

Original Research

Received 27 May 2026

Revised 11 July 2026

Accepted 23 July 2026

Natural Resources for Human Health 2026, Vol. 6 (13s): 1282–1300

KEYWORDS:

Scene Text Recognition; Large Language Model; Radial Spline Image; Transformation; Multi-scale Dilate; 2D Positional Encoding; Differentiable Binary;

Agentic AI Based Multiscale Differentiable Vision Transformer on LLM For Language Independent Scene Text Recognition

Vimuktha Evangeleen Salis¹, Md Shohel Sayeed^{*2}, Andrews Samraj³, Vineetha Edwina Jathanna⁴, Sharmila Chidaravalli⁵

¹Postdoc Fellow, Multimedia University, Malaysia, Global Academy of Technology, India.

^{*2}Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia.

³Department of Computer Science and Engineering, KGiSL Institute of Technology, India.

⁴Maniplal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

⁵Department of Information Science & Engineering, Global Academy of Technology, India.

ABSTRACT: Accurate text recognition from images is pivotal for several tasks in real-life scenarios and scene understanding like, autonomous driving, assistive mechanism for visually impaired, image-based translation and so on. Nevertheless, text recognition in natural scenes is challenged owing to the presence of noise, uneven text font distribution and deterioration of image quality under complicated acquisition circumstances. These features along with poor image quality, different text formats and complicated image backgrounds critically influence the text recognition accuracy, and these issues still requires immediate action in the field. To address these issues, an AI agent-based language dependent scene text recognition (STR) method named the Multi-scale Dilate Differentiable Binary Vision Transformer (MDDDB-ViT) supported on Large Language Model (LLM) is proposed. Initially, Radial Basis Thin Plate Spline Image Transformation-based pre-processing is employed to align original input image into a common coordinate system for recognition. Based on an encoder–decoder model, the proposed MDDDB-ViT method introduces a Multi-scale Dilate 2D Positional module in the encoder that by integrating agent with vision transformer boost multi-scale feature representation. In the decoder, a Differentiable Binary Position Visual Reinforced Large Language Model is designed to enable effective semantic correlation between encoder and decoder. The efficiency of the proposed MDDDB-ViT method is validated on ICDAR 2019 MLT dataset. Experimental results demonstrate that MDDDB-ViT method outperforms state-of-the-art methods across numerous performance metrics, therefore achieving a good balance between scene text recognition accuracy, training time and peak signal to noise ratio (PSNR).

1. INTRODUCTION

Scene text recognition makes it possible to stay connected to both the natural world and environmental issues. Advanced vision technology is necessary for improving communication, allowing users to interpret visual scenes in a manner that aligns with their own viewpoint. The main reason that computer vision has gotten so much better is because of the use of deep learning (DL) techniques. Vision Transformers (ViTs) have lately become a good alternative to traditional neural networks. ViT-based methods have shown competitive performance in a range of benchmark scene interpretation tasks when compared to

convolutional neural networks (CNNs) because of an attention mechanism. The Improved Multi-task Transformer based on Localization Mechanism Network (IMTLM-Net) was introduced in [1] as a new way to recognize English text. The IMTLM-Net approach has two parts: encoding and decoding. Encoding set up a dual-stream processing system for text and visuals, as well as a Permutation Language Model (PLM) to help organize words.

During the decoding process, two Multiple Head Attention (MHA) units were used, with a focus on visual and textual sequences. But the chance of mistakenly predicting a bad outcome or a good outcome makes false alarms happen more often. To improve focus on this element and cut down on false alarms, a Scene Text Recognition system based on Multi-scale Dilate 2D Positional Encoding is being made by combining the Visual Reinforcement Module (VRM) and the Position Reinforcement Module (PRM). Adaptive Patches Vision Transformer (APViT) was proposed in [2] for scene text recognition employing three components. They are Sparse Patches Selection (SPS), ViT-STR, and Token Code (TC). Initially using SPS images were segmented into suitable patches. Followed by which similar images were clustered to investigate different local patches in an adaptive fashion. Finally, the TC fine-tuned unnecessary parts employing self-attention mechanisms to intensify performance with improved accuracy. In spite of improved accuracy, however, the training time was not focused. To address this issue, spatial scale coverage permits neurons to perceive wider and relevant context for differentiating between similar texts therefore minimizing overall training time. While most present works apply data driven attention-based method, however, such methods neglect character geometric information. In [3], a novel shape-driven attention alignment method was proposed to obtain character-wise features with superior performance. But visual semantic relations were not focused. To concentrate on this research gap visual semantic relationships employing multi-modal text recognition was presented in [4] to make final predictions. A survey of transformer techniques in text recognition was investigated in [5].

STR has become a long-standing research area, and almost all the existing methods chiefly apply the connectionist temporal classification (CTC) technique. Nevertheless, these existing methods were not found to be much efficient for irregular STR, therefore generating huge volume of character level errors. To address on this aspect similarity-based sequence label corruption method was presented in [6] with the intent of detecting label errors showcasing its practical utility. Nevertheless, the memory demands of these methods were found to be significant, constraining their application in resource-constrained applications. To address this issue, cascaded transformers were applied in [7]. A method to focus on bilingual text employing pre-trained CNN models was proposed in [8] with improved accuracy. The main objective of STR is to recognize text in complex images with the intent of transforming it into editable text. Nevertheless, STR research has long been concentrated on English and there is a dearth of research on other small languages. A lightweight seal text recognition method was presented in [9] to predict label distribution to achieve the highest accuracy effect. However, due to variations in scene characteristics, scene text processing poses significant issues. To address this aspect quantization and contour-based mechanism was proposed in [10] for improved scene text processing. Yet another deep learning (DL) method was presented in [11] for complicated language set.

The remainder of the paper is structured as follows. Section 2 presents relevant research on the STR utilizing machine learning and deep learning methodologies. Section 3 provides a concise overview of the AI agent language-dependent STR, Multi-scale Dilate Differentiable Binary Vision Transformer (MDDB-ViT) methodology. Subsequently, Section 4 presents the case study together with the resultant inferences. Section 5 presents the experimental results and includes a thorough assessment study comparing the proposed MDDB-ViT approach with two current methods, including both tabular and graphical representations. Ultimately, Section 6 finishes the manuscript.

2. RELATED RESEARCH WORK AND LITERATURE REVIEW

Earlier STR methods generated a notable quantity of inaccurate yields because these images lose precise text content. To address this issue, Residual Multi-Feature Pyramid Network was designed in [12] to extract multi-level associations with improved performance accuracy. However, the noise rate was not focused. Yet another robust intelligent system was modelled in [13] for demanding weather conditions employing DL that in turn generated satisfied outcomes. The training time involved was not addressed. A method to provide semantic information via vision transformer was presented in [14] to ensure balance between accuracy and time. Urdu handwritten text recognition was performed in [15] by employing transformer-based method.

In the domain of artificial intelligence (AI), text reading from natural scene images also referred to as Scene Text Recognition (STR) recognizes text automatically. Nevertheless, STR poses notable issues owing to different and complexity of natural scene text. Large language Model (LLM) was incorporated in [16] via dual prediction resulting in improved accuracy. Nevertheless,

prevailing methods still pose issues when processing text images with low contrast and insufficient illumination. To address these issues noise robust STR was presented in [17]. Nowadays, most STR methods are based on the Transformer architecture and depend on pre-trained LLMs. Nevertheless, these pre-trained methods are specifically designed for resource-rich languages. A low resource based STR called, Correlation-Guided Decoding Strategy was presented in [18] with improved accuracy. Yet another method leveraging LLM using simultaneous localization and mapping was designed in [19] to provide accurate text information. Nevertheless, the deterioration of stone surfaces along with script evolution over centuries constitutes notable issues to accurate digitization. To address these aspects GAN augmented hybrid transformer architecture was presented in [20] achieving high translation accuracy. To address computational efficiency semantic relationships were explored in [21] for STR using vision transformer.

A method incorporating Convolution Neural Network (CNN) with Long Short-Term Memory (LSTM) for differing character shapes and contextual dependencies was proposed in [22] for multilingual text detection. A dictionary embedded technique was presented in [23] to address high-resolution issues ensuring visualizing pleasing results. Yet another Deep Learning method to focus on the cross linguistic issues using spatial reduction were applied in [24] to enhance efficiency. Yet a method to focus on accuracy employing attention-based CNN was proposed in [25] enabling end-to-end sequence learning. A method to focus on the low recognition accuracy incorporating convolution and recurrent neural network was designed in [26] to ensure model generalization ability therefore ultimately achieving the objective of text recognition accuracy. However, the cross-attention operation utilized in fine tuning visual and linguistic features while decoding was observed to be computationally expensive. To address this issue, cross-attention-free STR framework was presented in [27]. AI integrated with LLM was designed in [28] [29] for SRT with improved precision and accuracy.

In summary, the drawbacks associated with most of the systems reviewed here are associated with overfitting and full proof under robust conditions single dimension positional encoding model, which makes them inconvenient for multilingual scene text. Despite inclusion of portable solutions, several methods heavily rely on permutation language model that in turn do not focus on complex backgrounds, therefore compromising the training time. In spite of some research works fine-tuning the significance of feature channels in an adaptive fashion with improved accuracy, they have not shown how they can reduce the risk of false positive and false negative rate. Motivated by this research gap, in this work, Multi-scale Dilate Differentiable Binary Vision Transformer (MDDB-ViT) concentrating on both complex backgrounds and detection with risk of false positive and false negative is designed. The elaborate description of the MDDB-ViT method is provided in the following sub-sections

3. DESIGN OF AGENTIC AI BASED MULTISCALE DIFFERENTIABLE VISION TRANSFORMER ON LLM

3.1 Data Collection and Description

The ICDAR 2019 MLT OCR Dataset, which can be found at <https://www.kaggle.com/datasets/zubairlibhutto/mlt-19-ocr-dataset> [30], is a benchmark dataset for multilingual scene text recognition and detection systems. It is used a lot for training and testing these systems. The collection consists of pictures of natural scenes with text added, such as street signs, store names, and street marketing boards. There are 20,000 photos in the dataset, and each language has 2,000 images. A lot of pictures have text in more than one language. But there are at least 2,000 pictures for each language. The most important thing about the collection is that it includes scene text graphics from 10 various languages: Arabic, Bangla, Chinese, English, French, German, Italian, Japanese, Korean, and Latin. The ICDAR 2019 MLT OCR Dataset has annotations that include bounding boxes, language labels, and transcriptions. The dataset has 10,000 photos, with 1,000 for each language. Half of them are for training and half are for testing. Also, Ground Truth (GT) adds annotations to the words in the scene photos. This GT is a series of characters that don't have any spaces between them. A 4-corner bounding box (which stands for a total of 8 locations) is used to tag each GT-word, along with a script class and a Unicode transcription of that GT-word.

3.2 Problem Formulation

Provided with scene image containing text, the objective of scene text recognition (STR) in our work strives to translate each individual character within the text present in the image. Let ' $S \subseteq \bigcup_{H*W} \mathbb{R}^{H*W*3}$ ' denote the text images space and ' $Out = \{Out_1, Out_2, \dots, Out_N\}, Out_i \in CC$ ' denote the space of sequences in character set ' CC '. Next with the assumption of the subsistence of underlying data distribution ' DD ' over the set ' $S * Out$ ', let us define a loss function ' $\mathcal{L}: Out * Out \rightarrow \mathbb{R}$ ' to evaluate the difference between two sequences. Then, the objective of STR in our work is to identify a prediction function ' $f: S \rightarrow Out$ ' that reduces the expected loss ' $EL(S, Out) \sim DD[\mathcal{L}(f(S), Out)]$ '. The function ' f ' is implemented by a deep

network ' ϕ ' that is parameterized by ' θ ' and trained on a dataset ' $DS = \{(S_i, Out_i)\}_{i=1}^N$ ' that contains ' M ' samples from dataset ' DS ' from of underlying data distribution ' DD '.

3.3 Design

Text is an indispensable mode for humans to obtain information and take part in social communication. Scene text recognition (STR) is an indispensable subfield that concentrates on recognizing and detecting text embedded in natural images. In spite of the conventional OCR systems performing well on scanned documents however owing to variations in font, orientation and language fail in complex real-world environments. Recent evolutions in deep learning and ViT have notably boosted STR. However, most systems being language-dependent, limits generalization across multilingual scenarios. This work proposes an AI agent-based language independent STR method called Multi-scale Dilate Differentiable Binary Vision Transformer (MDDB-ViT) supported on LLM leveraging multimodal learning and reasoning capabilities. Figure 1 shows the architecture of MDDB-ViT method.

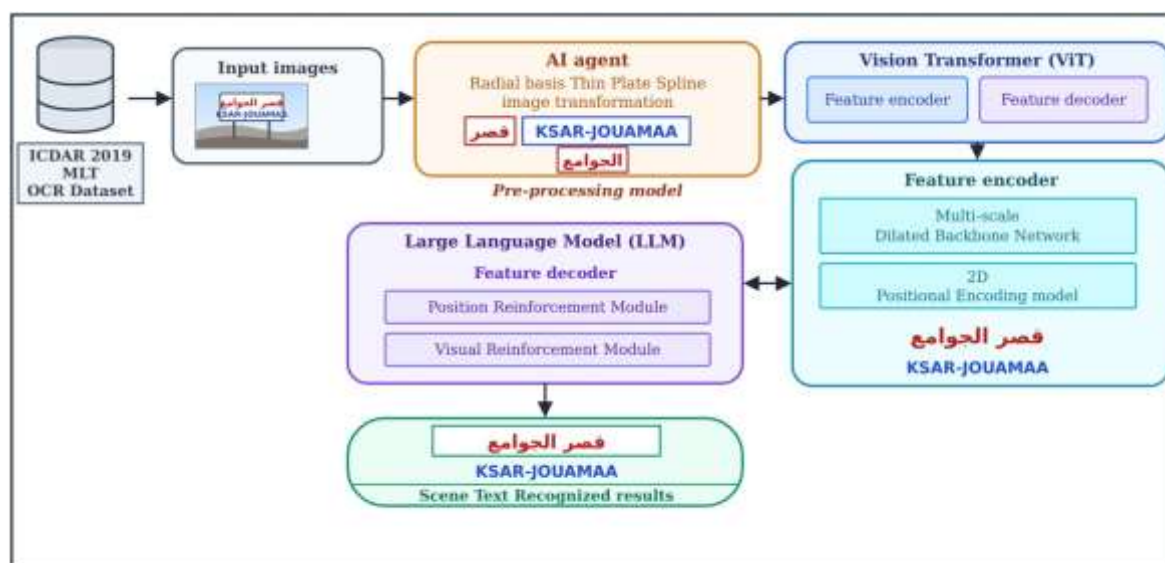


Figure 1. Architecture of MDDB-ViT method for scene text recognition.

As illustrated in the above figure the key components of the proposed method are AI agent, Vision Transformer (ViT), Large Language Model (LLM) and Language-Independent Scene Text Recognition (STR). The AI-agent will incorporate ViT's, LLMs, therefore making certain decision making adaptively to generalize across unseen languages using ICDAR 2019 MLT OCR Dataset extracted from <https://www.kaggle.com/datasets/zubairlibhutto/mlt-19-ocr-dataset>. AI Agent refers to the structure fuelled by Large Language Models (LLMs) that split the tasks of comprehending a sample input into intermediate steps like text detection, image rotation, script identification and character recognition. This is initially performed by the AI agent using Radial Basis Thin Plate Spline Image Transformation-based pre-processing model. Vision Transformer refers to a model that undertakes image patches to recognize text in complex backgrounds. The central components of novel Vision Transformer include Feature Encoder and Feature decoder. Scene Text Recognition (STR) is performed via Feature Encoder that with the aid of bounding boxes locate text in the sample input image and Multi-scale Dilate 2D Positional Encoding model. On the other hand, scene text detection discovers text regions employing detection head utilizing bounding boxes (simply say where the text) via Differentiable Binary Position Visual Reinforced Large Language based Decoder by mapping feature into characters in an extensive manner. The Large Language Model in decoder serves as an organizer using language knowledge with the intent of interpreting and obtaining the output from visual model, enhancing accuracy even in case of the text is partially blurred. Finally, the Language-Independent STR eliminates the requirement for language-specific classifiers by utilizing holistic vision-language model to recognize the text structure instead of depending on a predetermined dictionary of a single language.

3.3.1 Radial Basis Thin Plate Spline Image Transformation-based Pre-processing

Scene text recognizers are most efficient when their input images include firmly restricted regular text. This motivates us to perform a spatial transformation prior recognition process with the objective of converting input images into recognizable format. Despite though simple recognition tasks might skip the image transformation process, the STR framework employed in our work utilize transformation as a pre-processing step to reform text into a standardized format (e.g., straightening a skewed word) before feeding it into a feature extraction process. In our work, Radial Basis Thin Plate Spline Image Transformation-based pre-processing model is employed to transform an input text image into normalized image. Figure 2 shows the structure of the Radial Basis Thin Plate Spline Image Transformation-based pre-processing model.

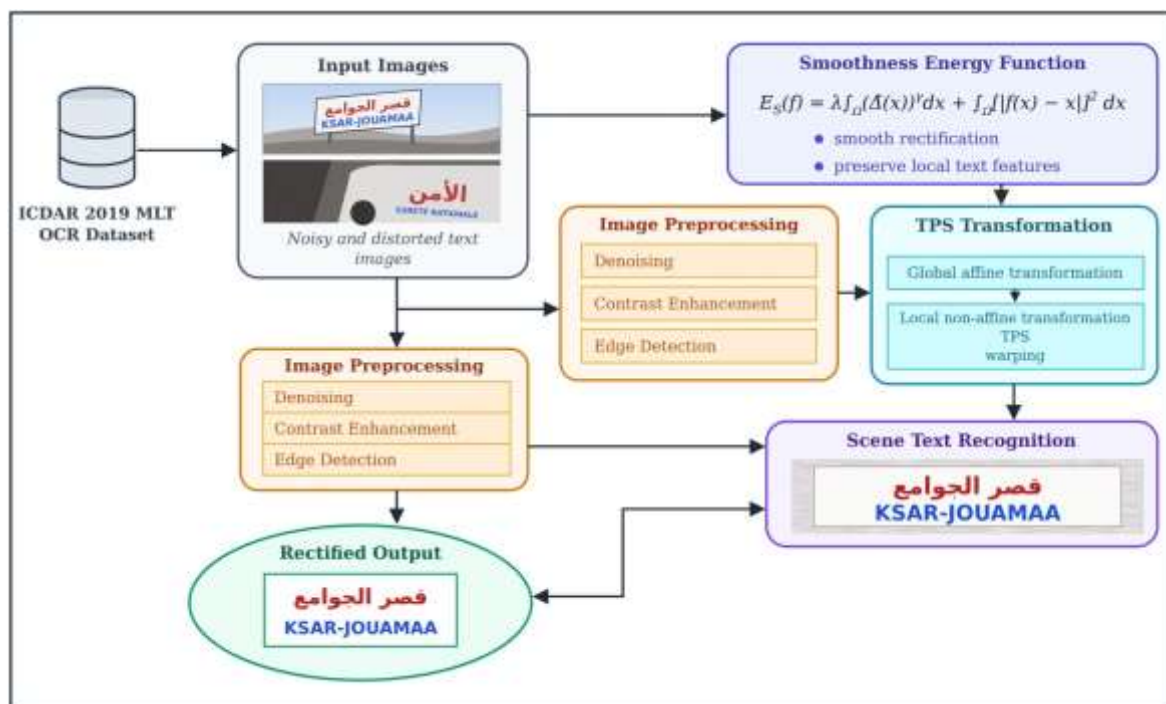


Figure 2. Structure of Radial Basis Thin Plate Spline Image Transformation-based pre-processing.

As illustrated in the above figure, the pre-processing model works by obtaining a set of fiducial or control points on the input text image obtained from ICDAR 2019 MLT OCR Dataset. The obtained fiducial points are mapped to subsequent points on a straight image via spatial transformation by mapping pixels between coordinates to reshape it for further processing. The transformation splits into a global affine component and a local non-affine component, therefore allowing handle irregular distortions. Initially, the corresponding ‘P’ fiducial points $(a_i, b_i), i = 1, 2, \dots, P$ are identified. With the ‘a’ being two dimensional in our work, the TPS fits a mapping function ‘f(a)’ between point sets $\{b_i\}$ and $\{a_i\}$ that minimizes the following energy function. TPS interpolates between collections of fiducial points utilizing smooth spline. This is due to the reason that with the text images coming in numerous shapes, i.e., tilted and curved texts, the energy function employed here acts as a controller to make certain that the rectification is smooth.

$$E_{TPS}(f) = \sum_{i=1}^l (b_i - f(a_i))^2 \quad (1)$$

Hence by minimizing the energy function according to (1) corrects orientation without creating disorganized distortions to the characters in use. Followed by which warping coefficients are measured by solving a spatial linear system emanated from distances between fiducial points via radial basis function as given below.

$$f(a) = \sum_{i=1}^l \alpha_i \phi(|z - CP_i|) \quad (2)$$

From the above equation () given a set of control points ‘ $CP_i, i = 1, 2, \dots, l$ ’ (i.e. acting as a center point for kernel), a radial basis function is defined by spatial mapping that maps pixel position ‘a’ to new pixel position denoted by ‘f(a)’. By applying the

above spatial linear system via radial basis function learn an invariant representation for several aspect ratios to reduce the complexity involved in complex-shaped text images. Finally the above transformation function is applied to all the pixels in the sample image to obtain processed image for further processing

3.3.2 Multi-scale Dilate 2D Positional Encoding-based Scene Text Recognition model

To achieve higher accuracy with lower false positive and false negative rate, a Multi-scale Dilate 2D Positional Encoding-based Scene Text Recognition model is designed as shown in figure 3.

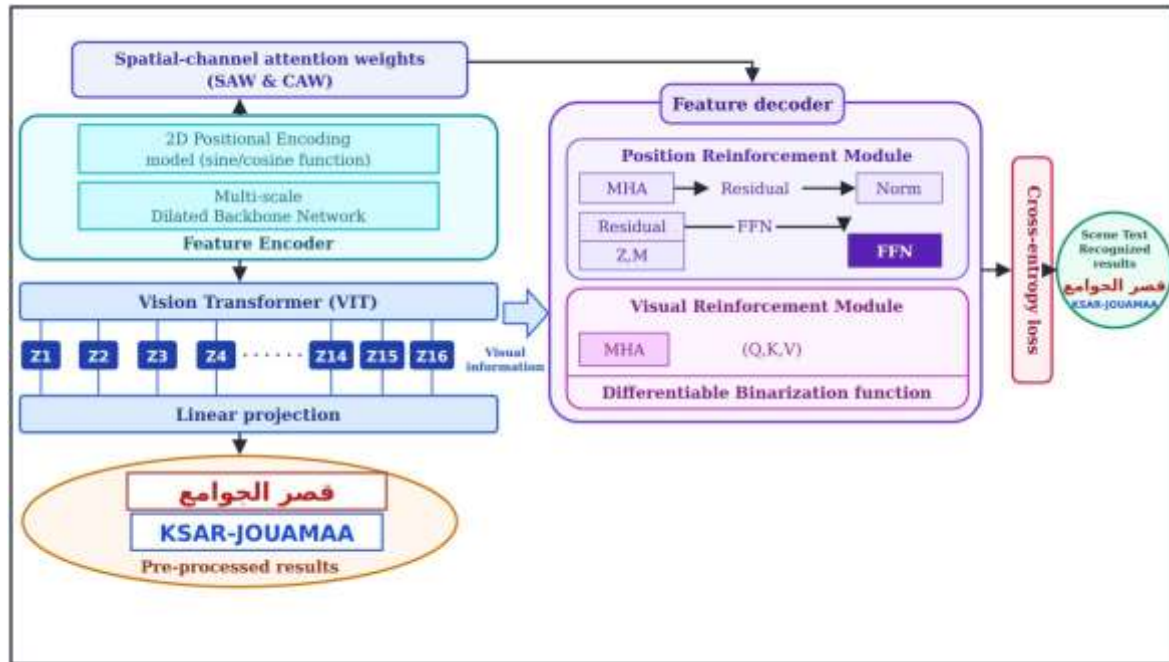


Figure 3. Structure of Multi-scale Dilate 2D Positional Encoding-based Scene Text Recognition.

In the encoder part, Vision Transformer along with Multi-scale Dilate is used as the backbone network for extracting visual information. We also design a 2D Positional Encoding model with the objective of learning the character position information that assists in optimizing spatial alignment between feature map and target character area, boosting the scene text recognition accuracy. In the decoder part, we introduce the Position Reinforcement Module (PRM) and Visual Reinforcement Module (VRM) to not only reinforce position information and integrate it with visual information and thus boost the model's accuracy with minimal false positive and false negative rate.

3.3.3 Multi-scale Dilated Backbone Network

The Multi-scale Dilated Backbone Network employed in our work being a specialized convolution neural network (CNN) enhances the feature extraction process by obtaining text at multiple scales without losing resolution. Let us consider the sample input image $S \in \mathbb{R}^{H \times W \times C}$, that is initially processed by patch embedding unit 16×16 , split into fixed-size image patches and mapped into a high dimensional feature space with the purpose of generating a sequence of patch features. A feature encoder employed in our work comprises of k encoder blocks, with $k = 12$ and the input of the feature encoder is represented as given below.

$$Z_0 = \{S_{class}; S_P^1 LM; S_P^2 LM; \dots; S_P^N LM\} + PE \quad (3)$$

From the above equation (3) S_{class} denotes the class embedding result generated via the linear projection matrix LM utilized in transforming input data from one vector space to another for the corresponding P patch dimension with respect to overall patch sequence length of N . This patch sequence is then provided as input to the backbone network where numerous layers capture global dependencies via multi-head self attention (MSA) and multi layer perceptron (MLP). The mathematical formula for MSA and MLP is as given below.

$$Z^{(k)} = Z^{(k)} + MHA(LN(Z^{(k)})) \quad (4)$$

$$Z^{(k+1)} = Z^{(k)} + MLP(LN(Z^{(k)})) \quad (5)$$

From the above equations (4) and (5) the MSA and MLP results are arrived at for ' k ' number of encoder blocks. The final output obtained above is embedded in ' $Z^{(k+1)}$ ', successive recognition forming as input to the Visual Reinforcement Model. Next, To efficiently encapsulate local stroke specimens and global contextual layout, a geometric dilation rate progression is employed ensuring balanced spatial scale coverage with ' $ss = 0$ ' denoting fine details, with ' $ss = 0.5$ ' for single-character context and ' $ss = 1$ ' for multi-character respectively. This configuration empirically achieves a favorable trade-off between representational capacity and model complexity employing multiple convolution branches with three different dilation rates. This is mathematically denoted as given below.

$$Fus_i = Conv_{3 \times 3}^{ss=ss_i}(Z^k) \quad (6)$$

From the above equation (6) with the spatial scale coverage represented as ' $ss = ss_i$, where $i = 0, 0.5$ or 1 ' extract features with different scale variations within the same input image.

3.3.4 2D Positional Encoding model

The 2D Positional Encoding employed in our work broadens the conventional 1D Positional Encoding by dividing the embedding dimension into two portions, one for x-coordinate employing sine function whereas another for y-coordinate using cosine function. Hence by applying the 2D Positional Encoding model with the feature of preserving spatial, top-down and left-right sequence improves the overall accuracy. Moreover, traditional vision transformers employed for STD lack recurrence convolution layers. Hence, it is paramount to impart relative location information of characters in the word for processing. With this objective, 2D positional encoding maps are employed ' $PE \in \mathbb{R}^{H \times W \times C}$ ', to represent the spatial information. Then the conventional 1D encoding suitable for 2D positional encoding maps is formulated as given below.

$$PE(HR, VR, 2i) = \sin(POS(HR). \alpha_i) \quad (7)$$

$$PE(HR, VR, 2i + 1) = \cos(POS(HR). \alpha_i) \quad (8)$$

$$PE(HR, VR, 2j + C_{size}/2) = \sin(POS(VR). \alpha_j) \quad (9)$$

$$PE(HR, VR, 2j + 1 + C_{size}/2) = \cos(POS(VR). \alpha_j) \quad (10)$$

From the above equations (7), (8), (9) and (10) the positional encoding map results are generated based on the ' $POS(HR)$ ', ' $POS(VR)$ ' denoting the horizontal, vertical position with ' α_i ', ' α_j ' denoting the learning factors for 2D positional encoding with respect to the channel size ' C_{size} ' respectively. With this, though the 1D positional encoding found to be suitable for sequential text, however using 2D positional encoding aids in capturing spatial relationship between characters, therefore providing higher accuracy in case of STR. Moreover, to integrate features for differing spatial scale coverage a spatial-channel attention weight mechanism is employed with the intent of notably improving the model's robustness against scale variations, distortions, therefore minimizing the false positive and false negative rate. Formulating spatial and channel attention weight models encompasses deriving relationships between specific information across pixels (i.e. spatial relations) and interdependencies between different feature channels. This procedure is initiated by obtaining a spatial similarity matrix, followed by which obtaining a channel similarity matrix via matrix multiplication. This is mathematically defined as given below.

$$Pos_{ij} = \frac{\exp(WM_i^T * WM_j)}{\sum_{i=1}^N \exp(WM_i^T * WM_j)} \quad (11)$$

$$C_{ij} = \frac{\exp(WM_i * WM_j^T)}{\sum_{i=1}^N \exp(WM_i * WM_j^T)} \quad (12)$$

From the above equations (11) and (12) ' Pos_{ij} ' refers to the influence of the ' $i - th$ ' position on ' $j - th$ ' position and ' C_{ij} ' denoting the influence of the ' $i - th$ ' channel on ' $j - th$ ' channel and finally ' WM_i^T ' denotes the weight matrix of ' $i - th$ ' row of ' WM ' and ' WM_j ' representing the ' $j - th$ ' column of ' WM ' respectively. Finally using both spatial and channel information

employing reshape tensor manipulation operation are reshaped for final feature enhancement. This is mathematically defined as given below.

$$AggRes = Reshape(WM * SAW) \oplus Reshape(CAW * WM) \quad (13)$$

From the above equation (13) the feature information of the spatial and channel attention model result '*AggRes*' is arrived at based on the alignment of feature representation results generated using '*Reshape*' function with respect to the spatial attention weight '*SAW*' and channel attention weight '*CAW*' activated via softmax function. The pseudo code representation of Multi-scale Dilated 2D Positional Encoding-based Scene Text Recognition is given below.

Algorithm 1: Multi-scale Dilate 2D Positional Encoding-based Scene Text Recognition

Input: Dataset '*DS*', Samples '*S* = {*S*₁, *S*₂, ..., *S*_{*N*}}

Output: computationally efficient and robust scene text recognition.

```

1: Initialize 'N = 6000'
2: Begin
3: For each Dataset 'DS' with Samples 'S' //pre-processing
4:     Generate Thin Plate Spline Image Transformation according to eq. (1) and eq. (2)
5:     Return pre-processed results
6: End for
//Multi-scale Dilated Backbone Network
7: For each Dataset 'DS' with pre-processed results
8:     Generate the input of feature encoder according to eq. (3)
9:     Generate the mathematical formula for MSA and MLP according to eq. (4) and eq. (5)
10:    Obtain spatial scale coverage result according to eq. (6)
//2D Positional Encoding model
11:    Generate positional encoding map results for horizontal and vertical position according to eq. (7), eq. (8), eq. (9) and eq. (10).
12:    Evaluate position similarity matrix results according to eq. (11)
13:    Evaluate channel similarity matrix results according to eq. (12)
14:    Combine feature information of spatial and channel attention model result according to eq. (13)
15:    Return scene text recognized results
16: End for
End of the Algorithm

```

As given in the above algorithm with the raw input samples obtained from ICDAR 2019 MLT OCR Dataset is initially subjected to pre-processing. Here, with the intent of reforming text into a standardized format utilizing smooth spline by minimizing energy function corrects orientation without creating disorganized distortions to the characters. Also, by applying radial basis function learn an invariant representation for several aspect ratios, therefore improving the overall peak signal to noise ratio (PSNR). Next the pre-processed sampled images are provided as input to feature encoder via Multi-scale Dilated Backbone Network and 2D Positional Encoding model. Here, by using dilated convolutions with varying rates via spatial scale coverage on contrary to standard convolutions, the network capture both fine local details and broad contextual information.

3.3.5 Differentiable Binary Position Visual Reinforced Large Language based Decoder

The feature decoder on the other hand interprets the extracted features to generate the final text sequence. Taking into consideration the baseline's restricted concentration on local detail information, a feature decoder is outlined to enhance the detail of text features by acquiring precise text boxes and enhance the model's feature learning potentiality. It is recognized that the forefront text instances make use of a very small dimension in the image, triggering equally rare useful information for scene text detection. In addition, the foremost scene text ration distribution suitably necessitates rich fine detail information to generate accurate and precise perceptions in scene text detection. Motivating at the above issues, this paper designs a position reinforcement module that utilizes the magnified position information as query vector to input visual reinforcement module. The linear layer finally outputs the detected text, therefore boosting the text feature detail information, providing finer features.

3.3.6 Position Reinforcement Module

The input of the Position Reinforcement Module refers to is the position embedding output of 2D Positional Encoding model in the encoder and the reinforced position information is obtained via Multi Head Attention (MHA), Residual Connection via Dropout, Layer Normalization, and Feed Forward Network (FFN) respectively via Query ' Q ', Key ' K ' and Value ' V '. This is mathematically defined as given below.

$$MHA(Q, K, V) = \text{Agg} \left(\text{Attention}(Qw_1^Q, Kw_1^K, Vw_1^V, \dots, Qw_N^Q, Kw_N^K, Vw_N^V) \right) \quad (14)$$

From the above equation (14), ' Q ', ' K ', ' V ' denotes the query, key, value with respect to corresponding weight matrices ' w_1^Q ', ' w_1^K ', ' w_1^V ' for the subsequent ' i - th attention' head aggregated via ' Agg ' function respectively. By applying the MHA boosts feature characterization by encapsulating long range dependencies therefore to concentrate on text regions while discarding complicated background portions. The result of the Multi Head Attention (MHA) function is superimposed to the input query to generate the residual connection result and then progressed via layer normalization. The result of residual connection and normalization is defined as given below.

$$RC = Q + \text{Dropout}(MHA(Q, K, V)) \quad (15)$$

$$NR = \text{Norm} \left(\frac{\beta_1}{\sigma} (RC - \mu) + \beta_2 \right) \quad (16)$$

From the above equations (15) and (16) the residual connection ' RC ' and normalization results ' NR ' are obtained by employing the mean ' μ ', standard deviation ' σ ', scaling point ' β_1 ', shifting point ' β_2 ' respectively. Here by making use of these two functions jointly permit training in turn detect varied size text. Followed by which the Feed Forward Network (FFN) result is generated as given below.

$$FFN(NR) = \gamma_i \text{ReLU}(\gamma_j(NR) + b_i) + b_j \quad (17)$$

From the above equation (17) utilizing the feed forward weight matrices ' γ_i ', ' γ_j ' and feed forward bias matrices ' b_i ', ' b_j ' the Feed Forward Network ' FFN ' results are obtained. By utilizing this network interpret pixel wise scores and map them back to the actual sample image coordinates employed to draw bounding boxes. Finally to obtain the scene text detected results the result of the ' FFN ' is appended to the normalized result as given below.

$$STD = \text{Norm} \left(NR + \text{Dropout}(FFN(NR)) \right) \quad (18)$$

3.3.7 Visual Reinforcement Model

The input of the Visual Reinforcement Model consists of the visual information extracted by the Vision Transformer (ViT) in the encoder and the position information enhanced by the Position Reinforcement Model. This input is processed with MHA, residual connections and layer normalization to generate the visually intensified features with improved positional information. In MHA the query ' Q ' denotes the position feature ' P ' that has been intensified by the Position Reinforcement Model with ' K ' and ' V ' denoting the visual features extracted using Vision Transformer. Finally, Differentiable Binarization function is employed that on contrary to traditional function convert text probability maps to binary maps (i.e. presence or absence of text) as given below.

$$DB = STD \left[\frac{1}{1 + e^{-k(PM_{ij} - TM_{ij})}} \right] \quad (19)$$

From the above equation (19) the differentiable binary map ‘ DB ’ results are generated, via probability map ‘ PM_{ij} ’ and threshold map ‘ TM_{ij} ’, for the corresponding sampled scene text detected result features ‘ STD ’ respectively. Finally using the cross-entropy loss function the loss is evaluated as given below.

$$\mathcal{L} = \sum_{i \in S_i} Pred_i \log Act_i + (1 - Pred_i) \log(1 - Act_i) \quad (20)$$

From the above equation (20) the cross-entropy loss function results ‘ $\mathcal{L}$ ’ are obtained based on the actual label value ‘ Act_i ’ while ‘ $Pred_i$ ’ denotes the predicted probability value of sample ‘ S_i ’ respectively. The pseudo code representation of Differentiable Binary Position Visual Reinforced Large Language based Decoder is given below.

Algorithm 2: Differentiable Binary Position Visual Reinforced Large Language based Decoder

Input: Dataset ‘ DS ’, Samples ‘ $S = \{S_1, S_2, \dots, S_N\}$ ’

Output: Intelligent and accurate scene text detection

- 1: Initialize $N = 6000$
- 2: Begin
- 3: For each Dataset ‘ DS ’ with Samples ‘ S ’
- //Position reinforcement module
- 4: Encapsulate long range dependencies by applying MHA according to eq. (14)
- 5: Detect varied size text via residual connection and normalization function according to eq. (15) and eq. (16)
- 6: Generate bounding boxes results using FFN according to eq. (17)
- 7: Generate scene detected results according to eq. (18)
- 8: Go to step 9
- //Visual reinforcement module
- 9: Evaluate Differentiable Binarization function result according to eq. (19)
- 10: Evaluate cross entropy loss function result according to eq. (20)
- 11: End for

End of the Algorithm

As given in the above algorithm by using high-level semantic information from a Large Language Model (LLM) in scene text detection refine the visual localization of text. By incorporating linguistic understanding with visual cues aids in better handling demanding scenarios that helps differentiate actual text from complex backgrounds or text-like textures. Also, by making the threshold process differential predicts precise text positions therefore improving the overall scene text detection accuracy.

3.4 Data Description

In this section case analysis of STC using ICDAR 2019 MLT dataset are simulated by applying MDDB-ViT method. Figure 4 given below provides the sample multilingual scene text images for performing simulation.



Figure 4. Sample images obtained from ICDAR 2019 MLT dataset.

With the above sample multilingual scene text images obtained from the ICDAR 2019 MLT dataset is initially subjected to the pre-processing by using Radial Basis Thin Plate Spline Image Transformation model. Figure 5 shows the pre-processed results.



Figure 5. Radial Basis Thin Plate Spline Image Transformation pre-processed results.

As illustrated in the above figure by obtaining the distances between fiducial points via radial basis function for several aspect ratios reduce the complexity involved in complex-shaped text images therefore improves the overall peak signal to noise ratio. The pre-processed samples are then subjected to scene text recognition by using Multi-scale Dilate 2D Positional Encoding model. Figure 6 shows the encoded results via Multi-scale Dilated Backbone Network and 2D Positional Encoding model.



Figure 6. Feature encoder results.

The first row in the above shows the pre-processed results. Following which Multi-scale Dilated Backbone Network was applied and the results are shown in the second row. By applying the by using dilated convolutions with varying rates with distinct spatial scale coverage aided in obtaining broad contextual information, therefore improving the overall training time. Next, the third row shows the results of the 2D Positional Encoding. Here, helps differentiate between similar but have different spatial locations, therefore minimizing the misinterpretation. With this resulted in the overall improvement of false positive and false negative rate in an extensive manner. Figure 7 given below shows the feature decoder results.

Bounding Boxes Results	Decoder Results
	



Figure 7. Decoder results.

As shown above feature decoder results comprise of bounding box along with the scene text detection results. By applying the position reinforcement and visual reinforcement model separately aids in better handling demanding scenarios that helps differentiate actual text from complex backgrounds or text-like textures. This in turn improves overall detection accuracy.

4. EXPERIMENTAL SETUP AND DISCUSSION

In this section, a comparison study is provided to analyze and validate the performance of three different vision transformer methods, AI agent-based language dependent STR method, Multi-scale Dilate Differentiable Binary Vision Transformer (MDDB-ViT) supported on LLM, Improved Multi-task Transformer based on Localization Mechanism Network (IMTLM-Net) [1], Adaptive Patches Vision Transformer (APViT) [2] in modeling language independent STR using ICDAR 2019 MLT OCR Dataset extracted from <https://www.kaggle.com/datasets/zubairlibhutto/mlt-19-ocr-dataset>. In the language independent STR with images obtained from ICDAR 2019 MLT dataset, the performance of the selected three is assessed via calculating the mean values of four quantitative measures using the testing set, including PSNR, training time, false positive rate, false negative rate and STR accuracy using Python general-purpose high-level programming language. The method was analyzed on a computer Intel(R) Core (TM) i7-6700HQ CPU@2.60GHz with a RAM of 32 GB running Windows. Similar training strategy was applied for all three methods and to make certain fair comparison is made between the proposed and two existing methods [1] and [2], input samples used for all training sessions are kept identical. Hence, with the objective of identifying the efficiency of the method, 6000 samples were selected arbitrarily. In addition, for each experiment performed, same dataset was employed for all the three methods.

4.1 Performance analysis of PSNR

As a quantitative performance metric, PSNR evaluates the reconstruction quality of low-resolution (LR) text images by measuring the pixel-wise difference between a restored image and its high-resolution (HR) ground truth. A higher PSNR represents better restoration of textual details that boosts the accuracy of STR methods. This is mathematically denoted as given below.

$$MSE = \frac{1}{m} \frac{1}{n} \sum_{x=1}^m \sum_{y=1}^n [Pre(a, b) - Act(a, b)]^2 \quad (21)$$

From the above equation (20) 'm' and 'n' refers to the rows and columns of the pixel matrix '(a, b)' with 'a' and 'b' denoting the row and column index respectively. Then, PSNR is defined as given below.

$$PSNR = 10. \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (22)$$

From the above equation (21) PSNR rate 'PSNR' is measured based on the mean square error values 'MSE' and maximum pixel value of MRI (i.e. 'MAX = 25'). In table 1 as listed below, the ten standard images from three different languages (i.e. Latin, Chinese and Arabic) are taken and compared with their performance based on PSNR as mentioned above. The value of PSNR shown in Table 1, irrespective of sample image size, the proposed MDDB-ViT test is improved than all discussed methods, i.e. IMTLM-Net [1] and APViT [2].

Table 1. Comparison results of PSNR.

Sample image size (KB)	PSNR (dB)		
	MDDB-ViT	IMTLM-Net [1]	APViT [2]
408	27.15	23.15	21.45
532	21.35	17.35	15.25
160.2	23.35	19.55	17.35
227	20.35	16.15	14.25
368	17.25	13.35	11.15
476	16.15	12.55	10.55
648	15.35	11.25	9.35
145	20.15	16.15	14.25
37.4	23.35	19.65	17.55
41.8	22.45	18.35	16.15

Figure 8 given above illustrates the results of PSNR using the proposed MDDB-ViT and two existing methods, IMTLM-Net [1] and APViT [2]. A mix of increase and decrease in PSNR values were observed using all the three methods irrespective of sample image size. However comparative analysis between three methods observed an improvement in terms of PSNR when using the proposed MDDB-ViT method. The reason was by reforming text into a standardized format using smooth spline minimized energy function correction orientation without creating disorganized distortions to the characters. Moreover, by employing radial basis function helps in learning invariant representation for several aspect ratios, therefore improving the overall peak signal to noise ratio (PSNR) of the MDDB-ViT method by 20% compared to [1] and 30% compared to [2].

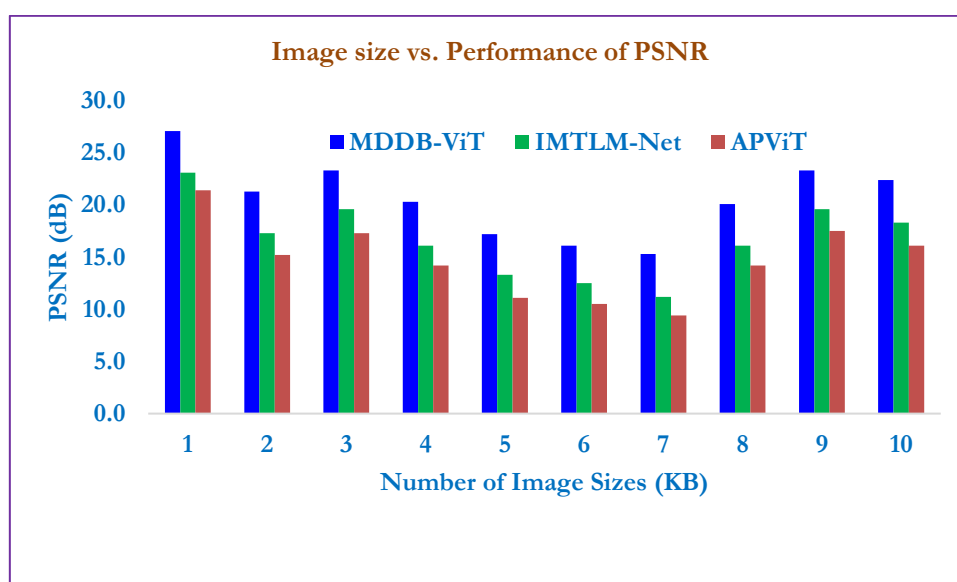


Figure 8. Sample image size versus PSNR.

4.2 Performance analysis of Training time

In this section the training time involved in analyzing language independent STR method is estimated as given below.

$$TT = \sum_{i=1}^N S_i * STD \quad (23)$$

From the above equation (23), training time ‘ TT ’ is measured based on the sample images ‘ S_i ’ obtained for experiments and the time consumed in scene text detected results ‘ STD ’ as output. It is measured in terms of seconds (sec). Table 2 given below lists the results of training time using three methods, MDDB-ViT, IMTLM-Net [1] and APViT [2] respectively.

Table 2. Comparison results of training time.

Samples	Training time (sec)		
	MDDB-ViT	IMTLM-Net [1]	MDDB-ViT
600	9	10.8	11.4
1200	10.35	11.46	13.15
1800	11.45	13.56	15.25
2400	12.55	14.66	16.85
3000	13.85	15.95	17.35
3600	14.35	16.46	18.55
4200	15.75	17.86	19.35
4800	16.85	18.96	20.45
5400	17.25	19.36	21.55
6000	18	20	22

Figure 9 given above illustrates the training time using the three methods, MDDB-ViT, IMTLM-Net [1] and APViT [2]. Samples ranging between 600 and 6000 were used for simulation and an overall ten simulation runs were performed to evaluate the training time. An increase in the training time was observed using all the three methods with the increase in the sample size. But simulations performed for average ten simulation runs showed minimal training time using MDDB-ViT method when compared to [1] and [2]. The reason was by using dilated convolutions with varying rates via spatial scale coverage upon comparison to standard convolutions both fine local details along with broad contextual information were captured. In addition by maintaining high-resolution feature maps permits neurons for differentiating between similar texts in turn aids in improving the overall training time of the MDDB-ViT method by 15% compared to [1] and 27% compared to [2].

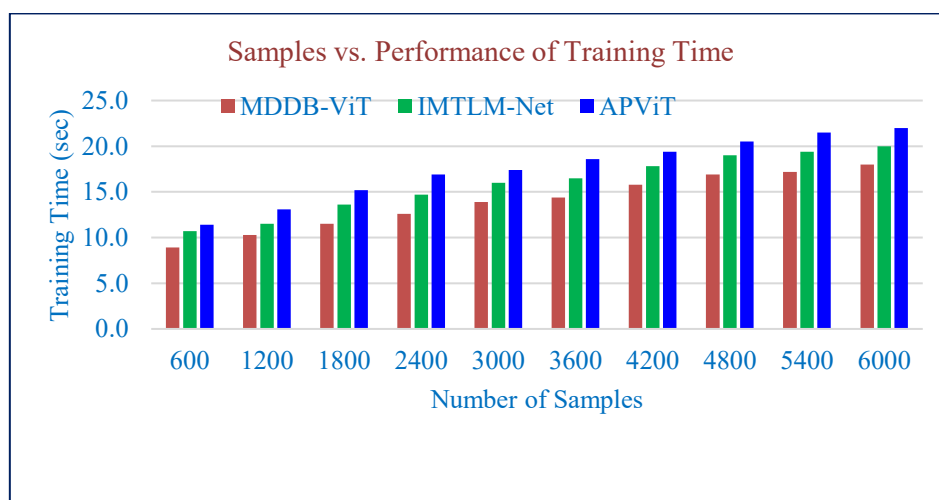


Figure 9. Samples Vs Training time.

4.3 Analysis of Performance Metrics: False Positive Rate, False Negative Rate, and Scene Text Recognition Precision

As far as Scene Text Recognition (STR) and detection is concerned, false positive rate and false negative rate are critical error modes that refer to the incorrect identification of non-text as text and the failure to detect actual text in natural images. The false positive rate and false negative rate is mathematically defined as given below.

$$FPR = \frac{FP}{FP+TN} \quad (24)$$

$$FNR = \frac{FN}{FN+TP} \quad (25)$$

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (26)$$

From the above equations (24), (25) and (26) false positive rate '*FPR*', false negative rate '*FNR*' and STR accuracy *Acc* '*Acc*' results are obtained based on the true positive rate '*TP*', false positive rate '*FP*', true negative rate '*TN*' and false negative rate '*FN*' respectively. False positive rate in other words denotes the probability that a true negative case is incorrectly identified as positive whereas false negative rate denotes the probability that a true positive case is incorrectly identified as negative. To be more specific a false positive is said to occur when the method identifies a portion as text though being part of background or noise. On the other hand a false negative is said to occur when the method fails in detecting text that is actually present in the sample image. Table 3 and table 4 lists the results of false positive rate, false negative rate and STR accuracy respectively.

Table 3. Comparison results of false positive rate and false negative rate.

Number of Samples	False Positive Rate			False Negative Rate			Accuracy (%)		
	MDDB-ViT	IMTLM-Net [1]	APViT [2]	MDDB-ViT	IMTLM-Net [1]	APViT [2]	MDDB-ViT	IMTLM-Net [1]	APViT [2]
600	0.54	0.68	0.76	0.005	0.01	0.014	97	95	93
1200	0.59	0.68	0.73	0.007	0.011	0.015	95.35	87.25	77.25
1800	0.62	0.71	0.76	0.009	0.012	0.016	93	85.15	75.35
2400	0.68	0.77	0.85	0.011	0.013	0.016	91	83.25	78.15
3000	0.73	0.83	0.88	0.012	0.014	0.017	90.35	82.15	72.55
3600	0.67	0.75	0.78	0.01	0.011	0.012	90	82.25	72
4200	0.64	0.72	0.75	0.009	0.01	0.011	88.25	80.45	70.45
4800	0.61	0.7	0.73	0.007	0.008	0.009	86.25	78.35	68.25
5400	0.58	0.65	0.71	0.007	0.008	0.007	84	76.15	66.35
6000	0.57	0.62	0.7	0.006	0.007	0.008	88.25	80.55	70.55

Figure 10 (a) and (b) shows results of false negative rate and false positive rate for samples ranging between 600 and 6000 using MDDB-ViT, IMTLM-Net [1] and APViT [2]. An average of ten simulation runs was performed, and values were plotted accordingly by substituting in equations (24) and (25). Neither increase nor decrease in false positive and false negative rate was observed using three methods. Also, false positive rate and false negative rate of MDDB-ViT method was found to be comparatively better than [1] and [2]. The reason was by applying the 2D Positional Encoding model upon comparison to traditional 1D model helps distinguish between similar but have different spatial locations, therefore reducing the misinterpretation. Moreover, by incorporating spatial and channel attention model for scene text recognition were observed to be less likely to miss text. This in turn reduced the false negative rate of MDDB-ViT method by 29% compared to [32], 56% compared to [33] and minimized false positive rate of MDDB-ViT method by 14% compared to [34] and 23% compared to [32].

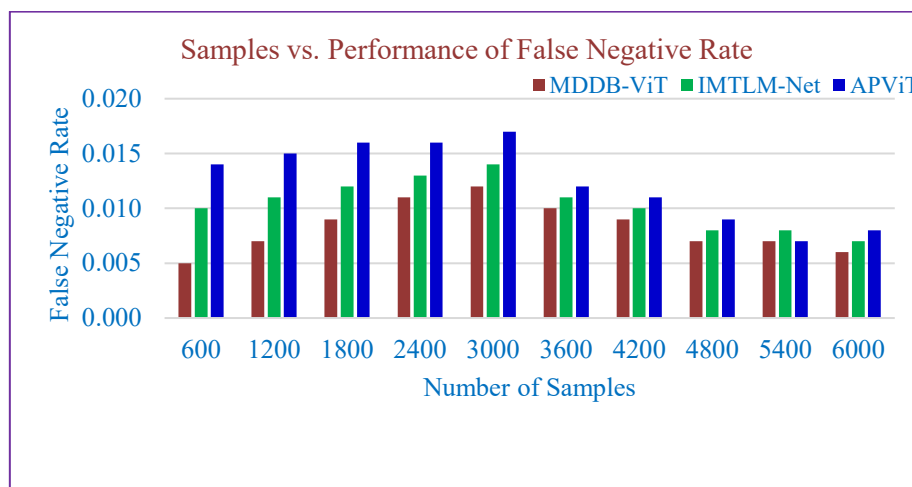


Figure 10. Samples versus False Negative Rate.

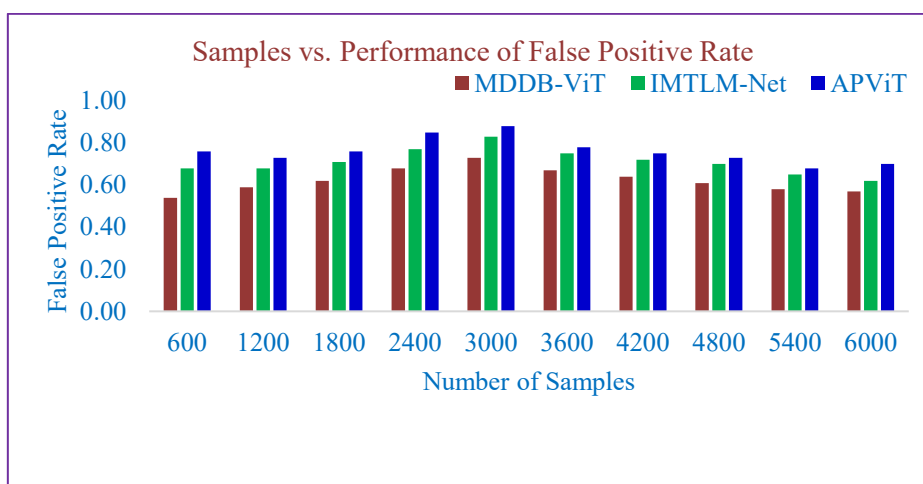


Figure 11. Samples versus False Positive Rate.

Finally figure 12 given above illustrates the accuracy using the three methods, MDDB-ViT, IMTLM-Net [1] and APViT [2]. The above results show improved accuracy when applied with MDDB-ViT method. The reason was by applying the Differentiable Binary Position Visual Reinforced Large Language based Decoder refine the visual localization of text via LLM. Moreover, by fusing linguistic understanding with visual cues helps in better handling demanding scenarios. This in turn differentiated actual text from complex backgrounds. Finally, by making threshold procedure differential predicts precise text positions therefore improving the overall scene text detection accuracy of MDDB-ViT method by 8% compared to [30] and 18% compared to [31].

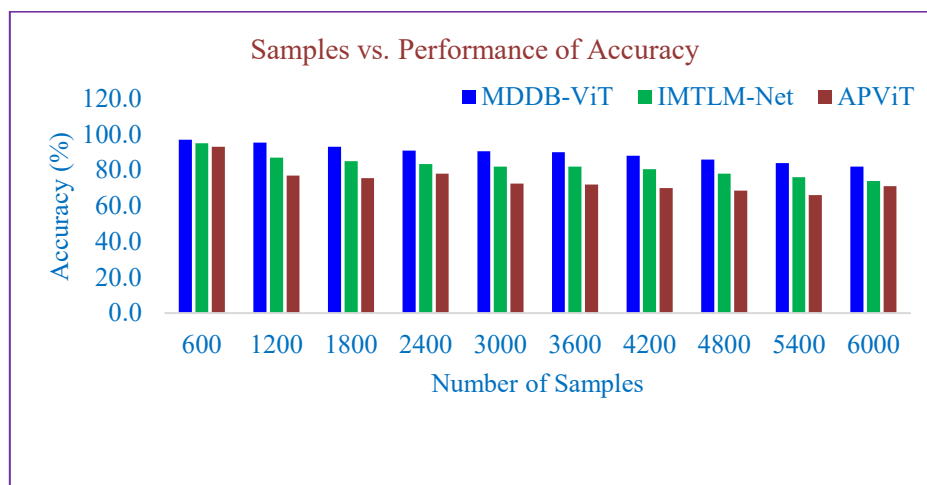


Figure 12. Samples versus Accuracy

5. CONCLUSION

Autonomous driving, robotic navigation necessitates to make certain the accurate identification of scene text with minimal response time. In this work an AI agent-based language dependent STR method called, Multi-scale Dilate Differentiable Binary Vision Transformer (MDDB-ViT) supported on LLM is proposed for making informed decisions. Here three processes, i.e., pre-processing for transforming an input text image into normalized image were designed. With the pre-processing being analytical and is straight forward Multi-scale Dilate 2D Positional Encoding-based Scene Text Recognition model was applied that in turn aids in minimizing both the false positive and false negative rate significantly with minimal training time. Obtained encoded results were fed into the Differentiable Binary Position Visual Reinforced Large Language based Decoder to perform fine-grained process for robust scene text detection with improved accuracy. The quantitative analysis and validation confirm that the proposed MDDB-ViT method was better than state-of-the-art methods, in terms of PSNR, training time, false positive rate, false negative rate and STR accuracy

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Centre for Intelligent Cloud Computing, Centre of Excellence for Advanced Cloud, Multimedia University, Malaysia, for their valuable guidance, support, and encouragement throughout the implementation of this research and Post-doc research work. The authors also gratefully acknowledge the Global Academy of Technology. The authors further extend their sincere appreciation to all individuals and institutions who contributed directly or indirectly to the successful completion of this research through their guidance, constructive discussions, encouragement, and support.

AUTHOR CONTRIBUTIONS

Dr. Vimuktha Evangeleen Salis: Involved in concept formation, formal analysis, problem formulation, methodology, implementation, result discussion, rough manuscript preparation.

Dr. Prof. Md Shohel Sayeed: Research supervision and revision of manuscript for scientific and intellectual contented.

Dr. Andrews Samraj: Research supervision, editing and correction of manuscript.

Dr. Vineetha Edwina Jathanna: Software analysis, performance evaluation of MDDB-ViT and result interpretation, editing and correction of manuscript.

Prof. Sharmila Chidaravalli: Conceptualization, Editing and correction of manuscript

CONFLICT OF INTEREST

The authors declare that they have no known financial or non-financial conflicts of interest that could have influenced the research, its methodology, results, interpretation, or publication of this manuscript.

ETHICS STATEMENT

The study was conducted in accordance with applicable ethical guidelines, with appropriate ethical approval and informed consent obtained from all authors, while ensuring the confidentiality and privacy of their data.

DATA AVAILABILITY STATEMENT

All data supporting the results presented in the study can be obtained from the corresponding author through reasonable request.

REFERENCES

- [1] Zhang, Q., Liu, F., Song, W., 2025. IMTLM-Net: Improved Multi-Task Transformer Based on Localization Mechanism Network for Handwritten English Text Recognition. *Complex & Intelligent Systems*, 11(1), 125. <https://doi.org/10.1007/s40747-024-01713-8>.
- [2] Zhang, N., Li, C., Wang, Z., Ma, J., Feng, Z., 2025. APViT: ViT with Adaptive Patches for Scene Text Recognition. *Discover Applied Sciences* 7, 245. <https://doi.org/10.1007/s42452-025-06570-9>.
- [3] Hu, Y., Dong, B., Huang, K., Ding, L., Wang, W., Huang, X., Wang, Q.-F., 2024. Scene Text Recognition via Dual-Path Network with Shape-Driven Attention Alignment. *ACM Transactions on Multimedia Computing, Communications and Applications* 20(4), 107. <https://doi.org/10.1145/3633517>.
- [4] Gao, X., Pang, Y., Liu, Y., Han, M., Yu, J., Wang, W., Chen, Y., 2024. Multimodal Visual-Semantic Representations Learning for Scene Text Recognition. *ACM Transactions on Multimedia Computing, Communications and Applications* 20(7), 1–18. <https://doi.org/10.1145/3646551>.
- [5] Afkari-Fahandari, A., Shabaninia, E., Asadi-Zeydabadi, F., Nezamabadi-Pour, H., 2025. A Comprehensive Survey of Transformers in Text Recognition: Techniques, Challenges, and Future Directions. *ACM Computing Surveys* 58(5), 111, 1–42. <https://doi.org/10.1145/3771273>.
- [6] Liu, W., Wu, Q., Hu, Y., Li, Y., 2025. SELECT: Detecting Label Errors in Real-World Scene Text Data. In: *Proceedings of the 7th ACM International Conference on Multimedia in Asia (MMAsia '25)*, Kuala Lumpur, Malaysia, 85, 1–7. <https://doi.org/10.1145/3743093.3771031>.
- [7] Ozkan, S., Maracani, A., Ozay, M., Kim, H., Cho, S., Noh, E., Min, J., Cho, J.M., 2025. Efficient and Accurate Scene Text Recognition with Cascaded-Transformers. In: *Proceedings of the 16th ACM Multimedia Systems Conference (MMSys '25)*, <https://doi.org/10.1145/3712676.3719267>.
- [8] Albalawi, B.M., Jamal, A.T., Al Khuzayem, L.A., Alsaedi, O.A., 2024. An End-to-End Scene Text Recognition for Bilingual Text. *Big Data and Cognitive Computing* 8(9), 117. <https://doi.org/10.3390/bdcc8090117>.
- [9] Huang, B., Bai, A., Wu, Y., Yang, C., Sun, H., 2024. DB-EAC and LSTR: DBnet Based Seal Text Detection and Lightweight Seal Text Recognition. *PLOS ONE* 19(5), e0301862. <https://doi.org/10.1371/journal.pone.0301862>.
- [10] Ademola, O.A., Petlenkov, E., Leier, M., 2023. Resource-Aware Scene Text Recognition Using Learned Features, Quantization, and Contour-Based Character Extraction. *IEEE Access* 11, 56865–56874. <https://doi.org/10.1109/ACCESS.2023.3283931>.
- [11] Fanjie, K., Yaqi, L., Miaomiao, X., Silamu, W., Yanbing, L., 2023. SUST and RUST: Two Datasets for Uyghur Scene Text Recognition. *IEEE Access* 11, 126209–126220. <https://doi.org/10.1109/ACCESS.2023.3331213>.
- [12] Mahadshetti, R., Lee, G.-S., Choi, D.-J., 2023. RMFPN: End-to-End Scene Text Recognition Using Multi-Feature Pyramid Network. *IEEE Access* 11, 61892–61900. <https://doi.org/10.1109/ACCESS.2023.3280547>.
- [13] Khalid, S., Shah, J.H., Sharif, M., Dahan, F., Saleem, R., Masood, A., 2024. A Robust Intelligent System for Text-Based Traffic Signs Detection and Recognition in Challenging Weather Conditions. *IEEE Access* 12, 78261–78274. <https://doi.org/10.1109/ACCESS.2024.3401044>.
- [14] Yang, X., Silamu, W., Xu, M., Li, Y., 2023. Display-Semantic Transformer for Scene Text Recognition. *Sensors* 23(19), 8159. <https://doi.org/10.3390/s23198159>.
- [15] Hamza, A., Ren, S., Saeed, U., 2024. ET-Network: A Novel Efficient Transformer Deep Learning Model for Automated Urdu Handwritten Text Recognition. *PLOS ONE* 19(5), e0302590. <https://doi.org/10.1371/journal.pone.0302590>.

- [16] Zhao, X., Xu, M., Silamu, W., Li, Y., 2024. CLIP-Llama: A New Approach for Scene Text Recognition with a Pre-Trained Vision-Language Model and a Pre-Trained Language Model. *Sensors* 24(22), 7371. <https://doi.org/10.3390/s24227371>.
- [17] Yue, H., Huang, Y., Vong, C.-M., Jin, Y., Zeng, Z., Yu, M., Chen, C., 2023. NRSTRNet: A Novel Network for Noise-Robust Scene Text Recognition. *International Journal of Computational Intelligence Systems* 16(1), 5. <https://doi.org/10.1007/s44196-023-00181-1>.
- [18] Xu, M., Zhang, J., Xu, L., Silamu, W., Li, Y., 2025. Correlation-Guided Decoding Strategy for Low-Resource Uyghur Scene Text Recognition. *Complex & Intelligent Systems* 11(1), 57. <https://doi.org/10.1007/s40747-024-01689-5>.
- [19] Zhang, W., Guo, Y., Niu, L., Li, P., Wan, Z., Shao, F., Nian, C., Farrukh, F.U.D., Zhang, D., Zhang, C., Li, Q., Zhang, J., 2024. LP-SLAM: Language-Perceptive RGB-D SLAM Framework Exploiting Large Language Model. *Complex & Intelligent Systems* 10, 5391–5409. <https://doi.org/10.1007/s40747-024-01408-0>.
- [20] Murugan, B., Visalakshi, P., 2025. GAN Augmented Hybrid Transformer Network (GHTNet) for Ancient Tamil Stone Inscription Recognition. *npj Heritage Science* 13, 604. <https://doi.org/10.1038/s40494-025-02097-9>.
- [21] Hassan, B.M., Gamel, S.A., Talaat, F.M., 2025. Attention Based Unified Architecture for Arabic Text Detection on Traffic Panels to Advance Autonomous Navigation in Natural Scenes. *Scientific Reports* 15, 34138. <https://doi.org/10.1038/s41598-025-04326-4>.
- [22] Bhattacharya, A., Basu, S., Chakraborti, T., 2025. DEPP: Dictionary Embedded Probabilistic Priors for Scene Text Image Super-Resolution. *Neural Computing and Applications* 37, 21259–21273. <https://doi.org/10.1007/s00521-025-11441-w>.
- [23] Zhao, M., Li, Y., Zhang, C., Du, Q., Peng, S., 2024. SVTR-SRNet: A Deep Learning Model for Scene Text Recognition via SVTR Framework and Spatial Reduction Mechanism. *Electronics* 13(23), 4756. <https://doi.org/10.3390/electronics13234756>.
- [24] 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>.
- [25] Yu, W., Ibrayim, M., Hamdulla, A., 2023. Scene Text Recognition Based on Improved CRNN. *Information* 14(7), 369. <https://doi.org/10.3390/info14070369>.
- [26] Buoy, R., Iwamura, M., Srun, S., Kise, K., 2023. ViTSTR-Transducer: Cross-Attention-Free Vision Transformer Transducer for Scene Text Recognition. *Journal of Imaging* 9(12), 276. <https://doi.org/10.3390/jimaging9120276>.
- [27] Alrasheed, H., Alghihab, A., Pentland, A., Alghowinem, S., 2025. Evaluating the Capacity of Large Language Models to Interpret Emotions in Images. *PLOS ONE* 20(6), e0324127. <https://doi.org/10.1371/journal.pone.0324127>.
- [28] Wang, Q., Liu, J., Zhu, Z., Deng, C., 2022. MA-CharNet: Multi-Angle Fusion Character Recognition Network. *PLOS ONE*, 17(8), e0272601. <https://doi.org/10.1371/journal.pone.0272601>.
- [29] Kaggle, n.d. MLT-19 OCR Dataset. <https://www.kaggle.com/datasets/zubairlibhutto/mlt-19-ocr-dataset>.
- [30] Huang, M., Peng, D., Li, H., Peng, Z., Liu, C., Lin, D., Liu, Y., Bai, X., Jin, L., 2025. SwinTextSpotter v2: Towards Better Synergy for Scene Text Spotting. *International Journal of Computer Vision*, 133, 5281–5301. <https://doi.org/10.1007/s11263-025-02428-0>.
- [31] Feng, Y., Li, C., 2025. Towards Robust Scene Text Recognition: A Dual Correction Mechanism with Deformable Alignment. *Electronics* 14(19), 3968. <https://doi.org/10.3390/electronics14193968>.
- [32] Ali, S., Gao, T., Zhu, Z., Hussain, Z., Ahmed, M., 2026. Mamba-STR: Efficient Context-Aware Scene Text Recognition Framework with Selective State Space Modeling. *Information Sciences*, 741, 123242. <https://doi.org/10.1016/j.ins.2026.123242>.
- [33] Sivasankaravel, V.S., Sreenivasan, S.R., Manoj, R., 2026. Rotation-Invariant Scene Text Extraction Using Transformer Networks. *International Journal of Machine Learning and Cybernetics* 17(5), 235. <https://doi.org/10.1007/s13042-026-03043-2>.
- [34] Chen, L., Lin, H., Chen, D., 2026. Character-Level Visual Guidance for Arbitrarily Shaped Scene Text Detection and Recognition. *Electronics*, 15(18), 4120. <https://doi.org/10.3390/electronics15184120>.