Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 12, No. 3, 2024, 455-464.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

A Transformer-Based Approach with Contextual Position Encoding for Robust Persian Text Recognition in the Wild

Zobeir Raisi* and Vali Mohammad Nazarzehi Had

Electrical Engineering Department, Chabahar Maritime University, Chabahar, Iran.

Article Info

Article History:
Received 15 July 2024
Revised 30 October 2024
Accepted 09 December 2024

DOI:10.22044/jadm.2024.14669.2569

Keywords:
Scene Text Recognition, Persian Scripts, Contextual Position Encoding, Transformers, Deep Learning.

**Corresponding author: zobeir.raisi@cmu.ac.ir (Z. Raisi).*

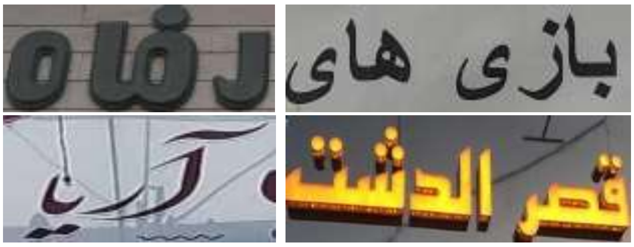
Abstract

The Persian language presents unique challenges for scene text recognition due to its distinctive script. Despite advancements in AI, recognition in non-Latin scripts like Persian still faces difficulties. In this study, we enhance the vanilla transformer architecture to recognize arbitrary shapes of Persian text instances. We apply Contextual Position Encoding (CPE) to the baseline transformer architecture to improve the recognition of Persian scripts in wild images, especially for oriented and spaced characters. The CPE utilizes position information to generate contrastive data pairs that help better in capturing Persian characters written in a different direction. Moreover, we evaluate several cutting-edge deep-learning models using our prepared challenging Persian scene text recognition dataset and develop a transformer-based architecture to enhance recognition accuracy. Our proposed scene text recognition architecture achieves superior word recognition accuracy compared to existing methods on a real-world Persian text dataset.

1. Introduction

Scene text recognition revolutionizes numerous tasks, ranging from document analysis to augmented reality and self-driving vehicles. It involves capturing an image of text, such as a sign or handwriting, and converting it into readable words or letters. The diverse and challenging conditions in images captured in natural settings make it difficult to develop an effective recognition method [7, 25, 26]. Deep learning frameworks, especially those for deep convolutional neural networks (CNNs) as well as recurrent neural networks (RNNs), form a foundation for STR methods, often coupled with techniques like connectionist temporal classification (CTC) for sequence prediction. However, these approaches face challenges with irregular text datasets [4, 6, 48].

Transformers and their variants like Performer have revolutionized deep learning architectures for scene text recognition (STR) and achieved superior performance in several benchmark datasets [32, 34].



(a) **(b)**

Figure 1. Current positional encoding methods face challenges in recognizing characters within a word when spaces are present between them.

They use a global attention mechanism for character encoding and decoding within the text image, enabling independent operation and successful application in sequential data tasks. The core strength of transformers is a multi-head self-attention mechanism, which allows for dynamic focus on relevant portions of the input data. Positional Encoding (PE) is essential for the transformer's success, mitigating permutation equivariance in self-attention and making the Transformer consider the sequence of the given input characters [10]. The most common approach is Sinusoidal PE (SPE), while other techniques like fully Learnable PE (LPE) and Relative PE (RPE) have been explored [32, 33]. These data-driven approaches learn positional encodings from training data but may limit generalizability [10, 24].

Most scene text recognition techniques are developed for left-to-right Latin text languages. However, a handful of research studies focus on recognizing text instances in images taken from real-world environments, particularly those written in right-to-left languages such as Persian. Applying or fine-tuning a pre-training Latin-text model may not lead to a good performance in Persian scripts because the characters and the way of writing of Persian scripts are completely different from Latin text like English, and this language has specific challenges different from English.

For example, as seen in Figure 1, the transformer architecture with Sinusoidal Positional Encoding that is used in many works [21, 31, 32] has difficulties in recognizing separate characters of word instances. The other problem in Persian scene text recognition is the lack of a real-world benchmark dataset. In contrast to several English image datasets (real or artificial) [17, 18, 27, 36], there are only a few Persian datasets prepared. Still, some of these datasets are not publicly available as in [19], or prepared for a specific task of scene text recognition as in [9, 30], or only used synthetic images as in [2].

To tackle these challenges, this paper initially prepared a comprehensive test set for Persian scene text recognition by considering various images found in natural environments. This dataset serves as a benchmark for future research in Persian scene text recognition.

Afterward, we evaluate different CNN, RNN-based methods [3, 5, 23, 37-39], and Transformer based deep learning techniques [29, 33] on the prepared dataset. We show that transformer methods achieve better results than the CNN and RNN models for many challenges of Persian text instances. To that effect, we consider the transformer-based method with SPE as a baseline. However, as shown in Figure 1, baseline methods still face many challenges. Many failures occur when text instances contain irregular text and spaces between characters.

To address this issue, we utilize CPE as positional encoding and demonstrate that this technique enhances the model's accuracy on images commonly found in Persian scripts. CPE enables positions to be conditioned on context by incrementing position values only for specific tokens as determined by the model. The present study introduces a text-alignment module within the framework to enhance the capture of spatial information for Persian characters. Furthermore, we propose modifications to the encoder module’s feed-forward network layer. These modifications aim to enhance the network’s robustness to extract meaningful features reliably from the self-attention outputs. Our work offers several novel contributions, including:

- We designed an encoder-decoder transformer-based architecture with spatial positional encoding to recognize irregular text in Persian scripts.
- We modified different parts of the baseline architecture, especially the PE, to make it more suitable for Persian scene text recognition.
- We are the first to apply the 2D-CPE to the transformer architecture to Persian scene text recognition.
- Compared to current best practices, our model achieves significantly higher recognition accuracy, as demonstrated by the experiments, including, [3, 5, 23, 37-39] on the collected test set of Persian text recognition in the wild images.
- For evaluation, we constructed a dataset comprising real-world cropped word images in the Persian script from various environments.
- We also created comprehensive synthetic images of Persian scripts to train the model.
- To comprehensively evaluate the efficacy of leading-edge models, we conducted an in-depth ablation study on our proposed dataset.
- On the Persian dataset, the proposed model surpasses all existing methods, achieving the highest WRA score.

2. Related Work

Cutting-edge STR approaches fall into two categories: RNN-CNN architectures use RNNs for sequential processing and CNNs for feature extraction, whereas transformers specialize in capturing long-range dependencies important for complex scene text recognition.

2.1. RNN and CNN-based STR

Deep learning approaches to scene text recognition typically involve a three-stage architecture.

Table 1. The Acronyms and definitions of different positional encodings used in transformer architectures.

Abbreviation	Description
PE	Positional Encoding
SPE [42]	Sinusoidal PE with 1D fixed frequencies
LPE [8]	Learnable PE with 1D weights
LSPE [44]	Learnable Sinusoidal Positional Encoding with 1D frequencies
2SPE [32]	Sinusoidal Positional Encoding in 2D
2LSPE [33]	Learnable Sinusoidal Positional Encoding in 2D
CPE(*Proposed*)	**Contextual Positional Encoding**

A CNN extracts image features, a recurrent neural network (RNN) models sequential dependencies and a prediction module generates the recognized character sequence. While RNN-based methods [3, 5, 23, 37-39] are effective for horizontal or near-horizontal text, they struggle with arbitrary shapes or distortions. Rectification modules [15] attempt to address highly curved text but may introduce errors due to distortion during perspective transformation.

Researchers have proposed various methods concerning the prediction head. Some methods [5, 23, 37] utilize a technique called Connectionist Temporal Classification for prediction and employ a VGG model as the feature extractor. This is followed by a Bidirectional Long Short-Term Memory (BLSTM) network [13]. Some Recent approaches [3, 38] use an attention mechanism to enhance recognition performance. While effective for some scenarios, these methods might not be the best choice for curved or rotated text. For instance, in [38], the authors proposed a spatial attention mechanism within an STN framework [15] to convert distorted text regions into a rectified format that is easy to recognize. In subsequent work, they employed a Thin-Plate Spline transformation using control points for improved rectification of curved text, leading to enhanced recognition performance on datasets containing irregular text. However, a limitation of these rectification-based methods (including [3, 38, 39, 46]) is leveraging one-dimensional (1D) features. This inherent characteristic makes them less suitable for directly recognizing irregular text examples, as they lose the essential spatial information encoded inside the two-dimensional image data.

2.2. Transformer based STR

The transformer architecture, a novel framework for natural language processing (NLP) introduced in [42], offers advantages over convolutional RNNs-CNNs neural networks. Due to its effectiveness in handling sequential data, it is now being used for lots of computer vision tasks.

Like language modelling, the sequential order of characters within a word and words within a sentence plays a critical role in scene text recognition. This dependency on character and word order has motivated scene text recognition's recent adoption of transformer-based architectures [16, 21, 32]. These methods, utilizing various positional encoding (PE) schemes, have surpassed the performance of prior cutting-edge approaches that relied on recurrent neural networks (RNNs) [3, 5, 23, 37-39]. This is evident on established benchmark datasets [17, 18, 27, 28, 36, 43]. For instance, He et al. [16] employed a fixed one-dimensional (1D) sinusoidal position encoding for recognizing horizontal, handwritten text. In [32], authors introduced a two-dimensional (2D) spatial PE for capturing the inherent spatial relationships between characters in irregular text, leading to improved performance. Alternatively, in [21], the authors introduced a 2D Spatial Pyramid Embedding (SPE) with adaptive amplitude for the transformer encoder, achieving state-of-the-art accuracy in scene text recognition across multiple benchmarks. This success is because of the model's capacity to learn positional information in horizontal and vertical directions. However, a limitation of these methods [21, 32] lies in their dependence on manually predefined PE frequencies. This inherent characteristic hinders their ability to handle inherent variability within text data [24].

Compared to recurrent architectures like RNNs and LSTMs, transformers exhibit reduced sensitivity to the order of input sequences. This advantage stems from the absence of inductive bias regarding positional information within the input set [24]. RNNs and LSTMs inherently encode position through their sequential processing nature.

Transformers leverage self-attention and feed-forward networks (FFNs) that operate in a permutation-equivariant manner. This means the model independently calculates each element's output in the input sequence, regardless of their order. While the 1D PE technique employed in the original transformer [42] effectively addresses the permutation equivariance issue for 1D sequences in natural language processing, it falls short in capturing the spatial information inherent in 2D image features extracted by convolutional neural networks (CNNs) [21].

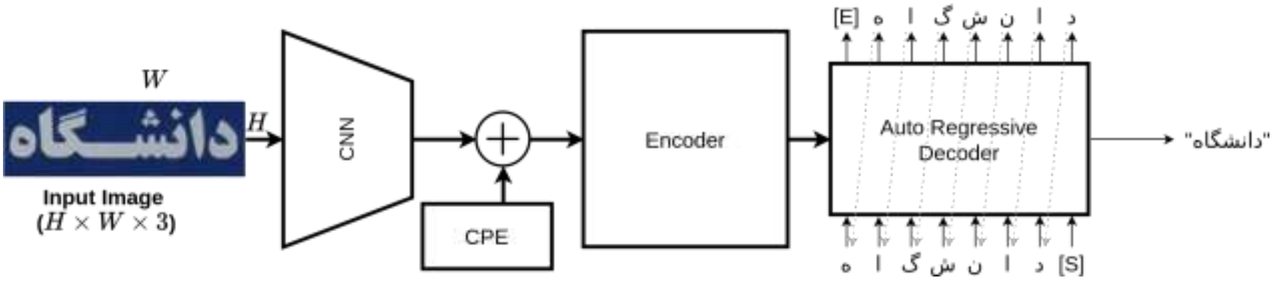


Figure 2. Transformer-Based Persian Scene Text Recognition with CPE, Res Net Backbone, and Modified FFN Sub-Blocks (vs. [42]).

The positional encoding (PE) scheme employed in transformers addresses the permutation equivariance issue by augmenting the input feature sequence with information regarding its order. This is crucial as the transformer architecture, unlike recurrent neural networks (RNNs), lacks inherent mechanisms like recurrence relations to capture sequential information within the data implicitly.

Positional encoding (PE) was first utilized in the original framework of the Transformer [42], which consists of predefined sinusoidal functions. The main drawback of this PE is its lack of learnable parameters, which restricts the maximum length of input sequences. To propose a solution to this issue, in [8], the authors suggested a learnable PE (data-driven), which learns the positions of input data during training.

This PE has two main drawbacks: (1) it is not inductive (i.e., in the testing time, it can not handle longer sequences seen in the training time), (2) it is not parameter efficient (i.e., it produces lots of trainable that later limit the generalization of the model). In [47], Shaw suggested a relative PE that reduces the number of trainable parameters, but it is not inductive. In a recent work by [24], the authors proposed a continuous dynamical positional encoding, which is learnable and takes advantage of parameter efficiency and the inductive properties of previously introduced positional encodings.

Prior studies have shown that adding position information inside each transformer block enhances performance [24, 42]. However, to ensure equitable comparison with established techniques that usually apply positional encoding (PE) solely in the first self-attention layer [21, 32], we limit our method to this initial block. Although learned PE provides comprehensive information integration [8], it lacks inherent bias. This issue arises because the requirement of a predetermined maximum sequence length during training [24], which may limit generalization to test sets of varying lengths.

3. Methodology

The overall architecture of the proposed model based on the standard transformer architecture introduced in [42] is shown in Figure 2. As shown, when fed to the proposed pipeline, the given input image goes through multiple modules to finally output the final string of words. The full details of these components are provided as follows: Generally, the transformer architecture consists of a sequence of N sub-blocks denoted as B_N , where n ranges from 1 to N.

Both encoder and decoder consist of three main sub-blocks: a position encoding *P*, a self-attention $A_N(\cdot)$, and a *FFN* layer $F_N(\cdot)$. For a given input set, $x = \{x_i\}_{i=1}^{t}$, these modules can be defined in the following manner [24]:

$$\begin{aligned} &\text{EN-DE}(x) = B_N \circ B_{N-1} \circ \cdots B_1(x), \\ &B_N(x) = F_N \circ A_N \circ [P + CNN](x) \end{aligned} \quad (1)$$

where *P* shows the positional encoding module, and CNN denotes the feature extractor.

3.1. CNN

To extract robust features capturing the 2D structure of the input image, we use a modified ResNet-31 architecture [12] as the feature backbone. The Res-Net feature extraction backbone is one of the most used deep-learning architectures for feature extraction, which captures the 2D structure of input word images. For more details, assume that $I \in R^{H \times W}$ is a given input word image, where W and H demonstrate the height and width. The CNN module extracts lower resolution features of $X \in R^{H' \times W' \times d}$, Here, d denotes the feature channel dimension, W' , and H' , represent the down-sampled height and width of 2-dimensional extracted features.

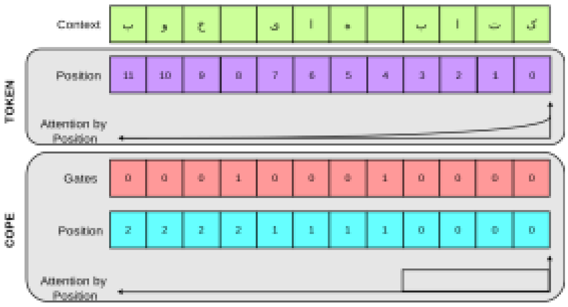


Figure 3. The proposed CPE and baseline SPE in attention of characters with spaces between them in a given word instance.

3.2. Proposed Positional Encoding

After extracting 2D features from the CNN, the resulting outputs are combined with positional encoding (PE) and then input into the Multi-Head Self-Attention (MHSA) subblock within the Encoder module in the Transformer architecture. The MHSA sub-block in the Transformer has a fascinating property known as permutation equivariance. This means that the model's output remains the same even if the elements' order within the input sequence is shuffled.

Positional encoding (PE) in Transformer architectures addresses the challenge posed by permutation equivariance. This encoding embeds positional information into the input features from CNN, enabling the model to differentiate between characters based on their relative order within the word string. Conventional position encoding methods, such as sinusoidal position encoding, assign a fixed numerical value to each position in the input. This approach can lead to challenges when the model is applied to sequences of different lengths, as the fixed encoding may not be appropriate.

To address the above problem, inspired by the recent work proposed in [10], we better utilize the Contextual Position Encoding (CPE) technique to capture the long and spaced characters in Persian scripts. The main advantage of CPE compared to previous generations of PE is that it learns to assign importance to different positions of characters based on surrounding characters, enabling the model to prioritize relevant parts of word instances and address issues with spaced characters. Figure 3 shows the proposed CPE that can be leveraged to the baseline Architecture. As illustrated in Figure 3, the CPE allows the model to identify and focus on the essential characters of a word instance with complex sequences.

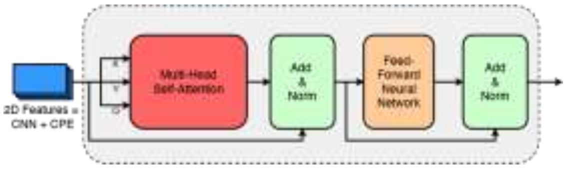


Figure 4. The encoder of the transformer-based pipeline.

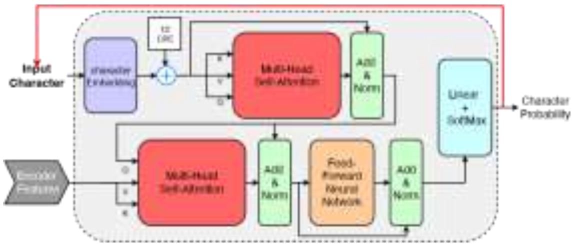


Figure 5. The decoder module of the transformer-based pipeline.

3.3. Encoder

The proposed architecture uses an encoder module for high-level feature extraction from the input image. As detailed in Figure 4, each sub-block within the encoder (Figure 2) is designed to capture specific aspects of the visual data. The encoder consists of three primary sub-blocks: Add-Norm, Multi-Head Self-Attention (MHSA), and Feed-Forward Network (FFN). A set of the inputs from adding 2D extracted features from CNN and the 2DCPE are fed into the MHSA to make the whole architecture attend more to different inputs. We employ a similar FFN sub-block as described in [32] to effectively handle the features produced by the multi-head self-attention mechanism of the encoder. The Add-Norm sub-block within the encoder module includes Layer Norm and residual connection, expressed as:

$$x = \text{LayerNorm}\left(x + F_s\left(x\right)\right) \tag{2}$$

where F_S shows the sub-layer module. The FFN in both the encoder and decoder comprises two stacked 1×1 convolutional layers with a ReLU activation layer followed by a residual connection. The output matrix $F_1(\cdot)$ has t rows. where the expression for the i^{th} row is given by:

$$F_1\left(x\right) = W_2\sigma\left(W_1 x_i + b_1\right) \tag{3}$$

where σ (·) denotes the activation function. b_1 and $W_{1,2}$ are biases and the weights of linear transforms.

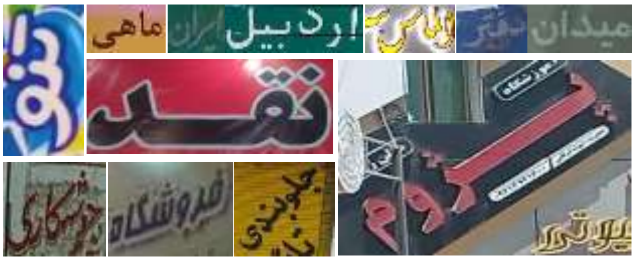


Figure 6. Sample real-world Persian cropped images of the collected dataset used for evaluation.

3.4. Decoder

Figure 5 shows the encoder part of the proposed model. The decoder utilizes these extracted feature maps to generate a sequence of characters. Similar to the architecture presented in [42], the decoder module incorporates MHSA and FFN layers. The only difference here is utilizing CPE as position encoding.

The decoder and encoder modules in a transformer function in a similar manner, retrieving information from a set of encoded representations. However, as shown in Figure 5, the decoder's inputs differ from the encoder's, incorporating an additional sub-block termed cross-attention. This sub-block is positioned after the self-attention layer within a single decoder block and followed by an addition and normalization step. This work uses an auto-regressive (AR) decoder to predict the output strings. The AR decoder takes input as a masked context sequence and extracts the visual features of the encoder sequence to generate the character sequence.

4. Experimental Results

This part presents a comparative analysis of the STR technique against established cutting-edge methods [3, 5, 23, 37-39], using the prepared Persian Script dataset.

4.1. Datasets

Two primary types of datasets are utilized to evaluate scene text recognition for Latin text. Regular-text datasets [17, 27, 45] contain mainly horizontal text, while irregular-text datasets [18, 28, 36, 43] include multi-oriented and curved text. Researchers often pre-train their models on synthetic images using the SynthText (ST) [11] and MJSynth (MJ) [14] datasets for higher accuracy.

Unlike Latin scene text recognition, finding public benchmark datasets for Persian scripts is difficult. PESTD is the primary resource, but it is limited to traffic sign *detection* and has restricted access.

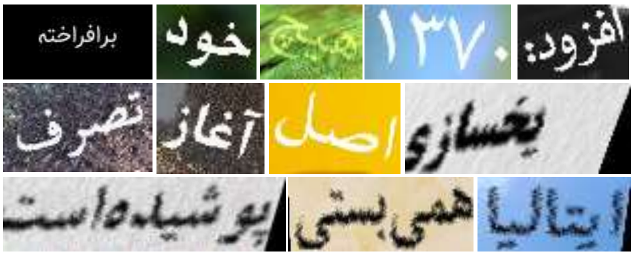


Figure 7. Sample synthetic Persian word images of the collected dataset used for retraining the models.

In contrast, ITDR-Synth offers 6,100 detection images and 40,220 recognition images for Persian text analysis [2]. However, ITDR-Synth consists only of synthetic images, unlike the challenging cases appearing in the wild.

Here, we suggest two datasets to tackle the aforementioned issue: a comprehensive real-world scene text recognition test set and a synthetic dataset specifically designed for training. The real-world dataset contains 3529 test images from diverse indoor and outdoor environments, capturing challenging scenarios. To visually represent the dataset, examples of real-world scene text images are displayed in Figure 6.

Furthermore, we provide a synthetic dataset comprising approximately 200,000 Persian word images. This dataset is particularly useful as it addresses the need for Persian words in existing datasets, making it ideal for training in this context. Figure 7 shows sample images from the created dataset.

4.2. Evaluation Metric

We assess the recognition performance of our scene text recognition system using two established metrics commonly employed in this domain: Word Recognition Accuracy (WRA) and Normalized Edit Distance (NED). WRA is particularly relevant due to its direct applicability in real-world scenarios, as opposed to character recognition accuracy, and has been widely used to evaluate text recognition schemes [3, 23, 37, 38]. WRA assesses the accuracy of scene text recognition schemes according to a collection of word images, and it is defined as follows:

$$WRA = \frac{\#\ of\ \ accurate\ recognized\ words}{All\ the\ words\ in\ the\ datasets} \tag{4}$$

WRA means that we only count words that are completely correct. The NED metric is defined as follows [40]:

$$Norm = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathrm{D}\left(w_i, \hat{w}_i\right) / \, max\left(w_i, \hat{w}_i\right) \tag{5}$$

Table 2. Experimental Results of the select scene text cognition models [3, 5, 23, 29, 37] on the proposed dataset. The best method in each trained dataset category is highlighted. The WRA and NED denote the word recognition accuracy and the normalized edit distance.

Model	WRA	NED
CRNN [37]	53.04	0.78
ROSETTA [5]	65.70	0.85
STARNET [23]	68.74	0.86
CLOVA [3]	69.24	0.87
UTRNet [29]	66.93	0.86
Baseline	66.11	0.86
Proposed	**73.94**	**0.91**

where Levenshtein Distance is shown by $D(:)$ [22]. w_i and w_i' are ground truths corresponding to the text regions and predicted word strings, respectively.

4.3. Implementation Details

Every model in this work is tested and trained on a system equipped with an NVIDIA RTX-3090 GPU. To ensure a fair and controlled comparison, we select object detection models trained on similar datasets. The models are trained on the prepared synthetic dataset. Only Persian characters are used during evaluation; The special letters and English characters are ignored. For the backbone feature extractor, we use a ResNet-31 [12]. The final model is selected according to the highest recognition accuracy achieved on these datasets. All recognition models, except UTRNeT [29], are trained for 200,000 iterations, whereas the UTRNeT model undergoes training for 50 epochs. The baseline model is an encoder-decoder-based transformer architecture with 1D SPE (See Table 1) with fixed frequency.

The proposed model is only different in the PE module than the baseline. We follow the setting of [3, 33] for adjusting the hyperparameters, optimizers, and augmentations during training and inference. We use 35 classes of Persian characters during training. These characters are as follows:

ا، آ، ب، پ، ت، ث، ج، چ، ح، خ، د ، ذ، ر، ز، ژ، س، ش، ص،
ض، ط، ظ، ع، غ، ف، ق، ک، گ، ل، م، ن، و، ه، ء، ی، ئ

4.4. Quantitative Results

Table 4 compares the effectiveness of the suggested approach (WRA) with several SOTA scene text recognition methods [3, 5, 21, 23, 37–39].

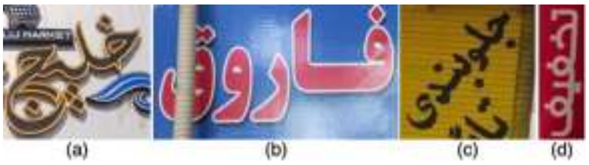


Figure 8. Sample qualitative results on some challenging images of the prepared real-world Persian dataset that are correctly recognized by the proposed model, where Other selected models [3, 5, 23, 29, 37] fail.

These models are well-known deep learning methods, including RNN-based methods, alongside newer transformer-based ones. The proposed method achieved better WRA (73.94) and ED (0.91) performances than the model considered in the evaluations. Furthermore, our suggested approach surpassed the baseline by a significant margin.

4.5. Qualitative Results

We tested the models in Table 4 to demonstrate their performance on real-world images. To that effect, we evaluated the qualitative results on various cropped word images from the prepared Persian dataset, as presented in Figure 8. The proposed model accurately recognized word images with different challenges, including irregular text, vertically oriented, partially occluded text, and complex font styles.

4.6. Ablation Study

We conducted numerous ablation experiments to analyze the influence of the synthetic dataset we collected and the proposed CPE in the architecture. The ablation study results are presented in Tables 3 and 4, comparing the effect of the created synthetic images and the utilized positional encoding, respectively.

First, as presented in Table 3, we trained the models on synthetic images in the ITDR-Synth dataset [2], which led to poor performance. However, when we utilized our proposed synthetic dataset, the models' WRA performance improved significantly. This confirms that training the models on the proposed synthetic dataset can significantly enhance recognition performance compared to using only the publicly available synthetic Persian dataset [2].

We also experimented with different positional encoding schemes to compare our proposed CPE. As shown in Table 4, applying CPE significantly improved the output results compared to other positional encoding techniques applied to the baseline methods by a large margin showing its effectiveness in Persian scene text recognition.

Table 3. Experimental Results of the select scene text recognition models [3, 5, 23, 29, 37] on the ITDR-Synth [2] and our proposed dataset. The WRA and NED denote the word recognition accuracy and the normalized edit distance.

Model	Dataset	WRA	NED
CRNN [37]	ITDR-Synth	22.28	0.45
ROSETTA [5]	ITDR-Synth	19.26	0.44
STARNET [23]	ITDR-Synth	27.17	0.48
CLOVA [3]	ITDR-Synth	27.68	0.50
UTRNet [29]	ITDR-Synth	24.72	0.49
Proposed	ITDR-Synth	31.52	0.52
Proposed	Our-Synth	**73.94**	**0.91**

Table 4. The effect of utilizing different positional encoding modules in the baseline architecture. the definitions of all abbreviations are presented in Table 1.

Model	Positional Encoding	WRA
Baseline	1DSPE	66.11
Baseline	2DPE	69.53
Baseline	SATRN	69.88
Baseline	2LSPE	70.45
Baseline	**CPE**	**73.94**

4.7. Limitation and Future Work

While the proposed model achieved SOTA performance compared to the selected models, there are still many cases in which the model has limitations in correctly capturing all the characters in some challenging samples. Figure 9 illustrates some of these failure cases. As shown, the model mostly misses one or two characters. This failure is due to not seeing these font styles during training. One solution may be training the model with real-world images of these images.

Recognizing text in scenes is challenging because annotating real-world images from scratch is time-consuming and costly. Recent AI advancements like DALLE [35], ChatGPT [1], and Gemini [41] offer potential solutions. This automation with cutting-edge models like [20] also addresses the bottleneck. Despite challenges in detecting and recognizing Latin text in uncontrolled environments, leveraging AI methodologies can effectively enhance model performance.

5. Conclusion

In this study, we have presented a new method for recognizing scene text in Persian. We have addressed the inherent challenges of the Persian script by extending a vanilla transformer architecture by utilizing a 2D version of Contextual Position Encoding (CPE).

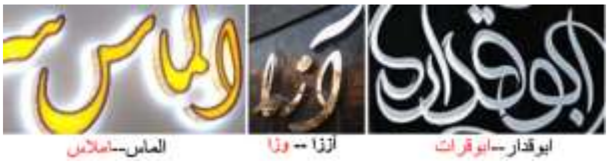


Figure 9. Failure example images of the proposed model. The black and red colors denote the ground truth and the output of the model.

CPE enables the model to better capture the challenging Persian text instances in wild images. Furthermore, the proposed model with CPE inherently mitigates the challenges posed by abundant spacing variations within Persian scene text instances. This architecture was evaluated on a comprehensive Persian scene text recognition dataset and compared to several well-known scene text recognition models. The findings from the experiments have demonstrated the efficacy of the suggested model on the prepared dataset with word recognition accuracy in comparison with the evaluated cutting-edge techniques. We believe the prepared dataset and proposed model are capable of serving as a benchmark for future research in Persian scene text recognition.

References

[1] J. Achiam et al., "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[2] F. Alimorad et al., "Synthesizing an Image Dataset for Text Detection and Recognition in Images," *Journal of Information and Communication Technology*, vol. 53, no. 53, pp. 78, 2023.

[3] J. Baek et al., "What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 4714-4722, doi: 10.1109/ICCV.2019.00481.

[4] D. Bautista and R. Atienza, "Scene Text Recognition with Permuted Autoregressive Sequence Models," *ECCV*, *Lecture Notes in Computer Science*, vol. 13688, Springer, Cham, doi: 10.1007/978-3-031-19815-1_11.

[5] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large Scale System for Text Detection and Recognition in Images," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*, Association for Computing Machinery, New York, NY, USA, pp. 71–79, doi: 10.1145/3219819.3219861.

[6] R. Buoy et al., "PARSTR: Partially Autoregressive Scene Text Recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 303-316, 2024, doi: 10.1007/s10032-024-00470-1.

[7] X. Chen et al., "Text Recognition in the Wild: A Survey," *ACM Comput. Surv.*, vol. 54, no. 2, Article 42, March 2022, doi: 10.1145/3440756.

[8] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Minneapolis, Minnesota, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[9] A. Fateh et al., "Persian Printed Text Line Detection Based on Font Size," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 2393–2418, 2023, doi: 10.1007/s11042-022-13243-x.

[10] O. Golovneva et al., "Contextual Position Encoding: Learning to Count What's Important," *13th International Conference on Learning Representations*, 2024.

[11] A. Gupta et al., "Synthetic Data for Text Localisation in Natural Images," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2315-2324, doi: 10.1109/CVPR.2016.254.

[12] K. He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[14] M. Jaderberg et al., "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition," arXiv preprint arXiv:1406.2227, 2014.

[15] M. Jaderberg et al., "Spatial Transformer Networks," *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15)*, MIT Press, Cambridge, MA, USA, pp. 2017–2025, 2015.

[16] L. Kang et al., "Pay Attention to What You Read: Nonrecurrent Handwritten Text-Line Recognition," *Pattern Recognition*, vol. 129, 2022, doi: 10.1016/j.patcog.2022.108766.

[17] D. Karatzas et al., "ICDAR 2013 Robust Reading Competition," *2013 12th International Conference on Document Analysis and Recognition*, Washington, DC, USA, pp. 1484-1493, doi: 10.1109/ICDAR.2013.221.

[18] D. Karatzas et al., "ICDAR 2015 Competition on Robust Reading," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, pp. 1156-1160, doi: 10.1109/ICDAR.2015.7333942.

[19] S. Kheirinejad et al., "Persian Text Based Traffic Sign Detection with Convolutional Neural Network: A New Dataset," *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran, pp. 060-064, doi: 10.1109/ICCKE50421.2020.9303646.

[20] A. Kirillov et al., "Segment Anything," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 3992-4003, doi: 10.1109/ICCV51070.2023.00371.

[21] J. Lee et al., "On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, pp. 2326-2335, doi: 10.1109/CVPRW50498.2020.00281.

[22] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Physics Doklady*, pp. 707–710, 1966.

[23] W. Liu et al., "STAR-Net: A Spatial Attention Residue Network for Scene Text Recognition," *Proc. Brit. Mach. Vision Conf. (BMVC)*, pp. 43.1–43.13, BMVA Press, 2016, available: https://api.semanticscholar.org/CorpusID:22482128.

[24] X. Liu et al., "Learning to Encode Position for Transformer with Continuous Dynamical Model," *Proceedings of the 37th International Conference on Machine Learning*, pp. 6327–6335, 2020.

[25] S. Long et al., "Scene Text Detection and Recognition: The Deep Learning Era," *International Journal of Computer Vision*, vol. 129, pp. 161–184, 2021, doi: 10.1007/s11263-020-01369-0.

[26] Z. Raisi and J. Zelek, "Visual Place Recognition from end-to-end semantic scene text features, *Frontiers in Robotics and AI*, Vol. 11, Article 1424883, 2024, doi: 10.3389/frobt.2024.1424883.

[27] A. Mishra et al., "Scene Text Recognition Using Higher Order Language Priors," *BMVC - British Machine Vision Conference*, Sep 2012, Surrey, United Kingdom, doi: 10.5244/C.26.127.

[28] T. Q. Phan et al., "Recognizing Text with Perspective Distortion in Natural Scenes," *2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, pp. 569-576, doi: 10.1109/ICCV.2013.76.

[29] A. Rahman et al., "UTRNet: High-Resolution Urdu Text Recognition in Printed Documents," *International Conference on Document Analysis and Recognition*, pp. 305–324, Springer, 2023, *Lecture Notes in Computer Science*, vol. 14191, doi: 10.1007/978-3-031-41734-4_19.

[30] M. Rahmati et al., "Printed Persian OCR System Using Deep Learning," *IET Image Processing*, vol. 14, no. 15, pp. 3920–3931, 2020, doi: 10.1049/iet-ipr.2019.0728.

[31] Z. Raisi and J. Zelek, "Occluded Text Detection and Recognition in the Wild," *2022 19th Conference on Robots and Vision (CRV)*, Toronto, ON, Canada, 2022, pp. 140-150, doi: 10.1109/CRV55824.2022.00026.

[32] Z. Raisi, M. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "2D Positional Embedding-Based Transformer for Scene Text Recognition," *Journal of Computational*

Vision and Imaging Systems, vol. 6, no. 1, pp. 1–4, 2021, doi: 10.15353/jcvis.v6i1.3533.

[33] Z. Raisi et al., "2LSPE: 2D Learnable Sinusoidal Positional Encoding Using Transformer for Scene Text Recognition," *2021 18th Conference on Robots and Vision (CRV)*, Burnaby, BC, Canada, 2021, pp. 119-126, doi: 10.1109/CRV52889.2021.00024.

[34] Z. Raisi, "Text Detection and Recognition in the Wild," PhD thesis, 2022, available: http://hdl.handle.net/10012/18453.

[35] A. Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv preprint arXiv:2204.06125, 2022.

[36] A. Risnumawan et al., "A Robust Arbitrary Text Detection System for Natural Scene Images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, 2014, doi: 10.1016/j.eswa.2014.07.008.

[37] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298-2304, 1 Nov. 2017, doi: 10.1109/TPAMI.2016.2646371.

[38] B. Shi et al., "Robust Scene Text Recognition with Automatic Rectification," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4168-4176, doi: 10.1109/CVPR.2016.452.

[39] B. Shi et al., "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035-2048, 1 Sept. 2019, doi: 10.1109/TPAMI.2018.2848939.

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

[41] R. Anil et al., "Gemini: A Family of Highly Capable Multimodal Models," arXiv preprint arXiv:2312.11805, 2023.

[42] A. Vaswani et al., "Attention is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010, 2017.

[43] A. Veit et al., "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," arXiv preprint arXiv:1601.07140, 2016.

[44] B.Wang et al., "On Position Embeddings in BERT," *International Conference on Learning Representations*, Austria, 2021.

[45] K. Wang and S. Belongie, "Word Spotting in the Wild," *ECCV 2010*, *Lecture Notes in Computer Science*, vol. 6311, Springer, Berlin, Heidelberg, 2010, doi: 10.1007/978-3-642-15549-9_43.

[46] F. Zhan and S. Lu, "ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 2054-2063, doi: 10.1109/CVPR.2019.00216.

[47] H. Zhang et al., "Self-Attention Generative Adversarial Networks," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 7354-7363, 09-15 Jun 2019, PMLR.

[48] S. Zhao et al., "CLIP4STR: A Simple Baseline for Scene Text Recognition with Pre-trained Vision-Language Model," arXiv preprint arXiv:2305.14014, 2023.

[49] F. Ariai et al., "Enhancing Aspect-based Sentiment Analysis with ParsBERT in Persian Language," *Journal of AI and Data Mining*, vol. 12, no. 1, pp. 1–14, 2024, doi: 10.22044/jadm.2023.13666.2482.

یک رویکرد مبتنی بر ترانسفورمر با استفاده از رمزگذاری موقعیت متنی برای شناسایی مقاوم متن فارسی در محیط‌های طبیعی چالش برانگیز

زبیر رئیسی* و ولی محمد نظرزهی

گروه مهندسی الکترونیک و مخابرات دریایی، دانشگاه دریانوردی و علوم دریایی چابهار، چابهار، ایران.

ارسال ۲۰۲۴/۰۷/۱۵؛ بازنگری ۲۰۲۴/۱۰/۳۰؛ پذیرش ۲۰۲۴/۱۲/۰۹

چکیده:

زبان فارسی به دلیل ویژگی‌های خاص نگارشی خود چالش‌های منحصربه‌فردی را برای شناسایی متن از محیط‌های طبیعی ارائه می‌کند. با وجود پیشرفت‌های هوش مصنوعی و یادگیری عمیق، شناسایی خط‌های غیر لاتین مانند فارسی، همچنان با چالش‌های فراوانی مواجه است. در این مطالعه، معماری اولیه یادگیری عمیق ترانسفورمر را برای تشخیص نوشته‌های فارسی با شکل‌ها و جهت‌های دلخواه تقویت کرده‌ایم. که برای این منظور، کدگذاری موقعیت متنی (CPE) را به معماری پایه ترانسفورمر اضافه کرده‌ایم تا شناسایی نوشته‌های فارسی در تصاویر طبیعی، به‌ویژه برای کاراکترهای جهت‌دار و فاصله‌دار، بهبود یابد. CPE از اطلاعات موقعیتی برای تولید جفت داده‌های متضاد استفاده می‌کند که به شناسایی بهتر کاراکترهای فارسی نوشته شده در جهت‌های مختلف کمک می‌کند. علاوه بر این، چندین مدل پیشرفته یادگیری عمیق را با استفاده از مجموعه داده آماده شده چالشی خود، برای شناسایی متن فارسی از مناظر طبیعی ارزیابی کرده و یک معماری مبتنی بر ترانسفورمر توسعه داده‌ایم تا دقت شناسایی را ارتقاء دهیم. معماری تشخیص متن صحنه پیشنهادی ما در مقایسه با روش‌های موجود در مجموعه داده آماده شده متن فارسی در تصاویر گرفته شده واقعی، به دقت تشخیص کلمه بالاتری دست می‌یابد.

کلمات کلیدی: شناسایی متن از مناظر طبیعی، نوشته‌های فارسی، رمزگذاری موقعیت متنی، ترانسفورمرها، یادگیری عمیق.