AHVE-CNER: Aligned Hanzi Visual Encoding Enhance Chinese Named Entity Recognition with Multi-Information

Xuhui Zheng[1,2], Zhiyuan Min[3], Bin Shi[1,2], Hao Wang[1,2]*,
[1]School of Information Management, Nanjing University, China
[2]Key Laboratory of Data Engineering and Knowledge Services in Jiangsu Provincial Universities (Nanjing University), China
[3]School of Software Technol, Zhejiang University, China
{zhengxuhui, binshi}@smail.nju.edu.cn; minzhiyuan@zju.edu.cn
Correspondence: ywhaowang@nju.edu.cn

Abstract

The integration of multi-modal information, especially the graphic features of Hanzi, is crucial for improving the performance of Chinese Named Entity Recognition (NER) tasks. However, existing glyph-based models frequently neglect the relationship between pictorial elements and radicals. This paper presents AHVE-CNER, a model that integrates multi-source visual and phonetic information of Hanzi, while explicitly aligning pictographic features with their corresponding radicals. We propose the Gated Pangu-π Cross Transformer to effectively facilitate the integration of these multi-modal representations. By leveraging a multi-source glyph alignment strategy, AHVE-CNER demonstrates an improved capability to capture the visual and semantic nuances of Hanzi for NER tasks. Extensive experiments on benchmark datasets validate that AHVE-CNER achieves superior performance compared to existing multi-modal Chinese NER methods. Additional ablation studies further confirm the effectiveness of our visual alignment module and the fusion approach.[1]

1 Introduction

Named entity recognition (NER) is one of the basic tasks of information extraction and an important NLP research. The results of NER will directly affect the downstream tasks (Liu et al., 2021b; Martins et al., 2019; Nasar et al., 2021)

Compared to English based on Latin, Chinese NER tasks face more problems. It does not have similar natural word separators and explicit word boundaries (Ma et al., 2019). On the other hand, Chinese does not contain roots or affixes similar to English words (Yadav et al., 2018) and computers understand Chinese texts in units of characters. The composition of Chinese characters(Hanzi) is derived from graphics, which itself is a recording language that evolves from the image characteristics of objects (Norman, 1988), so Chinese characters can also be **split into radicals**, **and each part has rich abstract graphic meanings**. Many studies have paid attention to the visual information of Hanzi et al. over NER tasks (Gu et al., 2023; Meng et al., 2019; Qi et al., 2023; Sehanobish and Song, 2019; Wu et al., 2021a; Xuan et al., 2021). In these studies, there were two ways to process glyph structures. One is to directly extract features from a single character image using a convolution neural network as additional inputs, and the other is to split characters into serialized radicals, based on the dictionary, those are mapped one-to-one to multiple vectors which can be learned.

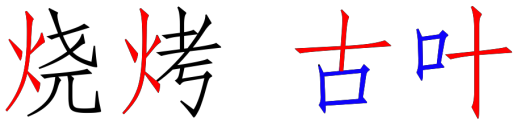
Radical	Hanzi	Words
钅 (gold)	钢(steel), 铁(iron)	钢铁(steel)
氵 (water)	漂(rinse), 流(flow)	漂流(drifting)
火 (fire)	烧(burn), 烤(roast)	烧烤(barbecue)

Table 1: Decomposition of radicals of Chinese characters: Words with similar structures and similar meanings form new words of the same type.

However, neither of these two processing methods directly connects the graphics of Chinese characters with their radical structures. For example, "银" (silver), "铜" (bronze), "铁" (iron), and "钢" (steel) all contain the radical "钅" (gold), which evolved from the character "金" (metal) and indicates a connection to metal. Other similar examples are shown in Table 1, **Chinese characters with the same radical often form words with related meanings**, such as "漂流" (drifting) and "烧烤" (barbecue). What's more, characters with the same radicals can have different glyphs due to their structural arrangements. There are 12 possible structural types, including left-right and up-down configurations. Therefore, characters with identical structural components can present different visual

*Corresponding author.
[1]The source code of the proposed method is publicly available at https://github.com/zxh20001117/AHVE-CNER.

Proceedings of the 31st International Conference on Computational Linguistics, pages 3391–3400
January 19–24, 2025.



(a) able to form. (b) unable to form.

Figure 1: Examples of different situations of words with same radicals. (a) 烧(burn) and 烤(roast) both have the radical of 火(fire) and they can form new word 烧烤(barbecue). But (b) 古(ancient) and 叶(leaf) are quite different in meaning thus can't form new word.

information. For instance, "囡" and "如", or "叶" and "古", share the components "女口" or "口十", but differ in their spatial relationships. So, **the spatial information of the graphics plays an important role with radicals to help us distinguish characters**. These cases are further illustrated in Figure 1.

To address the challenges in Chinese NER tasks of (1) **effectively utilizing the multimodal features of Chinese characters** and (2) **further aligning the pictographic features between character images and serialized radicals**, this article proposes a novel method called AHVE-CNER. The main contributions of this paper are:

- A novel Chinese character visual feature extraction framework that aligns pictographic information with radical information to create embeddings that encapsulate more structural rules.
- A cross-modal interactive fusion method called the Gated PanGu-π Cross Transformer, which integrates information across different modalities effectively.
- Comprehensive evaluation on multiple benchmark Chinese NER datasets, demonstrating superior performance and effectiveness compared to other multi-source information models.

2 Related Works

2.1 Chinese Character Visual Information Enhances NER

A common approach involves directly processing character images. Meng et al. (2019) proposed the Tianzige-CNN structure to extract visual features, while Sun's ChineseBERT (Sun et al., 2021) incorporates character images and pinyin information into training. Subsequent studies by Xuan et al. (2021), Gu et al. (2023), and Guo et al. (2022) used 3D convolution to capture relationships between adjacent characters. However, these methods are limited in processing the full pictographic visual information that glyphs offer, as they rely solely on images. Furthermore, they perform feature fusion via vector concatenation, which neglects the interaction between visual and contextual information. The second approach involves stroke-based decomposition. MFE-NER (Li and Meng, 2021) and StyleBERT (Lv et al., 2022) use the "Wubi" method to decompose Chinese characters into stroke sequences, providing additional visual information. Meanwhile, MECT (Wu et al., 2021b) and VisPhone (Zhang et al., 2023) break down characters into radicals using a Structure Component (SC) dictionary for radical-level features. However, these methods rely on rule-based decomposition and do not capture the visual structural features that are unique to Chinese character images with distinct compositional elements.

2.2 Multi-source Information Fusion Assists NLP Tasks

Researchers often integrate multi-source and multimodal information to enhance NLP tasks. Zaratiana et al. (2022) proposed GN-NER, which uses graph neural networks to enrich span representations. Gui et al. (2019b) utilized a graph neural network based on global vocabulary semantics for better local information retrieval. LEBERT (Liu et al., 2021a) deeply integrates character and lexical knowledge in encoders. Several studies also incorporate information from external lexicons for support (Ma et al., 2019; Gui et al., 2019a; Jia et al., 2020; Mai et al., 2022a; Mengge et al., 2020; Li et al., 2020; Zhang and Yang, 2018).

Combining pictographic and semantic features from different sources remains a challenge. MPMCNER (Mai et al., 2022b) introduced cross-modal attention to integrate semantics, glyphs, and phonetics. MECT (Wu et al., 2021a) designed a cross-transformer structure for modality interaction, while VisPhone (Zhang et al., 2023) added a selective fusion module to control the representation of characteristics. Similarly, CGR-NER (Gu et al., 2023) proposed a feature selection function.

To the best of our knowledge, AHVE-CNER is the first model to align features at both the glyph image and glyph radical structure levels to enhance Chinese NER tasks. This model combines contextual and Pinyin information for comprehensive

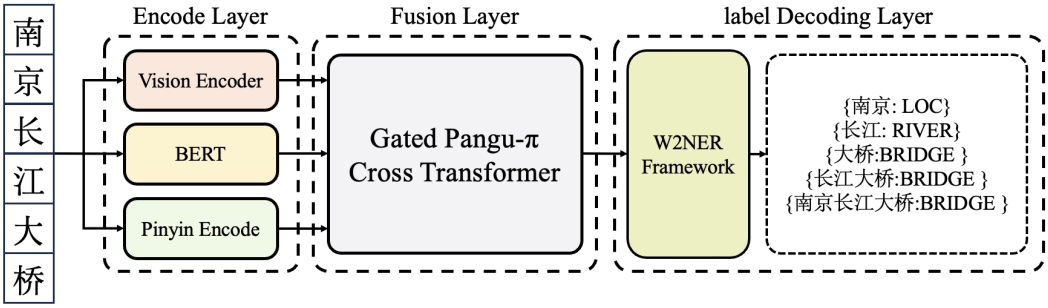


Figure 2: The overall framework of AHVE-CNER. There are three parts from left to right besides the input Chinese text. Encoder Layer transforms the input into three different modalities of information; Gated Pangu-π Cross Transformer then aims to merge them in Fusion Layer; W^{2}NER method is the key to decode previous features in a novel way.

feature extraction. We use ViT (Dosovitskiy et al., 2021) for the image feature extraction instead of CNN to directly map stroke structures to specific parts of Hanzi images. GRU (Cho et al., 2014) is employed to condense component sequence features of individual characters, and a multi-head self-attention mechanism facilitates global information interaction across characters. The Gated PanGu-π Cross Transformer fusion method is proposed to enhance non-linear interaction and optimize multi-modal information integration.

3 The Proposed AHVE-CNER Model

The framework of AHVE-CNER is shown in Figure 2. It mainly consists of three blocks. The input encoding part consists of the contextual encoder BERT, the Pinyin encoder and the Glyph encoder of the visual alignment structure. In the second part, two improved Cross Transformers based on Pangu-π (Wang et al., 2023) are used for the interaction between the three modal information. We use W^{2}NER (Li et al., 2022) framework for the final word to word relationship based label prediction.

3.1 Contextual Semantic Representation

Input the token converted from given text $\{C_1, C_2, \ldots, C_n\}$ into BERT, and obtain the contextual representation sequence $E_C = \{e_1, e_2, \ldots, e_n\}$, where e_i corresponds to the contextual representation of the characters C_i in the given text, n is the length of the characters contained in the text, $E_C \in \mathbb{R}^{(n \times d_c)}$, and d_c is the dimension of the output from BERT.

3.2 Visual Representation

As shown in Figure 3, the visual representation module takes two inputs: Hanzi images and radical information. Hanzi images are created in three writing styles using *.ttf* files, forming a three-channel image. Radical information is extracted from each character's Structure Component (SC) using a dictionary, which is available from the Online Xinhua Dictionary[2].

Step 1: The Hanzi image of size $I_i \in \mathbb{R}^{(3\times32\times32)}$, with three layers of Chinese characters, is processed using ViT (Dosovitskiy et al., 2021). The image is divided into 256 patches $I_p \in \mathbb{R}^{256\times(2^{2\cdot3})}$, each with a 2-pixel side length. These patches are arranged spatially and combined with 1D learnable positional embeddings to form the transformer input I_e. The output I_t is used for Hanzi visual feature alignment, with D_{patch} representing the dimension of patch embeddings after linear projection. We use E_p to represent the projection matrix in the following formula:

$$I_e = [I_p^1; I_p^2; \ldots; I_p^{256}]E_p + E_{ppos} \quad E_p \in \mathbb{R}^{(2^2\cdot3)\times D_{patch}} \tag{1}$$

$$I_t = TransformerEncoder(I_e) \tag{2}$$

Step 2: Using the structural component dictionary, each character is decomposed into a sequence of vectors $S_e \in \mathbb{R}^{l_S \times D_{patch}}$, where l_S represents the length of the structure component vector sequence. To align the image with the radicals, we use the SC embedding sequence as a query and compute its attention with the image embeddings, as described in Equation (3). This process enables the SC to align with patch embeddings that correspond to highly correlated parts of the image. The

[2] http://tool.httpcn.com/Zi/

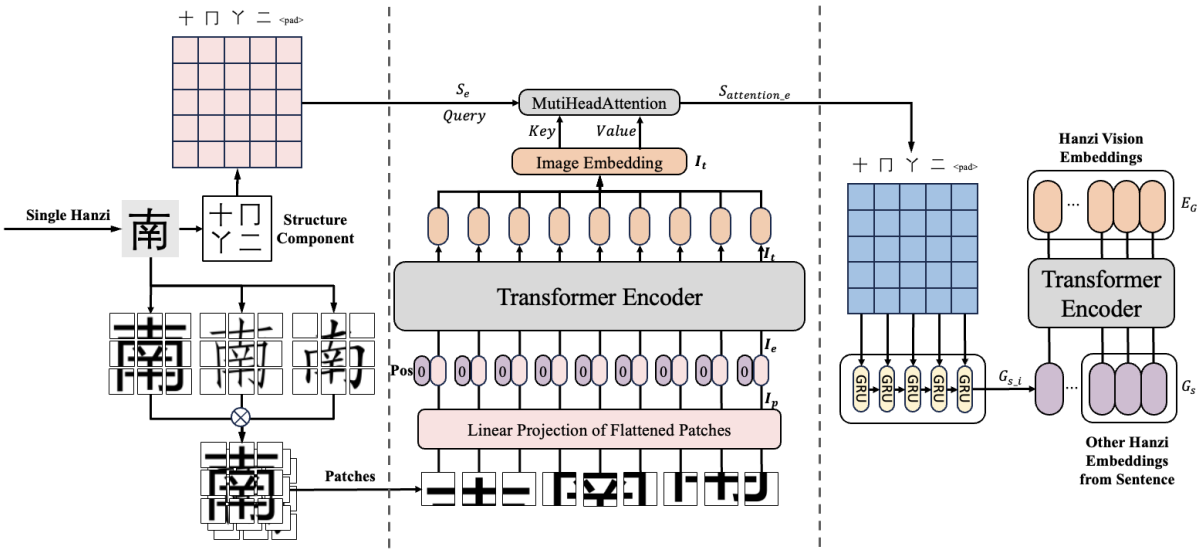


Figure 3: Aligned Hanzi Visual Encoder. The left part in this picture indicates two types of inputs including Hanzi image and the radical sequence; The middle part is the visual alignment encoder block; The right part aims to merge sequentia radical features.

resulting vector sequence, $S_{attention_e}$, integrates radical information with visual features from the Chinese character images.

$$\begin{aligned} S_{attention\_e} &= Attention(S_e, I_t, I_t) \\ &= Softmax(\frac{S_e I_t^T}{\sqrt{D_{patch}}})I_t \end{aligned} \tag{3}$$

Add GRU (Cho et al., 2014) to condense $S_{attention\_e}$ and only keep the output of its last layer as G_{s_i}. Stacked multiple GRU units is represented as $\overrightarrow{GRU}$ in Equation (4):

$$G_{s\_i} = \overrightarrow{GRU}(S_{attention\_e}) \tag{4}$$

To align the dimensions of the pronunciation representation with those of the contextual representation, we employ a linear layer for dimensionality adjustment. Additionally, position embeddings are added to G_S for all characters to incorporate positional information. The transformer encoder subsequently processes the enhanced pronunciation representations to produce the visual embedding $E_G = g_1, g_2, \ldots, g_n$.

3.3 Pinyin Representation

The pinyin sequence is processed following the method outlined in VisPhone (Zhang et al., 2023), which provides a systematic approach for extracting phonetic features of Chinese characters. Using the PyPinyin[3] toolkit, we extract the pinyin sequence and tones for each Chinese character. The toolkit supports multi-phonetic characters, such as "银行-h á ng (Bank)", "很行-x í ng (Absolutely OK)". A pinyin sequence is composed of three components in sequential order: initials (23 symbols), finals (38 symbols), and tones (5 symbols). In addition to the four standard tones, we include the light tone as a special pronunciation, encoding them numerically as 0-4.

Each Chinese character is decomposed into a pinyin sequence consisting of three components. Each component is mapped to a learnable vector representation. Subsequently, we apply one-dimensional convolution followed by max-pooling on the sequence of pronunciation embeddings to derive a single pronunciation embedding vector E_P.

3.4 Fusion Layer

To integrate the contextual E_C, visual E_G, and pronunciation E_P representations, we propose the Gated Pangu-π Cross Transformer. This approach combines cross-modal attention and gating mechanisms to facilitate interactions between different modalities, as illustrated in Figure 4.

The proposed method integrates visual and pronunciation features with contextual embeddings through a tri-modal gating unit, which facilitates effective fusion of the three modalities. In this framework, $C \rightarrow G$ and $C \rightarrow P$ denote the embeddings reconstructed from the contextual features (C), enriched with complementary informa-

[3] https://pypi.org/project/pypinyin, a toolkit employing deep learning to accurately determine the most appropriate pinyin based on context.

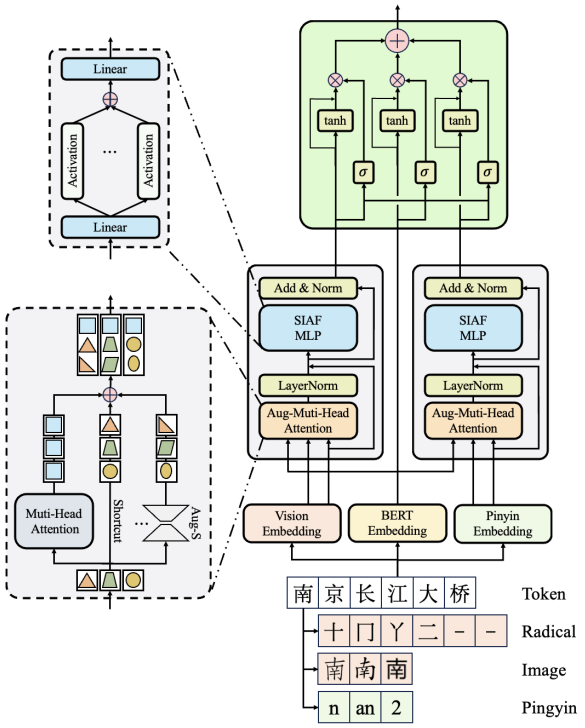


Figure 4: Gated Pangu-π Cross Transformer. Aug-S is the augmented shortcut method,and SIAF indicates the series activation function.

tion from Graphics (G) or Pronunciation (P). To enhance nonlinear interactions, parallel and augmented shortcut paths are introduced within the model. Additionally, a Series Activation Function (SIAF) is applied in the feed-forward network, utilizing multiple sequential activation functions to perform successive affine transformations. As detailed in Equation (5), the values Q_C, $K_G(P)$, and $V_G(P)$ are computed by multiplying the embeddings E_C, E_G, and E_P with distinct learnable parameter matrices. For simplicity, K, V, and E denote K_G, V_G, E_G and K_P, V_P, E_P, respectively, as used in parallel computations.

$$\begin{bmatrix} Q_{C,i} \\ K_i \\ V_i \end{bmatrix}^T = \begin{bmatrix} E_{C,i}W_{C,Q} \\ E_{K,i} \\ E_{V,i}W_V \end{bmatrix}^T \tag{5}$$

where each W is a learnable parameter matrix, and i is the i_{th} header after dividing the query, key and value. The augmented cross-modal multi-head attention mechanism is as Equation (6):

$$\begin{aligned} &AMH(Q,K,V) \\ =&MultiHead(Q,K,V)+V \\ &+\sum_{i=1}^{T}\tau_i(V;\Theta_i) \end{aligned} \tag{6}$$

$$\tau_i(V_{G(P)};\Theta_i)=\sigma(V_{G(P)}\Theta_i) \tag{7}$$

$$\begin{aligned} V^{'} =&LN(AMH(Q,K,V)) \\ &+V_G \end{aligned} \tag{8}$$

where τ_i indicates the i_{th} augmented shortcut function, T is the total number of parallel augmented shortcut, Θ_i represents the i_{th} matrix, and LN means LayerNorm method. Then it will be processed by the $SIAF\_MLP$ block as Equation (9):

$$\begin{aligned} &SIAF\_MLP(V^{'}) = \\ &(\sum_{i=1}^{n}\sigma_i(V^{'}W_{1_i}^{'})+b_{1_i}^{'})W_2^{'}+b_2^{'} \end{aligned} \tag{9}$$

$$\begin{aligned} &E_{C\to G(P)} = \\ &LN(SIAF\_MLP(V_{G(P)}^{'})+V_{G(P)}^{'}) \end{aligned} \tag{10}$$

where σ_i indicates the i_{th} nonlinear activation function, while $W_{1_i}^{'}$ and $W_2^{'}$ are both matrices of linear layer. The $G(P)$ in the formula subscript corresponds to the calculation of the complementary part of the Graphics(G) or Pronunciation(P).

After sufficient feature interaction, we use a multi-modal gating mechanism to selectively pass the information represented by three different modalities E_C, $E_{C\to G}$ and $E_{C\to P}$ to dynamically adjust their ultimate contribution:

$$\begin{aligned} \begin{bmatrix} E_C^{'} \\ E_{C\to G}^{'} \\ E_{C\to P}^{'} \end{bmatrix} =& \tanh\left(\begin{bmatrix} E_CW_{f_C}+b_{f_C} \\ E_{C\to G}W_{f_{C\to G}}+b_{f_G} \\ E_{C\to P}W_{f_{C\to P}}+b_{f_P} \end{bmatrix}\right) \\ &+\begin{bmatrix} E_C \\ E_{C\to G} \\ E_{C\to P} \end{bmatrix} \end{aligned} \tag{11}$$

$$\begin{aligned} Z =&\sigma(Concat(E_C,E_{C\to G},E_{C\to P})W_Z \\ &+b_{f_Z}) \end{aligned} \tag{12}$$

$$E_{fusion_i}=Z_i\begin{bmatrix} E_{C_i} \\ E_{C\to Gi} \\ E_{C\to Pi} \end{bmatrix} \tag{13}$$

where W_{f_C}, $W_{f_{C\to G}}$, and $W_{f_{C\to P}} \in \mathbb{R}^{D_{BERT}\times D_{BERT}}$ are learnable parameter matrices, while b_{f_C}, b_{f_G}, and b_{f_P} are biases used to perform affine transformations on the three different modal

features. $W_Z \in \mathbb{R}^{3 \cdot D_{BERT} \times 3}$ calculates the contribution of the three modal features, resulting in $Z_i \in \mathbb{R}^3$. A weighted sum is then used to produce the output $E_{fusion_i} \in \mathbb{R}^{D_{BERT}}$ based on these contributions.

The fused features are subsequently fed into the W^{2}NER decoding module (Li et al., 2022) for word-word relation prediction and entity recognition.

4 Experiments

4.1 Experiments Settings

4.1.1 Datasets

We evaluate the performance of AHVE-CNER model using two public Chinese NER datasets: Weibo (Peng and Dredze, 2015) and Resume (Zhang and Yang, 2018). The Weibo dataset consists of social media posts, while the Resume dataset contains resume data from Sina Finance. We use the span method to calculate F1-score (F1), precision (P), and recall (R). Statistical information for these datasets is provided in Table 2.

Dataset	Types	Train	Dev	Test
Weibo	Sentence	1.35 k	0.27 k	0.27 k
	Character	73.78 k	14.51 k	14.84 k
	Entity	1.90 k	0.39 k	0.42 k
Resume	Sentence	3.8 k	0.46 k	0.48 k
	Character	622.96 k	67.72 k	77.21 k
	Entity	13.33 k	1.63 k	1.49 k

Table 2: Datasets details.

Hyper-parameter	Range
Warm up	[0.1, 0.2, 0.3]
Batch size	[2, 4, 8]
Pinyin CNN kernel size	[1, 2, 3]
learning rate	[8e-4, 3e-3]
Pinyin/Radical lr	[2e-4, 4e-4, 8e-4]
Pinyin/Radical emb size	[64, 128, 256]
Pinyin/Radical emb drop	[0.1, 0.2, 0.3]
Font style	[楷体, 仿宋, 黑体]
ViT heads	[4]
ViT layers	[6]
Visual alignment head	[4]
Aug-shortcut matrices num	[4, 8, 12]
SIAF activate functions	[Sig, ReLU, Tanh, LogSig]

Table 3: Parameters range.

4.1.2 Parameters

Radical and pinyin embeddings are randomly initialized. BERT's hidden size is 768, with a weight decay of 0.1. We employ 30 1-D convolution kernels for CNN, similar to MECT and VisPhone. Dropout rates are 0.3 for output and 0.5 for the fusion layer. Other parameters are detailed in Table 3. We used the SMAC algorithm to search for the best hyper-parameters. Experiments are performed on NVIDIA Tesla A40 48GB GPUs using PyTorch.

4.1.3 Baselines

We compare the following baselines: ChineseBERT (Sun et al., 2021) and StyleBERT (Lv et al., 2022), which incorporate additional Hanzi information during pre-training. GlyNN (Song and Sehanobish, 2020), Glyce (Meng et al., 2019), and GLexicon (Qi et al., 2023) use CNN structures to capture Hanzi visual information. CGR-NER (Gu et al., 2023) and FGN (Xuan et al., 2021) use 3D-CNNs to enhance interaction between adjacent graphics. MFE-NER (Li and Meng, 2021) and MECT (Wu et al., 2021a) convert Hanzi Wubi or radicals into visual information. VisPhone (Zhang et al., 2023) integrates Pinyin as additional phonetic features alongside radical sequences.

4.2 Main Results

We compare AHVE-CNER against several representative multi-modal Chinese NER models, alongside the commonly used BiLSTM+CRF model, which utilizes outputs from a BERT encoder. To ensure consistency across experimental settings, we adopt the same BERT-wwm model (Cui et al., 2020) as utilized in MECT and VisPhone.

Comparison results on the Weibo and Resume datasets are summarized in Table 4 and 5. Models are categorized into: (1) classic BERT models, (2) models incorporating Hanzi glyph information during BERT pre-training, and (3) models using CNNs for Hanzi images or encoding Hanzi visual information from radical sequences.

Weibo: Table 4 shows that adding LSTM and CRF to pre-trained BERT yields an F1 score of 67.12%. Integrating Hanzi glyph processing into BERT increases this to 70.80% (ChineseBERT), demonstrating the benefit of Hanzi visual information. Combining CNN-processed visual representations with BERT further improves the F1 score to 71.25% (FGN), outperforming Glyce and GlyNN. MECT and VisPhone both use SC dictionary-based visual information, with VisPhone achieving 70.79%, 0.36% higher than MECT due to addi-

tional Pinyin representation. AHVE-CNER surpasses all with an F1 score of 73.09%.

Resume: Table 5 shows similar trends. Models processing Hanzi images for visual information are generally more effective. FGN, using CNN for Hanzi images, achieves the highest F1 score of 96.79%, 1.01% higher than BERT with LSTM and CRF. VisPhone, with an F1 score of 96.26%, is 0.28% better than MECT. AHVE-CNER, combining Hanzi images and SC, achieves the highest F1 score of 97.02%.

Model	Precision	Recall	F1
BERT-BiLSTM-CRF	66.88	67.33	67.12
StyleBERT	-	-	69.60
MFE-NER	70.36	65.31	67.74
ChineseBERT	68.75	72.97	70.80
Glyce	67.68	67.71	67.60
GlyNN	-	-	69.20
CGR-NER	70.23	71.70	70.70
GLexicon+BERT	71.04	70.29	71.24
FGN	69.02	73.65	71.25
MECT	-	-	70.43
VisPhone	-	-	70.79
AHVE-CNER	**72.27**	**73.92**	**73.09**

Table 4: Results obtained on Weibo(%).

Model	Precision	Recall	F1
BERT-BiLSTM-CRF	96.12	95.45	95.78
StyleBERT	-	-	-
MFE-NER	95.76	95.71	95.73
ChineseBERT	-	-	-
CGR-NER	-	-	-
GlyNN	-	-	95.66
Glyce	96.62	96.48	96.54
GLexicon+BERT	96.46	96.11	96.72
FGN	96.49	97.08	96.79
MECT	-	-	95.98
VisPhone	96.09	96.44	96.26
AHVE-CNER	**96.81**	**97.23**	**97.02**

Table 5: Results obtained on Resume(%).

Compared to BERT+LSTM+CRF, the results demonstrate that both visual and pronunciation information significantly enhance named entity recognition. Incorporating this information externally rather than during BERT's pre-training phase proves more effective. AHVE-CNER outperforms all other multi-source information Chinese NER models on the Weibo and Resume datasets.

4.3 Ablation Results

To evaluate the contribution of each component and the effect of Hanzi visual feature alignment, we performed an ablation study. The experimental settings involve the following configurations: **(1)** removing the entire visual encoding structure; **(2)** retaining only radical-based SC encoding by omitting ViT-based graphics processing; **(3)** excluding the Pinyin representation component; **(4)** substituting the fusion mechanism with simple concatenation; **(5)** replacing the Pangu-π cross transformer with a standard transformer encoder; and **(6)** replacing the gated multi-modal unit (GMU) with simple concatenation, with adjustments to GMU based on the number of input modalities.

Method	Precision	Recall	F1
AHVE-CNER	**72.27**	**73.92**	**73.09**
-W/o Vision	71.92	73.13	72.52 (-0.57)
-W/o ViT	72.08	73.29	72.68 (-0.41)
-W/o Pinyin	71.62	73.35	72.48 (-0.61)
-W/o Fusion	71.74	73.252	72.49 (-0.60)
-W/o Pangu-π	72.00	73.38	72.68 (-0.41)
-W/o GMU	71.71	73.77	72.73 (-0.36)

Table 6: Ablation with Weibo(%).

Method	Precision	Recall	F1
AHVE-CNER	**96.81**	**97.23**	**97.02**
-W/o Vision	96.76	96.65	96.71 (-0.31)
-W/o ViT	96.57	96.97	96.77 (-0.25)
-W/o Pinyin	96.60	96.75	96.67 (-0.35)
-W/o Fusion	96.54	96.69	96.61 (-0.41)
-W/o Pangu-π	96.65	96.87	96.76 (-0.26)
-W/o GMU	96.84	96.97	96.90 (-0.12)

Table 7: Ablation with Resume(%).

Results obtained on Weibo and Resume datasets are shown in Table 6 and 7. The content in the tables is divided into several parts according to the relationship between components. Compare all those F1 scores, it can be seen clearly that: (1) both **visual and phonetic information** of Hanzi significantly improve word presentation from BERT resulting in **increasing 0.57%-0.61% on Weibo** and **0.31%-0.35% on Resume**. (2) **phonetic information** obtained from Pinyin tends to be **more**

influential than visual information as removing it causes more declines than removing visual encoder. (3) **Gated Pangu-π cross transformer** mostly influence the final feature presentation leads to **0.60% and 0.41% improvement** on Weibo and Resume datasets.

4.4 Effectiveness of Visual Alignment

To demonstrate the effectiveness of the Hanzi visual alignment method, we conducted comparative experiments by replacing visual coding components with those from alternative models. The experiments included: **(1)** applying a Wubi radical dictionary for Hanzi decomposition without incorporating ViT-based visual embeddings; **(2)** encoding radical embeddings without integrating image-based visual features; **(3)** utilizing CNNs to encode Hanzi images while excluding radical information; and **(4)** employing 3D-CNNs for visual embedding generation. The results, as shown in Tables 8 and 9, are annotated as follows: '♠' denotes radical visual embeddings without Hanzi image integration, '♣' represents image-only encoding without radical information, and 'Only Contextual' corresponds to BERT embeddings without multimodal fusion, analogous to the W^{2}NER baseline.

Module	Precision	Recall	F1
AHVE-CNER	72.27	**73.92**	**73.09**
Only Contextual	70.84	73.87	72.32
Wubi♠	71.70	73.57	72.62 (-0.47)
SC♠	72.08	73.29	72.68 (-0.41)
CNN♣	72.08	73.41	72.74 (-0.35)
3D-CNN♣	**72.32**	73.46	72.89 (-0.20)

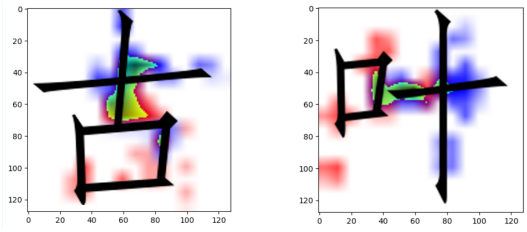
Table 8: Components Compare with Weibo(%).

Module	Precision	Recall	F1
AHVE-CNER	96.81	**97.23**	**97.02**
Only Contextual	96.96	96.35	96.65
Wubi♠	96.65	96.83	96.74 (-0.28)
SC♠	96.57	96.97	96.77 (-0.25)
CNN♣	96.85	96.83	96.84 (-0.18)
3D-CNN♣	**96.97**	96.86	96.91 (-0.11)

Table 9: Components Compare with Resume(%).

The results can be summarized as follows: **(1) Structure Component (SC)** encoding is superior to Wubi for visual information at the radical level on both datasets. **(2) 3D-CNN** provides better visual information extraction than simple CNN, with F1 score improvements of 0.15% and 0.07%, respectively. **(3) Using Hanzi** images for visual information outperforms radical sequence information alone, with gains of up to 0.27% on Weibo and 0.17% on Resume. **(4) Aligning Hanzi images** with radicals, as in AHVE-CNER, achieves the highest F1 scores, showing increases of 0.20% on Weibo and 0.11% on Resume compared to using only 3D-CNN. This demonstrates that the proposed Hanzi visual feature alignment method is currently the most effective for capturing pictographic features of Chinese characters.

We conducted a case study to visualize attention maps for two Chinese characters sharing identical SC sequences. As illustrated in Figure 5, the method successfully differentiates and aligns identical radicals in distinct Hanzi characters to their corresponding positions in the image, ensuring accurate integration of structural and visual features. It highlights the capability of the proposed method to effectively integrate pictographic information from Hanzi.



(a) Hanzi "古(ancient)". (b) Hanzi "叶(leaf)".

Figure 5: Case study: attention map of Hanzi pictographic visual feature alignment method. These two characters have exactly the same SC of "口(mouth)" and "十(ten)". Different colors represent attention of different radicals on the character and the yellow-green part is the attention overlap area.

4.5 Analysis in Efficiency

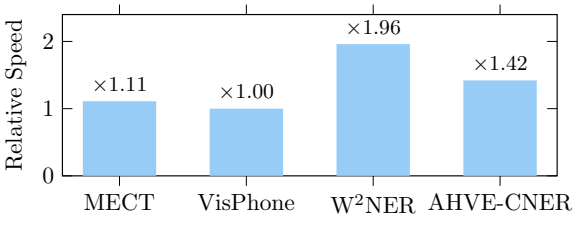


Figure 6: Relative inference speed of each model. The FLAT part in MECT and Visphone are set on non-parallel conditions.

We compare the non-parallel inference speed of MECT and Visphone with AHVE-CNER on a

NVIDIA Tesla A40 card, setting batch size = 4 and batch size = 1. We use the Visphone as the standard and calculate the other models' relative inference speed. The results are shown in Figure 6. The results indicate that MECT and VisPhone, both based on the FLAT structure, are slower than W2NER, which utilizes an LSTM structure, in non-parallel inference tasks. AHVE-CNER is 27% slower than W2NER, but it still outperforms VisPhone in terms of efficiency. However, the LSTM structure's inherent limitations prevent it from being optimized for parallel acceleration, leading to a slower reasoning speed compared to parallelable models.

5 Conclusion

This paper introduces AHVE-CNER, a novel Chinese NER model enhanced by visual information alignment. The proposed method integrates pronunciation and multi-source glyph information. Specifically, it employs a Vision Transformer (ViT) to process Hanzi images and aligns the extracted features with structural component encodings. The Gated PanGu-π Cross-Transformer is then used to fuse visual, pronunciation, and contextual information, with the PanGu-π module designed to enhance cross-modal interactions. Experimental results on the Weibo and Resume datasets demonstrate that multi-source alignment of Hanzi visual information significantly outperforms other visual feature extraction methods in Chinese NER. Future work will explore the integration of Chinese word-level information to further enhance model performance. Additionally, we aim to extend the application of AHVE-CNER to a broader range of Chinese NER datasets and other NLP tasks.

Acknowledgments

This article is the research result of the National Natural Science Foundation of China (No. 72074108) and the Special Fund for Basic Scientific Research Business of Central Universities project at Nanjing University, and is supported by the Jiangsu Young Talents in Social Sciences and Tang Scholar of Nanjing University.

References

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Preprint*, arXiv:1406.1078.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Ruiming Gu, Tao Wang, Jianfeng Deng, and Lianglun Cheng. 2023. Improving chinese named entity recognition by interactive fusion of contextual representation and glyph representation. *Applied Sciences*, 13(7):4299.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, volume 2019.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019b. A lexicon-based graph neural network for chinese ner. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1040–1050.

Xuchao Guo, Shuhan Lu, Zhan Tang, Zhao Bai, Lei Diao, Han Zhou, and Lin Li. 2022. Cg-aner: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition. *Computers and Electronics in Agriculture*, 194:106776.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Jiatong Li and Kui Meng. 2021. Mfe-ner: Multi-feature fusion embedding for chinese named entity recognition. *Cornell University - arXiv,Cornell University - arXiv*.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021a. Lexicon enhanced chinese sequence labeling using bert adapter. *arXiv preprint arXiv:2105.07148*.

Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and Philip S Yu. 2021b. Dense hierarchical retrieval for open-domain question answering. *arXiv preprint arXiv:2110.15439*.

Chao Lv, Han Zhang, XinKai Du, Yunhao Zhang, Ying Huang, Wenhao Li, Jia Han, and Shanshan Gu. 2022. Stylebert: Chinese pretraining by font style information. In *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 10, pages 646–652. IEEE.

Ruotian Ma, Minlong Peng, Qi Zhang, and Xuanjing Huang. 2019. Simplify the usage of lexicon in chinese ner. *arXiv preprint arXiv:1908.05969*.

Chengcheng Mai, Jian Liu, Mengchuan Qiu, Kaiwen Luo, Ziyan Peng, Chunfeng Yuan, and Yihua Huang. 2022a. Pronounce differently, mean differently: a multi-tagging-scheme learning method for chinese ner integrated with lexicon and phonetic features. *Information Processing & Management*, 59(5):103041.

Chengcheng Mai, Mengchuan Qiu, Kaiwen Luo, Ziyan Peng, Jian Liu, Chunfeng Yuan, and Yihua Huang. 2022b. Pretraining multi-modal representations for chinese ner task with cross-modality attention. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 726–734.

Pedro Henrique Martins, Zita Marinho, and André FT Martins. 2019. Joint learning of named entity recognition and entity linking. *arXiv preprint arXiv:1907.08243*.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. *Advances in Neural Information Processing Systems*, 32.

Xue Mengge, Bowen Yu, Tingwen Liu, Yue Zhang, Erli Meng, and Bin Wang. 2020. Porous lattice transformer encoder for chinese ner. In *Proceedings of the 28th international conference on computational linguistics*, pages 3831–3841.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Pengnian Qi, Peng Li, and Biao Qin. 2023. Glexicon: Glyph and lexicon-based embedding model for chinese ner. In *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 5–10. IEEE.

Arijit Sehanobish and Chan Hee Song. 2019. Using chinese glyphs for named entity recognition. *arXiv preprint arXiv:1909.09922*.

Chan Hee Song and Arijit Sehanobish. 2020. Using chinese glyphs for named entity recognition (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13921–13922.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*.

Yunhe Wang, Hanting Chen, Yehui Tang, Tianyu Guo, Kai Han, Ying Nie, Xutao Wang, Hailin Hu, Zheyuan Bai, Yun Wang, et al. 2023. Pangu-π: Enhancing language model architectures via nonlinearity compensation. *arXiv preprint arXiv:2312.17276*.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. 2021a. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. *arXiv preprint arXiv:2107.05418*.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. 2021b. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Zhenyu Xuan, Rui Bao, and Shengyi Jiang. 2021. Fgn: Fusion glyph network for chinese named entity recognition. In *Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence: 5th China Conference, CCKS 2020, Nanchang, China, November 12–15, 2020, Revised Selected Papers*, pages 28–40. Springer.

Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep affix features improve neural named entity recognizers. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 167–172.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2022. Gnner: Reducing overlapping in span-based ner using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 97–103.

Baohua Zhang, Jiahao Cai, Huaping Zhang, and Jianyun Shang. 2023. Visphone: Chinese named entity recognition model enhanced by visual and phonetic features. *Information Processing & Management*, 60(3):103314.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.