A vision transformer model for multilingual image-based text recognition

Abstract- Multilingual image-based text recognition is a tough problem with several practical applications. This work suggests an integrated ViT-YOLO model which integrates the strengths of the Vision Transformer (ViT) and You Only Look Once (YOLO) techniques to solve this challenge. The goal of the model is to correctly identify text in pictures with text in many languages. The ViT-YOLO model uses YOLO to locate text sections in pictures using patch extraction. Taking use of its robust image-understanding capabilities, the ViT model processes the derived patches for text recognition. To enhance the model's performance and robustness, a Generative Adversarial Network (GAN) is integrated for data augmentation. Experimental results demonstrate the superiority of the ViT-YOLO model over traditional methods and other deep learning models, achieving an impressive accuracy of 93.49%. These findings demonstrate that the proposed ViT-YOLO model holds significant promise in addressing multilingual text recognition challenges and paves the way for future advancements in multilingual image-based text recognition.

Keywords- Text recognition, Vision Transformer (ViT), You Only Look Once (YOLO), Generative Adversarial Network (GAN), multilingual-text recognition.

1 Introduction

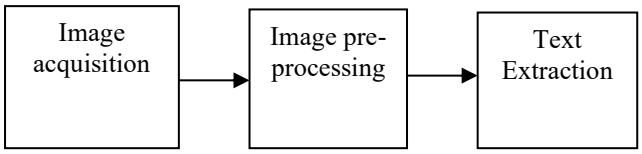
Recent developments in digital technology have made it possible to shoot photographs using various mobile devices. As a direct consequence of this, the quantity of images taken by users continues to rise daily. Although device-generated annotations are useful, they often remain the only option accessible for annotation purposes [1]. The text of images conveys crucial information about the image's meaning. Content-based picture retrieval, intelligent navigation, and automatic translation are examples of the many uses for annotated photos [2]. Many images include text, but the language of that writing is unknown in advance, or many languages are represented inside the same image's textual regions. The following genuine scientific topic must be answered: efficiently identifying and recognizing text in scene photographs. Text detection is a necessary first step that must be taken before moving on to text recognition. This study investigates the text detection issue within this study's scope.

Transformers have obtained extraordinary success in computer vision [3] for a variety of image-based operations, including identifying objects [4], segmentation based on semantics [5], and recognition of images [6,7]. Since transformers were first designed for Natural language processing (NLP) [8], it is labor-intensive to construct and train them for visual tasks [9, 10]. One way to drastically cut down on the amount of computing power needed is to apply a pure Vision Transformer (ViT) on a series of picture patches [11].

Recognizing text in a picture that contains more than one language is a formidable task, especially when the images involved are of varying scales, orientations, and resolutions. This study suggests an enhanced Transformer design to address the problems. The input for the supplemental pyramid transformer is not just single picture patches like those used by the main vision transformer but also representations of the nearby image patches [12]. In order to reduce the computational complexity of the proposed model, the main and secondary vision transformers learn the representation of multilingual scene text characteristics together by exchanging their parameters.

Text in natural situations is read by machines using area detection and recognition [13]. A deep learning model that could detect text in pictures with several languages is called a vision transformer (ViT) model. The model is built on top of the transformer framework, which has seen extensive usage in NLP applications. The ViT model's backbone is a convolutional neural network (CNN) responsible for feature extraction from the input picture. A transformer encoder then processes the features, which generates the recognized text [14]. A common encoder is responsible for the model's capacity to process input pictures in many languages. The encoder is pre-trained on a large multilingual text corpus with masked language modeling which aims to anticipate gaps in the text. This pre-training technique enables the encoder to acquire language-independent representations of text characteristics. These representations might then be used

to recognize text in pictures that comprise any of the languages that have been pre-trained. The procedure for translating text based on images into many languages is shown in Fig 1.



1. Multilingual Image-Based Translator [15].

The ViT model performs better on many benchmark datasets for multilingual image-based text recognition. Because of its potential utility in areas as diverse as document digitization, Image captioning, and multilingual text analysis, the ViT model for multilingual image-based text recognition is gaining attention as a promising solution to the problems associated with recognizing text in images that contain more than one language [16]. It is possible to recognize and transcribe the text in pictures comprising many languages using the Vision transformer (ViT) Model for multilingual image-based text recognition. Vision transformer (ViT) is a deep learning model that can extract and decode information from an image and decode them into text. The ViT Model provides businesses with a potent instrument for multilingual text recognition, which could conserve time and effort while enhancing accuracy and scalability.

To properly train a multilingual image-based text recognition system, a large amount of data in the form of text image samples and their annotations is required. Increasing the amount of training data through data augmentation techniques is one approach to addressing this problem. Using Generative Adversarial Networks (GANs) to augment data for image-related tasks has become common practice. However, unique challenges must be overcome before GANs could be used for data augmentation in text recognition. The proposed study employs Ghost Attention Module (GhoAM) to generate feature maps from the resulting images. Afterward, YOLO is used for object detection based on the final feature representation. Altogether, the Ghost Attention Module and Feature Fusion Approach boost the object detection system's accuracy through better feature representation of the images. This study introduces a ViT model for multilingual image-based text recognition systems and an adaptive data augmentation strategy based on GANs to address the class imbalance in text recognition problems and improve accuracy and performance. The accuracy of the resulting text recognition systems is higher than that of systems trained using more conventional methods because of the ViT model.

- To design and optimize loss functions for a GAN to improve the performance of the model by pre-processing the language-based image dataset and generating more training samples, thereby enhancing image generation and dataset processing.
- Design and optimize loss functions for YOLO to improve ROI extraction and automatically extract relevant features from images, enabling accurate object localization and object classification.
- To use the Ghost Attention Module to enhance the feature extraction step and enable the ViT model to classify the images with higher accuracy and lower loss.
- To classify pictures using the returned characteristics, the ViT method necessitates the development of a classification model.
- To use a variety of measures to evaluate how well the trained model performs on the testing dataset.

2 Literature of review

In this part, the author will take a quick look at the many literature reviews that have been offered by different writers. Text recognition has received a lot of attention and study in recent years, especially in the areas of computer vision and natural language processing. The literature review covers a range of works that provide novel approaches to text recognition and detection. Research in these areas is conducted to find solutions to problems including document analysis, scene-wide architecture, and object identification.

Cho et al. (2023) [17] suggested a paradigm for filter captioning that could create captions for photos by adapting a standard object language to the needs of certain fields. Individuals with visual impairments could benefit from the model's ability to translate visual material into text and voice. Instead of using the learning data to build unique picture captions, the author suggested shifting the dictionary's focus to the area of object vocabulary. However, the model is restricted in its caption generation since it uses a static object vocabulary.

Wu et al. (2022) [18] introduced an end-to-end adaptable video textual Detection, Evaluation, and Recognition (TransDETR) framework that recognizes each text via a particular query called "text query" over an extended period. The proposed TransDETR avoids the pitfalls of the adjacent-frame explicit match paradigm by implicitly following and recognizing each text. Extensive studies on four media text datasets showed that TransDETR outperforms the state-of-the-art systems on identification, tracking, and spotting jobs. TransDETR's

reliance on the precision and efficacy of the text queries itself is one of its drawbacks when it comes to text recognition using text inquiries. TransDETR's performance might be harmed, producing erroneous or insufficient recognition results, if the text queries are poorly specified or miss key aspects of the text.

Huang et al. (2022) [19] proposed a novel scene-wide architecture for text detection named SwinTextSpotter. The author employed a unique Recognition Conversion technique to directly guide text localization via recognition loss, which integrates the two tasks using a transform encoder with an evolving head as the detector. Experiments on the multilingual dataset and the arbitrary-shaped datasets showed that SwinTextSpotter greatly outperforms prior approaches. However, a limitation of this model is its dependency on accurate recognition results for effective text localization.

Li et al. (2022) [20] presented a digital imaging technician (DiT) model that uses a large collection of unlabeled text images, and a Text Image Transformer system was trained for Document Categorization Artificial Intelligence (AI) tasks. Several vision-based Documents AI tasks, such as picture classification, Optical Character Recognition (OCR), Layout Analysis Table, and Text Detection, are accomplished using DiT as the foundational network. The self-administered already trained DiT method outperforms state-of-the-art baselines on downriver tasks such as document image categorization, record layout analysis, table identification, and text identification for optical character recognition. The DiT model's effectiveness could be limited in scenarios where labeled data is scarce or difficult to obtain, hindering its applicability in certain real-world settings.

He et al. (2022) [21] proposed a Scene Text Recognition (STR) method that executes text argumentation based on visual semantics. The author created a subgraph for every instance, where nodes indicate the pixels, and edges are inserted between clusters based on their spatial similarity. The nodes at the bottom of each subgraph are linked to form a full graph. The author trained a convolutional graph system under cross-entropy loss supervision to use this Graph for Textual Reasoning (GTR). The model, S-GTR, emulates GTR within a language model within a segmentation-based STR baseline, which mutually learns to make the most of the visual-linguistic complementarity. S-GTR establishes new state-of-the-art on six difficult STR benchmarks and generalizes effectively to multilingual datasets. Despite its success in challenging STR benchmarks and multilingual datasets, a limitation of this model is its sensitivity to complex scene variations and occlusions, which can affect its recognition performance.

Li et al. (2021) [22] proposed the Transformer architecture for text recognition, which has shown superior performance over state-of-the-art models for all three types of text recognition (printed, handwritten, and scene). The TrOCR model is easy to understand and use and can be pre-trained on massive amounts of synthetic data before being refined with human-labeled datasets. However, a limitation of the proposed model is its high computational resource requirements, which might hinder its practical implementation in resource-constrained environments.

Kim et al. (2021) [23] suggested a new Visual Display Unit (VDU) model for document interpretation tasks that can be trained from start to finish without the need for an OCR framework. The proposed method exhibits state-of-the-art performance on a wide range of document interpretation tasks on publicly available benchmark datasets and proprietary industrial service datasets. The author also showed the efficacy of the suggested approach, especially when applied to a real-world scenario, through extensive experimentation and analysis. However, a limitation of the proposed model is its dependency on the quality and variability of the input documents, which might affect its performance in certain real-world scenarios.

Orihashi et al. (2021) [24] proposed a comprehensive contextual text analysis training approach for end-to-end scenario text recognition that allows for high recognition accuracy. The approach is based on the encoder-decoder paradigm based on Transformer, which requires large image-to-text matching datasets in the target language for substantial training and a highly accurate end-to-end model. The suggested approach gets around this problem by training the lacking in resources encoder-decoder model using clean, large datasets in languages with plenty of available data, such as English. The usefulness of the suggested strategy is demonstrated experimentally on a small publically data set for Japanese scenario recognition of text is currently available. The proposed model requires large datasets in the target language for substantial training and a highly accurate end-to-end model, which could be a limitation when such resources are not available.

Yang et al. (2020) [25] proposed a straightforward method with considerable potential for scene recognition of text. The author immediately links two-dimensional convolutional neural network (CNN) algorithms to an attention-based sequential decoder directed by holistic representation, eliminating the requirement for converting image inputs to sequence representations. The model improves acceleration and provides competitive or state-of-the-art recognition results across a range of regular and uneven scene textual benchmark datasets. However, it

might struggle with capturing long-range dependencies and context information across the text. This limitation can affect the model's ability to accurately recognize and understand complex textual patterns within scenes.

Hu et al. (2020) [26] proposed a text-based Visual Question Answering (TextVQA) model, which uses a multisensory transformer architecture and a deep representation of text within images. The author's use of self-attention to represent both the inter- and intra-modality context is what makes this model's mode-blending so natural. It also supports iterative response decryption using a dynamic points network, which allows for multi-step prediction rather than single-step classification. The suggested model achieves state-of-the-art performance on three TextVQA benchmark datasets. However, the model might struggle in real-world applications due to questions that are very complex or ambiguous for it to answer.

These studies show how different methods, such as the Transformer architecture, VDU model, contextual text analysis training approach, attention-based sequential decoder, and multisensory transformer architecture, can be effective in the field of text recognition, providing valuable insights into the field. The performance of text recognition has been greatly enhanced by these methods, and they offer promise for further improvement. Finally, this research provides unique and efficient approaches for text identification and detection that could aid in a variety of applications, such as aiding persons with visual impairments, analyzing documents, and monitoring traffic in real-time. These papers contribute significantly to the state of the art in text recognition and detection with their novel approaches.

2.1 Research Contribution:

- **Integrated ViT-YOLO Model:** This study proposes a unified model called ViT-YOLO to bridge the gap between the Vision Transformer (ViT) and You Only Look Once (YOLO) approaches. Combining the two methods allows the model to overcome the challenges of multilingual image-based text recognition and make use of the strengths of each more effectively.
- **Efficient Text Region Localization:** The ViT-YOLO model employs the YOLO method for fast patch extraction, which allows precise text area localization in pictures. The approach improves text recognition by pinpointing the relevant text sections.
- **Powerful Image Understanding:** The ViT-YOLO model makes use of the Vision Transformer (ViT) to efficiently process the extracted patches for text recognition. The ViT model outperforms its competitors because of its superior image-understanding skills, which help in the correct recognition and interpretation of multilingual text.
- **Robustness through GAN-based Data Augmentation:** The model's efficiency and reliability are improved by using a Generative Adversarial Network (GAN) to enrich the available data. Since the GAN creates synthetic data, the model could expand its training data and learn from additional scenarios.
- **Practical Applications:** Document translation, data retrieval, and auto-subtitling are just some of the many real-world uses for multilingual image-based text recognition. The suggested ViT-YOLO model is a powerful resource for various uses due to its high potential and high performance.

In conclusion, a unique integrated ViT-YOLO model is presented here for multilingual image-based word recognition. The model's ability to accurately localize text regions, comprehend complex images, remain stable despite GAN-based data augmentation and perform at a high level all point to its promise as a solution to the problems plaguing this area. This study paves the way for the creation of future practical applications that could make use of accurate and efficient multilingual image-based text recognition.

The Vision Transformer (ViT) model has difficulty with multilingual image-based text recognition because it requires large amounts of labeled training data. To obtain competitive performance, ViT models often need to pre-train on and fine-tune against enormous datasets. In multilingual settings, where labeled data could be poor or unavailable for certain languages, this might be a considerable difficulty. Since multilingual image-based text recognition requires identifying and interpreting text in several languages, it might be time-consuming, costly, and logistically challenging to gather and annotate varied and extensive datasets for each language. Due to its dependence on labeled training data, the ViT model struggles to accomplish accurate recognition in languages with sparse or asymmetrical data availability.

3 Background study

The ability to recognize text in natural environments, such as labels on objects, road signs, and instructions, is called scene text recognition (STR). Machines can use STR to make educated judgments like what to pick up, where to go, and what to do next with the use of this technology. Recognition precision has consistently been the target of the STR study. The importance of speed and computing efficiency, especially for mobile robots with limited resources, is often overlooked.

Rowel Atienza presented ViTSTR, a vision transformer (ViT)-based STR with a single-stage model architecture that is both computationally and parametrically efficient. Compared to a competitive, strong baseline approach like Thin-Plate-Spline ResNet-BiLSTM-Attention (TRBA), which achieves an accuracy of 84.3%, The compact ViTSTR obtains an accuracy of 82.6% (84.2% with data augmentation) at 2.4x speed up, while using only 43.4% as many parameters and 42.2% as many FLOPS. The compact implementation of ViTSTR outperforms the original by a wide margin, with an accuracy of 80.3% (82.1% with data augmentation) achieved at 2.5 the speed with only 10.9% as many parameters and 11.9% as many FLOPS.

The suggested baseline ViTSTR achieves 85.2% accuracy (83.7% without augmentation) at 2.3x the speed of TRBA, although requiring 73.2% more parameters and 61.5% more FLOPS. This is achieved by using information from external sources. Nearly all possible ViTSTR combinations are on the cutting edge of simultaneously optimizing accuracy, speed, and computing efficiency [27].

4 Preliminaries

In this section, the primary tools considered are the GAN, Ghost attention module (GhoAM), and integrated ViT-YOLO model that forms the basis of the proposed approach. The GAN is employed for data augmentation, GhoAM is used to enhance the representation of features in the augmented input images whereas the integrated ViT-YOLO model is responsible for training and testing the dataset, as networks with excellent detection accuracy can be built using the ViT network.

4.1 Generative Adversarial Network (GAN)

The use of GAN in the context of the Hybrid Model Layer-Wise Vision Transform can bring additional benefits to the YOLO model. GANs are a class of deep learning models consisting of a generator network and a discriminator network that work together competitively.

GANs might produce synthetic data samples that mimic actual pictures when used with the Hybrid Model Layer-Wise Vision Transform. Additional training data might be generated using a GAN's generator network, expanding the existing dataset and enhancing the model's generalizability to new situations.

4.2 Vision Transformer (ViT)

The Vision Transformer (ViT) is a state-of-the-art model for picture categorization tasks, and it has a stellar reputation for its efficacy. Instead of the conventional use of convolutional layers, it makes use of self-attention processes. The ViT uses a patch-based approach, partitioning the input picture into smaller patches that are then encoded using a transformer. The idea behind the ViT is to pay attention to each patch in a picture so that the author could learn about the image as a whole. The model's ability to consider interdependencies and connections between individual patches is what enables it to grasp the bigger picture. The ViT is capable of strong and precise image classification because it includes self-attention, which allows it to effectively acquire both local and global picture features.

4.3 GhostNet

GhostNet is a convolutional neural network (CNN) architecture that is incredibly efficient and lightweight, making it ideal for applications with limited resources. It was built with efficiency and low computational requirements in mind. GhostNet's most ground-breaking feature is the Ghost Module, which was developed specifically for it. There are two principal branches in the Ghost Module: the precise branch and the approximate branch. The main stem is accountable for discovering expressive representations that accurately reflect the facts. The goal of the approximation branch, on the other hand, is to get close to the results of the main branch while consuming fewer CPU cycles. GhostNet uses the Ghost Module to drastically reduce the model's computational complexity without sacrificing its performance. The main branch learns to accurately capture complex information, whereas the approximation branch learns to behave similarly but with less processing power. Because of this method, GhostNet can find a happy medium between model efficiency and performance, making it a great option when computing power is restricted.

4.4 Hybrid model

You Only Look Once (YOLO) object detection models might benefit from a new method that combines the best features of the Vision Transformer (ViT) with the GhostNet architecture: the Hybrid Model Layer-Wise Vision Transform. This method significantly enhances representation learning by introducing a layer-wise attention mechanism. Let's take a deeper look at the YOLO model's ghosted attention layers and how the Hybrid Model Layer-Wise Vision Transform is applied there to see how it impacts the model's performance.

- **Ghosted Attention Layer:** In the YOLO framework, the Ghost Module and the attention mechanism are combined in a single layer known as the "ghosted attention layer". Combining the benefits of convolutional operations with self-attention, this layer uses them on the input features.

- **Hybrid Model Integration:** The Hybrid Model Layer-Wise Vision Transform seamlessly integrates the principles of both the Vision Transformer (ViT) and GhostNet into the ghosted attention layer. By incorporating aspects of both architectures, it leverages the efficiency of GhostNet and the powerful representation learning of ViT.
- **Layer-Wise Attention:** The Hybrid Model Layer-Wise Vision Transform introduces a crucial enhancement through the layer-wise attention mechanism. Instead of globally attending to all patches within the image, the attention mechanism is applied individually for each ghosted attention layer. This layer-wise attention enables the model to capture local and global contextual information at multiple scales, enriching the representation learning process.

By applying the Hybrid Model Layer-Wise Vision Transform to each ghosted attention layer, the YOLO model becomes more proficient in detecting objects within images. This technique effectively combines the efficiency of the GhostNet architecture with the robust representation learning capabilities of the Vision Transformer. Consequently, the YOLO model as shown in Fig 2 exhibits improved object detection performance, offering enhanced contextual understanding and increased accuracy.

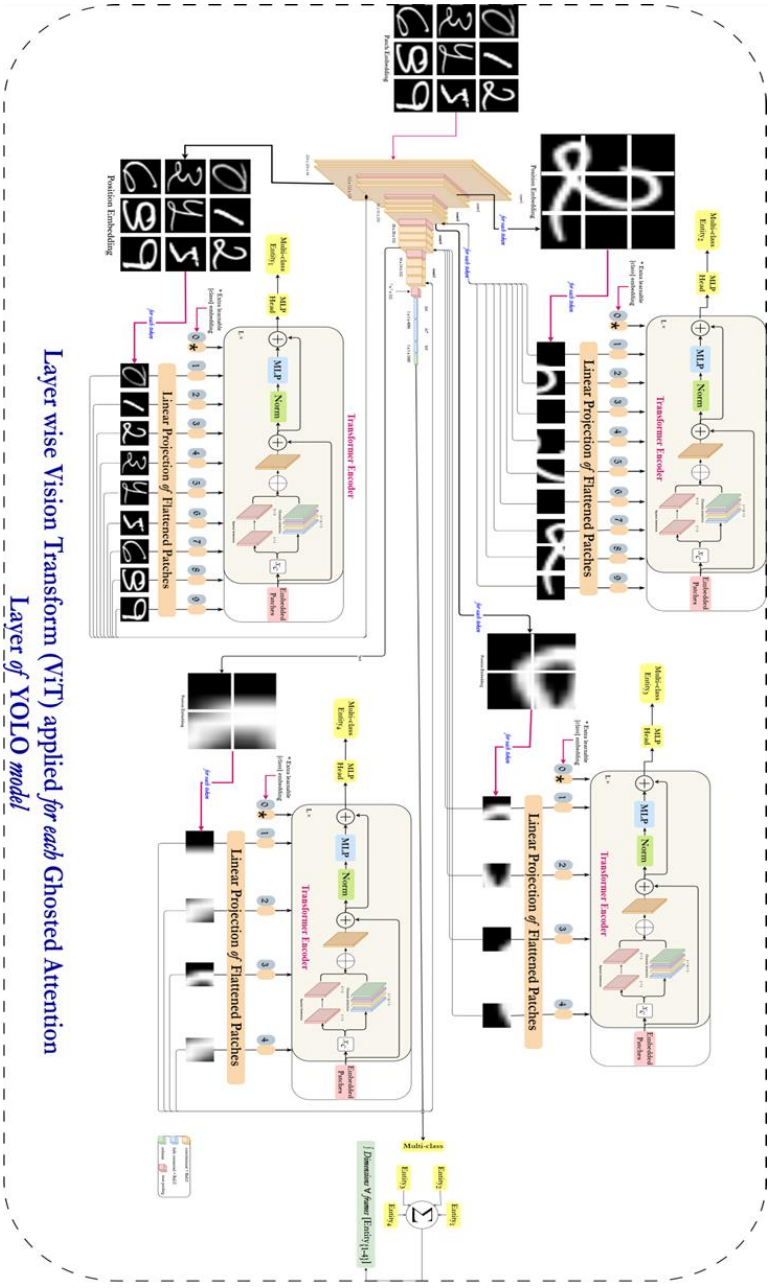


Fig. 2 2. Proposed architecture

5 Proposed methodology

The proposed methodology for multilingual image-based text classification consists of several phases. In the first phase, a dataset is collected, which includes images containing text in multiple languages. The second phase involves pre-processing the acquired images by resizing them, normalizing them, and applying various data augmentation techniques to increase the dataset's diversity. In the third stage, data enhancement is accomplished with the help of a Generative Adversarial Network (GAN). Using random noise as input, the GAN's generator network creates artificial pictures, while the discriminator network tells the difference. An objective function is used to optimize the GAN such that the gap between the two distributions of actual and produced images is as small as possible. Fourth, the enhanced pictures undergo processing by a Ghost Attention Module (GhoAM). This component creates two different feature maps, each of which captures important data and improves the portrayal of text areas in the photos.

In the fifth stage, the author employs the GhoAM-obtained fused feature maps to perform feature extraction. Images with text on them are fed into a mixture of the YOLO model and the Vision Transformer (ViT) to have their text characteristics extracted. The sixth step involves applying the extracted text sections to a classification problem involving several languages. A language-specific or multilingual text classifier is used for each language in the sample. The collected text features and their labels are used to train the classification models together with suitable loss functions like cross-entropy loss. The seventh stage involves checking the quality of the trained categorization models. Accuracy, precision, recall, and F1-score, among other measures, are computed for each language using a unique testing dataset. The accuracy of the multilingual text categorization system is evaluated as a whole by averaging the results across all languages.

The major steps included in designing the proposed methodology are given as follows:

- Data pre-processing: Make sure the images in the language-based image dataset can be used to train the model by cleaning and pre-processing them. This might require scaling, normalizing, and adding information.
- GAN for data augmentation: Additional training samples can be generated with the help of GAN. The model's efficiency can be boosted due to the increased variety of training data this provides.
- Ghost Attention Module: Compute feature maps using a Ghost Attention Module and fuse them using a feature fusion approach: F = fusion (GhoAM(I_aug, θ_a1), GhoAM(I_aug, θ_a2), method) where θ_a1 and θ_a2 are the Ghost Attention Module's parameters for the first and second set of feature maps, respectively.
- Image ROI extraction: Determine the most important elements of each Image in the collection and use that information to create a Region of Interest (ROI).
- Feature extraction using YOLO: Extract features from the images with the help of the YOLO algorithm. YOLO is an effective image feature extraction and real-time object detection system.
- Classification using ViT model: Image classification can be accomplished using a ViT model that has been trained on the training dataset. Adjust the parameters of the model to minimize the cross-entropy loss function.
- Performance evaluation: Evaluate the trained ViT model on the testing dataset. Use metrics like accuracy, precision, recall, and F1-score to assess the model's performance.

The flow chart given below in Fig 3 illustrates the working of the proposed methodology.

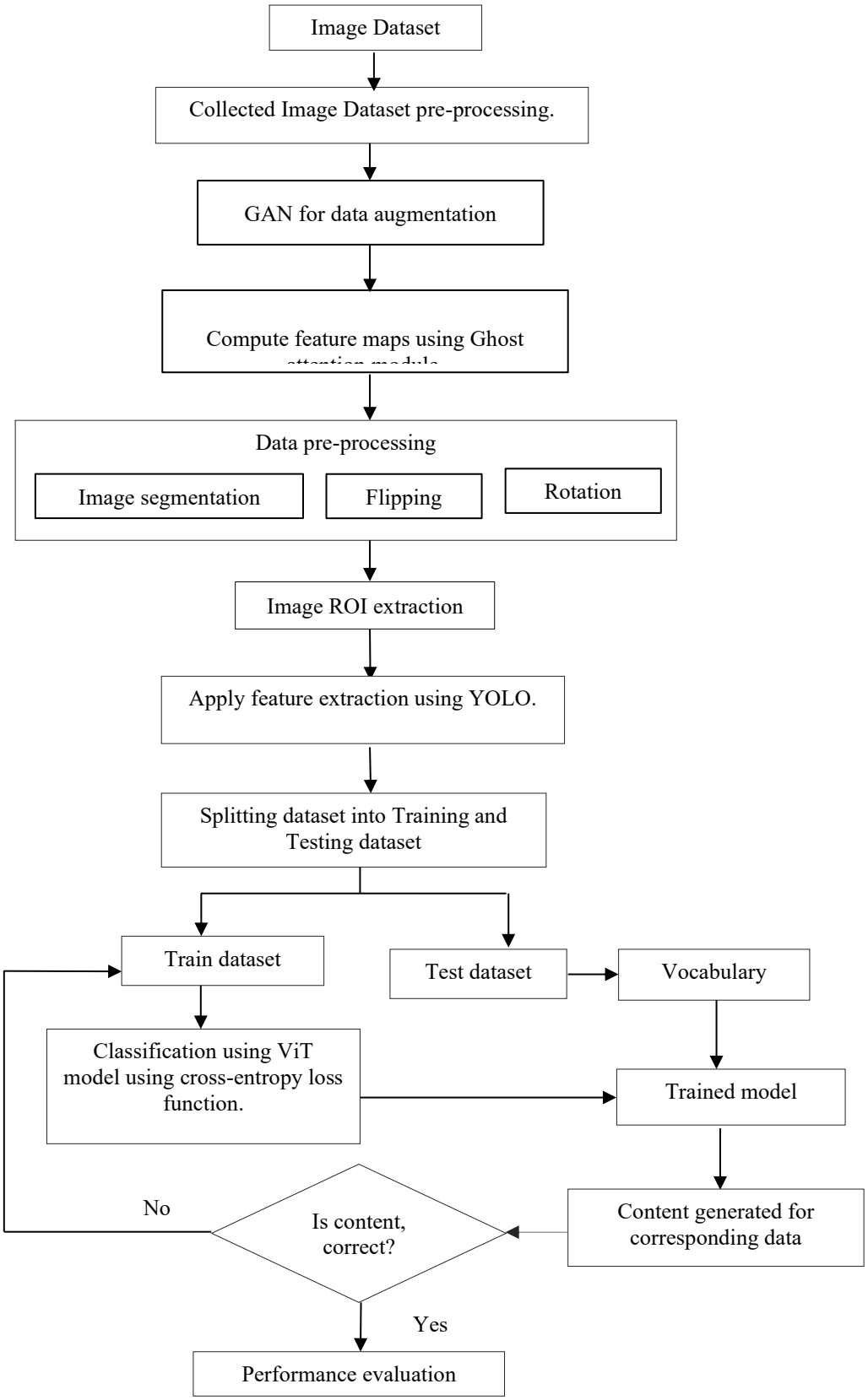


Fig. 3 3. Proposed methodology

6 Proposed algorithm

ALGORITHM: | **Multilingual Text Recognition**

Start

Phase I**:** Data Acquisition:

Step 1: Let D = {d1, d2, ..., dn} represent the dataset of multilingual image-based text data.

Phase II**:** Data pre-processing:

Step 2: Let R(d) denote the resized image of data d.

Step 3: Let N(R(d)) denote the normalized image obtained by subtracting the mean (μ) and dividing by the standard deviation (σ) of R(d).

Step 4: Let A(N(R(d))) represent the augmented image obtained through various techniques.

Phase III**:** GAN for data augmentation

Step 5: Let G(z; θ_g) represent the generator network that generates synthetic images given random noise z.

Step 6: Let D(x; θ_d) represent the discriminator network that distinguishes between real and generated images.

Step 7: Let J(G, D) represent the objective function used to optimize the GAN.

Phase IV**:** Ghost Attention Module:

Step 8: Let F1 = GhoAM(A(N(R(d))); θ_a1) represent the first set of feature maps obtained by applying the GhoAM to the augmented images.

Step 9: Let F2 = GhoAM(A(N(R(d))); θ_a2) represent the second set of feature maps obtained by applying the GhoAM to the augmented images.

Phase V**:** Feature extraction using YOLO through ViT

Step 10: Let FF = Fuse(F1, F2) represent the fused feature maps obtained by a feature fusion approach.

Step 11: Let FE(FF) represent the feature extraction process using YOLO with the Vision Transformer (ViT).

Step 12: Fuse the two sets of feature maps using a feature fusion approach: F = fusion(F1, F2, method).

Phase VI**:** Multilingual Text Classification:

Step 13: Let T = ExtractText(FE(FF)) represent the extracted multilingual text regions from the image features.

Step 14: For each language l in the dataset, let C_l(T) represent the classification model for language l, which is suitable for multilingual text classification.

Step 15: Let L_l(C_l(T)) represent the loss function used for training the classification model C_l(T) for language l.

Phase VIII**:** Performance evaluation:

Step 16: Let T_test represent the testing dataset for performance evaluation.

Step 17: For each language l in the dataset, calculate:

Step 18: Acc_l = (number of correctly classified samples for language l) / (total number of samples for language l)

Step 19: P_l = (true positives for language l) / (true positives + false positives for language l)

Step 20: R_l = (true positives for language l) / (true positives + false negatives for language l)

Step 21: F1_l = 2 * (P_l * R_l) / (P_l + R_l)

Step 22: To obtain an overall evaluation of the multilingual text classification system, calculate:

Step 23: Acc_overall = (sum of Acc_l for all languages) / (total number of languages)

End

7 Result and discussion

7.1 Dataset description

This study considered Devanagari handwritten digits from the MNIST dataset, the MJSynth (MJ), and the Arabic-Handwritten-Chars datasets for multilingual image-based text recognition.

- **MJSynth (MJ)-** MJSynth, sometimes known as MJ, is a synthetically generated dataset that contains 8.9

million images of words that look realistic. MJSynth was built with three layers in mind from the beginning: the background, the foreground, and an optional shadow or border. It utilizes a total of 1,400 unique fonts. Various variations exist in the typeface's kerning, weight, underlining, and other attributes. MJSynth additionally uses various backdrop effects, natural image mixing, rendering of borders and shadows, projected distortion, base coloring, and noise [28]. The sample images from the MJSynth dataset are given in Fig 4.

Fig. 4 4. Synthetic image samples from MJSynth

- **MNIST-** While resources for handwritten image data are abundant for languages like English and Japanese, they are scarce for many Indian languages like Hindi/Devanagari. To develop a customized version of MNIST, researchers explored the UCI (University of California) Machine Learning Repository and came upon the Hindi Handwritten characters dataset by Shailesh Acharya and Prashnna Kumar Gyawali [29]. However, only digits zero through nine (and no additional characters) remain in this collection. GrayScale Image is the data type, whereas PNG is the image format. The images have a 32 ×32-pixel resolution, with the actual character occupying 28 × 28 pixels with 2-pixel padding on all four sides. The Test set has roughly 300 images for each class, whereas the Train set contains roughly 1700 images [29]. The sample images from the MNIST dataset are given in Fig 5 as follows:

Fig. 5 5. Synthetic image samples from MNIST

- **Arabic Handwritten-Chars-** This dataset consists of 16,800 Arabic characters written by 60 participants aged 19 to 40, with a right-hand bias in 90% of the writers. Each person filled out two sheets with ten copies of each letter, from rom' alef' to' yeh.' A training set consisting of 13,440 text and 480 images per class and a separate test set consisting of the remaining data (3,360 characters to 120 images per class) make up the database. Both the training set and the test set are written by different authors. To make sure that the test set does not contain too many examples from any one source, the inclusion of writers to the test set was randomly arranged (to ensure variability of the test set) [30]. The sample images from the Arabic Handwritten-Chars dataset are given in Fig 6 as follows:

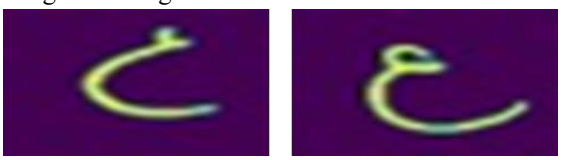


Fig. 6 6. Synthetic image samples from Arabic Handwritten-Chars

Based on the data given in Table 1, it appears that the integrated Vision Transformer (ViT) and YOLO model outperforms the other models in terms of accuracy. The proposed ViT-YOLO + GAN model achieved an accuracy of 93.49%, which is significantly higher than the other models, with the closest competitor being the TRBA model at 84.3%.

Table 1. Comparison table for various models

Model	Speed	FLOPS	Parameters	Accuracy
GRCNN [31]	11.2	1.8	4.8	78.3
Rosetta [32]	5.3	10.1	44.3	78.4
CRNN [33]	3.7	1.4	8.5	76.7
RARE [34]	18.8	2.0	10.8	82.1
ViTSTR-Base+Aug [27]	9.8	17.6	85.8	85.2
STAR-Net [35]	8.8	10.7	48.9	81.8
R2AM [36]	22.9	2.0	2.9	78.4
TRBA [37]	22.8	10.9	49.6	84.3
Proposed ViT-YOLO + GAN	10.37	6.8	89.4	93.49

- **Based on accuracy-** It could be observed from the graph given in Fig 7 that ViT-YOLO + GAN has achieved the highest accuracy of 93.49%, surpassing all the other models listed. The proposed ViT-YOLO + GAN has shown the highest accuracy among the listed models for multilingual image-based text recognition. The use of deep learning techniques such as ViT and YOLO, along with GAN for data augmentation, appears to be effective for achieving high accuracy in this task.

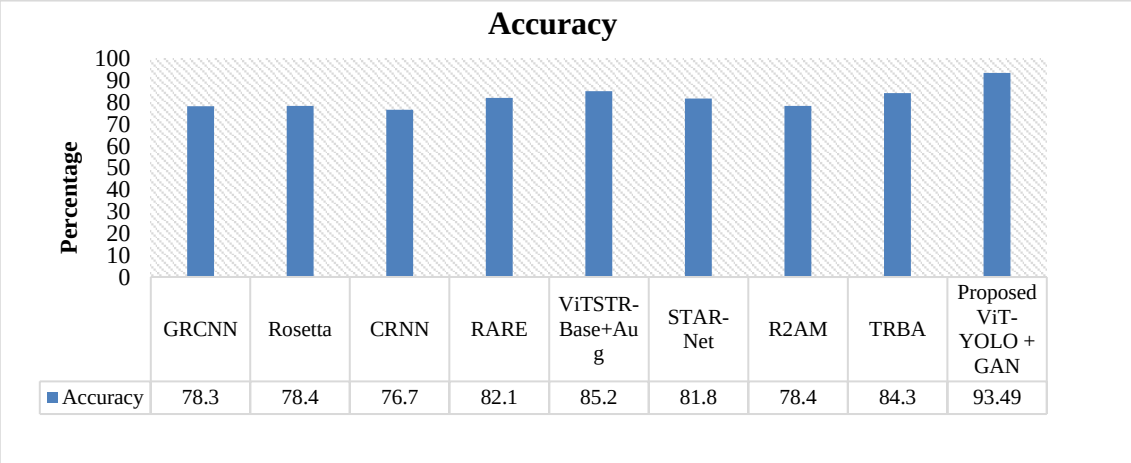


Fig. 7 7. The accuracy obtained by various traditional models and proposed models.

- **Based on Speed-** It could be seen from the graph given in Fig 8 that in terms of speed, the ViTSTR-Base+Aug model was the fastest, with a speed of 9.8. The proposed ViT-YOLO + GAN model had a speed of 10.37, which is not the fastest but still relatively fast compared to some of the other models.

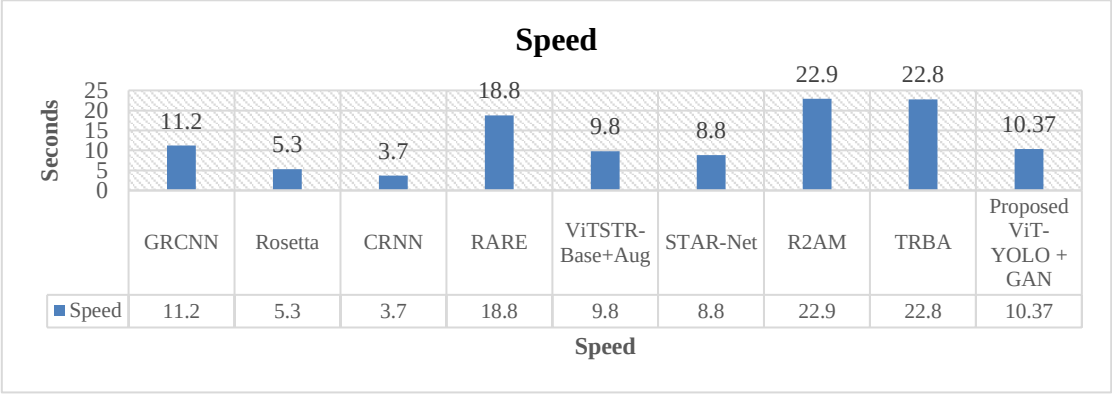


Fig. 8 8. The speed of various traditional models and proposed models.

- **Based on FLOPS-** When it comes to FLOPS (floating-point operations per second), the proposed ViT-YOLO + GAN model had the lowest at 6.8, while the Rosetta model had the highest at 10.1 as shown in Fig 9.

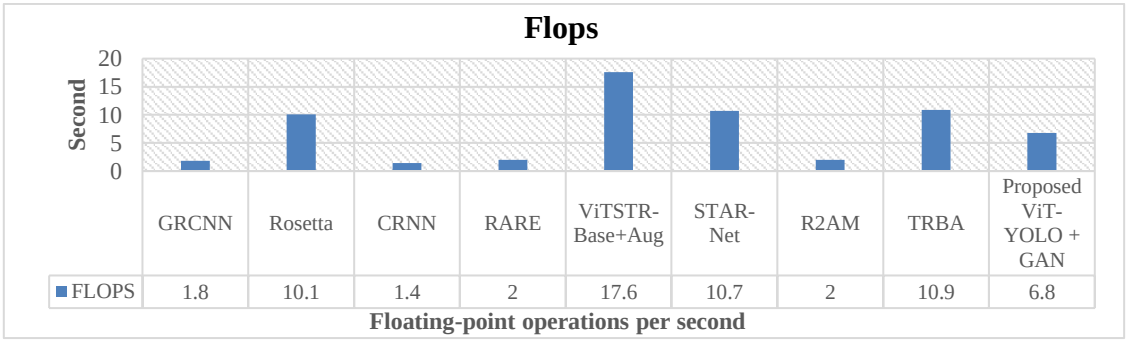


Fig. 9 Flops in various traditional models and proposed models.

Overall, it appears that the integrated proposed model has the best balance of accuracy and efficiency among the models considered in this comparison. It achieved the highest accuracy while still having a reasonable number of parameters and FLOPS, as well as a relatively fast speed. However, it's important to note that the choice of model ultimately depends on the specific requirements and constraints of the application.

The overall estimation of the proposed model (ViT-YOLO+GAN) parameters is in Table 2 with the graph shown below in Fig 10.

Table 2: Proposed model parameters

	Accuracy	Speed	Flops
Proposed Model (ViT-YOLO+GAN)	93.49%	10.37 m/s	6.8

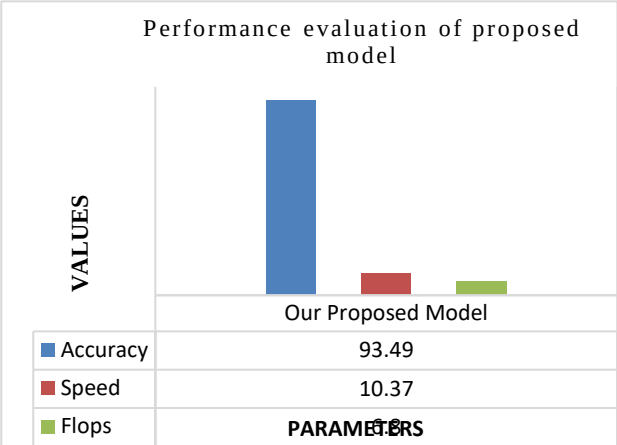


Fig. 10 Performance Evaluation Graph of Proposed Model

8 Conclusion and future scope

In conclusion, the suggested combined ViT-YOLO model has shown impressive accuracy in multilingual image-based word recognition. This model incorporates the best features of the ViT and YOLO methodologies into a single, effective tool. Using YOLO for patch extraction and ViT for text recognition, the model successfully identifies and localizes text portions inside pictures. The model's efficiency and reliability are both bolstered by the use of a GAN for augmented data. The ViT-YOLO model has a higher accuracy (93.49%) than both conventional approaches and other deep learning models, as shown by results from past searches. Thanks to ViT-YOLO's success, multilingual image-based text recognition now has some exciting new avenues to explore. Additional studies might concentrate on perfecting the integrated model's parameters for maximum efficiency and minimum computation. The model's performance in other languages might also be improved by exploring the viability of transfer learning approaches, in which the model is pre-trained on a large-scale multilingual dataset. Further research into the use of bigger and more diversified datasets tailored to multilingual text recognition might also aid in the improvement of model generalization. Additionally, the model's power to capture nuanced features and context in a multilingual text might be improved with the use of additional cutting-edge approaches such as attention mechanisms or transformer versions. In conclusion, the suggested combined ViT-YOLO model has a lot of potential and might lead to significant developments in the field of multilingual image-based text recognition.

References

[1]. Zarechensky, Mikhail, and Natalia Vassilieva. "Text detection in natural scenes with multilingual text." In CEUR Workshop Proceedings, pp. 32-35. 2014.

[2]. Liao, Wen-Hung, Yi-Hsuan Liang, and Yi-Chieh Wu. "An integrated approach for multilingual scene text detection." In 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 211-217. IEEE, 2015.

[3]. Han, Kai, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, et al. "A survey on vision transformer." IEEE Transactions on pattern analysis and machine intelligence 45, no. 1 (2022): 87-110.

[4]. Zhou, Bolei, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. "Scene parsing through ade20k dataset." In Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 633-641. 2017.

[5]. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022. 2021.

[6]. Zhang, Qiming, Yufei Xu, Jing Zhang, and Dacheng Tao. "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond." International Journal of Computer Vision (2023): 1-22.

[7]. Atienza, Rowel. "Vision transformer for fast and efficient scene text recognition." In Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland,

September 5–10, 2021, Proceedings, Part I 16, pp. 319-334. Springer International Publishing, 2021.

[8]. Yvon, François. "Transformers in natural language processing." In Human-Centered Artificial Intelligence: Advanced Lectures, pp. 81-105. Cham: Springer International Publishing, 2023.

[9]. Chen, Mark, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. "Generative pre-training from pixels." In International conference on machine learning, pp. 1691-1703. PMLR, 2020.

[10]. Zhu, Jiajun, and Guodong Wang. "TransText: Improving scene text detection via the transformer." Digital Signal Processing 130 (2022): 103698.

[11]. https://dida.do/blog/visual-transformers

[12]. Yan, X., Fang, Z., & Jin, Y. (2023). Augmented Transformers with Adaptive n-grams Embedding for Multilingual Scene Text Recognition. arXiv preprint arXiv:2302.14261.

[13]. Bautista, Darwin, and Rowel Atienza. "Scene text recognition with permuted autoregressive sequence models." In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII, pp. 178-196. Cham: Springer Nature Switzerland, 2022.

[14]. Tan, Yew Lee, Adams Wai-Kin Kong, and Jung-Jae Kim. "Pure Transformer with Integrated Experts for Scene Text Recognition." In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII, pp. 481-497. Cham: Springer Nature Switzerland, 2022.

[15]. Rithika, H., and B. Nithya Santhoshi. "Image text to speech conversion in the desired language by translating with Raspberry Pi." In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4. IEEE, 2016.

[16]. Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers." Advances in Neural Information Processing Systems 34 (2021): 23296-23308.

[17]. Cho, Suhyun, and Hayoung Oh. "Generalized Image Captioning for Multilingual Support." Applied Sciences 13, no. 4 (2023): 2446.

[18]. Wu, Weijia, Debing Zhang, Ying Fu, Chunhua Shen, Hong Zhou, Yuanqiang Cai, and Ping Luo. "End-to-end video text spotting with transformer." arXiv preprint arXiv:2203.10539 (2022).

[19]. Huang, Mingxin, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. "Swintextspotter: Scene text spotting via better synergy between text detection and text recognition." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4593-4603. 2022.

[20]. Li, Junlong, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. "Dit: Self-supervised pre-training for document image transformer." In Proceedings of the 30th ACM International Conference on Multimedia, pp. 3530-3539. 2022.

[21]. He, Yue, Chen Chen, Jing Zhang, Juhua Liu, Fengxiang He, Chaoyue Wang, and Bo Du. "Visual semantics allow for textual reasoning better in scene text recognition." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 888-896. 2022.

[22]. Li, Minghao, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. "Trocr: Transformer-based optical character recognition with pre-trained models." arXiv preprint arXiv:2109.10282 (2021).

[23]. Kim, Geewook, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. "OCR-free Document Understanding Transformer." arXiv preprint arXiv:2111.15664 (2021).

[24]. Orihashi, Shota, Yoshihiro Yamazaki, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Ryo Masumura. "Utilizing Resource-Rich Language Datasets for End-to-End Scene Text Recognition in Resource-Poor Languages." In ACM Multimedia Asia, pp. 1-5. 2021.

[25]. Yang, Lu, Peng Wang, Hui Li, Zhen Li, and Yanning Zhang. "A holistic representation guided attention network for scene text recognition." Neurocomputing 414 (2020): 67-75.

[26]. Hu, Ronghang, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9992-10002. 2020.

[27]. Atienza, Rowel. "Vision transformer for fast and efficient scene text recognition." In Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16, pp. 319-334. Springer International Publishing, 2021.

[28]. https://www.kaggle.com/datasets/garvitchaudhary/mjsynth

[29]. https://www.kaggle.com/datasets/anurags397/hindi-mnist-data

[30]. https://www.kaggle.com/datasets/rashwan/arabic-chars-mnist

[31]. Wang, Jianfeng, and Xiaolin Hu. "Gated recurrent convolution neural network for ocr." Advances in Neural Information Processing Systems 30 (2017).

[32]. Borisyuk, Fedor, Albert Gordo, and Viswanath Sivakumar. "Rosetta: Large scale system for text detection and recognition in images." In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 71-79. 2018.

[33]. Shi, Baoguang, Xiang Bai, and Cong Yao. "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." IEEE transactions on pattern analysis and machine intelligence 39, no. 11 (2016): 2298-2304.

[34]. Shi, Baoguang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. "Robust scene text recognition with automatic rectification." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4168-4176. 2016.

[35]. Liu, Wei, Chaofeng Chen, Kwan-Yee K. Wong, Zhizhong Su, and Junyu Han. "Star-net: a spatial attention residue network for scene text recognition." In BMVC, vol. 2, p. 7. 2016.

[36]. Lee, Chen-Yu, and Simon Osindero. "Recursive recurrent nets with attention modeling for ocr in the wild." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2231-2239. 2016.

[37]. Baek, Jeonghun, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. "What is wrong with scene text recognition model comparisons? dataset and model analysis." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 4715-4723. 2019.