计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(11):139−150 [doi: 10.15888/j.cnki.csa.009983] [CSTR: 32024.14.csa.009983]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

并行大模型驱动的多模态骨签文物分类①

范　涛[1], 王慧琴[1], 王　可[1], 刘　瑞[2], 王　展[3], 毛　力[1]

[1](西安建筑科技大学 信息与控制工程学院, 西安 710055)
[2](中国社会科学院 考古研究所, 北京 100101)
[3](陕西省文物保护研究院, 西安 710075)
通信作者: 王慧琴, E-mail: hqwang@xauat.edu.cn

摘　要: 汉朝未央宫遗址出土的约 6 万片骨签碎片中, 约 5.7 万片刻有释文, 多数骨签在出土时呈纵向断裂状态, 导致其上下部分分离, 对文物的数字化保护及系统化分类工作带来了挑战. 传统人工分类方法不仅效率低下且可能对骨签造成进一步的损伤, 为提升骨签文物的分类精度, 为后续考古研究提供支持, 本文提出了一种融合骨签图像与释文信息的并行多模态分类模型. 该方法采用 Vision-RWKV 大模型提取骨签图片的视觉特征, 利用 RWKV 大模型提取骨签上的释文信息, 通过动态交叉特征融合模块整合图像与文本特征, 并引入分类器进行精细化分类. 实验结果表明该方法达到了 92.85% 的准确率, 显著优于传统深度学习模型和其他多模态大模型. 研究成果为骨签文物的高效分类与整理提供了有力的技术支撑, 并为考古领域的智能化研究奠定了重要基础.

关键词: 骨签; 文物分类; 特征提取; 图像分类

引用格式: 范涛,王慧琴,王可,刘瑞,王展,毛力.并行大模型驱动的多模态骨签文物分类.计算机系统应用,2025,34(11):139–150. http://www.c-s-a.org.cn/1003-3254/9983.html

Parallel LLM-driven Multimodal Classification for Bone Stick Artifacts

FAN Tao[1], WANG Hui-Qin[1], WANG Ke[1], LIU Rui[2], WANG Zhan[3], MAO Li[1]

[1](College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)
[2](Institute of Archaeology, Chinese Academy of Social Sciences, Beijing 100101, China)
[3](Shaanxi Institute for the Preservation of Cultural Heritage, Xi'an 710075, China)

Abstract: Among the approximately 60 000 bone stick fragments excavated from the Weiyang Palace ruins of the Han Dynasty, around 57 000 are inscribed. Most bone sticks exhibit longitudinal fractures at the time of excavation, resulting in the separation of their upper and lower parts. This fragmentation poses significant challenges to the digital preservation and systematic classification of these cultural artifacts. Traditional manual classification methods are inefficient and may cause further damage to the bone sticks. To address these challenges, this study proposes a parallel multimodal classification model that integrates both bone stick images and inscription information. Visual features are extracted from the images using the Vision-RWKV large-scale model, while textual features are obtained from the inscriptions via the RWKV model. A dynamic cross-modal feature fusion module is introduced to integrate image and text features, followed by a classifier for fine-grained categorization. Experimental results demonstrate that the proposed method achieves an accuracy of 92.85%, significantly outperforming conventional deep learning models and other multimodal approaches. This study provides a robust technical foundation for the efficient classification and organization of bone stick artifacts and establishes a solid basis for the intelligent development of archaeological research.

Key words: bone stick; artifact classification; feature extraction; image classification

① 基金项目: 国家社科基金冷门绝学研究专项 (20VJXT001)
收稿时间: 2025-03-20; 修改时间: 2025-04-11, 2025-04-30; 采用时间: 2025-05-12; csa 在线出版时间: 2025-09-18
CNKI 网络首发时间: 2025-09-22

1 引言

汉长安城未央宫遗址出土了大量断裂的骨签文物, 骨签是由动物的肩胛骨加工而成, 呈长条形的签牌结构, 出土骨签中刻有释文的超过 5.7 万枚[1], 释文内容涵盖了整个西汉时期的官吏名称、武器类别以及计量单位等信息, 生动再现了西汉时期的社会结构与发展历程[2], 骨签对考古学研究具有重要的学术价值和文化意义. 然而骨签经历长期掩埋, 受泥土侵蚀流动与自身结构的影响, 大多数骨签出土时上下断裂并散落于不同的位置. 考古工作者主要依靠人工对骨签进行系统性分类整理, 此方法极为费时耗力, 同时在操作过程中可能会对骨签造成二次损伤. 将数字图像处理方法应用于骨签文物的分类, 可以显著提升分类效率.

随着深度学习与大模型技术的迅速发展与迭代, 当前图像分类领域已取得诸多重要进展. 例如以深度学习为代表的 YOLO (you only look once) 系列模型在图像分类任务中展现出优秀的性能, 特别是在对实时性要求较高的应用场景中被广泛采用. 诸多研究[3−10]在 YOLO 框架的基础上针对主干网络进行了改进, 以提升其在特定任务中的分类能力. 然而在骨签文物图像分类任务中, 由于 YOLO 等基于深度学习的模型主要依赖卷积神经网络进行特征提取, 其架构虽擅长捕捉全局目标及较大尺度的局部特征, 但在细粒度信息的提取方面存在一定局限性, 导致分类精度受限. 与 YOLO 等传统深度学习方法不同, CLIP (contrastive language-image pretraining) 等多模态大模型[11,12]通过联合学习文本与图像特征实现跨模态对齐, 并在零样本学习及开放词汇分类任务中展现出优异的泛化能力, 然而 CLIP 等多模态模型的核心架构通常基于 Transformer 架构进行特征提取, 该方法计算复杂度较高, 在处理大规模骨签文物图像数据时, 该模型的计算资源消耗极为庞大, 此外 CLIP 的图像编码部分主要采用基于自注意力机制的 ViT (vision Transformer) 结构, 该方法虽在全局特征建模方面表现突出, 但在捕捉细粒度的局部特征方面能力有限. 在骨签分类任务中, 这一特性导致其对细节特征的信息提取不足, 从而影响分类精度.

为应对上述挑战, 本文提出了一种基于多模态大模型并行驱动的骨签文物分类方法, 通过提取多模态特征信息, 增强模型在特征提取过程中的表现力, 并通过动态交叉特征融合策略, 利用特征间的非线性交互建模, 提高分类精度的同时提升模型的泛化能力. 该方法采用骨签的图像信息以及骨签上的释文信息作为模型的输入信息, 首先采用 Vision-RWKV[13]和 RWKV[14]大模型, 从不同模态的数据中分别提取特征, 而后利用动态交叉融合模块对各类特征进行有效融合, 最终通过分类器进行精细化分类. 采用对比学习方法, 通过迁移学习将预训练的大规模模型权重初始化到骨签分类任务中, 从而加速模型在该任务上的学习过程, 随后使用少量数据对网络进行微调, 进一步提升模型的泛化能力. 本文的主要贡献如下.

- 并行多模态分类方式的搭建: 为确保各类信息特征的充分提取, 兼顾系统整体的鲁棒性与模型运行效率, 本文提出了一种基于两种大模型并行驱动的集成分类系统, 提升特征提取能力以充分挖掘骨签数据的潜在信息. 通过依据数据特征的不同, 分别使用各自对应的大模型进行特征提取, 为骨签文物分类提供更为精确的分类依据.
- 图像数据预处理与噪声抑制: 为提升模型性能与泛化能力, 本文对图像数据集进行了高效预处理, 利用 DIS-Net 精确去除背景区域来降低噪声干扰对模型训练的影响.
- 动态交叉特征融合机制: 引入动态交叉特征融合机制, 利用交叉融合不同来源的特征信息, 使得分类模型能够从多角度、多维度对输入信息特征进行深层次理解.
- 精细化分类器的架构: 本文采用基于卷积层与池化层构建的精细化 CNN 分类器, 精准识别骨签图像的类别.

2 基本理论

2.1 RWKV

RWKV 是一种新型神经网络架构的大模型, 模型架构结合了 RNN[15]和 Transformer[16]的特点, RWKV 引入加权的键值对机制[17]来捕捉时间序列的上下文信息, 能够更好地捕捉不同时间步之间的关系, 在特征提取时保持高效性. RWKV 通过加权的方式调整记忆的长短, 优先保留与当前时间步相关的重要信息, 能更有针对性地提取特征. 同时, RWKV 架构中信息的流动方式使其能够更加高效地将信息从过去的时间步传播到当前时间步, 有助于提取更为精细的文字特征.

RWKV 模型以堆叠的残差块为基础构建, 如图 1(a)

所示. 每个残差块由时间混合子块与通道混合子块组成, 能够在特征维度和时间维度上同时建模相互依赖, 体现了递归结构的优势, 模型能够充分利用历史信息以提升当前时间步的特征提取能力. 在模型中独特的注意力操作使得分更新机制被引入, 特别是时间依赖的 Softmax 操作, 显著增强了数值稳定性, 缓解了梯度消失问题, 此机制可以使模型能够以更精细的粒度控制不同时间步之间的信息交互, 从而更加精准地捕获关键特征. 同时模型结合了层归一化技术[18], 对于稳定梯度流动至关重要, 也有效地解决了深度网络训练过程中常见的梯度消失和梯度爆炸问题. 稳定性不仅优化了模型的训练过程, 还进一步支持了深层网络的堆叠, 使模型能够捕获不同层次的抽象特征, 提取更加复杂的特征信息.

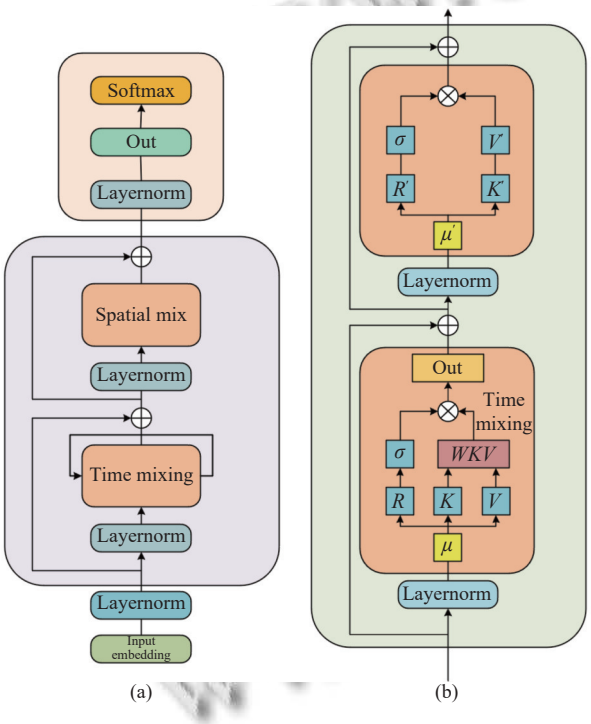


图 1 RWKV 架构图

RWKV 模型的核心架构由 4 个基本要素组成, 如图 1(b) 所示. 时间混合与通道混合模块的固有组成部分为: R (receptance vector): 作为接受向量, 用于接收来自历史时间步的信息; W (Weight): 表示位置权重衰减向量, 作为模型中的可训练参数; K (Key): 功能上类似于传统注意力机制中的“键”; V (Value): 承担传统注意力机制中的“值”的功能. 这些核心组件通过乘法交互在每个时间步中进行动态信息更新, 这种设计使得模型能够有效融合和提取时序信息, 同时保留了深度网络的表达能力和计算效率.

2.2 Vision-RWKV (VRWKV)

Vision-RWKV (VRWKV) 是将 RWKV 的注意力机制进行了修改以适应视觉任务, 与 CNN[19]和 ViT[20]相比采用了简化的注意力机制, 减少了计算复杂度, 并优化了高分辨率图像的处理能力, 捕捉图像的局部特征 (如边缘、纹理) 和全局信息 (如物体的上下文关系), 这使得其在处理复杂图像时能够提取更为精确的细节特征.

VRWKV 的整体架构包括L个相同的编码器层、平均池化层以及线性预测头, 如图 2(a) 所示. 每个 VRWKV 编码器层由两个主要模块构成: 空间混合模块和通道混合模块. 其中空间混合模块充当注意力机制, 执行具有线性复杂度的全局注意力计算.

通道混合模块则充当前馈网络[21], 在通道维度上进行特征融合, 如图 2(b) 所示. 该设计使得模型能够高效地整合来自不同卷积核或特征通道的表征信息, 从而增强了对图像多样化特征的提取能力, 提升了模型在复杂视觉任务中的表现. 在保留 RWKV 模型原有优势的基础上, VRWKV 对传统的注意力机制进行了多方面的优化, 具体包括双向注意力机制、相对偏差以及灵活衰减等设计.

双向注意力的引入使模型在进行图像特征提取时, 能够全面考虑图像中不同像素或区域之间的相互作用. 这极大地增强了模型在捕捉图像全局语义信息方面的能力, 模型能够更加细致地理解图像的全局结构和细节. 相对偏差的引入则帮助模型在处理具有不同尺寸的图像时, 灵活地调整对像素或区域间空间关系的捕捉能力, 这一设计特别有助于处理各种尺度的图像数据, 增强了模型的跨尺度特征提取能力. 灵活衰减机制的设计不仅强化了模型的全局注意力计算能力, 还能够在较大范围内捕捉图像的远程依赖关系, 通过自适应的衰减策略, 模型能够在计算注意力时, 聚焦于离当前标记较远的图像区域, 优化了远程特征的建模, 进一步提升了图像特征的深度表达.

2.3 模型选择

RWKV 与 VRWKV 作为近年来不断迭代和优化的大模型家族, 其不同版本在参数规模和适用场景上存在显著差异. 鉴于骨签数据的图像具有相对较小的

尺寸和较高的复杂性, 为平衡模型的系统鲁棒性与轻量化需求, 本文选择在 ImageNet-22K 数据集上预训练的 Vision-RWKV-L 模型提取骨签断裂处的关键视觉特征信息、在 Pile 数据集上预训练的 RWKV-6 World 模型提取骨签释文信息, 这些模型分别用于提取多模态数据中的特征信息, 保障模型适用性的同时显著提升训练效率, 并有效规避因模型参数过于庞大可能导致的过拟合问题. 此选择充分体现了在模型复杂性与性能需求之间的优化平衡, 能够为骨签数据分类提供稳健且高效的解决方案.

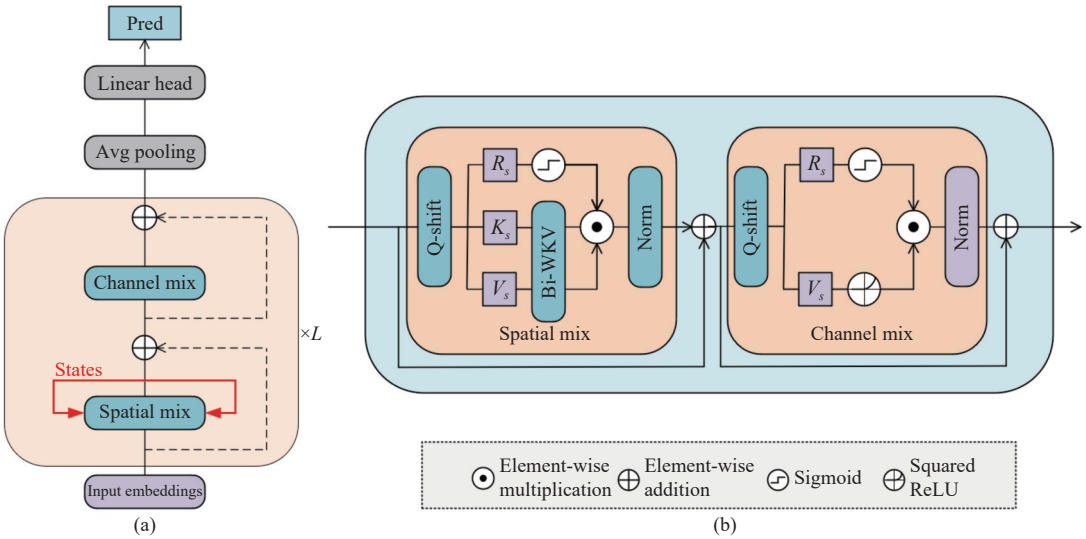


图 2 Vision-RWKV (VRWKV) 整体架构

3 方法提出

本文提出了一种基于多模态大模型驱动的骨签文物分类方法, 框架结构如图 3 所示. 该方法的核心包括以下几个模块: 骨签图像预处理模块 (IP 模块)、骨签特征提取模块 (VR 模块)、动态交叉特征融合模块 (DCFF 模块) 以及分类输出模块 (CC 模块). 整体工作流程如下: 首先 IP 模块对骨签图像数据集进行背景去除和裁剪预处理, 去除图像中的噪声和干扰. 接着将预处理后的骨签图片与骨签上的释文数据输入至 VR 模块, 通过特征提取网络提取多模态输入数据的特征向量表示. 随后 DCFF 模块对两类特征进行动态交叉融合, 生成统一的特征表示, 并将其传递至 CC 模块实现骨签文物的分类决策.

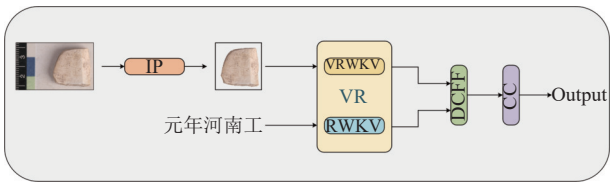


图 3 多模态大模型驱动的骨签文物分类方法框架图

3.1 骨签图像数据集预处理 (IP 模块)

原始骨签图像数据通常携带大量噪声信息, 如图 4(a) 所示. 若直接输入未经预处理的图像, 不仅会显著加重模型的计算复杂度, 还会引入冗余特征. 背景信息的特征冗余性会干扰模型的特征提取能力, 对模型的泛化性能产生负面影响. 为了有效提取图像中的关键匹配特征, 本文对图像数据集进行了预处理, 具体流程是分割图像中的主要信息区域, 之后对图片进行裁剪, 具体处理步骤如下.

图 4 骨签图像处理过程图

1) 使用 DIS-Net[22]对骨签主体进行分割. DIS-Net 是一种基于二类图像分割的深度网络, 由一个编码器、一个 U^2-Net[23]的图像分割模块和一个中间监督策略组成. DIS-Net 在高分辨率图像上能够精确地识别和分离目标物体. 由于其优越的二元分类能力和高精度分割特性, DIS-Net 在骨签图像分割中具有较高的适用性. 在本文的实验中, 遵循 DIS-Net 的训练策略, 利用预训练权重对网络进行微调, 实现对骨签图像的细粒度分割, 生成高精度的图像标签. 图 4(b) 展示了分割效果.

2) 将原始的骨签图像 (图 4(a)) 与分割结果图像 (图 4(b)) 进行逐像素的与运算, 保留骨签主体部分的像素值, 通过这一过程, 生成了标准化的骨签图像 (图 4(c)). 具体计算如式 (1), 其中 dst 表示输出图像, src_1 和 src_2 分别表示原始图像和掩模图像, (i, j) 表示像素位置.

$$dst(i,j) = src_1(i,j) \wedge src_2(i,j) + 255 \times (1 - src_2(i,j)) \quad (1)$$

3) 把分割后的结果图像 (图 4(c)) 进行主体的裁剪, 并确保裁剪后的图像尺寸为 224×224 像素, 这一尺寸标准化处理的目的是使得图像能够符合模型的输入要求, 如图 4(d) 所示, 从而保证数据的一致性并提高模型处理效率.

3.2 骨签特征提取模块 (VR 模块)

输入特征提取网络的数据分成两类, 分别是骨签上的释文信息、经过预处理的骨签图片. 整体结构如图 5 所示, 其中 RWKV 提取中文的骨签释文, 由于 RWKV 的架构结合了 RNN 的递归特性和 Transformer 的上下文建模能力, 在处理中文时具有更好的理解能力; VRWKV 提取预处理后的骨签图片特征信息.

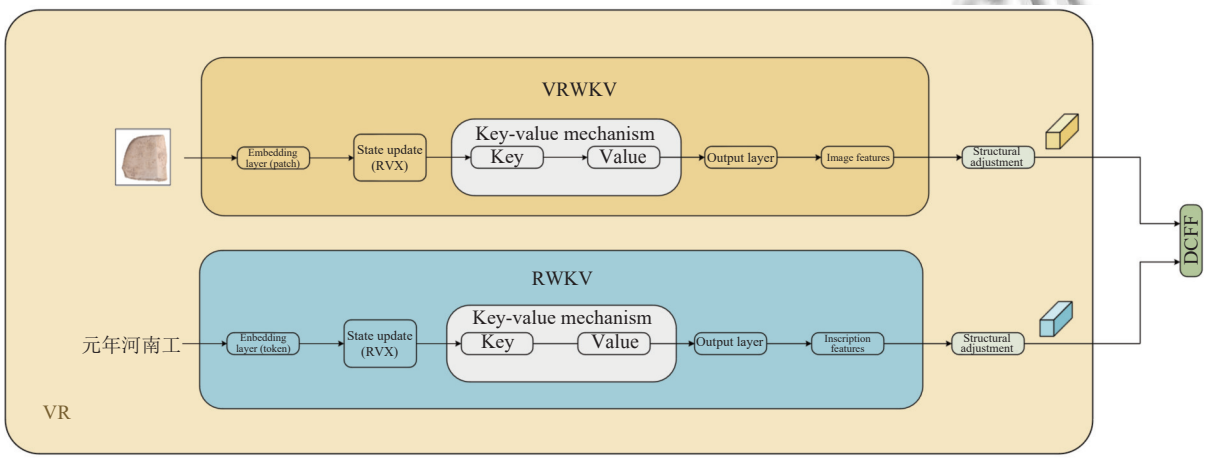


图 5 骨签特征提取模块结构图

1) RWKV 提取骨签释文特征

RWKV 特征提取机制分为以下几个步骤, 输入文本首先被分词为序列 $\{x_1, x_2, \cdots, x_T\}$, 其中 T 是序列长度, 每个词被映射到词汇表中的索引后, 通过嵌入矩阵 E 转化为嵌入向量 $e_t = E[x_t]$, $e_t \in \mathbb{R}^d$, 其中 $E \in \mathbb{R}^{V \times d}$ 是嵌入矩阵, V 为词汇表大小, d 为嵌入向量维度. RWKV 的核心是 Key-Value 递归累积机制, 用于逐步提取上下文信息, 每个时间步的嵌入向量 e_t 通过两个线性变换生成 Key 和 Value, 如式 (2) 所示.

$$K_t = W_k e_t,\ V_t = W_v e_t \quad (2)$$

其中, $W_k, W_v \in \mathbb{R}^{d \times d}$ 是可训练权重矩阵, 该模型通过递归公式将当前时间步的 Value 与前一时间步的状态 s_{t-1} 累积更新, 如式 (3) 所示.

$$s_t = \alpha_t \cdot s_{t-1} + (1 - \alpha_t) \cdot V_t \quad (3)$$

其中, $\alpha_t \in [0, 1]$ 是动态权重因子, 用于平衡历史信息和当前输入的影响. 当前时间步的最终输出隐藏状态 h_t 是 Key 和累积状态 s_t 的组合, 如式 (4) 所示. RWKV 的输出是上下文感知的隐藏状态矩阵 $H \in \mathbb{R}^{T \times d}$.

$$h_t = ReLU(K_t + s_t),\ h_t \in \mathbb{R}^d \quad (4)$$

其中, $d = 768$, 以最后一个时间步的向量 $h_T \in \mathbb{R}^{768}$ 作为整个输入序列的特征向量.

2) VRWKV 特征提取

特征提取过程包括图像分块、Key-Value 累积和上下文特征生成. 输入的图像首先被分割成若干图像块 (patch), 每个 patch 的大小为 $P \times P$ 像素, 分块数如式 (5) 所示.

$$N = \frac{H \times W}{P^2}, P_i \in \mathbb{R}^{R \cdot P \cdot C} \quad (5)$$

其中, $H \times W$ 是图像的大小, N 为总 patch 的数量, C 为通道数, 每个 patch 被展平为一维向量后, 通过线性变换映射到高维空间, 生成 patch 嵌入如式 (6) 所示.

$$E_i = W_p P_i, E_i \in \mathbb{R}^d \quad (6)$$

其中, $W_p \in \mathbb{R}^{(P \cdot P \cdot C) \times d}$ 是可训练权重矩阵, VRWKV 输出的全局特征为高维向量, 展平后维度为 150 528 (224×224×3). VRWKV 的核心是将 Key-Value 累积机制应用于图像 patch 的空间维度, 每个 patch 嵌入 E_i 通过线性变换生成 Key 和 Value 如式 (7) 所示.

$$K_i = W_k E_i,\ V_i = W_v E_i \quad (7)$$

状态递归公式与 RWKV 模型类似, 但作用于空间维度如式 (8) 所示.

$$s_i = \alpha_i \cdot s_{i-1} + (1-\alpha_i) \cdot V_i \tag{8}$$

最终每个 patch 的上下文感知特征如式 (9) 所示.

$$h_i = ReLU(K_i + s_i),\ h_i \in \mathbb{R}^d \tag{9}$$

VRWKV 的输出是每个 patch 的隐藏状态矩阵$H \in \mathbb{R}^{N\times d}$, 对于全局特征生成, 通常通过池化操作将所有 Patch 的特征整合为全局特征向量$h_{\text{global}} \in \mathbb{R}^{150528}$.

3) 特征向量调整

两个模型提取的特征维度存在显著差异, 其中文本与图像的特征维度差异较大, 若直接进行拼接, 可能导致特征融合过程的不平衡. 高维特征的相对权重在拼接时可能显著超过低维特征, 从而引发模型在训练过程中对某一模态的过度偏向, 削弱多模态学习的公平性与有效性. 为解决这一问题, 本文引入结构转换模块, 对各模态的特征向量进行统一映射, 确保特征维度的一致性. 具体而言, 本文采用全连接层对提取的特征进行线性变换, 将其映射至相同的维度$\mathbb{R}^{1024}$以增强模型在多模态特征融合时的均衡性与表现力. 由于全连接层的计算复杂度与输入和输出的维度成正比, 高维特征不仅增加计算量, 还会显著提升显存和内存需求, 因此本文将图像和文本特征分别映射到的统一维度 1 024 维, 这样可以显著降低计算复杂度和内存需求, 同时保留高维特征的主要信息. 这种变换的操作通常使用全连接层实现如式 (10) 所示, 其中$W \in \mathbb{R}^{1024\times d}$是用于变换维度的权重矩阵.

$$h' = ReLU(W \cdot h + b) \tag{10}$$

3.3 动态交叉特征融合 (DCFF 模块)

动态交叉特征融合可以充分挖掘多模态特征之间的交叉关系, 同时根据不同模态特征的重要性进行动态加权, 本文以文献[24]为理论基础, 将骨签的两类特征动态交叉特征融合. 传统的固定融合方法 (如静态加权或简单拼接) 无法根据输入数据的动态变化自适应调整权重, 而动态交叉特征融合通过动态生成每个模态的权重, 可以捕获模态重要性随上下文变化的特点. 动态交叉特征融合分为动态权重生成 (dynamic weight generation, DWG)、加权特征交互层 (weighted feature interactive, WFI)、模态内特征非线性变换层 (nonlinear transformation of in-modal features, NTIF) 以及最后的融合特征的生成层 (generation of fusion features, GFF), 其中 WFI 与 NTIF 共同构成特征交叉模态模块 FCM (feature cross-modal module),该模块在权重指导下实现跨模态交互与模态内非线性变换, 为最终融合提供更具判别性的表征. 具体结构如图 6 所示.

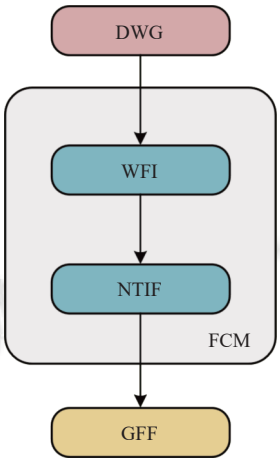


图 6 动态交叉特征融合模块层结构图

融合过程为特征向量首先输入动态权重生成层, 进行如式 (11) 的处理.

$$w_i = \sigma(W_i h_i + b_i),\ w_i \in \mathbb{R}^{1024} \tag{11}$$

其中, $W_i \in \mathbb{R}^{1024\times 1024}$是权重生成的可训练参数矩阵, $b_i \in \mathbb{R}^{1024}$是偏置向量, $\sigma(\cdot)$是 Sigmoid 激活函数, w_i输出是归一化权重, 而后为了保证权重的动态性和整体稳定性, 使用 Softmax 函数将权重向量归一化如式 (12) 所示.

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{2}\exp(w_j)},\ \alpha_i \in \mathbb{R}^{1024} \tag{12}$$

其中, $i \in \{1,2\}$对应两个模态即两类文本和图像. 在生成动态权重的基础上, 进行特征交叉机制的处理, 通过模态间的交叉机制实现特征融合, 通过该机制不同模态的特征通过逐维交叉操作进行相互作用, 以捕获模态间的高阶关系, 加权特征交互的目标是捕获两模态之间的协同关系. 首先是加权特征交互层, 为了增强模态内部的特征表示能力, 对每个模态的特征引入非线性变换, 即使用一个前馈网络对模态特征进行变换如式 (13) 所示.

$$h'_i = ReLU(W'_i h_i + b'_i),\ h'_i \in \mathbb{R}^{1024} \tag{13}$$

其中, $W'_i \in \mathbb{R}^{1024\times 1024}$为模态内非线性变换的可训练权重矩阵, $b'_i \in \mathbb{R}^{1024}$为偏置向量, 为非线性激活函数, 用

于引入非线性特性, 通过非线性变换, 可以强化模态内部的特征表达能力, 从而提高模态内和模态间信息交互的效果. 对于模态i和模态j的特征h_i, $h_j \in \mathbb{R}^{1024}$, 其加权交互特征计算如式 (14) 所示.

$$h_{\text{cross}}^{ij} = \alpha_i \odot h_i \cdot (\alpha_j \odot h_j),\ h_{\text{cross}}^{ij} \in \mathbb{R}^{1024} \tag{14}$$

其中, $\alpha_i \odot h_i$表示模态i的特征经过逐维加权调整后的值, h_{cross}^{ij}表示模态i与模态j的交互特征; 最后是融合特征的生成, 最终的融合特征由模态内加权特征 (通过动态权重调整后的单模态特征)、模态间交叉特征 (通过加权交互计算得到的模态间协同特征) 以及全局非线性变换 (融合后的综合特征通过一个前馈网络进行非线性映射) 组成, 最终的融合特征公式如式 (15) 所示.

$$h_{\text{fusion}} = ReLU\left[W_f\left(h'_{\text{text}} + h'_{\text{image}} + \sum_{i \neq j} h_{\text{cross}}^{ij}\right) + b_f\right] \tag{15}$$

其中, h'_{text}和h'_{image}是经过模态内非线性变换后的特征, h_{cross}^{ij}为模态间交叉特征, $W_f \in \mathbb{R}^{1024}$是融合后特征的可训练映射矩阵, $b_f \in \mathbb{R}^{1024\times1024}$为偏置向量, 通过该变换, 可以生成包含多模态信息的综合特征$h_{\text{fusion}} \in \mathbb{R}^{1024\times1024}$.

3.4 分类输出模块 (CC 模块)

本文构建了一种基于卷积层与池化层的高精度卷积神经网络分类器, 实现对骨签图像类别的精准识别, 具体结构如图 7 所示. 在分类过程中, 依托卷积层与池化层完成特征提取与降维优化. 卷积层主要通过滑动小型滤波器对输入特征进行加权卷积运算, 提取局部特征信息, 核心计算公式如式 (16) 所示.

$$y = f(W * x + b) \tag{16}$$

其中, x表示输入特征, W代表卷积核, y是输出特征, b为偏置项, f对应于激活函数, *表示卷积运算. 在卷积运算之后, 池化层用于降低特征图的维度, 减少计算复杂度的同时保留最具代表性的关键信息. 池化操作的数学表达式见式 (17).

$$y_{i,j} = \max_{(m,n)}\{x_{i+m,j+n}\} \tag{17}$$

其中, $y_{i,j}$表示池化操作后的输出, $x_{i+m,j+n}$代表输入特征对应的局部区域. 经过一系列卷积与池化操作后, 特征图的维度逐步缩减, 信息表达愈发紧凑, 此时输入数据的关键模式得以更有效地表征. 最后在全连接层中完成类别判别, 所得分类结果将通过 Softmax 层进行归一化处理, 输出各类别的概率分布. Softmax 函数的计算方式如式 (18) 所示.

$$P(y = j|x) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{18}$$

其中, z_j表示类别的得分, 而$P(y = j|x)$是对应类别j的概率. 为优化模型性能, 本研究采用交叉熵损失函数来指导参数调整, 计算方式详见式 (19).

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{19}$$

其中, y_i表示真实标签的概率分布, $\hat{y}_i$代表模型预测的概率分布, C是类别总数. 通过最小化交叉熵损失, 模型能够不断优化权重参数, 提升分类准确率.

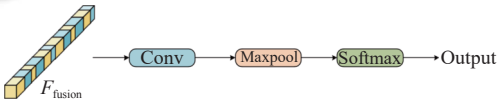


图 7 分类输出模块结构图

4 实验环境、数据集和评价指标

4.1 实验环境

本研究使用的计算平台配置为 64 位 Windows 11 操作系统, 搭载 32 GB 的内存, 中央处理器为英特尔第 14 代 Core i9-14900HX, 并配备 NVIDIA GeForce RTX 4070 显卡, 编程任务在 Python 3.9 环境中通过 PyTorch 框架进行实现.

4.2 数据集

本研究所使用的数据集来源于汉长安未央宫遗址出土的预处理后的骨签图像及其所对应的骨签释文, 共 5 490 组数据, 按照 6:2:2 的比例划分训练集、验证集、测试集. 每组数据均由匹配的骨签图像及其对应的释文构成. 为增强模型的泛化能力, 训练阶段对原始骨签图像数据集进行了随机增强操作, 包括平移和缩放变换. 本研究数据集按照骨签颜色 (黄色、黑色、白色) 与骨签部位 (上部分、中间部分、下部分) 组合, 共划分为 9 个类别, 分别为“黄色、上部分”“黄色、下部分”“黑色、上部分”“黑色、下部分”“白色、上部分”“白色、下部分”“白色、中间部分”“黄色、中间部分”“黑色、中间部分”. 每个类别均包含 366 组训练集、122 组测试集和 122 组验证集, 类总数为 610 组, 从而保证了不同类别之间的数据分布均衡. 数据集类型示例如图 8 所示.

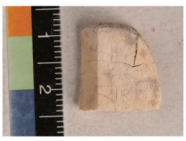
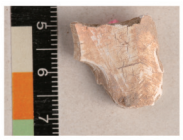
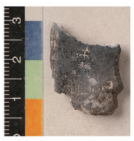
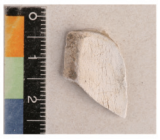
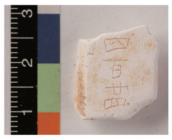
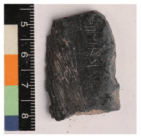
骨签数据集类型	骨签原始图片示例	骨签图片预处理数据集示例	骨签上释文数据集示例	数据集原始编号	数据集标签
黄色、上部分			服六	22691	0
黄色、下部分			六年河南工驠/ 廣作驠	16563	1
黑色、上部分			四年河南工官令定驠/ 元作府滿工驠	25764	2
黑色、下部分			驠七十五	20088	3
白色、上部分			甲千八百驠	27161	4
白色、下部分			驠四百廿四	17194	5
白色、中间部分			服六石	3654	6
黄色、中间部分			乙萬八百	4061	7
黑色、中间部分			丙二千二百卅六	5029	8

图 8 数据集类型示例

4.3 评价指标

本文分别采用计算准确率 (*Accuracy*)、精确率 (*Precision*)、召回率 (*Recall*) 和加权 *F*1 分数 (weighted *F*1 score, *wF*1) 作为评估模型的核心指标, 各个指标的取值范围均在区间[0, 1]内, 计算公式如式 (20)–(23) 所示, 其中TP_i, FP_i, FN_i分别表示第i类中的真正类中的真正类、假正类和假负类样本数, C为类别总数, N为所有样本总数, n_i为第i类的样本数, n_i/N为该类在整体样本中的占比权重. 其中准确率 (*Accuracy*) 表示模型预测正确的样本数量占总样本数量的比例; 精确率 (*Precision*) 是指分别计算每个类别的精确率后, 对所有类别结果取算术平均, 用于衡量模型在预测为某一类的样本中有多少实际属于该类, 是反映模型预测“正确性”的重要指标; 召回率 (*Recall*) 是对每个类别的召回率取算术平均, 用于衡量模型对真实为某类的样本中能够成功预测为该类的比例, 反映模型对该类样本的“覆盖能力”; 加权 *F*1 分数 (*wF*1) 是针对多分类任务的扩展版本, 每类的 *F*1 分数 (即该类精确率与召回率的调和平均) 计算后, 再取所有类别的算术平均值.

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{20}$$

$$Precision = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i} \tag{21}$$

$$Recall = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \tag{22}$$

$$wF1 = \sum_{i=1}^{C}\left(\frac{n_i}{N}\cdot\frac{2\cdot Precision_i \times Recall_i}{Precision_i + Recall_i}\right) \tag{23}$$

5 实验

5.1 消融实验

本研究采用对比学习策略对网络进行预训练, 模型能够初步学习并捕捉骨签数据的潜在特征. 在上述过程中, 利用预训练模型的权重对整个系统进行初始化, 确保模型在训练初期具备较强的表示能力. 具体来说, 本研究将批量大小设定为 16, 训练轮次设置为 200, 确保模型能够在训练过程中实现充分的收敛. 此外除了训练阶段的超参数调优外, 其余超参数保持不变. 通过这一预训练过程, 获得了适用于骨签数据的网络权重.

图 9 的训练结果展示了两种训练策略下网络性能的差异, 呈现了训练损失和验证损失随迭代轮次变化的趋势. 紫色与黄色曲线分别对应未使用预训练权重情况下的训练损失与验证损失曲线, 蓝色与绿色曲线反映了采用预训练权重时的训练损失与验证损失. 从结果可观察到, 使用预训练权重的网络在约 130 轮次后表现出损失趋于稳定的特性, 表明模型已完成收敛过程. 相较之下未使用预训练权重的网络在训练过程中损失波动显著, 并且损失下降速度显著较慢, 进一步反映了其训练性能的劣势. 上述结果表明了采用对比学习策略对网络进行预训练, 使模型能够初步学习并捕捉骨签数据的潜在特征的有效性.

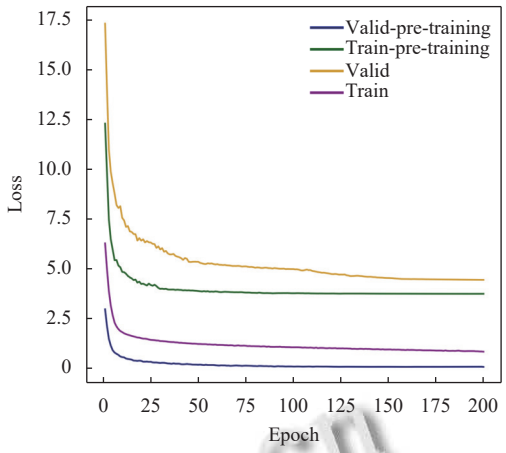


图 9 对比学习训练损失曲线

为验证骨签释文信息及动态交叉特征融合模块的有效性, 本文设计并开展了一系列实验, 分别包括: (1) 仅利用 VRWKV 模型提取骨签图像特征直接进行分类的实验; (2) 通过简单直接拼接 VRWKV 提取的图像特征与 RWKV 模型提取释文特征进行分类的实验; (3) 本文所提出的分类方法. 如表 1 实验结果显示, 单独依赖骨签图像特征作为分类依据的方案, 其分类精度相对来说较低. 在对骨签释文特征进行简单拼接后, 模型的各项评估指标均实现了 1.29%–2.68% 的提升, 其中准确率提高了 2.68%, 充分验证了释文信息对模型性能优化的有效性. 更为重要的是, 在引入动态交叉特征融合模块之后, 各项指标进一步提升 2.67%–3.33%, 其中精确率提升 2.74%, 精确率的提升意味着误报率进一步降低, 表明该模块在优化模型判别能力方面具有显著作用, 模型在类别划分时也更加谨慎, 有效减少了错误分类的样本. 此外, 加权 *F*1 分数 (*wF*1)

进一步提升 3.33 个百分点, 表明其在处理多类别复杂数据时展现出更优的泛化能力, *wF*1 的增长体现了模型在多类别任务中的预测稳定性与鲁棒性的增强, 从而从多维度印证了方法在实际应用场景中的可靠性与适应性.

表 1 消融实验结果表

模型结构	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	*wF*1 (%)	Params (M)	GFLOPs
VR	87.10	87.72	87.32	87.62	217.94	18.50
VR&R (normal contact)	89.78	89.16	89.64	88.91	339.47	31.53
VR&R&DCFF (ours)	**92.85**	**91.90**	**92.31**	**92.24**	**360.41**	**34.26**

为系统评估各模块叠加对模型计算效率的影响, 并精确量化不同结构配置下的资源开销, 本文在消融实验中对多种模块组合形式下的参数规模 (Params) 与计算复杂度 (GFLOPs) 进行了全面统计与对比分析. 如表 1 所示, VRWKV 与 RWKV 模块在整体模型中占据了主要的参数与计算资源份额. 在基础并行大模型架构中引入动态混合交叉模块后, 模型参数总量由原先的 339.47M 上升至约 360.41M, 推理计算开销则由 31.53 GFLOPs 增加至 34.26 GFLOPs. 整体而言, 动态混合交叉模块的引入增加了约 20.94M 的额外参数与 2.73 GFLOPs 的推理负担, 两者增幅均控制在 9% 以内. 这表明所提出的动态融合机制在保持模型结构轻量化的同时, 能够高效提升模型的表现与融合能力, 展示出良好的工程适应性与计算效率.

5.2 模型对比实验

为进一步验证本文所提出匹配方法的有效性, 本研究选取了 5 种当前主流的图像分类方法进行对比实验, 包括以深度学习为代表的 YOLOv10 和 YOLO11 图像分类模型, 以及 CLIP、ViLT、BLIP、VLMo、ALBEF 多模态大模型. 为了确保对比实验的公平性和模型选择的有效性, 本文分别选取了不同规模的模型进行实验, 具体包括 YOLOv10m、YOLO11m、ViLT-B、CLIP-ViT-L、BLIP-Large、VLMo-Base 和 ALBEF-Large. 实验过程中, 所有模型的批量大小均设定为 16, 训练轮次设置为 200, 确保模型在训练过程中能够充分收敛. 实验结果如表 2 所示.

实验结果表明, 以 YOLO 系列为代表的两种深度学习模型在骨签文物分类任务中表现相近, 这表明基于卷积神经网络的特征提取能力存在一定局限性, 导致模型在高阶特征表达方面能力不足, 进而影响分类精度. 此外, 实验结果还揭示了基于 Transformer 架构的多模态大模型在该任务中提升分类性能的前提是依赖更为复杂的特征提取网络, 即仅采用标准 Transformer 结构难以实现最佳性能. 针对上述问题, 本文所提出的方法融合 CNN、RNN 及 Transformer 架构的混合式架构大模型充分结合了三者各自的优势, 取得了显著的分类性能提升. 相较于参数规模最大的 ALBEF 模型, 本文提出的方法在各项评估指标均实现了 2.29%–3.52% 的提升, 其中分类准确率上提升了 3.17%, 这一结果充分验证了 RWKV 与 VRWKV 并行架构在准确度上的优越性, 证明了其在骨签分类垂直任务上的有效性. 同时本文提出的方法在加权 *F*1 分数 (*wF*1) 指标上较 ALBEF 模型提升了 3.52 个百分点, 显著优于基于 Transformer 架构的多模态模型. 该结果充分表明, 所设计的并行大模型架构在处理骨签图像分类的垂直任务中展现出更强的泛化能力, 在此特定领域内具备更高的适应性与性能优势.

表 2 对比实验结果表 (%)

模型类型	*Accuracy*	*Precision*	*Recall*	*wF*1
YOLOv10m	80.81	80.13	78.64	80.52
YOLO11m	81.93	82.87	82.95	82.59
ViLT	84.86	85.86	84.83	85.84
CLIP	86.82	85.94	86.44	86.73
BLIP	84.25	84.93	84.38	84.93
VLMo	85.71	86.81	83.84	85.17
ALBEF	89.68	89.61	89.93	88.72
VR&R&DCFF (ours)	**92.85**	**91.90**	**92.31**	**92.24**

5.3 大模型时效对比实验

为全面评估该架构在资源占用与时效性方面的表现, 本文在同一实验条件下, 与主流多模态大模型包括 CLIP (模型文件大小为 864 MB)、ViLT (模型文件大小为 870 MB)、BLIP (模型文件大小为 936 MB)、VLMo (模型文件大小为 1 100 MB)、ALBEF (模型文件大小为 1 223 MB) 以及本文模型 (模型文件大小为 1 250 MB) 进行了对比实验, 统计上述方法在训练、验证、测试阶段所需要的时间与其在全阶段 GPU 的利用率 (GPU utilization rate). 相关实验统计与可视化结果如图 10 所示.

从统计图中结果可以看出, 当前主流的基于 Transformer 架构的大规模视觉语言模型 (CLIP、ViLT、BLIP、VLMo 及 ALBEF) 在训练、验证与测试各阶段的时间消耗以及 GPU 资源的平均利用率, 整体呈现出

与模型结构复杂性正相关的趋势, 即随着模型参数规模和架构复杂度的增加, 其在不同阶段所需的计算时间显著增长, GPU 利用率亦相应提高, 反映出此类模型在性能提升的同时对硬件资源提出了更高要求. 相比之下本文提出的并行机制驱动的新型大模型架构在训练、验证及测试过程中的时间消耗和 GPU 利用率处于 CLIP 与 BLIP 之间. 尽管模型本身具备较大的参数规模, 其通过优秀的架构方式与高效的计算路径整合, 显著提升了推理阶段的执行效率, 在保持较高精度的前提下实现了与轻量级模型相近的推理速度. 该现象不仅充分体现了所提模型在保持表达能力和学习能力的同时具备出色的计算效率, 还进一步验证了其在实际应用场景中部署的可行性与广泛的推广潜力.

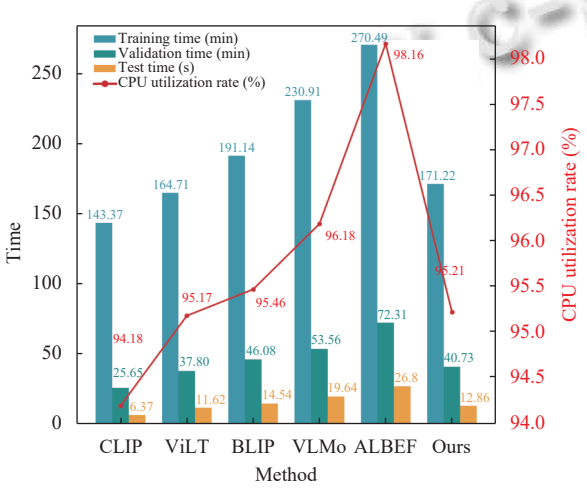


图 10 模型时效统计图

6 结论与展望

本研究中采用了两种并行的大模型对骨签数据进行全面的特征提取, 通过动态交叉特征融合策略有效整合了各模型所提取的信息. 使用精细化分类器对骨签图像进行分类, 进一步增强了分类结果的可靠性. 此外针对输入图像, 采取了高效的预处理方法以减少噪声干扰, 显著提高了数据质量. 本文所提出的方法解决了传统基于单一图像信息分类的不足, 也克服了多模态大模型特征提取不充分精度较低的问题. 通过引入对比学习策略进行预训练权重优化, 进一步提升分类精度. 实验结果表明, 所提方法达到了相对较高召回率、高加权 $F1$ 分数, 显著优于 YOLO 系列和多模态大模型, 提高了骨签图像分类的准确性.

本研究为考古学领域的专家提供高效的骨签图像分类整理的数字化方法. 但是由于骨签特征的高度复杂性、图像之间的相似性较强以及训练数据量的限制, 本文所提出的方法在分类准确度上尚未达到最优水平. 因此未来的研究将通过开展更多实验来进一步地精细化分类, 并探索其在实际应用场景中的部署潜力.

参考文献

1 于志勇. 汉长安城未央宫遗址出土骨签之名物考. 考古与文物, 2007(2): 48–62. [doi: 10.3969/j.issn.1000-7830.2007.02.008]
2 刘国能. 我国最早的专门档案馆库——汉代骨签档案馆库. 中国档案, 2007(7): 50–52.
3 朱定赟, 陈翼遥. 基于改进 YOLOv8 的火灾图像分类方法. 上海第二工业大学学报, 2025, 42(1): 59–65.
4 Ma H, Lei S, Celik T, *et al*. FER-YOLO-mamba: Facial expression detection and classification based on selective state space. arXiv:2405.01828, 2024.
5 Alhwaiti Y, Khan M, Asim M, *et al*. Leveraging YOLO deep learning models to enhance plant disease identification. Scientific Reports, 2025, 15(1): 7969. [doi: 10.1038/s41598-025-92143-0]
6 Elazab N, Gab-Allah WA, Elmogy M. A multi-class brain tumor grading system based on histopathological images using a hybrid YOLO and RESNET networks. Scientific Reports, 2024, 14(1): 4584. [doi: 10.1038/s41598-024-54864-6]
7 Raushan R, Singhal V, Jha RK. Damage detection in concrete structures with multi-feature backgrounds using the YOLO network family. Automation in Construction, 2025, 170: 105887. [doi: 10.1016/j.autcon.2024.105887]
8 Ramazhan MRS, Bustamam A, Buyung RA. Smart car damage assessment using enhanced YOLO algorithm and image processing techniques. Information, 2025, 16(3): 211. [doi: 10.3390/info16030211]
9 Ullah F, Ullah I, Khan RU, *et al*. Conventional to deep ensemble methods for hyperspectral image classification: A comprehensive survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 3878–3916. [doi: 10.1109/JSTARS.2024.3353551]
10 Archana R, Jeevaraj PSE. Deep learning models for digital image processing: A review. Artificial Intelligence Review, 2024, 57(1): 11. [doi: 10.1007/s10462-023-10631-z]
11 张新生, 陈鼎, 秦一冰. 基于 CLIP 文本特征增强的剪纸图像分类. 计算机应用研究, 2025, 42(7): 1994–2002. [doi: 10.19734/j.issn.1001-3695.2024.11.0485]

12 王皓嘉, 邓勇舰, 刘婷婷, 等. 基于 DINO 先验的混合特征面部表情识别模型. 计算机工程. https://link.cnki.net/urlid/31.1289.TP.20240807.1026.004. [2024-08-07]. [doi: 10.19678/j.issn.1000-3428.0069519]

13 Duan YC, Wang WY, Chen Z, *et al*. Vision-RWKV: Efficient and scalable visual perception with RWKV-like architectures. Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.

14 Peng B, Alcaide E, Anthony Q, *et al*. RWKV: Reinventing RNNs for the Transformer era. Proceedings of the 2023 Findings of the Association for Computational Linguistics. Singapore: ACL, 2023. 14048–14077.

15 Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020, 404: 132306. [doi: 10.1016/j.physd.2019.132306]

16 Han K, Xiao A, Wu EH, *et al*. Transformer in Transformer. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1217.

17 Xie SH, Li YD, Ma YL, *et al*. AutoGMM-RWKV: A detecting scheme based on attention mechanisms against selective forwarding attacks in wireless sensor networks. IEEE Internet of Things Journal, 2025, 12(4): 4403–4419. [doi: 10.1109/JIOT.2024.3484999]

18 Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450, 2016.

19 Alzubaidi L, Zhang JL, Humaidi AJ, *et al*. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 2021, 8(1): 53. [doi: 10.1186/s40537-021-00444-8]

20 Khan S, Naseer M, Hayat M, *et al*. Transformers in vision: A survey. ACM Computing Surveys (CSUR), 2022, 54(10s): 200. [doi: 10.1145/3505244]

21 Eldan R, Shamir O. The power of depth for feedforward neural networks. Proceedings of the 29th Conference on Learning Theory. New York: JMLR.org, 2016. 907–940.

22 Qin XB, Dai H, Hu XB, *et al*. Highly accurate dichotomous image segmentation. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 38–56.

23 Qin XB, Zhang ZC, Huang CY, *et al*. U^2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 2020, 106: 107404. [doi: 10.1016/j.patcog.2020.107404]

24 Xue ZH, Marculescu R. Dynamic multimodal fusion. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 2575–2584.

(校对责编: 张重毅)