scientific reports

Check for updates

OPEN

Linguistic-visual based multimodal Yi character recognition

Haipeng Sun[1,2,3], Xueyan Ding[3]✉, Zimeng Li[3], Jian Sun[3], Hua Yu[4] & Jianxin Zhang[3]✉

The recognition of Yi characters is challenged by considerable variability in their morphological structures and complex semantic relationships, leading to decreased recognition accuracy. This paper presents a multimodal Yi character recognition method comprehensively incorporating linguistic and visual features. The visual transformer, integrated with deformable convolution, effectively captures key features during the visual modeling phase. It effectively adapts to variations in Yi character images, improving recognition accuracy, particularly for images with deformations and complex backgrounds. In the linguistic modeling phase, a Pyramid Pooling Transformer incorporates semantic contextual information across multiple scales, enhancing feature representation and capturing the detailed linguistic structure. Finally, a fusion strategy utilizing the cross-attention mechanism is employed to refine the relationships between feature regions and combine features from different modalities, thereby achieving high-precision character recognition. Experimental results demonstrate that the proposed method achieves a recognition accuracy of 99.5%, surpassing baseline methods by 3.4%, thereby validating its effectiveness.

Keywords Deep learning, Character recognition, Transformer, Linguistic-visual model

The Yi ethnic group is among the oldest in China, possessing a rich literary heritage encompassing diverse fields, including medicine, astronomy, and religion. The study of Yi character recognition is of significant importance. Unlike the extensive research on Chinese and English text recognition[1,2], research on Yi character recognition remains relatively scarce. The Yi language boasts a rich and diverse vocabulary, comprising primarily monosyllabic and polysyllabic words, and consists of 1,165 distinct characters. The variability in character forms poses significant challenges for Yi character recognition.

The Yi character recognition task originates from a profound understanding of the challenges inherent in text recognition and the urgent demand for more advanced algorithms. Yi characters exhibit highly complex morphological structures, which hinder the ability of traditional manual feature definition methods to capture their deep semantic information adequately[3]. Consequently, advanced technologies, such as deep learning, are crucial for enabling automatic feature extraction. Additionally, single-character-based recognition methods frequently neglect the intrinsic relationships between characters, leading to a loss of contextual information and potential ambiguities. This emphasizes the necessity of considering character relationships throughout the recognition process. Furthermore, the complex structure and variable semantics of Yi characters pose significant challenges for character segmentation and localization, requiring meticulous preservation of character integrity to prevent forced segmentation[4–6].

In recent years, the outstanding performance of Transformer models in computer vision and natural language processing has led to novel insights and inspired innovative approaches[7,8]. Researchers recognize that integrating Transformer models into text recognition tasks has considerable potential to enhance performance significantly. Despite advancements in Yi character recognition technology, challenges persist, such as the complexity of character shapes, the high similarity between characters, and significant variation in inter-character spacing. Relying exclusively on the linguistic module may overlook crucial visual features, reducing recognition accuracy. Therefore, adopting an approach that integrates visual and linguistic information is essential for effectively addressing these challenges and enhancing the performance of Yi character recognition. The contributions of this paper are summarized as follows:

[1]Key Laboratory of Ethnic Language Intelligent Analysis and Security Management of MOE, Minzu University of China, Beijing 100081, China. [2]School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, China. [3]School of Computer Science and Engineering, Dalian Minzu University, Dalian 116600, China. [4]Yi Language Research Room, China Ethnic Languages Translation Centre, Beijing 100080, China. ✉email: dingxueyan@dlnu.edu.cn; jxzhang0411@163.com

(1) This paper proposes a multimodal method for Yi character recognition based on linguistic and visual features. This method effectively addresses the challenges in Yi character recognition by integrating visual and linguistic models to extract visual features and contextual information.
(2) To improve recognition accuracy, Deformable DETR[9] is employed to enhance multi-scale visual feature extraction, while the Pyramid Pooling Transformer[10] captures richer linguistic contextual features. Subsequently, a Cross Attention[11]-based fusion strategy is used to integrate multimodal information effectively.
(3) The experimental results confirm that the proposed method outperforms existing text recognition approaches in terms of accuracy, highlighting its robustness and effectiveness in addressing recognition challenges.

Related works

Recent advancements in deep learning techniques have significantly enhanced text recognition tasks across diverse application contexts[12,13]. Text recognition algorithms that leverage deep learning are generally classified into two categories: those based on connectionist temporal classification and those employing attention mechanisms.

Text recognition based on connectionist temporal classification

Graves et al.[14] proposed a text recognition algorithm based on Connectionist Temporal Classification (CTC) for training recurrent neural networks, thus enabling the direct annotation of unsegmented sequences.CTC algorithms have performed outstandingly in various domains, including speech and handwriting recognition. Building on the success of CTC algorithms in speech recognition, researchers have made significant strides in integrating these methods into text recognition to improve decoding performance. For instance, Shi et al.[15] proposed that recognizing natural scene text constitutes a sequence recognition task. They introduced an end-to-end trainable Convolutional Recurrent Neural Network (CRNN), eliminating the need for complex character segmentation. Instead, their approach leverages the strengths of deep Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), thereby significantly enhancing the effectiveness of natural scene text recognition algorithms. Subsequently, several text recognition algorithms[16–18] that utilize CTC decoding have shown enhanced recognition performance. However, some researchers have argued that CTC algorithms tend to produce overly precise prediction distributions, which could lead to overfitting.

To address these challenges, Liu et al.[19] incorporated a regularization term within maximum entropy and conditional probability, improving generalization and guiding the CTC algorithm to explore feasible and effective pathways. Feng et al.[20] integrated the CTC algorithm with a focal loss function, effectively addressing the challenge of text recognition affected by significant disparities among sample categories. Hu et al.[21] employed graph CNNs to enhance the precision and robustness of a text recognition algorithm based on CTC decoding. Although the CTC algorithm has shown promise and contributed significantly to advancements in text recognition, it still faces several limitations: (1) The theoretical foundation of the CTC algorithm is intricate, and its direct application in decoding incurs substantial computational overhead; (2) The CTC algorithm tends to generate overly precise and excessively confident prediction distributions[22], leading to degraded decoding performance, particularly in the case of character duplication; (3) Due to inherent structural and implementation limitations, the application of the CTC algorithm to 2D prediction tasks, such as irregular scene text recognition, presents significant challenges. Wan et al.[23] extended the foundational CTC algorithm by incorporating additional dimensions along the vertical axis, addressing its limitations in handling irregular text recognition tasks. Although this approach somewhat improves recognition performance, it does not fully resolve the core challenges associated with applying CTC algorithms to 2D prediction tasks. As a result, text recognition algorithms based on the CTC methodology have limitations in their applicability to diverse scenarios. Exploring the application of CTC algorithms to 2D prediction challenges presents a promising avenue for future research in this field.

Text recognition based on attention mechanism

Text recognition algorithms leveraging the attention mechanism have gained significant prominence in deep learning, particularly within Sequence-to-Sequence (Seq2Seq) frameworks. The core concept entails allowing the model to dynamically allocate attention to various segments of the input sequence during output generation, thereby enhancing its ability to focus on critical information during inference. To address challenges such as reduced recognition accuracy caused by text box irregularity, perspective distortion, and text deformation, Shi et al.[24] proposed the Robust Text Recognizer with Automatic Rectification (RARE). The model integrates a Spatial Transformer Network (STN) and a Sequence Recognition Network (SRN). During the training phase, image rectification is performed using the predictive Thin Plate Spline (TPS), ensuring that the images are suitable for SRN processing. To address these challenges, Wang et al.[25] introduced the Decoupled Attention Network (DAN), an end-to-end text recognition model that utilizes the attention mechanism. DAN aims to decouple the historical decoding outputs from the alignment process. The model uses a feature encoder to extract visual attributes from input images, applies a convolutional rectification phase to align visual features, and implements a decoupled text decoding phase that leverages feature maps and attention maps for prediction. The Transformer-based text recognition algorithm is inspired by its counterpart models in natural language processing and sequence modeling. Central to its design is the self-attention mechanism, which enables the model to assign varying degrees of importance to different elements within the input sequence, thereby enhancing contextual understanding. Yu et al.[26] enhanced the Semantic Reasoning Network (SRN) by incorporating the Global Semantic Reasoning Module (GSRM). This module operates through parallel pathways to capture comprehensive global contextual information, enhancing the overall model's effectiveness. To overcome limitations such as the slow training speeds of RNNs and the complexity of convolutional layers, Sheng et al.[27] introduced the No-

Recurrence Sequence-to-Sequence Text Recognizer (NRTR). This approach exclusively utilizes self-attention mechanisms in both the encoder and decoder stages, improving trainability through enhanced parallelization and reduced complexity. Given the challenges posed by complex backgrounds and noisy environments in scene text recognition, Atienza et al.[38] introduced the ViTSTR model, which leverages the Vision Transformer (ViT) architecture to achieve high accuracy while maintaining computational efficiency. The model processes images by dividing them into patches and utilizing transformer layers to capture global context, thereby enhancing its effectiveness for text recognition in challenging environments. To address the challenge of improving both recognition accuracy and efficiency in text recognition tasks, Jiang et al.[39] propose the Reciprocal Feature Learning (RFL) model. This model incorporates character counting as an auxiliary task, aiming to enhance the performance of the recognition process. Despite the advantages of linguistic expertise in scene text recognition, effectively integrating linguistic information into end-to-end deep networks remains a significant technical challenge. Fang et al.[28] proposed the Autonomous Bidirectional Iterative Network (ABINet) to address this challenge. It introduces an autonomous gradient-blocking mechanism between the visual and linguistic models to enhance linguistic modeling, a Bidirectional Cloze Network (BCN) for comprehensive linguistic modeling, and an iterative correction method to mitigate input noise. To address the challenges of improving both accuracy and efficiency in scene text recognition, Du et al.[40] proposed the SVTR model, simplifying the conventional hybrid architecture by removing sequential modeling. SVTR decomposes text images into small patches, referred to as character components, and processes them through hierarchical stages, which include mixing and merging at the component level. This approach enables the model to effectively capture intra-character and inter-character patterns, facilitating character recognition through a straightforward linear prediction.

In Yi character recognition, CNNs have been widely explored for their application in optical character recognition. However, two significant challenges remain: the manual annotation of training datasets, which is both time-consuming and labor-intensive, and the reliance on empirical experience for CNN architecture design and parameter tuning due to the absence of established theoretical frameworks. To address these challenges, Jia et al.[29] applied entropy theory to enhance a density-based clustering algorithm. Additionally, Jiejue[30] developed a deep learning-based Yi character recognition method. However, it suffers from low accuracy and limited robustness compared to traditional Yi character recognition algorithms, such as structural pattern recognition and artificial neural networks. Despite these efforts, challenges persist in improving both the efficiency and accuracy of Yi character recognition, highlighting the need for more advanced methods and the development of robust theoretical frameworks.

Methodology

The proposed method comprises three key components: a visual model, a linguistic model, and a fusion model. The visual model employs a CNN to extract features from Yi character images. In contrast, the linguistic model utilizes a transformer architecture to process the Yi character sequence and capture semantic relationships. The fusion model employs a cross-attention mechanism to integrate these modalities, aligning visual features with the linguistic context derived from the linguistic model. This approach integrates the spatial characteristics of the images with the semantic understanding of the text, thereby improving recognition performance. The architecture is shown in Fig. 1, and the workflow is described as follows:

- Firstly, the visual model employs the ResNet-45 network as the backbone to extract features from Yi character images. Composed of convolutional layers and residual connections, ResNet-45 effectively learns hierarchical visual representations. These features are subsequently passed to the Deformable DETR module, which addresses spatial variations in Yi characters, including size, position, and orientation. The visual model generates feature maps that depict the spatial distribution of the characters, which are utilized in subsequent stages for further processing.
- Secondly, the visual prediction results generated by the visual module are passed into the linguistic model. The linguistic model utilizes a Pyramid Pooling Transformer to process the Yi character sequence, capturing local and global contextual information by aggregating multi-scale features. The Pyramid Pooling mechanism extracts multi-scale features to capture diverse spatial patterns, while the transformer's self-attention mechanism models inter-part relationships, capturing the hierarchical structure of Yi characters.
- Finally, the fusion model utilizes a cross-attention mechanism to integrate feature information from the visual and linguistic models. Cross-attention integrates the stroke patterns and character shape features extracted by the visual model with the semantic context from the linguistic model, enabling more precise alignment and improving the accuracy of Yi character recognition. The mechanism computes attention scores to match the relevant visual and linguistic features, ensuring that spatial and semantic dimensions are thoroughly considered. This integration enables the model to effectively combine complementary visual and linguistic information, thereby improving the accuracy of Yi character recognition by capturing the intricate relationship between visual details and semantic context.

Visual model

Figure 2 illustrates the overall structure of the proposed visual model, consisting of the ResNet-45 backbone, Transformer encoder and decoder, a linear layer, and a softmax layer. ResNet-45 is adapted by modifying its depth and feature extraction capabilities to capture better the intricate, high-frequency features specific to Yi characters. To overcome the limitations of conventional Transformer attention, a deformable transformer structure is integrated into the encoder, enabling adaptive focus on spatial distortions and variability in Yi characters. Multi-scale deformable convolutions[31] are introduced to handle varying sizes and distortions, ensuring accurate recognition across both small and large characters. This deformable convolution approach replaces the standard Transformer in DETR with a multi-scale deformable attention module, which adapts to

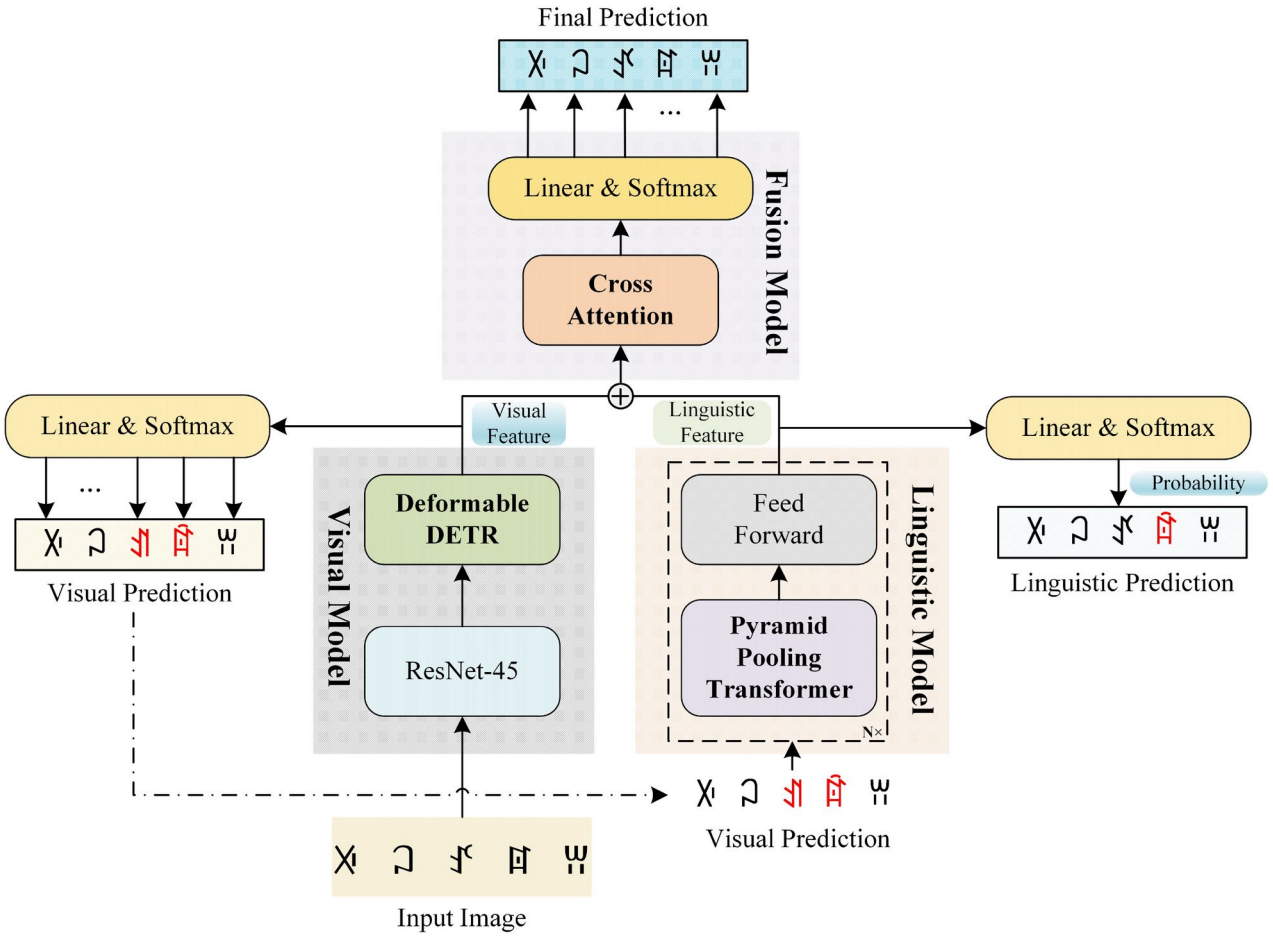

Figure 1. The overall architecture of the proposed method (the input image is processed through the visual module, where ResNet-45 extracts features related to the structure and strokes of Yi characters, while Deformable DETR captures spatial relationships and deformations, effectively addressing the distinctive characteristics of Yi characters. These features are passed through a Linear layer and a softmax layer to generate the visual prediction map, which is then input into the linguistic model. The Pyramid Pooling Transformer reduces sequence length and enhances feature representation. The linguistic features are then processed through feed-forward layers to generate the linguistic prediction map. Although the linguistic prediction map does not contribute to the final output, it is crucial for training the linguistic model, improving its ability to extract accurate features for Yi character recognition. Finally, Cross Attention aligns the visual and linguistic features, and linear and softmax operations generate the final prediction result).

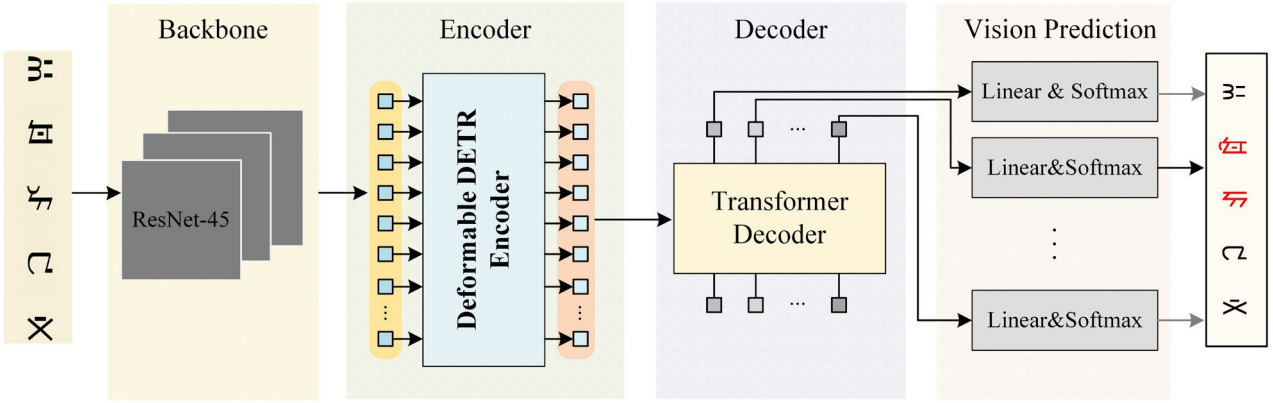

Figure 2. Overall structure of the visual model.

the varying scales and shapes of Yi characters. This approach reduces computational complexity and improves efficiency by eliminating the need for separate multi-scale feature fusion. Furthermore, the deformable attention module aggregates attention outputs across feature maps at multiple scales, enhancing detection performance by prioritizing relevant spatial patterns, such as stroke direction and character orientation.

The distinctive feature of the deformable attention Transformer lies in its focus on a designated set of key sampling locations, confined to the perimeter of reference points and independent of the dimensionality of the domain feature mapping. Additionally, assigning a fixed number of key points to each query effectively mitigates challenges related to convergence speed and low feature space resolution. The architecture of Deformable DETR is depicted in Fig. 3.

Given an input feature map $x \in \mathbb{R}^{C\times H\times W}$, let q index a query element with content feature z_q and a 2-d reference point p_q. The deformable attention feature is then calculated as follows:

$$DeformAttn(z_q, p_q, x) = \sum_{m=1}^{M} W_m \times \varphi(z) \tag{1}$$

where m signifies the index of attention heads, and $\varphi(z) = \sum_{k=1}^{K} A_{mqk} \cdot W'_m x(p_q + \Delta p_{mqk})$ in which k indexes the sampled keys, and K is the total sampled key number $(K \ll HW)$. Δp_{mqk} and A_{mqk} denote the sampling offset and attention weight of the k^{th} sampling point in the m^{th} attention head, respectively. The scalar attention weight A_{mqk}lies in the range [0, 1], normalized by $\sum_{k=1}^{K} A_{mqk} = 1$. $\Delta p_{mqk} \in \mathbb{R}^2$ are of 2-d real numbers with unconstrained range. As $p_q + \Delta p_{mqk}$ is fractional, bilinear interpolation is applied as in Dai et al. in computing $x(p_q + \Delta p_{mqk})$. Both Δp_{mqk} and ΔA_{mqk} are obtained via linear projection over the query feature z_q. In an implementation, the query feature z_q is fed to a linear projection operator of $3MK$ channels, where the first $2MK$ channels encode the sampling offsets Δp_{mqk} and the remaining MK channels are fed to a softmax operator to obtain the attention weights A_{mqk}.

Most modern object detection frameworks benefit from multi-scale feature maps[32]. Our proposed deformable attention module can be naturally extended for multi-scale feature maps. Let $\{x^l\}_{l=1}^{L}$ be the input multi-scale feature maps, where $x^l \in \mathbb{R}^{C\times H_l\times W_l}$. Let $\widehat{p}_q \in [0, 1]^2$ be the normalized coordinates of the reference point for each query element q. then the multi-scale deformable attention module is applied as:

$$MSDeformAttn(z_q, \widehat{p}_q, \{x^l\}_{l=1}^{L}) = \sum_{m=1}^{M} W_m \times \Gamma(z) \tag{2}$$

where m indexes the attention head, and $\Gamma(z) = \sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} \cdot W'_m x^l(\phi_l(\widehat{pq}) + \Delta pmlqk)$, where l indexes the input feature level, and K indexes the sampling point. Δp_{mlqk} and A_{mlqk} denote the sampling offset and attention weight of the k^{th} sampling point in the l^{th} feature level and the m^{th} attention head, respectively.

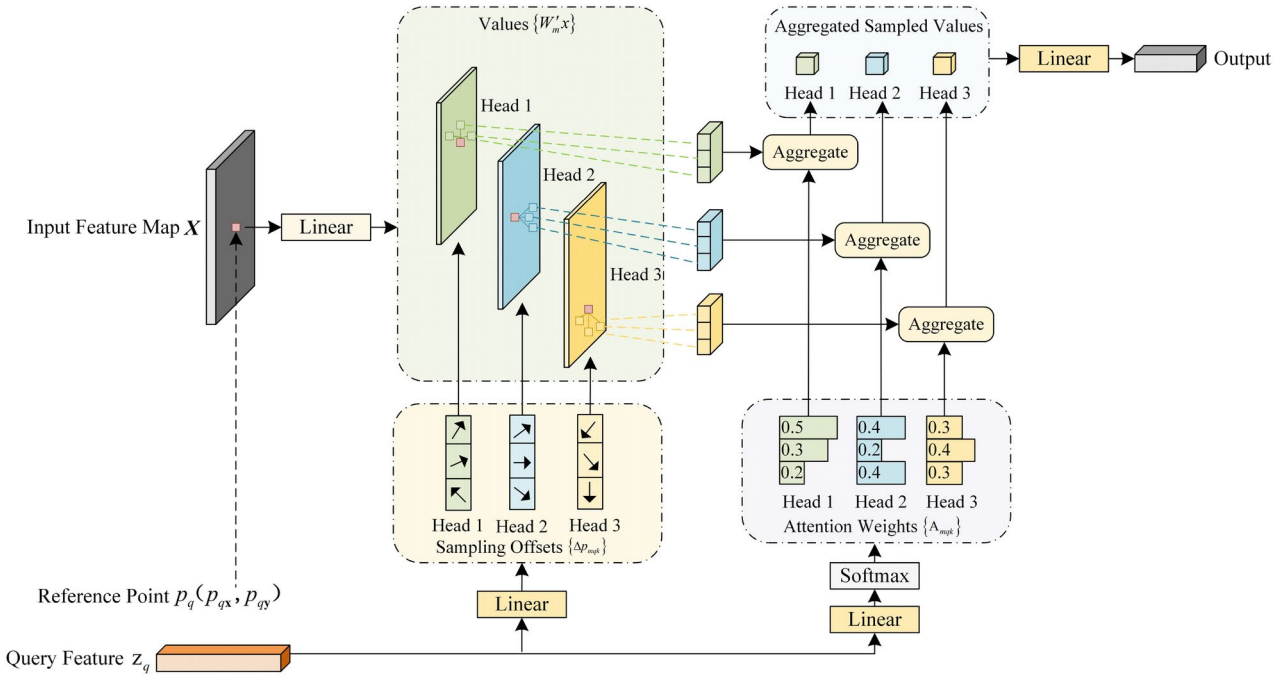

Figure 3. Architecture of the deformable DETR.

The scalar attention weight A_{mlqk} is normalized such that $\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} = 1$. Here, we use normalized coordinates $\widehat{p}_q \in [0,1]^2$ to clarify the scale formulation, where the normalized coordinates (0, 0) and (1, 1) indicate the top-left and bottom-right corners of the image, respectively. The function $\phi_l(\widehat{p}_q)$ in Eq. (2) re-scales the normalized coordinates $\widehat{p}_q$ to the input feature map of the l-th level. The multi-scale deformable attention is similar to the previous single-scale version, except it samples LK points from multi-scale feature maps instead of K points from single-scale feature maps.

Linguistic model

When processing linguistic features, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks, are often considered adequate due to their strong sequential modeling capabilities. These networks excel at capturing long-range dependencies in sequential data, which is crucial for understanding contextual information in language tasks. However, these methods may face limitations when applied to text recognition tasks involving complex spatial structures, such as Yi characters.

In contrast, Pyramid Pooling offers significant advantages for Yi character recognition, particularly in addressing multi-scale and spatial variability[33,34]. Unlike RNNs[35], LSTMs[36], and BiLSTMs[37], which may struggle with the complex spatial variations and multi-scale features inherent in Yi characters, The Pyramid Pooling mechanism aggregates information across multiple scales, enabling the model to capture fine-grained and broader structural features of Yi characters, such as stroke density, direction, and size variations. By enhancing spatial feature resolution and overcoming the limitations of traditional sequential methods, Pyramid Pooling improves recognition performance, especially under distortions and rotations. This approach reduces reliance on conventional methods, increasing the model's robustness against diverse writing styles and font variations. As a result, it significantly boosts both recognition accuracy and efficiency. This design, incorporating Pyramid Pooling, further optimizes the proposed linguistic model, as shown in Fig. 4.

Initially, the input features undergo processing through a pyramid pooling-based Multi-Head Self-Attention (P-MHSA) mechanism, then integrating the resulting attributes with the original input via a residual connection. This process facilitates the fusion of enhanced feature information with the original input. Subsequently, the features undergo normalization via a LayerNorm (LN) layer, aiding in model convergence. The features are subsequently subjected to linear transformation and non-linear mapping through a Feed-Forward Neural Network (FFN). Additionally, the residual fusion of FFN output features with their corresponding mappings, followed by normalization through the LN layer, further captures the relationships among Yi characters, enhancing overall model performance. The procedure described above can be summarized as follows:

$$X_{att} = \mathrm{LN}(X + \text{P-MHSA}(X)) \tag{3}$$

$$X_{out} = \mathrm{LN}(X_{att} + \mathrm{FFN}(X_{att})) \tag{4}$$

In this context, X, X_{att}, and X_{out} represent the inputs, the outputs of the MHSA, and the outcome of the transformer, respectively. Furthermore, P-MHSA denotes pooling-based multi-head attention.

The conventional transformer encounters limitations in processing Yi character images, including restricted sequence length, high computational overhead, and insufficient dynamic management of positional relationships within the input data. Therefore, we employ the Pyramid Pooling Transformer architecture, which processes multiple feature map scales and captures hierarchical contextual nuances specific to Yi characters. enhancing recognition accuracy. This innovation also reduces the input sequence length, thereby lightening the computational burden and improving performance and adaptability across various multimodal tasks. The configuration of the Pyramid Pooling Transformer is depicted in Fig. 5.

Firstly, the input variable X undergoes a dimensional transformation to conform to a 2-d spatial framework. Secondly, X undergoes several average pooling layers, resulting in a pyramidal feature map, as shown in Eq. (5):

$$P_i = \mathrm{Avgpool}_i(X), i \in (1, 2, ..., n) \tag{5}$$

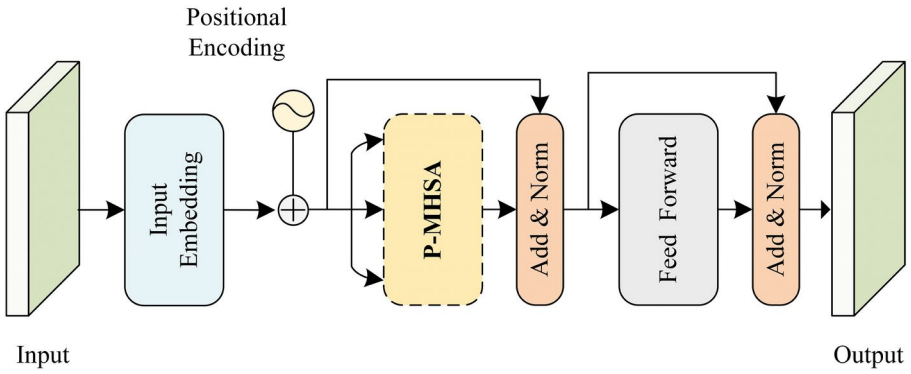

Figure 4. Overall structure of linguistic model.

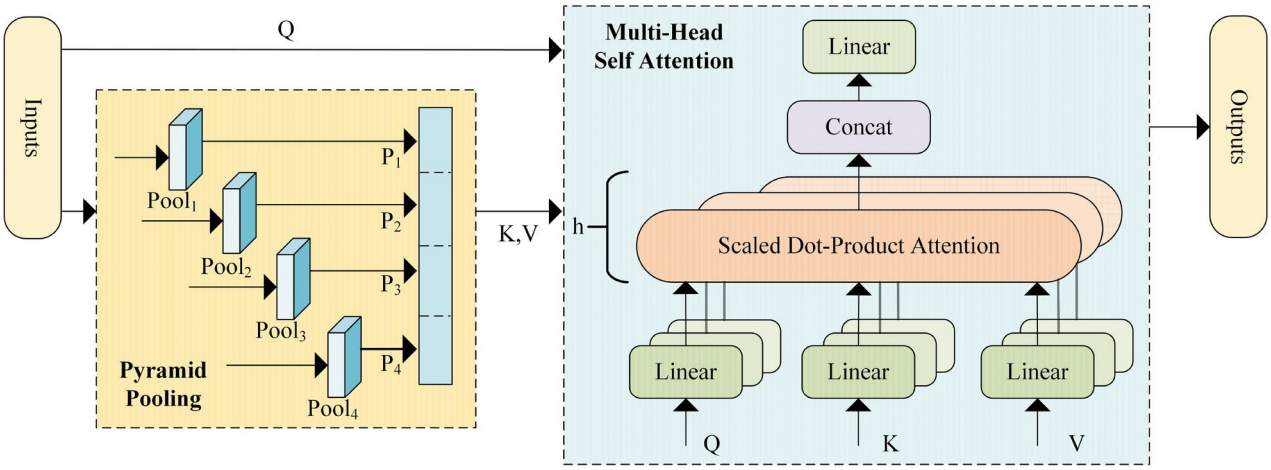

Figure 5. Architecture of the pyramid pooling transformer.

In this context, $P_1, P_2, \ldots, P_n$ represents the pyramid feature maps, while n denotes the number of pooling layers. Subsequently, these pyramid feature maps are passed through the depth-wise convolution layer, where relative position encoding is applied, as shown in Eq. (6):

$$P_i^{enc} = \mathrm{DWConv}(P_i) + P_i, i \in 1, 2, ..., n \tag{6}$$

where $\mathrm{DWConv}(\cdot)$ represents the depth-wise convolution kernel of size 3×3, while P_i^{enc} denotes the encoding of relative positional information within P_i. Since P_i represents a collection of ensemble features, it is essential to note that the computational cost for the operation defined in Eq. (6) is minimal. Subsequently, these pyramid feature maps undergo expansion and interconnection, as described in Eq. (7):

$$P = \mathrm{LN}(Concat(\zeta)) \tag{7}$$

where $\zeta = P_1^{enc}, P_2^{enc}, P_3^{enc}$.

The flattening operation is omitted for simplicity. Therefore, if the pooling ratio is sufficiently large, P can be a shorter sequence than the input X. Furthermore, P encapsulates the contextual abstraction features of input X and, as a result, can serve as a strong alternative to X when computing the MHSA.

Within the framework of MHSA, the tensors denoted as Q, K, and V represent the query, key, and value components, respectively. The approach outlined is presented in Eq. (8):

$$(Q, K, V) = (XW^q, PW^k, PW^v) \tag{8}$$

where W^q, W^k, and W^v represent the weight matrices involved in the linear transformations for the query, key, and value tensors, respectively. Subsequently, Q, K, and V are input into the attention module; thus, the calculation of the attention tensor A is represented as follows:

$$A = Softmax(\frac{Q \times K^T}{\sqrt{d_k}}) \times V \tag{9}$$

In this context, d_k represents the channel dimension of K, while $\sqrt{d_k}$ serves a dual purpose: facilitating approximate normalization and performing softmax operations across each matrix row.

Fusion model

In Yi character recognition tasks, the seamless integration of visual and linguistic models, which combines image and textual information, is crucial for enhancing model performance. Traditional multimodal fusion methods, such as feature concatenation or weighted summation, often fail to capture the intricate interdependencies between modalities, leading to suboptimal outcomes. These methods combine visual and linguistic features without adaptively considering their interrelationships, which is inefficient, especially for Yi characters' complex, spatially varied structures. To address this, we introduce a cross-attention mechanism that captures the interactions between visual features-such as stroke patterns and spatial arrangements-and linguistic features, encompassing Yi characters' structural and semantic properties. Unlike traditional methods, the cross-attention mechanism dynamically computes the relevance between modalities, adjusting each modality's contribution based on contextual importance. This enables efficient information transfer and aligns visual patterns with corresponding textual properties. Specifically, the CLS token from each branch mediates communication between the visual and linguistic components, enhancing the model's understanding of both. The exchanged information is then reintegrated into the original branches, enriching the token representation. Thus, the cross-

attention mechanism significantly improves multimodal integration by enabling dynamic, context-aware fusion, which is particularly beneficial for tasks like Yi character recognition involving complex, multi-scale structures and contextual dependencies.

Since the CLS token captures the abstract content of all patch tokens within its branch, its interaction with tokens from other branches facilitates the extraction of insights across multiple scales. This interaction is crucial for Yi character recognition, as the characters frequently display varying scales and intricate visual patterns. The CLS token learns to contextualize these variations through interaction with linguistic tokens that capture the structural properties of Yi characters. After assimilating with CLS tokens from other branches, each CLS token re-engages with patch tokens in its native branch in subsequent Transformer encoders. This step ensures that the model can refine the local visual context-such as stroke orientation and character size-using the broader linguistic understanding of character structure, thereby enhancing recognition accuracy. Consequently, this cross-attention mechanism enhances the model's ability to capture both local and global features of Yi characters, enabling the recognition of characters despite variations in stroke density, distortion, or rotation. By fostering collaboration across different branches, this mechanism enhances the model's robustness and enables it to adapt more efficiently to various challenging text scenarios. The cross-attention framework is illustrated in Fig. 6.

The CLS token from the expansive branch, represented by a circle, serves as a query and interacts with the patch token from the smaller branch through the attention mechanism. $f^l(\cdot)$ and $g^l(\cdot)$ represent projections that maintain consistent dimensions. The smaller branch follows the same procedure, with the exchange of CLS tokens and patch tokens forming an integral part of the process.

$$x'^l = [f^l(x^l_{cls})||x^s_{patch}] \tag{10}$$

where $f^l(\cdot)$ represents the projection function for dimensional alignment. Subsequently, the module performs a Cross Attention (CA) operation between x^l_{cls} and x'^l. Here, the CLS token serves as a unique query, facilitating information integration from the image block tokens into the CLS token. Mathematically, the CA operation can be expressed as follows:

$$q = x'^l_{cls}W_q, k = x'^l W_k, v = x'^l W_v \tag{11}$$

$$A = Softmax(qk^T/\sqrt{C/h}), CA(x'^l) = Av \tag{12}$$

where W_q, W_k, and $W_v \in \mathbb{R}^{C\times(C/h)}$ are learnable parameters, with C and h representing the number of embedded dimensions and attention heads, respectively.

Furthermore, like self-attention, CA employs multi-head attention, referred to as MCA. Specifically, for a given x^l, whose output from cross-attention is z^l, layer normalization and a residual connection are applied, as shown in Eqs. (13) and (14):

$$y^l_{cls} = f^l(x^l_{cls}) + MCA(LN(\xi)) \tag{13}$$

$$z^l = [g^l(y^l_{cls}||x^l_{patch})] \tag{14}$$

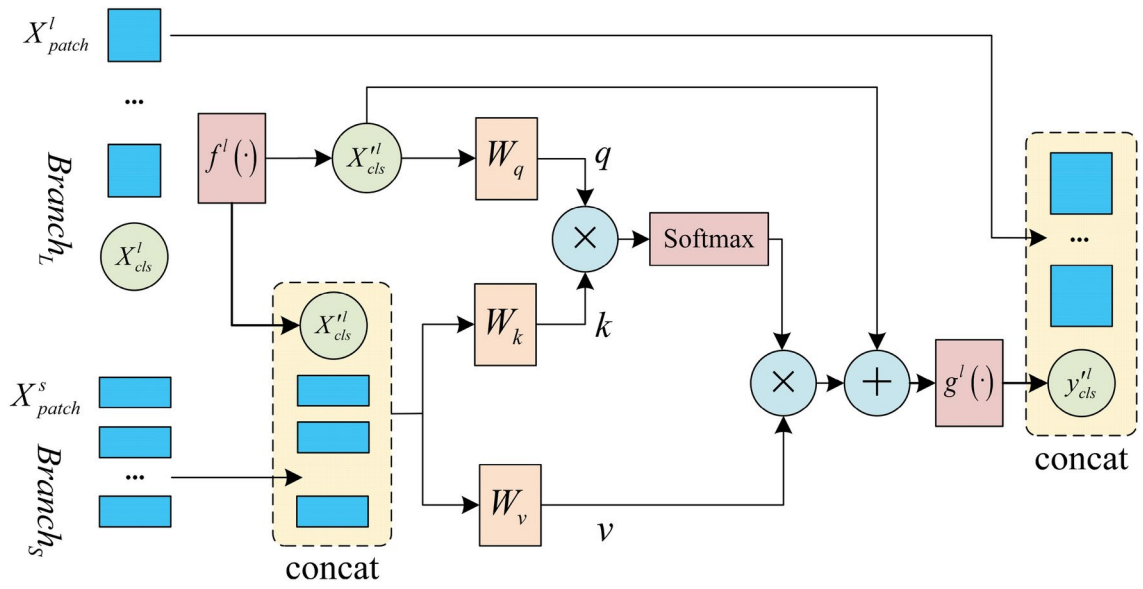

Figure 6. Architecture of the cross attention.

Datasets	Total size	Train size	Test size	Validation size
Original	200	160	20	20
Dataset_A	300	240	30	30
Dataset_B	300	240	30	30
Dataset_C	300	240	30	30

Table 1. Dataset description and sizes. Note: All sizes are in pages

Parameter	Version information
OS	Linux
CPU	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
Memory	64.0GB
GPU	NVIDIA GeForce GPU 3090Ti
Language	Python 3.8.0
Framework	Paddle 2.7.1

Table 2. Experimental environment.

In this context, $\xi = [f]^l(x^l_{cls})||x^s_{patch}$, where $f^l(\cdot)$ and $g^l(\cdot)$ represent the dimensional alignment projection and inverse projection functions, respectively.

Experiments

Datasets

The limited availability of publicly available datasets for Yi character images hinders research and practical applications in Yi character recognition tasks. Yi characters represent a significant aspect of cultural heritage; however, data scarcity hampers the effective processing of these characters by deep learning models. To address this challenge, we used image data from scanned copies of traditional Yi scriptures, including the ancient texts *Mamteyi*, *Leoteyi*, and *Guiding Meridian*, totaling 200 raw images. We employed several data augmentation techniques to enrich the dataset, simulating real-world variations and enhancing the model's ability to perform effectively across diverse conditions.

Specifically, this approach employs a multi-step background removal technique on the original Yi character images. It begins with contrast adjustment using Retinex, followed by fine-tuning with adaptive contrast enhancement. Subsequently, adaptive thresholding segments the foreground, and geometric features filter out the desired foreground, resulting in transparent backgrounds. This process isolates the characters, removes background noise, and effectively highlights their features. In the following step, patch-based synthesis and image quilting techniques merge the transparent Yi character images with various backgrounds, creating three augmented datasets: Dataset_A, Dataset_B, and Dataset_C. These synthetic backgrounds simulate different real-world scenarios, enhancing the dataset's diversity. Finally, Gaussian filtering is applied to both the original and augmented datasets to reduce noise and smooth the Yi character images, improving overall quality. Table 1 provides the dataset descriptions and sizes.

Experimental environment and training strategies

The Paddle deep learning framework is utilized to evaluate the effectiveness of the proposed method. Model training uses an NVIDIA GeForce GPU 3090 Ti paired with an Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz. A detailed configuration of the experimental setup is provided in Table 2. The Adam optimizer was employed during training with a momentum parameter of 0.9. The learning rate was subjected to exponential decay, starting at 0.0001. Each batch consisted of 16 samples, and training lasted for 500 epochs.

Performance metrics

In this study, recognition accuracy (ACC) is the primary metric to evaluate the performance of Yi character image recognition. It is calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{15}$$

TP (True Positive) refers to instances where the classifier correctly identifies a positive case; FP (False Positive) refers to instances where the classifier incorrectly classifies a negative example as positive; FN (False Negative) refers to instances where the classifier incorrectly classifies a positive example as unfavorable; and TN (True Negative) refers to instances where the classifier correctly identifies a negative case.

Method	+Deformable DETR	+Pyramid Pooling Transformer	+Cross Attention	ACC /%
				0.961
	✓			0.972
		✓		0.983
Proposed			✓	0.980
	✓	✓		0.985
		✓	✓	0.989
	✓		✓	0.981
	✓	✓	✓	**0.995**

Table 3. Ablation experiment results.

Method	Backbone	ACC /%
	ResNet-18	0.976
	ResNet-34	0.980
Proposed	**ResNet-45**	**0.995**
	ResNet-50	0.973
	ResNet-101	0.971

Table 4. Experimental results under different feature extraction networks.

Method	Module	ACC /%
	DETR	0.983
Proposed	Dynamic DETR	0.989
	Deformable DETR	**0.995**

Table 5. Experimental results with different DETR variants.

Experimental results analysis

Ablation experiment

We performed ablation experiments on the modified module to validate the effectiveness of the proposed technique for Yi character image recognition. These experiments were carried out on the Yi character recognition dataset, with the results in Table 3.

As shown in Table 3, the entries labeled +Deformable DETR, +Pyramid Pooling Transformer, and +Cross Attention integrate the deformable attention, Pyramid Pooling Transformer, and Cross Attention modules, respectively, into the base ABINet method. The original ABINet achieves a recognition accuracy of 96.1% on the Yi character recognition dataset. When implemented individually, these modules contribute substantial improvements of 1.1%, 2.2%, and 1.9% to the overall accuracy, respectively. This underscores the significant contribution of the three aforementioned modules to Yi character recognition. Additionally, these modules tackle challenges such as unclear backgrounds, varied font styles, and complex relationships between adjacent characters with similar semantics. Integrating Deformable DETR and the Pyramid Pooling Transformer increases recognition accuracy to 98.9%. Furthermore, combining the Pyramid Pooling Transformer and Cross Attention modules improves accuracy by two percentage points over the baseline. Integrating all three modules into ABINet's visual, linguistic, and fusion models results in a recognition accuracy of 99.5%. This reflects a 3.4% improvement over the original ABINet, confirming the effectiveness of the proposed enhancements.

As shown in Table 4, the proposed method utilizes different backbone networks in the visual model for Yi character recognition, yielding varied results. Specifically, ResNet-45 is employed as the backbone for comparison with other feature extraction networks. The results demonstrate a 1.5% increase in recognition accuracy compared to the following best-performing network. Additionally, the proposed method exhibits significantly improved feature extraction capabilities with ResNet-45.

To evaluate the effectiveness of the Deformable DETR module, we conducted ablation experiments with DETR[41] and Dynamic DETR[42]. The experimental results are shown in Table 5. DETR, using global self-attention, achieved an accuracy of 98.3%. Dynamic DETR, incorporating dynamic attention, improved the accuracy to 98.9%. In contrast, Deformable DETR, with a deformable attention mechanism, achieved the highest accuracy of 99.5%. These results highlight Deformable DETR's superior performance in recognition accuracy.

Comparison experiment

To validate the effectiveness of the proposed method, we compared it with several state-of-the-art text recognition approaches, including CRNN[15], SRN[26], ViTSTR[38], RFL[39], ABINet[28], and SVTR[40], all selected for their relevance

Method	ACC /%	Training time/h	Year
CRNN[15]	0.939	**28**	2016
SRN[26]	0.948	42	2020
ViTSTR[38]	0.916	45	2021
RFL[39]	0.920	30	2021
ABINet[28]	0.961	48	2021
SVTR[40]	0.952	52	2022
Proposed	**0.995**	50	

Table 6. Comparison experiment results.

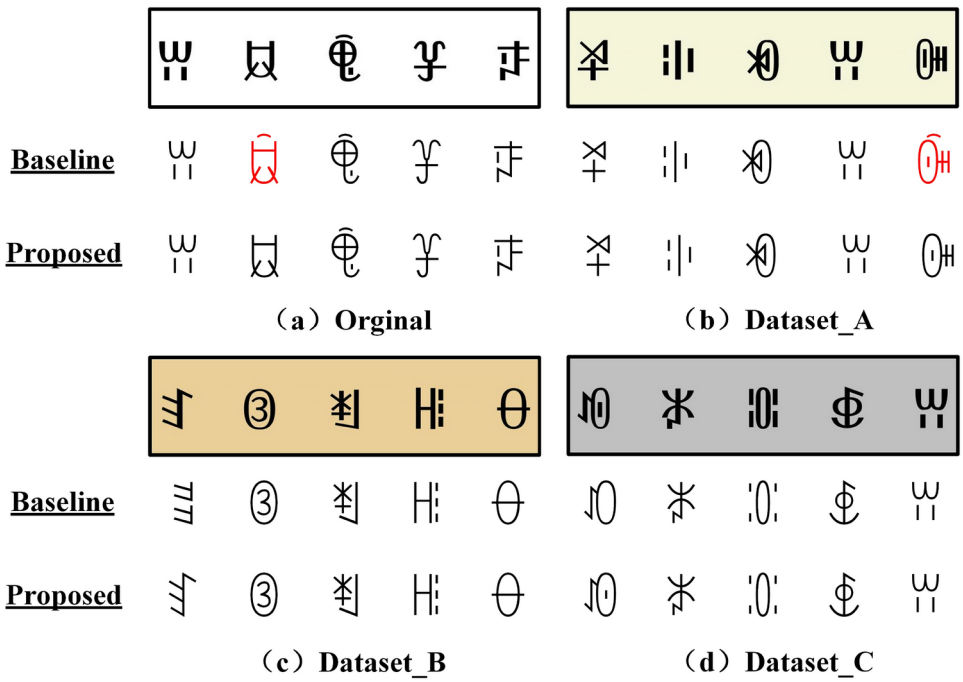

Figure 7. Visualization of text recognition results.

and distinct advantages in Yi character recognition. CRNN combines CNNs for feature extraction and RNNs for sequence modeling, making capturing complex stroke patterns practical. SRN captures local and global contextual relationships essential for Yi characters' context-dependent meaning. ViTSTR, leveraging a vision transformer architecture, excels at handling fine-grained structures. At the same time, RFL focuses on dense feature extraction, which is particularly useful for Yi characters' overlapping strokes and intricate components. ABINet, chosen as the baseline model, uses an attention-based mechanism to focus on relevant image regions and enhances Yi character recognition by integrating visual and linguistic models. Lastly, SVTR employs a patch-based approach that eliminates sequential modeling, making it well-suited for Yi script's complex, multi-component characters. As shown in Table 6, the proposed method outperforms these models, achieving accuracy improvements of 5.6%, 4.7%, 7.9%, 7.5%, 3.4%, and 4.3% over CRNN, SRN, ViTSTR, RFL, ABINet, and SVTR, respectively.

As shown in Table 6, the results highlight the superior performance of the proposed method on the proprietary Yi character recognition dataset. When compared to CRNN, SRN, ViTSTR, RFL, ABINet, and SVTR, the proposed approach achieves substantial accuracy improvements of 5.6%, 4.7%, 7.9%, 7.5%, 3.4%, and 4.3%, respectively.

Visualization and analysis

Figure 7 presents visualizations of the text recognition results on the Yi character datasets, comparing the baseline method with the proposed approach. The baseline method fails to capture finer details of the Yi characters, as evidenced by the red-marked areas indicating discrepancies with the original images. In contrast, our approach more effectively captures these details. The results demonstrate the stability and effectiveness of the proposed method across the datasets.

Conclusions

Recognizing Yi characters presents significant challenges due to their complex morphology and the intricate semantic relationships between characters, contributing to reduced recognition accuracy. We propose a multimodal approach that combines linguistic and visual information to address these challenges. In the visual modeling phase, we enhance the recognition model by incorporating Deformable DETR, which combines deformable attention and Transformer mechanisms to improve performance, particularly for images with deformations or complex backgrounds. The linguistic modeling phase employs a Pyramid Pooling Transformer that captures multi-scale contextual features, enhancing the representation of semantic information. In the fusion modeling phase, a cross-attention mechanism is used to refine the integration of visual and linguistic features, optimizing the overall recognition process. Experimental results demonstrate that the proposed method achieves a recognition accuracy of 99.5%, surpassing the baseline method by 3.4%, highlighting the effectiveness and precision of the approach.

Although our approach effectively improves accuracy, it still faces limitations, particularly regarding computational complexity. Deformable DETR introduces higher computational demands when processing high-resolution images, which may limit scalability in specific applications. Similarly, the Pyramid Pooling Transformer incurs significant computational costs as additional scales are introduced, potentially affecting efficiency in resource-constrained scenarios. To address these issues, future work will focus on developing lightweight architectures for Deformable DETR to reduce computational complexity without compromising performance. The Pyramid Pooling Transformer will also be optimized by implementing more efficient mechanisms for capturing multi-scale contextual features, ensuring its applicability in resource-limited environments. Additionally, we will conduct comprehensive cross-linguistic evaluations to assess the scalability and robustness of our method across a wide range of languages, testing its ability to handle larger datasets and perform effectively in real-world scenarios.

Data availability

The algorithm in this study has been implemented based on the PaddlePaddle framework, which is publicly and freely available at https://github.com/PaddlePaddle/PaddleOCR/tree/release/2.7.1.

Received: 26 November 2024; Accepted: 27 March 2025
Published online: 07 April 2025

References

1. Chinthaginjala, R., Dhanamjayulu, C., Kim, T.-H., Ahmed, S., Kim, S.-Y., Kumar, A. S., Annepu, V., & Ahmad, S. Enhancing Handwritten Text Recognition Accuracy with Gated Mechanisms. *Sci. Rep.* **14**(1), 16800 (2024).
2. Ptucha, R., Such, F. P., Pillai, S., Brockler, F., Singh, V., & Hutkowski, P. Intelligent Character Recognition Using Fully Convolutional Neural Networks. *Pattern Recogn.* **88**, 604–613 (2019).
3. Chen, S., Yang, Y., Liu, X. & Zhu, S. Dual discriminator gan: Restoring ancient yi characters. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2**(4), 1–23 (2022).
4. Yin, X., Min, D., Huo, Y. & Yoon, S.-E. Contour-Aware Equipotential Learning for Semantic Segmentation. *IEEE Trans. Multimedia* **25**, 6146–6156 (2022).
5. Yin, X., Im, W., Min, D., Huo, Y., Pan, F., & Yoon, S.-E. Fine-grained Background Representation for Weakly Supervised Semantic Segmentation. *IEEE Trans. Circ. Syst. Video Technol.* (2024).
6. Yin, X., Pan, F., An, G., Huo, Y., Xie, Z., & Yoon, S.-E. OpenSlot: Mixed Open-set Recognition with Object-centric Learning. *arXiv preprint* arXiv:2407.02386, (2024).
7. Li, L., Li, J., Wang, H. & Nie, J. Application of the Transformer Model Algorithm in Chinese Word Sense Disambiguation: a case study in chinese language. *Sci. Rep.* **14**(1), 6320 (2024).
8. Zhang, Z., Wang, L. & Cheng, S. Composed Query Image Retrieval based on Triangle Area Triple Loss Function and Combining CNN with Transformer. *Sci. Rep.* **12**(1), 20800 (2022).
9. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. Deformable detr: Deformable Transformers for End-to-end Object Detection. arXiv preprint arXiv:2010.04159, (2020).
10. Yu-Huan, W., Liu, Y., Zhan, X. & Cheng, M.-M. P2T: Pyramid Pooling Transformer for Scene Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(11), 12760–12771 (2023).
11. Chen, C.-F. R., Fan, Q., & Panda, R. Crossvit: Cross-attention Multi-scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 357-366 (2021).
12. Li, M. et al. Trocr: Transformer-based Optical Character Recognition with Pre-trained Models. *In Proceedings of the AAAI Conference on Artificial Intelligence* **37**, 13094–13102 (2023).
13. Zhang, H. et al. GLaLT: Global-Local Attention-Augmented Light Transformer for Scene Text Recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **35**(7), 10145–10158 (2024).
14. Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, 369–376 (2006).
15. Shi, B., Bai, X. & Yao, C. 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.* **39**(11), 2298–2304 (2016).
16. Gao, Y., Chen, Y., Wang, J., Tang, M., & Lu, H. Dense Chained Attention Network for Scene Text Recognition. In *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, 679–683 (2018).
17. Qi, X., Chen, Y., Xiao, R., Li, C.-G., Zou, Q. & Cui, S. A Novel Joint Character Categorization and Localization Approach for Character-level Scene Text Recognition. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 5: 83–90 (2019).
18. Bolan, S. & Shijian, L. Accurate Recognition of Words in Scenes Without Character Segmentation Using Recurrent Neural Network. *Pattern Recogn.* **63**, 397–405 (2017).
19. Liu, H., Jin, S., & Zhang, C. Connectionist Temporal Classification with Maximum Entropy Regularization. *Adv. Neural Inf. Process. Syst.* **31** (2018).
20. Feng, X., Yao, H., & Zhang, S. Focal CTC Loss for Chinese Optical Character Recognition on Unbalanced Datasets. *Complexity* (2019).
21. Wenyang, H., Cai, X., Hou, J., Yi, S. & Lin, Z. Gtc: Guided Training of ctc Towards Efficient and Accurate Scene Text Recognition. *In Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 11005–11012 (2020).

22. Miao, Y., Gowayyed, M., & Metze, F. EESEN: End-to-end Speech Recognition Using Deep RNN Models and WFST-based Decoding. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 167–174 (2015).
23. Wan, Z., Xie, F., Liu, Y., Bai, X. & Yao, C. 2D-CTC for Scene Text Recognition. arXiv preprint arXiv:1907.09705, (2019).
24. Shi, B., Wang, X., Lyu, P., Yao, C., & Bai, X. Robust Scene Text Recognition with Automatic Rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4168–4176 (2016).
25. Wang, T. et al. Decoupled Attention Network for Text Recognition. *In Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 12216–12224 (2020).
26. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., & Ding, E. Towards Accurate Scene Text Recognition with Semantic Reasoning Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12113–12122 (2020).
27. Sheng, F., Chen, Z., & Xu, B. NRTR: A No-recurrence Sequence-to-sequence Model for Scene Text Recognition. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, 781–786 (2019).
28. Fang, S., Xie, H., Wang, Y., Mao, Z., & Zhang, Y. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7098–7107 (2021).
29. Xiaodong, J., Wendong, G., & Jie, Y. Handwritten Yi Character Recognition with Density-Based Clustering Algorithm and Convolutional Neural Network. In *Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 1: 337–341 (2017).
30. Jiejue, Y. The Algorithm and Implementation of Yi Character Recognition based on Convolutional Neural Network. In *Proceedings of the 2022 2nd International Conference on Networking, Communications and Information Technology (NetCIT)*, 346–349 (2022).
31. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. Deformable Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 764–773 (2017).
32. Liu, L. et al. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vision* **128**, 261–318 (2020).
33. He, K., Zhang, X., Ren, S. & Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
34. Li, M. et al. Trocr: Transformer-based Optical Character Recognition with Pre-trained Models. *In Proceedings of the AAAI Conference on Artificial Intelligence* **37**, 13094–13102 (2023).
35. Zaremba, W. Recurrent Neural Network Regularization. arXiv preprint arXiv:1409.2329, (2014).
36. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **28**(10), 2222–2232 (2017).
37. Siami-Namini, S., Tavakoli, N., & Namin, A. S. The Performance of LSTM and BiLSTM in Forecasting Time Series. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, 3285–3292 (2019).
38. Atienza, R. Vision Transformer for Fast and Efficient Scene Text Recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*, 319–334 (2021).
39. Jiang, H., Xu, Y., Cheng, Z., Pu, S., Niu, Y., Ren, W., Wu, F., & Tan, W. Reciprocal Feature Learning via Explicit and Implicit Tasks in Scene Text Recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*, 287–303 (2021).
40. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., & Jiang, Y.-G. Svtr: Scene Text Recognition with a Single Visual Model. arXiv preprint arXiv:2205.00159, (2022).
41. Nicolas, C., Francisco, M., Gabriel, S., Nicolas, U., Alexander, K., & Sergey, Z. End-to-end Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision*, 213-229 (2020).
42. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L. & Zhang, L. Dynamic DETR: End-to-End Object Detection with Dynamic Attention. In *Proceedings of the IEEE International Conference on Computer Vision*, 2988-2997 (2021).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China Grants 61972062, 62402085, and 62306060, the Liaoning Basic Research Project 2023JH2/101300191, the Liaoning Doctoral Research Start-up Fund 2023-BS-078, and the Dalian Youth Science and Technology Star Project 2023RQ023, and the Open Project of Key Laboratory of Ethnic Language Intelligent Analysis and Security Management of MOE under Grant ORP-202401.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: H.P. Sun, X.Y. Ding, Z.M. Li; data collection: H.P. Sun, J. Sun, H. Yu; analysis and interpretation of results: H.P. Sun, X.Y. Ding, J. Sun, J.X. Zhang, H. Yu; draft manuscript preparation: H.P. Sun, X.Y. Ding, J. Sun, J.X. Zhang, H. Yu. All authors reviewed the results and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.D. or J.Z.