Received: July 7, 2023. Revised: September 10, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Improving Text Recognition in Natural Images Using Contextual Modules and Transformer-based Decoding

Kapil Kumar[1]* **Abhishek Kumar Mishra[2]**

[1]*School of Computer Science & Applications, IFTM University, Moradabad, U.P., India*
[2]*Department of Computer Science & Engineering, IFTM University, Moradabad, U. P., India*
* Corresponding Author Email: kapilchauhan.svm@gmail.com

Abstract: This research study proposes a comprehensive approach to improve text recognition in natural photographs. The approach addresses challenges such as curved text, low-resolution images, and efficient recognition algorithms by integrating rectification, visual feature extraction, semantic context modeling, and global context modeling. The model comprises several modules for text recognition. The rectification module normalizes non-uniform text areas using a spatial transformation network to recognize curved text accurately. Visual feature extraction captures complicated patterns and improves picture discrimination with ResNet50. The global context module addresses text sequence dependencies, whereas the semantic context module gathers semantic information. Transformer-based text decoding uses masked multi-head attention. Evaluating benchmark datasets (TotalText, CTW1500, and ICDR15) demonstrates promising results, with the ResNet50 backbone achieving impressive F-measures of 89.3%, 86.4%, and 93.7%, respectively. This research successfully combines rectification, visual feature extraction, semantic context modeling, and global context modeling techniques to address challenges in text recognition for natural photographs. The proposed strategies and components contribute to the model's overall improvement of text recognition.

Keywords: Semantic contextual module, Global contextual module, ResNet50, Transformer, Multihead attention.

1. Introduction

Identifying text within natural images poses a considerable obstacle due to a range of environmental limitations, such as lighting circumstances and the existence of curved and diminutive characters. Researchers have classified text recognition methods into two distinct categories: segmentation-based and regression-based approaches.

Segmentation-based techniques employ the full convolution network (FCN) [1] as a fundamental tool for detecting text. Nevertheless, it is common to require supplementary post-processing measures to enhance the accuracy of the identified text regions. In contrast, regression-based approaches utilize well-established methodologies such as single shot detectors (SSD) [2], faster R-CNN [3]. These techniques aim to perform a direct regression of the bounding boxes of text regions. Both methodologies have been thoroughly examined to enhance efficiency while minimizing model intricacy.

Optical character recognition (OCR) is a multifaceted methodology utilized to identify and classify textual data within images. The process involves two distinct stages, namely detection and recognition. During the detection stage, the text region is segmented, while in the recognition stage, the text is extracted from the segmented region. The efficacy of optical character recognition (OCR) may be impeded by suboptimal document quality and restricted quantities of text[4].

The popularity of deep learning techniques in text recognition has increased. However, the focus has been primarily on accuracy metrics, with other significant factors such as memory usage and inference time being overlooked. Edge computing has emerged as a popular trend, particularly for mobile applications. This technology facilitates text detection directly on edge devices such as smartphones[5, 6]. Developing efficient and appropriate models for edge devices is crucial to attaining maximum computational efficiency.

Researchers have made significant advancements in developing text identification models that prioritize speed and cost-effectiveness. This has been achieved by reducing the number of floating-point operations per second (FLOPS). The EAST [7] algorithm is a prominent method for detecting scene text based on the PvaNet framework[8]. The system utilizes a convolutional neural network to extract visual characteristics and produce per-pixel forecasts of text regions. The optimization process comprises three main steps: removal of candidate selection, text area construction, and word segmentation.

This research study focuses on enhancing text recognition in natural photographs. Recognizing text in images presents challenges such as curved text, low-resolution images, and the need for efficient recognition algorithms. This study aims to propose an integrated approach combining different techniques to effectively address the challenges associated with text recognition, ultimately achieving high accuracy and efficiency.

Our Contribution to the research paper

1. A novel rectification network corrects image text misalignment. Innovative rectification network. The rectification module requires accurate control point identification. Image resizing with 32×64 dimensions enabled this capability.
2. ResNet50 uses a CNN module to extract 2D characteristics from images. The technique departs from 1D-feature-based approaches. A final layer was added to the ResNet50 module to generate a 4 × 25 picture with 25 width. The layer maintains horizontal pixel data and improves text recognition in longer photos.
3. The semantic context module (SCM) is designed to provide contextual information to other software systems. Semantic content management (SCM) was created to collect and organize textual semantic data. The SCM system created several semantic context components, improving the model's grasp of textual information.
4. The utilization of the global context module (GCM) ensured precise representation of sequence context dependencies in the text.
5. We are proposed a decoding module based on transformer. The system comprises the masked multi-head attention mechanism, feed-forward network (FFN), and layer normalization.

Collaborations that handle misalignment, visual changes, and semantic context have enhanced text recognition. The suggested methods and components improve the model's text recognition and comprehension across many contexts.

Structured as follows, the paper's succeeding sections: Scene text detection literature is reviewed in section 2. Section 3 proposes scene text detection using ResNet50 with a multi-head attention network. Experimental setups, technical details, and model evaluation metrics are described in section 4. Section 5 presents the indicated approaches' results. Section 6 concludes the suggested research.

2. Related works

The domain of text recognition encounters three fundamental challenges in its current state:

1. The prevalence of curved text images within natural scenes presents a notable obstacle that must be overcome.
2. A considerable abundance of low-resolution text images further complicates the task of accurate recognition.
3. The demand for enhanced recognition speed in practical applications necessitates a focus on optimizing efficiency.

Numerous studies have actively concentrated on devising innovative approaches and strategies to address these challenges.

When dealing with curved text in natural environments, it is crucial to consider that submitting the entire text area to text recognition systems may yield suboptimal recognition outcomes because of numerous invalid regions. In their study, Shi et al. [9] introduced an automated rectification method designed to tackle the problem of curved text. The module employs a spatial transformation network to execute the image conversion process into a more uniform and readable format. The transformation process effectively corrects different forms of non-standard text, thereby enhancing the capability to detect curved text. In their publication, Shi et al. [10] presented a novel rectification network aimed at reducing the intricacy of the rectification module. The network has a thin-plate spline transformer that incorporates a specialized attention mechanism designed to handle non-uniform text. During the training process, the curved text points are manipulated to align them horizontally. In their study, Zhan et al. [11] suggested using a repetitive rectification framework to improve text recognition's efficacy. The approach employed in this methodology involves the iterative rectification of text regions, as opposed to the singular rectification technique utilized in references[9, 10]. Despite the advancements in curved text recognition techniques, addressing low-quality text images still needs to be solved.

To address the low quality text recognition, Zhang et al. [12] improves text recognition by addressing hazy pictures. Similarity units in the network synchronize attention information with sequence data, improving feature representation and gated attention. Wan et al. [13] improved recognition of scene text using attention and segmentation models. Visual and semantic information was integrated more effectively. Attention and segmentation models improve the technique. Attention models improve low-resolution pictures, whereas segmentation models enhance visual features. Yu et al. [14] proposed a semantic inference network that uses actual semantic text data to categorize and horizontally arrange characters using visual and semantic context information. This method improves public dataset performance. Zhang et al. [15] developed a search-automated text recognition system. The system adapts to neural architecture exploration datasets. This approach recognizes low-resolution text using many robust modules. Several studies have sped up text recognition. Zhu et al. [16] created a unique text recognition system using transformer-based NLP and parallel computing. A strong backbone and transformer network encode contextual information via a hierarchical attention method with four self-attention blocks. The models mentioned above integrated various modules and were generally characterized by their substantial dimensions, enabling them to process low-resolution text for recognition purposes effectively. Several studies have also underscored the significance of enhancing the efficiency of text recognition.

Li et al. [17] propose a transformer model that uses self-attention processes to improve text recognition. Locality-sensitive hashing compresses this model. This method simplifies softmax regression and speeds processing by reducing parameters. The model can detect low-resolution text well, although it may need to. Lee et al., adevised adaptive 2D positional encoding [18]. Transformer networks improve feature extraction and solve irregular picture challenges. This approach works well for analyzing open datasets, especially ones with abnormalities. The typical transformer's encoding component has six self-attention levels; our technique has 12 [19]. Kim et al. [20] altered the transformer model's encoder. To boost its performance, they replaced it with a squeeze-and-excitation feedforward neural network (FFN). The transformer's primary goal is faster logistics-related inference. Self-attention modules help transformer-based text recognizers outperform CNN-based ones. The transformer concept is very effective yet computationally intensive. Text recognition technology is increasing in logistics, making velocity important. Ren et al. [21] used transformer networks to recognize shopping receipt text. They also solved attention drift. A transformer-based decoupled attention network partitions prediction processes using an attention mechanism, improving efficacy.

TextSnake [27] proposes a new approach that utilizes discs to represent text. They predict centerlines and scores using the FPN network, resulting in accurate segmentation. TextField [28] employs a convolutional neural network to encode direction fields, specifically targeting irregular text and effectively resolving challenges associated with separating adjacent text instances. The CA-STD [31] utilizes a feature refinement module to enhance feature representation and incorporates a conditional attention mechanism to handle spatial-textual relationships. Contour information aggregation enhances feature representations, thereby facilitating accurately identifying diverse text shapes. JMNET [32] utilizes lightweight features and incorporates the scale spatial perception module (SSPM) and attention spatial perception module (ASPM) to enhance feature representation. The proposed unsupervised embedding spatial perception loss function improves robustness by effectively handling uncertain boundaries in text. The TransText [33] model improves scene text detection with arbitrary shapes by utilizing parallel branches and a modified Transformer decoder. The CTPN [34] demonstrates superior performance in precisely localizing text lines. This is achieved by directly analyzing text proposals at a fine-scale level within convolutional feature maps. The integration of recurrent neural networks and their efficacy renders them a valuable asset in text detection. TPLAANet [35] is a comprehensive method that handles text with various orientations and shapes through a unified architecture. TPLAANet addresses the detection task by employing central mask prediction, bounding box regression, and mask accuracy. In addition, the location-awareness-attention network improves character information by employing two-dimensional attention weights for recognition.

Previous studies have shown that transformer networks can effectively process low-quality text images. However, their performance in real-world applications was lacking in terms of speed. The transformer-based approaches did not investigate simplification strategies for speeding up text recognition. This study introduces a new text

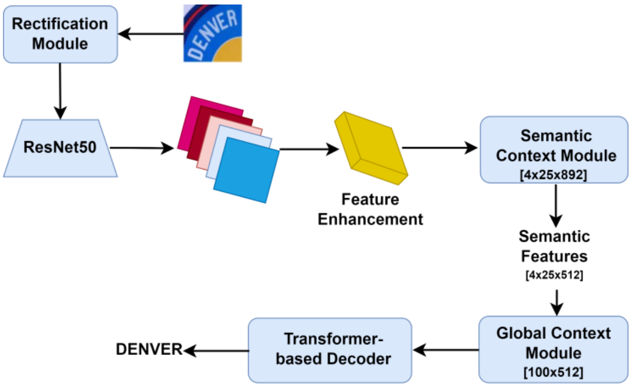
Rectification Module
DENVER
ResNet50
Feature Enhancement
Semantic Context Module [4x25x892]
Semantic Features [4x25x512]
Global Context Module [100x512]
Transformer-based Decoder
DENVER

Figure. 1 Proposed framework

recognizer that simplifies the transformer network and includes two additional small modules. These modules are used to complement and improve the modeling of context and long-range dependencies in order to handle low-quality text images effectively.

3. Proposed methodology

The research paper's model includes the rectification network, visual feature extraction using ResNet50 [22], contextual module (Semantic and Global), and transformer-based decoding module. We reduced non-uniform images to 32×100 to reduce their effect. To align the text horizontally, we resize the images to 32x64 to help the rectification module identify control points. Our CNN-based module extracts two-dimensional (2D) characteristics, unlike typical recognition methods that use one-dimensional (1D) features. The CNN module's final layer outputs a 4 × 25 image with a width of 25 to maintain horizontal pixel information and improve text recognition in lengthy images. SCM captures semantic information and generates several semantic context aspects. We then use the GCM to represent sequence context dependencies. We construct a simplified transformer decoding component with N = 3 blocks. Each block has a masked multi-head attention mechanism, an FFN, and layer normalization. Fig. 1 shows the network structure.

3.1 Rectification module

The proposed model integrates a spatial transformation network (STN) [23] into its rectification network to mitigate the problem of non-uniform or distorted images. The model mentioned above employs a modified version of the spatial transformer network (STN), specifically the thin-plate spline (TPS) [9] technique, to normalize non-uniform text regions to detect scene text. The rectification module utilizes fiducial points and seamless spline interpolation to accurately align and standardize text regions' shapes. Using accurate feature extraction methodologies yields improved pipeline efficiency for text recognition

3.2 ResNet50 for visual features extraction

Academic study praises ResNet50 for its capacity to solve overfitting, vanishing gradients, parameter efficiency, and feature representation. "Identity shortcut connection" enhances gradient propagation and reduces vanishing gradients.

VFE was based on ResNet50 in our study. The feature extraction module's output layer is 4x25xC, reflecting the image's height, width, and channels. A feature improvement module that integrates high- and low-level semantic information improves data comprehension and feature representations.

A CNN merges traits from layers 3, 4, and 5 to create a cohesive feature map. This amalgamation captures and depicts subtle and layered properties necessary for later processing.

Our method extracts unique features from visual data by fully leveraging ResNet50 and adding a feature improvement module. This enhances image categorization and item recognition.

3.3 Semantic contextual module

Fig. 2 demonstrates the integration of the residual crisscross attention module, which enhances feature extraction in our framework. Two convolutional modules are included to improve this process further. We intentionally chose not to incorporate recurrent crisscross attention [16] in our methodology. To overcome feature degradation and establish reliable dependencies on textual data, we employ residual networks inspired by ResNet50 [3]. This approach is particularly advantageous when dealing with lengthy text sequences. The cross-attention technique, proposed initially for semantic segmentation, effectively captures contextual information. Fig. 2 visually represents how cross-attention operates within our framework. Our approach significantly enhances text recognition performance by employing advanced methodologies. These methodologies include the integration of the residual crisscross attention module, convolutional modules, and cross-attention techniques. They optimize feature extraction, mitigate degradation issues, and effectively capture contextual information.

The crisscross attention is first applied on feature map X of size H x W x C, to compute the crisscross attention, we can define a set of learnable parameters, such as weight matrices, for the attention mechanism.

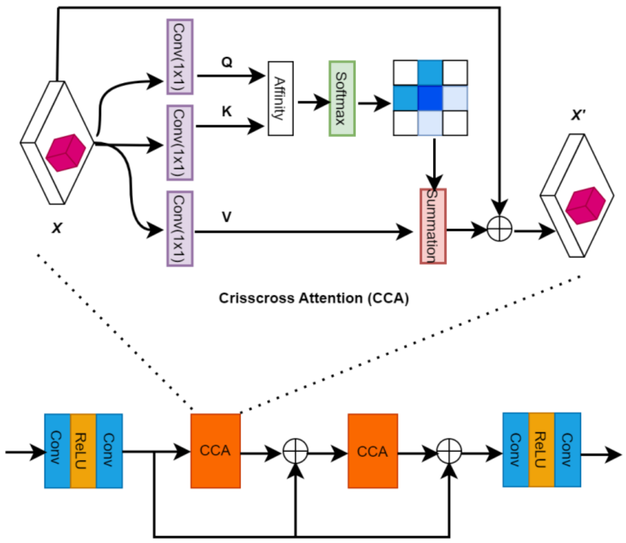

Figure. 2 Semantic contextual module

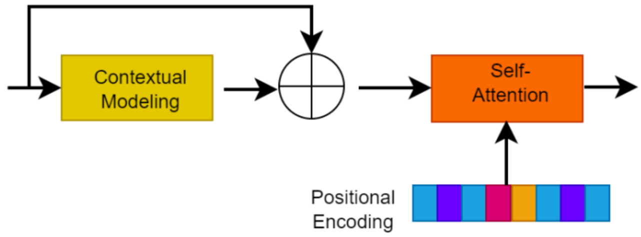

Figure. 3 Global context module

First, we reshape the input X into a sequence of feature vectors.

$$X_{Reshape} = Reshape\big(X, (H \times W, C)\big) \qquad (1)$$

Next, we perform linear transformations on $X_{Reshape}$ to obtain query (Q), key (K), and value (V) matrices:

$$Q = X_{Reshape} \times W_q \qquad (2)$$

$$K = X_{Reshape} \times W_k \qquad (3)$$

$$V = X_{Reshape} \times W_v \qquad (4)$$

Here, W_q, W_k, W_v are weight matrices specific to the crisscross attention layer.

Now, we compute the attention scores by taking the dot product between query (Q) and key (K) matrices:

$$Attention = softmax\big(Q \times K^T / \sqrt{d_k}\big) \qquad (5)$$

d_k represents the dimensionality of the key matrix.

Next, we apply the attention scores to the value (V) matrix to obtain the attended values Ag):

$$Attended\ (Ag) = Attention \times V \qquad (6)$$

Finally, we reshape the attended values back to the original spatial dimensions:

$$Ag_{Reshape} = Reshape(Ag, (H, W, C)) \qquad (7)$$

The output of the crisscross attention layer would be Attended reshape, which captures the dependencies between different spatial locations in the input feature map. The SCM typically accepts X' as an input and generates the feature map X" as an output, which can be calculated using the following procedure.

$$X'' = CCA\big(X' \oplus CCA(X')\big) \oplus X' \qquad (8)$$

where CCA denotes the crisscross attention; ⊕ represents element-wise addition

3.4 Global context module

The GCM (global context module) incorporates context modeling, self-attention, and location encoding components to capture text-dependent dependencies. We build upon previous research [30] to effectively distribute attention values through element-wise addition, emphasizing character features for accurate text recognition.

To address challenges related to long text sequences and attention drift, we enhance the GCM self-attention module by introducing learnable character weights. This improvement allows for dynamic weighting of characters, resulting in improved accuracy and focused attention.

The global context block compresses the feature map's height to a single value, producing a precise feature sequence representing the input text. This ensures effective understanding and encoding of fundamental text attributes for further processing.

Fig. 3 depicts the GCM's context modeling, self-attention, and location encoding modules, designed to capture distant relationships, reduce attention drift, and generate an accurate feature sequence reflecting the input text.

3.5 Decoder

The GCM (global context module) significantly advances our research by facilitating global data collection. Our transformer architecture's decoder component demonstrates improved streamlining and efficiency compared to the reference work [7]. It consists of three layers that effectively reduce complexity while maintaining efficiency.

We employ a masked multi-head attention mechanism in the decoder to capture interdependencies among decoding positions. Alongside the conventional multi-head attention mechanism and feed-forward network (FFN), we propose an additional element to simulate better and depict interconnections.

Residual connections [29] are implemented in each sublayer of the decoder to enhance gradient flow and accelerate convergence during training. Layer normalization ensures consistent and homogeneous transformations.

Our study incorporates eight parallel attention mechanisms within the multi-head attention mechanisms. This approach allows us to acquire a broader range of comprehensive and intricate contextual information.

4. Experimental setups

4.1 Dataset

The proposed method's effectiveness was evaluated through experimentation on three benchmark datasets that are widely recognized, as shown in Table 2.

4.2 Implementation details

In our experimental configuration on Kaggle, we utilized PyTorch and ResNet50. The model was pre-trained on the SynthText dataset and then fine-tuned on real datasets. The training process consisted of 200 epochs with a batch size of 64. The initial learning rate was set to 1×10^{-4}

To enhance the variety of training samples, we applied data augmentation techniques. These included random rescaling of images between 0.5 and 2.0, horizontal flipping, and random rotations ranging from -10° to 10°. Random cropping was performed on the transformed images to extract size 640 x 640 samples. The input images were standardized during the inference stage by scaling the shorter side to a predetermined length while maintaining the aspect ratio. The integration of pre-training, fine-tuning, and data augmentation techniques aimed to improve model's performance and generalization.

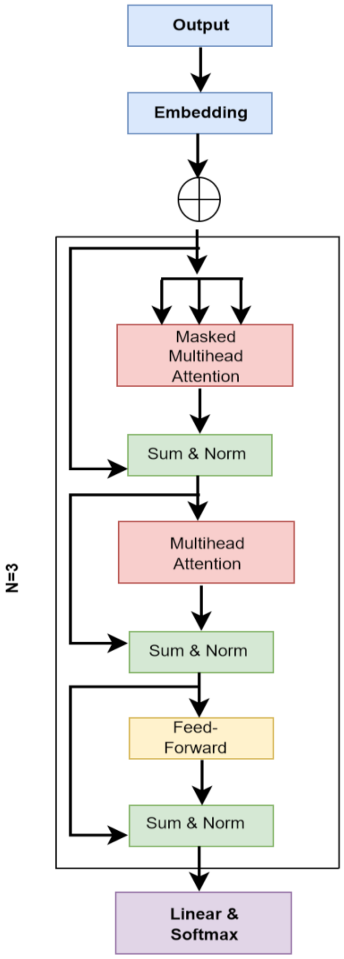

Figure. 4 Transformer based decoder module

Table 1. Dataset description

DataSets	Train Image	Test Image	Focus
ICDR15[24]	1000	500	Incidental scene text
CTW1500[25]	1000	500	Horizontal, Multi-oriented, curved texts
TotalText[26]	1225	300	Curved texts, horizontal, multi-oriented

4.3 Evaluation metric

The evaluation metric for our proposed methods are;

$$Precision\ (P) = t_p\ /(t_p + f_p) \tag{9}$$

$$Recall\ (R) = t_p/(t_p + f_n) \tag{10}$$

$$F1 - Score\ (F) = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{11}$$

Table 2. CNN backbones, highlighting the modified ResNet50's ability to capture well-defined features with a balanced model size and accuracy

Backbone Network	**Accuracy**		
	ICDR15	**CWT1500**	**TotalText**
VGG16	91.2	76.5	91.4
ResNet18	93.8	83.8	94.5
ResNet34	94.6	86.5	96.7
ResNet50	**98.1**	**92.7**	**98.6**
EfficientNet	97.3	90.1	97.1

Table 3. Comparing the performance of different decoder block and attention head configurations in the proposed framework reveals a marginal decrease in performance when increasing the number of heads (N=3, H=8) due to overfitting

# of Decoder Block	**# of Head**	**Accuracy**		
		ICDR15	**CTW1500**	**TotalText**
1	8	93.8	86.3	96.7
1	16	94.0	87.4	97.1
2	16	94.2	87.8	97.5
3	**8**	**97.5**	**91.7**	**98.2**
3	16	96.3	88.7	97.8
4	16	96.8	89.0	97.5

Table 4. Comparison with other state-of-arts method in TotalText dataset

References	**TotalText**		
	P	**R**	**F**
TextSnake[27]	82.7	74.5	78.4
TextField[28]	81.2	79.9	80.6
CRAFT[29]	87.6	79.9	83.6
PSENet[30]	84.0	78.0	80.9
CA-STD[31]	82.9	82.1	82.5
JMNet[32]	90.3	81.2	85.2
TransText[33]	90.8	**83.5**	87.0
Ours	**94.7**	83.4	**88.7**

P: Precision, R: Recall, F: F1-score

Table 5. Comparison with other state-of-arts method in CTW1500 dataset

References	**CTW1500**		
	P	**R**	**F**
CTPN[34]	60.4	53.8	56.9
EAST[7]	78.7	49.1	60.4
TextSnake[27]	67.9	**85.3**	75.6
PSENet[30]	84.8	79.7	82.2
CA-STD[31]	83.1	84.5	83.8
Ours	**90.8**	82.4	**86.4**

P: Precision, R: Recall, F: F1-score

where t_p, f_n, and f_p represent the true positive, false negative, and false positive values, respectively.

5. Results

5.1 Ablation study

We conducted experiments using different CNN models in the feature extraction process, as detailed in Table 3. The models used in the study were VGG16, ResNet18, ResNet34, ResNet50, and EfficientNetB0. ResNet50 is distinguished among these models for its more profound architecture, incorporation of residual connections to address the degradation issue, and a well-balanced compromise between model size and accuracy. The pre-trained representations facilitate efficient transfer learning.

Our experimental results indicate that employing 8 attention heads in both the encoder and decoder components yielded positive results. Based on the favorable outcomes shown in Table 4, we intentionally opted to modify the number of decoder blocks to 3 per iteration. The decision was made to strike a balance between model complexity and efficiency. Including three decoder blocks improved information processing and reduced the risk of overfitting.

5.2 Comparison with other state-of-art methods

Our proposed methodology demonstrated exceptional performance on the TotalText dataset, surpassing TextSnake[27], TextField[28], CRAFT[29], PSENet[30], and CA-STD[31] in terms of efficacy. Our methodology outperformed the competition with a precision of 94.7%, recall of 83.4%, and F1 score of 88.7%. TransText[33], JMNet[32], and CA-STD[31] also showed promising results with precision rates of 90.8%, 90.3%, and 89.3%, respectively, along with commendable recall values, as shown in Table 5.

Assessing text detection techniques on the CTW1500 dataset has provided valuable insights for our academic research. Our proposed approach has undergone rigorous evaluation, resulting in significant achievements with a precision rate of 90.8%, a recall rate of 82.4%, and an exceptional F1 score of 86.4%. These results surpassed alternative methodologies such as CTPN [34], EAST [7], TextSnake [27], PSENet [30], and CA-STD [31], highlighting the effectiveness of our method. CA-STD [31] and PSENet [30] demonstrated commendable performance with precision-recall-F1 scores of 86.4%, 81.2%, 83.7%, 84.8%, 79.7%, and 82.2%, respectively, as shown in Table 6.

Table 6. Comparison with other state-of-arts method in ICDR15 dataset

References	ICDR15		
	P	R	F
CTPN[34]	74.2	51.6	60.9
EAST[7]	83.6	73.5	78.2
PSENet[30]	86.9	84.5	85.7
JMNet[32]	87.4	81.9	84.6
TPLANET[35]	94.2	81.3	87.3
TransText[33]	92.1	88.5	90.3
Ours	**97.5**	**90.2**	**93.7**

Table 6 presents the results of various text detection methodologies applied to the ICDR15 dataset, which are crucial for our scholarly investigation. Through a comprehensive evaluation, our proposed methodology has showcased exceptional efficacy with a precision rate of 97.5%, a recall rate of 90.2%, and an impressive F1 score of 93.7%. Comparative analysis reveals that our approach outperforms other methods such as CTPN [34], EAST [7], PSENet [30], JMNet [32], TPLANET [35], and TransText [33]. TPLANET and TransText demonstrate commendable precision, recall, and F1 scores, with TPLANET [35] achieving 94.2%, 81.3%, and 87.3%, and TransText [33] achieving 92.1%, 88.5%, and 90.3%, respectively.

6. Conclusion

This study focuses on the challenges of detecting text in natural images, particularly in curved text and low-resolution images. The methodology proposed in this study utilizes multiple strategies to enhance text recognition in various scenarios. The rectification module utilizes a spatial transformation network to rectify curved text regions. By utilizing the ResNet50 model, the process of extracting visual features is enhanced, leading to improved recognition performance, specifically in elongated images. A specialized module facilitates the capture of semantic context, while a separate module accounts for contextual dependencies within sequences. The decoding module in transformers consists of masked multi-head attention, a feed-forward network, and layer normalization. This module effectively combines and presents textual information, resulting in improved accuracy and efficiency in text recognition. This approach optimizes text recognition by combining rectification, feature extraction, semantic context, global context, and transformer-based decoding. The efficacy of the proposed method has been experimentally validated on benchmark datasets, namely ICDR15, CTW1500, and TotalText. The results demonstrate its superiority over various state-of-the-art precision, recall, and F1-score methods. The ablation study and method comparison provide additional validation of the efficiency of our approach. Our proposed methodology shows promise as a potential solution for addressing challenging curved text, low-quality images, and recognition speed, particularly in practical applications, particularly on edge devices. The achievements of this study highlight the effectiveness of an integrated approach in addressing complex challenges related to text recognition. Furthermore, these findings provide a basis for future advancements in this field. This study provides insights into the advancement of text recognition technology by addressing the urgent need for efficient and accurate text recognition in various contexts.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

The first author conducted investigations, collected datasets, implemented the study, analyzed the results, and prepared the original draft. The second author provided supervision, completed a work review, and performed validation.

References

[1] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 4, pp. 640–651, 2017.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C.Berg “SSD: Single Shot MultiBox Detector”, In: *Proc of European Conference, Amsterdam*, The Netherlands, pp. 21–37, 2016.

[3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137–1149, Jun. 2017.

[4] K. Manjari, M. Verma, and G. Singal, “A survey on Assistive Technology for visually impaired”, *Internet of Things*, Vol. 11, p. 100188, 2020.

[5] S. Bharat, Y. Jahongir, G. Abdulaziz, and K. T. Hyong, “Development of a Low-cost Industrial OCR System with an End-to-end Deep Learning Technology”, *대한임베디드공학회논문지*,

Vol. 15, No. 2, pp. 51–60, Apr. 2020.

[6] M. B. Revanasiddappa and B. S. Harish, "A New Feature Selection Method based on Intuitionistic Fuzzy Entropy to Categorize Text Documents", *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 5, No. 3, p. 106, 2018.

[7] J. Wernersbach and C. Tracy, "East", *Swim. Holes Texas*, pp. 45–68, 2021.

[8] K. H. Kim, Y. Cheon, S. Hong, B. S. Roh, and M. Park, "PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection", *CoRR*, Vol. abs/1608.0, 2016.

[9] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust Scene Text Recognition with Automatic Rectification", In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4168–4176, 2016.

[10] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 9, pp. 2035–2048, 2019.

[11] F. Zhan and S. Lu, "ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification", In: *Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2054–2063, 2019.

[12] Y. Zhang, S. Nie, W. Liu, X. Xu, D. Zhang, and H. T. Shen, "Sequence-To-Sequence Domain Adaptation Network for Robust Text Image Recognition", In: *Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2735–2744, 2019.

[13] Z. Wan, J. Zhang, L. Zhang, J. Luo, and C. Yao, "On Vocabulary Reliance in Scene Text Recognition", In: *Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11422–1143, 2020.

[14] D. Yu, X. Li, C. Zhang, T. Liu, J. Han, J. Liu, and E. Ding, "Towards Accurate Scene Text Recognition With Semantic Reasoning Networks", In: *Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12110–12119, 2020.

[15] H. Zhang, Q. Yao, M. Yang, Y. Xu, and X. Bai, "AutoSTR: Efficient Backbone Search for Scene Text Recognition", In: *Proc of Computer Vision - ECCV2020*, pp. 751–767, 2020.

[16] Y. Zhu, S. Wang, Z. Huang, and K. Chen, "Text Recognition in Images Based on Transformer with Hierarchical Attention", In: *Proc of IEEE International Conference on Image Processing (ICIP)*, pp. 1945–1949, 2019.

[17] B. Li, X. Tang, X. Qi, Y. Chen, and R. Xiao, "Hamming OCR: A Locality Sensitive Hashing Neural Network for Scene Text Recognition", *CoRR*, Vol. abs/2009.1, 2020.

[18] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention", In: *Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2326–2335, 2020.

[19] A. Vaswani et al., "Attention is All You Need", In: *Proc. of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[20] Y. G. Kim, H. Kim, M. Kang, H. J. Lee, R. Lee, and G. Park, "Analysis of the Novel Transformer Module Combination for Scene Text Recognition", In: *Proc of IEEE International Conference on Image Processing (ICIP)*, pp. 1229–1233, 2021.

[21] L. Ren, H. Zhou, J. Chen, L. Shao, Y. Wu, and H. Zhang, "A Transformer-Based Decoupled Attention Network for Text Recognition in Shopping Receipt Images", *Neural Computing for Advanced Applications*, pp. 563–577, 2021.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", In: *Proc of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks", *Advances in Neural Information Processing Systems*, Vol. 2015, pp. 2017–2025, 2015.

[24] D. Karatzas, L. G. Bigroda, A. Nicolaou, S. Ghosh, A. Bagdanov, and M. Iwamura, "ICDAR 2015 competition on Robust Reading", In: *Proc. of 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1156–1160, doi: 10.1109/ICDAR.2015.7333942, 2015.

[25] Y. Sun, J. Liu, W. Liu, J. Han, E. Ding, and J. Liu, "Chinese Street View Text: Large-Scale Chinese Text Reading With Partially Supervised Learning", In: *Proc of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[26] C. K. Ch'ng and C. S. Chan, "Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition", In: *Proc of 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01, pp. 935–942, 2017.

[27] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A Flexible Representation

for Detecting Text of Arbitrary Shapes", *Computer Vision – ECCV 2018*, pp. 19–35, 2018.

[28] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "TextField: Learning a Deep Direction Field for Irregular Scene Text Detection", *IEEE Transactions on Image Processing*, Vol. 28, No. 11, pp. 5566–5579, Nov. 2019.

[29] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection", In: *Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2019-June, pp. 9357–9366, 2019.

[30] W. Wang E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape Robust Text Detection With Progressive Scale Expansion Network", In: *Proc. of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2019-June, pp. 9328–9337, 2019.

[31] X. Wu, Y. Qi, J. Song, J. Yao, Y. Wang, Y. Liu, Y. Han, and Q. Qian, "CA-STD: Scene Text Detection in Arbitrary Shape Based on Conditional Attention", *Information*, Vol. 13, No. 12, p. 565, 2022.

[32] Z. Lin, Y. Chen, P. Chen, H. Chen, F. Chen, and N. Ling, "JMNET: Arbitrary-shaped scene text detection using multi-space perception", *Neurocomputing*, Vol. 513, pp. 261–272, 2022.

[33] J. Zhu and G. Wang, "TransText: Improving scene text detection via transformer", *Digital Signal Processing*, Vol. 130, p. 103698, 2022.

[34] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting Text in Natural Image with Connectionist Text Proposal Network", In: *Proc of Computer Vision -- ECCV 2016*, pp. 56–72, 2016.

[35] D. Zhong, S. Lyu, P. Shivakumara, U. Pal, and Y. Lu, "Text proposals with location-awareness-attention network for arbitrarily shaped scene text detection and recognition", *Expert System Application*, Vol. 205, No. 2021, 2022.