A Refined Multilingual Scene Text Detector Based on YOLOv7

Houssem Turki[1,3][a], Mohamed Elleuch[2,3][b] and Monji Kherallah[3][c]
[1]*National Engineering School of Sfax (ENIS), University of Sfax, Tunisia*
[2]*National School of Computer Science (ENSI), University of Manouba, Tunisia*
[3]*Advanced Technologies for Environment and Smart Cities (ATES Unit), University of Sfax, Tunisia*

Keywords: Multilingual Scene Text Detection, YOLOv7, Specific Data Augmentation, Deep Learning.

Abstract: In recent years, significant advancements in deep learning and the recognition of text in natural scene images have been achieved. Despite considerable progress, the efficacy of deep learning and the detection of multilingual text in natural scene images often face limitations due to the lack of comprehensive datasets that encompass a variety of scripts. Added to this is the absence of a robust detection system capable of overcoming the majority of existing challenges in natural scenes and taking into account in parallel the characteristics of each writing of different languages. YOLO (You Only Look Once) is a highly utilized deep learning neural network that has become extremely popular for its adaptability in addressing various machine learning tasks. YOLOv7 is an enhanced iteration of the YOLO series. It has also proven to be effective in solving complex image-related problems thanks to the evolution of its 'Backbone' responsible for capturing the features of images to overcome the challenges encountered in a natural environment which leads us to adapt it to our text detection context. Our first contribution is to over-come environmental variations through the use of specific data augmentation based on improved basic techniques and a mixed transformation method applied to "RRC-MLT" and "SYPHAX" multilingual datasets which both contain Arabic scripts. The second contribution is the refinement of the 'Backbone' block of the YOLOv7 architecture to better extract the small details of the text which particularly stand out in Arabic scripts in punctuation marks. The article highlights future research directions aimed at developing a generic and efficient multilingual text detection system in the wild that also handles Arabic scripts, which is a new challenge that adds to the context, which justifies the choice of the two datasets.

1 INTRODUCTION

Text detection is vital in the field of computer vision, utilizing various Machine Learning (ML) and deep learning models to improve the effectiveness of text detection and associated tasks. Detecting text in the wild within diverse environments poses several challenges, including issues such as mixed languages, complex character designs, and various types of distortions like shape, size, multi-scale characters, orientation, gradient illumination, blur, noise, low contrast, complex background, etc. (see Figure 1). The complexity is further heightened when dealing with multilingual scripts (Bai et al., 2018). These challenges exist in datasets specific to our context, but there are few multilingual datasets (Saha et al., 2020) and particularly limited studies that contain Arabic text (Boujemaa et al., 2021) which also presents totally different challenges from Latin or other scripts; it comprises calligraphy, punctuation marks, ligatures, skewed, multi-level baselines, multi-position joining, etc. (Boukthir et al., 2022) (see Figure 2). As well only a few studies concentrate on the central issue of multilingual and varied scripts which can exist in the same image with the aim of developing a generic and resilient system capable of taking into account this varied multitude of challenges (Turki et al., 2023a). Among the many algorithms developed for object and text detection, the YOLO framework distinguishes itself as an extensively utilized algorithm (Nourali et al., 2024). During its evolution, the YOLO series of algorithms

[a] https://orcid.org/0009-0001-8472-3622
[b] https://orcid.org/0000-0003-4702-7692
[c] https://orcid.org/0000-0002-4549-1005

Turki, H., Elleuch, M. and Kherallah, M.
A Refined Multilingual Scene Text Detector Based on YOLOv7.
DOI: 10.5220/0013157100003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 512-519
ISBN: 978-989-758-737-5; ISSN: 2184-433X

has undergone iterative enhancements, its use in the context of text detection in natural scenes images has given remarkable results going from version 5 to version 7 (Turki et al., 2023b). This inspires us to employ the YOLOv7 and evaluate the detection capabilities of this neural network architecture when used for text detection in natural environments. Our selection of YOLOv7 is based on the latest advancements in the field and the promising outcomes achieved (Khan et al., 2023). We make two notable contributions:

Firstly, our primary contribution entails the novel experimental study utilizing two multilingual datasets comprising text in natural scenes images featuring the Arabic scripts. The contribution involves selecting and enhanced a specific data augmentation techniques tailored for natural scenes, incorporating a mixed transformation method designed specifically for detecting text in dark area and low-light conditions. In the second contribution, we will be improving the architecture of YOLOv7 by refining the ‘Backbone’ block through replacing in the first step two ELAN modules with SwinV2_TDD modules (Yang et al., 2023), which includes adding a convolutional layer and an upsampling layer at the beginning of each stage in the Swin Transformer v2 (Liu et al., 2022), the model is enhanced. This modification strengthens the extraction of local features and avoids over-compressing feature maps, thereby enhancing the accuracy of detecting text and small details, such as those found in Arabic writing. In the second step we introduce a BiFormer Attention Mechanism (Zhu et al., 2023) to direct the network's attention toward more important features without increasing the model complexity and to ensure the model focuses more on text information during the detection process.

The paper is organized into four sections as follows. In Section 2, an overview of various related works and their methodologies are presented, highlighting frequently used methods. Section 3 presents the methodology and provides the experimental assessment. Finally, Section 4 concludes the paper.

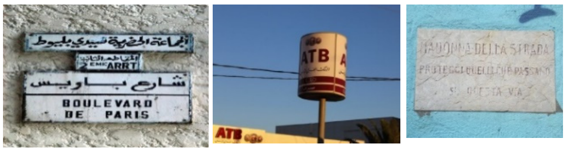


Figure 1: Samples of different challenging problems of text in the wild.

Figure 2: Samples of different challenges in Arabic scripts.

2 RELATED WORK

Many applications take advantage of the capabilities offered by deep learning and computer vision algorithms to improve text detection in learning processes (Turki et al., 2017). Object detection encompasses the recognition of specific object instances in images, videos, or live applications, including the detection of text in diverse environments of the wild (Amrouche et al., 2022). YOLO is a commonly utilized technology for various applications involving the detection of objects and text (Redmon et al., 2016). It simplifies the detection process by partitioning the input image into a grid, enabling the prediction of bounding boxes and class probabilities for each cell in the grid. YOLO has become highly popular in the field of object detection because of its simple architecture and low complexity, ease of implementation (Ravi et al., 2022). This algorithm adeptly recognizes objects and their spatial positions by leveraging the information provided by the bounding boxes (Diwan et al., 2023). Our main goal is to adapt a recent YOLO architecture for text detection, for this reason our contributions focus on the development of a new refined architecture of the YOLOv7 framework, effective for multilingual text detection capable of overcoming the different challenges of a natural environment as well as the variety of scripts including those of the Arabic language.

Implementing YOLO for text detection involves improving its initial architecture and training the algorithm to detect and locate text regions within an image. In natural scenes, such as identifying text on road signs or billboards in images related to autonomous driving applications, can be achieved using YOLO (Cui et al., 2024). This is valuable for guidance and decision-making processes in autonomous vehicles. In video surveillance systems, YOLO can be used for real-time text detection (Azevedo et al., 2024). Its application extends to recognizing text in surveillance videos, such as detecting license plates (Ramajo-Ballester et al., 2024), identification of vehicle identification numbers (VIN), traffic sign recognition (Kunekar et al., 2024). In medical Image Analysis, YOLO can assist in locating and extracting text from medical images, including radiology reports, pathology slides, and diagnostic imagery (Julia et al., 2024). In the realm of augmented reality applications, YOLO can contribute to text detection (Chu et al., 2024). It has the capability to identify text in frames captured in real-time. In Natural Language Processing (NLP) and Chatbot Applications, YOLO can be adapted to detect and extract text from images in the context of NLP applications (Karakaya et al., 2024). In chatbots or virtual assistants, YOLO-based text detection can be used to understand and respond to textual information present in images. In mobile Applications, YOLO can be integrated into mobile applications to enable on-device text detection for various purposes, such as translating text in real-time or extracting information from images (Hendrawan et al., 2024). Liu et al. (Liu et al., 2022) improved YOLOv7 through the incorporation of an attention mechanism into its main structure, which allows extracting crucial features. Chen et al. (Chen et al., 2022) improved YOLOv7 and then used it effectively in the recognition field to develop an automated selection technique. Zheng et al. (Zheng et al., 2022) improves the backbone of YOLOv7 by accelerating model convergence and making improvements including enhancing the non-maximum suppression technique through the application of grouped anchor frames during ensembles of training data.

Through the training of YOLO on adequately annotated datasets tailored for text detection, the algorithm becomes proficient in accurately identifying and localizing text regions within images. This ability opens up opportunities for the automation of tasks related to text in diverse domains.

3 METHODOLOGY

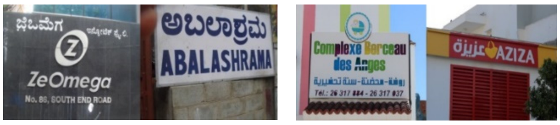
This study includes various experiments to obtain the best improved model of YOLOv7 to ensure the best detection of multilingual text in images of natural scenes containing the Arabic script. The first experiments of the work focus on the preparation of the two selected multilingual datasets not chosen randomly, through the use of targeted specific data augmentation techniques. In this context, our second refinement experiments focus on strengthening the extraction of features related to the 'Backbone' block. the details of the improvements will be described in the following sections following a balanced strategy aimed at obtaining a better generic detection system.

3.1 Datasets Employed in Experiments

For our experiments we chose the two datasets RRC-MLT (Nayef et al., 2017) and SYPHAX (Turki et al., 2023c). The first dataset RRC-MLT (Robust Reading Challenge on Multi-lingual Text) is associated with the ICDAR (International Conference on Document Analysis and Recognition) competitions, specifically the ICDAR 2019 Robust Reading Challenge on Multi-lingual Text Detection and Recognition (RRC-MLT). This dataset is designed to evaluate and benchmark text detection and recognition methods on multi-lingual and multi-script text in a variety of scenarios, including different languages, fonts, and text orientations. The challenges in the RRC-MLT competition typically include tasks related to locating and recognizing text in complex, real-world images.

The second dataset SYPHAX, was gathered in the city of "Sfax," the second-largest city in Tunisia after the capital. It encompasses a total of 2008 images, with 401 allocated for testing and 1607 for training across 7 different classes. The dataset was compiled with a focus on addressing the dynamic challenges associated with detecting text in natural scenes. The images predominantly feature text in both Arabic and Latin scripts. The SYPHAX dataset introduces a novel and significant challenge in the realm of text detection in the wild.

we chose the two datasets RRC-MLT and SYPHAX because they contain, compared to other benchmarks in the field, more scripts varied in terms of language, challenges and text exposure. RRC-MLT dataset contains texts in 9 different languages with an equal number of images with varying degrees of complexity at the script level and contain 2000 images with Arabic scripts, moreover this database is part of an ICDAR2019 competition from which we can compare our results obtained. SYPHAX dataset contains three languages but focuses more on Arabic language scripts and these related challenges; all the images contain Arabic scripts. Table 1 summarizes the comparison between the two selected datasets and Figure 3 illustrates samples of their multilingual text in the wild.



(a) RRC-MLT (b) SYPHAX dataset

Figure 3: Samples of multilingual text in the wild from RRC-MLT, and SYPHAX dataset.

Table 1: Comparison between the two datasets RRC-MLT and SYPHAX.

Dataset	RRC-MLT	SYPHAX
Size	18000	2008
Script	Arabic, French, Englis, Germa, Bangla, Chines, Italian, Japanese and Korean	Arabic, English and French
Images with Arabic scripts	2000	2800
Type of content	Text line, Words	Text line, Words
Availability	Public	Public
Source	Camera, Web Camera	Camera
Text shape	Linear, bending	Linear, bending

We carefully select each specific data augmentation technique with consideration, ensuring its alignment with the images found in natural settings and its relevance to the characteristics of the chosen dataset. The improvement of the chosen values of each transformation is based mainly on experimentation. This meticulous approach aims to achieve effective results when applying text detection methods. We choose 5 principal data augmentations techniques based on basic image manipulations (Shorten et al., 2019) (see Figure 4) and we add a mixed transformation method.

3.1.1 The Basic Data Augmentations Techniques

- Applying rigid geometric transformations (Schaefer et al., 2006) is employed to generate two additional skew angles, simulating difficult camera angles on both the right and left sides.
- Applying a horizontal directional blur to replicate the effect of capturing images in motion from the camera (Naveed et al., 2024).
- Conducting color space transformations through white balance adjustments to produce three distinct color temperatures that simulate the brightness variations during different times of the day (He et al., 2024).
- Application of median filtering, coupled with top-hat and bottom-hat transforms, serves to eliminate Salt and Pepper Noise in images. Additionally, the top-hat and bottom-hat transforms are employed to enhance image sharpness (Dong et al., 2024).
- Histogram Equalization (HE): Given that image pixels typically cluster within specific intensity ranges, histogram equalization is employed to enhance contrast (Tiendrebeogo et al., 2024).

3.1.2 The Mixed Transformation Method

Certain images depicting natural scenes present a challenge where text is situated in a dark region with varying light and shadow, rendering the text unreadable and challenging to locate. This difficulty is particularly pronounced when dealing with close color tones and grayscale distinctions between the text and the background (refer to Figure 5). To achieve these objectives, the mixed transformation method serves a dual purpose. Firstly, it aims to enhance image information by maximizing the visibility of text in relation to the background. Secondly, it artificially extends the size of the training dataset. From a technical standpoint, the primary function of the mixed transformation method relies on the fusion of three maps: the original grayscale image adding a sharpen filter, the Inversion map, and the CLAHE map (refer to Figure 6). The Inversion step facilitates the processed image to closely resemble the grayscale map of the visible image, leading to a notable improvement in recognition performance. This adaptation improves the network's capability to analyze the image, particularly boosting details in white and gray within dark areas, thereby easing the extraction of dark features (Joshi et al., 2023). On the other hand, the CLAHE operation works to create a more uniform grayscale distribution, simultaneously boosting contrast and suppressing noise. This results

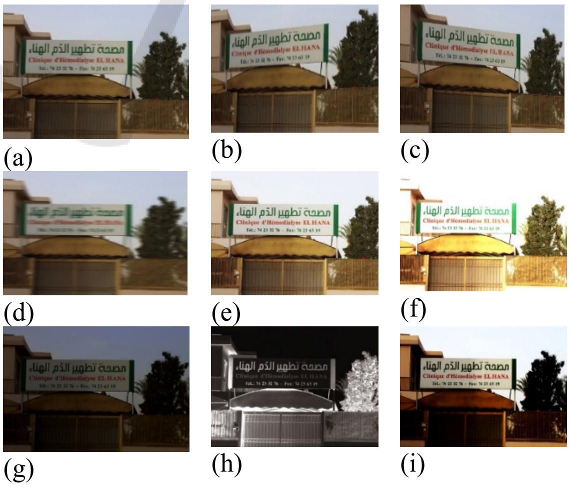


Figure 4: Samples of the basic data augmentation techniques: (a): original image (b): Right skew (c): left skew (d): horizontal directional blur (e): color temperatures 1 (f): color temperatures 2 (g): color temperatures 3 (h): median filtering & bottom-top-hat transforms (i): histogram equalization.

in an increased level of detail information in the image, with the processed pixel area becoming finer. Consequently, the approach not only suppresses noise but also enhances contrast (Mangal et al., 2024).

Figure 5: Samples of images with dark areas.

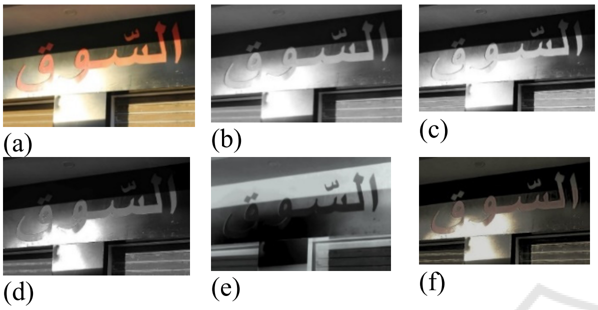


Figure 6: Sample of a mixed transformation method: (a): original image (b): grayscale image (c): Sharpen image (d): Inversion map (e): CLAHE map (f): mixed combination.

3.2 Proposed Method for Backbone Enhancement

3.2.1 The Network Architecture of YOLOv7

YOLOv7 is a real-time object detector belonging to the YOLO family, known for its single-stage architecture (Wang et al., 2023). The architecture of YOLO is based on Fully Connected Neural Networks. The YOLO framework comprises three primary components: the Backbone, the Neck, and the Head (see Figure 7**)**. The main purpose of the Backbone is to capture significant features present in an image and transmit them to the Head via the Neck. Initially, the images to be detected are processed in the 'Backbone' section. Then, they undergo three conventional convolutional operations as part of the following process: convolution, batch normalization, and SiLU activation. The MP module is a dual-channel component responsible for performing maximum pooling and convolution on the input feature images. The SPPCSPC module represents a spatial pyramid pooling structure designed to overcome image distortion issues that may arise during the scaling of input images. Following the features extraction process within the Backbone section, the image will advance to the Neck section for the purpose of features fusion. The architecture design of YOLOv7 is based on ELAN (Efficient Layer Aggregation Network). ELAN strives to establish a streamlined network by effectively handling both the shortest and longest gradient paths. The Neck section utilizes an FPN (Feature Pyramid Network) network structure for this purpose. The bottom-up features fusion process combines feature images of various scales, resulting in a multi-scale target detection effect (Patel et al., 2022). Ultimately, the Head comprises output layers responsible for producing the final detections. In YOLOv7, the head that generates the final output is referred to as the Lead Head. Figure 7 illustrates the complete architecture of the YOLOv7 network structure. In contrast to earlier versions, YOLOv7 was trained solely on the MS COCO dataset (Redmon et al., 2018), without the use of pre-trained backbones.

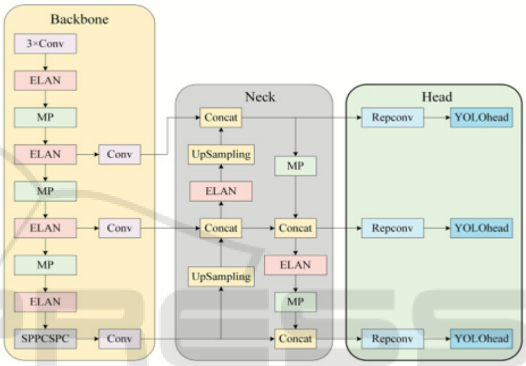


Figure 7: YOLOv7 architecture (Wang et al., 2023).

3.2.2 Replace ELAN with SwinV2_TDD Modules

Our initial enhancement involves substituting the first and last ELAN module within the Backbone block with SwinV2_TDD (Yang et al., 2023). To enhance the capture of local information and precisely detect small areas, Swin Transformer v2 (Liu et al., 2022) has been upgraded to SwinV2_TDD (Yang et al., 2023). Since punctuation marks, ligatures and small characters at multiple levels are often focused in particular regions, it is essential to pay more attention to local information. Among the advantages of integrating SwinV2 TDD are the reduction in size of feature maps and improving the efficiency of calculations. This results in quicker feature extraction while preserving high-quality outcomes and decreasing memory usage. Secondly, it prevents the excessive compression of feature maps by restoring them to their original sizes and enhances their expressive capability.

3.2.3 Integration of BiFormer Attention Mechanism

The BiFormer attention mechanism (Zhu et al., 2023) is a novel visual Transformer model that combines a two-layer routing attention approach with Bidirectional Routing Attention (BRA). It operates by iteratively transferring information to compute attention weights, thereby achieving dynamic and query-adaptive sparsity. This capability allows it to handle intricate feature relationships while minimizing computational costs. Through multiple iterations, the two-layer routing attention can dynamically adjust the attention weights. The advantages of BiFormer attention mechanism are multiple. First, the model's capacity to detect essential features and contextual information within the image can be enhanced, enabling it to concentrate more effectively on small scripts. Second, The BiFormer attention mechanism can capture the relationship between the target and the background, helping to produce more distinct feature representations. Figure 8 illustrates the final improvement of the Backbone block of YOLOv7.

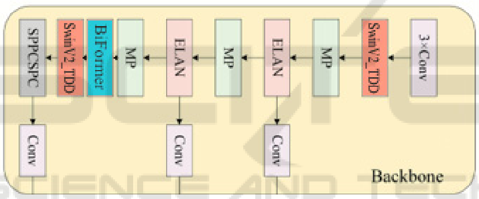


Figure 8: Final improved Backbone of YOLOv7.

3.3 Experiments and Results

It is crucial to note that the effectiveness of the text detection algorithm using YOLOv7 may vary depending on several factors involved; features of the dataset utilized and the training methods employed. Thus, in our study, the reported precision, recall, and F-scores differ depending on the improved architecture of YOLOv7, the inclusion of multilingual text with Arabic script and the data augmentation methods applied. The evaluation metrics are the precision, recall and F-score defined as:

$$Precision = \frac{\sum_i^N \sum_j^{|D^i|} M_D(D_j^i, G^i)}{\sum_i^N |D^i|} \quad (1)$$

$$Recall = \frac{\sum_i^N \sum_j^{|G^i|} M_G(G_j^i, D^i)}{\sum_i^N |G^i|} \quad (2)$$

$$F\text{-}score = 2\ x \frac{Precision\ X\ Recall}{Precision + Recall} \quad (3)$$

N represents the complete count of images within a given dataset. $|D^i|$ and $|G^i|$ are the number of detection and ground true rectangles in i^{th} image. $M_D(D_j^i, G^i)$ and $M_G(\ G_j^i, D^i)$ are the matching scores for detection rectangles D^i and ground true rectangle G^i.

The experimental outcomes of the proposed method are promising as evidenced by the observed values in Table 2. Moreover, the F-score rankings remain consistent across the two multilingual datasets, which confirms the efficacy of the method, providing additional validation. Table 2 displays the experimental results tested on SYPHAX dataset, we can note that the enhanced YOLOv7 obtains the best result in precision (87.3%) and the F-score (82.41%); the language of the scripts is distributed equally between all the images in the database and the improvement of the architecture of YOLOv7 is well suited especially to the Arabic language. On the other hand, the tested on RRC-MLT dataset, note that we obtain the best result in the recall (79.12%); which also proves the impact of strengthening feature extraction even at the script diversity level. From the point of view of result values, this enhancement is especially remarkable in recognizing texts of diverse sizes and various scripts. These findings validate the credibility of the method introduced in this article. From the point of view of classification of results, this study is encouraging and warrants further exploration and investigation especially in this context of multiple challenges, which is shown in Table 3 to compare the results obtained with those of the ICDAR2019 competition (Nayef et al., 2019).

Table 2: Experimental result of the proposed method.

Dataset	Precision (%)	Recall (%)	F score (%)
SYPHAX	**87.3**	78.05	**82.41**
RRC-MLT	83.75	**79.12**	81.36

Table 3: Results compared to the ICDAR 2019 competition (challenge for task-1: multi-lingual text detection) (Nayef et al., 2019).

RRC-MT Dataset	Precision (%)	Recall (%)	F score (%)
Tencent-DPPR Team	87.52	80.05	83.61
Proposed method	83.75	79.12	81.36
MM-MaskRCNN	84.73	70.21	76.79
MLT2019 ETD	78.71	54.44	64.36
Cyberspace	69.48	35.61	47.09

3.4 Discussion

The ‘Backbone’ is one of the most crucial components in YOLOv7, significantly influencing the performance of the entire detection model. By improving this block by replacing ELAN modules and incorporating the BiFormer attention mechanism, the model can more effectively utilize spatial and semantic information in the image, thereby enhancing the representation of text during feature extraction. This enhances the model's generalization ability and robustness while decreasing its reliance on other stages. Therefore, these modifications to the YOLOv7 backbone can enhance the model’s detection accuracy for small targets and improve the feature extraction stage, effectively leveraging key features and contextual information in the image. This design choice can enhance text detection performance and offer more precise feature representation for subsequent processing stages, resulting in improved detection outcomes (see Figure 9).

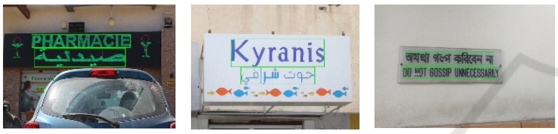


Figure 9: Text detection samples based on the proposed method.

4 CONCLUSION

This research paper introduces a novel method for multilingual text detection in the wild, utilizing the YOLOv7 model optimized through specific data augmentation techniques. The proposed method improves the YOLOv7 model by enhancing the features extraction capabilities of the backbone network and takes into account the characteristics of Arabic language scripts rarely treated in the context of multilingual detection. The effectiveness of this method is showcased through experiments performed on two multilingual datasets. The obtained results are promising, indicating that the enhanced YOLOv7 model has the potential to further advance by incorporating targeted enhancements to improve text detection performance across various multilingual scripts. This may involve modifications to the network architecture and refining training strategies. As YOLOv7 is a relatively recent development, it holds significant promise for future research and exploration in the field of text detection.

REFERENCES

Amrouche, A., Bentrcia, Y., Hezil, N., Abed, A., Boubakeur, K. N., & Ghribi, K. (2022, November). Detection and localization of arabic text in natural scene images. In 2022 First International Conference on Computer Communications and Intelligent Systems (I3CIS) (pp. 72-76). IEEE.

Azevedo, P., & Santos, V. (2024). Comparative analysis of multiple YOLO-based target detectors and trackers for ADAS in edge devices. Robotics and Autonomous Systems, 171, 104558.

Bai, X., Yang, M., Lyu, P., Xu, Y., & Luo, J. (2018). Integrating scene text and visual appearance for fine-grained image classification. IEEE Access, 6, 66322-66335.

Boujemaa, K. S., Akallouch, M., Berrada, I., Fardousse, K., & Bouhoute, A. (2021). ATTICA: a dataset for arabic text-based traffic panels detection. *IEEE Access*, *9*, 93937-93947.

Boukthir, K., Qahtani, A. M., Almutiry, O., Dhahri, H., & Alimi, A. M. (2022). Reduced annotation based on deep active learning for arabic text detection in natural scene images. *Pattern Recognition Letters*, *157*, 42-48.

Cui, Y., Guo, D., Yuan, H., Gu, H., & Tang, H. (2024). Enhanced YOLO Network for Improving the Efficiency of Traffic Sign Detection. Applied Sciences, 14(2), 555.

Chu, C. H., & Liu, S. (2024). Virtual Footwear Try-On in Augmented Reality Using Deep Learning Models. Journal of Computing and Information Science in Engineering, 24(3), 031002.

Chen, J., Liu, H., Zhang, Y., Zhang, D., Ouyang, H., & Chen, X. (2022). A Multiscale Lightweight and Efficient Model Based on YOLOv7: Applied to Citrus Orchard. Plants, 11(23), 3260.

Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. Multimedia Tools and Applications, 82(6), 9243-9275.

Dong, J., Wang, N., Fang, H., Lu, H., Ma, D., & Hu, H. (2024). Automatic augmentation and segmentation system for three-dimensional point cloud of pavement potholes by fusion convolution and transformer. Advanced Engineering Informatics, 60, 102378.

Hendrawan, A., Gernowo, R., Nurhayati, O. D., & Dewi, C. (2024). A Novel YOLO-ARIA Approach for Real-Time Vehicle Detection and Classification in Urban Traffic. International Journal of Intelligent Engineering & Systems, 17(1).

He, W., Zhang, C., Dai, J., Liu, L., Wang, T., Liu, X., ... & Liang, X. (2024). A statistical deformation model-based data augmentation method for volumetric medical image segmentation. Medical Image Analysis, 91, 102984.

Julia, R., Prince, S., & Bini, D. (2024). Medical image analysis of masses in mammography using deep learning model for early diagnosis of cancer tissues. In Computational Intelligence and Modelling Techniques for Disease Detection in Mammogram Images (pp. 75-89). Academic Press.

Joshi, G., Natsuaki, R., & Hirose, A. (2023). Neural Network Fusion Processing and Inverse Mapping to Combine Multisensor Satellite Data and Analyze the Prominent Features. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 2819-2840.

Khan, N., Ahmad, R., Ullah, K., Muhammad, S., Hussain, I., Khan, A., ... & Mohamed, H. G. (2023). Robust Arabic and Pashto Text Detection in Camera-Captured Documents Using Deep Learning Techniques. *IEEE Access*, *11*, 135788-135796.

Kunekar, P., Narule, Y., Mahajan, R., Mandlapure, S., Mehendale, E., & Meshram, Y. (2024). Traffic Management System Using YOLO Algorithm. Engineering Proceedings, 59(1), 210.

Karakaya, M., Ersoy, S., Feyzioğlu, A., & Ersoy, S. (2024). Reading Gokturkish text with the Yolo object detection algorithm. Journal of Mechatronics and Artificial Intelligence in Engineering.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... & Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12009-12019).

Liu, S., Wang, Y., Yu, Q., Liu, H., & Peng, Z. (2022). CEAM-YOLOv7: Improved YOLOv7 Based on Channel Expansion and Attention Mechanism for Driver Distraction Behavior Detection. IEEE Access, 10, 129116-129124.

Mangal, A., Garg, H., & Bhatnagar, C. (2024). A Robust Co-saliency Object Detection Model by Applying CLAHE and Otsu Segmentation Method. International Journal of Intelligent Systems and Applications in Engineering, 12(1s), 481-490.

Nourali, K., & Dolkhani, E. (2024). Scene text visual question answering by using YOLO and STN. International Journal of Speech Technology, 1-8.

Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., ... & Ogier, J. M. (2017, November). Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR) (Vol. 1, pp. 1454-1459). IEEE.

Naveed, H., Anwar, S., Hayat, M., Javed, K., & Mian, A. (2024). Survey: Image mixing and deleting for data augmentation. Engineering Applications of Artificial Intelligence, 131, 107791.

Nayef, N., Patel, Y., Busta, M., Chowdhury, P. N., Karatzas, D., Khlif, W., ... & Ogier, J. M. (2019, September). Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In 2019 International conference on document analysis and recognition (ICDAR) (pp. 1582-1587). IEEE.

Patel, K., Bhatt, C., & Mazzeo, P. L. (2022). Improved Ship Detection Algorithm from Satellite Images Using YOLOv7 and Graph Neural Network. Algorithms, 15(12), 473.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Ravi, N., & El-Sharkawy, M. (2022). Real-Time Embedded Implementation of Improved Object Detector for Resource-Constrained Devices. Journal of Low Power Electronics and Applications, 12(2), 21.

Ramajo-Ballester, Á., Moreno, J. M. A., & de la Escalera Hueso, A. (2024). Dual license plate recognition and visual features encoding for vehicle identification. Robotics and Autonomous Systems, 172, 104608.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Saha, S., Chakraborty, N., Kundu, S., Paul, S., Mollah, A. F., Basu, S., & Sarkar, R. (2020). Multi-lingual scene text detection and language identification. *Pattern Recognition Letters*, *138*, 16-22.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

Schaefer, S., McPhail, T., & Warren, J. (2006). Image deformation using moving least squares. In ACM SIGGRAPH 2006 Papers (pp. 533-540).

Turki, H., Elleuch, M., Kherallah, M., & Damak, A. (2023a, September). Arabic-Latin Scene Text Detection based on YOLO Models. In *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)* (pp. 1-6). IEEE.

Turki, H., Elleuch, M., & Kherallah, M. (2023b, November). Multi-lingual Scene Text Detection Containing the Arabic Scripts Using an Optimal then Enhanced YOLO Model. In *International Conference on Model and Data Engineering* (pp. 47-61). Cham: Springer Nature Switzerland.

Turki, H., Elleuch, M., Kherallah, M. (2023c). SYPHAX Dataset. IEEE ataport. https://dx.doi.org/10.21227/ydqd-2443.

Turki, H., Halima, M. B., & Alimi, A. M. (2017, November). Text detection based on MSER and CNN features. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR) (Vol. 1, pp. 949-954). IEEE.

Tiendrebeogo, A. (2024). Identification of plants from the convolutional neural network. Multimedia Tools and Applications, 1-11.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7464-7475).

Yang, Y., & Kang, H. (2023). An enhanced detection method of PCB defect based on improved YOLOv7. *Electronics*, *12*(9), 2120.

Zhu, L., Wang,X., Ke, Z., Zhang, W., & Lau, R. W. (2023). Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10323-10333).

Zheng, J., Wu, H., Zhang, H., Wang, Z., & Xu, W. (2022). Insulator-Defect Detection Algorithm Based on Improved YOLOv7. Sensors, 22(22), 8801.