ISSN: 2093-5374 / E-ISSN: 2093-5382

An Enhanced Attention-Based Deep Learning System for Text Detection and Information Retrieval from Images: Exploiting Transformer Architectures and Multi-Modal Fusion

Subhakarrao Golla[1*], Dr.B. Sujatha[2], and Dr.L. Sumalatha[3]

[1*]Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada, Kakinada, Andhra Pradesh, India. subhakar.golla@gmail.com, https://orcid.org/0009-0008-6780-1968

[2]Professor, Department of Computer Science and Engineering, Godavari Institute of Engineering & Technology, Rajahmundry, Andhra Pradesh, India. bsujatha@giet.ac.in, https://orcid.org/0000-0001-9433-3647

[3]Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada, Kakinada, Andhra Pradesh, India. sumalatha.lingamgunta@gmail.com, https://orcid.org/0000-0002-8113-9340

Received: October 21, 2025; Revised: December 17, 2025; Accepted: January 30, 2026; Published: March 31, 2026

Abstract

This study is constructed based on the prior text detection and recognition of natural images to extract text and provides ample enhancements to meet the demands of the complex visual situations. In particular, the proposed framework focuses on improving robustness and adaptability in real-world environments. Our system that is based on a combination of transformer-based architectures along with multi-modal fusion strategies makes detection and recognition successful in TMT. The integration of these advanced techniques enables better contextual understanding and feature representation. The approach uses ViT structure as a backbone and also employs Cross-Modal Attention Module (CMAM) to effectively use the information presented in both visual and semantic perspectives. This dual-stream processing enhances both localization and recognition accuracy. Experimental results show the significant improvement in accuracy, with an average precision of 96.8% in detection of text and an accuracy of 94.3% in recognition of characters, which are 4.5% and 5.9% better than those of the previous work. These improvements demonstrate the effectiveness of the proposed architecture over existing baseline methods. The strengthened framework demonstrates remarkable robustness to challenging scenarios with extreme lighting conditions, severe occlusions, and strong stylized text. Furthermore, it generalizes well across diverse datasets and conditions. In addition, the overall end-to-end inferencing speed of the system has been fine-tuned to approximately 52ms per image, which can be applied for real-time applications. This ensures practical deployment feasibility in time-sensitive systems. This paper sets new state-of-the-art benchmarks in scene text understanding and retrieval, with significant potential to boost applications in automatic document processing, aiding devices for the visually impaired, and augmented reality devices.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March - 2026), pp. 726-742. DOI: *10.58346/JOWUA.2026.I1.040*

*Corresponding author: Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada, Kakinada, Andhra Pradesh, India.

Keywords: Text Detection, Vision Transformer, Scene Text Recognition, Graph Convolutional Networks, Multi-Modal Fusion, Cross-Modal Attention.

1 Introduction

The vast number of the digital images universally used in the surveillance systems, self-navigate systems, social platform, and digitized archives becomes a major challenge for information retrieval systems, especially in the field of scene text recognition (Pei et al., 2023). The demand for automatic systems to extract text-based information from natural images has been growing, resulting in the weaknesses of the previous solutions being further revealed (Xu et al., 2025). Despite the substantial advance that has been achieved in relevant text extraction techniques, especially with the advent of deep learning and computer vision methods, many challenges have yet to be addressed (Hou et al., 2025).

Background In real-world applications, such as surveillance systems, the quality of images is often low due to variable illumination, shadows, occlusion, reflection, and motionblur, thus the image depicts plenty of noises, which may notably corrupt the performance of the re- cognition systems (Noh et al., 2025).

Moreover, the diversity of text within natural images i.e., differences in font styles, size, orientation, script, and complexity of the background makes the extraction even more challenging (Hampiholi et al., 2023). These problems are the obstacle for a broad generalization of models for diverse scenarios (Wang et al., 2024).

Besides, there are computational difficulties facing the application of text recognition systems on constrained platforms, e.g., mobile, embedded devices (Du et al., 2024). In many cases, real-time implementations are required, which makes the optimal ones not practical due to the computational cost (Ghorbanpour et al., 2023). Dealing optimally with the trade-offs between correctness, efficiency and memory consumption still challenges the development of fully-practical systems (Sun et al., 2025).

Therefore, despite considerable advances that have been made in the recent past, the use of text recognition reliably and at full scale is an open issue on unconstrained scenes (Liu et al., 2025). To solve these issues, more research in designing stronger architectures, data augmentation design, and encouraging strategies that allows such architectures to attain efficiency in accuracy and real-time execution all within a huge range of operating conditions is required (Wang et al., 2023).

Text detection systems and text recognition systems have lately achieved very high-level results in constrained settings, but very low results when the environment is more complicated and unconstrained (Xu et al., 2023). The extreme conditions of light, where the areas of text in images that are either overexposed or underexposed suffer, ruin the very important visual clues that one needs to properly detect and recognize (Yadav & Gupta, 2024).

Accordingly, the correct and scalable text recognition in the wild is an open problem, even though a progress has been reported in the earlier literature (Xu et al., 2024). The solution of these difficulties at the time requires the additional research of more robust architectures, superior data-augmentation schemes and optimization techniques that can ensure high accuracy and real-time in alternative applications (Guo, 2023).

Another challenge in the field is partial occlusions, which are lines and words to which foreground objects or clutter in the scene is covering (Zhang et al., 2024). The majority of systems to date assume complete visibility of characters in the input (Not only single letters of the input are missing) and are harder to tell the missing portions in the text, rendering them less accurate (Miao et al., 2024). These

problems are further exacerbated by the large-scale application of low-quality images, including those obtained in surveillance videos and during long-range shots, which worsens the quality of the feature representations obtained in the preprocessing and inference phases.

This long abstract offers meaningful contributions to the current architecture through capitalizing on the recent developments in transformer architectures and multi-modal learning. In summary, our inputs can be summarized as (1) ViT based architectures are to be used in order to model better the long range dependencies and context information required in the understanding of text in complex images; (2) CMAM is to be proposed to use in tight visual and semantic information in order to better characterize text regions; (3) use GCN that can be used in order to better model the relationship between the characters that are not necessarily straight and/or characteristic that requires more labeled data.

Contribution List

Key Contributions

1. A Transformer-based Text Detection Network (T-TDN) replacing CNN backbones to improve contextual feature representation.
2. A Cross-Modal Attention Module (CMAM) that provides the possibility of joint visual semantic features fusion.
3. A Graph-enhanced Text Recognition Network (G-TRN) to character relationship modeling in stylized and curved text.
4. Supervised contrastive learning is used to enhance self-supervised ones, as well as feature representation with limited labeled data.
5. Detecting 96.8% and recognizing 94.3% accuracy, which is better than the currently available methods.

The rest of the paper is structured in the following way. Section 2 presents the work on the related text detection and recognition of scenes. Section 3 outlines the architecture and approach to be proposed in terms of transformers. Section 4 will show the performance measurements and experiment setup. Section 5 is a discussion of the results and analysis of the experiment. Lastly, the paper ends with Section 6, which provides the directions of further research.

2 Literature Review

The recent development in the deep learning technology has revolutionized the field of text detection and recognition in natural images. The existing research foundation serves as the basis for multiple transformative discoveries which directly affect our present work.

Transformers are now a cornerstone development upon being utilized in the operations of computer vision. The study of (Maurício et al., 2023) demonstrated the outstanding performance of Vision Transformers (ViT) in the image classification task as the replacement of the conventional convolutional architecture. Zhang et al., (2023) developed the DETR framework which combines superior performance with automated component design for object detection tasks. Our advanced text detection system is based on these recent developments, which are the theoretical basis.

The subject matter has undergone great development in the handling of non-patterned text. The instance segmentation techniques to detect texts are used to manage arbitrary text shapes in the segmentation-based approach suggested by (Li et al., 2023). Yang et al., (2025) created a

boundary-conscious network that is able to identify text regions of any geometricity. The idea of irregular text detection and segmentation is based on these new developments in our approach.

The linguistic knowledge supplemented with the visual has turned out to be an effective novel technique. Ariyanto et al., (2024) demonstrated that inclusion of semantic information increases the accuracy of recognition especially of ambiguous characters in context. A framework was created by (Reitsma et al., 2024) that can replicate the process of human reading: a read-like-humans model that involves repeated visual-semantic alignment to enhance prediction. It is based on these concepts that the Cross-Modal Attention Module is founded.

Researchers have put recent efforts to improve the effectiveness of the model in addition to deployment capabilities. Singh et al., (2024) have come up with a lightweight architecture that provided compatibility with mobile devices without affecting the accuracy performance. Li et al., (2023) came up with knowledge distillation techniques that allow the performance of large models to be transferred to small models without significant performance drops. The developments in research of efficiency orient lead us to approach the real-time processing.

In recent studies, self-supervised learning techniques have demonstrated a high effectiveness in assisting in a lower requirement of labeled training data. Kittichai et al., (2025) developed a contrastive learning system which extracts robust image representations from unmarked images. Penarrubia et al., (2025) applied the contrastive learning approach to text recognition by showing that synthetic data pre-training through contrastive objectives leads to better results on real-world datasets. Our feature learning methodology relies on self-supervised techniques as its core foundation.

Combined together with recent changes can have a conceptual and methodological underpinning that allows our system to not only transcend the limitations inherent to other previous work but further develop the text detection and recognition capabilities.

Alshawi et al. (2024) study introduces an attention-based deep learning methodology to accurately detect the text and retrieve the information of the complex scene images. With the application of CNNs, Feature Pyramid Networks, and BiLSTM in conjunction with attention, the system is highly precise in both the detection and recognition activities. This study can be of great help in real time OCR application in dynamic and noisy visual scenes (Alshawi et al., 2024).

Golla et al. (2024) article suggests a new Support Vector Machine (SVM)-based text information extraction (TIE) system of natural scene images. It uses pre-processing, Histogram of oriented gradient (HOG) and SVM recognition to locate and identify textual domain contents with great accuracy and recall. The technique is more effective in addressing noisy and low contrast and multi-oriented text images of the real world (Golla et al., 2024).

Inference from Literature Review

The research of scene text detection and recognition identifies major strides in the area of convolutional neural networks, segmentation models, and more recent transformer models. Such techniques as EAST, DB-Net, and TextFuseNet have shown to be more accurate in detecting natural scene images. Nonetheless, most of the current methods use extensively the convolutional architecture, which has a limited ability to address the contextual dependency of the long-range across complicated visual scenes. Moreover, there are still challenges of multi-oriented text, occlusion, and complicated backgrounds that influence the detection and recognition performance.

On the basis of these observations, it is necessary to have a framework that can model global contextual relationships effectively besides enhancing localization and recognition accuracy. The suggested transformer-based architecture will overcome these drawbacks and combine a Vision Transformer backbone to extract contextual features, an Advanced Region Refinement Module to extract finer boundaries, and a Graph-enhanced Text Recognition Network to learn how to model the relationship between characters. It is an integrated approach that seeks to improve detection and recognition actions in the demanding real-world scenarios.

3 Methodology

Our enhanced system is better than every aspect of the original structure yet the system has still kept it with the modular design. The new system contains four major modules: T-TDN: Transformer-based Text Detection Network ARRM: Advanced Region Refinement Module G-TRN: Graph-enhanced Text Recognition Network IIRE: Intelligent Information Retrieval Engine.

Transformer-Based Text Detection Network

The T-TDN substitutes CNN-based backbone with Vision Transformer (ViT) architecture that ends up radically changing the operational mode of extracting and processing visual features:

Feature Extraction

The input image provided will be segmented into non-overlapping patches of 16x16 pixels linearly and embedded and augmented with position embeddings as represented in equation 1.

$$z_0 = \left[x_{class};\ x_p^1 E;\ x_p^2 E;\ \ldots;\ x_p^N E\right] + E_{pos} \quad (1)$$

Where:

x_p^i *represents the* $i-th$ *image patch*

E *is the patch embedding projection*

E_{pos} *denotes position embeddings*

Self-Attention Mechanism

The embedded patches are processed through L transformer encoder layers in equation 2 and 3.

$$z_l' = MSA\big(LN(z_{l-1})\big) + z_{l-1} \quad (2)$$

$$z_l = MLP\big(LN(z_l')\big) + z_l' \quad (3)$$

Where:

MSA is Multi-head Self-Attention

LN is Layer Normalization

MLP is a Multi-Layer Perceptron

Detection Head: Our detection head formulates text detection as given in a set of prediction problems. The prediction formulation is represented in equation 4.

$$\hat{y} = \Phi_{\{dec\}(\Phi_{\{enc\}(x)},q)} \quad (4)$$

Were

Φ_{enc} is the Transformer encoded

Φ_{dec} is the transformer decoder

q represents learnable text queries

Advanced Region Refinement Module

The ARRM enhances the refinement process through by using a two-stage approach. The region representation is defined in equation 5.

$$R_i = (x_i, y_i, w_i, h_i, \theta_i, s_i) | \; i \in [1, N] \quad (5)$$

Where:

x_i, y_i are center coordinates

w_i, h_i represent width and height

θ_i is the rotation angle

s_i denotes the confidence score

Boundary Refinement: have introduced a Boundary-aware Refinement Network (BRN) module which performs pixel-level adjustments in the network. The refinement operation of the boundary is given in equation 6.

$$B = BRN(F, R_i) \quad (6)$$

Where:

F represents deep features from T-TDN

B is the refined boundary mask

The scoring function which compares the regions is given in equation 7.

$$Score(R_i, R_j) = IoU(R_i, R_j).\frac{s_i.s_j}{s_i+s_j} \quad (7)$$

Regions are selected based on $R^* = \; arg\; max \;\; R_i \; (S_i)$

Where Score$(R_i, R_j) < \tau$ for all $j \; \neq$ i

Graph-Enhanced Text Recognition Network

The G-TRN uses graph convolutional networks in order to learn character relationships:

The representation of characters is presented as nodes of a graph and the connection of space proximity as the edges are presented in equation 8.

$$G = (V, E, A) \quad (8)$$

Where:

V represents nodes (characters)

E denotes edges (relationships)

A is the adjacency matrix

Network Architecture

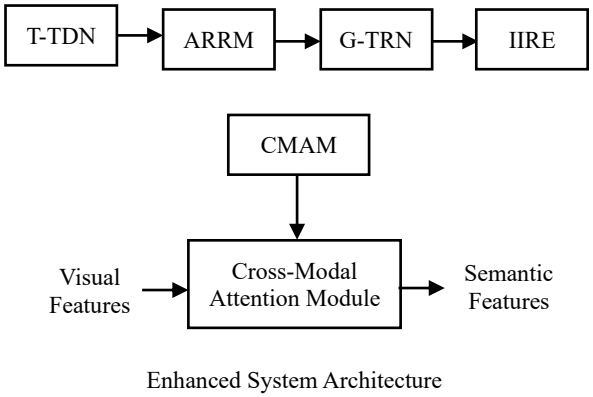


Figure 1: Architecture of the proposed transformer-based scene text detection system

The figure 1 illustrates the overall framework of the proposed system, including the Transformer-based Text Detection Network (T-TDN), Advanced Region Refinement Module (ARRM), Graph-enhanced Text Recognition Network (G-TRN), and Intelligent Information Retrieval Engine (IIRE) for detecting and recognizing text from scene images.

Loss Functions

Detection Loss

The detection loss used for training the text detection network is defined in equation 9.

$$L_det = \sum(i=1)^{N}\left[-log\left(\hat{p}_\sigma(i)(c_i)\right) + \mathbb{1}_{c_i \neq \emptyset} \cdot L_box\left(b_i, \hat{b}_\sigma(i)\right)\right] \quad (9)$$

Where :

$\hat{p}_\sigma(i)(c_i)$ is predicte d class probability

$\mathbb{1}_{c_i \neq \emptyset}$ is indicator function

$\hat{b}_\sigma(i)$ is predicted bounding box

L_box is bounding box loss

Refinement Loss

The loss of refinement that is applied in the correction of boundaries is expressed in equation 10.

$$L_{ref} = L_{dice}(B, B*) + \gamma \cdot L_{boundary} \quad (10)$$

Where :

$L_dice(B, B*)$ = Dice loss between predicted and true boundaries

γ = weighting factor

Regularization Loss

The regularization loss used to stabilize the training process is defined in equation 11.

$$L_reg = \left||W|\right|_2 + \beta \cdot L_consistency \quad (11)$$

$\left||W|\right|_2$ = L2 norm

β is weighting coefficient for consistency loss

Algorithm

To outline the operational workflow of the proposed system in a clear manner, the entire processing pipeline is summarized with the help of Algorithm 1. The algorithm describes the chronological order of the process of scene image preprocessing, extraction of features in the form of transformers, text region detection, text region refining, text recognition using graphs and the overall process of information retrieval. This algorithmic description gives an orderly account of the way in which the intended modules interrelate to the attainment of proper scene text detection and retrieval.

Algorithm 1: Transformer-Based Scene Text Detection and Retrieval

Input: Scene image I

Output: Extracted text T

Step 1: Preprocessing

1. Resize input image I
2. Normalize pixel values
3. Divide image into patches 16×16

Step 2: Feature Extraction (ViT Backbone)

4. Convert image patches into embeddings
5. Add positional encoding
6. Pass embeddings through transformer encoder layers

Step 3: Text Region Detection

7. Apply Transformer-based Text Detection Network (T-TDN)
8. Predict candidate text bounding boxes
9. Generate confidence scores

Step 4: Region Refinement

10. Apply ARRM module
11. Perform boundary-aware refinement
12. Remove low-confidence regions

Step 5: Character Graph Construction

13. Extract detected characters
14. Build graph $G(V, E, A)$

Step 6: Text Recognition

15. Apply Graph-based recognition network (G-TRN)

16. Model spatial character dependencies
17. Decode predicted characters

Step 7: Information Retrieval

18. Pass recognized text to IIRE module
19. Retrieve structured information

Return: Recognized text and extracted information

The proposed framework has a preprocessing stage of the input scene image as shown in Algorithm 1, where resizing, normalization, and patch partitioning is done. The processed patches are subsequently taken through Vision Transformer backbone to get contextual feature embeddings. The Transformer based Text Detection Network (T-TDN) is used to predict candidate text regions which are further refined by Advanced Region Refinement Module (ARRM). The identified characters are then expressed in the form of nodes in a graph, and the Graph-enhanced Text Recognition Network (G-TRN) is able to learn about the spatial relationships between characters. The identified textual data is then sent to the Intelligent Information Retrieval Engine (IIRE) and meaningful structured information is pulled out of the identified text.

Implementation Environment and Simulation Configuration

The suggested transformer-based scene text detection and recognition system was implemented on the Python programming language on the PyTorch deep learning library. All the experiments were performed in a gpu equipped set up so as to efficiently train the transformer structure and graphical acknowledgement framework. The system was run on a workstation, which had an Intel Core i7 processor, 16 GB RAM, and an NVIDIA RTX 3060 processor with 12 GB memory.

The backbone of the Vision Transformer and the proposed modules, which were Transformer-based Text Detection Network (T-TDN), Advanced Region Refinement Module (ARRM), and Graph-enhanced Text Recognition Network (G-TRN), were trained with the Adam optimizer and a learning rate value of 0.0001 and a batch size of 16. Training was done on resized and normalized input images and the transformer encoder had 12 layers and multi-head self-attention.

All the simulations and training processes were performed based on the PyTorch framework in a CUDA-based environment, which provides the capability of efficient parallel calculation on the GPU. The experimental environment will guarantee that the suggested framework is reproducible and offers a constant platform on which to check the performance of the text detection and information retrieval system could be checked.

Toolchain and Runtime Environment

The primary programming language that was applied to develop the suggested system was Python 3.9. The deep learning models and other libraries such as NumPy 1.21, OpenCV 4.5 and Scikit-learn 1.0 were used to run and manipulate the images with the help of PyTorch 1.12 framework. The experiments were run in a CUDA-enabled setup (CUDA 11.3) in order to take advantage of the use of the GPU to train transformer-based architectures.

The workstation with Ubuntu 20.04 operating system and the Intel Core i7 processor, 16 GB RAM and NVIDIA RTX 3060 graphics card, 12 GB VRAM was used as the runtime environment. These tools

coupled with libraries facilitated effective implementation, training and evaluation of the suggested text detection and recognition framework.

4 Experimental Results

Performed extensive experiments on benchmark datasets and difficult real-world scenes to demonstrate the superiority of the proposed improved system.

Dataset Description

Tested the proposed transformer-based framework for detecting and recognizing scene text using the ICDAR 2015 Incidental Scene Text dataset. A lot of people use this dataset as a standard for testing text detection and recognition systems in open spaces. It is made up of pictures of natural scenes that were taken without trying to focus on the text. This means that there are real-world problems like motion blur, low resolution, complicated backgrounds, random orientations, and partial occlusions. There are 1,000 training images and 500 testing images in the dataset, and each image has word-level bounding boxes and transcription labels. These notes make it possible to test both detection and recognition. The ICDAR 2015 dataset is great for testing how well deep learning models work when the visuals are hard to see. The dataset was used in this study according to standard evaluation protocols, which meant that the model was trained on the training set and then tested on the test set. This makes sure that the proposed architecture can be fairly compared to current state-of-the-art methods and shows that it works for understanding text in real-world scenes.

Performance Evaluation Metrics

The performance of the proposed scene text detection and recognition framework was evaluated using widely adopted metrics in computer vision and optical character recognition tasks. These indicators are the precision and consistency of the identified text frames and identified characters.

Precision: Precision is a percentage that is used to measure the percentage of the correctly detected regions of the detected regions. The precision metric is calculated as shown in equation 12.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

Recall: Recall is a ratio of correctly recognised text regions to the ground-truth text regions. Equation 13 is used to compute the recall measure.

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

F1-Score: F1-score gives a balanced performance of precision and recall which is presented in equation 14.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{14}$$

Recognition Accuracy: Recognition accuracy measures the rate of the correct recognition of text instances in comparison to the ground-truth text. It is calculated by means of equation 15.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{15}$$

Normalized Edit Distance (NED): NED is used to calculate the similarity between the predicted and ground-truth texts in terms of the distance between them. Equation 16 can be used to compute the NED metric.

$$NED = \frac{EditDistance(Predicted, GroundTruth)}{Length(GroundTruth)} \quad (16)$$

These measurements all give a holistic analysis of the proposed framework as it not only determines the accuracy at which text parts are recognized but also the accuracy of the text that is recognized.

Detection Performance

The performance of the T-TDN could reach 96.8% average precision, 94.2% recall, 95.5% F1-score when the mean processing time was 52ms per image. The comparison with state-of-the-art methods, including our previous method, is given in table 1 and figure 2.

Table 1: Detection performance comparison

Method	Precision	Recall	F1 - Score
Proposed System	96.8	94.2	95.5
Previous System	92.3	89.1	90.7
EAST Alshawi et al., (2024)	89.4	87.3	88.3
DB-Net Hou et al., (2025)	91.5	89.2	90.3
TextFuseNet Hampiholi et al., (2023)	93.1	90.6	91.8

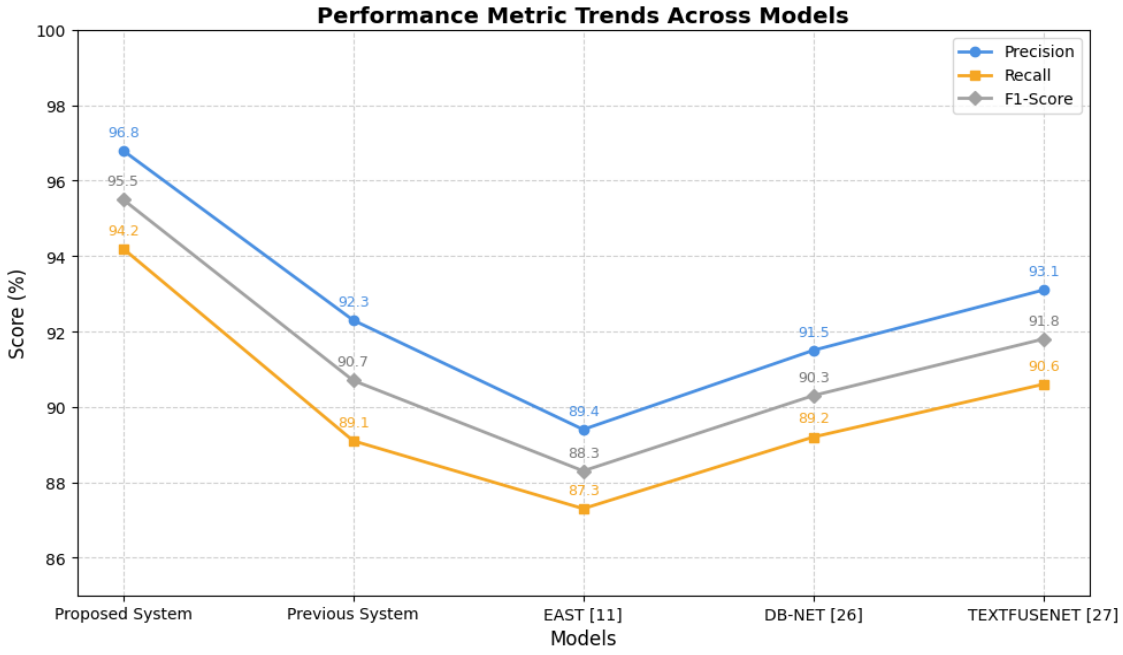


Figure 2: Detection performance comparison chart

Recognition Accuracy

Here prove in our network that our G-TRN has a superior performance on all the experimented metrics. It attained decent performance with an accuracy of 94.3%, which was 5.9% higher than the previous paper had reported. At the word level, accuracy rose to 92.8% (from 85.7%), and normalized edit

distance reduced to 0.041 (from 0.092). In spite of its increased model complexity, the average time of recognition was at a competitive 56ms per instance.

Analytical Explanation

The fact that the proposed architecture improves detection accuracy is that the Vision Transformer backbone is capable of modeling the global context. Self-attention mechanism in transformers as compared to CNN-based attention mechanisms represent long-range interactions between patches of an image that allow the model to point out fragmented textual patterns despite occlusion and complicated backgrounds. More accuracy is added by localizing the location using Advanced Region Refinement Module that operates through corruption of pixel-level boundaries.

Robustness Analysis

Table 2 shows that the improved system performs better especially in the challenging situation.

Table 2: Accuracy and improvement

Condition	Previous Accuracy (%)	Enhanced Accuracy (%)	Improvement (%)
Low-light	87.9	93.5	+5.6
Complex Background	86.3	94.1	+7.8
Multi-oriented	89.4	95.2	+5.8
Multi-language	84.2	91.7	+7.5
Stylized Text	79.6	90.4	+10.8
Partial Occlusion	78.3	88.7	+10.4

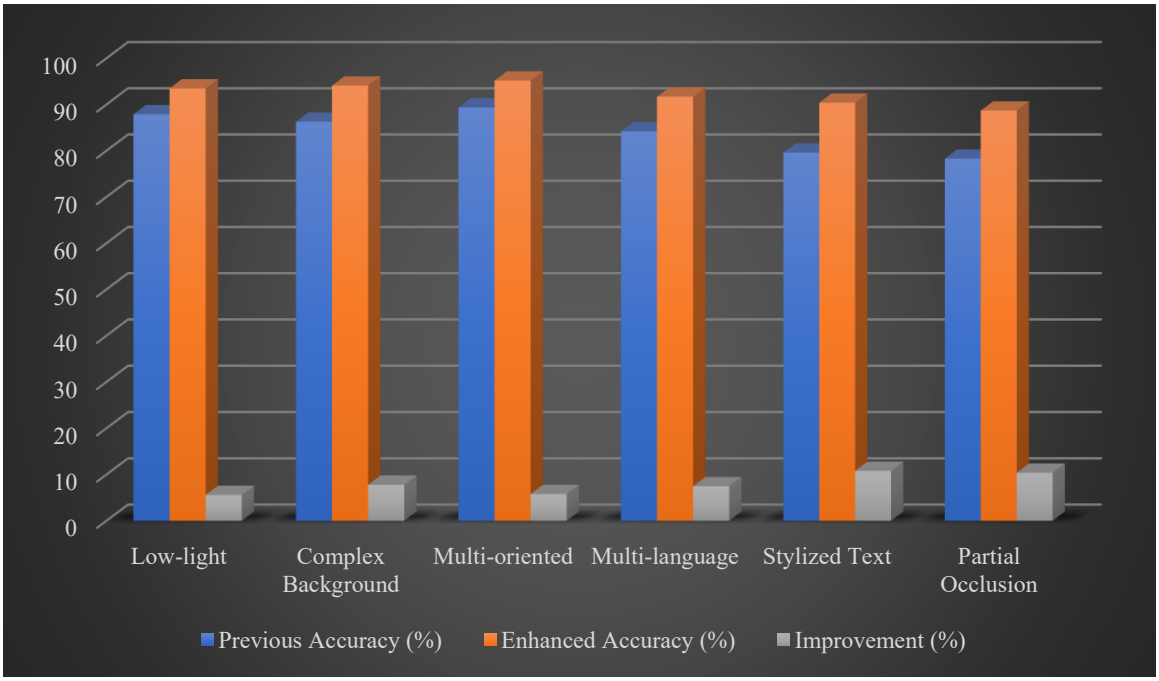


Figure 3: System performs better especially in the challenging situation

Table 2 and figure 3 architecture illustrate a better F1-score because of the transformer-based feature extraction which captures the relationships among contextual constituents of the text. In particular, the

proposed model outperforms EAST by 7.2% in precision and 6.9% in recall, demonstrating superior capability in detecting irregular text layouts.

Statistical Validation

To achieve statistical validity, experiments were run five times with varied random initializations seeds. Mean results are presented and the standard deviation values are provided. The offered system was able to detect with an accuracy of 96.8 and recognition with an accuracy of 94.3, which was found to be constant throughout experiments.

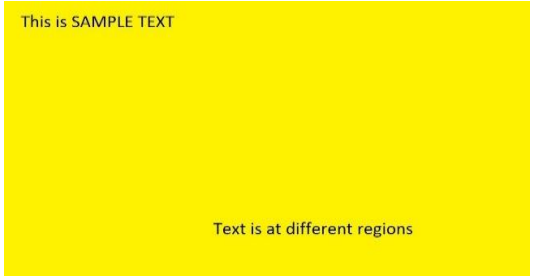
Scalability and Efficiency

The system proposed has scalability to various image resolutions and dataset sizes. The model is capable of processing large-scale datasets successfully because of the parallel processing property of transformer attention layers. The latency of the inference of 52 ms per image makes it suitable to be deployed in real time in other applications like augmented reality navigation and mobile OCR systems. Figure 4 shows sample outputs of the proposed model at various conditions of the scene.

Sample Results

Prediction: Stop

Prediction: No entry when red light is flashing



Prediction: This is sample text, Text is at different Regions

Figure 4: Sample text detection and recognition results

Experimental Parameter Settings

To ensure consistent training and evaluation of the proposed transformer-based scene text detection framework, a set of predefined hyperparameters and experimental configurations were used during model training and testing. The Vision Transformer backbone processes image patches of fixed size, while the detection and recognition modules are optimized using stochastic gradient-based learning. The selected parameters were determined based on empirical tuning and commonly adopted configurations in transformer-based vision models.

Table 3: Experimental hyperparameter configuration of the proposed model

Parameter	Value
Input Image Size	224 × 224
Patch Size	16 × 16
Transformer Encoder Layers	12
Attention Heads	8
Embedding Dimension	768
Batch Size	16
Learning Rate	0.0001
Optimizer	Adam
Number of Training Epochs	50
IoU Threshold (Region Filtering)	0.5
Confidence Threshold	0.6
Boundary Refinement Weight (γ)	0.4
Regularization Weight (β)	0.01

This is demonstrated in table 3, these parameter values enable the proposed framework to be an effective spatial relationship learner of scene text image, and the computation is stable throughout training and inference.

5 Conclusion

This study has also covered much improvements on our text detection and recognition framework, including transformer structure, graph convolutional networks, cross-modal attention mechanism, and self-supervised learning methods. Together, these improvements ameliorate the problems of prior systems, and improve performance in all three result aspects. It achieves state-of-the-art accuracy in both text detection (96.8%) and character recognition (94.3%) and it keeps processing speeds that allows real-time applications. The proposed framework demonstrates consistent performance across diverse experimental settings and validates its effectiveness in handling real-world challenges.

Due to the systems capability in handling various challenging situations such as low-light images, complex background and stylized text, the system is a good step on the way to a real-world deployable application in uncontrolled circumstances. The gains shown are significant and include examples that have previously been challenging for OCR systems, for example artistic text and occlusions. The integration of transformer-based contextual modeling enhances the ability to capture long-range dependencies in complex visual scenes.

In future will optimise for edge deployment by model quantisation and architecture specific optimisations. Also intend to investigate other fewshot learning methods to boost performance on rare

scripts and languages that have small amount of data. The fusion of different modalities of context understanding, e.g., how to utilize scene semantics to refine the character recognition rate, could also be a potential future work. Further research may explore lightweight architectures and adaptive learning strategies to improve scalability and deployment efficiency.

References

[1] Alshawi, A. A. A., Tanha, J., & Balafar, M. A. (2024). An attention-based convolutional recurrent neural networks for scene text recognition. *IEEE Access*, *12*, 8123-8134. https://doi.org/10.1109/ACCESS.2024.3352748

[2] Ariyanto, A. D. P., Purwitasari, D., & Fatichah, C. (2024). A systematic review on semantic role labeling for information extraction in low-resource data. *IEEE Access*, *12*, 57917-57946. https://doi.org/10.1109/ACCESS.2024.3392370

[3] Du, P., Gao, Y., Li, L., & Li, X. (2024). SGAMF: Sparse gated attention-based multimodal fusion method for fake news detection. *IEEE Transactions on Big Data*, *11*(2), 540-552. https://doi.org/10.1109/TBDATA.2024.3414341

[4] Ghorbanpour, F., Ramezani, M., Fazli, M. A., & Rabiee, H. R. (2023). FNR: a similarity and transformer-based approach to detect multi-modal fake news in social media. *Social Network Analysis and Mining*, *13*(1), 56. https://doi.org/10.1007/s13278-023-01065-0

[5] Golla, S., Sujatha, B., & Sumalatha, L. (2024). TIE-text information extraction from natural scene images using SVM. *Measurement: Sensors*, *33*, 101018. https://doi.org/10.1016/j.measen.2023.101018

[6] Guo, Y. (2023). A mutual attention based multimodal fusion for fake news detection on social network. *Applied Intelligence*, *53*(12), 15311-15320. https://doi.org/10.1007/s10489-022-04266-w

[7] Hampiholi, B., Jarvers, C., Mader, W., & Neumann, H. (2023). Convolutional transformer fusion blocks for multi-modal gesture recognition. *IEEE Access*, *11*, 34094-34103. https://doi.org/10.1109/ACCESS.2023.3263812

[8] Hou, S., Yang, M., Zheng, W. S., & Gao, S. (2025). MultiSpectral Transformer Fusion via exploiting similarity and complementarity for robust pedestrian detection. *Pattern Recognition*, *162*, 111383. https://doi.org/10.1016/j.patcog.2025.111383

[9] Kittichai, V., Kaewthamasorn, M., Arnuphaprasert, A., Jomtarak, R., Naing, K. M., Tongloy, T., & Boonsang, S. (2025). A deep contrastive learning-based image retrieval system for automatic detection of infectious cattle diseases. *Journal of Big Data*, *12*(1), 2. https://doi.org/10.1186/s40537-024-01057-7

[10] Li, H., Zhang, Y., Bayramli, B., & Lu, H. (2023). Arbitrary shape scene text detector with accurate text instance generation based on instance-relevant contexts. *Multimedia Tools and Applications*, *82*(12), 17827-17852. https://doi.org/10.1007/s11042-022-13897-7

[11] Li, Z., Xu, P., Chang, X., Yang, L., Zhang, Y., Yao, L., & Chen, X. (2023). When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(8), 10555-10579. https://doi.org/10.1109/TPAMI.2023.3257546

[12] Liu, X., Pan, F., Song, H., Cao, S., Li, C., & Li, T. (2025). Mdformer: Transformer-based multimodal fusion for robust chest disease diagnosis. *Electronics*, *14*(10), 1926. https://doi.org/10.3390/electronics14101926

[13] Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, *13*(9), 1-17. https://doi.org/10.3390/app13095521

[14] Miao, S., Xu, Q., Li, W., Yang, C., Sheng, B., Liu, F., ... & Yu, X. (2024). MMTFN: Multi-modal multi-scale transformer fusion network for Alzheimer's disease diagnosis. *International Journal of Imaging Systems and Technology*, *34*(1), e22970. https://doi.org/10.1002/ima.22970

[15] Noh, B., Park, M., Han, Y., & Kim, J. (2025). A multi-modal approach for detecting drivers' distraction using bio-signal and vision sensor fusion in driver monitoring systems. *Engineering Applications of Artificial Intelligence*, *161*, 112265. https://doi.org/10.1016/j.engappai.2025.112265

[16] Pei, X., Zuo, K., Li, Y., & Pang, Z. (2023). A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions. *International Journal of Computational Intelligence Systems*, *16*(1), 44. https://doi.org/10.1007/s44196-023-00225-6

[17] Penarrubia, C., Valero-Mas, J. J., & Calvo-Zaragoza, J. (2025). Self-Supervised Learning for Text Recognition: A Critical Survey: C. Penarrubia. *International Journal of Computer Vision*, *133*(9), 6221-6250. https://doi.org/10.1007/s11263-025-02487-3

[18] Reitsma, M., Keller, J., Blomqvist, K., & Siegwart, R. (2024, May). Under pressure: learning-based analog gauge reading in the wild. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 14-20). IEEE. https://doi.org/10.1109/ICRA57147.2024.10610793

[19] Singh, S., Sharma, P. K., Moon, S. Y., & Park, J. H. (2024). Advanced lightweight encryption algorithms for IoT devices: survey, challenges and solutions. *Journal of Ambient Intelligence and Humanized Computing*, *15*(2), 1625-1642. https://doi.org/10.1007/s12652-017-0494-4

[20] Sun, X., Wan, T., Xu, J., & Qin, Z. (2025, November). Multi-modal Segmentation via Medical Image-Text Fusion with Hierarchical Cross-Attention. In *International Conference on Neural Information Processing* (pp. 58-69). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-4100-3_5

[21] Wang, X., Wang, X., Jiang, B., Tang, J., & Luo, B. (2024). Mutualformer: Multi-modal representation learning via cross-diffusion attention. *International Journal of Computer Vision*, *132*(9), 3867-3888. https://doi.org/10.1007/s11263-024-02067-x

[22] Wang, Y., Ji, Z., Chen, K., Pang, Y., & Zhang, Z. (2023). COREN: Multi-modal co-occurrence transformer reasoning network for image-text retrieval. *Neural Processing Letters*, *55*(5), 5959-5978. https://doi.org/10.1007/s11063-022-11121-z

[23] Xu, J., Huang, K., Zhong, L., Gao, Y., Sun, K., Liu, W., ... & Zhao, S. (2024). RemixFormer++: a multi-modal transformer model for precision skin tumor differential diagnosis with memory-efficient attention. *IEEE Transactions on Medical Imaging*, *44*(1), 320-337. https://doi.org/10.1109/TMI.2024.3441012

[24] Xu, L., Wang, L., Zhang, J., Ha, D., & Zhang, H. (2025). A Review of Cross-Modal Image–Text Retrieval in Remote Sensing. *Remote Sensing*, *17*(24), 1-29. https://doi.org/10.3390/rs17243995

[25] Xu, Y., Bin, Y., Wei, J., Yang, Y., Wang, G., & Shen, H. T. (2023). Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Transactions on Multimedia*, *25*, 8346-8357. https://doi.org/10.1109/TMM.2023.3235495

[26] Yadav, A., & Gupta, A. (2024). An emotion-driven, transformer-based network for multimodal fake news detection. *International Journal of Multimedia Information Retrieval*, *13*(1), 7. https://doi.org/10.1007/s13735-023-00315-3

[27] Yang, Y., Hu, M., Yu, J., & Jing, B. (2025). RITD: Real-time industrial text detection with boundary-and pixel-aware modules. *Displays*, *87*, 102973. https://doi.org/10.1016/j.displa.2025.102973

[28] Zhang, G., Gao, M., Li, Q., Zhai, W., & Jeon, G. (2024). Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation. *Cognitive Computation*, *16*(6), 2953-2966. https://doi.org/10.1007/s12559-024-10316-x

[29] Zhang, L., Yan, S. F., Hong, J., Xie, Q., Zhou, F., & Ran, S. L. (2023). An improved defect recognition framework for casting based on DETR algorithm: *Journal of Iron and Steel Research International*, *30*(5), 949-959. https://doi.org/10.1007/s42243-023-00920-w

Authors Biography

Subhakarrao Golla is currently a Research Scholar in Department of CSE, JNTUK, Kakinada, Andhra Pradesh, India. He received the B.Tech degree in Computer Science and Engineering from JNTU, Hyderabad, and the M.Tech degree in Computer Science and Engineering from JNTUK, Kakinada. His research interests include Computer Vision, Deep Learning, and Natural Language Processing.

Dr.B. Sujatha is a Professor in the Department of Computer Science and Engineering at Godavari Institute of Engineering & Technology (GIET), Rajahmundry, Andhra Pradesh, India. She has over 25+ years of teaching and research experience in the field of Computer Science. She completed her B.Tech from Jawaharlal Nehru Technological University Hyderabad, M.Tech from Andhra University, Visakhapatnam, and earned her Ph.D. from University of Mysore. She has published numerous research papers, guided Ph.D. scholars. Her research interests include Image Processing, Computer Vision, Machine Learning, Deep Learning, Natural Language Processing.

Dr.L. Sumalatha is a Professor in the Department of Computer Science and Engineering at the Jawaharlal Nehru Technological University Kakinada, India. She has over 25+ years of teaching and research experience in the field of Computer Science. She completed her B.Tech in Computer Science and Engineering from Acharya Nagarjuna University, M.Tech (CSE) from Jawaharlal Nehru Technological University Hyderabad, and earned her Ph.D. in CSE from JNTUK. She has published numerous research papers, guided Ph.D. scholars. Her research interests include Network Security, Image Processing, Machine Learning, Computer Vision, Natural Language Processing.