

双视角并行视觉架构驱动的骨签分类^①



马珂欣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: 汉长安城未央宫三号建筑遗址出土 6 万余枚骨签, 多为长条状肩胛骨片, 其数量庞大、形制高度相似且信息承载复杂, 系统化分类是厘清骨签属性与构建文物信息体系的关键前提。然而, 长期埋藏导致骨签普遍断裂、分散, 人工整理与分类效率低且存在文物损坏风险, 亟需高效自动化分类方法。针对上述问题, 本文提出一种基于双视角并行 Vision-RWKV 特征建模的骨签分类方法。首先, 在 UNet++ 框架基础上引入通道注意力机制、轻量化特征生成策略与边缘引导约束, 实现骨签主体的高精度分割, 并将分割结果作为分类网络输入以突出关键区域。随后, 在分类阶段采用两个并行的 Vision-RWKV 模型分别提取骨签正反面特征, 并通过协同融合策略实现双视角信息的统一建模, 增强特征表达的完整性与判别力。实验结果表明, 该方法在骨签分类任务中取得 94.76% 的准确率, 优于现有主流图像分类模型, 为骨签文物的数字化整理与考古信息管理提供了有效技术支撑。

关键词: 骨签; 图像分类; 图像分割; 文物分类; 双视角

引用格式: 马珂欣, 王慧琴, 王可, 刘瑞, 王展, 毛力. 双视角并行视觉架构驱动的骨签分类. 计算机系统应用, 2026, 35(9): 239-254. <http://www.c-s-a.org.cn/1003-3254/10242.html>

Bone Stick Classification via Dual-view Parallel Vision Architecture

MA Ke-Xin¹, WANG Hui-Qin¹, WANG Ke¹, LIU Rui², WANG Zhan³, MAO Li¹

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: More than sixty thousand bone sticks excavated from Building No. 3 of the Weiyang Palace site in the Han Chang'an City are predominantly elongated scapular bone fragments, whose large quantity, similar morphology, and complex information content make systematic classification essential for archaeological data organization. However, long-term burial has caused extensive fragmentation and dispersion, rendering manual classification inefficient and increasing the risk of artifact damage, thereby necessitating automated solutions. To address this challenge, this study proposes an intelligent bone stick classification method based on dual-view parallel Vision-RWKV feature modeling. High-precision segmentation of bone stick regions is first achieved by incorporating channel attention mechanisms, lightweight feature generation strategies, and edge-guided constraints into the UNet++ framework, and the segmented regions are used as inputs to emphasize key structural features. Subsequently, two parallel Vision-RWKV models are employed to extract features from the front and back views of bone sticks, and a cooperative fusion strategy is adopted to unify dual-view information and enhance discriminative capability. Experimental results demonstrate that the proposed method achieves an accuracy of 94.76%, outperforming existing image classification models and providing effective technical support for archaeological data organization.

Key words: bone stick; image classification; image segmentation; cultural artifact classification; dual-view

① 基金项目: 国家社科基金冷门绝学研究专项 (20VJXT001)

收稿时间: 2026-01-20; 修改时间: 2026-02-09, 2026-03-05; 采用时间: 2026-03-09; csa 在线出版时间: 2026-07-17

CNKI 网络首发时间: 2026-07-20

1 引言

汉长安城未央宫三号建筑遗址出土了总量达 6 万余枚的骨签文物,其形制多为由动物肩胛骨加工而成的长条状骨片,长度约为 5.8–7.2 cm,宽度约为 2.1–3.2 cm^[1]。骨签中部通常带有半月形缺口,并刻有文字记录,其中保存有释文的数量超过 5.7 万枚,文字内容涵盖官署名称、官吏与工匠信息、兵器类型及相关技术参数,如编号、弓弩张力和射程等^[2],较为完整地反映了西汉时期中央对兵器生产与质量管理的制度体系。这些文字资料不仅为研究汉代经济运行和手工业生产提供了重要依据,也为理解当时的社会结构与政治治理方式提供了直观证据。由于长期埋藏于地下环境中,骨签受到土壤侵蚀、地层扰动及材质老化等多重因素影响,出土时普遍存在断裂、分散等问题,增加了整理与研究难度。现阶段考古工作主要依赖人工方式对骨签进行分类和整理,不仅效率较低,而且在操作过程中存在一定的文物安全风险。因此,引入数字图像处理等计算机技术手段,对骨签文物开展自动化或半自动化分类研究,有助于提高工作效率并降低人为损伤的可能性。

随着深度学习方法的不断演进,图像分类技术在模型结构设计与特征表达能力方面均取得了显著进展。以 ResNet^[3]、DenseNet^[4]和 EfficientNet^[5]等为代表的卷积神经网络模型,通过引入残差连接、特征复用及复合尺度缩放策略,有效缓解了网络退化问题并提升了特征提取效率,已在多类视觉识别任务中取得稳定表现。近年来,研究者围绕这些基础网络不断对主干结构进行轻量化或注意力增强改进,使模型在复杂背景和小样本条件下具备更强的判别能力。然而,在骨签文物图像分类场景中,由于目标表面风化严重、纹理对比度低且类别间差异主要体现在刻痕形态与局部细节上,单纯依赖卷积神经网络^[6]的层级特征提取机制,往往难以充分刻画具有判别意义的细粒度信息,从而限制了分类性能的进一步提升。

在目标检测与相关分类任务中,YOLO (you only look once) 系列模型^[7–9]作为典型的一阶段深度学习框架,以端到端的训练方式和较高的推理速度在实际应用中得到广泛关注。已有研究通过改进 YOLO 的主干网络结构、特征融合策略及损失函数设计,提升其在特定场景下的分类与识别能力。然而,YOLO 模型本质上仍以卷积神经网络作为特征提取核心,更侧重于对

目标整体结构和显著区域的快速感知,对于依赖微小刻痕、弱纹理差异进行区分的骨签文物图像而言,其对细粒度局部特征的建模能力相对有限,因而在精细化分类任务中仍存在一定局限。

除卷积模型与 YOLO 框架外,近年来基于注意力机制的视觉模型逐渐成为图像分类研究的重要方向。Vision Transformer^[10]、Swin Transformer^[11]等模型通过自注意力机制对全局上下文关系进行建模,在整体语义表达和复杂结构理解方面展现出明显优势。部分研究尝试将此类方法引入文物与文化遗产图像分析领域,以提升对整体形态与空间关系的刻画能力。然而,这类模型通常对训练样本规模和计算资源依赖较高,在骨签文物图像数量有限且分辨率较高的条件下,容易面临训练成本偏高或对局部细节敏感性不足的问题。

与此同时,随着深度学习相关技术的发展,基于预训练的视觉或多模态模型^[12]逐渐被应用于图像分类研究中。此类模型通常依托大规模跨领域数据进行预训练,在特征泛化能力和语义表达能力方面具有显著优势,并在零样本或少样本分类任务中展现出良好潜力。然而,在骨签文物图像分类任务中,这类模型往往更关注高层语义信息,对刻痕深浅、边缘破损形态等细粒度视觉特征的刻画仍存在不足;同时,其模型规模庞大、计算与存储开销较高,在实际考古场景中部署和应用仍面临一定限制。

为应对上述挑战,本文提出了一种基于双视角特征融合的并行视觉架构的骨签分类方法,采用两个视觉模型分别对骨签正面与反面图像进行并行特征建模,并通过特征融合实现对正反两面视觉信息的协同利用,用于骨签文物的精细化分类。该方法以经过高精度分割的骨签正反面图像作为输入,利用 Vision-RWKV^[13]模型对正反面图像分别进行特征提取,从不同视角编码骨签的全局语义与局部纹理信息。提取的多层次特征随后通过动态交叉融合模块进行整合,该模块能够自适应捕捉正反面之间的互补关系,并增强局部刻痕和纹理细节的表达能力。在训练过程中,模型先借助对比学习对骨签特征进行结构化表示,同时通过迁移学习将预训练权重初始化到骨签分类任务中,加速模型收敛并提升对少量样本的学习能力。最终,经过微调的模型在分类器上实现对骨签的精细化分类,同时动态调整不同视角特征的权重,以兼顾正反两面信息的判别价值和整体鲁棒性。本文的主要贡献如下。

● 构建双视角并行视觉架构: 针对骨签文物正、反两面信息互补的特点, 分别对其正、反面图像进行独立特征建模. 并通过动态特征融合实现两面信息的协同利用, 为骨签细粒度分类提供了一种高效的多视角建模方案.

● 设计高精度图像预处理机制: 提出一种语义分割网络 SGUNet++对骨签图像进行分割, 提取骨签主体区域并去除背景冗余信息, 为后续特征提取提供规范输入, 同时增强微小刻痕和局部纹理的可辨识度.

● 设计动态交叉特征融合机制: 提出一种动态交叉特征融合模块, 通过自适应方式融合来自正反两面及多层次的特征信息, 使模型能够在多视角和多维度下充分捕捉骨签图像的细粒度纹理和结构特征, 从而增强判别能力.

● 精细化分类器架构: 本文采用多分支卷积路径结合注意力加权机制的分类器. 各分支分别聚焦不同尺度的局部特征, 注意力机制自适应地调整各分支的输出权重, 将融合后的特征映射为类别概率, 以实现高精度的细粒度骨签分类.

2 基本理论

2.1 UNet++

UNet++^[14]是语义分割领域中对经典 U-Net^[15]架构的重要改进, 其核心创新在于构建了密集嵌套的跳跃连接机制, 其结构图如图 1 所示. 该架构通过系统性的结构重构, 在保持编码器-解码器基本框架的同时, 显

著提升了特征传递的效率和分割边界的精度.

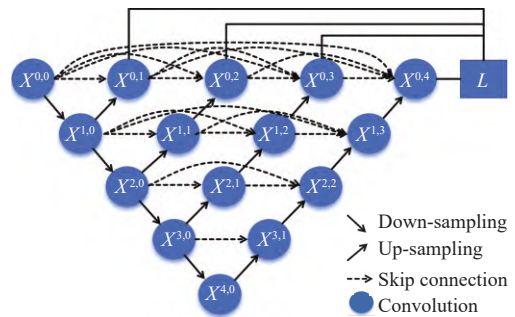


图 1 UNet++架构图

相比原始 U-Net 仅在同尺度间进行一次特征传递, UNet++在编码器与解码器之间增加多个中间连接节点, 使解码端能够融合来自不同层级的特征信息, 从而减小高低层特征之间的语义差异, 提升分割结果的细节表达能力. 此外, UNet++还采用深度监督机制, 在多个解码层输出辅助结果, 以增强训练稳定性并提高分割精度. 整体而言, 该结构在保持计算复杂度相对可控的前提下, 有效提升了特征融合能力与边界分割效果, 为后续结构优化提供了良好基础.

2.2 Vision-RWKV

Vision-RWKV 是在 RWKV (receptance weighted key-value) 架构基础上发展而来的前沿视觉模型, 最早由 Duan 等人^[16]于 2025 年提出, 其设计目标是兼顾高分辨率图像处理和长序列特征建模的能力, 同时在保持全局依赖捕获能力的前提下优化计算效率. 其结构如图 2 所示.

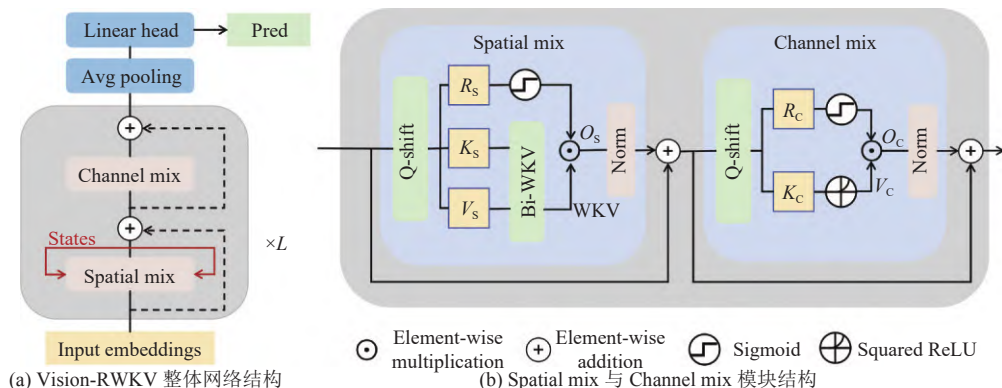


图 2 Vision-RWKV 整体架构

RWKV 将循环神经网络 (recurrent neural network, RNN) 的递归状态机制与 Transformer^[17]风格的序列建模思想相结合, 通过递归加权机制替代传统自注意力

结构, 实现线性复杂度的全局依赖建模, 克服了 Transformer 在长序列输入上的计算和内存瓶颈. Vision-RWKV 在此基础上针对视觉任务进行扩展, 引入空间特征编

码与多层次递归特征建模,使其能够同时捕捉全局语义信息与局部细节特征.从结构上看,Vision-RWKV主要包括输入编码层、递归加权特征提取层、特征聚合输出层这3个核心模块.

在输入编码阶段,高分辨率图像被划分为固定大小的图像块或序列单元,每个单元通过线性映射或卷积操作生成特征向量,并加入位置编码以保持空间信息和序列顺序,为递归建模提供基础表示.递归加权特征提取层是 Vision-RWKV 的核心,其内部采用树状或递归结构,对序列化输入进行逐层加权累积.该机制通过对每个位置的特征进行动态加权整合,实现长距离依赖建模,并在多层次上形成层次化的特征表示,使低层局部纹理信息与高层全局语义信息能够在同一特征空间中有效融合.

递归加权特征提取层在每个递归单元中通过线性变换、非线性激活及状态累积,将当前输入特征与前一递归状态结合,形成渐进式特征表示.这种层级递归设计不仅保证了全局信息的持续传播,还强化了对局部细节和微观结构的表达能力.通过树状的多路径加权结构,模型能够实现不同层级、不同空间位置特征的互补整合,形成稳定且丰富的多尺度表示.

在特征聚合输出阶段,多层次递归特征通过跳跃连接或多尺度融合策略整合,生成最终的全局表示向量,为后续的视觉任务提供结构化特征.这种融合机制能够充分利用不同层级的信息,使模型在表达能力和判别力之间取得平衡.同时,Vision-RWKV的模块化设计允许引入卷积操作、残差连接以及注意力增强模块,以进一步强化局部特征捕捉能力和全局依赖建模能力.

Vision-RWKV 通过递归加权机制与 Transformer 风格序列建模的结合,在全局依赖建模、局部细节感知与计算效率之间取得了平衡,其递归树状结构在层级特征表达、多尺度融合和长距离信息传播方面具有明显优势,为现代视觉模型提供了不同的理论基础和模块化设计思路,适合作为高分辨率图像与长序列视觉任务的核心建模框架.

2.3 模型选择

作为近年来不断演进的视觉序列建模方法,Vision-RWKV 系列在网络结构复杂度、特征表达能力及任务适配性等方面表现出显著差异.针对骨签图像尺度相对较小、局部结构细节丰富且特征分布复杂的特点,本文在模型选择过程中重点权衡了特征表达能力与计

算复杂度之间的关系.基于上述考虑,选用在 ImageNet-22K 数据集^[18]上完成预训练的 Vision-RWKV-L 模型,对骨签图像中的断裂边缘、表面纹理及局部结构等关键视觉特征进行高层次表征,以在保证建模能力的同时提升训练效率并增强模型的整体适用性.

3 方法提出

本文提出了一种基于双视角信息协同建模的并行视觉分类框架,用于骨签文物的精细化类别判别,其整体结构如图3所示.该方法由骨签分割 (bone stick segmentation, BSS) 模块、双视角特征编码 (dual-view feature encoding, DVFE) 模块、自适应交互融合 (adaptive interactive fusion, AIF) 模块以及分类输出 (classification output, CO) 模块构成,形成一套端到端的分类流程.

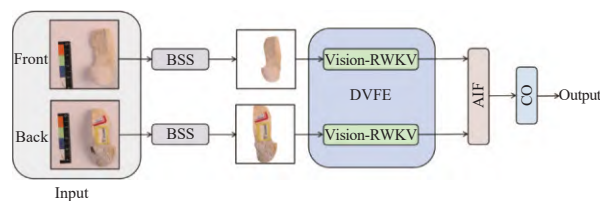


图3 双视角并行视觉架构骨签文物分类方法框架图

在数据处理阶段,BSS模块分别对骨签正面与反面图像进行背景抑制与尺度对齐,通过裁剪与噪声削弱操作,减少成像条件差异对后续特征学习的影响,从而保证双视角输入在空间结构上的一致性.随后,规范化后的正反面图像被并行输入至DVFE模块.该模块采用双分支结构,对两种视角信息进行独立建模,通过层级化特征提取策略分别学习各自具有判别意义的纹理与结构特征,避免单一视角主导特征表达.在特征融合阶段,引入AIF模块对双分支输出的特征进行交互式整合.该模块通过构建视角间的特征关联关系,在保持各自特征完整性的同时,实现跨视角的信息补充与强化,使融合结果能够更加全面地反映目标的整体判别特征.最终,融合后的特征表示被送入CO模块,通过判别性映射完成类别推理,实现对骨签文物的稳定分类输出.

3.1 骨签图像分割 (BSS) 模块

由于原始骨签图像通常包含复杂背景、光照不均及非目标区域干扰,直接将其输入后续特征建模网络,容易引入冗余信息并削弱模型对关键结构的关注能力,

从而影响整体分类性能. 因此, 在特征提取之前对骨签区域进行准确定位与分离, 是提升模型判别能力和鲁棒性的必要前提. 基于上述考虑, 本文构建了骨签图像分割 (BSS) 模块, 其结构如图 4 所示, 用于对输入图像进行规范化处理并提取骨签主体区域.

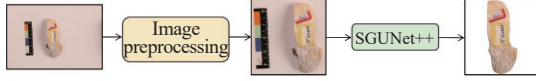


图 4 BSS 模块结构图

该模块由图像预处理与语义分割两个阶段组成: 首先通过预处理操作对原始图像进行背景抑制与尺度统一, 减少噪声干扰并提升图像质量; 随后利用 SGUNet++ 分割网络对骨签区域进行精确分割, 获得结构完整、边界清晰的骨签主体图像. 骨签图像分割结果示例如图 5 所示.

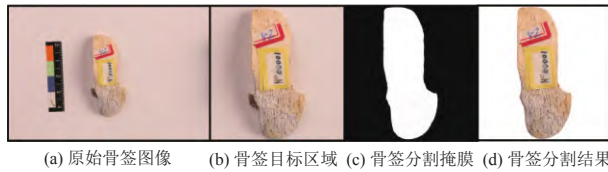


图 5 骨签图像分割结果示例图

(1) 图像预处理阶段

设原始输入图像为 $I \in \mathbb{R}^{H \times W \times C}$, 其中 H 、 W 分别表示图像高度与宽度, C 为通道数. 首先将原始骨签图像 (图 5(a)) 裁剪至骨签主体区域范围, 并调整至统一尺寸 224×224 像素, 以满足后续模型对输入尺度的固定要求. 尺度规范化过程能够有效消除样本间尺寸差异, 保证输入数据在空间维度上的一致性, 同时提升模型在训练与推理阶段的计算效率与稳定性. 随后对图像进行像素强度标准化, 使每通道的均值为 0、标准差为 1, 计算过程如式 (1) 所示. 其中 μ 和 σ 分别为图像像素均值和标准差. 经过该阶段处理后, 得到规范化骨签图像 I_p (图 5(b)), 为分割网络提供一致且高质量的输入.

$$I_p = \frac{I_C - \mu}{\sigma} \quad (1)$$

(2) 骨签图像分割阶段

该阶段提出一种面向骨签图像的高效背景分割方法, 即在标准 UNet++ 结构的基础上, 进一步融合通道注意力机制、高效特征生成与残差增强机制, 构建了一个轻量化且高效的深度神经网络模型 SGUNet++, 其结构图如图 6 所示, 通道注意力机制通过 SE (squeeze-

and-excitation) 模块^[19]实现, Residual Ghost block 结构包含了高效特征生成 (采用 Ghost 模块^[20]实现) 与残差增强机制 (通过 residual connection 完成). 具体结构如图 7 所示.

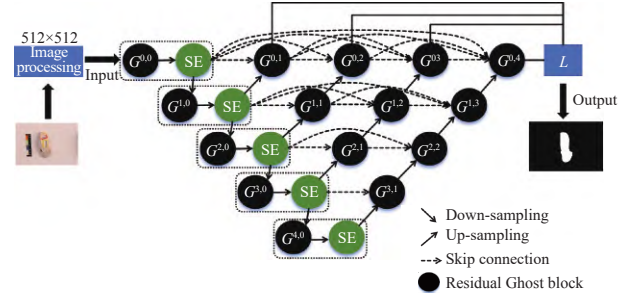


图 6 SGUNet++ 网络结构图

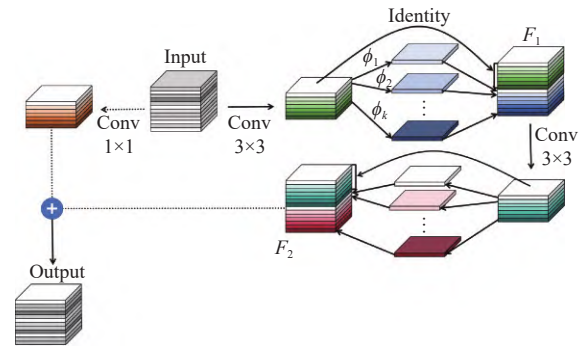


图 7 Residual Ghost block 结构图

编码路径由多级残差 Ghost 卷积层与通道注意力层共同构成. 对于输入特征图 $I_p \in \mathbb{R}^{H \times W \times C}$, 首先, 通过一个标准卷积生成成本征特征 F_{primary} , 如式 (2) 所示. 其中, r 为压缩比 ($r=2$).

$$F_{\text{primary}} = \text{Conv}_{k \times k}(I_p) \in \mathbb{R}^{C' \times H \times W}, C' = \left\lfloor \frac{C_{\text{out}}}{r} \right\rfloor \quad (2)$$

其次, 利用深度可分离卷积 (即逐通道卷积) 廉价地扩展出幻影特征 F_{ghost} , 如式 (3) 所示:

$$F_{\text{ghost}} = \text{DWConv}_{d \times d}(F_{\text{primary}}) \quad (3)$$

最终拼接得到完整输出, 如式 (4) 所示, 从而在显著减少参数数量的同时保留丰富的特征表达能力.

$$F = [F_{\text{primary}}, F_{\text{ghost}}] \in \mathbb{R}^{C_{\text{out}} \times H \times W} \quad (4)$$

为进一步提升训练稳定性与优化效率, 模块引入残差连接机制, 其前向过程如式 (5) 所示:

$$Y = \begin{cases} H(X) + X, & C_{\text{in}} = C_{\text{out}} \\ H(X) + \text{BN}(\text{Conv}_{1 \times 1}(X)), & C_{\text{in}} \neq C_{\text{out}} \end{cases} \quad (5)$$

其中, $H(\cdot)$ 为主干路径, 即双 Ghost 模块串联. 所有卷积操作均采用反射填充 (reflect padding) 以缓解边界伪影, 并使用 LeakyReLU 激活函数增强非线性建模能力, 如式 (6) 所示, 形成完整的非线性输出.

$$\text{LeakyReLU}(x) = \max(0, x) + \alpha \min(0, x), \alpha = 0.01 \quad (6)$$

为进一步增强特征选择性, 在每一级残差 Ghost 卷积层的输出端引入通道注意力机制, 即 SE 模块. 首先通过全局平均池化对通道特征进行压缩, 得到每个通道的统计描述, 如式 (7) 所示:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Y_c(i, j) \quad (7)$$

其中, H 与 W 分别为特征图的高与宽. 接着通过两层全连接网络实现非线性映射, 如式 (8) 所示:

$$s = \sigma(W_2 \delta(W_1 z)) \quad (8)$$

其中, $\delta(\cdot)$ 为 ReLU 激活函数, $\sigma(\cdot)$ 为 Sigmoid 函数, $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$, r 为通道压缩率. 得到的通道权重 s 用于重新标定原始特征, 如式 (9) 所示:

$$Y' = s \otimes Y \quad (9)$$

其中, $\otimes$ 表示通道间逐元素相乘操作.

解码路径通过密集式跳跃连接与多层次特征融合策略, 实现了低层细节信息与高层语义表征之间的充分交互. 每一级解码节点不仅与其对应尺度的编码特征建立直接连接, 还与同层级的前序解码节点形成多分支级联关系, 从而构建出嵌套式的信息流传递结构. 以第 i 层第 j 个节点特征为例, 其输入由当前层的直接编码特征 $X_{i,j}$ 与上一级解码结果共同组成, 如式 (10) 所示:

$$X_{i,j} = f_{i,j}([X_{i,0}, X_{i,1}, \dots, X_{i,j-1}, u(X_{i+1,j-1})]) \quad (10)$$

其中, $[\cdot]$ 表示通道拼接操作, $u(\cdot)$ 表示双线性上采样函

数, $f_{i,j}(\cdot)$ 表示由残差 Ghost 卷积层实现的非线性特征变换. 该多级嵌套结构使得同一解码层能够同时接收来自不同语义深度的特征, 从而建立跨层级的上下文关联. 此外, 特征融合模块中仍保留 SE 机制以增强语义一致性, 使融合过程能够自适应地分配不同来源特征的权重, 如式 (11) 所示:

$$\tilde{X}_{i,j} = s_{i,j} \cdot X_{i,j}, s_{i,j} = \sigma(W_2 \delta(W_1 \text{GAP}(X_{i,j}))) \quad (11)$$

在完成多层次特征融合后, 解码路径顶层输出特征被送入二分类的分类头, 用于生成最终的分割预测结果. 为实现高维特征向类别空间的投影, 使用一个 1×1 卷积执行逐像素线性变换, 将输入特征图通道从 C 压缩至 1, 如式 (12) 所示:

$$L = \text{Conv}_{1 \times 1}(\tilde{X}_{0,4}) \quad (12)$$

其中, 卷积核形状为 $1 \times 1 \times C \times 1$, C 为当前特征的通道数. 卷积输出经过 Sigmoid 激活函数转化为概率域预测将每个像素的值压缩至区间 (0, 1), 从而解释为属于前景类别的概率, 得到分割掩模.

将预处理后的骨签图像 (图 5(b)) 与其对应的分割结果 (图 5(c)) 在像素层面进行逐点相乘, 仅保留属于骨签主体区域的像素信息, 如式 (13) 所示:

$$\text{out}(i, j) = \text{img}(i, j) \wedge \text{lab}(i, j) + 255 \times (1 - \text{lab}(i, j)) \quad (13)$$

其中, out 表示输出图像, img 表示输入图像, lab 表示掩模图像, (i, j) 表示像素位置. 从而有效去除背景干扰, 得到结构规范、背景一致的骨签图像 (图 5(d)).

3.2 双视角特征编码 (DVFE) 模块

针对汉长安城出土骨签文物在正、反面成像过程中所呈现的显著结构差异, 本文在特征编码阶段采用正反面并行的建模策略, 利用 DVFE 模块完成骨签特征提取, 如图 8 所示.

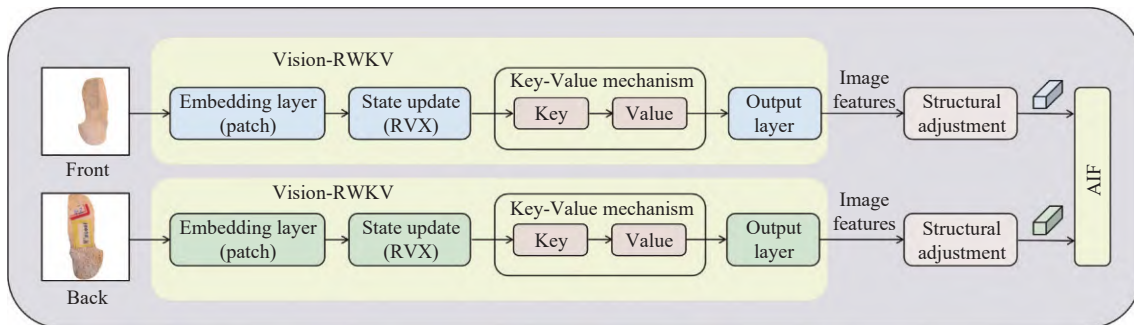


图 8 DVFE 模块结构图

由于骨签正反两侧在刻痕分布、表面风化形态以及局部纹理连续性等方面存在系统性差异,若采用单一共享编码器进行特征提取,模型需要在同一参数空间内同时适配两类差异显著的视觉模式,容易导致特征表示在判别维度上的相互干扰,从而削弱高层语义建模的稳定性与区分能力.因此,本文构建两个结构一致但参数独立的 Vision-RWKV 视觉编码分支,分别以分割后的骨签正面图像 I_f 与反面图像 I_b 作为输入,实现特征的并行提取与独立建模.

对于输入图像 $I \in \mathbb{R}^{H \times W \times C}$, 首先将其划分为 N 个大小为 $P \times P$ 的非重叠图像块 (Patch), 总块数由式 (14) 确定:

$$N = \frac{H \times W}{P^2} \quad (14)$$

其中, $H \times W$ 是图像分辨率, P 为 Patch 边长, C 为通道数 (对于 RGB 图像, $C=3$). 在 Patch 表示学习阶段, 对每个 Patch 进行展平操作, 将其从三维张量转换为一维向量, 即每个 Patch 的原始维度 $P \times P \times C$ ($224 \times 224 \times 3$) 将被转换为一个长度为 $P^2 C$ 的一维向量 $p_i \in \mathbb{R}^{P^2 C}$. 通过线性映射将展平后的 Patch 向量投影至高维特征空间, 得到 Patch 级嵌入表示, 如式 (15) 所示:

$$x_i = W_e p_i, \quad i = 1, 2, \dots, N \quad (15)$$

其中, $W_e \in \mathbb{R}^{D \times P^2 C}$ 为可训练权重矩阵, D 表示 Patch 嵌入的特征维度. 由此, 输入图像被转换为长度为 N 的视觉 token 序列 $x = \{x_1, x_2, \dots, x_N\}$.

在此基础上, Vision-RWKV 将 RWKV 模型中的 Key-Value 累积机制引入至 Patch 的空间序列中. 与基于自注意力的全局建模方式不同, 该机制通过线性递归的形式在空间上逐步聚合上下文信息, 从而避免显式构建高维注意力矩阵. 每个 Patch 嵌入分别通过线性映射生成对应的 Key 与 Value 表示, 如式 (16) 所示:

$$k_i = W_k x_i, \quad v_i = W_v x_i \quad (16)$$

模型沿 Patch 的空间顺序对历史 Key-Value 信息进行递归累积, 其状态更新过程如式 (17) 所示:

$$s_i = \lambda s_{i-1} + k_i \odot v_i \quad (17)$$

其中, s_i 表示第 i 个 Patch 的空间上下文累积状态, $\lambda \in (0, 1)$ 为衰减系数, 用于调节不同空间位置 Patch 对当前状态的影响程度. 当 λ 趋近于 1 时, 模型更倾向于整合远距离 Patch 信息, 从而强化整体结构建模; 当 λ 趋近于 0 时, 模型更强调局部邻域特征, 有利于细粒度刻

痕与纹理细节的表达. 在参数初始化阶段, 采用对数参数化策略 (即 $\lambda = \sigma(\log \tau)$, 其中 τ 为可学习参数), 并将初始值设定为对应中等记忆长度的状态 (即 $\lambda_0 \approx 0.9$), 以在训练初期平衡全局收敛速度与局部特征学习能力. 由此, λ 实际上控制了空间感受野的衰减速度与上下文记忆长度. 通过该递归形式, 模型能够在空间维度上持续整合来自不同区域的视觉信息, 使特征表示逐步引入更大范围的上下文结构. 基于累积状态 s_i , 第 i 个 Patch 的上下文感知特征进一步的定义如式 (18) 所示:

$$h_i = \delta(k_i + s_i), \quad h_i \in \mathbb{R}^D \quad (18)$$

其中, k_i 表示当前 Patch 的局部语义信息, s_i 表示跨区域整合后的上下文信息, 二者相加实现局部与全局特征的融合, ReLU 激活函数增强了非线性表达能力. 最终得到的 Patch 级隐藏状态矩阵 $H = [h_1, h_2, \dots, h_N] \in \mathbb{R}^{N \times D}$ 构成 Vision-RWKV 的主要输出, 其在 Patch 粒度上同时编码了局部纹理信息与全局空间上下文关系.

在图像级表示生成阶段, 本文采用全局平均池化 (global average pooling, GAP) 对所有 Patch 的上下文特征进行整合, 得到单一的全局特征向量 $f_G \in \mathbb{R}^{224 \times 224 \times 3}$, 如式 (19) 所示:

$$f_G = \frac{1}{N} \sum_{i=1}^N h_i \quad (19)$$

选择均值池化的原因在于其能够稳定整合各区域响应, 避免少量高响应 Patch 主导整体表示, 从而提高特征表达的鲁棒性与平滑性, 更适合骨签表面连续结构与整体形态的建模需求.

正、反面图像经过两个独立 Vision-RWKV 编码器提取的特征向量维度可能存在差异或在数值分布上不完全一致. 若直接对其进行拼接或融合, 高维特征可能在计算中占据主导, 从而导致模型在训练时对某一视角的图像特征产生偏向, 影响多视角特征融合的均衡性与表达效果. 为解决这一问题, 本文对正、反面图像特征进行线性映射, 使其统一到相同的维度, 以保证融合过程的平衡性和特征表示的稳定性. 具体而言, 通过全连接层对每个分支的图像特征进行变换, 将其映射为统一维度 D_u , 如式 (20) 所示:

$$\tilde{F}^v = W^v F^v, \quad v \in \{f, b\} \quad (20)$$

其中, $W^v \in \mathbb{R}^{D_u \times D_v}$ 为可训练权重矩阵, D_v 为原始特征维度, $\tilde{F}^v$ 为映射后的统一特征表示. 在本文实验中, 统

一维度设置为 1024, 这既能够降低计算量和内存开销, 又保留高维特征的核心信息, 从而确保正反面特征在融合阶段的均衡性和判别力。

3.3 自适应交互融合 (AIF) 模块

为充分刻画骨签正反面特征之间的互补关系, 本文引入动态交叉特征融合策略, 对来自不同视角的图像特征进行联合建模, 采用 AIF 模块实现。该方法基于文献[21–23]的相关思想, 通过动态生成融合权重, 自适应调节正反面特征在融合过程中的贡献程度。相较于静态加权或简单拼接等固定融合方式, 该策略能够根据输入样本特征分布的变化灵活调整融合权重, 从而更有效地反映正反面信息在不同骨签样本中的相对重要性。设骨签正、反面图像经并行 Vision-RWKV 编码并完成维度统一后的特征分别为 $\tilde{F}^f, \tilde{F}^b \in \mathbb{R}^{D_u}$, 动态交叉特征融合过程主要包括视角相关性建模 (view relevance modeling, VRM)、双向跨视角交叉建模 (bidirectional cross-viewpoint modeling, BCVI)、视角重要性动态调节 (dynamic adjustment of view weight, DVW) 及自适应融合表示 (adaptive fusion representation, AFR) 生成这 4 个阶段, 其整体结构如图 9 所示。

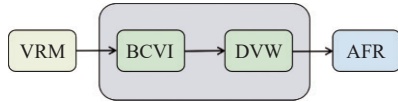


图 9 AIF 模块结构图

在正反面特征对齐之后, 首先对两种视角在全局语义层面的相关性进行建模。具体而言, 将正反面特征拼接形成联合表示, 并通过线性映射与非线性激活生成初步权重响应如式 (21) 所示:

$$r = \sigma(W_r [\tilde{F}^f \parallel \tilde{F}^b] + b_r) \quad (21)$$

其中, W_r 采用 Xavier Uniform 初始化, b_r 初始化为 0, $[\cdot \parallel \cdot]$ 表示特征拼接, $\sigma(\cdot)$ 为 Sigmoid 激活函数。随后, 采用 Softmax 函数对权重进行归一化, 得到正面对应的动态融合系数, 如式 (22) 所示:

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^2 \exp(w_j)}, \alpha_i \in \mathbb{R} \quad (22)$$

其中, $i \in \{f, b\}$ 分别对应正面与反面特征。该过程保证融合权重在样本层面自适应变化, 同时满足数值稳定性与可解释性。本阶段不直接生成融合结果, 而是用于

刻画正反面特征在语义空间中的整体关联程度, 为后续交叉建模提供全局引导信息。

在双向跨视角交叉建模阶段, 为显式刻画两视角特征之间的关联关系, 本文引入跨视角注意力机制, 对正反面特征进行双向交互建模。以正面特征 $\tilde{F}^f$ 作为查询向量, 反面特征 $\tilde{F}^b$ 作为键和值, 构建正面到反面的跨视角响应 Z_1 , 如式 (23) 所示:

$$Z_1 = \text{Softmax} \left(\frac{(W_q \tilde{F}^f)(W_k \tilde{F}^b)^T}{\sqrt{d}} \right) (W_v \tilde{F}^b) \quad (23)$$

其中, W_q 、 W_k 、 W_v 为可学习映射参数, 均遵循 Xavier Uniform 初始化策略, 对应偏置项设为 0, d 为特征维度归一化因子。该过程能够引导正面特征在更新时重点关注反面图像中具有判别意义的结构与纹理区域。对称地, 以反面特征为查询、正面特征为键和值, 可得到反面到正面的交叉响应 Z_2 , 从而实现正反面信息的双向交互建模。

考虑到不同骨签样本在正反面成像质量与信息完整性方面存在显著差异, 固定权重的融合方式难以适配复杂场景。为此, 本文引入视角重要性动态调节机制, 根据正反面特征的联合表示自适应生成融合权重, 如式 (24) 所示:

$$\alpha = \text{Softmax}(W_\alpha [\tilde{F}^f \parallel \tilde{F}^b] + b_\alpha), \alpha \in [\alpha_f, \alpha_b] \quad (24)$$

其中, W_α 与 b_α 为可训练参数。该权重向量用于刻画正反面视角在当前样本中的相对重要性, 使模型能够根据输入特征的整体分布特性, 自适应选择更具判别力的视角信息参与融合。

在完成双向交叉建模与视角重要性调节之后, 本文利用动态权重对双向交叉特征进行逐元素加权求和, 得到最终融合结果如式 (25) 所示:

$$F_{\text{fusion}} = \alpha_f \odot Z_1 + \alpha_b \odot Z_2 \quad (25)$$

其中, $\odot$ 表示逐元素乘法。该融合过程在特征维度层面对正反面交互信息进行自适应平衡, 使最终表示同时保留两种视角的互补结构特征与细粒度纹理信息, 从而为后续判别任务提供更加稳定且具有判别力的高层语义表示。

3.4 分类输出 (CO) 模块

设输入骨签图像经预处理后表示为特征张量 X , 本文构建的卷积神经网络分类器通过多层卷积与池化

操作对其进行逐级特征抽象,其整体网络结构如图10所示。

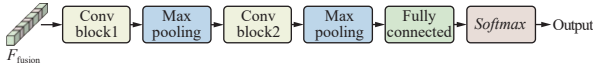


图10 分类输出模块结构图

在特征提取阶段,网络采用标准卷积层对输入特征进行局部建模,通过线性卷积运算与非线性激活函数提取骨签图像中的纹理与结构信息,如式(26)所示:

$$Y^{(l)} = \sigma(W^{(l)} \times X^{(l)} + b^{(l)}) \quad (26)$$

为降低特征冗余并提高计算效率,在卷积层后引入常规池化操作进行空间下采样,其过程如式(27)所示,用于压缩特征分辨率并增强表示的稳定性。

$$P_{i,j}^{(l)} = \max_{(u,v) \in \Omega_{i,j}} Y_{u,v}^{(l)} \quad (27)$$

在分类阶段,经过多层特征提取后,网络通过全连接层将高维特征映射至类别空间,并采用 Softmax 函数进行归一化处理,如式(28)所示:

$$P_k = \frac{\exp(s_k)}{\sum_{c=1}^C \exp(s_c)} \quad (28)$$

模型训练时使用交叉熵损失函数作为优化目标,用以衡量预测结果与真实标签之间的差异,其定义见式(29):

$$L_{CE} = - \sum_{k=1}^C y_k \log P_k \quad (29)$$

4 实验环境、数据集和评价指标

4.1 实验环境

实验 PC 为 64-bit Windows 11 操作系统,配备 16 GB RAM、Intel Core i9-14900HX CPU 和 NVIDIA GeForce RTX 5070 GPU。所有实验均在 Python 3.10 环境下采用 PyTorch 框架编程完成。

4.2 数据集

本研究使用的数据集来源于汉长安城未央宫遗址出土的骨签图像,共包含 8 640 对经过预处理及分割的骨签正、反面图像样本。所有样本依据其物理特征被划分为 12 个类别(见表1)。数据集按 6:2:2 的比例划分为训练集、验证集和测试集。为增强模型泛化能力,训练阶段对图像进行了随机数据增强。

表1 数据集划分表(对)

数据集类别	训练集	测试集	验证集	类别总数	
完整0	黄色00	432	144	144	720
	白色01	432	144	144	720
	黑色02	432	144	144	720
	灰色03	432	144	144	720
上部1	黄色10	432	144	144	720
	白色11	432	144	144	720
	黑色12	432	144	144	720
	灰色13	432	144	144	720
下部2	黄色20	432	144	144	720
	白色21	432	144	144	720
	黑色22	432	144	144	720
	灰色23	432	144	144	720
总计	5 184	1 728	1 728	8 640	

4.3 评价指标

为全面评估模型在骨签图像多分类任务中的整体性能,本文选取准确率(Accuracy, Acc)、精确率(Precision, $Prec$)、召回率(Recall, Rec)和加权 $F1$ 分数(weighted $F1$ -score, $wF1$)作为主要评价指标,各评价指标的具体计算形式统一定义如式(30)–(33)所示:

$$Acc = \frac{1}{N} \sum_{i=1}^K T_i \quad (30)$$

$$Prec = \frac{1}{K} \sum_{i=1}^K \frac{T_i}{T_i + P_i} \quad (31)$$

$$Rec = \frac{1}{K} \sum_{i=1}^K \frac{T_i}{T_i + N_i} \quad (32)$$

$$wF1 = \sum_{i=1}^K \left(\frac{n_i}{N} \cdot \frac{2 \cdot Prec_i \times Rec_i}{Prec_i + Rec_i} \right) \quad (33)$$

其中, T_i 、 P_i 、 N_i 分别表示第 i 类中的真正类样本数、假正类样本数和假负类样本数, K 为类别总数, N 表示样本总数, n_i 为第 i 类样本数量, $\frac{n_i}{N}$ 为该类别在整体样本中的权重。

上述指标的取值范围均限定在 $[0, 1]$ 区间内,数值越大表示模型性能越优。其中,准确率用于刻画模型在全部样本上的整体判别能力,反映预测结果与真实标签一致的样本在总样本中的比例,是衡量模型总体分类效果的基础指标。精确率关注模型预测结果的可靠性,通过统计模型预测为某一类别的样本中真实属于该类别的比例,衡量模型在“预测为正”时的准确程度。为适应多分类场景,本文对各类别分别计算精确率,并对结果

进行平均处理,以避免单一类别样本分布不均对评价结果产生偏置.召回率则侧重于模型对真实样本的识别能力,通过衡量某一类别中被模型成功识别出来的样本比例,反映模型对该类别样本的覆盖情况.同样地,在多分类任务中,本文采用对各类别召回率取平均的方式,综合评估模型对不同类别样本的整体检出能力.此外,为进一步兼顾精确率与召回率之间的平衡性,本文引入加权 $F1$ 分数作为综合评价指标.加权 $F1$ 分数首先在类别层面计算精确率与召回率的调和平均,以获得每一类别对应的 $F1$ 分数,随后根据各类别在整体样本中的占比进行加权求和,从而在样本分布不均衡的情况下,能够更加客观地反映模型的综合分类性能.

为定量评价骨签背景分割模型的性能,本文采用 $Dice$ 系数和平均交并比 ($mIoU$) 作为评价指标. $Dice$ 系数用于衡量预测分割区域与真实标注区域之间的重叠程度,其计算公式为:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (34)$$

平均交并比 ($mIoU$) 通过计算各类别预测区域与真实区域交集和并集的比值,并对所有类别取平均,其计算公式为:

$$mIoU = \frac{1}{C} \sum_{i=1}^C \frac{TP_i}{TP_i + FP_i + FN_i} \quad (35)$$

其中, TP 、 FP 和 FN 分别表示真阳性、假阳性和假阴性像素数, C 表示参与评价的类别数量.两项指标的取值范围均为 $[0, 1]$,数值越大,表明模型对骨签区域与背景的分割结果越准确.

5 实验

5.1 双视角并行协同建模骨签分类模型消融实验

本研究在正式监督训练之前,采用基于对比学习的预训练策略对网络进行初始化,使模型能够提前学习骨签数据中的结构与纹理特征,从而获得更合理的参数初始状态.在实验设置中,批量大小统一设为 16,训练轮次设定为 200,其余超参数在不同训练策略下保持一致,以保证实验结果的公平性与可比性.

图 11 展示了在是否引入预训练权重条件下,模型训练损失与验证损失随训练轮次变化的趋势.黄色与绿色曲线分别对应未加载预训练权重时的训练损失与验证损失曲线,红色与蓝色曲线分别对应加载预训练

权重时的训练损失与验证损失曲线.

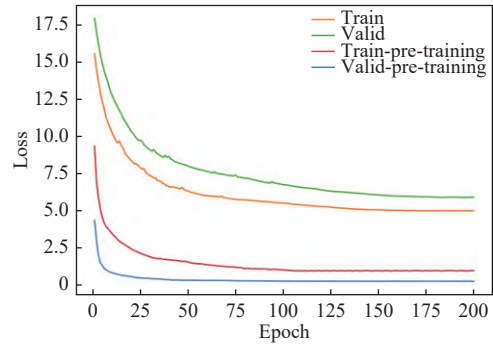


图 11 不同训练策略下模型损失曲线

可以观察到,采用预训练权重初始化的模型在训练初期即表现出较低的损失水平,并在约 115 轮后进入稳定收敛阶段,此后损失变化幅度较小,表明模型已完成有效优化.相比之下,未使用预训练权重的模型在整个训练过程中损失下降速度较慢,且在训练末期仍存在一定幅度的波动,未表现出与预训练模型相当的稳定收敛特性.

从整体趋势来看,采用预训练策略的模型在训练集与验证集上均取得了更低的损失收敛水平,验证损失曲线也表现出更好的平滑性与稳定性,表明该模型在骨签颜色与部位分类任务中具备更优的泛化能力.上述结果充分验证了基于对比学习的预训练方法在提升模型训练效率与性能方面的有效性.

为了评估骨签特征提取模块与动态交叉特征融合模块的有效性,本研究进行了 4 组不同配置的实验.第 1 组实验使用 Vision-RWKV 模型提取骨签图像的正面特征进行分类,简记为 VR-Front;第 2 组实验使用 Vision-RWKV 模型提取骨签图像的反面特征进行分类,简记为 VR-Back;第 3 组实验使用两个并行的 Vision-RWKV 模型分别提取骨签图像的正、反面特征,再将提取的特征进行简单拼接,随后进行统一分类,简记为 VR-F&VR-B;第 4 组实验则采用本文提出的分类方法进行分.实验结果如表 2 所示.

从实验结果可以看出,仅依靠单一骨签图像正面或反面特征进行分类时,模型在各项指标上的表现相对有限,Accuracy 分别为 88.68% 和 88.17%,wF1 分别为 88.31% 和 87.84%,表明单一视角下的特征难以全面刻画骨签信息.通过将正、反面特征进行简单拼接并输入统一分类网络,模型的各项评估指标均有所提

升,其中 Accuracy 提升至 91.56%, wF1 提升至 90.70%,表明多视角特征的联合利用能够增强特征表达的完整性与判别能力. 进一步引入本文提出的特征交互与加权融合策略后,模型在 Accuracy 和 wF1 上分别达到 94.76% 和 93.56%,相较于简单拼接方法实现了约 3.2 个百分点的提升. 尤其是在精确率方面的提升,表明模型在分类过程中误报率明显下降,特征交互模块有效增强了模型对类别边界的判别能力,同时 wF1 的提高也反映了模型在多类别任务上的稳定性和泛化能力显著增强. 综合来看,动态交叉特征融合不仅提升了模型分类性能,也优化了模型在复杂骨签样本下的可靠性与适应性.

表 2 双视角协同建模的并行视觉分类模型消融实验结果表

模型结构	Accuracy (%)	Precision (%)	Recall (%)	wF1 (%)	Params (M)	GFLOPs
VR-Front	88.68	88.29	88.43	88.31	217.94	18.52
VR-Back	88.17	87.61	87.90	87.84	217.94	18.52
VR-F&VR-B	91.56	90.31	91.28	90.70	435.88	37.53
Ours	94.76	93.85	93.43	93.56	456.82	39.80

注: 加粗数据表示最佳结果

为量化各模块对模型资源消耗的影响,本文对不同结构配置下的参数规模 (Params) 与计算量 (GFLOPs) 进行了统计分析. 结果显示,正反面特征编码器在整体模型中占据主要计算资源. 引入交叉特征融合模块后,模型参数由 435.88M 增至 456.82M, GFLOPs 由 37.53 增至 39.80. 尽管增加了一定计算量,但提升幅度在可控范围内,同时带来了显著的性能增益,验证了该模块在保持模型可用性与计算效率的同时,有效强化了特征融合与判别能力.

5.2 SGUNet++骨签分割模型消融实验

本研究所使用的数据集来源于汉长安城未央宫遗址出土的骨签图像,共 500 张. 使用 Adobe Photoshop 2021 (PS) 图像处理软件的物体自动提取算法对骨签主体进行初步分割,再进行手动修正骨签边缘和细节,生成高质量的分割标签图像. 最终,得到 500 组实验数据,每组数据均由匹配的骨签图像及其对应的蒙版构成. 图 12 展示了 3 组实验数据样本. 设 100 组骨签图像作为独立测试集,其余按照 8:2 的比例划分为训练集和验证集.

为验证 SGUNet++模型对骨签高精度分割的有效性,本研究设计并开展了一系列实验. 实验将 batch size 设定为 2, epochs 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.

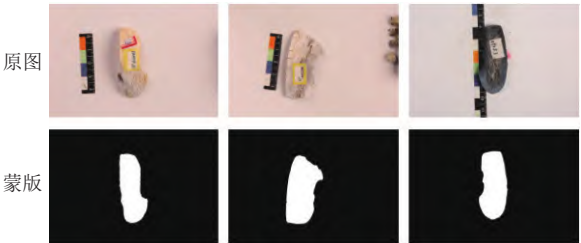


图 12 实验数据样本示例

从图 13 中可以看出,原始 UNet++ 的损失下降速度相对较慢,在约 80 个 Epoch 后趋于稳定,最终收敛在 10^{-2} 左右. 引入 SE 模块后,模型的收敛曲线整体下移,损失下降速度加快,这证明了通道注意力机制有助于提升特征表达质量. 同样地, Ghost 模块的应用通过减少冗余卷积计算,使网络在去除大量参数的情况下仍能保持更好的特征建模能力,其曲线最终收敛至约 3×10^{-3} . UNet++-RG 模型在训练初期表现出良好的下降趋势,体现出该结构对梯度传播和训练稳定性的积极作用. 而本研究提出的 SGUNet++模型收敛效果最为突出,在约 50 个 Epoch 内损失即降至约 10^{-4} 并保持稳定. 其具备最快的收敛速度和最低的收敛损失,这充分说明所提结构显著增强了特征提取与优化效率,验证了模型改进机制的有效性. 上述结果表明,SGUNet++不仅在准确性方面超越了他模型,而且在训练效率上也表现出色,是骨签高精度分割的一个有效解决方案.

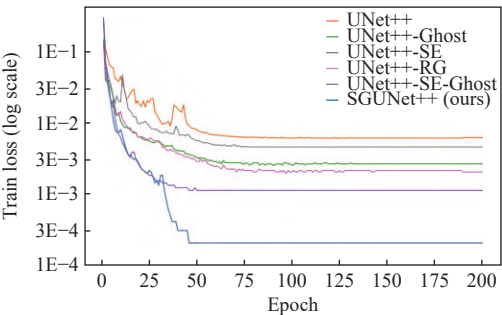


图 13 SGUNet++模型消融实验损失曲线图

实验结果如表 3 所示,实验结果表明,各模块均在不同维度上对模型性能产生了正向作用. 基础 UNet++ 模型的 Dice 系数为 0.9760, mIoU 为 0.9653,为后续改进提供了基准参考. 引入 SE 注意力机制后,模型 Dice 提升至 0.9892, mIoU 提升至 0.9780, Recall 达到 0.9878,表明通道注意力机制能够有效增强特征表达能力,但

参数量增至 36.85M, 推理速度小幅下降至 28.13 f/s. 将标准卷积替换为轻量化 Ghost 模块后, 模型参数量由 36.62M 降至 18.37M, GFLOPs 减半至 69.76, 同时 *Dice* 和 *mIoU* 分别提升至 0.9823 和 0.9746, 推理速度提高至 32.51 f/s, 验证了 Ghost 模块在降低计算开销的同时仍能保持较高分割性能. 进一步引入基于 Ghost 的 RG 模块, 模型参数量虽增加至 20.59M, 但 *mIoU* 提升至 0.9842, Recall 达到 0.9879, 表明 RG 结构在减少冗余计算的同时具备更强的特征生成能力. 在 SE 模块基础上融合 Ghost 模块, 模型在保持较低参数量 (18.6M)

和较高推理速度 (32.77 f/s) 的同时, *Dice* 提升至 0.9891. 最终, 本文提出的 SGUNet++通过在 SE 模块基础上进一步引入 RG 模块, 最终使 *Dice* 达到 0.9951, 较原始 UNet++提升约 1.95%; *mIoU* 提升至 0.9902, 提升约 2.49%; Recall 达到 0.9931, 均为所有模型中的最高值, 同时 GFLOPs 降至 83.23, 较原始模型降低约 39.7%. 整体而言, 本文提出的 SGUNet++在显著提升分割精度的同时有效降低了模型计算量, 这表明该结构在精度与效率之间取得了优异的平衡, 验证了所提设计策略的有效性和优越性.

表3 SGUNet++模型消融实验结果表

模型结构	<i>Dice</i>	<i>mIoU</i>	Recall	Params (M)	FPS (f/s)	单张预测时间 (ms)	GFLOPs
UNet++	0.9760	0.9653	0.9740	36.62	28.48	36.61	138.10
UNet++-SE	0.9892	0.9780	0.9878	36.85	28.13	35.53	138.13
UNet++-Ghost	0.9823	0.9746	0.9862	18.37	32.51	27.78	69.76
UNet++-RG	0.9875	0.9842	0.9879	20.59	31.23	31.96	81.19
UNet++-SE-Ghost	0.9891	0.9839	0.9906	18.6	32.77	29.24	69.80
SGUNet++ (ours)	0.9951	0.9902	0.9931	21.74	29.23	33.81	83.23

注: 加粗数据为本文方法的结果, 即分割效果最佳时对应的指标结果

5.3 双视角并行协同建模骨签分类模型对比实验

为系统评估本文方法在骨签图像分类任务中的有效性, 本文选取了多类主流标准图像分类模型作为对比, 包括传统卷积神经网络 ResNet-50、DenseNet-121, 高效网络 EfficientNet-B3, 深度学习分类模型 YOLOv13-m, 基于 Transformer 架构的 ViT-B/16、Swin-Base 以及 CLIP-ViT-L/14 (fine-tuned)^[24], 上述版本均为各自架构中应用最广、结构成熟的代表性版本, 同时能够体现本文方法通过正反面双视角特征并行提取与协同融合实现细粒度建模的优势. 所有模型批量大小统一设置为 16, 训练轮次为 200, 以确保充分收敛. 实验结果如表 4 所示.

表4 双视角协同建模的并行视觉分类模型对比实验结果 (%)

模型类型	Accuracy	Precision	Recall	wF1
ResNet-50	82.47	81.96	82.21	82.08
DenseNet-121	82.81	82.33	82.54	82.42
EfficientNet-B3	83.69	83.18	83.41	83.29
YOLOv13-m	85.32	84.76	85.01	84.88
ViT-B/16	86.74	86.29	86.48	86.38
Swin-Base	88.21	87.82	88.04	87.93
CLIP-ViT-L/14 (fine-tuned)	89.36	88.97	89.14	89.02
Ours	94.76	93.85	93.43	93.56

注: 加粗数据表示最佳结果

从表 4 中可见, 传统卷积神经网络在骨签分类任务中的准确率集中在 82%–84%, 难以捕捉跨区域判别性信息; YOLOv13-m 虽提升至 85.32%, 但在高类别相

似度样本上判别能力仍有限. 基于 Transformer 架构的模型 ViT-B/16 和 Swin-Base 通过全局建模将准确率提升至 86.74% 和 88.21%, 虽优于卷积网络, 但在复杂场景下仍存在误判. CLIP-ViT-L/14 经微调后达到 89.36% 的 Accuracy 和 89.02% 的 wF1, 表现最佳, 但提升幅度有限, 反映通用视觉语义模型在专业化文物图像任务中的适配性仍有空间.

相比之下, 本文方法在各项指标上均取得显著优势, Accuracy 和 wF1 分别达到 94.76% 和 93.56%, 相较 CLIP-ViT-L/14 分别提升 5.40 和 4.54 个百分点. 结果表明, 本文方法能够更有效地建模骨签图像中的判别特征, 在多类别任务中表现出更强的稳定性与鲁棒性, 验证了其在骨签分类场景中的有效性与实用价值.

5.4 SGUNet++骨签分割模型对比实验

本研究选取了 5 种图像分类方法进行对比实验, 包括以深度学习为代表的 U-Net^[15]、U²-Net^[25]、TransUNet^[26]和 HRNet^[27]图像分割模型以及本文所提 SGUNet++模型. 为了确保对比实验的公平性, 实验过程中, 所有模型的批量大小均设定为 2, 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 5 所示.

在主要评价指标中, SGUNet++的 *Dice* 系数、*mIoU* 和 Recall 均为所实验模型中的最高水平, 高达 0.9951、0.9902 和 0.9931. 在模型开销方面, SGUNet++的参数

量为 21.74M, 显著低于多个对比模型, 同时单张预测时间仅为 33.81 ms, 在保持高精度的同时具备更高的推理效率. 综合而言, SGUNet++在精度、计算量与推理速度上均实现了优于现有方法的性能, 在骨签图像分割任务中具备较强的适配性和应用性.

表 5 SGUNet++骨签分割模型对比实验结果表

模型结构	Dice	mIoU	Recall	Params (M)	单张预测时间 (ms)
U-Net	0.9802	0.9515	0.9820	95.06	76.87
U ² -Net	0.8210	0.6970	0.8978	173.6	59.19
TransUNet	0.9083	0.8320	0.8523	404.2	172.32
HRNet	0.8740	0.7762	0.8006	37.57	133.54
SGUNet++	0.9951	0.9902	0.9931	21.74	33.81

5.5 骨签分割结果可视化对比分析

为直观评估 SGUNet++模型在复杂骨签图像条件下的分割稳健性与边界保持能力, 本文构建了多场景可视化对比分析. 如图 14 所示, 各组样例分别展示了不同骨签图像在分割任务中的典型预测效果, 包括标准清晰样本 (图 14(a))、邻近干扰样本 (图 14(b))、边界不规则样本 (图 14(c))、复杂背景样本 (图 14(d)), 并对每类样本同时展示原始图像、真实标注掩码及模型预测掩码, 以并列方式进行视觉对照分析.

从实验结果可见, SGUNet++模型在多数情况下能够准确识别骨签的主体区域, 与真实标注掩码在形状与轮廓上保持高度一致, 尤其在整体结构完整、对比度清晰的样本中预测结果几乎重合; 对于存在局部凸起、缺口或形态不规则的样本, 预测轮廓仍能较好贴合真实边界, 仅在细微凹陷或尖锐转折处出现轻微偏差; 在含有复杂背景或邻近干扰物的样本中, 模型依然能够有效抑制背景噪声, 实现主体区域的稳定分割. 整体而言, SGUNet++模型在不同形态、尺度及成像条件下均具备良好的鲁棒性与边界刻画能力, 其通过多尺度特征融合机制, 在保持全局结构连贯的同时精准锁定局部细粒度特征, 显著降低了复杂考古场景下的分割误差.

为评估不同语义分割模型在考古骨签图像上的分割性能, 本文选取 3 组具有代表性的典型样本 (涵盖不同形态复杂度与背景干扰程度) 进行可视化对比分析, 如图 15 所示. 图中第 1 列为原始骨签图像, 第 2 列为人工标注的真值掩码 (label), 后续各列依次展示了 U-Net、U²-Net、TransUNet、HRNet 及本文所提 SGUNet++ 的预测结果; 红色方框标出了各模型预测出现误差的

关键区域, 以直观揭示其局限性. 从可视化细节来看, U-Net 虽在主体轮廓提取上整体稳定, 但在边缘细节处极易出现断裂, 尤其是对骨签表面的微小刻痕等细小结构识别能力不足, 导致轮廓连续性受损; U²-Net 则表现出严重的过分割倾向, 不仅将背景噪点误判为目标区域, 还在主体内部生成了虚假孔洞, 致使几何结构完整性严重受损, 预测掩码呈现明显的破碎化特征; TransU-Net 凭借全局建模优势使边界平滑度较好, 然而在骨签尖端、窄颈等细粒度区域却出现了边界模糊与局部缺失, 其边缘贴合度较真实掩码存在明显偏差, 未能精准捕捉细微形态变化; HRNet 在处理粗大目标时表现尚可, 但对纤细结构的分割存在系统性偏离, 往往导致关键特征丢失. 相比之下, SGUNet++在保持主体轮廓完整性的同时, 对局部缺损与边界过渡的处理更为鲁棒, 有效抑制背景干扰, 增强了刻痕、凹槽等细粒度结构的特征表达, 在 3 组样本中均无明显伪影或缺失, 预测掩码与真实标注高度吻合. 该对比直观揭示了当前主流架构在处理细粒度文物分割任务中的局限性, 同时验证了 SGUNet++在复杂背景与不规则形态下的更强泛化能力与有效性.

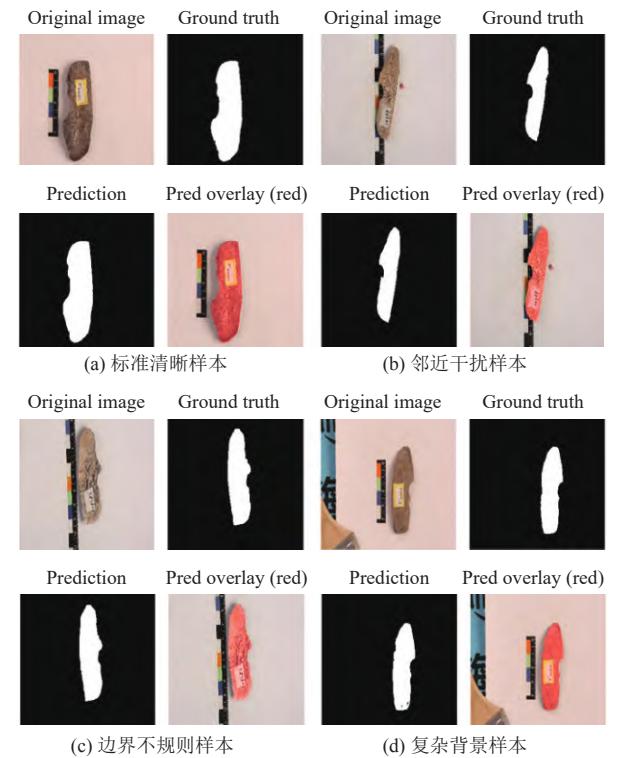


图 14 SGUNet++模型多场景可视化对比图

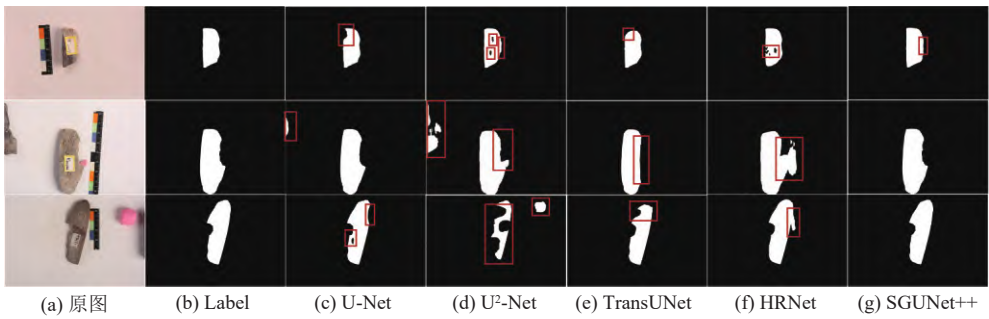


图 15 不同语义分割模型进行骨签分割对比可视化图

5.6 骨签分类结果可视化对比分析

为直观评估本文提出的骨签分类方法性能, 研究设计了骨签分类结果的可视化对比分析, 图 16 展示了部分样本的分类结果. 实验结果表明, 绝大多数样本的

模型预测标签与人工标注结果完全一致, 较高的匹配度充分验证了该分类方法在区分骨签表面颜色特征方面的有效性与可靠性, 能够为出土文物状态评估提供稳定依据.

数据集 原始编号	骨签正面原始 图片示例	经预处理骨签 正面图片示例	骨签反面原始 图片示例	经预处理骨签 反面图片示例	真实标签	模型预测标签
0540					完整黄色 (00)	00
3017					上部黄色 (10)	10
8478					上部白色 (11)	11
4427					上部黑色 (12)	12
0556					上部灰色 (13)	13
8231					下部白色 (21)	21
4710					下部黑色 (22)	22
3295					下部灰色 (23)	23
1897					下部黑色 (22)	20
3974					完整白色 (01)	00

图 16 部分骨签分类结果可视化对比图

尽管该方法分类结果整体表现良好,但个别样本仍存在分类偏差,其主要原因可能与地下埋藏环境引起的复杂风化效应有关。例如,编号1897的样本(真实标签为“下部黑色”)被误判为“下部黄色”,推测源于骨质长期受土壤矿物质沁染,原本黑色区域呈现深褐或暗黄色过渡色调,从而干扰模型的颜色判别;编号3974(真实标签为“整体白色”)被误判为“整体黄色”,则可能由于骨签表面轻微氧化或土沁作用,使视觉表现偏离标准白色定义。上述现象说明,虽然该方法整体具备较高可信度,但在处理长期埋藏所导致的非典型色泽变化时,仍需结合考古专业经验进行人工复核,以弥补单纯依赖视觉特征判别可能产生的误差。

6 结论与展望

本文针对汉长安城出土骨签图像的多类别分类任务,提出了一种基于骨签主体分割与双视角并行 Vision-RWKV 特征建模的分析方法。在骨签主体分割模块中,基于 UNet++ 框架引入通道注意力机制、轻量化特征生成策略与边缘引导约束,实现对骨签主体区域的高精度分割,有效抑制背景噪声和无关区域干扰,使模型能够聚焦骨签关键结构与纹理特征。而后采用骨签正反面图像的并行 Vision-RWKV 特征提取与协同融合策略,不仅充分挖掘两面信息的互补性,还增强了特征表示的判别性和完整性,从而显著提升分类模型在复杂骨签样本上的识别能力、稳定性及泛化性能。实验结果表明,该分类模型在 Accuracy 与加权 F1 分数指标上均明显优于传统卷积神经网络和 Transformer 类模型,验证了所提方法在骨签多类别分类任务中的有效性与稳定性。

本研究在骨签分类任务中取得了较好的性能,为考古学领域的文物数字化保护提供了重要技术支持。但由于骨签特征复杂、样本间高度相似且训练数据量有限,其在精细分类上的表现仍有提升空间。未来研究将通过扩大数据集、优化特征建模与融合策略,并探索多模态信息的融合应用,进一步增强模型在复杂骨签场景下的判别能力。

参考文献

- 刘贤鹏. 汉未央宫中央官署类骨签研究. 文物, 2025(7): 47–56.
- 李毓芳. 骨签的发掘、整理与出版. 中国书法, 2025(5): 54–80. [doi: 10.3969/j.issn.1003-1782.2025.05.007]
- Liao J, Guo LP, Jiang L, *et al.* A machine learning-based feature extraction method for image classification using ResNet architecture. Digital Signal Processing, 2025, 160: 105036. [doi: 10.1016/j.dsp.2025.105036]
- Arulananth TS, Prakash SW, Ayyasamy RK, *et al.* Classification of paediatric pneumonia using modified DenseNet-121 deep-learning model. IEEE Access, 2024, 12: 35716–35727. [doi: 10.1109/ACCESS.2024.3371151]
- Sharma K. Cardiomegaly disease classification using convolutional neural networks and EfficientNet. Vascular and Endovascular Review, 2025, 8(2): 167–180.
- Chuquimarca LE, Vintimilla BX, Velastin SA. A review of external quality inspection for fruit grading using CNN models. Artificial Intelligence in Agriculture, 2024, 14: 1–20. [doi: 10.1016/j.aiia.2024.10.002]
- Ramos LT, Sappa AD. A decade of you only look once (YOLO) for object detection: A review. IEEE Access, 2025, 13: 192747–192794. [doi: 10.1109/ACCESS.2025.3630988]
- Hussain M. YOLOv1 to v8: Unveiling each variant—A comprehensive review of YOLO. IEEE Access, 2024, 12: 42816–42833. [doi: 10.1109/ACCESS.2024.3378568]
- Zhang SH, Guo XX, Tan M, *et al.* DE-YOLOv13-S: Research on a biomimetic vision-based model for yield detection of Yunnan large-leaf tea trees. Biomimetics, 2025, 10(11): 724. [doi: 10.3390/biomimetics10110724]
- Wang YL, Deng YJ, Zheng YJ, *et al.* Vision Transformers for image classification: A comparative survey. Technologies, 2025, 13(1): 32. [doi: 10.3390/technologies13010032]
- Üzen, Firat H, Atila O, *et al.* Swin Transformer-based fork architecture for automated breast tumor classification. Expert Systems with Applications, 2024, 256: 125009. [doi: 10.1016/j.eswa.2024.125009]
- Lan L, Wang FX, Zheng XT, *et al.* Efficient prompt tuning of large vision-language model for fine-grained ship classification. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 5600810. [doi: 10.1109/tgrs.2024.3509721]
- He ZX, Tang J, Zhao ZT, *et al.* RWKVmatch: Vision RWKV-based multi-scale feature matching network for unsupervised deformable medical image registration. Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025. 1–5.
- 付豪, 张振利, 陈源. 改进 UNet++ 模型的脑肿瘤图像分割算法. 河北大学学报(自然科学版), 2025, 45(4): 398–407.
- Yuan YQ, Cheng Y. Medical image segmentation with

- UNet-based multi-scale context fusion. *Scientific Reports*, 2024, 14(1): 15687. [doi: [10.1038/s41598-024-66585-x](https://doi.org/10.1038/s41598-024-66585-x)]
- 16 Duan Y, Wang W, Chen Z, *et al.* Vision-RWKV: Efficient and scalable visual perception with RWKV-like architectures. *Proceedings of the 2025 International Conference on Learning Representations*. OpenReview.net, 2025.
- 17 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017. 6000–6010.
- 18 Deng J, Dong W, Socher R, *et al.* ImageNet: A large-scale hierarchical image database. *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami: IEEE, 2009. 248–255.
- 19 Ding J, Chang J, Han RR, *et al.* CDSE-UNet: Enhancing COVID-19 CT image segmentation with canny edge detection and dual-path SENet feature fusion. *International Journal of Biomedical Imaging*, 2025, 2025: Article ID 9175473. [doi: [10.1155/ijbi/9175473](https://doi.org/10.1155/ijbi/9175473)]
- 20 Chu XS, Zhou ZP, Liao WP, *et al.* Ghost-YOLO: A lightweight traffic sign detection framework via GhostNetV3. *Proceedings of the 4th Conference on Fully Actuated System Theory and Applications*. Nanjing: IEEE, 2025. 1557–1562.
- 21 Yan PF, Gong WS, Li ML, *et al.* TDF-Net: Trusted dynamic feature fusion network for breast cancer diagnosis using incomplete multimodal ultrasound. *Information Fusion*, 2024, 112: 102592. [doi: [10.1016/j.inffus.2024.102592](https://doi.org/10.1016/j.inffus.2024.102592)]
- 22 Li XY, Xiao SQ, Li Q, *et al.* The bearing multi-sensor fault diagnosis method based on a multi-branch parallel perception network and feature fusion strategy. *Reliability Engineering & System Safety*, 2025, 261: 111122. [doi: [10.1016/j.res.2025.111122](https://doi.org/10.1016/j.res.2025.111122)]
- 23 Wu X, Huang TZ, Deng LJ, *et al.* Dynamic cross feature fusion for remote sensing pansharpening. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*. Montreal: IEEE, 2021. 14667–14676.
- 24 Guo YM. Multimodal multilabel classification by CLIP. *arXiv:2406.16141*, 2024.
- 25 程鸿, 刘坚, 李雪, 等. U²-Net 在建筑物提取中的边缘精度分析. *价值工程*, 2024, 43(29): 89–91.
- 26 Chen JN, Mei JR, Li XH, *et al.* TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of Transformers. *Medical Image Analysis*, 2024, 97: 103280. [doi: [10.1016/j.media.2024.103280](https://doi.org/10.1016/j.media.2024.103280)]
- 27 Wang CS, Zhu PL, Yang S, *et al.* A semantic segmentation method for winter wheat in north China based on improved HRNet. *Agronomy*, 2024, 14(11): 2462. [doi: [10.3390/agronomy14112462](https://doi.org/10.3390/agronomy14112462)]
- (校对责编: 张重毅)