**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 15, No. 1, February 2026, pp. 831~840
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp831-840

# Adaptive deformable feature augmentation and refinement network for scene text detection and recognition

**Ratnamala S. Patil, Geeta Hanji, Rakesh Hudud**
Department of Electronics and Communication, Poojya Doddappa Appa College of Engineering, Kalaburagi, India

**Article Info**

***Article history:***

Received Aug 17, 2024
Revised Dec 26, 2025
Accepted Jan 10, 2026

***Keywords:***

ADFARN
Deep feature augmentation module
Deep feature refinement
Scene text recognition
Text recognition

## ABSTRACT

Scene text recognition (STR) is the task of detecting and identifying text within images captured from natural scenes, a challenging process due to variations in text appearance, orientation, and background complexity. The proposed methodology, adaptive deformable feature augmentation and refinement network (ADFARN), is designed to address these challenges by combining deformable convolutional networks for robust enhanced feature extraction with a novel deep feature refinement (FRE) that leverages refinement for precise text localization. This approach enhances the differentiation between text and background, significantly improving recognition accuracy. The ADFARN methodology includes a comprehensive process of feature extraction, deep feature augmentation module (DFAM), and the generation of score and threshold maps through differentiable binarization. The adaptive nature of the model allows it to handle low-resolution and partially occluded text effectively, further increasing its robustness. Additionally, the proposed method aligns visual and textual features seamlessly. Extensive performance evaluation on the common objects in context (COCO)-Text dataset demonstrates that ADFARN outperforms existing state-of-the-art methods in terms of precision, recall, and F1-scores, establishing it as a highly effective solution for STR in real-world applications.

***Corresponding Author:***

Ratnamala S. Patil
Department of Electronics and Communication, Poojya Doddappa Appa College of Engineering
Kalaburagi, India
Email: ratnamala_12@rediffmail.com

## 1. INTRODUCTION

Texts play an important role in the cultural transmission process since they are a storehouse of human wisdom. Spoken language has become a more powerful medium for the development of human civilization because it has broken down boundaries related to time and space. Nowadays, a large amount of textual data is stored digitally as documents, movies, or photos. As a result, using computer technology for scene text picture detection and end-to-end identification is more essential than ever [1]. Therefore, it is more important than ever to use computer technology for end-to-end identification and scene text picture detection [1]. Applications for text recognition can be found in many different areas, such as assistive technology for the blind, driving assistance, and handwriting recognition. Scanned document recognition and scene text recognition (STR) are the two main subcategories of text recognition. Despite the impressive advancements in software-defined radios (SDR), STR is still a difficult task. Numerous factors, such as slanted lettering, linguistic variances, poor image quality, varied typefaces, and unique text forms, contribute to this issue. STR, a branch of optical character recognition (OCR), seeks to precisely find and identify characters in images of situations, including text that appears on billboards. Text that is twisted or deformed, complex backgrounds, and various font styles are among the factors that lead to computer vision issues.

Text representation techniques are continually evolving to keep up with advances in information technology. More specifically, effective communication and information access depend more and more on the ability to recognize text from start to finish and to identify it in context. As a result, much more research in this particular field is required. Deep learning techniques are being used because STR researchers are being used frequently [2]. Text in natural scene photos can be challenging to identify due to its inconsistent nature, extreme blurring, perspective distortions, and diverse character. Following the resurgence of neural networks and improvements in publicly accessible vision datasets, the computer vision community has demonstrated a strong interest in the research topics of text recognition from low-resolution images and scene text identification of irregular text from naturalistic photos. The first phases of sophisticated deep learning algorithms have been demonstrated by the most recent International Conference on Document Analysis and Recognition Robust Reading (ICDAR) challenging reading challenges. These days, the most widely used deep learning recognition techniques are photo rectification, feature extraction, and sequence prediction. The accuracy of text recognition in real-world situations has significantly improved with the use of deep learning in STR [3].

Convolutional neural networks (CNNs) use local spatial information in the input to efficiently uncover hidden patterns. However, in the field of STR, recurrent neural networks (RNNs) are thought to be the best method for capturing context and dependency in sequential data [4]. To make predictions, RNNs, may retain and utilize past data from earlier time steps, as a result, they work well with sequential input, such as text data. RNNs efficiently capture the contextual relationships between elements in STR tasks, allowing precise text identification and comprehension. Text is frequently displayed in STR as a patch or string of characters. Conversely, CNNs show competence in identifying important visual characteristics in input images. CNNs are capable of hierarchically developing complicated representations and capturing local spatial patterns. Convolutional layers and pooling techniques are employed for this [5]. The following methods can be used to identify features from textual images and to classify or identify objects in the short tandem repeats field. Although deep learning works incredibly well, it suffers greatly from partially obscured or poor-quality images. The public databases contain a range of image types, such as regular, low-resolution, and partially occluded photos. There are several reasons why text graphics with low resolution might exist. One cause might be that the image was compressed to reduce storage space [6]. Another possibility is that the picture was taken using a camera that has a limited amount of focus points. In recognition systems, low-resolution pictures are often handled with bicubic and bilinear interpolation methods. The up-sampled pictures are still out of focus. Furthermore, although these techniques greatly enhance performance on typical scene text, they are unable to yield satisfactory outcomes on difficult irregular text, which has long been a problem for STR.

The incorporation of deep learning methodologies to improve text identification and recognition in natural photos is highlighted in this paper's thorough analysis of advanced techniques in STR. The study investigates the efficacy of CNNs and RNNs in enhancing text identification accuracy in order to address the difficulties presented by irregular text shapes, low image quality, and complicated backdrops. A new method is presented that incorporates a deep feature augmentation module (DFAM) and deep feature refinement module (DFRM) for accurate text localization, along with a deformable convolutional network for improved feature extraction. The methodology includes a complex feature extraction process, the DFAM, and the use of differentiable binarization to create score and threshold maps. The effectiveness of the suggested STR techniques in practical applications is demonstrated by extensive experiments carried out on the common objects in context (COCO)-Text dataset, which show notable gains in precision, recall, and F1-scores when compared to current state-of-the-art methods.

The main contributions of this paper can be summarized as follows:

i) Enhanced text localization: the proposed adaptive deformable feature augmentation and refinement network (ADFARN) methodology introduces a novel deep feature refinement (FRE) that significantly improves text localization by leveraging refinement.

ii) Robust enhanced feature extraction: ADFARN utilizes a deformable convolutional network to perform enhanced feature extraction, capturing intricate text patterns across various scales and resolutions.

iii) State-of-the-art performance: ADFARN outperforms current state-of-the-art techniques in terms of precision, recall, and F1-scores after thorough testing on the COCO-Text dataset. A strong and effective text recognition system is produced by combining improved feature extraction and boundary augmentation approaches, establishing a new standard in the field of STR.

This paper's research is divided into four sections: a quick summary is covered in the section 1, and related work is covered in the section 2. Creating a suggested methodology is the focus of the section 3. The performance evaluation is covered in the section 4, where the findings are displayed as tables and graphs.

## 2. RELATED WORK

The identification of text in any format can be achieved by utilizing an encoder that leverages local dependency modeling, as proposed by Lee *et al*. [6]. The encoder was integrated with an adaptive 2D self-attention mechanism to efficiently capture spatial interactions. The limitation of training spatial transformer network (STN)-based irregular text recognition systems is discussed by Cheng *et al*. [7]. The method utilizes weight combinations to construct sequences and incorporates feature extraction in four text directions. The approach used by the robust scanner [8] to reduce erroneous recognition of semantic-free data involves the utilization of position-enhanced and hybrid branches in the decoder. The merging of these branches to produce prediction results is accomplished using a dynamic fusion module. The utilization of merging modules and mixing blocks was implemented by Du *et al*. [9] in their study to enhance the process of multi-granularity feature extraction in pure virtual machine architectures. The utilization of this particular approach resulted in an improved trade-off between accuracy and performance.

The thin-plate spline (TPS)++ transformation for text correction, known as TPS++, was first introduced by Zheng *et al*. [10]. The attention technique is employed by TPS++ to enhance the precision and adaptability of text correction. The TPS++ system employed a simultaneous assessment of attention scores and foreground control points to enhance the readability and naturalness of text repairs. The sharing of the recognizer's feature backbone results in a decrease in both the inference time and the parameter overhead. The graph-based modeling approach was introduced by Yan *et al*. [11] as a method for acquiring basic representations of text graphics from scenes. To train these representations, the researchers developed weighted aggregators and pooling techniques. The input representations undergo a transformation process using graph convolutional networks, resulting in the generation of more intricate visual text representations. The following work proposes a systematic approach to addressing misalignment problems in the field of text recognition. The proposed technique, referred to as primitive representation learning network with 2D attention (PREN2D), is an encoder-decoder model that utilizes a 2D attention mechanism and visual text representations. The technique employed in this approach utilizes character-by-character identification to decrease the speed of processing. The decoupled attention network was introduced by Wang *et al*. [12] to address the challenges of alignment and historical decoding in STR. The deep alignment network consists of three primary components: a feature encoder, a decoupled text decoder, and a convolutional alignment module. The detachment alignment network enhances the accuracy and flexibility of text recognition by isolating the alignment procedure. The experiments conducted on text-like sound patterns revealed that the method encountered difficulties in accurately aligning the text. Deelaka *et al*. [13] developed a new model architecture that incorporated various visual feature encoding and feature projection techniques. The model produced a predetermined set of item labels by considering the restricted character count in the training images. However, the system was not capable of accurately forecasting the positions of the items. Federated learning systems aim to minimize parameter spaces and computational complexity to achieve efficient training and real-time inference. The model utilized a feature localization unit and an encoder that relied on geometric shapes to predict ground-truth label sequences. The model assumed that the input photos were arranged horizontally and contained only one row of text. The technique is specifically designed to handle numerical data. However, the use of unexpected or irregular language can significantly impact the effectiveness and efficiency of the technique. The proposed methodology [14] aims to achieve two main objectives: enhancing the model's sensitivity to latent features and expediting end-to-end sequence learning for Persian digit identification. The incorporation of a convolutional-based model that combines the excitation gate with squeezing enables the achievement of this objective.

For STR, or visual collaboration and dual-stream fusion (VOLTER), it is strongly advised to employ dual-stream fusion and visual augmentation approaches. To overcome visual constraints and enhance predictive capabilities, the first step is to develop a multi-stage local-global collaboration visual model (LGC-VM) [15]. Integrating local and global elements at various scales is this paradigm's main goal. A vision-language contrastive (VLC) module is our system's second feature. By making it possible to compare the representations of both languages, this module aims to facilitate successful links between vision and language. Accurately aligning the feature spaces of the language-model (LM) and vision-model (VM) is the main goal. In addition, we propose the creation of a dual-stream feature enhancement (DSFE) module to solve the problem of synchronizing several modalities and offer a smoother integration. Facilitating one-way communication between verbal and visual elements is the aim of this module.

The approach for text recognition is referred to as prototype-based unsupervised domain adaptation (ProtoUDA) [16]–[18]. The class prototypes are computed using the source, target, and mixed (source-target) domains in this approach. The ProtoUDA technique utilizes pseudo-labels to extract character features while simultaneously offering word-level monitoring. Additionally, we provide two complementary parallel modules for alignment at both the instance and class levels. The purpose of these modules is to facilitate the transfer of data from source domain to destination domain, utilizing specific character features as criteria.

## 3. METHOD

The process of extracting and fusing enhanced features begins with feeding an input picture into the deformable feature extractor network (DFEN) module. The feature $fe$ with 128 channels is produced by concatenating the fused enhanced features after they have been upsampled to 1/4 of the original image's size. Next, to extract refinement and get a feature $fe_x$, by implementing the precise DFRM. $fe$ and $fe_x$ are added together element by element to get the $fe_{usage}$ usage. A prediction head is fed with $fe_{usage}$ to forecast text and non-text score maps. With $fe$ input, a second prediction head creates the threshold map. Ultimately, the score map and threshold map use differentiable binarization to compute the approximate binary map. The proposed model is connected to the network during the training phase to improve the feature representations. Every network node is trained from start to finish. Figure 1 shows the proposed ADFARN architecture.

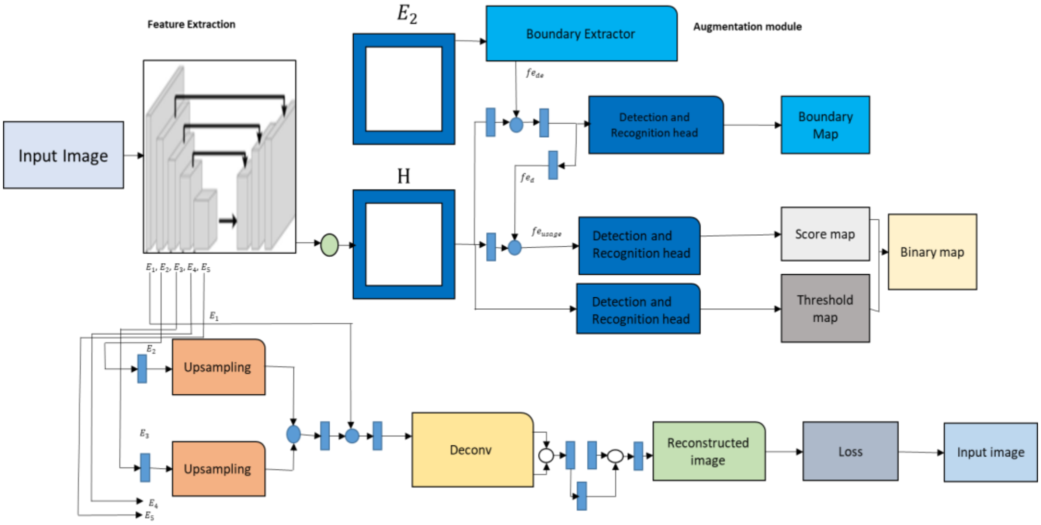

Figure 1. Proposed ADFARN architecture

### 3.1. Deformable feature extractor network module

This model employs a deformable convolutional network to extract enhanced features as $C_1, C_2, C_3, C_4$, and $C_5$ represents various feature maps wherein the resolutions recorded as given as 1/2, 1/4, 1/8, 1/16, and 1/32 for the input size of the image with the corresponding channels as given by 64, 256, 512, 1024, and 2048. The model provides an altering field for the model which benefits the text instances for varied aspects and scales. The convolutions are applied in all the three stages. The enhanced features are then further fused by up-sampling the sum element-wise. Further, the fused enhanced features of 1/4, 1/8, 1/16, and 1/32 resolution are generated with 128 channels.

### 3.2. Deep feature augmentation module

Robust CNNs are used in scene text identification algorithms to extract improved features and boost overall performance. However, when creating feature maps of various sizes using basic sampling or convolutional approaches, the textures and borders of text instances are compromised. This insight leads to the development of a lightweight, pluggable module for DFAM enhancement. $fe_{ps}$ enhances feature representation. However, given $E_k$ with $k = \{1,2,3,4,5\}$, the proposed model focuses on the prediction of the reconstructed image $K_{rec}$ for the reconstruction of the input image is given by (1).

$$\begin{aligned} K_{rec} &= fe_{ps}(E_k) \\ k &= 1,2,3,4,5 \end{aligned} \tag{1}$$

To feed the input to the proposed model for conducting the ablation experiments the inputs fed are $E_1, E_2$, and $E_3$. The features derived from the input $E_2$ and $E_3$ are upsampled to $E_1$ through linear interpolation. These are concatenated and processed through a convolution block followed by an element sum of $E_1$, to sample feature maps that are fixed to the original size within the input image, a deconvolutional layer is modified. The expected outcomes are produced using different $conv_{3\times3}$ convolutional layers. In order to enable text detection across sceneries and traffic panels, the network learns and acquires comprehensive information on feature representation of texts.

### 3.3. Deep feature refinement module

In general, it is a complex task to classify pixels in annotations that are far from the boundary accurately. Our study and testing suggest that enhanced feature fusion might lead to confusion between background and border data. Current scene text detectors disregard the significance of text borders and always consider every pixel in a proposal identically. Despite making up a relatively minor portion of the image, the pixels in text borders are crucial for text localization. For accurate localization, we thus suggest an FRE that specifically uses text refinement. The convolutional networks in orthogonal directions $conv_{1\times3}$ $conv_{3\times1}$ in three dilations which capture text refinement. Through element-wise sum, concretely, an element-wise list is generated as a feature $fe_{de}$, the $fe$ contains information that combines $fe$ and $fe_{de}$, in this one branch uses a boundary map prediction head. To acquire a feature $fe_{d}$ for text-based feature improvement, the other branch goes through a $conv_{3\times3}$ and rectified linear unit (ReLU); the boundary map indicates the boundary/non-boundary classification process.

### 3.4. Optimized label generation and loss function

Every text occurrence is designated as a polygon in the score map, threshold map, and estimated binary map, which are all the same. Different datasets are used to differentiate the vertexes. Each pixel (x,y) in the binary map is downsized to a pixel whose value is summarized to 0, and the shortest distance is calculated $F_{x,y}$. The mapping distance for each text is formulated as shown in (2). The distance is mapped from each text distance which is evaluated as given in (3). This is evaluated as given in (4).

$$F = \{F_{x,y}\}; \qquad x, y \in V \tag{2}$$

$$I = \{1 \quad if\ F_{x,y} < 2 \;\; 0 \qquad else \tag{3}$$

$$N = N_U + \alpha_1 N_V + \alpha_2 (N_D + N_{PS}) + \alpha_3 N_{loss} \tag{4}$$

Here $N_U$, $N_V$, $N_D$ , $N_{PS}$ and $N_{loss}$ depicts border maps, binary maps, score maps, threshold maps, and reconstructed images. The parameters are set to 2, 0.2, and 0.02. The binary cross entropy loss value is used to represent the cross-entropy loss. for $N_U$ ,$N_1$ loss for $N_V$ and dice loss as $N_D$. Algorithm 1 shows the enhanced boundary-enhanced STR (ADFRN) algorithm.

```
Algorithm 1. Enhanced boundary-enhanced STR (ADFARN)
Input:   An input image
Step 1:  DFEN:
         i)   Feed the input image into the DFEN module.
         ii)  Apply deformable convolutional networks to extract enhanced features $C_1, C_2, C_3, C_4$, and $C_5$
              where:
              –  $C_1$ has resolution 1/2
              –  $C_2$ has resolution 1/4
              –  $C_3$ has resolution 1/8
              –  $C_4$has resolution 1/16
              –  $C_5$ has resolution 1/32
              Each feature map has channels 64, 256, 512, 1024, and 2048 respectively.
         iii) Fuse the enhanced features by upsampling and summing element-wise, resulting in a fused
              deformable feature fe with 128 channels.
Step 2:  FRE:
         i)   Implement the FRE:
              –  Extract refinement to get $fe_x$ feature using convolutional networks in orthogonal directions
                 $conv_{1\times3}$ $conv_{3\times1}$ in three dilations.
              –  Combine fe and $fe_x$ element-wise to get $fe_{usage}$
         ii)  Use a prediction head-on $fe_{usage}$ to forecast text and non-text score maps.
         iii) Use a second prediction head-on fe to create the threshold map.
Step 3:  Differentiable binarization: use differentiable binarization to calculate the estimated binary map
based on the score map and threshold map.
Step 4:  DFA:
         i)   For each feature map $E_k$ with $k = \{1,2,3,4,5\}$: predict reconstructed image $K_{rec} = fe_{ps}(E_k)$.
         ii)  Upsample features derived from $E_2$ and $E_3$ are upsampled to $E_1$ through linear interpolation.
```

iii) Concatenate and process these features through a convolution block, followed by an element sum of $\mathrm{E}_1$.
iv) Sample feature maps that are fixed to the original input image size should be subjected to a deconvolutional layer.
v) Generate predicted results through various $\mathrm{conv}_{3\times3}$ convolutional layers.

Step 5: Optimized label generation:
i) Label each text occurrence as a polygon for the score map, threshold map, and estimated binary map.
ii) Compute the shortest distance $\mathrm{F}_{\mathrm{x,y}}$ from each pixel $\mathrm{x,y}$ and shrink it in the binary map to a pixel value of 0 if $\mathrm{F}_{\mathrm{x,y}}$ <2, otherwise to 1.

Step 6: Optimized loss function: calculate the loss as:
$\mathrm{N} = \mathrm{N_U} + \alpha_1\mathrm{N_V} + \alpha_2(\mathrm{N_D} + \mathrm{N_{PS}}) + \alpha_3\mathrm{N_{loss}}$
i) Where $\mathrm{N_U}$ represents the score map loss, $\mathrm{N_V}$ represents the threshold map loss, $\mathrm{N_D}$ represents the binary map loss, $\mathrm{N_{PS}}$ represents the reconstructed image loss, and $\mathrm{N_{loss}}$ represents the boundary map loss.
ii) Set parameters to α1=2, α2=0.2, and α3=0.02.
iii) Use cross-entropy loss for binary cross-entropy value $\mathrm{N_U}$, loss $\mathrm{N_1}$ for $\mathrm{N_V}$, and dice loss for $\mathrm{N_D}$.

Step 7: Training: connect the proposed model to the network and train every network node from start to finish to improve feature representations.

Output: approximate binary map indicating detected and recognized text.

## 4. PERFORMANCE EVALUATION

The assessment metrics utilized for text detection encompass precision, recall, and F1-score. The ratio of recognized text regions to all text regions is measured by the recall metric. The F1-score, sometimes referred to as F1-score, is a statistic that uses harmonic average to combine recall and accuracy. It is frequently used to assess how well detection algorithms work. One crucial criterion for assessing a model's performance is its computational complexity. It takes into account elements like inference time, computational complexity, and parameter count. The robustness metric, which is seen to be of the utmost relevance, is frequently used to assess a model's performance. The capacity of model to perform consistently across many datasets and contexts is referred to as model stability the format of tables and graphs.

### 4.1. Dataset details

A large-scale dataset called COCO-Text was created to improve text identification and detection in natural photos. It adds more than 63,686 photos with more than 173,589 text instances to the COCO dataset. Bounding boxes, transcriptions, and characteristics like language and readability are added to each text instance. The dataset is perfect for creating and evaluating reliable text detection and identification algorithms because of the variety of text appearances, intricate backgrounds, and multilingual content. Widely used for benchmarking, COCO-Text helps push the boundaries of scene understanding by incorporating textual information, offering a comprehensive resource for researchers and practitioners aiming to enhance text analysis in real-world scenarios.

### 4.2. Results

A comparison of different approaches based on precision, recall, and F1-score is shown in Table 1. The proposed segmentation (PS) method outperforms all other methods with the highest precision of 96.89%, recall of 96.76%, and an F1-score of 96.5%. Notably, the ensemble segmentation (ES) method also demonstrates strong performance, achieving a precision of 94.28%, recall of 93.84%, and an F1-score of 94.05%, indicating its effectiveness in the given context.

Table 1. Results

| Method | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|
| Manjari *et al.* [19] | 81.36 | 79.84 | 80.59 |
| Prabu and Sundar [20] | 82.57 | 80.65 | 81.59 |
| Larbi [21] | 76.89 | 77.98 | 77.43 |
| Tarride *et al.* [22] | 75.41 | 76.32 | 75.86 |
| Bhatt *et al.* [23] | 89.63 | 87.81 | 88.71 |
| Vishwakarma *et al.* [24] | 91.37 | 86.29 | 88.75 |
| ES [25] | 94.28 | 93.84 | 94.05 |
| PS | 96.89 | 96.76 | 96.5 |

Figure 2 indicates that the PS methodologies achieve the highest precision, around 94%, suggesting superior performance in correctly identifying relevant instances compared to other methods. The research in [23], [24] also demonstrate high precision, around 90% and 91% respectively, showing strong performance but slightly lower than ES and PS. Prabu and Sundar [20] achieved a precision of approximately 83%, which is moderate but still significantly higher than the remaining methods. The research in [19], [21] show precision values of approximately 81% and 77% respectively, indicating reasonable but lower performance. Tarride *et al*. [22] present the lowest precision value at around 75%, suggesting room for improvement. This analysis highlights the effectiveness of the PS and ES methods in achieving high precision in text detection tasks on the COCO-Text dataset.

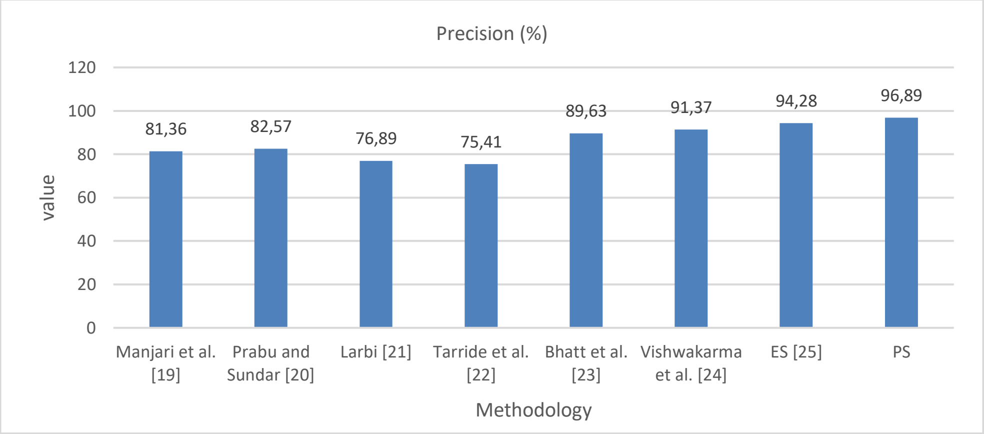

Figure 2. Precision measure

Figure 3 depicts the recall (%) values for various methodologies applied to the COCO-Text dataset. The methods compared are from studies in [19]–[25], respectively. The PS and ES methodologies achieve the highest recall rates, both around 94%, indicating their superior ability to identify relevant instances. The research in [23], [24] also demonstrate strong recall values at approximately 88%, showing effective performance. The research in [19], [20] achieve moderate recall rates of around 81%, while the research in [21], [22] show lower recall rates at approximately 77% and 76% respectively. This analysis highlights the effectiveness of the PS and ES methods in achieving high recall in text detection tasks on the COCO-Text dataset, outperforming other methodologies in terms of identifying maximum number of relevant instances.

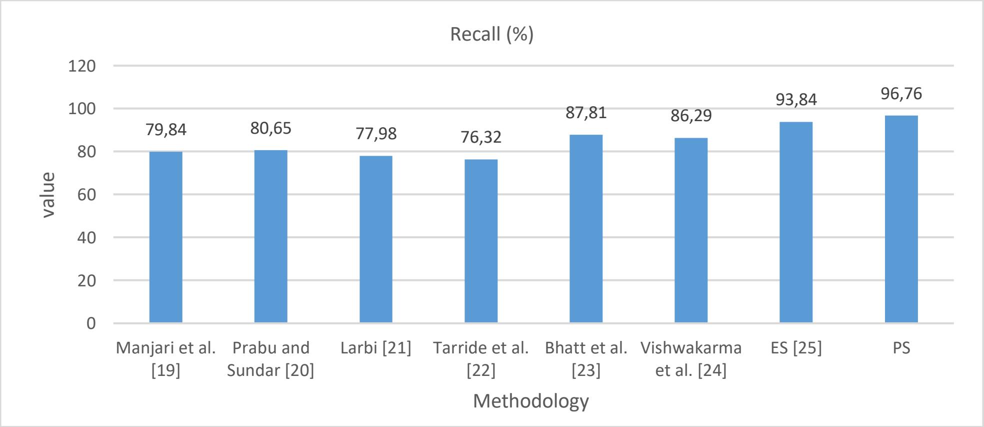

Figure 3. Recall measure

Figure 4 presents a comparative analysis of various methods based on their F1-scores. The methods evaluated include those proposed in [19]–[25], respectively. The highest F1-score is achieved by the PS method with 96.5%, followed closely by the ES method with 94.05%. The research in [23], [24] also show strong performances with F1-scores of 88.75% and 88.71%, respectively. The methods in [19], [20] yield moderate performances, with F1-scores of 81.59% and 80.59%. In contrast, the research in [21], [22] have the lowest F1-scores at 77.43% and 75.86%, respectively. This analysis highlights that the PS and ES methods significantly outperform the others, indicating their superior efficacy in the evaluated context. The results suggest that while several methods show competitive performance, there is a clear distinction in effectiveness among the top-performing and lower-performing methods.

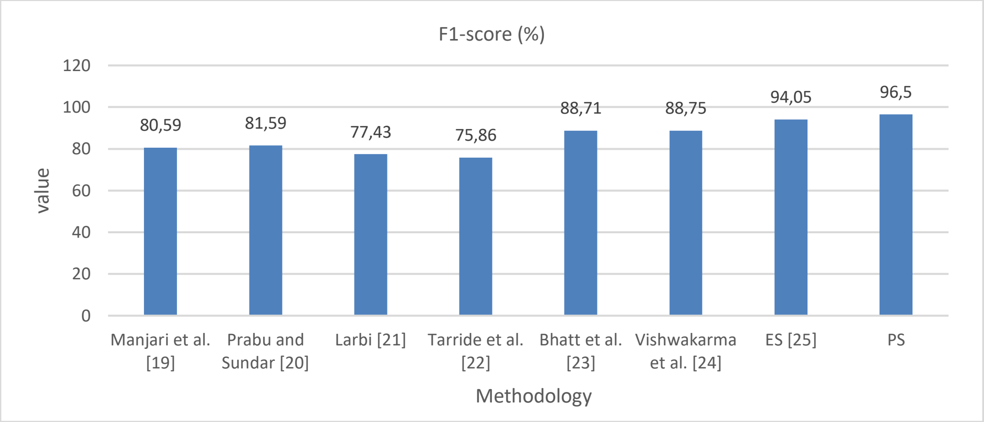

Figure 4. F1-score measure comparison of existing state-of-art techniques with PS

### 4.3. Comparison analysis

The ADFARN approach significantly outperforms the current state-of-the-art technique, ES, according to performance measures. In particular, a greater rate of correctly identified text instances was shown by a 2.77% improvement in the precision statistic. The model's improved capacity to recognize all pertinent text instances was reflected in the recall metric, which experienced an even bigger improvement of 3.11%. Additionally, the F1-score—which strikes a compromise between recall and precision—rose by 2.60%, indicating a comprehensive improvement in the text recognition system's overall performance. These enhancements highlight how well the suggested ADFARN approach handles the challenges of STR and produces better outcomes than other approaches. The comparison analysis is displayed in Table 2.

Table 2. Comparison analysis

| Metric | ES | PS | Improvization in % |
|---|---|---|---|
| Precision (%) | 94.28 | 96.89 | 2.768349597 |
| Recall (%) | 93.84 | 96.76 | 3.111679454 |
| F1-score (%) | 94.05 | 96.5 | 2.604997342 |

## 5. CONCLUSION

The ADFARN methodology presents a significant advancement in the field of STR. By integrating deformable convolutional networks for deformable feature extraction and a novel FRE, ADFARN effectively addresses the challenges posed by variations in text appearance, orientation, and background complexity. The comprehensive process of DFEN, DFAM, DFRM, and the use of differentiable binarization enhances the precision and accuracy of text detection and recognition in natural scenes. The adaptive nature of the model allows for robust handling of low-resolution and partially occluded text, making it highly versatile. The incorporation of an integrated module further improves the alignment of visual and textual features. Performance evaluations on the COCO-Text dataset demonstrate that ADFARN significantly outperforms existing state-of-the-art methods, achieving higher precision, recall, and F1-scores. This research establishes ADFARN as a robust and efficient solution for real-world text recognition applications, paving the way for further advancements in this domain.

## ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our guide for his constant direction, priceless insights, and support during the research process.

## FUNDING INFORMATION

The authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ratnamala S. Patil | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |  |  | ✓ |  |
| Geeta Hanji | ✓ | ✓ |  | ✓ | ✓ | ✓ |  | ✓ | ✓ | ✓ | ✓ | ✓ |  |  |
| Rakesh Hudud | ✓ | ✓ | ✓ |  | ✓ | ✓ | ✓ |  | ✓ | ✓ |  |  | ✓ |  |

| | | | | | |
|---|---|---|---|---|---|
| C | : **C**onceptualization | I | : **I**nvestigation | Vi | : **Vi**sualization |
| M | : **M**ethodology | R | : **R**esources | Su | : **Su**pervision |
| So | : **So**ftware | D | : **D**ata Curation | P | : **P**roject administration |
| Va | : **Va**lidation | O | : Writing - **O**riginal Draft | Fu | : **Fu**nding acquisition |
| Fo | : **Fo**rmal analysis | E | : Writing - Review & **E**diting | | |

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The data that support the findings of this study are available in [IEEE Transactions on Geoscience and Remote Sensing] at http://doi.org/10.1109/TGRS.2024.3404605, reference [3].

## REFERENCES

[1] M. Agrawal, A. S. Jalal, and H. Sharma, "A deep learning based strategies for scene-text VQA system used in industrial applications: a critical analysis," in *2024 1st International Conference on Innovative Sustainable Technologies for Energy, Mechatronics, and Smart Systems (ISTEMS)*, Dehradun, India: IEEE, Apr. 2024, pp. 1–5, doi: 10.1109/ISTEMS60181.2024.10560122.

[2] Y. Yan, N. Cooper, O. Chaparro, K. Moran, and D. Poshyvanyk, "Semantic GUI scene learning and video alignment for detecting duplicate video-based bug reports," in *IEEE/ACM 46th International Conference on Software Engineering*, New York, NY, USA: ACM, Apr. 2024, pp. 1–13, doi: 10.1145/3597503.3639163.

[3] F. Wang, X. Zhu, X. Liu, Y. Zhang, and Y. Li, "Scene graph-aware hierarchical fusion network for remote sensing image retrieval with text feedback," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024, doi: 10.1109/TGRS.2024.3404605.

[4] T. do, T. Tran, T. Nguyen, D.-D. Le, and T. D. Ngo, "SignboardText: text detection and recognition in in-the-wild signboard images," *IEEE Access*, vol. 12, pp. 62942–62957, 2024, doi: 10.1109/ACCESS.2024.3395374.

[5] X. Yang, Z. Qiao, J. Wei, D. Yang, and Y. Zhou, "Masked and permuted implicit context learning for scene text recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 964–968, 2024, doi: 10.1109/LSP.2024.3381893.

[6] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 2326–2335, doi: 10.1109/CVPRW50498.2020.00281.

[7] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: towards arbitrarily-oriented text recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 5571–5579, doi: 10.1109/CVPR.2018.00584.

[8] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, "RobustScanner: dynamically enhancing positional clues for robust text recognition," in *Computer Vision – ECCV 2020 (ECCV 2020)*, Glasgow, UK: Springer, Cham, 2020, pp. 135–151, doi: 10.1007/978-3-030-58529-7_9.

[9] Y. Du *et al.*, "SVTR: scene text recognition with a single visual model," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 884–890, doi: 10.24963/ijcai.2022/124.

[10] T. Zheng, Z. Chen, J. Bai, H. Xie, and Y.-G. Jiang, "TPS++: attention-enhanced thin-plate spline for scene text recognition," in *Thirty-Second International Joint Conference on Artificial Intelligence*, Aug. 2023, pp. 1777–1785, doi: 10.24963/ijcai.2023/197.

[11] R. Yan, L. Peng, S. Xiao, and G. Yao, "Primitive representation learning for scene text recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 284–293, doi: 10.1109/CVPR46437.2021.00035.

[12] T. Wang *et al.*, "Decoupled attention network for text recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12216–12224, Apr. 2020, doi: 10.1609/aaai.v34i07.6903.

[13] P. N. Deelaka, D. R. Jayakodi, and D. Y. Silva, "Geometric perception based efficient text recognition" *arXiv:* 2302.03873, 2023

[14] A. A. A. Alshawi, J. Tanha, and M. A. Balafar, "An attention-based convolutional recurrent neural networks for scene text recognition," *IEEE Access*, vol. 12, pp. 8123–8134, 2024, doi: 10.1109/ACCESS.2024.3352748.

[15] J.-N. Li, X.-Q. Liu, X. Luo, and X.-S. Xu, “VOLTER: visual collaboration and dual-stream fusion for scene text recognition,” *IEEE Transactions on Multimedia*, vol. 26, pp. 6437–6448, 2024, doi: 10.1109/TMM.2024.3350916.

[16] X.-Q. Liu, X.-Y. Ding, X. Luo, and X.-S. Xu, “ProtoUDA: prototype-based unsupervised adaptation for cross-domain text recognition,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 9096–9108, Dec. 2024, doi: 10.1109/TKDE.2023.3344761.

[17] P. Pujar, A. Kumar, and V. Kumar, “Efficient plant leaf detection through machine learning approach based on corn leaf image classification,” *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 1139–1148, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp1139-1148.

[18] S. H. Sreedhara, V. Kumar, and S. Salma, “Efficient big data clustering using adhoc fuzzy C means and auto-encoder CNN,” in *Inventive Computation and Information Technologies*, New Delhi, India: Springer, Singapore, 2023, pp. 353–368, doi: 10.1007/978-981-19-7402-1_25.

[19] K. Manjari, M. Verma, G. Singal, and S. Namasudra, “QEST: quantized and efficient scene text detector using deep learning,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1–18, May 2023, doi: 10.1145/3526217.

[20] S. Prabu and K. J. A. Sundar, “Enhanced attention-based encoder-decoder framework for text recognition,” *Intelligent Automation & Soft Computing*, vol. 35, no. 2, pp. 2071–2086, 2023, doi: 10.32604/iasc.2023.029105.

[21] G. Larbi, “Two-step text detection framework in natural scenes based on Pseudo-Zernike moments and CNN,” *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10595–10616, Mar. 2023, doi: 10.1007/s11042-022-13690-6.

[22] S. Tarride *et al.*, “Large-scale genealogical information extraction from handwritten Quebec parish records,” *International Journal on Document Analysis and Recognition*, vol. 26, no. 3, pp. 255–272, Sep. 2023, doi: 10.1007/s10032-023-00427-w.

[23] R. Bhatt, A. Rai, S. Chanda, and N. C. Krishnan, “Pho(SC)-CTC—a hybrid approach towards zero-shot word image recognition,” *International Journal on Document Analysis and Recognition*, vol. 26, no. 1, pp. 51–63, Mar. 2023, doi: 10.1007/s10032-022-00407-6.

[24] D. K. Vishwakarma, P. Meel, A. Yadav, and K. Singh, “A framework of fake news detection on web platform using ConvNet,” *Social Network Analysis and Mining*, vol. 13, no. 1, Jan. 2023, doi: 10.1007/s13278-023-01026-7.

[25] T. Geng, “Transforming scene text detection and recognition: a multi-scale end-to-end approach with transformer framework,” *IEEE Access*, vol. 12, pp. 40582–40596, 2024, doi: 10.1109/ACCESS.2024.3375497.

## BIOGRAPHIES OF AUTHORS

**Ratnamala S. Patil** received her bachelor’s degree in Electronics and Communication Engineering from the Visvesvaraya Technological University, Belgaum, India in 2014 and master degree in Digital Communication and Networking from same university in 2016. She is currently pursuing her Ph.D. degree from the same university. She is presently working as assistant professor in Departments of Electronics and Communication Engineering Sharnbasva University Kalaburagi, Karnataka, India. Her primary area of interest is image processing, machine learning, and pattern recognition. She can be contacted at email: ratnamala_12@rediffmail.com.

**Geeta Hanji** working presently as professor in Department of Electronics and Communication Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi. She has 18 years of teaching and 10 years of research experience, and completed her B.E., M.Tech., and Ph.D. in Electronics and Communication Engineering. Her research area includes digital image processing and pattern recognition. She published more than 55 research papers in above mentioned areas. She has 30 years of teaching experience and 18 years of research experience. She can be contacted at email: geetanjalipatil123@gmail.com or geetahanji@pdaengg.com

**Dr. Rakesh Hudud** have earned an engineering degree in Electronics and Communication from SDM College of Engineering in Dharwad, affiliated with Visvesvaraya Technological University, Belagavi, in 2011. Followed by an M.Tech. degree from Poojya Doddappa Appa College of Engineering in Kalaburagi, also affiliated with Visvesvaraya Technological University, Belagavi, in 2013, and culminating with a Ph.D. in Image Processing from Sri Satya Sai University of Technology and Medical Sciences in Sehore in 2019, currently serves as an assistant professor at Poojya Doddappa Appa College of Engineering in Kalaburagi, Karnataka. With over five years of experience, and currently guiding research scholar under Visvesvaraya Technological University in Belagavi. He can be contacted at email: rhuded@gmail.com.