A Comparative Study of Deep Learning Algorithms for Multilingual Scene Text Detection and Recognition

Taher Ali Mahmood [(1)]
Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq .
taher.24csp35@student.uomosul.edu.iq
0009-0003-9454-4630

Yusra Faisal Mohammad [(2)]
Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq .
yusrafaisalcs@uomosul.edu.iq
0000-0002-8504-8243

Abstract

This survey presents a systematic and analytical review of deep learning algorithms for multilingual scene text detection and recognition. We critically analyze over 50 publications from 2013 to 2025, categorizing methods into regression-based, segmentation-based, and hybrid detection approaches, alongside CTC-based, attention-based, Transformer-based, and vision-language recognition models. Our analysis reveals that segmentation-based detectors such as DB++ achieve 87.9% F-measure on ICDAR2015, while vision-language models like CLIP4STR reach 94.1% average recognition accuracy. End-to-end methods with detection-recognition synergy, particularly DNTextSpotter, achieve 85.3% on Total-Text. We systematically evaluate 14 benchmark datasets across 10+ scripts, quantify annotation cost trade-offs, and identify critical gaps in low-resource language support. This survey provides comparative analysis through quantitative benchmarking, identifies key architectural trends, and outlines future research directions including unified multi-task frameworks, LLM integration, and efficient edge deployment.

Keywords: Deep Learning, Scene Text Detection, Text Recognition, Multilingual OCR, Transformer, CLIP, End-to-End Text Spotting .

1.Introduction

Detecting and recognizing scene text in natural images is a major challenge in the field of computer vision, given its direct uses in applications for converting document text to digital text, self-driving vehicles. [1]. and virtual reality, where scene text shows a significant difference in font type, size, shape, and color, in addition to complex text backgrounds, compared to text in clear digital documents.[2]. Fig . 1 shows the general structure of the text discovery and recognition process, although modern systems and methods combine the two stages into one, which is called end-to-end.[3].

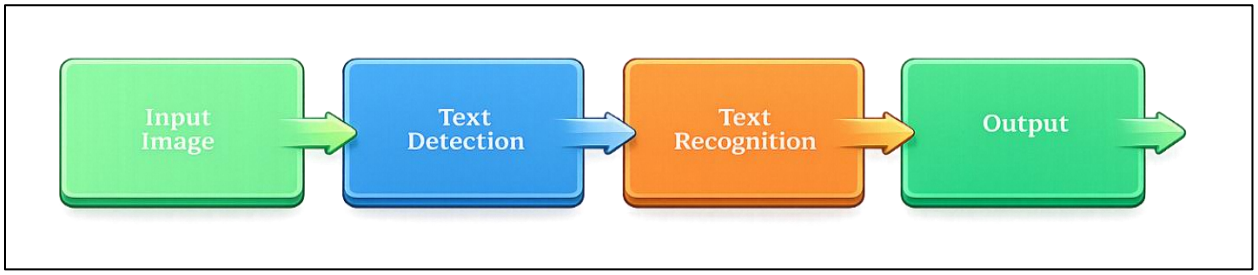

Fig. 1. General pipeline of scene text detection and recognition.

Deep Learning (DL) created a qualitative leap in this area. Convolutional Neural Networks (CNNs) allowed automatic inference of features using DL techniques, which was a superior technique to the traditional methods that required manual feature design [4]. Transformer models are also being developed to create new ways of modelling long-range

correlations, and hybrid models combining vision and language have been created to effectively combine and align multiple forms of media, such as CLIP [5]. Additionally, differences in visual characteristics of writing systems lead to challenges with multilingual environments; for example, Arabic script is connected, Chinese characters are graphic-dense, and Devanagari script has an inherent structural nature and thus require flexible computer architectures to adapt to the fundamental differences in these scripts [6].

Although much has been accomplished, most research to date either predates the existence of Transformer based methods [2], is focused on single language cases [7], or does not include any quantitative comparisons between different methodologies [8]. This research contains new information that fills these gaps: (1) A complete taxonomy (shown in Fig. 2) which includes Detection, Analysis and End-To-End Spotting Methodologies; (2) Quantitative Analysis of Performance across Benchmark Data sets; (3) Critical Evaluation of Multilingual Capability; and (4) Open Challenge Identification. The complete taxonomy is illustrated in Fig . 2.

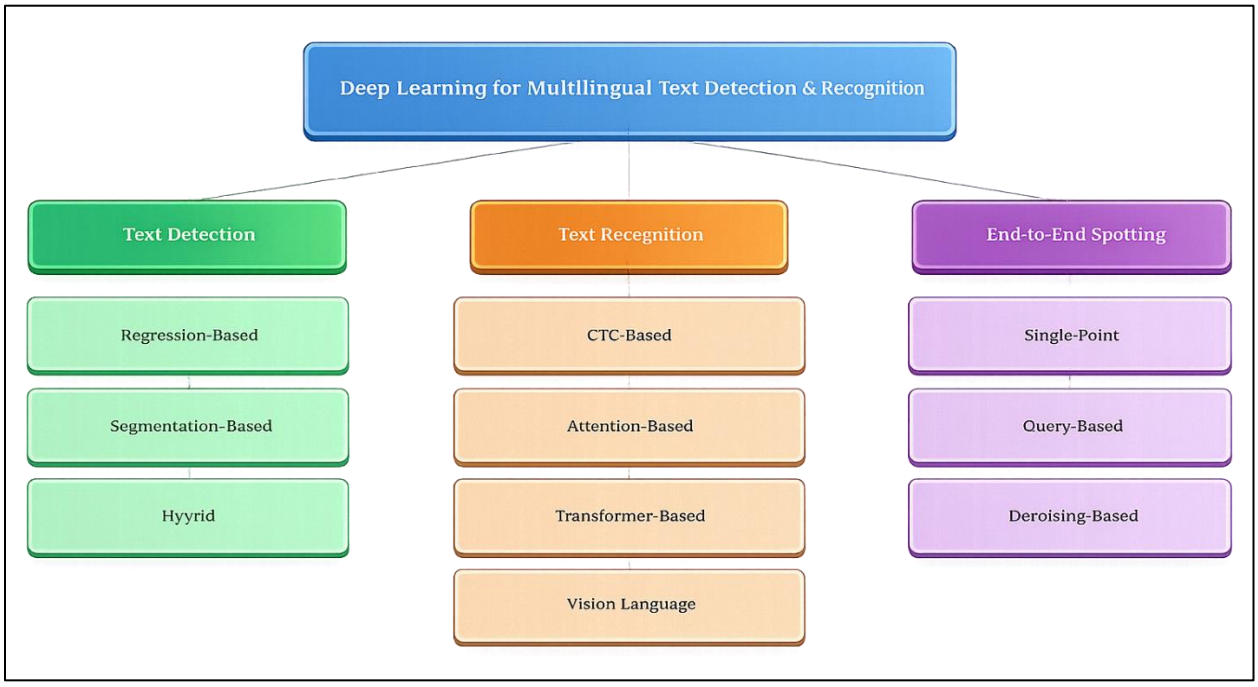

Fig. 2. Taxonomy of deep learning methods for multilingual text detection and recognition.

2. Deep Learning Methods for Text Detection

The process of identifying text in an image is known as text detection. Text detection methods can be divided into three main groups: regression-based, segmentation-based, and hybrid. The next step will be to systematically evaluate all the available architectural designs for each of the three types of text detection methods, then compare their corresponding performance trade-offs for each of the different design approaches.

2.1 Regression-Based Methods

eure maps are used by regression-based detectors to directly predict bounding box coordinates. The EAST detector was the first to provide an efficient single-network detection method without requiring complex post-processing [9]. In the paper by He et al., they presented a method for applying regression to both multi-oriented and multi-lingual detection using fully convolutional networks via a multi-task learning strategy. Their work resulted in an 82.0% F-measure on ICDAR2015 [10]. The authors problem from the work of Guan et al. proposed RFAN using multi-scale attention for detecting industrial text, provided the MPSC dataset, and achieved an F-measure of 85.2% on ICDAR2015 [11]. Therefore, regression methods may have difficulty identifying arbitrary-shaped text because they rely on fixed bounding boxes.

2.2 Segmentation-Based Methods

Segmentation-based techniques define our detection problem as a pixel-wise classification issue, allowing for arbitrary shapes to be handled. Liao et al.'s DB/DB++ architecture introduced a differentiable binarization method where two probabilities (and a thresholding map) are combined together in order to perform pixel-wise classification. At the time of publication, an F-measure of 87.9% on the ICDAR2015 dataset, and 88% F-measure on the Total-Text dataset was achieved at nearly real-time speed [12]. Wang et al. proposed the PAN++ architecture, which utilizes a kernel-based representation along with progressive expansion of kernels, achieving an F-measure of 82.9% on the ICDAR2015 dataset [13]. Zheng et al. introduced KACNet which utilizes kernels that are adjusted based on the distance map prediction, achieving F-measures on ICDAR2015 and Total-Text datasets of 87.1% and 88.5% respectively.[14]

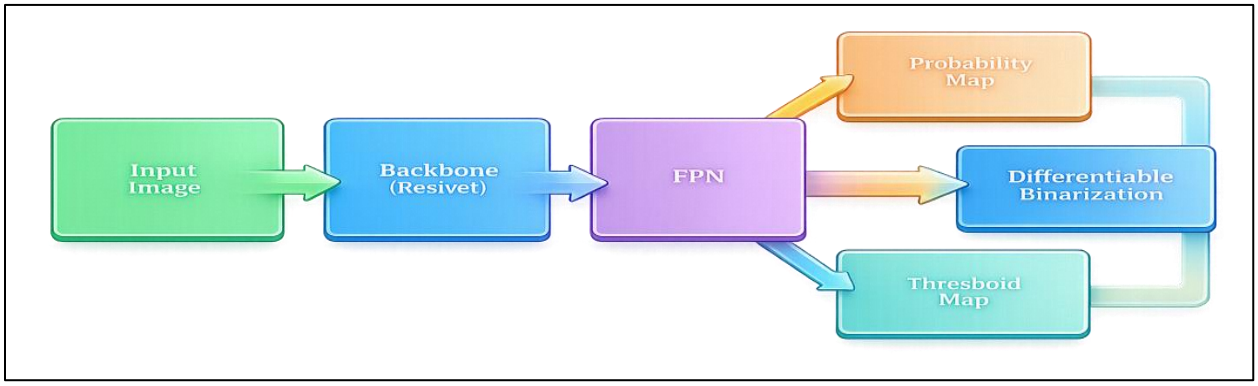

Fig. 3. DB/DB++ segmentation-based architecture with differentiable binarization.

2.3 Hybrid and Advanced Methods

According to Tang et al., using only a small number of "representative" features to represent an entire sample was more efficient than using dense representations, which resulted in an 86.7 percent success rate on the ICDAR2015 dataset [15]. Xie et al. utilized an approach called MTP that involves CLIP distillation and masked image modeling to achieve an overall accuracy of 89.3 percent based on self-supervised pre-training methods [16]. Wan et al. created DAT to combine word/line/paragraph/page level detection into one model using interactive attention, which produced the highest level of accuracy from the multiple levels of granularity [17]. A full comparison of all types of detection methods can be found in Table 1 as well as in Fig . 4.

TABLE 1: Text Detection Methods Performance Comparison

Method	Year	Approach	Backbone	IC15 F%	TT F%	Speed
EAST[9]	2017	Regression	PVANet	78.3	-	Fast
He et al. [10]	2018	Regression	ResNet-50	82.0	-	Med
PAN++ [13]	2022	Kernel	ResNet-18	82.9	78.5	Fast
RFAN [11]	2022	Attention	ResNet-50	85.2	-	Med
FSG [15]	2022	Sampling	ResNet-50	86.7	87.3	Med
DB++ [12]	2023	Segment.	ResNet-50	87.9	88.0	Fast
KACNet [14]	2024	Adaptive	ResNet-50	87.1	88.5	Med
DAT [17]	2024	Multi-Gran.	Swin-T	88.2	89.1	Slow
MTP [16]	2025	Pre-train	ResNet-50	89.3	-	Med

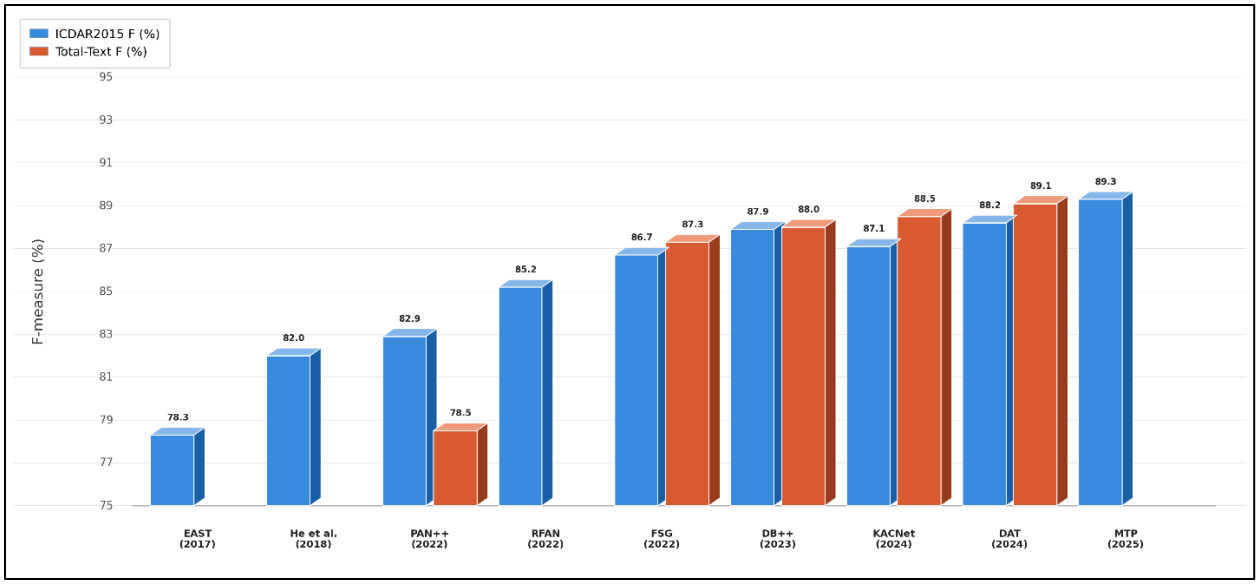

Fig. 4. Detection performance comparison on ICDAR2015 and Total-Text.

3. Deep Learning Methods for Text Recognition

Text recognition turns identified areas into strings of characters. We examine four paradigms: CTC-based, attention-based, Transformer-based, and vision-language methodologies.

3.1CTC and Attention-Based Methods

The alignment-free training of CTC based features and character sequences enables CTC-based methods. Using CTC on Arabic handwriting text and Transformers to deal with cursive text dependencies, exploiting self-attention, is what Mostafa et al. did when they introduced OCFormer [18]. The most popular and useful OCR, Tesseract, does not currently work on difficult scene text [19]. In addition to focusing attention on the features, attention based methods are useful; the Canonical glyph masks used in CAM by Yang et al., create an environment that will reduce background noise, thus improving performance by 4.1% on both of the datasets [20].

3.2Transformer and Vision-Language Models

The development that has been achieved so far on the area of vision-language integration is the most significant. The paper by Wang et al., CLIP-OCR; presented a symmetric version of distillation on the CLIP encoders to show 93.8% average accuracy [21]. In their article entitled CLIP4STR, Zhao et al. used CLIP and tuned it in order to select specifically STR; that obtained an average accuracy of 94.1% [22]. Xu et al. developed a method of implementing compaction of the text into single token representations to facilitate easy recognition and retrieval and gave them the name OTE [23]. Lastly, IGTR made a contribution through the contribution of Du et al. who used linguistic instructions to define their recognition based on attribute [24]. All these studies may be compared and followed to demonstrate the uniformity of increased performance as a result of the vision-language integration as illustrated in Table 2 and Fig . 5.

TABLE 2: Text Recognition Methods Accuracy Comparison

Method	Year	IIIT5K	SVT	IC13	IC15	Avg
CLIP-OCR [21]	2023	97.5	95.2	97.8	87.2	93.8
CAM [20]	2024	97.2	95.0	97.5	86.8	93.2
CLIP4STR [22]	2024	97.7	95.8	97.9	88.1	94.1
OTE [23]	2024	97.8	96.1	98.2	88.5	94.5
IGTR [24]	2025	97.6	96.4	98.0	89.2	94.8

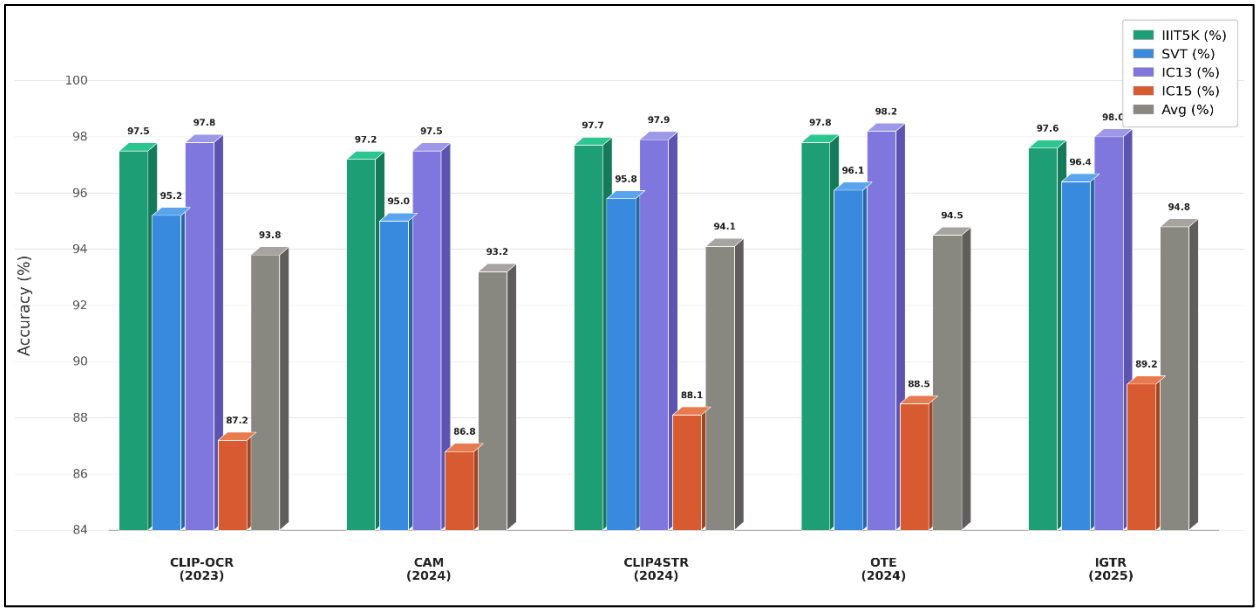

Fig. 5. Recognition accuracy comparison across IIIT5K, SVT, and IC13 benchmarks.

4. End-to-End Text Spotting

End-to-end spot origins combine both detection and recognition; thus, there is no longer an accumulation of mistakes from each stage in the two-stage process. This relationship is demonstrated in the overall framework depicted in Fig. 6.

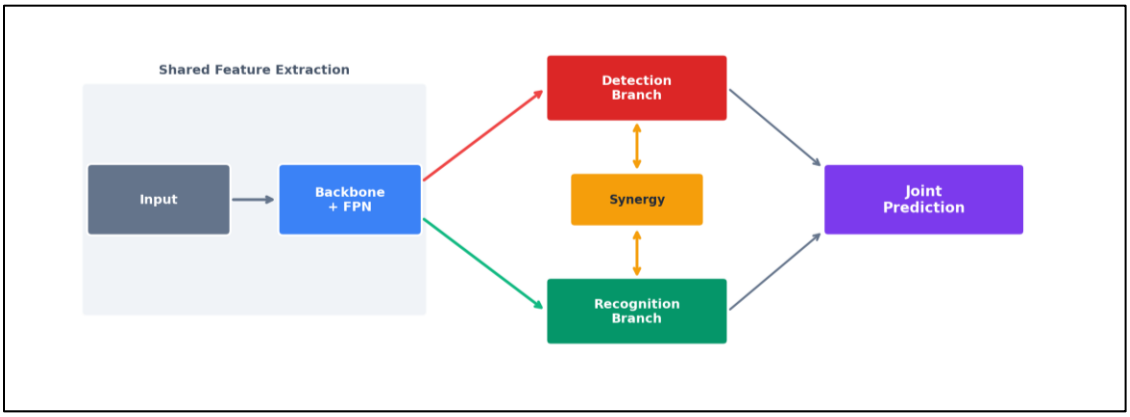

Fig. 6. End-to-end text spotting framework with detection-recognition synergy.

4.1 Transformer-Based Spotting

Through the utilisation of the Swin Transformer backbone and the recognition conversion function, SwinTextSpotter by Huang et al. achieved synergy of 83.9% on Total-Text [5]. ESTextSpotter also introduced explicit synergy in this regard using the concept of joint query formulations [25]. Lastly, Huang et al. connected end-to-end and two-stage paradigms by unifying modular strength with shared representations [26].

4.2 Novel Supervision and Instruction-Guided Approaches

SPTS by Peng et al. demonstrated single-point annotation suffices for text spotting via autoregressive Transformer prediction [27]. Tang et al. eliminated location annotations entirely using transcription-only supervision with cross-attention localization [28]. DNTextSpotter by Qiao et al. addressed bipartite matching instability through denoising training with masked character sliding, achieving 85.3% on Total-Text and 11.3% improvement on Inverse-Text [29].

InstructOCR by Duan et al. leveraged human language instructions, achieving state-of-the-art results and transferring to VQA tasks with 2.6% gain on TextVQA [30]. The comprehensive comparison in Table 3 and Fig. 7 quantifies these advances.

TABLE 3: End-to-End Text Spotting Performance

Method	Year	TT-None	TT-Full	IC15-S	Ann.	Innovation
Tang et al. [28]	2022	65.8	76.3	-	Trans.	Voice
PAN++ [13]	2022	69.2	78.5	68.6	BBox	Kernel
SwinTextSpotter [5]	2022	74.3	83.9	77.0	BBox	Swin
SPTS [27]	2022	74.1	82.4	73.5	Point	Auto-reg.
ESTextSpotter [25]	2023	76.2	83.5	77.3	BBox	Synergy
DNTextSpotter [29]	2024	79.8	85.3	80.2	BBox	Denoise
InstructOCR [30]	2025	78.5	85.1	79.8	BBox	Instruct.

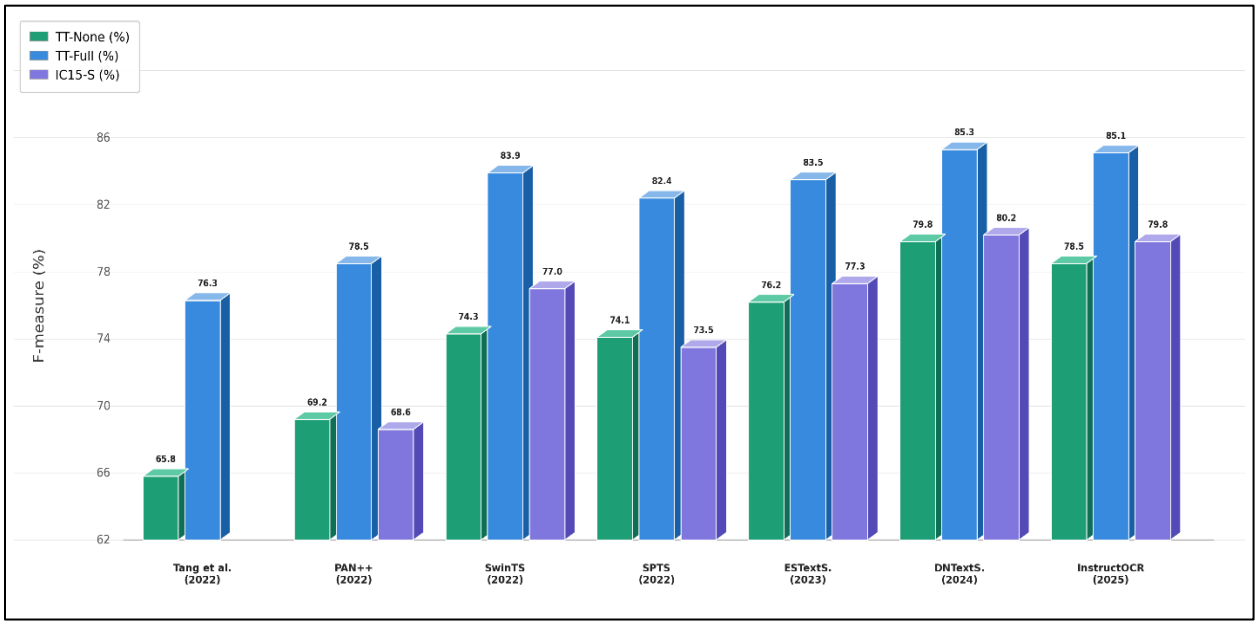

Fig. 7. End-to-end spotting performance on Total-Text (None and Full lexicon).

5. Benchmark Datasets for Multilingual Text

The advancement of both text detection and recognition lessons is derived from benchmark datasets, therefore we have performed a systematic study of fourteen large benchmark multi- language datasets to evaluate their script coverage, the types of annotations they provide, and their relative sizes. Fig. 8 presents an illustration of how scripts are distributed across different multilingual benchmark datasets.

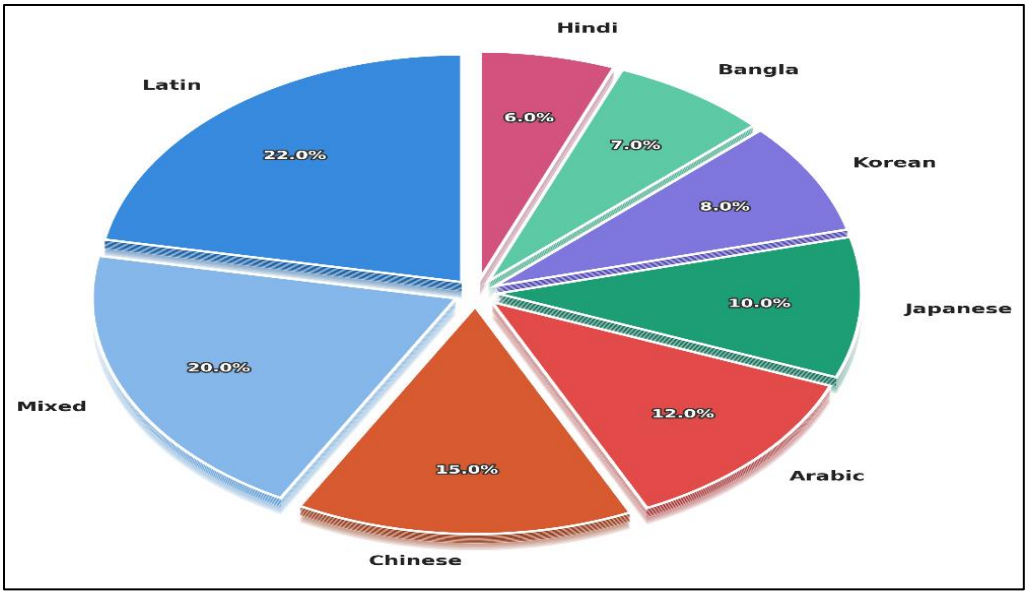

Fig. 8. Script distribution in multilingual text detection datasets.

ICDAR 2015 delivered 1500 incidental scene pictures taken through Google Glass [31]. ICDAR 2017 MLT added 9 distinct scripts including Arabic, Chinese and Korean [6].The datasets released during ICDAR 2019 include three important datasets: RRC MLT 2019, which includes 10 scripts [32], RRC LSVT, which contains 450,000 instances of Chinese text that have only been partially labeled [33], as well as RRC ArT, which includes arbitrary shaped text [34].COCO-Text included 63,686 images that each have 173,000 annotations for incidental text [35].The TextOCR dataset included 900,000 instances of polygons and annotation [36].SynthText allowed researchers to pretrain on 800,000 images, which were all artificially generated [37].

RCTW-17 provided a Chinese text reading evaluation for multilingual/specific script evaluation [38]. Hassan et al. created benchmarks for Arabic for deep learning applications using cursive writing [39]. SignboardText research evaluated Vietnamese signboard text [40]. Nath et al.'s IndicDLP provides an evidence base for crossscript generalisation with 11 Indic languages, and 12 document types [41]. DOST collected 32,147 Japanese scene image and 935k written text regions for evaluation [42]. Summary of all datasets presented in Table 4.

TABLE 4: Major Benchmark Datasets for Scene Text

Dataset	Year	Images	Scripts	Task	Annotation	Instances
ICDAR2015 [31]	2015	1,500	Latin	Det+Rec	Word BBox	6.5K
SynthText [37]	2016	800K	Latin	Det	Char BBox	8M+
DOST [42]	2016	32,147	Japanese	Det+Rec	Region	935K
COCO-Text [35]	2016	63,686	Multi	Det+Rec	BBox	173K
MLT 2017 [6]	2017	18,000	9 Scripts	Det+Script	Word BBox	108K
RCTW-17 [38]	2017	12,263	CN+EN	Det+Rec	Line BBox	44K
MLT 2019 [32]	2019	20,000	10 Scripts	Det+Rec+E2E	Word BBox	133K
RRC-LSVT [33]	2019	450K	Chinese	Det+E2E	Partial	450K
RRC-ArT [34]	2019	10,166	CN+EN	Det+Rec+E2E	Polygon	69K
TextOCR [36]	2021	28,134	Multi	Det+Rec	Polygon	900K
SignboardText [40]	2024	6,890	Vietnamese	Det+Rec	Polygon	35K
IndicDLP [41]	2025	18K+	12 Scripts	Layout	Region	90K+

6. Multilingual and Script-Specific Analysis

Text processing in multiple languages requires working with a wide range of different types of scripts, such as Arabic, which is a right-to-left, flowing, cursive script; Chinese, which has many tightly packed logographic characters; and various Indian scripts, which have character pairs linked together by visible horizontal lines [6].Abdessamad et al. used Deformable DETR to find Arabic text in an image [43]. OCFormer used Transformer-CTC for recognising handwritten Arabic text [18]. He et al. created a method that was able to simultaneously perform multi-angled predictions across several languages, and were able to achieve very good cross-script results [10]. As illustrated on the heatmap in Fig. 9, there is no method that dominates across all benchmarks. This means that it is necessary to utilize a specific architecture designed to accommodate differences in the fundamental structures of the various script systems. Table 5 provides an overview of the multi-lingual methods discussed.

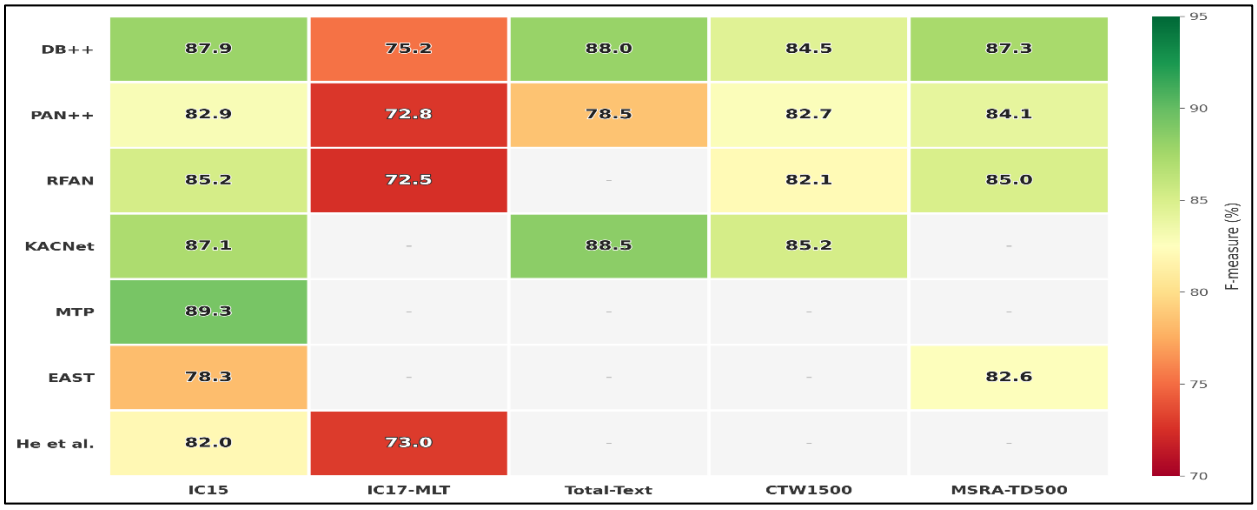

Fig. 9. Detection performance heatmap across major benchmarks.

TABLE 5: Multilingual and Script-Specific Methods

Method	Year	Target Script(s)	Architecture	Task
RRC-MLT [6]	2017	9 Scripts	Various	Detection
He et al. [10]	2018	Multi-oriented	FCN Regression	Detection
OCFormer [18]	2021	Arabic HTR	Transformer+CTC	Recognition
Hassan et al. [39]	2021	Arabic STR	CNN+Attention	Recognition
Abdessamad et al. [43]	2024	Arabic	Deformable DETR	Detection
IndicDLP [41]	2025	11 Indic+EN	Layout Models	Layout

7. Emerging Trends and Advanced Techniques

The discipline is being transformed by a number of trends. Peng et al. developed UPOCR, which combines text removal, segmentation, and detecting if a document has been edited through the use of a ViT encoder-decoder model with task prompts that can be learned [44]. Zhang et al. created DocKylin, which uses multimodal LLM to compress pixels and tokens to improve document understanding [45]. Also, Zhang et al. improved how MLLM can use similarities to reduce image tokens as a method of improving reasoning with their Simignore approach [46]. Zhu et al. generated realistic text images using an LLM-guided conditional diffusion method to create images that could then be used to augment other datasets [47].

Gu et al. proposed using meta-learning through MetaWriter to implement personalized handwriting recognition with an improved ability to learn from experience and develop accurate prompts for recognizing handwriting style [48]. Chakraborty et al. designed a system for creating recognition of contextually relevant content without training, all while achieving excellent performance, through state-of-the-art techniques such as using less computational power

[49]. Ma et al. addressed the challenge of recognizing low resolution text onboard through super-resolution using text prior [50]. Fang et al. created methods that included editing of text whose recognition would provide visual consistency between recognition and edited forms [51], illustrated by Fig . 10.and As illustrated in Table 6: Emerging Methods and Techniques .

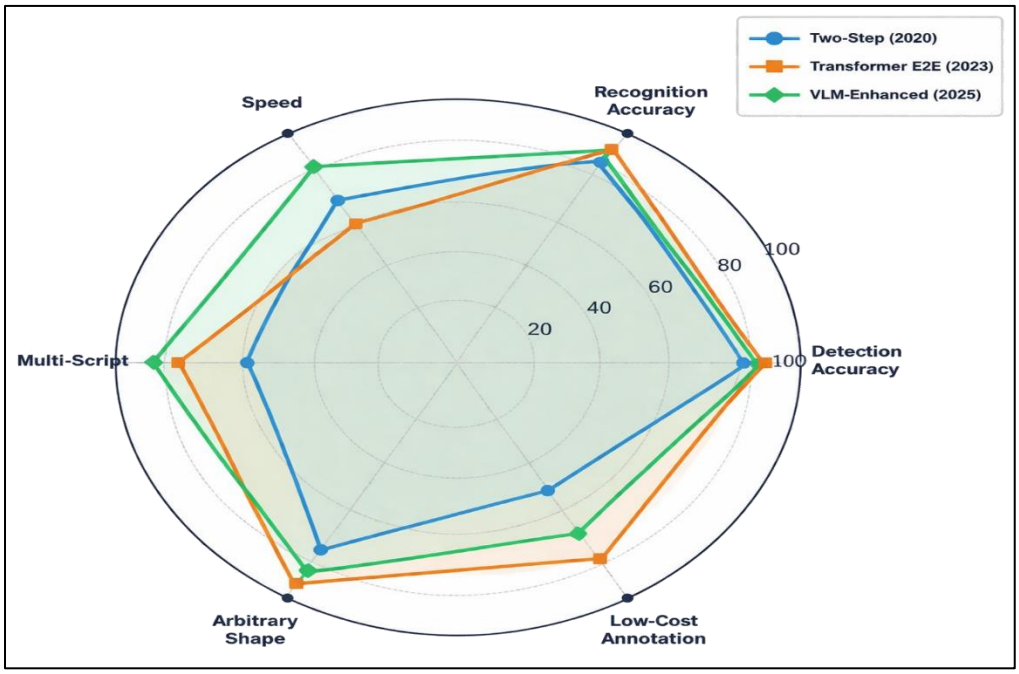

Fig. 10. Radar chart comparing capabilities of two-step, Transformer E2E, and VLM-enhanced approaches.

TABLE 6: Emerging Methods and Techniques

Method	Year	Category	Key Innovation	Application
TPGSR [50]	2023	Super-Resolution	Text Prior	Enhancement
UPOCR [44]	2024	Unified Multi-Task	Task Prompts	3 OCR Tasks
DocKylin [45]	2025	Document LLM	Token Slimming	VDU
SceneVTG [47]	2025	Text Generation	MLLM+Diffusion	Data Aug.
MetaWriter [48]	2025	Personalized HTR	Meta-Learning	Handwriting
Context-STR [49]	2025	Training-Free	Context-Driven	Efficient STR

8. Challenges, Analysis, and Future Directions

8.1 Critical Analysis of Current Limitations

There are a number of important limitations with our quantitative analysis. The first limitation is that the cost of annotation is too high. Annotations at the character level take 30 seconds each, but single point annotations cost only 1.5 seconds each (See Fig . 11). This creates a cost difference of 20 times between character level annotated datasets and single point annotated datasets; therefore, the ability to create large datasets based on character level annotations [27].Another limitation relates to the gap in performance between benchmark and real-world deployments, as shown in Table 7.

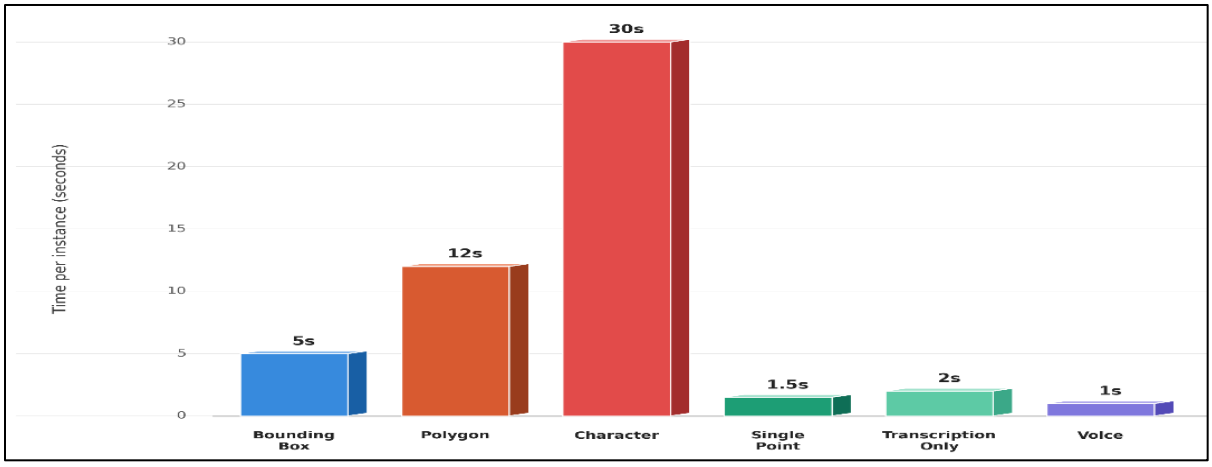

Fig. 11. Annotation cost comparison across supervision strategies.

TABLE 7: Critical Challenges Analysis

Challenge	Quantified Impact	Current Best Approach	Gap
Script Diversity	15-25% drop cross-script	Script-specific training	No universal model
Annotation Cost	20x cost difference	Single-point (SPTS)	5-8% accuracy gap
Real-time Speed	>100ms latency for E2E	DB++ (real-time det.)	E2E still slow
Low-Resource Lang.	<50% of scripts covered	Transfer learning	Script mismatch
Arbitrary Shapes	10-15% drop on curved	Bezier/polygon repr.	Complex post-proc.

8.2 Future Research Directions

According to research there are five main focal points (Fig. 12). The first is the unified/multi-purpose models for performing all detection, recognition, layout analysis and understanding tasks across all scripts and to use the same model for both UPOCR and DAT, therefore extending UPOCR and DAT capabilities [44].Second, unify DocKylin framework using LLM's to perform zero-shot recognition of new scripts in given contexts and reduce error rate by combining context and semantics to analyze errors in newly processed scripts based only on LLM output [45].Third, build multilingual self-supervised pre-training models using MTP [16].Fourth, use VLM's for fully unsupervised detection of script annotation costs to reduce annotation costs[28] .Fifth, provide training-free edge deployment solutions [50].

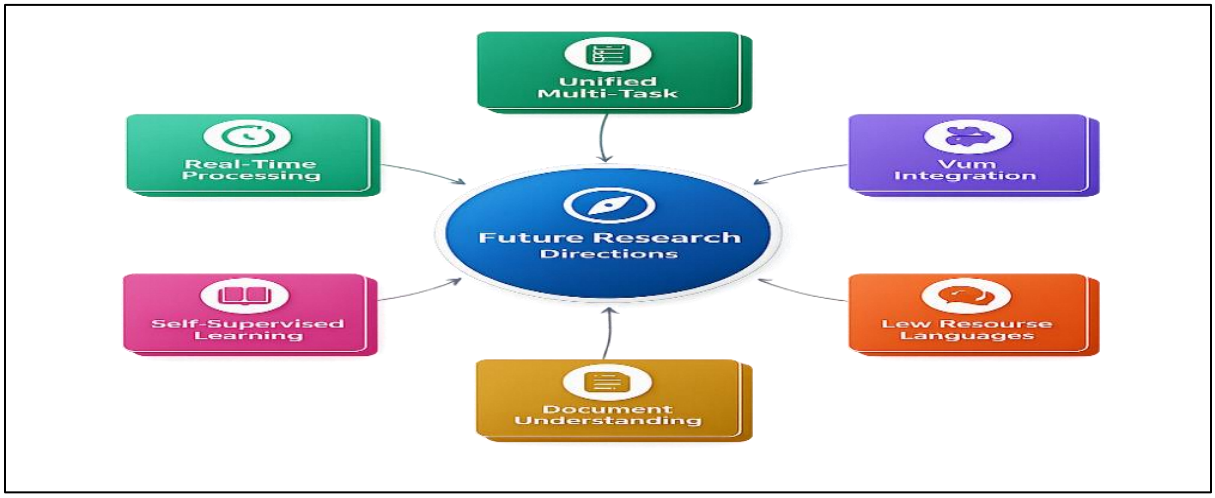

Fig. 12. Future research directions in multilingual text detection and recognition.

9. Conclusion

In this survey paper, we have performed a systematic analyze of deep learning algorithms that have been used for the detection and recognition of multilingual text found in 50 publications published from 2013 to 2025. This year's findings show that detection has been achieved by two leads with segmentation-based detection having an F-Measure of 87.9% on ICDAR2015 (DB++) and vision language recognition with an average accuracy of 94.1% by CLIP4STR, both of which represent the best currently available methods. Two methods that combine detection and recognition have been developed that also include features for edge computing and real time application. Furthermore, many of these new techniques using transformers combined with existing models of vision and language, and also x-large language models create a new opportunity for developing systems capable of understanding text (multilingual) for the first time. Future work should also address the use of joint models, self-supervised multilingual pre-training, and annotation methods with efficient implementations, and to allow the achievement of useable in practical and scalable applications regardless of the writing system.

10.References

[1] A. Hussein and M. S. M. Altaei, "Deep Learning Techniques for Detecting and Segmenting Text in Natural Scene Images: Review," *Al-Nahrain Journal of Science*, vol. 27, no. 2, pp. 133–144, Jun. 2024, doi: 10.22401/ANJS.27.2.14.

[2] X. Liu, G. Meng, and C. Pan, "Scene text detection and recognition with advances in deep learning: a survey," *Int. J. Doc. Anal. Recognit.*, vol. 22, no. 2, pp. 143–162, Jun. 2019, doi: 10.1007/s10032-019-00320-5.

[3] Z. Liu, R. Song, K. Li, and Y. Li, "From Detection to Understanding: A Systematic Survey of Deep Learning for Scene Text Processing," Sep. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app15179247.

[4] H. Zhang, K. Zhao, Y.-Z. Song, and J. Guo, "Text extraction from natural scene image: A survey," *Neurocomputing*, vol. 122, pp. 310–323, 2013, doi: https://doi.org/10.1016/j.neucom.2013.05.037.

[5] M. Huang *et al.*, "SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4583–4593. doi: 10.1109/CVPR52688.2022.00455.

[6] N. Nayef *et al.*, "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)*, 2017, pp. 1454–1459. doi: 10.1109/ICDAR.2017.237.

[7] P. Blanco-Medina, E. Fidalgo, E. Alegre, and V. González-Castro, "A survey on methods, datasets and implementations for scene text spotting," *IET Image Process.*, vol. 16, no. 13, pp. 3426–3445, 2022, doi: 10.1049/ipr2.12574.

[8] Md. A. Uddin *et al.*, "Deep learning-based human activity recognition using CNN, ConvLSTM, and LRCN," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 259–268, 2024, doi: https://doi.org/10.1016/j.ijcce.2024.06.004.

[9] S. Mahajan, R. Rani, and A. Kamboj, "Deep learning-based modified-EAST scene text detector: insights from a novel multiscript dataset: Deep learning-based modified-EAST...," *Int. J. Doc. Anal. Recognit.*, vol. 28, no. 1, pp. 97–119, Jul. 2024, doi: 10.1007/s10032-024-00491-w.

[10] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-Oriented and Multi-Lingual Scene Text Detection With Direct Regression," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5406–5419, 2018, doi: 10.1109/TIP.2018.2855399.

[11] T. Guan *et al.*, "Industrial Scene Text Detection With Refined Feature-Attentive Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6073–6085, 2022, doi: 10.1109/TCSVT.2022.3156390.

[12] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 919–931, 2023, doi: 10.1109/TPAMI.2022.3155612.

[13] W. Wang *et al.*, "PAN++: Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5349–5367, 2022, doi: 10.1109/TPAMI.2021.3077555.

[14] J. Zheng, H. Fan, and L. Zhang, "Kernel Adaptive Convolution for Scene Text Detection via Distance Map Prediction," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5957–5966. doi: 10.1109/CVPR52733.2024.00569.

[15] J. Tang *et al.*, "Few Could Be Better Than All: Feature Sampling and Grouping for Scene Text Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4553–4562. doi: 10.1109/CVPR52688.2022.00452.

[16] H. Xie *et al.*, "Masked Text Pre-Training for Scene Text Detection," *IEEE Trans. Multimedia*, vol. 27, pp. 9429–9443, 2025, doi: 10.1109/TMM.2025.3613181.

[17] X. Wan *et al.*, "Towards unified multi-granularity text detection with interactive attention," in *Proceedings of the 41st International Conference on Machine Learning*, in ICML'24. JMLR.org, 2024.

[18] A. Mostafa *et al.*, "OCFormer: A Transformer-Based Model For Arabic Handwritten Text Recognition," in *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 2021, pp. 182–186. doi: 10.1109/MIUCC52538.2021.9447608.

[19] V. Rajmod, G. Derkar, P. Nagrale, N. Awari, and P. Lokhande, "Text Extraction from Image using OCR," *6th International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2025 - Proceedings*, pp. 113–116, 2025, doi: 10.1109/ICMCSI64620.2025.10883061.

[20] M. Yang, B. Yang, M. Liao, Y. Zhu, and X. Bai, "Class-Aware Mask-guided feature refinement for scene text recognition," *Pattern Recognit.*, vol. 149, p. 110244, 2024, doi: https://doi.org/10.1016/j.patcog.2023.110244.

[21] Z. Wang, H. Xie, Y. Wang, J. Xu, B. Zhang, and Y. Zhang, "Symmetrical Linguistic Feature Distillation with CLIP for Scene Text Recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, in MM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 509–518. doi: 10.1145/3581783.3611769.

[22] S. Zhao, R. Quan, L. Zhu, and Y. Yang, "CLIP4STR: A Simple Baseline for Scene Text Recognition With Pre-Trained Vision-Language Model," *IEEE Transactions on Image Processing*, vol. 33, pp. 6893–6904, 2024, doi: 10.1109/TIP.2024.3512354.

[23] J. Xu, Y. Wang, H. Xie, and Y. Zhang, "OTE: Exploring Accurate Scene Text Recognition Using One Token," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 28327–28336. doi: 10.1109/CVPR52733.2024.02676.

[24] Y. Du, Z. Chen, Y. Su, C. Jia, and Y.-G. Jiang, "Instruction-Guided Scene Text Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2723–2738, 2025, doi: 10.1109/TPAMI.2025.3525526.

[25] M. Huang *et al.*, "ESTextSpotter: Towards Better Scene Text Spotting with Explicit Synergy in Transformer," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19438–19448. doi: 10.1109/ICCV51070.2023.01786.

[26] M. Huang, H. Li, Y. Liu, X. Bai, and L. Jin, "Bridging the Gap Between End-to-End and Two-Step Text Spotting," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15608–15618. doi: 10.1109/CVPR52733.2024.01478.

[27] D. Peng *et al.*, "SPTS: Single-Point Text Spotting," in *Proceedings of the 30th ACM International Conference on Multimedia*, in MM '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 4272–4281. doi: 10.1145/3503161.3547942.

[28] J. Tang, S. Qiao, B. Cui, Y. Ma, S. Zhang, and D. Kanoulas, "You Can even Annotate Text with Voice: Transcription-only-Supervised Text Spotting," in *Proceedings of the 30th ACM International Conference on Multimedia*, in MM '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 4154–4163. doi: 10.1145/3503161.3547787.

[29] Q. Qiao *et al.*, "DNTextSpotter: Arbitrary-Shaped Scene Text Spotting via Improved Denoising Training," in *Proceedings of the 32nd ACM International Conference on Multimedia*, in MM '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 10134–10143. doi: 10.1145/3664647.3680981.

[30] C. Duan *et al.*, "InstructOCR: instruction boosting scene text spotting," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, in AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. doi: 10.1609/aaai.v39i3.32286.

[31] D. Karatzas *et al.*, "ICDAR 2015 competition on Robust Reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1156–1160. doi: 10.1109/ICDAR.2015.7333942.

[32] N. Nayef *et al.*, "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)*, 2019, pp. 1582–1587. doi: 10.1109/ICDAR.2019.00254.

[33] Y. Sun *et al.*, "ICDAR 2019 Competition on Large-Scale Street View Text with Partial Labeling - RRC-LSVT," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1557–1562. doi: 10.1109/ICDAR.2019.00250.

[34] C. K. Chng *et al.*, "ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text - RRC-ArT," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1571–1576. doi: 10.1109/ICDAR.2019.00252.

[35] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1601.07140

[36] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8798–8808. doi: 10.1109/CVPR46437.2021.00869.

[37] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic Data for Text Localisation in Natural Images," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2315–2324. doi: 10.1109/CVPR.2016.254.

[38] B. Shi *et al.*, "ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17)," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1429–1434. doi: 10.1109/ICDAR.2017.233.

[39] H. Hassan, A. El-Mahdy, and M. E. Hussein, "Arabic Scene Text Recognition in the Deep Learning Era: Analysis on a Novel Dataset," *IEEE Access*, vol. 9, pp. 107046–107058, 2021, doi: 10.1109/ACCESS.2021.3100717.

[40] T. Do, T. Tran, T. Nguyen, D. D. Le, and T. D. Ngo, "SignboardText: Text Detection and Recognition in In-the-Wild Signboard Images," *IEEE Access*, vol. 12, pp. 62942–62957, 2024, doi: 10.1109/ACCESS.2024.3395374.

[41] O. Nath, S. Kukkala, M. Khapra, and R. K. Sarvadevabhatla, "IndicDLP: A Foundational Dataset for Multi-lingual and Multi-domain Document Layout Parsing," in *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part I*, Berlin, Heidelberg: Springer-Verlag, 2025, pp. 23–39. doi: 10.1007/978-3-032-04614-7_2.

[42] T. and M. N. and S. H. and I. Y. and K. K. Iwamura Masakazu and Matsuda, "Downtown Osaka Scene Text Dataset," in *Computer Vision – ECCV 2016 Workshops*, H. Hua Gang and Jégou, Ed., Cham: Springer International Publishing, 2016, pp. 440–455.

[43] A. Abdessamad, A. Habib, and A. Abdellah, “Arabic Text Detection From Natural Images Using Transformers,” in *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, 2024, pp. 1–6. doi: 10.1109/ISIVC61350.2024.10577807.

[44] D. Peng *et al.*, “UPOCR: towards unified pixel-level OCR interface,” in *Proceedings of the 41st International Conference on Machine Learning*, in ICML’24. JMLR.org, 2024.

[45] J. Zhang, W. Yang, S. Lai, Z. Xie, and L. Jin, “DocKylin: a large multimodal model for visual document understanding with efficient visual slimming,” in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, in AAAI’25/IAAI’25/EAAI’25. AAAI Press, 2025. doi: 10.1609/aaai.v39i9.33076.

[46] X. Zhang, F. Zeng, Y. Quan, Z. Hui, and J. Yao, “Enhancing multimodal large language models complex reason via similarity computation,” in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, in AAAI’25/IAAI’25/EAAI’25. AAAI Press, 2025. doi: 10.1609/aaai.v39i10.33107.

[47] J. and G. F. and L. W. and W. X. and W. P. and H. F. and Y. C. and Y. Z. Zhu Yuanzhi and Liu, “Visual Text Generation in the Wild,” in *Computer Vision – ECCV 2024*, E. and R. S. and R. O. and S. T. and V. G. Leonardis Aleš and Ricci, Ed., Cham: Springer Nature Switzerland, 2025, pp. 89–106.

[48] W. Gu, L. Gu, C. Y. Suen, and Y. Wang, “MetaWriter: Personalized Handwritten Text Recognition Using Meta-Learned Prompt Tuning,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 23494–23504. doi: 10.1109/CVPR52734.2025.02188.

[49] R. Chakraborty, P. Shivakumara, U. Pal, and C.-L. Liu, “A Lightweight Context-Driven Training-Free Network for Scene Text Segmentation and Recognition,” in *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part II*, Berlin, Heidelberg: Springer-Verlag, 2025, pp. 253–272. doi: 10.1007/978-3-032-04617-8_15.

[50] J. Ma, S. Guo, and L. Zhang, “Text Prior Guided Scene Text Image Super-Resolution,” *IEEE Transactions on Image Processing*, vol. 32, pp. 1341–1353, 2023, doi: 10.1109/TIP.2023.3237002.

[51] Z. Fang *et al.*, “Recognition-Synergistic Scene Text Editing,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 13104–13113. doi: 10.1109/CVPR52734.2025.01223.