable balance between detection performance and inference speed, making it a promising choice for practical applications.

5. Conclusion

In summary, we have presented an efficient framework for real-time detection of medical and arbitrary-shaped text using a vision-language pre-training model. Our approach incorporates the Efficient Feature Enhancement Module and Multi-scale Feature Fusion Module to enhance feature extraction without significant computational overhead. Through extensive experiments on multiple datasets, we have demonstrated the effectiveness of our method, achieving notable improvements in both speed and accuracy compared to previous state-of-the-art text detectors. The integration of visual and language understanding tasks enhances the model's ability to capture rich visual and semantic representations, making it suitable for various applications like medical image analysis and document processing.

Future work can focus on further optimizing computational efficiency and extending the framework's capabilities to handle more challenging text detection scenarios.

Acknowledgement

This work was supported by the Scientific Research Funds of Northeast Electric Power University (No. BSZT07202107).

References

Medical Inventory Invoice OCR Element Extraction Task (CMedOCR). URL https://tianchi.aliyun.com/dataset/131815.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019.

Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017.

Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European conference on computer vision (ECCV)*, pages 186–202, 2018a.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018b.

Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8554–8564, 2019.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016.

Minghang He, Minghui Liao, Zhibo Yang, Humen Zhong, Jun Tang, Wenqing Cheng, Cong Yao, Yongpan Wang, and Xiang Bai. Most: A multi-oriented scene text detector with localization refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8813–8822, 2021.

Pan He, Weilin Huang, Tong He, Qile Zhu, and Xiaolin Li. Single shot text detector with regional attention. *IEEE*, 2017a.

Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE international conference on computer vision*, pages 745–753, 2017b.

Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding. Wordsup: Exploiting word annotations for character based text detection. *arXiv e-prints*, 2017.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018a.

Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5909–5918, 2018b.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481, 2020.

Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):919–931, 2022.

Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, and Wang Ling Goh. Learning markov clustering networks for scene text detection. *arXiv preprint arXiv:1805.08365*, 2018.

Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018.

Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE transactions on multimedia*, 20(11):3111–3122, 2018.

Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 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*, pages 56–72. Springer, 2016.

Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4234–4243, 2019.

Pengfei Wang, Chengquan Zhang, Fei Qi, Zuming Huang, Mengyi En, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. A single-shot arbitrarily-shaped text detector based on context attended multi-task learning. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019a.

Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9336–9345, 2019b.

Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8440–8449, 2019c.

Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. 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*, pages 11753–11762, 2020.

Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li. Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9038–9045, 2019a.

Lele Xie, Yuliang Liu, Lianwen Jin, and Zecheng Xie. Derpn: Taking a further step toward more general object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9046–9053, 2019b.

Chuhui Xue, Wenqing Zhang, Yu Hao, Shijian Lu, Philip Torr, and Song Bai. Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. *Springer, Cham*, 2022.

Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012.

Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017.

Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding. 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*, pages 10552–10561, 2019.

Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. In *Proceedings of the Asian Conference on Computer Vision*, pages 1161–1177, 2022.

Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9699–9708, 2020.

Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Adaptive boundary proposal network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1305–1314, 2021.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019a.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019b.