BicTd-Brightening Intercultural Communication: Text Detection with Enhanced Light

Vidya, Manjula G R and Manjula C Belavagi

Abstract—**Identifying text in natural scenes poses significant challenges due to variations in illumination, complex backgrounds, and diverse text styles. This paper introduces novel method aimed at improving the performance over enhancing the detection of low light images with text. Proposed method integrates traditional computer vision techniques with deep learning models to achieve robust and accurate text detection results. Initially, input images are processed to evaluate light values, determining the need for contrast enhancement using adaptive threshold to improve text visibility under different lighting conditions. Following this convolutional neural network architecture is employed to extract features from the video frame. To further refine text detection, proposed work introduce a novel attention mechanism that prioritizes regions with high text likelihood, thereby enhancing the discriminative power of the model. Additionally, a post-processing step based on non-maximum suppression to filter out false positives and improve the final detection results. The success of the implemented technique is demonstrated by an experimental evaluation on an in-house dataset consisting of 20 video clips with multilingual text. It outperforms state-of-the-art algorithms in terms of robustness and detection accuracy in difficult lighting conditions. BicTd method exhibits promising results in real-world scenarios, offering practical solutions for applications such as autonomous driving, document analysis, and augmented reality.**

Index Terms—**Information Retrieval, Text Detection, Text Extraction , Machine Learning, Deep Learning.**

I. Introduction

The evolution of computer vision has brought text detection into the spotlight due to its diverse array of practical applications. From facilitating seamless translation to aiding in image searching, scene interpretation, geographic mapping, and assisting individuals with visual impairments, text detection has become integral to numerous fields. In the realm of text recognition in videos, a series of four stages typically guide the process: detection, localization, extraction, and recognition. Each stage in detail is explained below:

1) Detection: The first stage involves detecting areas within the video frames where text is present. This is typically done using techniques such as edge detection, contour detection, or more advanced methods like convolutional neural networks (CNNs) trained specifically for text detection. The goal is to identify regions of interest that potentially contain text.
2) Localization: Once text regions are detected, the next step is to precisely localize the boundaries of each individual text instance within these regions. This involves techniques such as bounding box regression or segmentation to accurately outline the extent of each text element. Localization ensures that text is isolated from the background and other elements in the scene.
3) Extraction: After localization, the identified text regions need to be extracted from the video frames. This step often involves cropping the regions of interest and applying preprocessing techniques to enhance the quality of the extracted text. Preprocessing may include image enhancement, noise reduction, or perspective correction to improve the readability of the text.
4) Recognition: The final stage is text recognition, where the extracted text regions are converted into machine-readable format. Optical Character Recognition (OCR) algorithms are commonly employed for this purpose. These algorithms analyze the extracted text and attempt to convert it into digital text data that can be further processed or analyzed. Advanced OCR techniques, including deep learning-based approaches, have significantly improved the accuracy of text recognition in videos.

Each of the above mentioned stages plays a crucial role in the overall process of text detection and recognition in videos, enabling various applications such as automatic subtitle, content indexing for video search, real-time translation, and accessibility features for visually impaired individuals. As computer vision technology continues to advance, we can expect further improvements in the accuracy, speed, and versatility of text recognition systems in videos.

Following are the research contributions : An innovative approach to instant and accurate text identification within pipelines is proposed. The method comprises two fundamental steps, each contributing to the efficiency and precision of the overall process. Firstly, we leverage a Fully Convolutional Network (FCN) model, which provides predictions for words or text lines within the video frames. This step serves as the foundation for subsequent localization efforts. Building upon the predictions generated by the FCN model, the pipeline incorporates the use of rotated rectangles to produce multiple predicted text locations within the frames. While this step increases the granularity of text localization, it also introduces redundancy in the form of multiple bounding boxes.

To address this redundancy and refine the localization to a single, highly accurate box, Non-Maxima (NMS) Sup-

Manuscript received June 6, 2024; revised May 23, 2025.

This work is supported by the Vision Group on Science and Technology (VGST) under Grant 749, Karnataka Science and Technology Promotion Society (KSTePS), Government of Karnataka, India.

Vidya is Assistant Professor in the Department of Robotics and AI, JNN College of Engineering, Shivamogga, Karnataka, India. (e-mail: vidya.rashinkar@gmail.com)

Manjula G R is Professor in the Department of Computer Science and Engineering, JNN College of Engineering, Shivamogga, Karnataka, India. (e-mail: grmanjula@jnnce.ac.in)

Manjula C Belavagi is Assistant Professor-Selection Grade in the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Karnataka, India, 576104 (Corresponding Author, phone: +91 9448-627-687 e-mail: manjula.cb@manipal.edu)

pression is employed . This technique effectively reduces the number of predicted boxes to the most relevant and precise localization, ensuring optimal performance and accuracy. Through rigorous testing on standard datasets such as the Robust Reading Competition sets and ICDAR2015. [1], BicTd solution has demonstrated very good performance and detection capabilities compared to existing algorithms. By streamlining the text detection process and enhancing accuracy, Intended research aims to advance the field of computer vision and unlock new possibilities for practical applications.

The paper is organized as follows in Section II state of the art techniques are discussed. Section III provides methodology and Section IV gives Results and discussions. Finally paper concludes with conclusion Section.

II. Literature survey

In several applications we find text in images, videos. In many automated applications, the goal is to confirm whether the region is text or non-text. If text exists, then progress goes through detect, localize, identify script, recognition of language as stages depending on the domain and work chosen. There are various methods, approaches, algorithms, and technologies used for text detection, identification, and recognition. Mainly the approaches divided into classical text detection approach, Machine learning, Deep learning-based approach, and Mixed approach as given in Fig. 1. Classical techniques often involve a step-by-step process using a pipe line, with each step being finished in a separate block. These techniques focus on local characteristics, establish grouping criteria, extract related components, and categorize textual elements. The focus is primarily on feature extraction. Classical methods sections describe some such approaches.

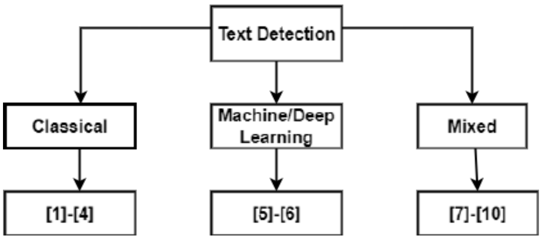

Fig. 1. classification of Text Detection approaches

A. Classical methods:

Lyu et al. in 2005 [2] conducted a thorough investigation of the features of multilingual texts, including Chinese and English. They provide a thorough, effective method for video text extraction, localization, and detection based on the analysis, emphasizing multilingual capacity throughout the entire processing process. Additionally resistant to different background complexity and text appearances is their suggested strategy.

Text is detected using a combination of edge detection, local thresholding, and hysteresis edge recovery. To identify text regions accurately, the coarse-to-fine localization scheme is applied. The text extraction consists of adaptive thresholding, dam point labeling, and inward filling. Experimental results on a large number of video images and comparisons with other methods are reported better results. Extraction of text from natural scene photographs, both Indian and non-Indian, was examined by Mahajan and Rani in 2018 [3]. Processing scenic photos involves a number of steps, including text recognition, segmentation, detection, and location. The initial stage, which involves extracting textual material, is crucial to subsequent processing. Researchers have a hurdle when text extraction becomes challenging due to variations in text font, size, skewness, and noise in the acquired image. Basavaraju etal. in 2021 [4] worked on the multilingual textual data that contains a variety of geometric shapes, which makes discovering the text data even more challenging. In their work, an effective model is presented to locate the multilingual arbitrary oriented text. The method developed a neighborhood structure model to locate the text region. Initially, the maxmin cluster is applied along with 3X3 sliding window to sharpen the text region. The neighborhood structure creates the boundary for every component using normal deviation calculated from the sharpened image. Finally, the double stroke structure model is employed to locate the accurate text region. The presented model is analyzed on five standard datasets such as NUS, arbitrarily oriented text, Hua's, MRRC and real-time video dataset with performance metrics such as recall, precision, and f-measure. Khalid Ali et al [5] in 2017 used zone features of candidate characters based text detection method, by resizing the candidate regions into w x w (w=25) and performed zoning process for feature extraction and applied text region construction and selection rules to find the text regions. Using complementary candidate characters, they are groping and classifying the regions. Experimental results show that recall, precision, F-measure are 0.63, 0.84, 0.72 respectively.

B. Machine learning-Deep learning Method:

Recent methods and approaches use machine learning (ML)/Deep Learning(DL). Here the main aim of the algorithm and method is to make the machine learning to identify the feature to be extracted using training. For text localization these approaches use basically a window/Kernel technique, the classifier rule applied to classifying and forming groups in machine learning, in deep learning these stages combined to make the learning faster. The following papers employed DL and ML techniques for text detection, extraction, and recognition. Using a fully connected network (FCN) for feature learning and text classification, along with non- maximum suppression (NMS) to construct a final result region, text detection in a scene is developed by Zhou et. al [6] and [7]. Two geometric forms were used in the network's design for the text sections. They are called QUAD(quadrangle)and RBOX(rotatedbox). Suggested methods, which utilize the precise same EAST algorithm

C. Mixed Methods:

To yield more precise and accurate findings, some methods use mixed mode approaches, which integrate machine learning/deep learning techniques with classical methods at certain points.. An end-to-end system for scene text detection, localization, and language identification is presented in Multilingual Scene Text Detection and Language Identification

[8] by Saha et al. There are three main steps in this approach. creation of a potential proposal, improvement of the proposal, and linguistic classification of the improved proposal . A K Bhunia et al [9] [10] attention-based convolutional-LSTM network for script identification in natural scene pictures and film frames patches the images before feeding them into a CNN-LSTM framework. After the LSTM, the SoftMax layer is used to determine attention-based patch weights. Local features are obtained by multiplying these weights patch-wise using the matching CNN. Additionally, global features are taken from the last cell state of the LSTM. They used a fusion method that dynamically adjusts the weights of the local and global characteristics for each patch. Port container number detection based on improved EAST algorithm [11] is proposed by XingQi Feng et al. They developed a new algorithm by which eliminates multiple intermediate stages, but also improves the regression method and loss function of the box number area

Wafa Khlif et al [12] has evolved method for detecting text in scene images based on multi-level connected component analysis and learning text component features using CNN, followed by a graph-based grouping of overlapping text boxes and When evaluated on different datasets there method achieved better detection results compared to state-of-the-art methods. Shi-Xue et al [13] proposed an iterative, coarse-to-fine boundary transformer to directly predict precise text boundaries by simplifying the post-processing. The method consists of visual encoding module with feature extraction, a module that generates coarse boundary proposals using prior information (direction field, distance field and classification map), a boundary transformer module, and a novel boundary energy loss module. This module ensures stable optimization via energy minimization and monotonically decreasing constraints.

Ong Yi Ling et al [14] invented as app ”Let Me Read for You” shows 86% accuracy in recognizing types of vertically oriented scene texts(VOS), including top-to-bottom(T2B) vertical texts, botton-to-top (B2T)vertical texts and vertival stroke(VSt) vertical texts, when evaluated with the VOD dataset.Yuanbin Wang et al [15] implemented defogging method. First, the non Sub sampled contour transform (NSCT) is employed in their method to partition the image into low and high frequency elements. In order to enhance the preservation of complex information, the Retinex approach utilises a bootstrap filter instead of a Gaussian filter. The denoising method entails the application of a threshold to the high-frequency sub-band pictures which involves boosting coefficients that surpass the threshold and attenuating coefficients that are below it. The ultimate image is recreated following dehazing by employing the inverse transformation of NSCT. Then, the grey wolf optimization (GWO) algorithm is employed to optimise the regularisation factor of the kernel in the guided filter, hence improving the defogging process. Xinling Feng et al [16] proposed a low resolution digital image reconstruction method based on Rational Function Model. With sequence of bilateral filtering, Rational Function Model applied.Data obtained from the model is taken as the input information, and the generative adversarial network is used to extract image features.Next, SIFT algorithm and Difference of Gaussians function are used for accurate feature extraction.Finally, the processes of feature point direction matching, wavelet transform, bilateral regularization processing, pixel correction, edge adaptive processing, etc. are carried out for ortho correction of the image function model. On this basis, image reconstruction is eventually established.

III. Methodology

Flowchart in figure 2 illustrating the working process of BicTd, a novel approach for enhancing light and text detection in images. The flowchart delineates the sequential steps involved, including preprocessing of input, evaluation of light values using when HSL (Hue, Saturation, and Light) values fall below the 30% threshold of light.The mean value of light below 30% threshold of light. Frames exceeding this threshold are forwarded to the CNN for feature extraction, utilization of a novel attention mechanism, post-processing with non-maximum suppression (NMS), and generation of final results. The ResNet (Residual Network) architecture depicted in Figure 3 incorporates skip or shortcut connections to facilitate faster training without the risk of vanishing gradients. This is achieved by using residual blocks, which consist of convolutional layers along with identity mappings. While the basic building blocks remain consistent across different ResNet architectures, such as ResNet-50, the number of these blocks in each stage may vary. In the proposed work, ResNet-50 v1 is utilized. This particular architecture accepts input images of size 224x224x3. After passing through the five stages of the network, which comprise a series of residual blocks, the output is a single image with 1000 classes. Each stage in ResNet-50 consists of a varying number of residual blocks, with deeper stages typically having more blocks to capture increasingly abstract features. Overall, ResNet-50 v1 offers a balance between depth and computational efficiency, making it suitable for a wide range of image classification tasks. Image enhancement does not adhere to a singular restoration method. Various techniques are employed based on specific requirements, particularly for classifying images as low-light images. In this regard, brightness is a crucial factor, and the average mean intensity of the image’s lightness serves as a determinant for further processing. When the average falls below 30% of light value, the IMG-ENH module accepts the image as a low-light input. Subsequently, the image undergoes three stages of processing to enhance its quality. Figure 5 delineates the working process of the IMG-ENH network, illustrating its functionality in addressing low-light image enhancement The extracted feature at different stages of resnet-50 is used in feature extraction and merging module with Feature Pyramid network as given in Figure 5. Output from this is given to text detection layer, to decide text and non-text regions. With dice loss gives the result about how much it is similar to ground truth. RBOX and text rotation module gives the bounding box coordinate values for the given image after applying Non-Maxima Suppression algorithm. Detecting text from Low light images is a challenge with Light enhancer IMG-ENH module implemented based on the working of retinex theory help the proposed work to detect text from low light images too. Low light enhancer IMG-ENH module work with three different network. They are Reflectance map, Illumination map and illumination enhancement network. Reflectance map concentrate only on reflectance of color,

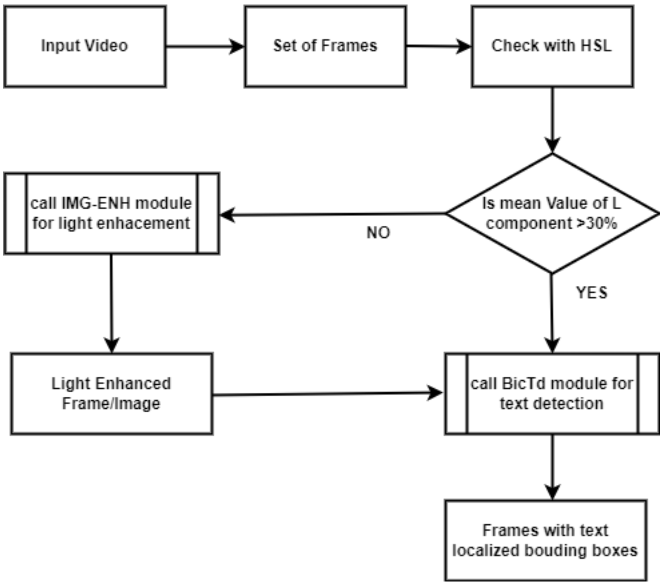
Input Video
Set of Frames
Check with HSL
Is mean Value of L component >30%
NO
YES
call IMG-ENH module for light enhacement
Light Enhanced Frame/Image
call BicTd module for text detection
Frames with text localized bouding boxes

Fig. 2. Working process flowchart

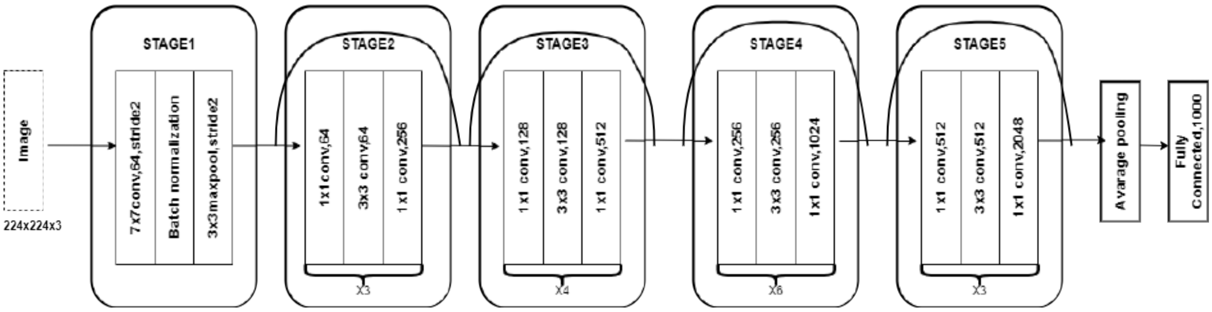

Fig. 3. Resnet-50 v1 architecture

Illumination map look into the brightness component, both module handle these component of an image, then both combined together and given to illumination enhancement network module, which look into the multiscale elements of the obtained image and adjust the illumination suitably.

A. Label Generation

- Geometry Map Generation The geometry map is either RBOX or QUAD as per [5] in our experiment QUAD is used to find ground truth of 8 channels & RBOX calculate rotated rectangle, that cover the minimal area. If the pixel has positive score put into the RBOX channel of ground truth.
- Loss Function Equation 1 is used to calculate the loss:

$$L = L_s + kL_g \tag{1}$$

where L$_s$ is score map loss and L$_g$ is geometry loss respectively. The value of k in our experiment is 1 and is used to balance the loss.

- Dice loss shown in equation 2 is used to know the performance and accuracy of the score map.

$$L_{dice} = 1 - 2 * IoU \tag{2}$$

Where IoU is intersection of Union.

- Loss for geometries: Here the challenge is to find suitable bounding box out of many overlay of bounding boxes, and to find most accurate one our proposed method use Non-Maxima suppression(NMS) which uses Intersection of Union (IoU) and the loss is given by equation 3.

$$L_{bb} - logIoU = -log\frac{|\acute{R} \cap R^*|}{|\acute{R} \cap R^*|} \tag{3}$$

where$\acute{R}$ represents predicate axis aligned bounding box geometry and R^* is corresponding ground truth Height and width is computed using wi=min($\acute{d}2$,$d2^*$)+min($\acute{d}4$,$d4^*$) hi=min($\acute{d}1$,$d1^*$)+min($\acute{d}3$,$d3^*$) where d1, d2, d3, d4 as the distance from pixel to top, right, bottom and left of rectangle. The union area is given by Equation 4.

$$|\acute{R} \cup R^*| = |\acute{R}| + |R^*| - |\acute{R} \cap R^*| \tag{4}$$

and the final geometry loss is computed as sum of equation 3 and angle loss L θ + k with a constant 'k' and is set to 10 to match the scaling with bounding box geometry shown in equation 5.

$$L_g = L_{bb} + kL_\phi \tag{5}$$

- Training used pre trained model built using Resent version 1 with 50 layers depicted in Figure 3 of architecture with pre assigned weight and bias used as checkpoints with Tensorflow2 framework and tested on CPU systems on Windows 10 operating system.

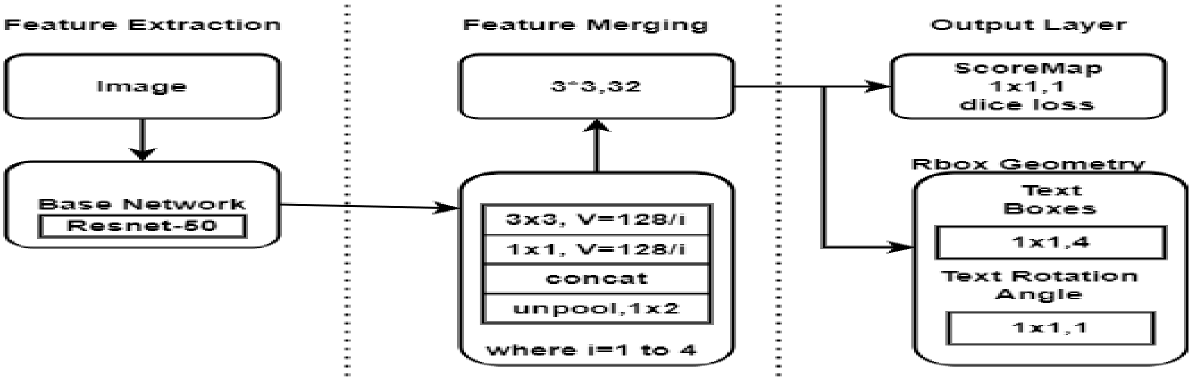

Fig. 4. Text Detection Module

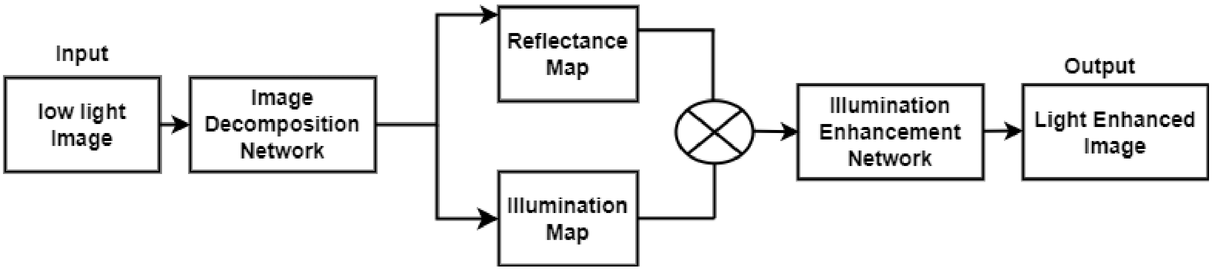

Fig. 5. Low Light Enhancer Module

B. Implementation Details:

In the proposed work, the power of Python is used along with the TensorFlow2 framework and OpenCV to explore computer vision tasks on a Windows platform. The paper explores the functionality of the text localization and light enhancement, respectively. The algorithm 1 and 2 outline the steps for the text localization process.

Algorithm 1 Feature Extraction & Merge

```
Input: video frames
Output: score map, geometry map and loss
Preprocess the input frame,into suitable format (in multiples of 32) Resize the image to 224 * 224 * 3 (height * width * channel)
Pass these images to Resnet-50 v1 for feature extraction at stage2, stage3, stage4 and stage5 as shown in Figure 3. They capture both low-level (spatial information) and high level (semantic information)
extract the features from different layers , map the features in pipeline at merging stages ($g_i$) to form one larger text region from set of smaller regions ($h_i$) [given in equations 6 and 7 respectively]
after last final mergeing a convolutional layer of 3 * 3 produces final feature maps of likely text regions, and is given as input to output layer. As shown in Figure 3
In the output layer 4, feature maps are fed to score map to calculate score map and loss, and the same fed to RBOX a multichannel geometry maps to potentially estimate the orientation (geometry) of the text shown in Table I.
```

$$g_i = \begin{cases} \text{unpool}(h_i), \text{if } i \leq 3 \\ \text{conv } 3 \times 3(h_i), \text{if } i = 4 \end{cases} \tag{6}$$

TABLE I
MULTICHANNEL GEOMETRY MAPS

Geometry	channels	description
AABB	4	$G = R = \{d_i \mid i \in \{1,2,3,4\}\}$
RBOX	5	G={R, θ}

$$h_i = \begin{cases} f_i, \text{if } i = 1 \\ \text{conv } 3 \times 3(\text{conv } 1 \times 1([gi - 1 : f_i])), \text{if } i \neq 1 \end{cases} \tag{7}$$

As shown in Table 1, RBOX encoding involves the use of 5 channels. The first four channels correspond to the distances from pixel i to the top, right, bottom, and left edges, respectively. The fifth channel is specifically utilized for encoding the rotation angle. In comparison, the axis-aligned bounding box employs 4 channels, where each channel represents the distance of pixel i from the top, right, bottom, and left edges

Low light Restoration

In the context of low-light images, the proposed endeavour aims to optimize light restoration methods to enhance detection probability. And the work is divided into three modules. 1)Layer decomposition mentioned in Algorithm 3 2) Reflectance restoration mentioned in Algorithm 4 and 3) illumination adjustment is given in Algorithm 5.

Output image from algorithm 4 and 5 merged together to get final output image, this image fed to text detection module gives more bounding boxes compared to normal low-light frame. And the experimental result on RoadText1k video's frame is detailed in Figure 12 detect one bounding box before enhancement and detects two bounding boxes after light enhancement.

Algorithm 2 Local Non-Maxima Suppression

```
procedure LocalNonMaximaSuppression(Bounding boxes with geometry confidence scores)
    Input: Accept set of bounding boxes with geometry confidence scores
    Output: Final set of bounding boxes after non-maxima suppression

    Arrange the bounding boxes in descending order based on confidence scores
    while Set of bounding boxes is not empty do
        Select the bounding box with the highest confidence score
        Calculate Intersection over Union as per Equation (3)
        Remove all bounding boxes with overlap greater than 0.5 with the selected bounding box
        Add the selected bounding box to the final set
    end while
end procedure
```

Algorithm 3 Layer Decomposition Algorithm

```
procedure LayerDecomposition(Input image $I$)
    Step 1: Accept input image $I$
    Step 2: Prepare images for high and low values depending on illumination $L$ and reflectance $R$
        $I \rightarrow I_{\text{land}}, I_h \quad R \rightarrow R_l, R_h \quad L \rightarrow L_l, L_h$
    Step 3: Regularize the similarity of reflectance:
        $L_{\text{rs}}^{\text{LD}} := \|R_l - R_h\|_{\text{mse}}^2$ where $\| \, \|_{\text{mse}} \rightarrow$ mean square error (MSE) or $l2$ norm (9)
    Step 4: Derive illumination smoothness:
        $L_{\text{is}}^{\text{LD}} := \|\nabla L_l / \max(\nabla I_l, \epsilon)\|_1 + \|\nabla L_h / \max(\nabla I_h, \epsilon)\|_1$ where $\nabla \rightarrow$ first-order derivative and $\| \, \|_1 \rightarrow l1$ norm (10)
    Step 5: Mutual consistency for normalizing edges and flat regions:
        $L_{\text{mc}}^{\text{LD}} := \|\mathcal{M} \circ \exp(-c, \mathcal{M})\|_1$
        where $\mathcal{M} := |\nabla L_l| + |\nabla L_h|$ and $\| \, \|_1^{\leftarrow(@)} \rightarrow l1$ norm (11)
    Step 6: Reconstruct the two images into one single image $OI$, and the reconstruction loss:
        $L^{\text{LD}} := E + 0.01 L_{\text{rs}}^{\text{LD}} + 0.08 L_{\text{is}}^{\text{LD}} + 0.1 L_{\text{mc}}^{\text{LD}}$ (12)
end procedure
```

Algorithm 4 Reflectance Restoration

```
procedure ReflectanceRestoration(Reflectance mapped image, Illumination mapped image)
    Step 1: Accept input as reflectance mapped and illumination mapped images
    Step 2: Calculate the structural similarity and restored reflectance

        $L_{\text{RR}} := \|\hat{R} - R_h\|_{\text{mse}}^2 - \text{SSIM}(\hat{R}, R_h) + \|\nabla \hat{R} - \nabla R_h\|$    (13)

end procedure
```

Algorithm 5 Illumination Adjustment

```
procedure IlluminationAdjustment(Illuminated mapped image as input)
    Step 1: Accept illuminated mapped image as input
    Step 2: Calculate the ratio of strength: [$\alpha = \frac{Lt}{Ls}$ where $Ls \rightarrow$ source light and $Lt \rightarrow$ target light ]
    Step 3: To adjust illumination, perform gamma correction using: [$\gamma = \frac{\| \log(\hat{L}) \|}{\| \log(Ls) \|}$ where $\hat{L}$ is the adjustment illumination map ]
end procedure
```

IV. Results and discussions

To assess the effectiveness of the proposed approach, we evaluated its performance on well-established image datasets, including ICDAR15+13, MLe2e [17], and MSRA-td500 [18] and the details of the datasets are given in Table II. The obtained accuracy rates for these datasets were 78%, 79.5%, and 79.5%, respectively. To visually demonstrate the results, representative output images from each dataset are presented in Figure 6, Figure 7 and Figure 8 respectively, showcasing the input image, ground truth, and the obtained results.

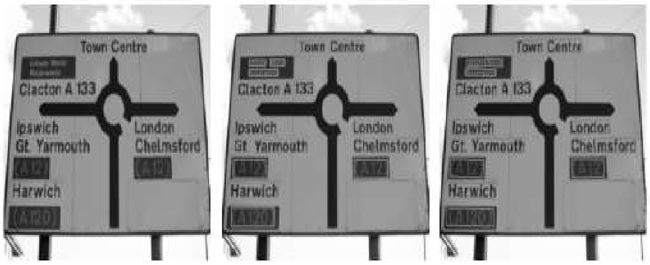

Fig. 6. ICDAR15+ICDAR13 Dataset

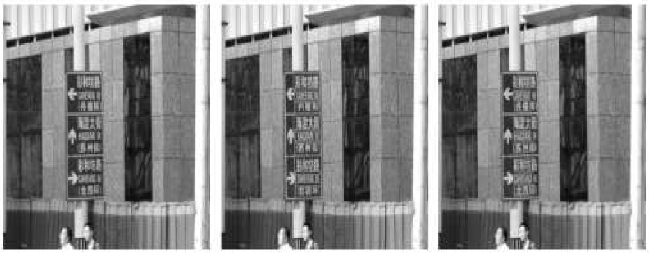

Fig. 7. MLe2e Dataset

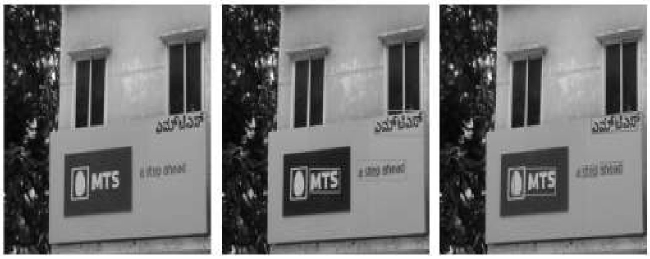

Fig. 8. MSRA-td500 Dataset

The proposed method achieved an impressive accuracy of 98.3% on inhouse multilingual video datasets, demonstrating

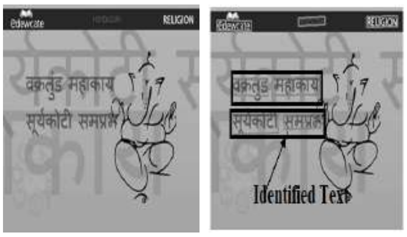

Fig. 9. Enhancement Not needed

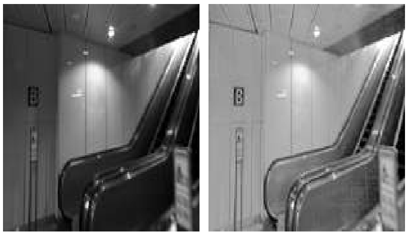

Fig. 10. Enhancement applied

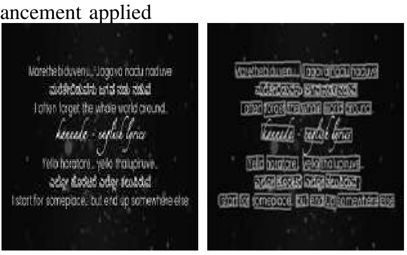

Fig. 11. Enhancement Not Needed

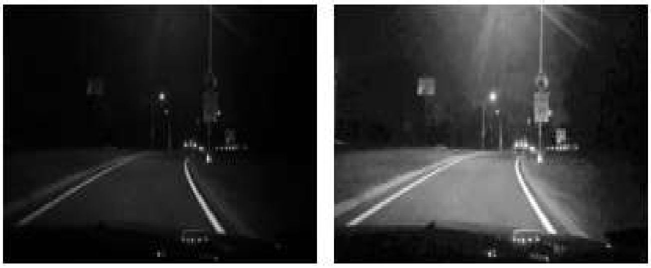

Fig. 12. Result of low light enhancement

robust performance on diverse content that includes images with combinations of Telugu and English, English-only, Sanskrit and Kannada, and Kannada and English.

A. Datasets

Text Detection: Datasets are indispensable in Deep Learning for training models, learning features, generalizing to new data, facilitating transfer learning, augmenting training data, evaluating model performance, addressing bias and fairness concerns, and gaining domain understanding. Table III provides information on various datasets utilized for text detection and text recognition tasks, with a focus on evaluating the proposed methodology using MSRA-td500, ICDAR13, ICDAR15, MLe2e, and Bi-lingual Text datasets [19].

Light-Enhancement:
To test and evaluate the light enhancement module, datasets considered are Low-Light (LOL), See-in-Dark (SID) [28], and Ex-dark [29]. Additionally, videos also incorporated from the RoadText1k dataset [18], which comprises 1000 video clips divided into 300 for testing, 500 for training, and 200 for validation. Due to its size, three videos are selected from the test set for analysis.

Performance Measurement:
In the realm of machine learning and deep learning, the assessment of model performance is accomplished through the analysis of obtained results. This evaluation involves the use of a confusion matrix, which classifies outcomes into four categories: True-Positive, True-Negative, False-Positive, and False-Negative. This classification is particularly relevant in the context of binary or multiclass classification tasks. In our specific scenario, we are engaged in binary classification for detecting text and non-text regions. The regions identified by the model are juxtaposed with the ground truth-values, and entries are allocated to the respective categories within the confusion matrix based on the comparison results. Inhouse dataset, consisting of 20 short videos, with the frame count varying across video clips, as specified in the table. All calculations were performed manually. The evaluation of text localization involved categorizing the outcomes into four groups.

- True-Positive (TP): Text regions with bounding boxes matching the Ground Truth
- True-Negative (TN): Accurate identification of non-textual regions consistent with the Ground Truth
- False-Negative (FN): Text regions undetected by the model.
- False-Positive (FP): Non-text regions erroneously identified as text by the model

By utilizing a video dataset consisting of 20 clips, with each clip containing between five to ten frames. The aim was to assess how effective our proposed approach is. In total, there were 114 frames in the dataset, and the overall total of ground truth values was 2538.

Experimental results shown in Table III has 2014 samples predicted as positive (TP) and 501 samples predicted as negatives(TN) and 37 samples are predicted as positive(FP). The accuracy of the model is computed using the formula 8:

$$Accuracy = \frac{(TP + TN)}{TotalSamples} \tag{8}$$

Applying Equation 8 to the provided data yields an accuracy of (2014 + 501) / (2014 + 37 + 501 + 5), which simplifies to 2515 / 2557, resulting in 0.9835 or 98.3%. According to the problem statement misclassification refers to the occurrence of overlapping bounding boxes during the detection of text regions, particularly those failing to recognize the no-space area between characters of different words. In our test dataset, the misclassification rate is determined as 0.16%

$$Miscalssification = \frac{(FP + FN)}{TotalSamples} \tag{9}$$

Misclassification is calculated using the formula 9, resulting in (37 + 5) / 2557, which simplifies to 0.016 or 0.16%. Precision, represents the accuracy of positive predictions which is computed using the equation 10 , Recall depicts the proportion of actual positives correctly predicted and it is computed using the equation 11 and harmonic mean of precision and recall, is determined by the equation 12.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

TABLE II
COMPARISON AMONG RECENT TEXT DETECTION AND RECOGNITION DATASET

Dataset	Detection Images			Orientation			Properties		Task	
	Train	Test	Total	H	Ar	Cu	Language	annotation	D	R
MSRA-td500 [20]	300	200	500	✓	✓	-	EN,CN	TL	✓	-
ICDAR13 [21]	229	233	462	✓	-	-	EN	w	✓	✓
ICDAR15 [1]	1000	500	1500	✓	✓	✓	ML	w	✓	✓
ICDAR17-MLT [22]	7200	9000	18000	✓	✓	✓	ML	w	✓	✓
ICDAR17-RCWT [13]	8034	4229	12263	✓	✓	✓	EN,CN	w	✓	✓
ICDAR19-ArT [23]	5603	4563	10166	-	✓	-	EN	W	✓	✓
MLe2e [22]	-	-	711	✓	✓	✓	ML	w,c	✓	-
IIT 5K [24]	2000	3000	5000	✓	-	-	EN	w	✓	✓
Bi-lingual Text [17]	583	54	637	✓	✓	✓	ML	w	✓	✓
KAIST scene text [25]	-	-	711	✓	✓	✓	ML	w	✓	✓
CUTE80 [25]	-	-	288	✓	✓	✓	EN	w	✓	✓
Text-ocrv0.1 [26]	21778	3232	25808	✓	✓	✓	ML	w	✓	✓
Total-Text [27]	1255	300		✓	✓	✓	EN	w	✓	✓

H- Horizontal, AR-arbitrary, MO-Multi-oriented, Cu-Curved, EN-English, CN-Chinese, ML-Multilanguage, w-word, c-character, d-detection, R-recognition and TL-TextLine

TABLE III
DETAILS OF IN-HOUSE DATASET AND FINDINGS

Flicks	Number of Frames	Ground Truth	True Positive	True Negative	False Positive	False Negative
1	11	249	200	43	11	0
2	8	132	84	50	0	0
3	6	271	129	142	0	0
4	14	104	82	26	4	0
5	4	56	44	12	0	0
6	4	28	28	0	0	0
7	3	123	112	11	0	0
8	3	22	18	3	2	0
9	5	65	40	15	10	0
10	5	170	58	112	0	0
11	5	20	17	5	0	0
12	5	12	10	2	0	0
13	3	45	39	6	0	0
14	5	690	685	0	0	5
15	7	70	60	0	10	0
16	9	196	158	38	0	0
17	6	210	204	6	0	0
18	6	11	9	3	0	0
19	2	46	19	27	0	0
20	3	18	18	0	0	0
Total	114	2538	2014	501	37	5

TABLE IV
COMPARATIVE ANALYSIS FOR TEXT DETECTION

METHOD	DATASET	PRECISION	RECALL	F SCORE
Ztext [5]	ICDAR15 [1]	0.84	0.63	0.72
	ICDAR13 [21]	0.835	0.783	0.782
EAST [6]	ICDAR15 [1]	0.835	0.783	0.782
	MLe2e [22]	0.835	0.783	0.782
Method by [8]	ICDAR13 [21]	0.882	0.884	0.883
Attention based convolution [9]	own dataset	0.95	0.89	0.9
Proposed BicTd	ICDAR13 [21]	0.786	0.7	0.79
	MLe2e [22]	0.887	0.776	0.761
	Bi-Lingual text [17]	0.957	0.871	0.911
	MSRA-td500 [20]	0.98	0.84	0.9
	inhouse	0.98	0.99	0.98

$$F-Score/Harmonic \quad Mean = \frac{2*Precision*Recall}{Precision+Recall} \tag{12}$$

The model gave obtained 98.2% precision, 99.7% recall and 98.5% harmonic mean.

Zero-Reference Deep Curve Estimation (Zero-DCE) The Zero-DCE [30] model available with Keras used for low-light image enhancement. The model aims to enhance the brightness and contrast of images by learning a set of pixel-wise adjustment curves without using reference or ground-truth images during training.Same model used in this work for comparison of the obtained result.

Related to light enhancement, metric used for comparison of low light are PSNR, SSIM, LOE and NIQE. A higher value of PSNR & SSIM indicated better quality. LOE and NIQE the lower the better. Peak Signal-to-Noise Ratio (PSNR) Equation 14, Structural Similarity Index(SSIM) Equation 13, Loss of Enhanceent (LoE), and Natural Image Quality Evaluation (NIQE) represented in the Equation 16, are terms in different fields. PSNR is a measure of the quality of an image or video restoration task, SSIM is used to measure the similarity of two images, LOE is a metric used to measure the degree to which the original lightness or brightness of an image is altered or lost after enhancement techniques are appliedin the implemented work this metric not used as it require reference image, and Naturalness Image Quality Evaluator (NIQE) is a widely used no-reference image quality assessment metric, often applied in light enhancement tasks to measure the quality of the enhanced image.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

- x and y are the two images being compared.
- μ_x andμ_y are the mean pixel values of x and y respectively.
- $\sigma_x^2\sigma_y^2$ are the variances of x and y respectively.
- σ_{xy} is the covariance of x and y
- C_1 and C_2 are constants to stabilize the division when the denominator is small.

$$PSNR = 10 \cdot \log_{10}\left(\frac{\mathrm{L}^2}{\mathrm{MSE}}\right) \tag{14}$$

- L is the maximum possible pixel value of the image.
- MSE is the Mean Squared Error between the two images given by the equation 15.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(I(i,j) - K(i,j))^2 \tag{15}$$

m and n are the dimensions of the images and I and K I(i,j) and K(i,j) are the pixel values at position (i,j).

$$NIQE = \sqrt{(x-\mu)^T \sum^{-1}(x-\mu)} \tag{16}$$

- x is the feature vector of the test image
- μ is the mean vector of the reference model
- $\sum$ is the covariance matrix of the refrence model

Visual comparison of sample image from LOL data set is shown in Figure 15 shows that IMG_ENC performed better in both qualitative and quantitative results and are determined in Table V and Figure 14. Low SSIM indicates better restoration quality whereas enhanced image looks more natural in case of low NIQE values. From the Table V it can be observed that Zero-DCE over-enhances brightness using deep curve estimation but does not reconstruct fine image details namely edges and textures. This causes a significant drop in SSIM and higher NIQE. Whereas IMG_ENC likely focuses on edge preservation or mild enhancement, preserving image structure i.e. SSIM but compromising brightness with the impact that the image looks more natural but lacks visibility enhancement. The proposed light enhancement method outperforms other low-light enhancement techniques in terms of NIQE as shown in Figure 13. The proposed method for text detection is contrasted with alternative algorithms in the table provided below. From Table IV, it is determined that our suggested approach works better when assessing the overall results of several algorithms, as per the table When evaluating on the ICDAR13 dataset used by EAST [6], proposed method achieves a higher F-score, indicating its effectiveness in overall performance. Compared to MLe2e, the proposed method achieved higher precision. Additionally, the recent bilingual text datasets produced impressive results when analyzed using our approach, highlighting its adaptability and effectiveness in handling diverse data types. /Proposed method also outperforms on in-house dataset comprising of 20 video clips. The out come of the Which , our method

TABLE V
QUANTITATIVE COMPARISON OF RESULTS OF LIGHT ENHANCEMENT USING PROPOSED TECHNIQUE

Dataset	Metrics	PSNR	SSIM	NIQE
LOL dataset(1018)	IMG_ENC	34.96	0.4783	5.301
	Zero_DCE	40.29	0.1956	7.33
Ex-Drak (78)	IMG_ENC	34.65	0.4557	3.435
	Zero_DCE	40.74	0.206	3.709

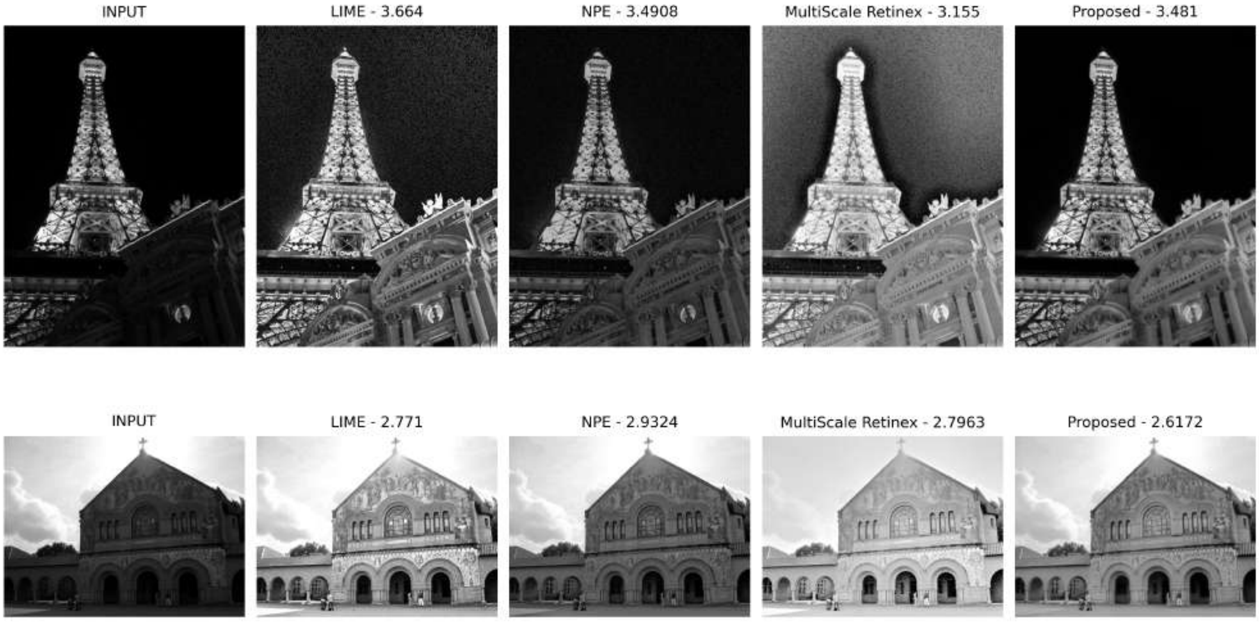

Fig. 13. Comparative analysis of LIME, NPE, MultiScale Retinex and Proposed Technique

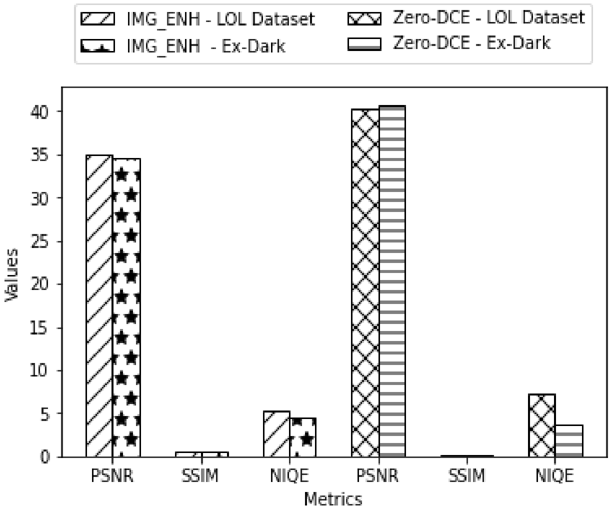

Fig. 14. Comparison of PSNR, SSIM and NIQE on IMG-ENC, ZERO-DCE on LoL and EX-DARK datasets

outperforms other algorithms with precision, recall, and F-score obtained values of 0.98, 0.99, and 0.98 respectively. This signifies a substantial quantitative improvement of 0.9% over the inspired method.Thus quantitatively performing 0.9% better than the inspired method [6]

Limitations

The suggested approach performs well on many datasets, but it has limitations when dealing with text containing curved or vertical elements, such as Arabic or Hebrew scripts. These scripts often have characters that are longer or more complex, with overlapping strokes and diacritics, making it challenging to accurately localize words. As a result, false positives may occur. To address this issue, the geometric shape calculation for bounding boxes, especially for RBOXes, needs to be improved, and better methods to handle overlapping bounding boxes need to be developed. Additionally, refining the word-level localization process can further enhance the accuracy of text detection.

Fig. 15. Visual comparison of IMG_ENH and Zero-DC

V. CONCLUSION

The experimental results demonstrate the effectiveness of the proposed approach across multiple languages, including English, Hindi, and Kannada. This paper focused on enhancing the algorithm's performance with low-light images by incorporating low-light restoration techniques. By mitigating the impact of flow light to improve text localization accuracy, particularly in challenging lighting conditions. It's important to note that our current approach does not specifically address issues related to blurred images or extremely small text. Future enhancements will involve refining proposed algorithms to minimize false positives and improving the geometric shape calculations for RBOXes. Additionally, ongoing experimentation involves integrating low-light restoration directly into the algorithm, enabling automatic light adjustment for locating text in low-light frames within videos.

ACKNOWLEDGMENT

Authors would like to express gratitude to staff and management of JNN College of Engineering and Manipal Academy of Higher Education for their support.

REFERENCES

[1] D. Karatzas and G.-B. et al., "Icdar 2015 competition on robust reading," in *Proc. 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1156–1160.

[2] M. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 243–255, 2005.

[3] S. Mahajan and R. Rani, "Text detection and localization in scene images: a broad review," *Artificial Intelligence Review*, vol. 54, no. 6, p. 4317–4377, aug 2021.

[4] H. T. Basavaraju, V. N. Manjunath Aradhya, and D. S. Guru, "Neighborhood pixel-based approach for arbitrary-oriented multilingual text localization," in *Proc. Intelligent Systems, Technologies and Applications*, Singapore Springer, 2020, pp. 1–12.

[5] A. e. a. Iqbal, K., "Ztext: zone based text localization in natural scene images," *International Journal of Computer Science and Network Security*, vol. 17, no. 4, pp. 306–314, 2017.

[6] X. Zhou and C. Y. et al., "East: An efficient and accurate scene text detector," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, 2017, pp. 2642–2651.

[7] L. Gómez and D. Karatzas, "A fine-grained approach to scene text script identification," in *Proc. 12th IAPR Workshop on Document Analysis Systems*, Greece, 2016, pp. 192–197.

[8] S. Saha and N. C. et al., "Multi-lingual scene text detection and language identification," *Pattern Recognition Letters*, vol. 138, pp. 16–22, 2020.

[9] A. K. Bhunia, A. Konwer, A. K. Bhunia, A. Bhowmick, P. P. Roy, and U. Pal, "Script identification in natural scene image and video frames using an attention based convolutional-lstm network," *Pattern Recognition*, vol. 85, pp. 172–184, 2019.

[10] X. Q. Feng, Q. Liu, and Z. W. Wang, "Port container number detection based on improved east algorithm," *Journal of Physics: Conference Series*, vol. 1651, no. 1, p. 012088, 2020.

[11] S.-X. Zhang, C. Yang, X. Zhu, and X.-C. Yin, "Arbitrary shape text detection via boundary transformer," *IEEE Transactions on Multimedia*, vol. 26, no. 1, pp. 1747–1760, 2024.

[12] W. Khlif, "Multi-lingual scene text detection based on convolutional neural network," Ph.D. dissertation, Universit'e de La Rochell, France, Jun 2022.

[13] B. Shi and Y. al., "Icdar2017 competition on reading chinese text in the wild (rctw-17)," in *Proc. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, USA, 2017, pp. 1429–1434.

[14] L. B. T. e. a. Ong Yi Ling, "Interpreting texts in natural scene images," *IAENG International Journal of Computer Science*, vol. 49, no. 2, pp. 595–605, 2022.

[15] B. W. Yuanbin Wang and Z. Duan, "Image dehazing based on sigmoid-guided filtering-retinex in nsct domain," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 4, pp. 641–650, 2024.

[16] Z. G. Xinling Feng, Cuiqing Zhu, "Research on low resolution digital image reconstruction method based on rational function model using neural networks," *IAENG International Journal of Computer Science*, vol. 51, no. 2, pp. 75–82, 2024.

[17] Veronica, "Bi-lingual text detection dataset," 2023, visited on 2025-04-13. [Online]. Available: Available at https://universe.roboflow.com/veronica/bi-lingual-text-detection

[18] S. Reddy and M. et al., "Roadtext-1k: Text detection & recognition dataset for driving videos," in *Proc. IEEE International Conference on Robotics and Automation*, France, 2020, pp. 11 074–11 080.

[19] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM International Conference on Multimedia*, United States, 2019, p. 1632–1640.

[20] J. Tetazoo, "Msra text detection 500 database tc11," 2012. [Online]. Available: http://www.iapr-tc11.org/mediawiki/index.php/MSRA-Text-Detection-500-Database-(MSRA-TD500)

[21] K. et al., "Icdar 2013 robust reading competition," in *Proc. 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493.

[22] N. Nayef, F. Yin, I. Bizid, and e. a. Choi, "Robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt," in *Proc. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, Japan, 2017, pp. 1454–1459.

[23] C. K. Chng and e. a. Liu, "Icdar-2019 robust reading challenge on arbitrary-shaped text - rrc-art," in *Proc. International Conference on Document Analysis and Recognition*, Australia, 2019, pp. 1571–1576.

[24] A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, 2012, pp. 2687–2694.

[25] T. Q. Phan and S. al., "Recognizing text with perspective distortion in natural scenes," in *Proc. IEEE International Conference on Computer Vision*, USA, 2013, pp. 569–576.

[26] A. Singh and P. et al., "Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8798–8808.

[27] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Japan, 2017, pp. 935–942.

[28] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. EEE/CVF Conference on Computer Vision and Pattern Recognition*, Utah, 2018, pp. 3291–3300.

[29] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[30] C. Guo and L. et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, 2020, pp. 1777–1786.