WTR: Wild Text Recognition Framework Through Vision Transformer Dual Attention Mechanism

Shilpi Goyal[1], Dr. Deepak Motwani[2]
[1]*Research Scholar, Amity University,Gwalior, India*
[2]*Associate Professor, Amity University,Gwalior, India*

***Abstract*—Recognition and identification of text in an image is quite difficult due to the prevailing distortions, curved texts, asymmetrical alignments and irregular character spacing. Currently transformer-based methods are used for scene text recognition. It has to place reliance on global self-attention which can recognize horizontally aligned text but unable to identify fine grained characters in complicated wild scenes. This paper explores a new framework called WTR (Wild Text Recognition) which proposes to be a set of dual attention tool which uses global attention for sequence coherence and local attention for fine-grained character features. Two-dimensional learnable positional encodings, within the Vision Transformer encoder, has been used in order to augment spatial adaptability. It further enhances recognition accuracy using language-aware post-processing module which is based on pre-trained language models. Under standard experimental conditions, it has been found that WTR method exceeds other existing state of the art methods as it achieves 96.4% accuracy on IIIT5K, 94.2% on Total-Text, and 96.4% on SCUT-CTW1500, showing robustness against curved, rotated, and perspective-distorted text.**

***Index Terms*—Scene Text Recognition, Vision Transformer, Dual Attention, 2D Positional Encoding, Language Model.**

I. INTRODUCTION

Contrary to Optical Character Recognition (OCR) Scene Text Recognition (STR) is a method which intends to extract textual content from various images, such as street sign boards, digital displays, and billboards etc., drawn from natural environment. Due to perspective distortions, curved layouts, and continually changing fonts, this method faces multiple challenges in scene text recognition. Above stated discrepancies make character recognition difficult in the wild situations. It is yet another difficult crucial problem for applications in autonomous navigation, assistive technologies, and visual information retrieval [1].

Earlier rectification modules were used to extract irregular text. Shi et al. [2] familiarized the world with a Spatial Transformer Network (STN) with a CNN-RNN decoder to regulate distorted text, while ASTER [3] integrated rectification with an attention-based decoder.

These earlier approaches involved in text recognitions were reasonably effective in moderate distortions. However, these methods have been found to be less effective in highly curved or arbitrarily shaped text.

Introduction of attention mechanism into transformer-based models has displayed better results. Vision Transformer (ViT) [4] encodes images into patches, and thus, it facilitates long-range dependency modeling. ViTSTR [5] used this concept to STR by considering cropped text images as patch sequences. Similarly, PARSeq [6] used parallel decoding, while ABINet [7] unified visual and linguistic features. These above stated models chiefly depend on global attention which is an effective tool in extracting structured text. However, these models face difficulty while dealing with fine-grained variations such as stroke differences and irregular character spacing.

Modern research has discovered local or deformable attention to enhance adaptableness. Swin Transformer [8] restricted attention to local windows for efficiency, while Deformable ViT [9] dynamically adjusted receptive fields. In the context of STR, SText-DETR [10] engaged scalable probes for arbitrary-shaped text detection. Despite of modern research and advances in the field of scene text detection, certain drawbacks exist. Current models either emphasize global context at the cost of character-level detail or focus too narrowly, missing long-range coherence.

To fill the gap, we propose Wild Text Recognition (WTR), a transformer-based method featured as a dual attention tool, is propounded with the intention to integrate (1) Local Attention, that captures fine-grained character-level details by correlating neighboring patches. (2) Global Attention, which models long-range coherence across the entire text sequence.
Moreover, WTR uses two-dimensional learnable positional encoding with the intention to increase spatial adaptability and a language-aware post-processing module to refine predictions at the word level. This hybrid method improves robustness against curved, rotated, and distorted text, enabling accurate recognition in complex wild scenes.

The main features of WTR can be summed up as:

- It presents a new dual-attention transformer-based methodology to facilitate wild scene text recognition.
- It uses 2D learnable positional encoding that adjusts itself to asymmetrical spatial arrangements.
- A language-aware post-processing segment is added to improve sequence accuracy by leveraging contextual semantics.
- WTR, having gone through various benchmarks, achieves state-of-the-art performance. It outclasses existing STR models on both regular and irregular datasets.

II. RELATED WORK

A. Rectification-Based Methods

Early Scene Text Recognition (STR) systems struggled with distortions such as perspective warping and curved baselines. To address this, rectification modules were integrated into recognition pipelines. Shi et al. [2] introduced the Spatial Transformer Network (STN) with a CNN-RNN decoder to normalize text before recognition. ASTER [11] extended this by combining STN with an attention-based decoder, achieving strong results on moderately curved text. However, rectification is limited when text exhibits extreme irregularities, as the transformation itself may distort fine-grained features.

B. Attention-Based Models

To bypass explicit rectification, researchers explored attention mechanisms. SAR (Show-Attend-Read) [12] applied 2D attention directly over feature maps, localizing irregular characters without preprocessing. Later, SRN [13] incorporated semantic reasoning to improve sequence modeling. These approaches improved flexibility but often struggled with capturing long-range dependencies and preserving fine spatial details, especially in cluttered or noisy backgrounds.

C. Transformer-Based Recognition

Transformers revolutionized STR by enabling direct modeling of sequence dependencies. ViTSTR [5] applied Vision Transformer (ViT) [4] to scene text, treating cropped word images as patch sequences. PARSeq [6] introduced parallel decoding to accelerate inference, while ABINet [7] incorporated both visual and linguistic information into a unified framework. More recently, CLIP4STR [14] leveraged vision-language pre-training to enhance recognition. While these methods demonstrate strong results, they primarily depend on global self-attention, which effectively models overall context but may fail to capture subtle character-level differences such as stroke variations or irregular inter-spacing.

D. Hybrid and Deformable Attention

To overcome the rigidity of global self-attention, hybrid and deformable attention mechanisms have been proposed. Swin Transformer [8] restricted attention to local windows, improving efficiency, while Deformable ViT [9] allowed dynamic receptive fields that adapt to object geometry. In the STR domain, SText-DETR [10] introduced scalable queries for arbitrary-shaped text detection. Although these techniques enhance adaptability, they still face a trade-off: local methods preserve detail but lose global coherence, while global methods capture sequence structure but miss fine granularity.

E. Our Contribution

Building on these observations, we propose Wild Text Recognition (WTR), which introduces a dual attention mechanism within a ViT-like framework. WTR integrates (1) Local Attention, to capture fine-grained details across neighboring patches, and (2) Global Attention, to ensure long-range sequence coherence.
Additionally, WTR incorporates 2D learnable positional encoding for improved spatial awareness and a language-aware post-processing module for contextual refinement. This design addresses the shortcomings of

prior works by simultaneously modeling fine-grained character features and global dependencies, leading to superior recognition performance in wild, irregular text conditions.

III. PROPOSED METHOD: A WTR FRAMEWORK

The proposed Wild Text Recognition (WTR) framework leverages a Vision Transformer backbone enhanced with a dual attention mechanism, 2D learnable positional encoding, and a language-aware post-processing module. The overall architecture is shown in Fig. 1. WTR aims to address both fine-grained local details (e.g., character strokes) and long-range dependencies (e.g., inter-character coherence) for robust recognition of irregular text in natural scenes.

A. Patch Embedding with 2D Learnable Positional Encoding

Given an input cropped text image, $I \in \mathbb{R}^{H\times W\times C}$, we divide it into non-overlapping patches of size $P \times P$. Each patch is flattened and projected into an embedding vector using a learnable linear projection, as in (1):

$$E = \text{Linear}(\text{Flatten}(I)) \in \mathbb{R}^{N\times D} \quad (1)$$

where $N = \frac{HW}{P^2}$ is the number of patches and D is the embedding dimension.

Unlike standard ViT, which uses fixed sinusoidal encodings, WTR combines a two-dimensional (2D) learnable positional encoding with patch embedding as in (2):

$$\hat{E} = E + P_{2D}, \quad P_{2D} \in \mathbb{R}^{N\times D} \quad (2)$$

This encoding adapts with irregular character outlines and enhances spatial awareness for both curved and rotated text.

B. Dual Attention Block

To capture both local character details and global sequence coherence, WTR introduces a dual attention block composed of:

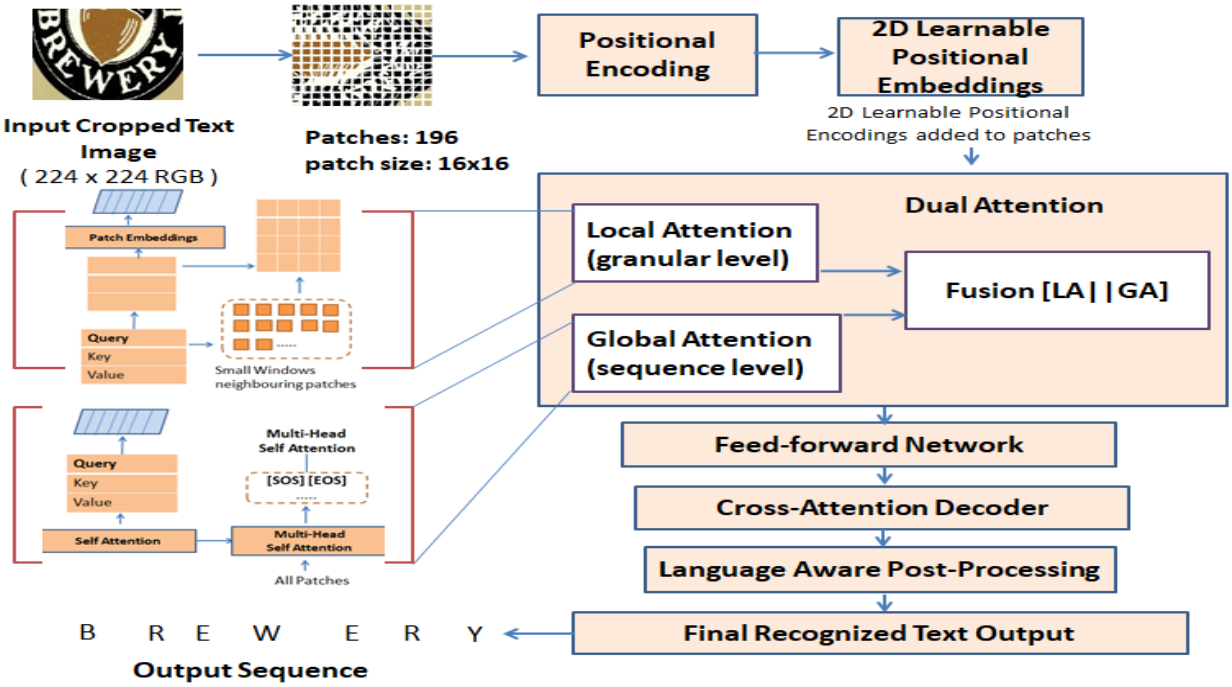

Fig. 1. Overall architecture of the proposed Wild Text Recognition (WTR) framework. The input image is divided into patches and embedded with 2D positional encoding. A Dual Attention Block combines Local Attention (character-level details) and Global Attention (sequence coherence). The fused features are decoded via cross-attention with character embeddings, followed by language-aware post-processing for semantic refinement of the final output. Dual Attention Blocks in WTR. (a) Local Attention captures fine-grained details by restricting attention to a neighborhood of each patch. (b) Global Attention models long-range dependencies across the entire sequence of patches. Outputs from both modules are fused to preserve local precision and global coherence.

1) Local Attention (LA)

Local Attention focuses on fine-grained spatial details within a restricted neighborhood of each patch. Given query, key, and value representations for a patch and its neighborhood, as in (3) and (4).

$$LA(Q_i, K, V) = \sum_{j \in \mathcal{W}(i)} \alpha_{ij} V_j \tag{3}$$

$$\alpha_{ij} = \frac{\exp(Q_i K_j^\top / \sqrt{d_k})}{\sum_{j' \in \mathcal{W}(i)} \exp(Q_i K_{j'}^\top / \sqrt{d_k})} \tag{4}$$

This mechanism allows the model to distinguish subtle variations in strokes, edges, and character components.

2) Global Attention (GA)

Global Attention aggregates contextual information across the entire sequence of patches, as in (5), using standard multi-head self-attention (MHSA). This ensures coherence across characters, capturing dependencies such as ligatures, spacing, and sequence order.

$$GA(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{5}$$

3) Fusion of LA and GA

Outputs from local (3) and global (5) branches are fused to form a comprehensive representation in (6):

$$Z = \text{FFN}(\text{LayerNorm}([LA \| GA])) \tag{6}$$

where $[LA \| GA]$ denotes concatenation. The fusion preserves both local precision and global consistency.

C. Cross Attention Decoder

We use cross-attention, as in (7), to decode character sequences. Character embedding serves as queries, while patch embedding act as key-value pairs. The decoder outputs are then passed through linear layers for classification over 57-character classes (A-Z, a-z, 0-9, [SOS], [EOS], [pad]) and ignoring the common upper cases (indicating through (-) sign) as shown in Table 1.

$$Z_j = Attention\ (C_j, K_{img}, V_{img}) \tag{7}$$

where K_{img}, V_{img} are image patch features and C_j is character

TABLE 1 CHARACTER CLASSIFIER LISTED FOR TEXT CLASSIFICATION

Classes	Description	Total classes
0-9	Numerals	(+)10
A-Z	Upper case English characters	(+)26
a-z	Lower case English characters	(+)26
C, O, S, V, W, X, Z, U	Common geometrical shapes as lower case characters	(-)8
[SOS], [EOS],[pad]	Special tokens	(+)3

D. Language-Aware Post-Processing

Characters might be spatially located across multiple patches leading to ambiguous embedding of text. Language-aware post-processing helps to precise the sequence of characters by de-duplicating the characters. Existing language model (LM) is used as post-processing step to de-duplicate the characters and enhance the quality of recognition of text as a word. Further we improve recognition by re-ranking or correcting based on word-level context using BERT language models. E.g., if the classifier outputs "BREWERR", LM might correct to "BREWERY". The decoded sequence is aligned and converted into a word and decode using a language transformer.

IV. EXPERIMENTAL SETUP

A. Datasets

To evaluate WTR, we use a combination of synthetic and real-world benchmarks commonly used in Scene Text Recognition (STR).

- Pretraining datasets
  - MJSynth (Synth90k) [15]: 9M synthetic word images covering 90k English words.
  - SynthText [16]: 8M synthetic images generated by embedding text into natural scenes with geometric transformations.

- Fine-tuning and evaluation datasets
  - IIIT5K-Words (IIIT5K) [17]: 5,000 word images from websites.
  - Street View Text (SVT) [18]: 647 cropped images from Google Street View.

- ICDAR 2013 (IC13) [19]: 1,015 cropped text images.
- ICDAR 2015 (IC15) [20]: 1,811 word images with strong perspective distortions.
- SVTP [21]: 645 text images with perspective distortion.
- CUTE80 [22]: 288 curved text samples.
- SCUT-CTW1500 (SCUT) [23]: 500 images with arbitrarily curved text.
- Total-Text [24]: 500 images containing multi-oriented and curved text.

Together, these benchmarks cover regular text (IIIT5K, SVT, IC13) and irregular text (IC15, SVTP, CUTE80, SCUT, Total-Text).

B. Evaluation Metric

We adopt Word Recognition Accuracy (WRA) as defined in (8). A prediction is correct if the entire word matches the ground truth.

$$WRA(\%) = \frac{No.of\ Correctly\ Recognized\ Words}{Total\ number\ of\ Words} \times 100 \quad (8)$$

C. Data Augmentation

To improve generalization, we apply the following transformations as shown in Fig. 2.

- Gaussian blur and additive noise.
- Perspective warp and rotation (±30°).
- Non-linear distortions (curvature, stretching).
- RandAugment [25] for automated augmentation policy search.

Fig. 2. Illustration of augmented text images designed for WTR.

D. Training Details

- Framework: PyTorch, trained on NVIDIA Tesla T4 (Colab Pro).
- Image size: 224×224.
- Optimizer: AdamW with learning rate 5×10−3
- Gradient clipping: 5.0.
- Loss: Cross-Entropy over 57-character classes.
- Epochs: 2 (pre-train) + 20 (fine-tuning).
- Batch size: 192 (Tiny), 128 (Small), 64 (Base).
- Models trained: WTR-Tiny, WTR-Small, WTR-Base with varying depth and embedding sizes as shown in Table 2.

TABLE.2 WTR CONFIGURATIONS

WTR Version	Patch Size	Depth	Embedding Size	No. of Heads	Sequence Length
Tiny	16	12	192	3	27
Small	16	12	384	6	27
Base	16	12	768	12	53

IV. RESULTS AND DISCUSSION

A. Quantitative Results

WTR-Base consistently outperforms most baselines, especially on **irregular datasets** (CUTE80, SCUT, Total-Text), demonstrating its robustness to distortions in Table 3.

Fig.3 shows the Comparison of Accuracy on three datasets on State-of-the-art methods and Fig.4 represents the Comparison of Accuracy parameter on different three datasets for proposed methods.

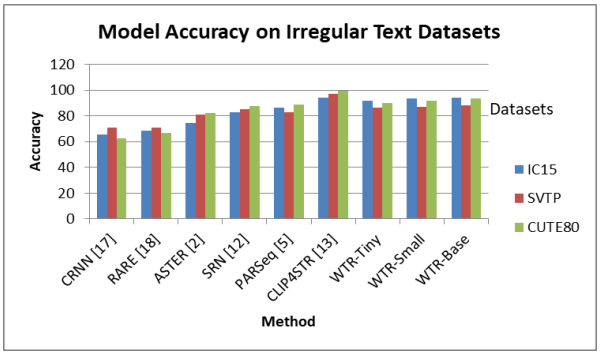

Fig. 3. Comparison of parameter Accuracy on three datasets on State-of-the-art methods.

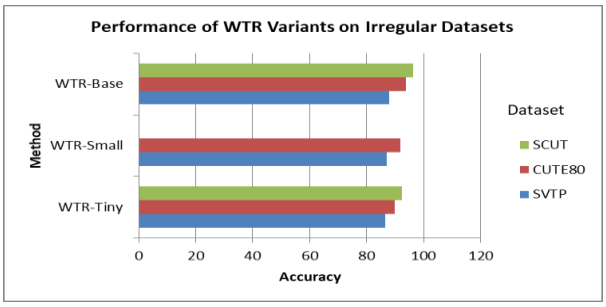

Fig. 4. Comparison of parameter Accuracy on three datasets for proposed methods.

B. *Ablation Studies*

To validate each component of WTR, we conducted ablation experiments on Total-Text and SCUT as shown in Table 4. Both Local Attention (LA) and Global Attention (GA) contribute significantly, while Language Model (LM) provides additional refinement.

TABLE 4. ABLATION STUDY RESULTS (WORD ACCURACY %)

Model Variant	Total-Text	SCUT
WTR without LA	90.4	92.1
WTR without GA	89.7	91.8
WTR without LM	91.1	92.5
Full WTR (LA+GA+LM)	95.4	96.4

C. *Qualitative Results*

Fig. 5 illustrates sample predictions. WTR successfully recognizes:

TABLE 3. WORD RECOGNITION ACCURACY (%) ON REGULAR AND IRREGULAR BENCHMARKS

Method	IIIT 5K	SVT	IC 13	IC 15	SVTP	CUTE80	SCUT	Total-Text
CRNN [26]	81.8	80.1	88.4	65.8	70.8	62.7	–	–
RARE [27]	86.0	85.4	93.5	68.6	71.1	66.8	–	–
ASTER [3]	91.8	89.5	90.9	74.4	80.9	81.9	60.7	–
SRN [13]	94.8	91.5	–	82.7	85.1	87.8	–	–
PARSeq [6]	97.0	93.6	96.2	86.5	82.9	88.9	92.2	–
CLIP4STR [14]	97.7	95.2	96.1	94.1	97.2	99.3	81.1	81.1
WTR-Tiny	94.3	93.2	92.6	91.9	86.5	90.0	92.5	92.8
WTR-Small	95.6	94.3	93.9	93.4	87.1	91.9	94.0	93.5
WTR-Base	96.4	95.8	95.2	94.2	87.9	93.7	96.4	95.4

- Curved text (“ALABAMA” in Total-Text).
- Rotated text (“FUTUTESTARS” text in SCUT).
- Distorted characters (“MEXOUT” in Total-Text)

These examples confirm WTR’s robustness to spatial variations.

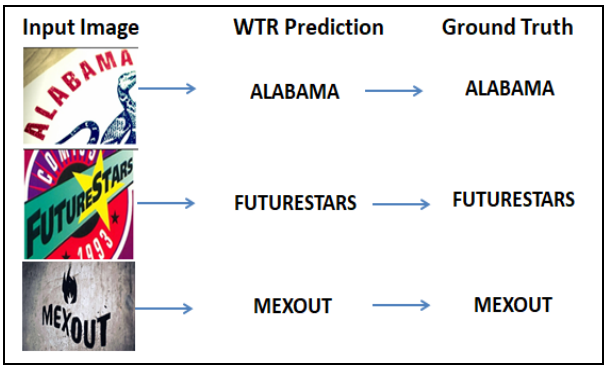

Fig. 5. Qualitative results of WTR on challenging samples. Top row: Curved text (“ALABAMA” from Total-Text). Middle row: Rotated text (“FUTURESTARS” from SCUT). Bottom row: Distorted characters (“MEXOUT” from Total-Text). These examples demonstrate the robustness of WTR against curvature, rotation, and character distortion.

D. *Discussion*

The proposed WTR framework demonstrates several prominent strengths. The introduction of the dual attention mechanism provides a balance between fine-grained local precision and long-range global coherence, allowing the model to effectively capture both character-level details and sequence-level dependencies. The incorporation of two-dimensional learnable positional encoding further enhances the adaptability of the model to irregular layouts, such as curved or rotated text, which are common in natural scenes. Additionally, the language-aware post-processing module significantly improves semantic correctness by refining raw predictions and correcting ambiguous outputs based on contextual information.

Despite these advantages, WTR also has certain limitations. First, training requires large-scale synthetic pre-training datasets such as MJSynth and SynthText, which may not always be accessible or efficient for every research setting. Second, compared to lightweight CNN-RNN models, WTR has a higher memory and computational cost, which could limit its deployment in resource-constrained environments.

Looking forward, there are several promising directions for future research. One important direction is to optimize WTR for real-time deployment on mobile and embedded devices, ensuring efficiency without compromising accuracy. Another is to extend the framework to multilingual scene text recognition, addressing diverse scripts and languages beyond English. These enhancements would broaden the applicability of WTR and strengthen its role as a practical solution for robust text recognition in unconstrained environments.

V. CONCLUSION

This paper presents a new method as Wild Text Recognition (WTR) which facilitates robust and accurate scene text detection from complicated natural environment.

Existing transformer-based methods rely on global attention; on the other hand, WTR uses a dual attention framework that combines Local Attention for fine-grained character-level details and Global Attention for long-range sequence coherence. Moreover, two-dimensional learnable positional encoding has been integrated so as to enhance spatial adaptability and a language-aware post-processing module to refine recognition at the word level.

Wide research on multiple public benchmarks reveals that WTR outperforms other available regular and irregular tools such as IIIT5K, SVT, IC13 IC15, SVTP, CUTE80, SCUT. Ablation studies confirm that each component of the framework—local attention, global attention, and language modeling—contributes significantly to the overall accuracy.

Despite of giving better accurate results in scene text recognition, WTR has certain limitations. The dependency on large-scale synthetic pre training escalates computational cost, and the transformer backbone needs extra memory compared to lightweight CNN-RNN models. However, these challenges pave way for further possible areas of research.

In the future, it is intended to:

- Develop lightweight variants of WTR for mobile and real-time applications.
- Extend the WTR tool to multilingual text recognition, addressing scripts beyond English.
- Investigate self-supervised and semi-supervised training strategies to reduce reliance on large labeled datasets.

By addressing the inherent anomalies of wild scene text through a hybrid attention design, WTR proves to be a beneficial foundation for next-generation text recognition systems in unrestricted environments.

REFERENCES

[1] S. Goyal and D. Motwani, “A Study of Text Extraction Algorithms for Natural Scene Images,” SN Comput. Sci., vol. 5, no. 6, p. 731, Jul. 2024, doi: 10.1007/s42979-024-03068-w.

[2] B. Shi, X. Bai, and C. Yao, “An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, 2017, doi: 10.1109/TPAMI.2016.2646371.

[3] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, “ASTER: An Attentional Scene Text Recognizer with Flexible Rectification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2035–2048, 2019, doi: 10.1109/TPAMI.2018.2848939.

[4] A. Dosovitskiy et al., “an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale,” ICLR 2021 - 9th Int. Conf. Learn. Represent., 2021.

[5] R. Atienza, “Vision Transformer for Fast and Efficient Scene Text Recognition,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12821 LNCS, pp. 319–334, 2021, doi: 10.1007/978-3-030-86549-8_21.

[6] D. Bautista and R. Atienza, “Scene Text Recognition with Permuted Autoregressive Sequence Models,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 13688 LNCS, pp. 178–196, 2022, doi: 10.1007/978-3-031-19815-1_11.

[7] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, “Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern

Recognit., pp. 7094–7103, 2021, doi: 10.1109/CVPR46437.2021.00702.

[8] Z. Liu et al., "Swin Transformer," 2021 IEEE/CVF Int. Conf. Comput. Vis., pp. 9992–10002, 2021, [Online]. Available: https://ieeexplore.ieee.org/document/9710580/.

[9] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision Transformer with Deformable Attention," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2022-June, pp. 4784–4793, 2022, doi: 10.1109/CVPR52688.2022.00475.

[10] P. Liao and Z. Wang, "SText-DETR: End-to-End Arbitrary-Shaped Text Detection with Scalable Query in Transformer," in Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2024, pp. 481–492, doi: 10.1007/978-981-99-8546-3_39.

[11] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2035–2048, Sep. 2019, doi: 10.1109/TPAMI.2018.2848939.

[12] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, pp. 8610–8617, 2019, doi: 10.1609/aaai.v33i01.33018610.

[13] D. Yu et al., "Towards accurate scene text recognition with semantic reasoning networks," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, pp. 12110–12119, 2020, doi: 10.1109/CVPR42600.2020.01213.

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

[15] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition," arXiv Prepr. arXiv1406.2227, pp. 1–10, 2014, [Online]. Available: http://arxiv.org/abs/1406.2227.

[16] A. Gupta, Ankush and Vedaldi, Andrea and Zisserman, "Synthetic data for text localisation in natural images," in Gupta, Ankush and Vedaldi, Andrea and Zisserman, Andrew, 2016, pp. 2315--2324.

[17] C. Mishra, Anand and Alahari, Karteek and Jawahar, "Scene text recognition using higher order language priors," in BMVC-British machine vision conference, 2012.

[18] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," Proc. IEEE Int. Conf. Comput. Vis., no. 4, pp. 1457–1464, 2011, doi: 10.1109/ICCV.2011.6126402.

[19] D. Karatzas et al., "ICDAR 2013 robust reading competition," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 1484–1493, 2013, doi: 10.1109/ICDAR.2013.221.

[20] D. Karatzas et al., "ICDAR 2015 competition on Robust Reading," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015-Novem, pp. 1156–1160, 2015, doi: 10.1109/ICDAR.2015.7333942.

[21] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," Proc. IEEE Int. Conf. Comput. Vis., pp. 569–576, 2013, doi: 10.1109/ICCV.2013.76.

[22] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 9357–9366, 2019, doi: 10.1109/CVPR.2019.00959.

[23] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting Curve Text in the Wild: New Dataset and New Solution," 2017, [Online]. Available: http://arxiv.org/abs/1712.02170.

[24] C. K. Ch'Ng and C. S. Chan, "Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 1, pp. 935–942, 2017, doi: 10.1109/ICDAR.2017.157.

[25] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," Adv. Neural Inf. Process. Syst., vol. 2020-Decem, 2020.

[26] I. Campiotti and R. Lotufo, "Optical character recognition with transformers and CTC,"

DocEng 2022 - Proc. 2022 ACM Symp. Doc. Eng., 2022, doi: 10.1145/3558100.3563845.

[27] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust Scene Text Recognition with Automatic Rectification," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 4168–4176, 2016, doi: 10.1109/CVPR.2016.452.