文章编号：2096-1383（2025）03-0258-05

基于仿生视觉聚焦的西夏文识别网络

敖 敏[a,b]，王存睿[a,b]
（大连民族大学 a. 计算机科学与工程学院；b. 大连市汉字计算机字库设计技术创新中心，辽宁 大连 116605）

摘 要：为解决大量的西夏文古籍没有被数字化的问题，提出了基于仿生视觉聚焦的西夏文识别网络。使用四角号码对西夏文单字标注的方法构建西夏文数据集，通过使用该数据集进行西夏文单字识别的实验与分析。该网络通过引入仿生视觉聚焦注意力机制，改善Swin-Transformer因滑动窗口注意力模块产生的伪影情况，提高西夏文的识别准确度。引入卷积GLU门控模块作为Swin-Transformer的前馈网络层，增强模型的鲁棒性，提高对于西夏文古籍单字的识别准确率。实验结果表明：该网络对西夏文的识别正确率达94.54%，优于其他网络，解决了西夏文在实际应用场景中的识别问题。自建的西夏文数据集也为后续的西夏文相关工作提供了数据支持。

关键词：西夏文；Swin-Transformer；仿生视觉聚焦注意力；卷积GLU
中图分类号：TP391　　**文献标志码：**A
DOI:10.13744/j.cnki.cn21-1431/g4.2025.03.009

Bio-Inspired Visual Attention-Based Tangut Character Recognition Network

AO Min[a,b]，WANG Cunrui[a,b]
（a. School of Computer Science and Engineering; b. Dalian Chinese Character Computer Font Design Technology Innovation Center, Dalian Minzu University, Dalian Liaoning 116605, China）

Abstract: To address the issue that a large number of extant Tangut ancient books have not been digitized, this paper proposes a Tangut script recognition network based on biomimetic visual focus. A Tangut script dataset is constructed by annotating individual Tangut characters using the four-corner numbering method. Experiments and analyses are conducted on the recognition of single Tangut characters using this dataset. The proposed network introduces a biomimetic visual focus attention mechanism to mitigate artifacts caused by the sliding window attention module in Swin-Transformer, thereby improving the recognition accuracy of Tangut script. Additionally, a convolutional GLU gating module is incorporated as the feed-forward network layer in Swin-Transformer to enhance the model's robustness and improve the recognition accuracy of individual characters in Tangut ancient texts. Experimental results demonstrate that the network achieves a recognition accuracy of 94.54% for Tangut script, outperforming other networks in recognition capability and addressing the challenges of Tangut script recognition in practical application scenarios. Furthermore, the self-constructed Tangut script dataset provides valuable data support for subsequent research related to Tangut script.

Key words: Tangut; Swin-Transformer; biomimetic visual focus attention; convolutional GLU

西夏文又称河西文字、番文，是记录西夏党项族语言的文字，属表意体系。西夏文结构仿汉字但又有其特点，其主要用点、横、竖、撇、捺、拐、拐钩等组字，斜笔较多，没有竖钩。单纯字

收稿日期：2024-11-29；**最后修回日期：**2025-03-18
基金项目：大连市创新基金项目（2023JJGX026）；民族教育信息化教育部重点实验室联合基金项目（EIN2024B002）。
作者简介：敖敏（1999-），女，达斡尔族，内蒙古呼伦贝尔人，大连民族大学计算机科学与工程学院硕士研究生，主要从事字体辅助设计研究。

较少，合成字占绝大多数[1]。西夏文不仅是中华民族历史文化遗产的重要组成部分，更是全人类共享的历史瑰宝。国际学术界普遍认同，深入研究西夏文化，对于理解中华民族历史的多元性和丰富性，以及人类文明的交流与交融，具有无可估量的学术价值。

西夏文字符在 2016 年 6 月收录进第 9.0 版 Unicode 编码中，采用四角号码也就是字符左上，左下，右上，右下的笔画号码作为输入来确定是哪一个西夏文字符如图 1。

图 1　西夏文四角号码

1　研究现状

从形制与数量来看，西夏文字符与汉字字符都是方形的字符，但西夏文字符的笔画结构更加复杂，大部分的笔画都在 14 划以上，并且西夏文包含了更多的曲线、角度和组合的状态，这都使得当前对于西夏文的识别比较困难。现阶段对于西夏文的识别研究所使用的数据集多是通过裁切部分古籍并进行人工标注。这也使得数据量无法覆盖全部西夏文字，并且数据分布不均匀。而且古籍上截取的西夏文字符会有模糊、不完整等问题。

按照识别技术划分可知，西夏文识别方法主要分为机器学习与深度学习两个阶段。第一阶段的研究多采用传统的机器学习方法对西夏文进行识别研究，如孟一飞[2]采用基于 ASM 算法对西夏文的必行特征进行提取基于结构进行识别，这种方法对于简单的字体识别较快，但是复杂字体的识别较为困难。刘兴长[3]采用 HOG 特征提取及模糊支持向量机进行识别，该方法适合小样本数量的学习识别，有着较好的鲁棒性但不适合大量样本数据，多分类的训练识别。孟一飞[4]使用 MLAE 和 MLP 对西夏文识别工作进行研究，这类神经网络方法有着较好的自学习能力，有着较高的容错率，但训练收敛速度慢，而且会出现过拟合的情况。

随着深度学习的发展，西夏文的识别也进入了第二阶段。杨文慧[5]、ZHANG G[6]、刘佟[7]、张光伟[8]等分别 CNN、DCNN、DenseNet、GoogleNet 等经典 CNN 网络进行识别，这种方法无须手动提取西夏文特征信息，效果好，分类效率高。但当样本量大的时候就需要调参。ZHANG G[9]、MA J[10]采用 TCRNet、STCRNet 等基于生成模型端到端识别进行研究。这种方法有效地解决了样本数量小、分布不均匀的问题，但对于网络模型的可解释性不好。张光伟采用基于 RNN 的方法进行识别工作，该方法有着较好的记忆性，无限制的输入长度，同时拥有 CNN 的较好的效果，但训练速度较慢，且容易发生梯度消失或爆炸。

综上所述，现如今的西夏文识别大多使用深度学习分类网络进行，因为互联网上缺少公开的字符量全及分布均匀的西夏文数据集，因此西夏文的识别还有较大的进步空间。

2　研究方法

为解决西夏文字符的识别问题，本文将建立一个来源可靠、标注准确的西夏文数据集，根据这一数据集提出基于仿生视觉聚焦的西夏文识别网络如图 2。该网络采用 Swin-Transformer 为基础识别网络。通过对西夏文数据集图像分块传入 Transformer 的自注意力学习机制中，利用 Transformer 网络的全局感受特性对西夏文数据集进行学习，利用滑动窗口多头自注意力机制（Windows Multi-head Self-Attention，W-MSA）增强西夏文不同块之间的交互达到全局建模的能力，返回识别后对应的结果。

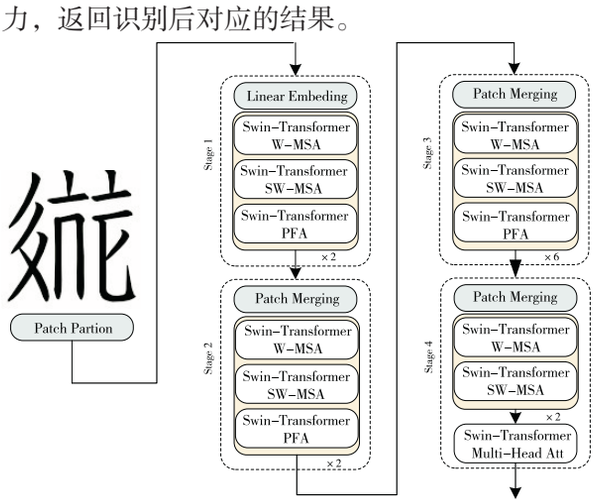

图 2　模型整体架构图

由于西夏文数据量不大，所以采用 Swin-Transformer-Base 进行识别工作，过深的网络结构不但会增加训练的参数量，消耗更长的时间，

同时结果的准确率也不会进一步提升。

为了能够更好地提升模型对于西夏文的识别性能，增强模型对于西夏文全局特征的学习，引入仿生视觉聚焦模块，以全局坐标信息的关注度来提取西夏文的全局特征。为了增强模型的鲁棒性，提升对于多情况下西夏文的识别准确度，引入卷积门控注意力单元（Convolutional Gated Linear Unit，CGLU），使模型更好地捕捉每个标记具有特殊门控信号，增强改进后的 Swin-Transformer 具有更高的稳健性。

2.1 仿生视觉聚焦融合模块

Swin-Transformer 是对图像文字进行分块识别，对文字进行切割自学习，这导致缺少块与块之间的学习。哪怕用过滑动窗口的模式进行块之间的通信，无论叠多少层都会出现难以消除的块级伪影如图 3。并且由于西夏文字符单字笔画较多且密集，所以这对西夏文的识别产生影响。

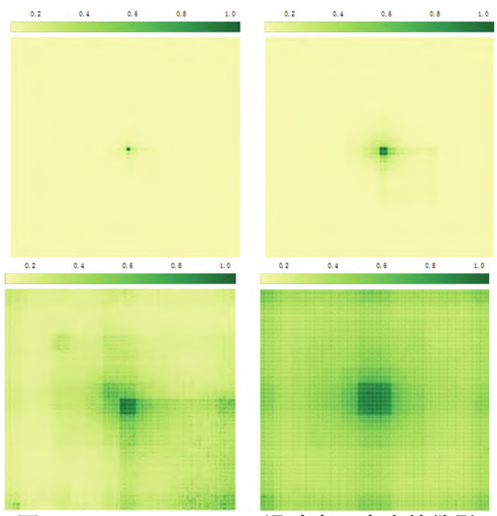

图 3 Swin-Transformer 滑动窗口产生的伪影

本文引入基于仿生视觉的聚焦注意力机制（pixel-focused attention，PFA），在滑动窗口层之后增加带有 PFA 的 Transformer 层。通过 PFA 的全局感知特点，对分块识别产生的全局感知能力弱问题进行改善。提高模型对于西夏文笔画的识别与其对于全局所处位置的感知能力。

为了模拟这一特性，该机制将分为两个路径进行实现如图 4。

给定输入图像为$X \in R^{C\times H\times W}$，对于聚焦中心点将以像素为中心的滑动窗口内的一组像素定义为 $\rho(i, j)$，中心聚焦实像的固定窗口大小为 $k\times k$，因此滑动窗口对应的公式为

$$\|\rho(i,j)\| = k^2 。 \tag{1}$$

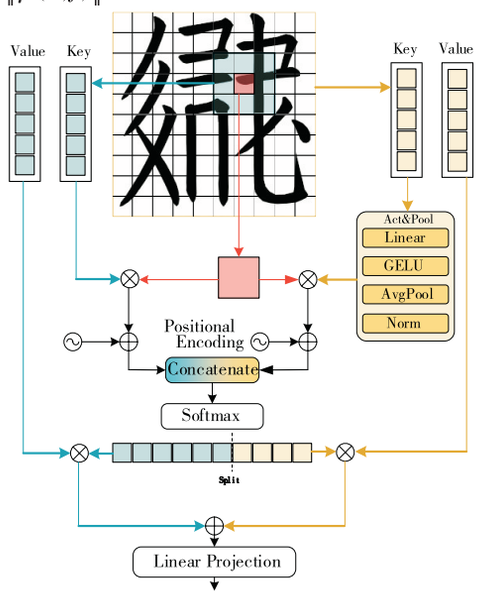

图 4 仿生视觉的聚焦注意力模块

同时，对于焦外虚像部分，定义其特征映射池化为$\sigma(X)$，焦点外窗口大小为 $H_p\times W_p$，因此焦点外池化窗口的公式为

$$\|\sigma(X)\| = H_p W_p 。 \tag{2}$$

焦点内窗口通过降维展平，生成对应的 Key 与 Value 用于后续与焦点所在窗口 Query 进行自注意力操作。首先将 Key 值$K^T_{\rho(i,j)}$与焦点 Query 值$Q_{(i,j)}$进行向量积操作，生成焦内与焦点的特征权重$S_{(i,j)\sim\rho(i,j)}$，对应公式为

$$S_{(i,j)\sim\rho(i,j)} = Q_{(i,j)} K^T_{\rho(i,j)} 。 \tag{3}$$

焦点外窗口通过激活池化层进行下采样，压缩提取有用的信息，从而提高下采样后的信息压缩率，将压缩后的信息降维展平生成其对应的 Key 与 Value 用于与 Query 进行自注意力操作。将焦点外生成的 Key 值$K^T_{\sigma(i,j)}$与焦点生成的 Query 值$Q_{(i,j)}$进行向量积操作生成焦外虚像与焦点的特征权重$S_{(i,j)\sim\sigma(i,j)}$，对应的公式为

$$S_{(i,j)\sim\sigma(i,j)} = Q_{(i,j)} K^T_{\sigma(i,j)} 。 \tag{4}$$

将生成后的$S_{(i,j)\sim\rho(i,j)}$与$S_{(i,j)\sim\sigma(i,j)}$添加位置编码后进行拼接，与可训练矩阵 $B_{(i,j)}$ 相加，并通过 Softmax 进行归一化处理，使焦点内外图像得到融合，并使模型更好地学习焦点内外的特征重要性，公式为

$$A_{(i,j)} = \mathrm{softmax}\left(\frac{\mathrm{Concat}(S_{(i,j)\sim\rho(i,j)}, S_{(i,j)\sim\sigma(X)})}{\sqrt{d}} + B_{(i,j)}\right) 。 \tag{5}$$

将归一化后的融合特征按照 k^2，$H_p \times W_p$ 的尺寸进行分割生成$A_{(i,j)\sim\rho(i,j)}$，$A_{(i,j)\sim\sigma(X)}$，分别用于与焦点内外的 Value 值$V_{\rho(i,j)}$和$V_{\sigma(X)}$进行向量相乘并相加，得出 PFA 注意力特征权重 $PFA(X_{(i,j)})$ 为

$$A_{(i,j)\sim\rho(i,j)}, A_{(i,j)\sim\sigma(X)} = \text{Split}(A_{(i,j)}) \text{with size}[k^2, H_p W_p]; \tag{6}$$

$$PFA(X_{(i,j)}) = A_{(i,j)\sim\rho(i,j)} V_{\rho(i,j)} + A_{(i,j)\sim\sigma(X)} V_{\sigma(X)} 。 \tag{7}$$

在滑动窗口路径中，位于特征图边缘的像素点不可避免地在边界外以零填充计算相似度。为了防止这些零相似度影响 softmax 归一化操作，使用填充掩码将这些结果设置为 $-\infty$。

PFA 的计算复杂度公式为

$$\Omega(PFA) = 5HWC^2 + 2H_p W_p C^2 + 2HWH_p W_p C + 2HWk^2 C 。 \tag{8}$$

2.2　卷积门控注意力模块

前馈神经网络（Feed-Forward Network，FFN）作为 Transformer 的重要组成部分，其作用是能作为 channel-mixer 为增强对模型的特征提取能力。由于西夏文字符复杂的字形设计，在古籍中难免出现的变形破损等问题，就更需要模型有强大的鲁棒性，加强对于西夏文重要特征的提取能力，因此引入 CGLU 门控机制作为 Swin-Transformer 的 FFN 层如图 5。增强其对于重要特征的筛选能力。

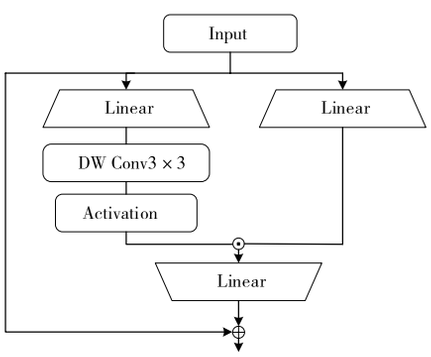

图 5　CGLU 门控单元

CGLU 通过在激活函数前添加最小的深度可分离卷积，使其实现符合最近邻特征的注意力机制的门控单元。该门控通道使每个标记具有自己独特的门控信号同时能够捕捉来自零填充的位置信息，满足对于西夏文位置编码重要性的需求。

3　数据集

为使识别网络更加全面地对西夏文特征进行学习，选取西夏文 TTF 字库包作为数据集来源。

3.1　数据制备

通过解包并逐字将其转化为“.png”格式的图片。通过 rime 输入法中的四角号码与西夏文 Unicode 对应表，对西夏文字符进行标注，使每一个西夏文字符名称都是所对应的四角号码。

西夏文现共有 6 145 个字符，通过四角号码对数据集进行分类，得到 3 103 种西夏文字符。

3.2　数据扩增

对预处理后的西夏文数据集进行翻转、旋转、变暗、遮挡等操作。

增强后的部分效果如图 6。增强后最终的西夏文图像数据共计 122 900 个字符样本量，单张图像的宽 × 高为 224 × 224。

图 6　数据增强后的效果

4　实验与结果分析

为了评测本文模型对于西夏文的识别能力。采用自建的西夏文数据集与现有先进识别网络进行对比实验。

实验软硬件平台包括：CPU 为 Intel i7-9700k，GPU 为 NVIDIA GeForce Titan，操作系统为 Ubuntu 22.16，深度学习框架为 PyTorch。

4.1　数据集设置

为了提高模型的训练效果，将西夏文图像数据的 80% 用作训练集，20% 用作测试集进行模型训练和检测。验证方法采用随机增强后的西夏文数据集作为验证集，以用于评估模型在识别西夏文的能力。由于西夏文数据量较小，选用 Swin-Transformer-Base 版本进行西夏文特征学习。对于 Swin-Transformer-Base 各层的超参数设置基本一致。

4.2　结果与分析

为了验证本文算法的最优性与可靠性，本文使用上述处理的西夏文数据集分别对 EfficientNet、DenseNet、Resnet-50、MobileNet、Vision Transformer 与本实验设计的模型分别进行训练，对比以上算法的预测正确率，对比结果见表 1。

表1 实验结果对比

模型	正确率/%
DenseNet	82.00
Resnet-50	81.79
Vision Transformer	87.00
EfficientNet	90.48
MobileNet	86.68
Our	94.54

从表1的实验结果可以看出，本文提出的方法在与其他四种算法的比较中，获得了最高的识别正确率94.54%。与传统的Vision Transformer算法相比，本文模型的识别正确率提高了7.54%。实验中发现，Vision Transformer训练时间过长，需要超其他训练方式一倍多的时间才能完全收敛。

本实验对《大方广佛华严经》西夏文版进行了裁切，随机选取其中的西夏文字符，进行裁切、去噪、二值化等处理，输入进模型进行预测，并人工判断正确率，显示模型对古籍上西夏文单字的识别正确率为84%，显示出较高的泛化能力，可见本模型有着较高的鲁棒性。

4.3 消融实验

本文中模型用于西夏文识别的消融实验结果见表2。其中表中模块A为改进后的增加PFA的模型，模块B为CGLU门控模块。

表2 消融实验结果对比

网络结构	正确率/%
Swin-Transformer	91.19
Swin-Transformer +A	93.26
Swin-Transformer +B	91.80
Swin-Transformer + A+B	94.54

结果表明，原模型加入A模块后，西夏文的识别正确率提升了1.36个百分点，这证明了A模块在提升西夏文识别正确率方面具有显著效果，更改的注意力机制能够更好地识别和学习西夏文对于重点特征。加入B模块后，西夏文的识别正确率提升了0.61个百分点，但与A模块同时相加，西夏文识别正确率提升2.64%，可见模块B配合模块A可以获得更高的正确率。此外还对古籍西夏文进行加B模块前后的测试，正确率提高了8%，可见B模块对模型的鲁棒性有了较好的提升。

5 结 论

针对西夏文识别问题，提出的基于仿生视觉聚焦的西夏文识别网络，通过修改Swin-Transformer-Base在滑动窗口注意力之后添加PFA注意力机制，并改进FFN层添加CGUL门控模块，显著提升了对西夏文的识别能力，并增强了模型的鲁棒性，能够更好地识别为训练过的古籍西夏文。通过消融实验和对比实验，验证了本方法相较于其他方法的有效性和优势，显示出较大的实用性。该方法不仅对西夏文识别具有重要意义，也为更了解中华文化保护古籍并加快西夏文数字化提供了技术支持，是古文识别领域具有前景的方法。

参考文献

[1] 王云庆，唐敏. 论西夏王陵的遗产价值与申遗之路[D]. 济南：山东大学历史文化学院，2013.

[2] 孟一飞，张晓彪，杨小花. 基于ASM算法的特征提取与匹配在文字识别中的应用[J]. 广西大学学报（自然科学版），2017，42（6）：2183-2190.

[3] 刘兴长，孟昱煜. 基于HOG特征提取和模糊支持向量机的西夏文字识别[J]. 西北师范大学学报(自然科学版)，2019，55（5）：39-43.

[4] 孟一飞. 西夏文字数字信息化若干问题研究[D]. 北京：北京交通大学，2019.

[5] 杨文慧. 西夏古籍文字样本数据库的创建及应用技术研究[D]. 银川：宁夏大学，2018.

[6] ZHANG G，HAN X.Deep learning based tangut character recognition[C]//proc of the 2017 4th International Conference on Systems and Informatics（ICSAI），HangZhou：IEEE，2017：437-441.

[7] 刘佟. 基于深度学习的西夏文古籍文献识别研究与实现[D]. 银川：宁夏大学信息工程学院，2022.

[8] 张光伟. 基于深度学习的西夏文献数字化[J]. 西夏学，2020（2）：206-213.

[9] ZHANG G，ZHAO Y.Target-directed mixup for labeling tangut characters[C]//proc of the 2019 International Conference on Document Analysis and Recognition（ICDAR），Sydney：IEEE，2019：202-207.

[10] MA J，CAO Y，MA Z，etal.End-to-end tangut character database building and recognition method[J].Iet Image Processing，2022，16（8）：2087-2100.

（责任编辑 王楠楠）