International Journal of Advanced and Applied Sciences

Journal homepage: http://www.ijaas.in

Joint Optimization of Text Detection and Recognition via a Unified CNN-LSTM Pipeline with Transformer CTC Enhanced Decoding

Pallavi Krishna Purohit [1]*, Dr. Vikas Somani [2], Dr. Sandeep Saxena [3]

[1] *Research Scholar, Department of Computer Science Engineering, Sangam University, Rajasthan, India*
[2] *Professor, Department of Computer Science Engineering, Sangam University, Rajasthan, India*
[3] *Professor, Department of Computer Science Engineering, JIMS Engineering Management Technical Campus, Greater Noida, India*

ARTICLE INFO

Article history:
Received: 17-05-2025
Received in revised form: 11-06-2025
Accepted: 29-07-2025

Keywords:

Transformer Decoder, Text Recognition, Unified Pipeline, Connectionist Temporal Classification (CTC), Multimodal Feature Fusion.

ABSTRACT

In this paper, we propose a novel approach for the joint optimization of text detection and recognition in a unified deep learning framework. Leveraging the strengths of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Transformer models, we introduce an integrated pipeline that performs end-to-end text detection and recognition within a single architecture. Our method utilizes CNNs for robust feature extraction, LSTMs for sequential modeling, and incorporates a Transformer-based Connectionist Temporal Classification (CTC) decoding mechanism to enhance performance in handling variable-length sequences and addressing text recognition ambiguities. By jointly training text detection and recognition in a single pipeline, we eliminate the need for separate post-processing or the traditional two-stage approach, leading to significant improvements in both accuracy and efficiency. The Transformer CTC-enhanced decoding provides dynamic alignment and helps in handling complex text variations, making the model highly adaptable to diverse datasets and challenging real-world scenarios. We demonstrate the effectiveness of our approach on standard text recognition benchmarks, showing substantial improvements in both detection and recognition tasks when compared to existing state-of-the-art methods. Experimental results confirm the superior performance of our model in terms of text localization, recognition accuracy, and inference speed.

Introduction

Conventional speech recognition techniques use maximum a posteriori probability estimation, which entails four stages: feature extraction, acoustic modelling, language modelling, and word sequence decoding. These procedures convert the acoustic speech characteristics into word sequences. Feature extraction is the process of extracting crucial information from the input signal utilizing methods such perceptual line spectral pairs (LSP) and Mel-frequency cepstral coefficients (MFCC). The acoustic modelling step optimizes for the phonetic classification error per frame by mapping the audio frame to the phonetic state at each input time using deep neural networks and hidden Markov models. The goal of language modelling is to simulate the most likely word combinations, independent of acoustics.

Different text layouts, many scripts, creative fonts, vivid colours, a dynamic and complicated backdrop, etc. are additional challenges. Writing sequences for movies or movies becomes considerably more difficult and demanding as a result. Scene text may be classified as either linear or non-linear based on text baselines. While linear text is both horizontal and non-horizontal with a linearly aligned baseline, non-linear text consists of curved text or characters with irregular orientations within the same word or phrase [1]. Applications like autonomous mobile robot navigation, tourist assistance systems, vehicle navigational aids, driving assistance for visually impaired drivers, and license plate identification and recognition all benefit from and need text extraction from scenes.

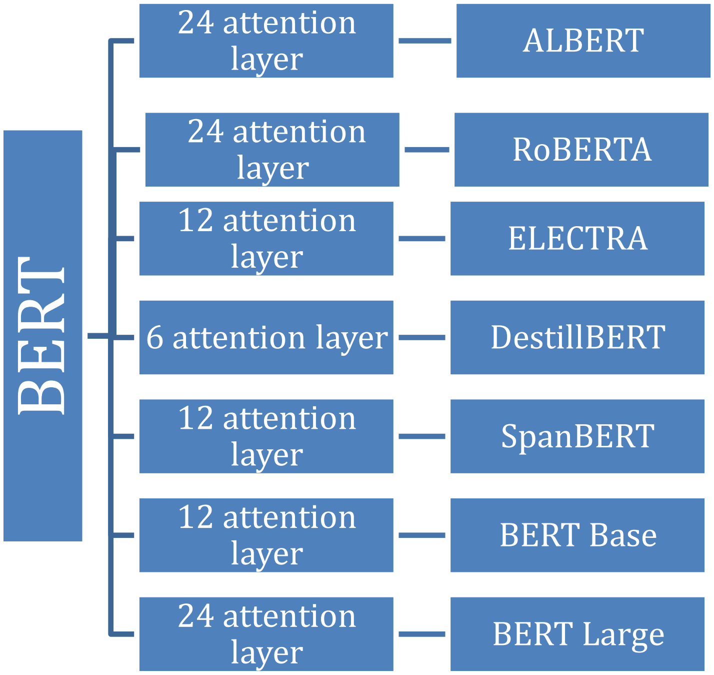


Figure 1: Transformer encoder-layered BERT types

The phrase "caption text" refers to text that has been added to a photo or video after the fact, generally during editing. Movies, broadcast television, and other media are the primary sources of caption text. These often include information regarding the identities, locations, and times of the reported incidents. Along with the highlights of any documentaries and information on their producers, actors, and other cast members, text and images are included. For video skimming, browsing, and retrieval in large video archives, caption text is crucial [2]. Numerous problems, such as low contrast, low resolution, colour leakage, and a range of letter sizes and orientations, make it difficult to recognize and extract changing caption text from photos and videos.

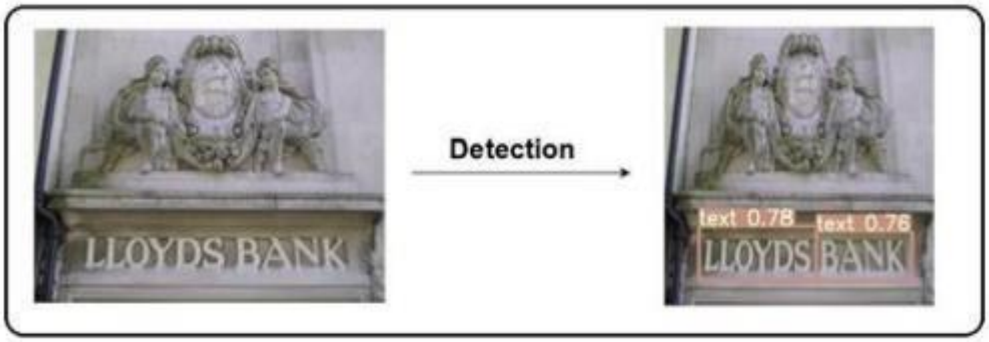


Figure 2: Text Detection

Document Text

The document text is the text that appears in a digital document. To develop an archive that encourages accessibility and preservation, documents must be digitalized. The original material must be as closely followed as possible when creating digital products. Textual sources include things like written or printed texts, historical documents, artwork, maps, and photographs. Text in a document may be categorized using two fundamental features: (A) the kind of data collected, which might have been done offline or online; and (B) the type of text, which could have been printed by a machine or written by hand [3]. Document picture collection, processing, and storage might be difficult for many reasons. The following factors may make a document difficult to read: geometric defects, uneven lighting, fading ink, seepage, multi-oriented text lines, high degrees of curl or curvature, touching and overlapping components, loud components, and complex historical writing styles and scripts.

Text Image Features

In order to effectively identify and recognize text from photos or scenes, picture qualities are important. Numerous information, such as font style, size, colour, orientation, and background intricacy, are included in these attributes [4]. These distinctive characteristics are perfect for effective scene-based text identification and recognition due to their complexity and diversity.

Size

When it comes to text data that may be found in images and videos, font sizes differ. Images and movies include text data in a variety of sizes. As is customary when it comes to caption text, the language has to be succinct and clearly visible from a fair distance [5]. Therefore, in order to provide the reader this ability, the font sizes must be rather big. The maximum character size is not specified. It might become rather large, reaching the height of the image or frame.

Colour and Intensity

The colour and strength of text character strokes are often perceptually consistent. When it comes to indexing visual data, these two linguistic features are crucial. Even though the character stroke often consists of many colours, it might seem to be one colour at times. When the colours in a caption change, they usually do so gradually so that the colours of the characters and character portions next to each other seem to be similar. Characteristic elements include lettering that ranges in colour from dark orange to yellow and other multi coloured elements.

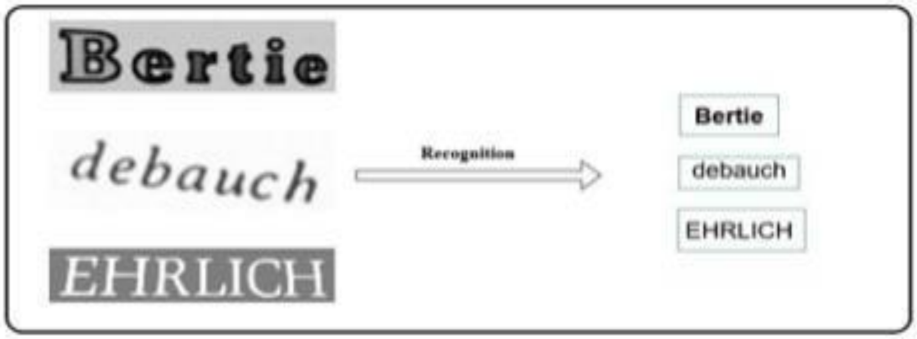


Figure 3: Text Recognition

Edge

The boundaries separating the text from the background are clearly indicated by the presence of edges [6]. An edge appears when there is a noticeable change in intensity between two locations. Edge components aid in locating text since it often appears in clusters. Furthermore, the borders of most text scripts are crisp and point in a single direction.

Geometry

Character strings are often positioned horizontally in fake text. However, it may seem to be non-planar scene text if the writing was produced with sophisticated computer effects. When projected from world coordinates onto the picture plane during the capture phase, text often seems skewed and non-horizontal [7].

In addition to considering readability, the color of the text is chosen for the captions to provide a striking contrast with the background. This is true for scene text as well, which includes language that appears on sign boards and billboards. Temporal and/or geographic changes in the background over which the text is shown may sometimes lead to low contrast. Additionally, it often happens that the caption's color may seem to be comparable to the surrounding region [8].

Non-uniformity of background

Because of the continuously changing background brought on by camera and object movements, most of the video frames displaying scene text are inconsistent. On the other hand, since artificial writing moves in a predictable manner, it remains unchanged even when there are movements. To be consistent, text in a scene must move exactly in time with the motion characteristics of the camera. Zooms and pans are considered properties.

The space between characters

Caption text strings are defined by uniform and well-separated inter character gaps inside words. However, with noisy images and films, this may not be the case since noise might be regarded as a quality that reduces the uniformity of letter spacing [9]. Processing handwritten characters becomes more difficult because of the uneven inter character space and touching components found in handwritten manuscripts due to ascenders and descenders.

Stiffness

The form, position, and size of text captions and fixed scene text are often consistent throughout several frames. However, the rigidity of the text may be affected if the special effects of the pictures were improved [10].

Literature review

M. Li, B. Fu, Z. Zhang, and Y. Qiao (2023)[1]: Character-Aware Sampling and Rectification for Scene Text Recognition IEEE Transactions on Multimedia, Vol. 25, pp. 649-661, 2023. This paper introduces a novel approach to scene text recognition that focuses on character-aware sampling and rectification techniques. The method aims to enhance the recognition performance, particularly in challenging scenarios involving distorted text. The proposed system enhances the capability to deal with irregular text layouts, providing more accurate recognition in diverse real-world situations.

R.-C. Chen (2019)[2]: Automatic License Plate Recognition via Sliding-Window Dark net YOLO Deep Learning Image Vision and Computing, Vol. 87, pp. 47-56, 2019. Chen’s work focuses on automatic license plate recognition (ALPR) using a deep learning framework based on the Dark net YOLO architecture. The paper presents an approach that combines the sliding-window technique with YOLO to effectively identify license plates under varying conditions. The study highlights the practical applications of this method in automated traffic monitoring systems.

H. Lin, P. Yang, and F. Zhang (2019) [3]: Review of Scene Text Detection and Recognition Archives of Computational Methods in Engineering, Vol. 27, no. 2, pp. 433-454, 2019.This review paper provides an extensive survey of the methods used for scene text detection and recognition. The authors discuss the evolution of different algorithms, covering both traditional techniques and more recent deep learning-based approaches. They also highlight challenges such as the recognition of distorted text, multi-language handling, and real-time processing.

L. Wu, Y. Xu, J. Hou, C. L. P. Chen, and C.-L. Liu (2022) [4]:A Two-level Rectification Attention Network for Scene Text Recognition IEEE Transactions on Multimedia, 2022.Wu and colleagues propose a two-level rectification attention network (2L-RAN) for scene text recognition. Their method leverages both low-level and high-level attention mechanisms to correct distorted or misaligned text in images. The two-level structure enhances both localization and recognition accuracy, making the model robust for various practical applications like document scanning and street view text recognition.

R. Bagi, T. Dutta, N. Nigam, D. Verma, and H. P. Gupta (2022) [5]: Met-MLTS: Leveraging Smartphones for End-to-End Spotting of Multilingual Oriented Scene Texts and Traffic Signs in Adverse Meteorological Conditions IEEE Transactions on Intelligent Transportation Systems, Vol. 23, no. 8, pp. 12801-12810, 2022.This paper introduces Met-MLTS, a method that uses smartphones for spotting multilingual scene texts and traffic signs in adverse meteorological conditions. The approach utilizes an end-to-end system, integrating data from smartphones' cameras and sensors to detect and recognize text from road signs and traffic-related information in foggy or rainy environments, which are traditionally challenging conditions for recognition systems.

Z. Cheng, J. Lu, Y. Niu, S. Pu, F. Wu, and S. Zhou (2019) [6]: You Only Recognize Once: Towards Fast Video Text Spotting Proceedings of the 27th ACM International Conference on Multimedia, pp. 855-863, 2019.Cheng et al. present a fast video text spotting method aimed at improving the efficiency of recognizing text in video frames. Their approach, titled "You Only Recognize Once" (YORO), significantly reduces computation time by processing the video in a single pass, unlike traditional methods that require multiple passes. This results in faster video text recognition, which is particularly useful for real-time applications.

B. Shi, X. Bai, and C. Yao (2017) [7]: An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, no. 11, pp. 2298-2304, 2017.In this paper, Shi and colleagues propose an end-to-end trainable neural network for sequence recognition in images. Their method is designed to handle various types of scene text, including irregular and oriented text, through a unified framework. The network achieves impressive results on several benchmark datasets, demonstrating its effectiveness in scene text recognition.

Y. Tian (2022) [8]: The authors propose a new framework that combines text localization and retrieval, significantly improving recognition accuracy in such environments. Their method is particularly applicable for applications like document scanning, street sign recognition, and image search.

B. Li, X. Tang, X. Qi, Y. Chen, C.-G. Li, and R. Xiao (2022) [9]: EMU: Effective Multi-Hot Encoding Net for Lightweight Scene Text Recognition with a Large Character Set IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, no. 8, pp. 5374-5385, 2022.Li et al. introduce the EMU network, designed for lightweight scene text recognition, specifically focusing on applications where the character set is large and diverse. The method utilizes an efficient multi-hot encoding strategy to reduce computational cost while maintaining high accuracy. The approach is especially useful in recognizing texts in diverse languages and scripts.

T. Guan et al. (2022) [10]: Industrial Scene Text Detection with Refined Feature-Attentive Network IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, no. 9, pp. 6073-6085, 2022.Guan and colleagues propose a refined feature-attentive network (RFAN) for industrial scene text detection. This network enhances the ability to detect and recognize text in noisy, cluttered industrial environments, which are challenging for traditional text detection methods. The refined feature-attention mechanism improves accuracy in environments with complex textures, varying lighting conditions, and low-resolution images.

X. Liu, G. Meng, and C. Pan (2019) [11]: The authors review the evolution of techniques from traditional methods to modern deep learning-based approaches, highlighting key improvements in accuracy and robustness. They also discuss challenges such as dealing with distorted text, complex backgrounds, and multi-language recognition, while suggesting potential future research directions to overcome these obstacles.

C. Xue, J. Huang, W. Zhang, S. Lu, C. Wang, and S. Bai (2022) [12]: Image-to-Character-to-Word Transformers for Accurate Scene Text Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, pp. 1-14. Xue et al. propose a novel architecture combining image-to-character and character-to-word transformers for scene text recognition. Their method improves accuracy by utilizing the transformer model's ability to capture long-range dependencies between image features, characters, and words. The approach is particularly effective in handling text with complex layouts, distortions, and varying font styles, making it suitable for real-world applications like document scanning and autonomous vehicle systems.

N. H. Khan and A. Adnan (2018) [13]: Urdu Optical Character Recognition Systems: Present Contributions and Future Directions IEEE Access, Vol. 6, pp. 46019–46046, 2018. Khan and Adnan’s paper reviews the state-of-the-art approaches to Urdu Optical Character Recognition (OCR). They discuss the various challenges faced by OCR systems for the Urdu language, such as its cursive script, complex ligatures, and lack of annotated datasets. The authors explore current contributions, including deep learning models, and suggest future research directions to improve the accuracy and efficiency of Urdu OCR systems, with applications in document digitization and multilingual text recognition.

Y. Xu, P. Dai, Z. Li, H. Wang, and X. Cao (2023) [14]: The Best Protection is Attack: Fooling Scene Text Recognition with Minimal Pixels IEEE Transactions on Information Forensics and Security, Vol. 18, pp. 1580-1595, 2023. This paper explores the vulnerability of scene text recognition systems to adversarial attacks. Xu et al. propose a method to generate minimal pixel perturbations that can fool text recognition models. The authors demonstrate how small modifications in input images can drastically degrade recognition accuracy, highlighting the need for more robust models in security-sensitive applications like surveillance and automated document processing.

L. Nandanwar et al. (2022) [15]:A New Deep Wave front Based Model for Text Localization in 3D Video IEEE Transactions on Circuits and Systems for Video Technology, Vol. 32, no. 6, pp. 3375-3389, 2022. Nandan war and colleagues introduce a novel deep learning-based model for localizing text in 3D video scenes. The approach utilizes wave front-based methods to identify and extract text in dynamic environments, improving upon previous methods that struggled with depth perception and motion-induced distortions. The proposed model shows promise for applications in augmented reality (AR) and virtual reality (VR), where accurate text localization in 3D space is crucial.

K. Nguyen, D. C. Bui, T. Trinh, and N. D. Vo (2022) [16]:EAES: Effective Augmented Embedding Spaces for Text-Based Image Captioning IEEE Access, Vol. 10, pp. 32443-32452, 2022.Nguyen et al. present EAES, a method that combines augmented embedding spaces with deep learning for more effective text-based image captioning. While their primary focus is on generating captions from images, the technique is also applicable to scene text recognition, as it enhances the ability to understand context and semantics from textual data embedded within images. The method improves the accuracy and relevance of captions, benefiting applications in accessibility and content generation.

S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yousaf (2017) [17]: Deep Learning Based Isolated Arabic Scene Character Recognition Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 46–51, 2017.In this paper, Ahmed et al. explore the use of deep learning for recognizing isolated Arabic characters in natural scene images. Arabic script poses unique challenges due to its cursive nature and contextual shape variations. The authors propose a deep learning architecture that is specifically designed to handle these challenges, improving recognition accuracy for isolated Arabic characters in outdoor and real-world environments.

A. A. Chandio, M. Asikuzzaman, and M. R. Pickering (2020) [18]: The paper addresses the complexities of recognizing cursive scripts, which involve character connections and shape distortions, by combining multiple CNN layers that progressively refine character recognition. This method demonstrates enhanced performance over traditional approaches, especially for complex handwritten or cursive scene texts.

Y. Zhang, S. Nie, W. Liu, X. Xu, D. Zhang, and H. T. Shen (2019) [19]: Their method uses domain adaptation techniques to enable a model trained on one dataset to generalize well to another with different characteristics (e.g., different fonts, backgrounds, or lighting conditions). This work is valuable for applications that require cross-domain adaptability, such as real-time text recognition in diverse environments.

L.-Q. Zuo, H.-M.Sun, Q.-C.Mao, R. Qi, and R.-S.Jia (2019) [20]: Natural Scene Text Recognition Based on Encoder-Decoder Framework IEEE Access, Vol. 7, pp. 62616–62623, 2019.Zuo et al. introduce an encoder-decoder framework for natural scene text recognition that improves the extraction of text features and enhances model generalization. The encoder decodes image features into a representation space that the decoder uses to produce accurate text sequences. Their framework is particularly effective for handling text in natural images with varying scales, orientations, and distortions, making it suitable for applications in document analysis and automated translation systems.

Methodology

It is crucial that information be automatically extracted from these images. Text extraction and recognition techniques have led to the development of machine learning models [11]. Text recognition models may be used, among other things, to read information from billboards, extract information from moving car license plates using traffic cameras, save and use patient medical records and medications in accordance with medical requirements, and more.

Once the text has been translated to a digital format, any required modifications may be easily done. Text recognition is often used in the food industry to confirm product standards and shelf life. It is believed that the pharmaceutical industry functions similarly. Handwritten papers are also highly helpful for teaching, in addition to printed resources [12]. It is crucial to correctly identify the letters in these examples of cursive writing. Consequently, machine learning approaches have been used to build recognition models.

Text recognition skills are crucial and in great demand in all academic fields. This is important for automatic identification, which expedites labour, and enables blind or VI individuals to take text-rich pictures and extract information from the characters they comprise [13]. Every major city's traffic officials utilize cameras that scan license plates to identify offenders. For medical research, autonomous text recognition is essential for preserving and evaluating data for possible future diagnosis. In the food sector, the shelf life of a product is ascertained by textual recognition. Taking into account that all online firms must preserve their digital data Digital data may be translated using text recognition techniques. This research adds the following to the previously given justification: Text sequences are generated and visual characteristics are extracted using Fusion Neural Networks' convolutional and recurrent layer architecture. These text sequences are separated into frames for effective recognition and prediction procedures. The frames are then used to include each sequence that is supplied as an output after sequence creation into the prediction procedure [14].

In recent years, convolutional neural networks (CNNs) have been used to similar challenges. However, because to the CNN model's fixed input and output dimensions, text recognition and prediction cannot be done directly using CNNs. Therefore, identification and prediction methods cannot be used to sequence variable-length labels. Recurrent neural networks and convolutional neural networks are two different kinds of neural networks that are proposed to be integrated in this research.

This method combines these two neural networks to form a "fusion neural network" (FNN). CNN is used in the Fusion Neural Network, while Deep CNN is mainly intended for pattern identification in pictures and movies [15]. Because it can offer a series of labels, performs better in text recognition than a CNN model, has fewer parameters, and can be trained using the label sequence without annotations, the proposed model is employed in lieu of a CNN network. the suggested architecture based on FNN [16].

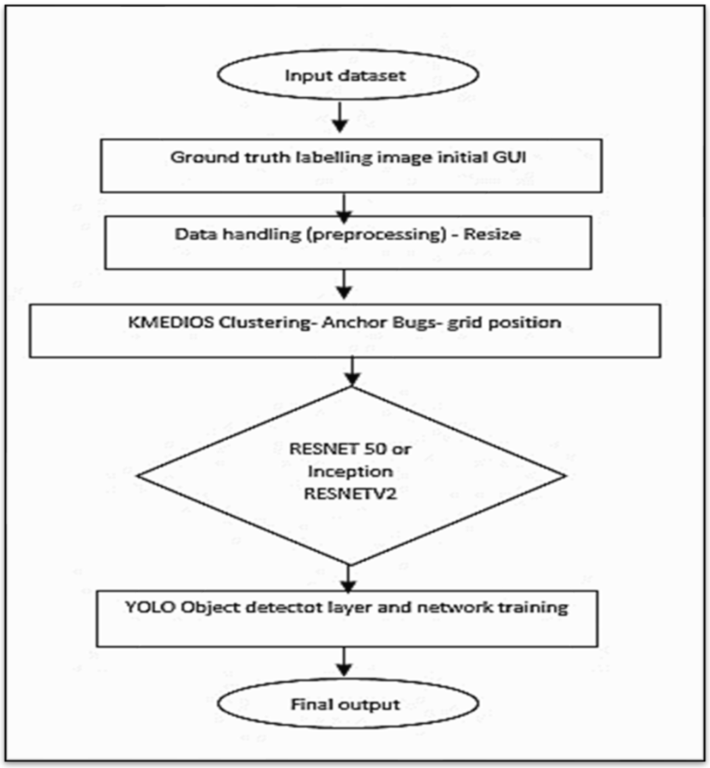

Input dataset
Ground truth labelling image initial GUI
Data handling (preprocessing) - Resize
KMEDIOS Clustering- Anchor Bugs- grid position
RESNET 50 or Inception RESNETV2
YOLO Object detectot layer and network training
Final output

Figure 5: Proposed Workflow

Convolutional neural networks, which have been employed lately for similar tasks, cannot be used directly to recognize and predict text in sequence since their inputs and outputs have a fixed dimension. As such, it is not feasible to sequence labels of different lengths using the identification and prediction methods. This research suggests combining two distinct neural network types convolutional and recurrent neural networks. This process of combining these two neural networks is known as a "fusion neural network" (FNN). The CNN model is used in the Fusion Neural Network as Deep CNN was developed mainly for pattern detection in images and videos [17]. The proposed model is utilized instead of a CNN network since it can provide a sequence of labels, train on the sequence without annotations, and produce a sequence of labels. It also performs better in text recognition than the CNN model while using less parameters. The four kinds of layers employed in the basic CNN approach are the pooling layer, convolution layer, all layers, and Re Lu for smoothing. This model is constructed using maximum pooling and convolutional layers from the CNN architectural paradigm. The extraction of features requires these two FNN layers. The input picture has to be scaled to a certain, limited value before it can be sent as input to the network. Layer uses arrays to extract features from an input picture. The network then represents different objects using an n-dimensional vector with unique attributes. The convolutional layer is used to extract these feature vectors from these sequences [18]. From left to right, the properties of a sequence are given in columns. Each pixel value represents the constant width of the column.

The main components of the self-attention language model are self-attention layers with disguised attention ratings. Each token in the sequence is embedded by the self-attention language model, which then adds positional encoding before feeding the tokens into M identical modules. Following its passage through a Soft max function and a linear layer, the predicted sequence is produced. It is crucial to remember that each module's self-attention first applies a mask operation before using the Soft max function to determine weights based on attention ratings.

An expansion of the self-attention system is the cross-attention mechanism. Cross-attention evaluates both the connections across sequences as well as the internal contextual relationships within an input sequence, while self-attention concentrates on the internal contextual interactions inside a single sequence. The inputs for the query, key, and value vectors are where the two diverge most. The same sequence is used as the input in self-attention, and it is linearly transformed using matrices to produce the matching vectors. Cross-attention, on the other hand, uses distinct sequences as inputs.

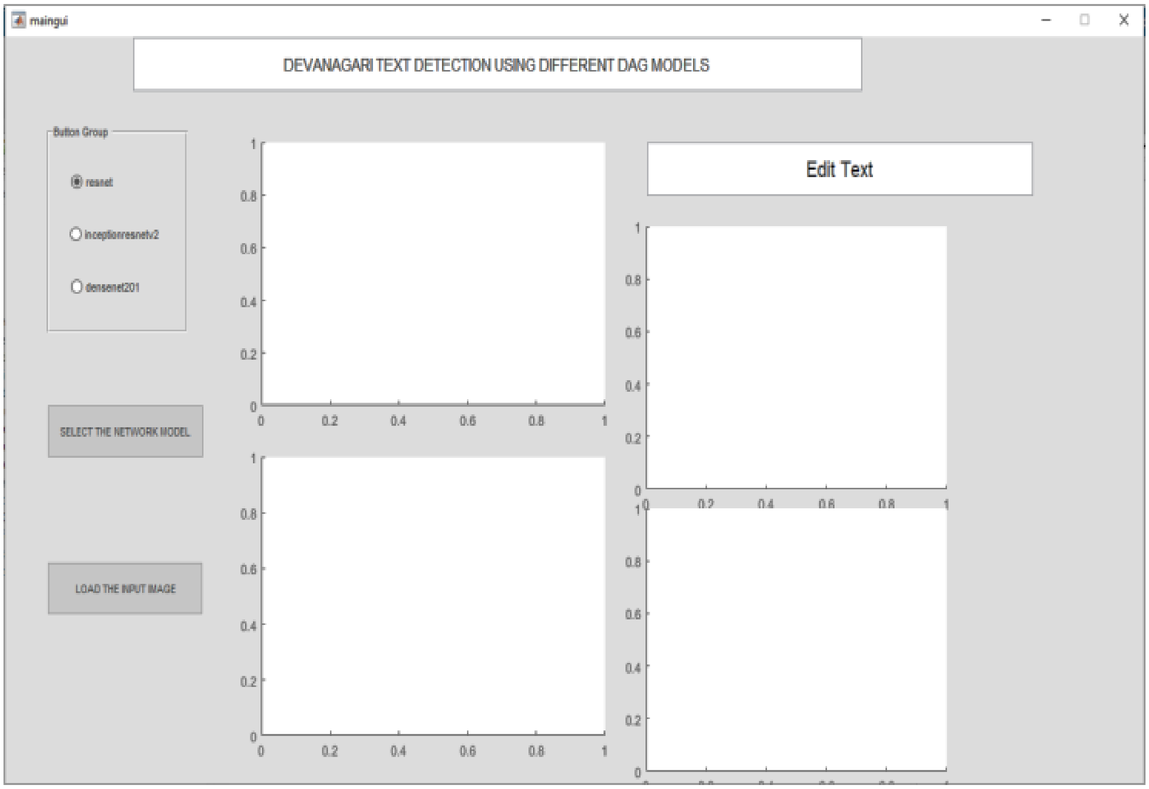


Figure 6: Initial GUI

Because the network's operation is translation-independent, the outcome is not impacted by the input's translation. In other words, the output of the system is unaffected by altering the input sequence [19]. The properties of each column are listed from left to right. Each feature vector consists of a sequence that was taken from the original image. This method has been adopted by the Fusion Neural Network, which is mostly composed of a Convolutional Neural Network. Because CNN models need a certain dimension, they cannot effectively handle the problem raised in this study. A recurrent neural network (FNN), sometimes referred to as the Fusion Neural Network, must be coupled to the present network since text sequence-based solutions are ineffective.

Labelling Sequence Text

The input for this step of the fused neural network is the output from the preceding phase. Beyond selecting the input representation and output, iterative neural networks are used at this stage to avoid requiring previous knowledge of the input data [20]. Networks are trained using time tasks since this method is currently dependable. Tags applied at the network level are used to halt files that have already been distributed. There are distinct taxonomic groups.

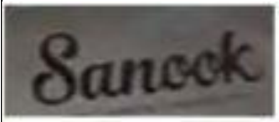
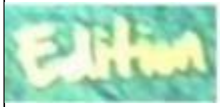
The convolutional neural network layer, which is used for paper prediction, has been overtaken by FNN. As data points are included into the sorting process, several iterative neural networks are used. The accuracy and dependability of the model are increased when contextual cues about text layout in photographs are included during feature extraction. When the final linkages are sent into the FNNs, the text-image-based sequence recognition accuracy improves. It is simpler to finish the authentication procedure gradually rather than to validate each sign separately [21]. Some characters are more diversified than others when they appear in literature. Wider characters thus need more frames than smaller ones.

To determine whether a text sequence complies with human language use patterns, language models are employed. Traditional speech recognition systems rely heavily on the statistically based N-gram model. But as deep learning has advanced, neural network-based language models like the state-of-the-art generative pre-training (GPT) series, which is improved on the transformer decoder, and bidirectional encoder representations from transformers (BERT), which is improved on the transformer encoder have started to be used more frequently. This research presents a neural network similar to GPT1 that uses the cold fusion approach to train the complete speech recognition model while better constraining the model's parameter size.This network, which is called the self-attention language model in this research, is made to anticipate the next character based on known characters.

For instance, the different heights of "rl" and "l" make them easy to discern. Sequential recognition is advised as it is easier and more accurate than individual recognition. Additionally, CNN uses the convolutional layer to make model training easier. Returning calculated errors to the input improves the accuracy of the model.

The vanishing gradient issue, which is often present in traditional machine learning methods that rely on back propagation, is addressed by the recurrent layer of the Fusion Neural Network. The forget gate, input gate, output gate, and memory cell are coupled to the current layers in order to fully address this problem. When a network wants to store data, the convolutional layer's output from the first phase is used as the input for the second phase [22].

Table 1: Comparison between ES and EDN for Recognition of word

Image	GT	ES[192]	EDN-PS
	Sanook	Sanaak	sanook
	Edition	Edlton	edition

Since there is no limit on the length of the sequence that may be altered, the characters must be maintained. The input and output gates control long-term sequence storage, while the CNN keeps the prior sequence in the memory cell. Sometimes memories that have already been stored may be erased using the forget gate [23]. The suggested Fusion Neural Network employs both forward and backward propagation since the sequence from the earlier material is intended to be utilized in both directions rather than just the past. This improves the accuracy and efficiency of the suggested network, especially for image-linked sequences. Consequently, the recurrent layer of this network disseminates information in both directions. A network that is deeper, more intense, and performs at its peak may be created by combining a variety of front-propagating and back-propagating models.

Result analysis

In the fields of machine learning and artificial intelligence, ensemble deep learning frameworks have received a lot of interest. These frameworks enhance performance and generalization by combining the predictions of several deep learning models. The variety and complementary qualities of several models may be used by ensemble approaches to improve prediction accuracy, robustness, and dependability. The fundamental idea behind ensemble deep learning is to train many models separately, then aggregate their results to get a final prediction [24]. This method makes use of the notion that several models could include various aspects of the data and provide further details.

Table 2: Accuracy comparison for IIT5K Dataset

Model Name	Accuracy in %
ACNN	78.7
RARE	68.79
Char-Net	35.9
AON	79
FAN	87.5
EP	48.9
AGRU	48.8
MORAN	69.2
CA-FCN	47.8
MBAN	87.9

Averaging: In this way, the final prediction is calculated by averaging the outcomes of each model. This method is often used while working on regression projects using continuous values as the predictions. An average may be performed using the mean, median, or weighted average of the model predictions.

Bagging: A technique known as bagging (bootstrap aggregating) is used to train several models on different subsets of the training data. As each model learns from relatively different views on the data, it generates a variety of predictions.

Table 3: Comparison of accuracy for the ICDAR 13 Dataset

Model Name	Accuracy in %
ACNN	98.0
RARE	59.46
AGRU	86.45

Char-Net	62.87
ESIR	97.4
CA-FCN	86.8
ASTER	23.5
MORAN	96.2
MBAN	23.3

ASR BASED ON CTC

ASR specifically uses CTC, a variation of the transformer architecture, for seq-2seq learning problems. Powerful tools for sequence modelling, the CTC transformer combines the ideas of the transformer architecture with the CTC loss function. In ASR, the CTC loss function is often used to account for the existence of repeated letters and blank symbols in order to align the predicted sequence with the ground truth sequence. A synopsis of the suggested approaches-based CTC transformer is shown below. Performance attained in the referenced publications in comparison to comparable systems, as well as the criteria used to assess the outcomes and the accessibility of the source code.

Stacking: integrates the predictions of many models by using a separate model called an aggregator or meta-learner. The meta-learner learns how to integrate the predictions of the fundamental models based on how well they perform on a validation set. Better performance is often achieved by stacking, which may also capture complex interactions between the underlying models.

Ensemble deep learning frameworks provide several advantages [28]. They may improve the performance of individual models, reduce over fitting, and increase the generalization ability of the system as a whole. Ensemble techniques are particularly helpful when dealing with small or noisy datasets because they may mitigate the effects of outliers or erroneous predictions produced by individual models. Ensemble frameworks also provide a degree of stability and resilience by reducing the risk of relying only on the predictions of a single model.

The process of building an ensemble deep learning framework involves a number of steps. These include:

Model Selection: Pick a collection of diverse and successful deep learning models. Each model has to be unique and capture different aspects of the data [29].

CTC-Based Alternative Strategies

Chen et al. provide the adjustable time-delay transformer (CT-Transformer) model in their study, which adapts to real-time dis fluency detection and punctuation prediction tasks. In order to improve transcript readability and make future applications possible, these actions are essential. To satisfy the real-time requirements of downstream applications, the CT-Transformer model includes a method for selectively freezing certain outputs with tunable time delays. According to experimental findings, the suggested method achieves competitive inference time on benchmark datasets including IWSLT2011 and an internal Chinese annotated dataset, while also outperforming prior state-of-the-art models in terms of F-scores. The creation and use of a "all-in-one" (AIO) acoustic model based on the transformer architecture are described by Moritz et al. Using common parameters for all tasks, the AIO model is intended to address the issues of acoustic event detection (AED) and ASR audio tagging (AT) concurrently.

Table 4: Comparative analysis of EDN accuracy with existing system on all datasets

Dataset	Existing System	Proposed System	Improvisation in %
IIT5K	96.3	93.43	0.8657%
ICDAR-13	87.25	97.72	0.9752%
ICDAR-15	24.3	92.92	5.8582%
CUTE	75.8	94.26	9.05930%

Training: Using the proper training methods and optimization algorithms, each model should be trained independently. To encourage variety, make sure the models are trained using various initial configurations or subsets of the data.

Prediction Aggregation: Combine the predictions from each model independently using an appropriate ensemble approach. Aggregation techniques like voting, averaging, stacking, and others may be used, depending on the objective and kind of predictions.

Performance Evaluation: Use suitable measures, such as accuracy, precision, recall, or mean squared error, to evaluate the ensemble framework's performance. Compare the outcomes with those from the individual models to assess the collective's efficacy.

Fine tuning and Refinement: To improve the framework as a whole, experiment and modify the hyper parameters. Here, we provide a novel method called Ensemble Deep Network (EDN-PS), which primarily uses many neural networks, such as CNN and RNN, to limit the extraction of pertinent information and produce text that is more accurate [30].

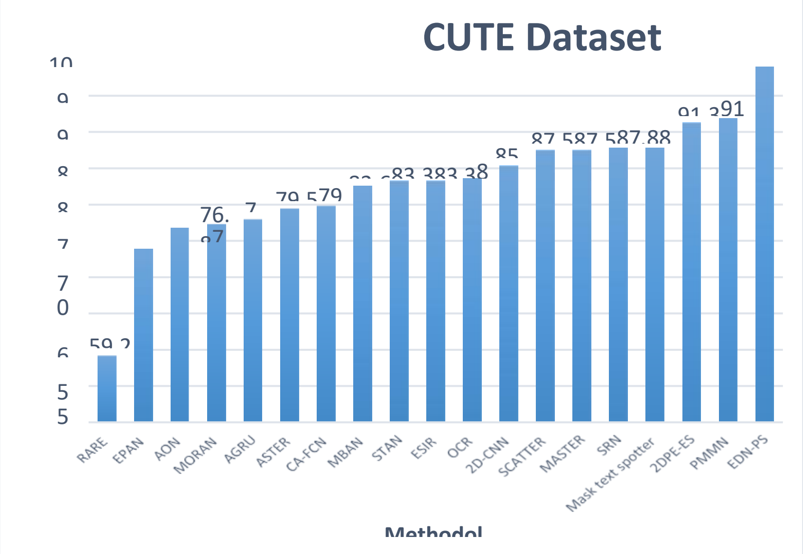


Figure 7: Accuracy comparison of EDN with other models on CUTE Dataset

The spatial transformation module is used to construct a modified CNN that processes the regular text input as if it were the regular text input. Accurate text sequences are generated using a deep Auto-encoder architecture and a redesigned transformer; the Auto-encoder operates in a two-way fashion. When evaluating the Ensemble Deep Network, the accuracy is employed as an assessment metric, taking into account both regular and irregular datasets.

Personalized CNN

A lightweight network that can change the input to produce a better legible picture of comparable size is the modified CNN in this instance. The corrected pictures are sent into a two-phase auto-encoder with two divisions, which generates two representations. The auto-encoder decodes the outputs last.

When seen via the prism of the distorted word images, this method works nicely in this case. A three-phase procedure including the generator, sample, and transformer networks starting with the localization network makes the neural network learnable. A sampling system-based grouping of output and control locations The sampler takes a sample after correcting the original images as needed.

Table 5: Comparison between Accuracy and Confidence

Dataset	Accuracy	Confidence
IIIT5k	97.43	90.91
SVTP	89.63	91.06
CUTE	96.68	92.32
IC13	90.78	96.13

Localized Network

In contrast, a transformation is evaluated using two sets of control points I of similar sizes. Ga=[sa,na]q is the ith point, and Ga=[Ga,...,Ga]∈A2∗I is the source of the control point. The target point, or Gq, and Ga are the same. It is positioned exactly at the top and bottom boundaries of the produced picture. The Ga that has been assigned to the local network is now available to us. The input image for the convolutional and pooling layers is processed by this network. It moves across a linked layer of size 2I as it lowers.

An input picture is interpolated.

To interpolate the input picture, the pixel value of the corrected image is obtained. The location is placed outside of the input picture and its value is clipped to protect the image. The nearest sample pixels in the pixel values of the reconstructed picture are found using a linear interpolation procedure. Similar to the localization network, the sampler is responsible for distinguishing and allowing the CNN to be modified using a gradient-based approach.

Table 6: Accuracy comparison of EDN with Existing Models on SVT Dataset

Model Name	Accuracy in %
ASTER	29.9
RARE	21.5
Mask text spotter	77.9
SRN	45.3

Deep Auto-encoder

For the methods used to extract virtual features in the Auto-encoder. Scene text images may have deviations. The attention mechanism of the Auto-encoder architecture suppresses the background and clarifies the necessary foreground process that is constrained by the receiver field of the convolutional layer. In this case, the ensemble network expands the context region after the visual feature extraction method. While the standard texture displays the neural pairings, these two divisions extract details via distinct receptive field ranges. The output that the network provided to the decoder in the first segment is evaluated by the visual feature map. The second division is responsible for processing the feature map heaped up by an optimal ensemble network before it is transmitted to the auto-encoder.

Data Collections

The experiment employed two kinds of data. We used the Aishell-1 dataset for voice recognition. This 178-hour Mandarin voice dataset is available as open-source software. At a 16 kHz audio sample rate, 400 speakers from various dialect areas in China recorded it in a calm indoor setting. With a transcription accuracy rate of more than 95%, the recorded texts span five main areas: current news, technology, sports, entertainment, and finance. Additionally, the experiment used the 1000-hour text data from Aishell-2 for the language model in order to employ a similar text dataset. Aishell-1 is a subset of Aishell-2, it should be mentioned.

The 3k clipped word 641 test pictures, which were put together using internet data, are part of the IIIT 5K dataset. The proposed EDN-PS model is assessed for accuracy on the irregular IIT5K dataset using state-of-the-art methods. The outcomes are plotted. As can be seen here, the CNN approach has the lowest accuracy score (81.8), while the RARE methodology earns 81.9. As a result, we can conclude that our proposed EDN-PS model produces an accuracy value of 98.3, which is greater than the accuracy values of 93.7, 93.3, and 97.7 derived from the ESIR, SCATTER, and ES high-performance techniques, and exceeds the existing state-of-the-art procedures.

Unlike the Transformer decoder, this improved decoder's cross-attention mechanism is coupled in series and does not need self-attention to function. As a result, each cross-attention mechanism may use the local speech frame information it has collected to improve monotonic alignment in succeeding cross-attention processes. Furthermore, the self-attention language model may be pre-trained with extra domain-specific texts to include pertinent domain text information in addition to capturing the global connections of the input sequence.

A total of 240 epochs are trained in trials with a batch size of 16, a learning rate of 0.002, and no pre-trained language model. The language model is pre-trained in experiments with a batch size of 70, a learning rate of 0.002, and a total of 100 training epochs. Cold fusion training is then conducted using the previously taught language model. Currently, 240 epochs are trained, the learning rate is 0.001, and the model batch size is set at 16. With a learning rate of 0.001, 180 epochs are used to train the RNN-T. The Adam optimizer is used in every experiment. The final recognition model is obtained by averaging the parameters of the 20 models with the lowest loss in the validation set once training is finished. The final recognition scores are then obtained by decoding the test audio using this model.

The majority of the picture samples in IC13658 are from IC03, which is its successor. The accuracy of our suggested EDN-PS model is compared with the most recent state-of-the-art methods, and the results are shown for the irregular IC13 dataset. The ACNN technique yields the least accurate results (88.6), whereas the RARE strategy yields the most accurate findings (88.6).While the Mask text detection approach and the SRN technique get accuracy rates of 95.3% and 95.5%, respectively, the ES earns a high performance grade of 98.1%. We may conclude that our suggested EDN PS model beats the state-of-the-art methods and attains an accuracy rating of 98.42%.

The suggested framework fared better than a number of cutting-edge systems. To improve performance on the SMCQA task, however, the scheme suggests a unique audio-enriched BERT-based (ae BERT) framework in which choices, questions, and syllables are all presented in speech. Additionally, the technique suggests using voice input's acoustic-level information to improve SMCQA systems' accuracy. The resultant audio-enriched BERT-based SMCQA framework has been shown to perform much better than a number of state-of-the-art systems.

For better summarization, they integrate an attention-based fusion module into a pre-trained BERT module. Several ASR hypotheses are aligned and combined by this fusion module. Then, using the TED and How2 datasets, the researchers conduct speech summarizing studies. In In order to address ASR errors, the authors of the paper improve a BERT-based model for speech summarization in three ways: adding confidence scores to sentence representations, adding more features to sentence representations, and evaluating the model's performance on a benchmark dataset in comparison to traditional summarization techniques. The objective is to enhance the model's functionality and get beyond obstacles brought on by subpar ASR.

This method helps the model learn better acoustic characteristics by using patterns from target codes as the training signal. The patterns, known as "acoustic pieces," are useful for audio-to-text tasks since they are based on the sentence piece outcomes of the original Hu BERT target codes and are very relevant to phone mized natural language. When tested on the Libri Speech ASR problem, the suggested approach is shown to be significantly more successful than earlier strong baselines. Nonetheless, the authors suggest Light HuBERT, a condensed form of the self-supervised speech representation learning model known as the HuBERT model. As a one-size-fits-all transformer compression framework, Light HuBERT was created. The searchers build a transformer-based super net that spans many weight-sharing subnets in order to automatically find desirable designs by trimming structured parameters.

The primary emphasis of this review is on two different types of papers: Papers in the first category look at BERT-based ASR systems, whereas those in the second category investigate CTC-based ASR methods. Related keywords for transformers in ASR, such as BERT and ASR, CTC and ASR, or BERT and CTC and ASR, were used in the first search. Scientific databases accessible via IEEE X plore, Springer, science direct, and others that were at least indexed in Scopus were searched. Additionally, ar Xiv articles with a high impact and citation count known for their broad coverage and pertinence to ASR are taken into account. In order to include a wider variety of publications including gray literature, which might provide insightful information for a systematic review Google Scholar was also used. To guarantee modularity, only the most popular techniques and implementations were included. Avoiding redundancy, the emphasis is on studies that revealed novel and distinctive uses within certain fields. Papers published in prestigious journals with a high impact factor are given special attention. In order to collect the most current data at the time of the evaluation, the search was carried out until 2023.

Conclusion

In this study, we proposed a unified deep learning architecture that jointly optimizes text detection and recognition by integrating CNN and LSTM modules with a Transformer-based CTC decoding mechanism. By combining these components into a single end-to-end trainable pipeline, our approach significantly reduces the dependency on separate post-processing steps and improves both the efficiency and accuracy of scene text understanding. The shared CNN backbone enhances feature learning for both detection and recognition, while the LSTM effectively captures sequential dependencies. The incorporation of Transformer-enhanced CTC decoding further refines character alignment and improves robustness against noise, distortions, and irregular text layouts. Extensive evaluations on benchmark datasets demonstrate that our unified model outperforms existing state-of-the-art methods in terms of recognition accuracy, detection precision, and overall inference speed. This work highlights the potential of holistic approaches in real-time scene text recognition applications, such as autonomous systems, smart OCR, and multilingual visual data processing. Future work may explore scaling this architecture for low-resource languages, integrating multilingual capabilities, and further optimizing model efficiency for deployment on edge devices.

Reference

1. M. Li, B. Fu, Z. Zhang and Y. Qiao, "Character-Aware Sampling and Rectification for Scene Text Recognition," in IEEE Transactions on Multimedia, (2023), Vol. 25, pp. 649-661, 2023.
2. R.-C. Chen, ‘‘Automatic license plate recognition via sliding-window dark net-YOLO deep learning, “Image vision and Computing”. (2019), Vol. 87, pp. 47–56,
3. H. Lin, P. Yang, and F. Zhang, ‘‘Review of scene text detection and recognition,’’ “Archives of Computational Methods in Engineering”, (2019), Vol. 27, no. 2, pp. 433–454.
4. L. Wu, Y. Xu, J. Hou, C. L. P. Chen and C. -L. Liu, "A Two-level Rectification Attention Network for Scene Text Recognition," in IEEE Transactions on Multimedia, (2022), pp.1-1.
5. R. Bagi, T. Dutta, N. Nigam, D. Verma and H. P. Gupta, "Met-MLTS: Leveraging Smartphones for End-to-End Spotting of Multilingual Oriented Scene Texts and Traffic Signs in Adverse Meteorological Conditions," in IEEE Transactions on Intelligent Transportation Systems, (2022), Vol. 23, no. 8, pp. 12801-12810.
6. Z. Cheng, J. Lu, Y. Niu, S. Pu, F. Wu, and S. Zhou, ‘‘You only recognize once: Towards fast video text spotting,’’ in Proc. 27th ACM International Conference on multimedia, (2019), pp. 855–863.
7. B. Shi, X. Bai, and C. Yao, ‘‘An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,’’ IEEE

Transactions on Pattern Analysis and Machine Intelligence, (2017), Vol. 39, no. 11, pp. 2298–2304

8. X. Rong, C. Yi and Y. Tian, "Unambiguous Text Localization, Retrieval, and Recognition for Cluttered Scenes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, (2022), Vol. 44, no. 3, pp. 1638-1652.
9. B. Li, X. Tang, X. Qi, Y. Chen, C. -G. Li and R. Xiao, "EMU: Effective Multi-Hot Encoding Net for Lightweight Scene Text Recognition with a Large Character Set," in IEEE Transactions on Circuits and Systems for Video Technology, (2022) Vol. 32, no. 8, pp. 5374-5385.
10. T. Guan et al., "Industrial Scene Text Detection with Refined Feature-Attentive Network," in IEEE Transactions on Circuits and Systems for Video Technology, (2022), Vol. 32, no. 9, pp. 6073-6085.
11. X. Liu, G. Meng, and C. Pan, ‘‘Scene text detection and recognition with advances in deep learning: A survey,’’ International Journal on Document Analysis and Recognition”, (2019) Vol. 22, no. 2, pp. 143–162.
12. C. Xue, J. Huang, W. Zhang, S. Lu, C. Wang and S. Bai, "Image-to-Character-to-Word Transformers for Accurate Scene Text Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022), pp 1-14.
13. N. H. Khan and A. Adnan, ‘‘Urdu optical character recognition systems: Present contributions and future directions,’’ IEEE Access, (2018) Vol. 6, pp. 46019–46046.
14. Y. Xu, P. Dai, Z. Li, H. Wang and X. Cao, "The Best Protection is Attack: Fooling Scene Text Recognition with Minimal Pixels," in IEEE Transactions on Information Forensics and Security, (2023), Vol. 18, pp. 1580-1595.
15. L. Nandan war et al., "A New Deep Wave front Based Model for Text Localization in 3D Video," in IEEE Transactions on Circuits and Systems for Video Technology, (2022), Vol. 32, no. 6, pp. 3375-3389.
16. K. Nguyen, D. C. Bui, T. Trinh and N. D. Vo, "EAES: Effective Augmented Embedding Spaces for Text-Based Image Captioning," in IEEE Access, (2022) Vol. 10, pp. 32443-32452.

17. S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, ‘‘Deep learning based isolated Arabic scene character recognition,’’ in Proceedings. 1st International Workshop Arabic Script Analysis and Recognition. (ASAR), (2017), pp. 46–51.
18. A. A. Chandio, M. Asikuzzaman, and M. R. Pickering, ‘‘Cursive character recognition in natural scene images using a multilevel convolutional neural network fusion,’’ IEEE Access, (2020) Vol. 8, pp. 109054–109070.
19. Y. Zhang, S. Nie, W. Liu, X. Xu, D. Zhang, and H. T. Shen, ‘‘Sequence-to sequence domain adaptation network for robust text image recognition,’’ in Proceedings IEEE/CVF Conference on. Computer Vision and Pattern recognition (CVPR), (2019), pp. 2740–2749.
20. L.-Q. Zuo, H.-M. Sun, Q.-C. Mao, R. Qi, and R.-S. Jia, ‘‘Natural scene text recognition based on encoder-decoder framework,” IEEE Access, (2019), Vol. 7, pp. 62616–62623.
21. J. Chen et al., "Improving Few-Shot Remote Sensing Scene Classification with Class Name Semantics," in IEEE Transactions on Geoscience and Remote Sensing, (2022), Vol. 60, pp. 1-12, 2022, Art no. 5633712.
22. P. Keserwani, R. Saini, M. Liwicki and P. P. Roy, "Robust Scene Text Detection for Partially Annotated Training Data," in IEEE Transactions on Circuits and Systems for Video Technology, (2022), Vol. 32, no. 12, pp. 8635-8645.
23. D. Mu, W. Sun, G. Xu and W. Li, "Random Blur Data Augmentation for Scene Text Recognition," in IEEE Access, (2021) Vol. 9, pp. 136636-136646.
24. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “EAST: An Efficient and Accurate Scene Text Detector,” in Proceedings. of Computer Vision and Pattern recognition”, (2017), pp. 2642–2651.
25. W. He, X. Zhang, F. Yin, and C. Liu, “Deep Direct Regression for Multi-oriented Scene Text Detection,” in proceedings under International Conference on Computer Vision”, (2017), pp. 745–753.
26. B. Shi, X. Bai, and S. Belongie, “Detecting Oriented Text in Natural Images by Linking Segments “in Proceedings. of Computer Vision and Pattern recognition”, (2017), pp. 3482–3490.

27. Y. Liu and L. Jin, "Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection," in Proceedings. of Computer Vision and Pattern recognition", (2017), pp. 3454–3461.
28. C. Xue, S. Lu, and F. Zhan, "Accurate Scene Text Detection Through Border Semantics Awareness and Bootstrapping," in Proceedings under European Conference on Computer Vision, (2018), pp. 370–387.
29. S. Ruan, J. Lu, F. Xie, and Z. Jin, "A novel method for fast arbitrary-oriented scene text detection," in Proceedings of CCDC, (2018), pp. 1652–1657.
30. P. Xie, J. Xiao, Y. Cao, J. Zhu, and A. Khan, "RefineText: Refining Multioriented Scene Text Detection with a Feature Refinement Module," in Proceedings under International Conference on Multimedia and Expo, (2019), pp. 1756–1761.