分　类　号：TP391　　单位代码：10183
研究生学号：2021562011　　密　　级：公　开

吉 林 大 学

硕士学位论文

(学术学位)

基于深度学习的甲骨文字检测方法研究

A Research on Oracle Bone Inscriptions Detection based on Deep Learning Method

作者姓名：　付新然
专　　业：　计算机科学与技术
研究方向：　计算机视觉
指导教师：　杨溪
培养单位：　人工智能学院

2024 年 5 月

基于深度学习的甲骨文字检测方法研究

A Research on Oracle Bone Inscriptions Detection based on Deep Learning Method

作者姓名：付新然

专业名称：计算机科学与技术

指导教师：杨溪 副教授

学位类别：工学硕士

答辩日期：2024 年 5 月 27 日

摘　　要

基于深度学习的甲骨文字检测方法研究

作为古文字的一种，甲骨文被认为是现代文字的直接祖先，记录了当时社会的许多重要信息，对现代考古研究具有不菲的价值。但是，由于在电子资料中不规范的文字处理方式使得对古文字的检索工作十分困难；而且甲骨文的考释工作需要大量的已知的甲骨图像数据，即便是经验丰富的专业人员也要消耗很多精力才能通过人工对比和分析完成标注的任务，相关人工智能辅助较为薄弱。当前对于古文字数字化的研究工作大多集中在对拓片图像的去噪或是对单个文字图片的分类，仅有少数针对甲骨文进行检测与定位的研究。基于深度学习的文本检测作为近年来计算机视觉的一个热门方向，因其设计相对简单和应用场景广泛备受研究者关注，将文本检测与甲骨文研究相结合，可以简化甲骨文数据收集和整理工作以及后续处理流程，加快古文字数字化的进程，为考古研究提供便利。但在场景文字图片中，文字区域和背景区域拥有鲜明的区别，而在甲骨文拓片图像中，古文字区域与划痕等噪声区域区别不明显，这样的情况导致直接将场景文字检测的方法应用于甲骨文数据集中无法得到令人满意的结果。基于上述分析，本文通过将甲骨文字形态先验特征引入到深度学习模型中，提出了两个改进办法，通过先验信息使深度学习模型在优化过程中学习到区分文字区域的内在特征，从而获得更准确的预测结果。

本文首先提出了基于交叉注意力机制的特征融合检测模型，该模型同时将拓片图像和甲骨文字先验特征作为输入，通过交叉注意力机制将文字先验特征与骨干网络提取的图像卷积特征相融合，使得模型中间层的高维特征图具有更强的表征能力，从而在解码过程中产生准确的边界，进而提升多个基础模型的检测效果。此外，本文还提出了基于多类别伪标签预测的检测模型，通过加入多类别预测的任务，使模型显式地学习可以代表文字区域与背景区域的不同特征，将该特征与高维卷积特征图融合后，可以丰富高维特征图的向量在特征空间中的局部信息，达到区分文字区域与背景区域的目的。

最后，本文在分别甲骨文数据集和其他古文字数据集中评估了所提出的方法。在甲骨文数据集的实验结果表明，本文所提出的方法在不引入过多参数的情况下

提升基线模型的检测效果。在其他古文字数据集的检测结果说明了模型在优化过程中学习到了拓片上文字的内在特征，所以迁移到其他数据集也可以取得准确的检测结果，也证明了将迁移学习的方法应用于古文字检测任务的潜力，为古文字数字化研究提供了有力支持。

关键词：

深度学习，文字检测，甲骨文

Abstract

A Research on Oracle Bone Inscriptions Detection based on Deep Learning Method

As a type of ancient script, Oracle Bone Inscriptions (OBIs) are considered direct ancestors of modern writing systems, recording many significant aspects of ancient society and holding considerable value for contemporary archaeological research. However, due to the lack of standardized handling of text in electronic data, retrieval of oracle bone inscriptions proves to be extremely challenging. Moreover, the interpretation of OBIs requires a substantial amount of known oracle bone image data. Even experienced professionals expend considerable effort to annotate tasks through manual comparison and analysis, highlighting the relatively weak support from AI assistance in this field. Currently, research on the digitization of ancient texts mainly focuses on denoising of rubbings or classification of individual character images, with only a few studies addressing detection and localization of OBIs. Text detection based on deep learning has emerged as a hot topic in computer vision in recent years, drawing significant attention from researchers due to its relatively simple design and wide range of application scenarios. Combining text detection with researches on OBIs can simplify the collection and organization of OBIs data, as well as streamline the subsequent processing workflow,

accelerating the digitization process of ancient texts and providing convenience for archaeological research. In scene text images, there is a clear distinction between text regions and background regions. However, in oracle bone script rubbings, the distinction between ancient text regions and noise such as scratches is not as evident. This situation leads to unsatisfactory results when directly applying scene text detection methods to oracle bone script datasets. Based on the analysis above, this paper proposes two improvement methods by incorporating prior knowledge of inscriptions into deep learning models. By introducing prior information, the deep learning models learn intrinsic features to distinguish text regions during the optimization process, thereby achieving more accurate prediction results.

This paper first introduces a feature fusion detection model based on cross-attention mechanism. The model simultaneously takes image and text prior feature vectors as input. By employing a cross-attention mechanism, the model integrates the text prior features with the convolutional features extracted by the backbone network enhancing the representational power of the intermediate feature maps, thereby improving the detection performance of multiple base models. This paper also proposes a detection model based on multi-class pseudo-label prediction. The model explicitly learns distinct features that represent text regions and background regions by incorporating a multi-class prediction

task, after integrating this feature with high-dimensional convolutional feature maps, it enriches the local information of vectors in the feature space, achieving the goal of distinguishing text regions from background regions.

Finally, this paper evaluates the proposed methods on both oracle bone script datasets and other ancient text datasets. The experimental results on the oracle bone script dataset demonstrate that the proposed method improves the detection performance of the baseline models without introducing excessive parameters. The detection results on another ancient text datasets show that the model has learned the intrinsic features of the inscriptions during the optimization process, therefore accurate detection results can also be achieved when transferred to another dataset and demonstrate the potential of applying transfer learning methods to the task of ancient text detection and provide robust support for the digitization of ancient texts research.

Keywords:

Deep Learning, Text Detection, Oracle Bone Inscription

目　录

第 1 章 绪　论

1.1 研究背景与意义

甲骨文被认为是现代汉语的直接先祖，是我国目前为止被发现的年代最早的成熟文字体系，与汉语的发展和演变密切相关。不仅如此，甲骨文作为中国古代政治制度、经济状态和社会生活的重要载体，记录了战争，农业和占卜等多种主题的信息。古文字常出现在考古遗址中，作为重要的文化遗存，提供了丰富的考古学证据，通过对甲骨文的解读，可以还原古代社会的历史事实，帮助我们理解历史事件，传承并发扬传统文化。古文字还是是研究语言演变的重要依据。通过研究古文字，可以了解语言的发展变化，探讨文字、词汇、语法等方面的演变规律。研究古文字有助于理解不同语言和文字系统的相互影响和融合过程，揭示语言的多样性和共性。不仅如此，古文字的碑刻中常包含大量的文学作品，通过古文字研究，可以挖掘和欣赏古代文学的独特魅力，古文字本身也具有书法和美学价值，研究古文字可以领略古代书法艺术的发展和风格演变。因此，对甲骨文的理解与分析是研究我国历史发展与文字演变的重要一环，对甲骨文的发现与考释工作也在考古学研究领域得到的关注越来越多[1]。甲骨文作为文献资料的一种，对现代考古研究具有不菲的价值，目前的学术研究越来越重视对出土文献的利用，利用甲骨文进行的语言学与科技研究也越来越多[48]。作为最早出现的文字体系，甲骨文因雕刻在兽骨或龟甲上得名[49]，因此其形状，尺寸和方向都变化各异，长相相似的文字可能对应的现代含义完全不同[50]。出于对文物保护的目的，出土的甲骨会被拓印成拓片，并以图片的形式存储，如图 1-1 所示。因此已出版的一些古文字电子资料中的文字处理方式也直接采用图像剪贴，图文混排等形式，使得对古文字的检索工作十分困难[53]；此外，已经出土并解析的甲骨文字数量占全部甲骨文字数量的一半不到，却已经数量庞大，对甲骨文的考证和解释工作需要大量的已标注的甲骨图像数据，即便是经验丰富的考古人员也要消耗很多精力才能通过人工对比和分析完成标注的任务，在该环节中的相关人工智能辅助较为薄弱[2]。因此，通过智能化的方法检测并分析甲骨拓片上的文字在语言研究和考古研究中都具有重要意义。

为了准确地识别甲骨文，首先需要确定每一个古文字在甲骨上的位置。在最

早对于甲骨文字形建模的研究中，研究人员将其完全视为由顶点和曲线组成的联通图形[54]，并通过拓扑属性或数学形态学处理提取甲骨文字的笔划形态和字形特征[56]。当前一些工作都是为了解决单个甲骨文字的分类问题[55]，或是利用传统手工特征提取的方法[30], [31]对甲骨文定位，由于甲骨本身破损的缘故，甲骨拓片上出现了许多如破损或划痕的噪声区域，一些研究者关注于对甲骨拓片去噪[63]等工作，而将文本检测的方法应用于甲骨文数据集并解决甲骨文字检测问题的工作相对较少。基于深度学习的文本检测作为近年来计算机视觉的一个热门方向，因其设计相对简单和应用场景广泛备受研究者关注。将文本检测与甲骨文研究相结合，可以简化甲骨文数据收集和整理工作以及后续处理流程，加快甲骨文数字化的进程，为考古研究提供便利。但因为刻印形式各异，刻印过程随意以及保存完整性等历史原因，拓片上的甲骨文字形式与场景文字有较大区别，直接将场景文字检测的方法应用于甲骨文中无法得到理想的效果。综上所述，如何挖掘甲骨文字数据独有的内在特征，同时利用深度学习和特征融合的方法实现甲骨文字检测是一件重要的工作，不论是在计算机科学领域还是对于考古研究领域都具有研究意义和应用价值。

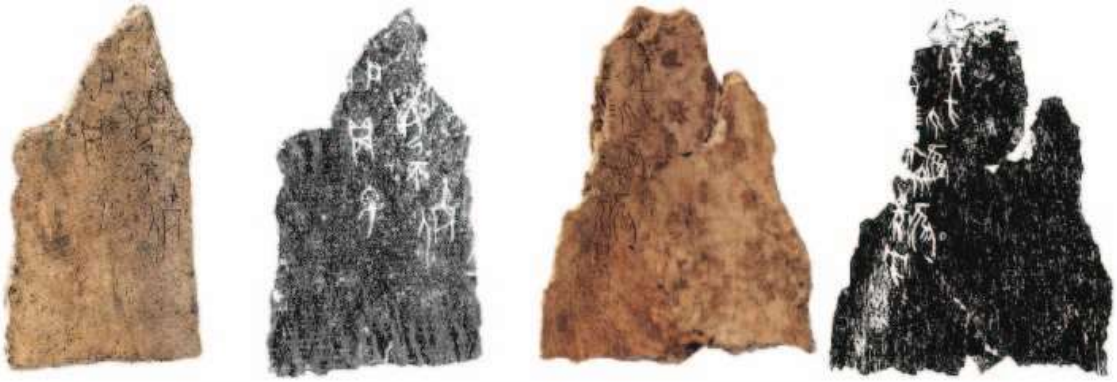


图 1-1 出土甲骨及对应拓片[67]

1.2 国内外研究现状

在甲骨文检测工作中，研究者们在各个方面做了一些开创性工作。对于古文字数据的收集和检索，王慧慧[18]等人设计了一款甲骨文标注软件，采用纯手标的方式对 7800 余张甲骨图片进行标注，但在标注完成后，还需经过古文字专业人员校对和非技术校对才得到了质量较高的甲骨文数据集，这项工作为之的研究工作的开展提供了充足的数据保证。韩等人[61]设计了一款手写甲骨文字信息系统，

包括了手写甲骨文检索和古文字与当代文字对应查找等功能。为了能够充分利用机器学习的方法来自动处理甲骨文数据，李等人[19]利用相邻的逆矩阵对每个甲骨文字进行编码，并基于图同构理论对甲骨文字进行识别，林等人[62]通过判断甲骨文字的部首与结构对单个甲骨文字进行识别。这些传统方法仅解决了单个文字的分类任务，而没有完成确定拓片中的甲骨文字的位置的任务。邢等人[20]首次对甲骨文位置检测这一任务进行分析，它们将甲骨文字检测视为目标检测任务，并采用了多个传统的目标检测模型并按照目标检测的流程设计实验，他们的实验证明，基于深度学习的检测类模型可以用于甲骨文检测，但仍然需要对模型的改进措施才能取得更好的检测效果。

本工作跟随文字检测的设计方式完成甲骨文检测任务。文字区域的常用标注方法可以划分为两种：边界区域表示和蒙板（Mask）表示。现有的甲骨文字数据集多采用字符级别的边界区域表示，即分别对每个文字的位置进行标注。边界区域表示方法用一组有序的点的坐标集合来表示包围文本区域的外边界框。蒙板表示是像素级别的标注方法，该方法将所有属于文字区域的像素点归为一类，不属于文字区域的背景像素点归为一类，生成一张与原图相同大小的蒙板作为标签。边界区域表示和蒙板表示之间可以通过算法相互转化。由于标注方法的不同，完成文本检测任务的流程也可以划分为两个方向：基于参数回归的检测方法和基于图像分割的检测方法。

1.2.1 基于回归的文字检测研究现状

基于回归的方法将文本检测任务类比于目标检测任务，图片中的文字即是待检测的目标，模型需要对候选预测框的位置信息和类别信息进行预测，并用相应的边界框将文字圈出。例如，廖等人[21]的 TextBoxes 系列模型采用基于锚框的方法解决文字检测的任务，他们应用 VGG 模块[5]作为骨干网络结构提取图像信息，通过非极大值抑制[59]后处理输出包围文字的矩形边界框的坐标信息，来预测图片中的文字，通过修改锚框的尺寸使得网络模型更适用于文本检测。虽然这种方法易于理解，但却受到文本形状差异性的限制，无法给出倾斜的文本或是任意形状的文本的准确预测。为了解决预测任意形状文本的问题，研究人员将曲线的隐式表示方法引入到回归的方法中，即预测文本边界曲线隐式表示的参数。龙等

人[64]提出“蛇形文本”的文本参数化形式，利用序列化的圆形来表示文本区域，最后通过圆形的圆心坐标，半径和方向向量的参数表示即可唯一确定图片中的文本的位置，利用神经网络端到端预测该参数表示，即可预测文本在图片中位置，由于覆盖圆的半径是可变的且行进方向是各异的，所以这种表示形式可以适用于弯曲文本检测。但这种弯曲文本参数化的方式过于复杂，为了更直观地参数化弯曲文本的边界与区域，刘等人[27]的 ABCNet 系列模型使用三次贝塞尔曲线来参数化表示图片中弯曲的文本的外包围曲线，并通过他们提出的 Bezier Align 模块回归预测生成弯曲文本保卫曲线的贝塞尔曲线控制点，然后根据控制点生成文本的包围区域。与之相近的算法思想还有朱等人[28]的 FCENet，在这个方法中，作者利用离散傅里叶变换拟合的方式参数化文本的包围区域。通过定义文本区域的起点，终点、采样方向以及采样速度将弯曲文本的边界曲线用傅里叶变换参数化表示，再利用神经网络回归傅里叶变换的参数达到弯曲文本预测的目的。叶等人利用贝塞尔曲线对弯曲文本中心线区域进行采样，进而将其编码为代表文字区域的嵌入表示，通过交叉注意力的查询机制与骨干网络融合，配合分类损失、中心预测损失和边界损失，端到端完成文字检测的任务。

1.2.2 基于分割的文字检测研究现状

回归方法的局限包括手工设计特征和非极大值抑制后处理等复杂的设计。为了避免这些情况，研究人员将重点放在利用分割方法进行文本区域检测。使用分割方法时，一般采用如下的流程：通过文本区域的边界框标注得到边界框内的像素的类别标注；利用神经网络对图像提取特征并预测与图像等大小的模板；对模板进行后处理，获得包围文字区域的边界框。周等人[23]的 EAST 模型利用经典的卷积神经网络提取特征，采用全卷积网络作为骨干网络，不同的预测头分别预测任意形状的蒙板，文本旋转角以及包围框，通过联合训练分割与回归任务实现了倾斜文本框的准确预测。廖等人[24]的 DBNET 对图片中文字区域的预处理和后处理模块进行改进，他们设计了一种可微分的函数，利用区域图和边界图来模拟像素预测值二值化的过程，并基于此设计了新的后处理过程，从而使得整个预测过程都可以被优化。王等人[16]同时利用水平卷积核检测和竖直卷积核预测文本的边界区域而非中心区域的预测图，实现了对不规则形状的文本的精确检测。

王等人[26]设计了步进式扩张模块，利用多尺度的预测结果得到准确的边界信息，克服了文字检测数据集中待检测目标距离较近和形状各异的问题，取得了良好的结果。唐等人[22]将注意力机制与文本检测任务结合，卷积神经网络依旧作为特征提取模块，随后通过弱监督损失获得表示文字的高分区域并依此稀疏化特征图；注意力层的输入也是文本区域的稀疏点表示，注意力层的输出则变成了预测框的坐标。为了实现字符级别的预测效果，Baek 等人提出了 CRAFT[25]，CRAFT 在训练初期利用带有字符级标注的合成数据训练网络，当网络模型具有一定预测能力后逐步加入真实图片。在训练过程中，CRAFT 还利用真实标注中的字符数量作为置信度来权衡损失大小，从而巧妙地实现了字符级别的预测模。刘等人[66]利用了 CRAFT 算法的标签生成策略和沙漏网络模型实现了基于分割的甲骨文字检测，为了避免边界区域粘连的问题，他们还利用步进式扩张模块对预测结果进行后处理。当前对于文字检测方面的研究大多集中在利用更合理的文本标注表示产生更精准的预测，其检测流程和标注生成方式虽然可以迁移到甲骨文字检测任务当中，但由于甲骨文字的标注相对简单，故复杂的标注形式并不适用。谢等人[35]同时将角点图的特征信息和文字图片的特征信息送入网络，通过交叉注意力机制学习文字图片的全局特征并通过对比损失学习文字间的特征，实现了对艺术字的精准预测，角点图作为一种有效的表示帮助模型准确识别形状各异的艺术字。这对于探索如何有效地利用先验知识，提高甲骨文字检测的准确性具有指导意义。

1.3 研究内容

本工作主要针对甲骨拓片上的甲骨文字检测任务进行研究，具体而言，本工作将以基于语义分割的甲骨文字检测模型为研究基础，并利用甲骨文字库先验信息和高维特征融合进行模型改进创新。本工作首先对当前应用于甲骨文字检测的任务模型进行分析，甲骨文字检测任务要求模型对于输入的拓片图像输出拓片中古文字所在的位置，常用的输出形式为文字的边界框表示。基于回归的检测方法必须绑定先验锚框生成算法和非极大值抑制后处理算法，不仅设计复杂，而且对超参数敏感的特性使得难以获得满意的结果。基于语义分割的甲骨文字检测方法则可认为是完成一个对像素进行分类的任务，可以实现更精准的预测和更直观的展示，因此本工作选择基于语义分割的方法设计甲骨文字检测模型。接下来本工

作对甲骨文数据集进行整理分析，与场景文本检测数据集相比，甲骨文数据集有其独特的特征。第一，场景文本检测数据集中的文本区域与背景区域差别较大，例如广告牌上的文字和背景的天空等颜色信息，语义信息差异巨大；而甲骨文拓片中的文字区域与拓片上其他噪声区域的差异较小，噪声区域由破损或划痕组成，与所刻文字的形态十分相似，同时划痕还会覆盖某些甲骨文字，出现遮挡的情况。第二，与文字检测数据集相比，甲骨拓片上的甲骨文字被单独标注，一张拓片上的各个文字间没有语法关联，不像场景文字具有明确的语义涵义和上下文关联。第三，已有记录的甲骨文字数量远远小于汉字的数量，同时其刻印在拓片上的形状特征较为固定，因此可以通过分别提取甲骨文单字的形态特征作为先验信息来指导模型得到更好的检测结果。根据以上分析，本工作对甲骨文字检测模型进行研究并提出了引入甲骨文单字先验信息的模型改进策略，设计了甲骨文单字先验信息的提取方法和甲骨文单字先验信息与检测骨干网络特征融合的方法。

本工作主要包括以下两个部分：

1. 基于注意力机制特征融合的甲骨文字检测模型。本方法通过有监督学习方法预先获得了甲骨文单字形态先验信息的高维向量，接下来利用交叉注意力模块将预先得到局部先验信息与骨干网络提取的全局特征信息融合，为深度学习模型提供代表文字区域形态特征的先验知识，从而提高模型学习过程的稳定性和模型的准确率。引入先验信息后的特征在解码阶段可以更准确地区分文字区域。

2. 基于伪类别标签机制的甲骨文字检测模型。针对如何利用甲骨文单字先验信息的问题，本方法设计了伪标签预测模块，该模块生成文字区域多分类别蒙板，并利用伪标签进行监督。随后该蒙板作为先验信息与骨干网络特征图融合，用于最终文字区域蒙板预测。通过伪标签预测模块，获取代表文字区域的先验信息和预测最终结果可以在同一个模型中完成，在不引入过多参数的前提下，提升了网络对文字区域的识别能力。最后，本工作还将训练好的甲骨文字检测模型应用于另一个古文字数据集完成检测任务，检测结果证明了甲骨文字检测模型的迁移学习能力，为开展古文字检测任务预训练与迁移学习研究提供了可能。

1.4 组织结构

本文的结构分为五章，主要内容如下：

第一章：首先介绍本研究的研究背景和对于人工智能领域以及考古学领域的意义，然后介绍本研究涉及的相关工作，最后介绍论文的主要研究内容和论文整体组织结构。

第二章：介绍本文方法部分涉及到的理论基础，具体包括：基于深度学习的计算机视觉，光学文字识别和注意力机制。

第三章：介绍本文提出的基于注意力机制多模态特征融合的甲骨文字检测模型。通过实验证明本文所提出模型的在古文字检测任务中的有效性，同时进行消融实验分析影响实验结果的原因。通过可视化结果展示模型的效果，对模型结构进行分析，对模型设计进行讨论。此外，还展示了在甲骨文字数据集中预训练完成的模型迁移至其他古文字拓片数据集的检测效果，证明了模型对于古文字检测的泛化性和神经网络模型在古文字领域迁移预测的能力

第四章：首先介绍本文提出的基于伪标签机制的甲骨文字检测模型，然后通过实验验证本文所介绍方法的有效性，同时进行检测结果可视化分析，对模型设计的合理性进行分析和讨论。最后，通过展示预训练甲骨文字检测模型迁移至其他古文字拓片的检测效果，证明模型对于古文字检测的有效性并且展示了神经网络模型在古文字领域迁移学习的潜力。

第五章：总结本文所研究的课题，并对未来人工智能的算法与考古学结合的研究方向进行展望和设想。

第 2 章 相关理论知识介绍

本章将介绍基于文字识别方法完成甲骨文字检测任务的相关技术和理论基础。首先介绍深度学习算法理论基础，基于优化的深度学习算法已经被广泛应用于高维隐征提取，类别预测[5], [6]和内容生成[7]等计算机视觉任务中，本文所采用的实例分割方法可以被认为是一种像素级别的类别预测任务。之后介绍光学文字识别领域对文字图像的建模方式和对图片中文字的定位算法。最后，介绍本文涉及到的注意力机制的相关理论。

2.1 基于深度学习的计算机视觉

2.1.1 卷积运算与卷积神经网络

卷积运算经常被应用在图像处理中，是提取图像隐含特征的有效方法。在计算机中存储的图片可以看作一个多通道的二维数组结构，所以对图像进行处理的卷积核也是一个二维数组。给定一张图像$\boldsymbol{I} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{n}\times\boldsymbol{3}}$ 和一个卷积核 $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{3}\times\boldsymbol{U}\times\boldsymbol{V}}$，卷积运算的定义为：

$$\boldsymbol{Y} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{3}} \boldsymbol{I_i} \bullet \boldsymbol{W_i} \in \mathbb{R}^{\boldsymbol{m'}\times\boldsymbol{n'}} \quad \cdots\cdots\cdots\cdots (\boldsymbol{2.1})$$

不失一般性地，对于输入图片的每一个通道，卷积运算的具体计算方法可以表示为：

$$\boldsymbol{y}_{ij} = \sum_{u=1}^{U}\sum_{v=1}^{V} \boldsymbol{w}_{uv}\boldsymbol{x}_{i-u+1,j-v+1} \quad \cdots\cdots\cdots\cdots (2.2)$$

其中，$\boldsymbol{w}_{uv}$为卷积核中可以被优化的参数，$\boldsymbol{x}$为输入图像某个通道具体位置的像素值，$\boldsymbol{y_{ij}}$ 为每一个通道内计算得到的对应位置像素的值，$\boldsymbol{Y}$为卷积操作得到的结果，由各个通道的计算结果按位置加和得到，也称为特征图（feature map）。但是，对于每一个空间位置，采用一组隐藏表示而不是一个隐藏表示是更为合理的，这样一组隐藏表示可以想象成一些互相堆叠的二维数组，可以蕴含更多的隐藏信息。因此，在具体实现时，卷积核往往是一个四维的数组，每个维度分别代

表了输出通道，输入通道，卷积核的长度和卷积核的高度，其计算过程示意图如图 2-1 所示。

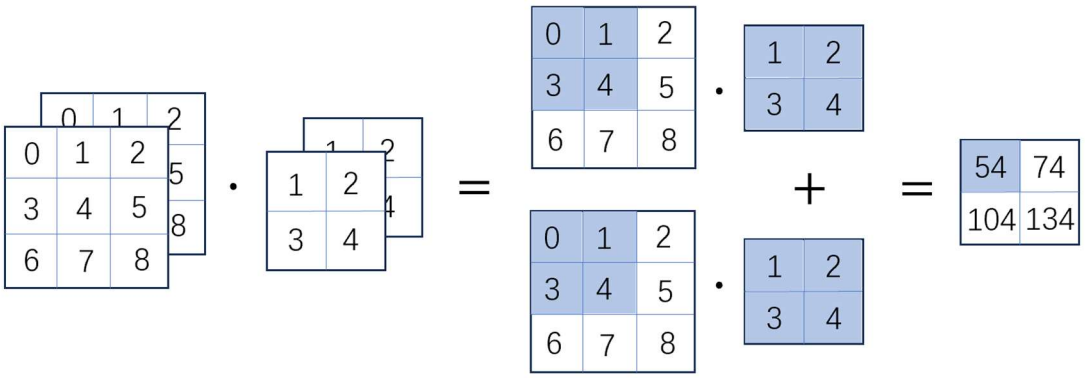


图 2-1 卷积运算示意图

卷积操作满足平移不变性和局部性两个特点，平移不变性是指待检测对象出现在图像中任何位置，卷积层应该对相同的图像区域具有相似的反应。局部性指卷积层应关注输入图像中的局部区域，而不过度在意图像中相隔较远区域的关系。正是因为这两个特性，使得卷积操作成为计算机视觉领域提取特征的首选。

卷积神经网络是以卷积操作为主的神经网络模型。用于帮助识别处理支票的数字的 LeNet 被认为是最早的卷积神经网络[9]，仅仅由少数几层卷积操作和池化操作堆叠而成，但在手写数字分类任务中取得了出色的结果。随后出现的 AlexNet[10]和 GoogleNet[11]增加了卷积神经网络的深度，在更困难的数据集和更泛化的任务中取得了更好的结果。AlexNet 证明深层神经网络卓有成效，但它没有提供一个通用的模板来指导设计新的网络。Simonyan 等人[5]提出了更为抽象但通用的卷积模块设计，他们将卷积、池化和激活函数等操作打包成块，只需要修改块内的超参数便可以很容易地在任何现代深度学习框架的代码中实现这些重复的架构。为了提升深度神经网络的性能，何等人[6]提出了残差块的设计，残差块通过给非线性的卷积层增加直连边的方式来提高信息的传播效率，使得神经网络可以学习更简单的残差函数。通过将使用不同卷积核大小的残差块串联起来，就得到了含有不同参数量的残差网络 (ResNet)。残差网络很好的解决了随着网络深度增加效果却退化的问题。

卷积神经网络一般作为骨干网络将图像编码为高维的特征图，在卷积神经网络之后连接若干全连接层便可以把特征图映射为长度固定的（与任务相关）置信度预测值。这种方法只适用于图像级别的分类和回归问题，无法解决图像像素级

别的分类问题[35]。针对该问题， Long 等人[37]提出了全卷积网络，实现了对图像进行像素级别的语义分割。全卷积网络将中间层特征图的高和宽通过反卷积[38]上采样回输入图像的尺寸，因此，输出的类别预测与输入图像在像素级别上具有一一对应关系。全卷积网络是语义分割深度学习网络中最重要的框架之一，为后续众多语义分割模型提供了借鉴。在后续的工作中，Dumoulin 等人[39]提出了转置卷积运算，转置卷积通过卷积核“广播”输入元素，从而产生大于输入的输出，既满足了上采样的要求，也使得模块拥有可以学习的参数，因此可以被优化。转置卷积运算至今依旧是各种语义分割模型的基本操作。

2.1.2 模型优化与梯度下降

神经网络模型可以认为是一个可微分的函数，模型的参数可以通过优化的方式，使得在数据集上的经验风险最小化。在深度学习中，最简单和常用的优化算法是小批量梯度下降法，该方法沿着梯度的负方向对神经网络参数进行迭代，逐步找到使得神经网络梯度最小的一组参数，其更新过程为：

$$\theta_t \leftarrow \theta_{t-1} - \omega \nabla \theta_t \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.3)$$

其中，ω为学习率，是在优化过程中的一个重要的超参数，影响收敛结果的速度与好坏；$\nabla\theta_t$为第 t 次更新时的梯度，通过链式求导法则计算得到。还有很多基于梯度下降所改进的优化算法，如动量法和 Adam 法[16]等，这些算法作为优化器，是神经网络模型训练过程的重要组成部分。

2.2 光学文字识别

光学文字识别是指利用计算机视觉模型检测图片中文字的位置与内容的一种技术。该技术在文档数字化、电子商务、身份验证、医疗领域和教育领域具有广泛的应用。与目标检测常用的矩形边界框标注方式不同，由于文本设计样式的不同，文字区域的标注经常是不规则的多边形，因此基于深度学习的光学文字识别方法主要采用语义分割网络模型，即先理解图像中每个像素是否为文字区域，再生成文字区域的包围框。在预处理阶段，需要将标签的多边形坐标表示转化成文字区域的蒙板表示。在廖等人[24]的工作中，他们分别为文字区域生成边界蒙板

和区域蒙板，作者通过对多边形边界框进行等比例缩小，得到了仅仅覆盖文字中心区域的多边形，在该区域内的像素均用 1 填充，其他背景区域的像素均用 0 填充，由此产生文本区域蒙板；将多边形区域进行等比例扩大和缩小，中间相差的部分作为文本的边界模板。在一些之前的工作中，研究者们使用一个连续的二维高斯分布来生成图像蒙板标注，位于字符框中心的像素点即为二维高斯分布的均值，具有较高的位置分，位于字符框边缘的像素点位置分较低，从而充分利用了像素点的位置信息。在后处理阶段，需要根据模型预测的蒙板得到文字区域的包围框，这一步骤可以通过平面区域求凸包算法[40]求解，该算法可以获得规则的矩形包围框，也可以获得不规则的边界。随着注意力机制与视觉结构模型的出现，一些工作通过利用 Transformer 架构可以省略对于蒙板的后处理流程而准确获得边界框的参数表示。例如，Yair 等人[33]直接在卷积神经网络后面加入 Transformer 模块，将 Transformer 的输出送入转置卷积分支和循环神经网络分支分别完成蒙板预测和文字识别的任务。为了更端到端地完成文字识别任务，叶等人提出的 DeepSolo 模型[34]利用卷积神经网络提取文本图像特征，基于图像特征生成贝塞尔曲线控制点和文本中心采样点，随后 Transformer 的解码器接收采样点特征和图像特征，输出最终的文本分类预测、边界点预测、中心线预测和文字类别预测。

2.3 注意力机制与 Transformer 模型

2.3.1 注意力机制

在自然界中，人们往往需要有选择性地关注重要信息并忽略次要信息才能以有限的经历做出更适应自身生存的决策[52]。受此启发，研究人员提出了可应用于深度学习模型的注意力机制。注意力机制的计算过程可以分为两步：首先是在所有输入信息上计算注意力分布，接下来是根据注意力分布来计算输入信息的加权平均。为了从 N 个输入向量 $[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ 中选择出和某个特定任务相关的信息，需要引入一个和任务相关的表示，称为查询向量 $\boldsymbol{q}$，并通过一个打分函数来计算每个输入向量和查询向量之间的相关性，即计算输入之间的注意力分布[65]，或者说是将查询向量和键值对向量的内在关系通过打分函数映射到输出结果中。

比较常用的两种打分函数是加性注意力打分函数[12]和缩放点积注意力打分函数，在本文中使用缩放点积计算注意力分布，这种模型在实现上可以更好地利用矩阵乘积，从而计算效率更高。其计算公式如下：

$$\boldsymbol{s(x,q)} = \frac{\boldsymbol{x^T q}}{\sqrt{\boldsymbol{D}}} \cdots\cdots (2.4)$$

其中 $\boldsymbol{D}$ 为输入向量的维度，$\boldsymbol{s(x,q)}$是计算注意力分布的打分函数的结果。获得了打分函数的计算结果后，通过计算打分函数结果的加权平均来改变不同值向量的权重，即注意力层的输出结果，其计算公式如下：

$$\boldsymbol{Attention(Q,K,V)} = Softmax\big(\boldsymbol{s(Q,K)}\big)\cdot \boldsymbol{V} \cdots\cdots (2.5)$$

如果作为查询 $\boldsymbol{Q}$ 和键值对 $(\boldsymbol{K},\boldsymbol{V})$ 的矩阵都来自于同一个输入，则称之为自注意力 (Self Attention)[13]，自注意力可以建立输入序列之间的长距离依赖关系[14]，或是反映对当前像素进行编码时应给予其他位置像素的关注程度。如果作为查询和键值对的向量来自于不同的向量，则称之为交叉注意力 (Cross Attention)，交叉注意力中的查询可以是被优化学习的嵌入表示 (Embedding)，也可以是带有特定标注信息的向量特征，通过与键值向量计算注意力分数，在不同的任务中可以起到不同的查询作用[15]，从而建模不同的任务场景。

2.3.2 多头注意力

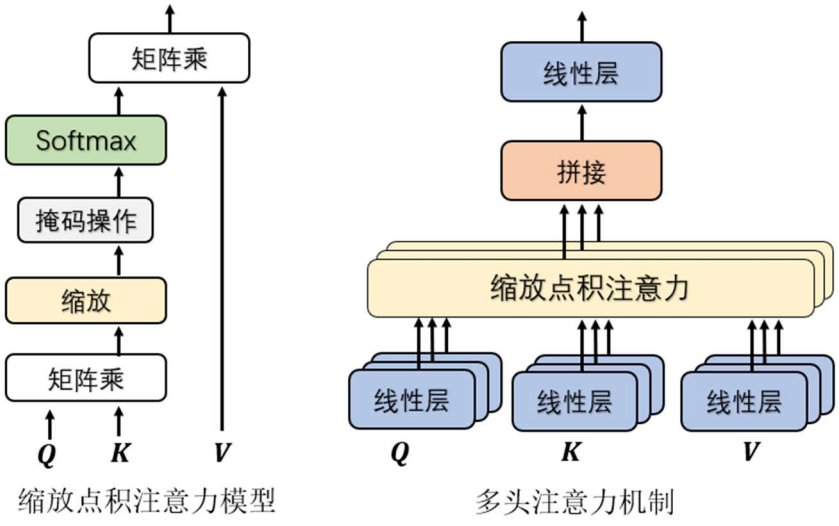


图 2-2 注意力机制计算示意图[14]

与使用单一的注意力打分函数相比，更好的办法是分别将作为查询、键和值的矩阵通过多个可以被优化学习的并列的线性层 (Linear Layer) 映射到不同的

高维的隐特征[14]，从而获得更强的表征能力，即多头注意力机制：

$$Head = Attention(\boldsymbol{Q}\cdot\boldsymbol{W^Q}, K\cdot\boldsymbol{W^K}, V\cdot\boldsymbol{W^V}) \cdots\cdots\cdots (2.6)$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Concat(Head_1,\ldots,Head_n)\cdot\boldsymbol{W} \cdots\cdots (2.7)$$

其中，不同的“头”即代表了不同的线性层，其中的参数可以将输入的查询，键和值映射到独特的特征空间中，允许模型在不同位置关注来自不同表征空间的信息，提取更丰富的特征。计算注意力机制的示意图如图 2-2 所示。

2.3.3 Transformer 模型

Transformer 模型没有使用任何卷积层或循环神经网络层，是完全基于注意力机制实现的编码器－解码器架构[14]，其详细架构见图 2-3。Transformer 的输入首先会通过嵌入层（Embedding）变成向量嵌入表示，之后与正弦余弦位置编码相

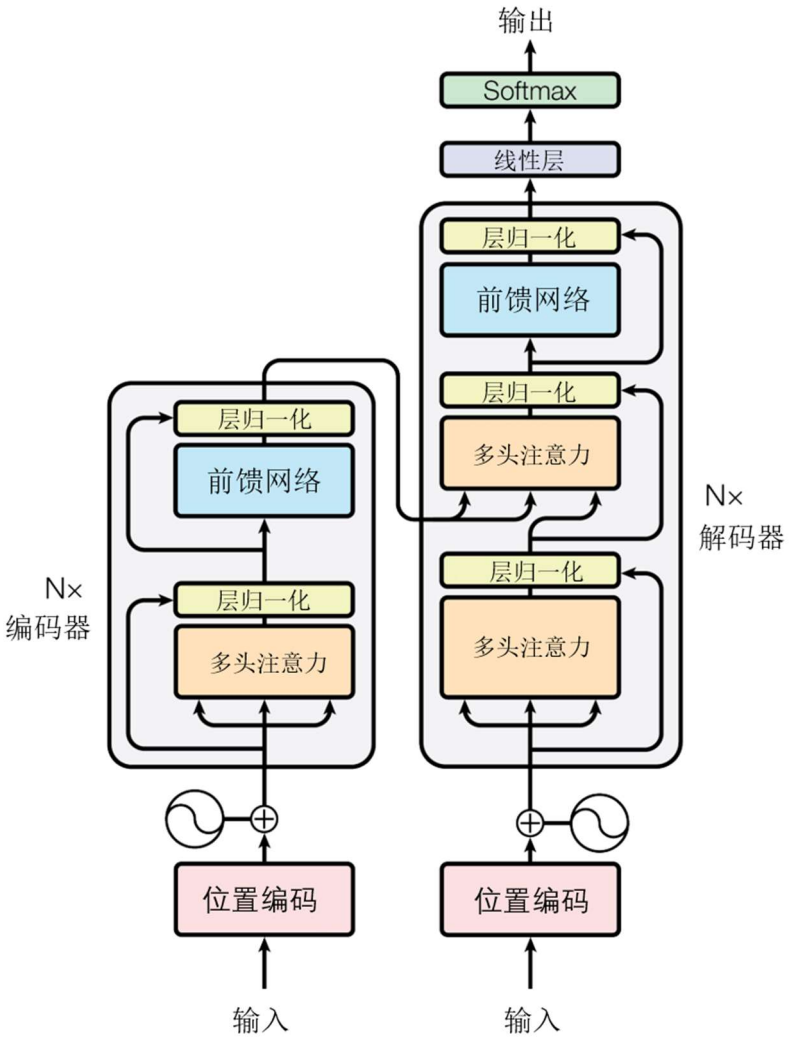


图 2-3 Transformer 结构图[14]

加[41]，位置编码操作的目的是为了使用序列绝对的或相对的顺序信息，使模型容易寻找输入序列的空间或时间关联性。Transformer 的编码器由多个相同的层堆叠而成，每个层包括两个子模块：多头自注意力模块和基于位置的前馈网络（Position Wise Feed-Forward Network），在子模块之间还包括残差连接和层规范化 (Layer Norm) 操作。Transformer 的解码器也是由多个相同的层堆叠而成，解码器每个层包含了三个子模块：解码器多头自注意力层模块、“编码器-解码器”交叉注意力模块和基于位置的前馈网络，与编码器相同，残差连接和层规范化操作也会作用于子层的输出向量。在 Transformer 模型中，多头自注意力模块和多头交叉注意力模块中的线性层参数和前馈网络的参数都是可以学习的。

2.4 本章小结

本章节详细介绍了应用于本工作的基础算法理论，包括在计算机视觉和深度学习中常用的一些卷积神经网络与梯度下降优化方法；光学文字识别领域中常用的算法设计以及注意力机制和以注意力机制为主的 Transformer 模型。

第 3 章 基于注意力特征融合的甲骨文字检测模型

3.1 基于注意力机制的特征融合检测模型

本章提出的基于注意力机制的特征融合检测模型整体结构如图 3-1 所示，该模型由像特征编码解码模块、基于交叉注意力机制的特征融合模块和动态损失权重模块三个部分组成，三个部分各自承担着关键的任务。

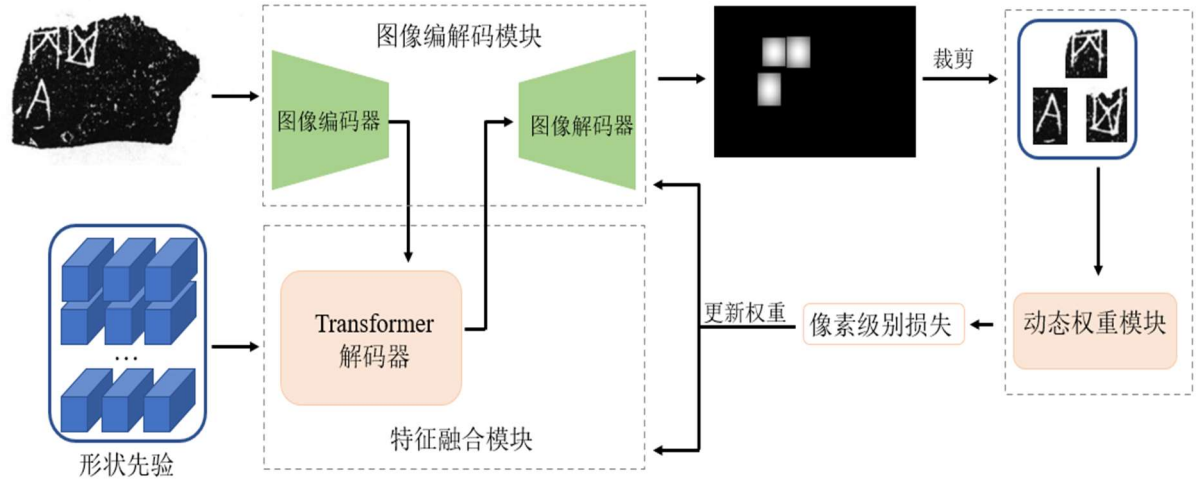


图 3-1 网络整体结构图

1. 图像特征编码解码模块：

这一部分可以被视为整个模型的“骨架”，它负责将输入的图像数据转换成中间特征图，并进一步将这些中间特征图映射到最终的预测蒙板。在这个过程中，图像经过编码得到了高维的隐空间表示，这些表示包含了图像中的重要信息。解码过程则将这些高维隐编码预测成蒙板形式，以便进行最终的预测和分析。

2. 基于交叉注意力机制的特征融合模块：

这个模块的关键在于它引入了注意力机制，将甲骨文单字的先验信息与图像特征信息进行融合。这种注意力机制可以使模型更加关注图像中与甲骨文相关的部分，从而提高预测的准确性和效率。通过交叉注意力，模型能够在不同层次上对输入进行更有针对性的处理，从而更好地捕捉到图像中与甲骨文相关的特征。

3. 动态损失权重模块：

该模块利用甲骨文单字的先验信息来平衡单次训练的损失。在训练过程

中，模型需要不断地调整权重以使损失最小化，而动态损失权重模块通过考虑甲骨文先验信息，可以使模型更加有效地进行这一优化过程。该模块只在训练时使用，在预测时并不需要。

除了这三个主要部分之外，本章还介绍形状先验的构造方法和整体损失函数的设计。形状先验的构造方法可以帮助模型更好地理解甲骨文的形态特征，从而提高预测的准确性；而整体损失函数的设计则是确保模型能够在训练过程中得到有效的指导，从而达到更好的训练效果。综合来看，这些部分共同构成了一个完整的模型框架，使其能够有效地应用于甲骨文预测的任务中。

3.1.1 甲骨文单字形态先验特征

本工作的目标是通过借助甲骨文单字的形态先验信息来提升模型对甲骨拓片上的文字区域的关注能力，从而产生更精准的预测结果。一个良好的先验信息应该是一类包含了文字基本形状特征的向量编码。在本工作中，所有的先验结构特征均来自于局部的文字区域同时拥有有限的数量，因此可以通过特征融合机制，将局部文字区域的形态学特征与模型的输入数据进行融合。这种融合方式使得模型在处理每个像素时能够考虑到文字的基本形状特征，从而更准确地判断其属于文字、划痕还是背景区域。

为了获得形状先验信息，本工作使用有监督学习的方式来训练一个由提取特征的骨干网络和类别预测分支组成的轻量级分类网络。其流程如图 3-2 所示。对于输入的单字图像 $I \in \mathbb{R}^{H\times W\times 3}$，骨干网络首先计算其结构隐特征 $\boldsymbol{z} = f_\theta(I) \in \mathbb{R}^C$，并且通过该隐特征预测其类别$\boldsymbol{c} = h_\theta(\boldsymbol{z}) \in \mathbb{R}^N$，随后根据每个单字图像标

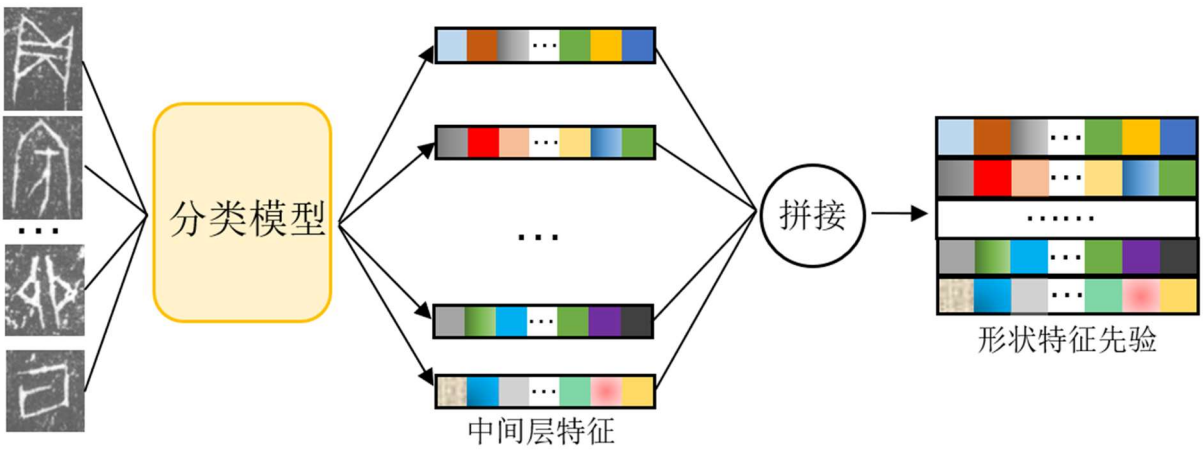


图 3-2 形状特征先验生成过程

注的类别标签优化整个分类网络的参数。训练过程结束后，该分类网络可以通过提取到的特征准确地区分不同甲骨文单字，即网络的参数可以有效地提取文字区域的特征信息，同时隐特征很好地编码了甲骨文单字区域的隐藏信息。将所有骨干网络输出的隐特征拼接在一起并保存下来便构成了一个保留甲骨文单字形态先验信息的特征字典 $\boldsymbol{Z} = \{\boldsymbol{z}^i \mid i = 1, 2, \ldots, N\} \in \mathbb{R}^{N \times C}$ 。该字典起到了两个作用，既作为注意力特征融合模块的一个输入通过交叉注意力查询机制与图像特征信息进行信息融合，也作为动态权重损失模块的先验信息与预测结果进行匹配，产生损失的权重。

3.1.2 网络架构

本章所提出的检测网络由两个主要部分组成，分别是图像特征编解码模块和基于注意力的特征融合模块。对于图像特征编解码模块，本工作选择使用广泛应用于各种语义分割任务的编码器解码器架构网络。对于输入图像 $I \in \mathbb{R}^{H \times W \times 3}$，以卷积操作为主的图像特征编码模块先将其编码成中间层特征图 $\boldsymbol{F}^e = E_\theta(I) \in \mathbb{R}^{H' \times W' \times d}$ 。 在另一个分支中，中间层特征图和预先得到的特征字典作为两种不同模态的信息被一同送入到堆叠的 Transformer 解码器模块中，该解码器由位置编码，交叉注意力层和前馈神经网络层组成。具体来说，对于输入的特征字典 $\boldsymbol{Z} \in \mathbb{R}^{N \times C}$，通过交叉注意力查询机制将其与卷积网络提取的特征相融合，具体的计算过程如下：

$$\boldsymbol{Q} = Linear_q(\boldsymbol{Z}) \quad (3.1)$$

$$\boldsymbol{K} = Linear_k(\boldsymbol{F}^e) \quad (3.2)$$

$$\boldsymbol{V} = Linear_v(F^e) \quad (3.3)$$

$$\boldsymbol{F}^{fus} = FFN\left(softmax\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{D}}\right) \cdot \boldsymbol{V}\right) \quad (3.4)$$

其中 $Linear_q, Linear_k, Linear_v$ 为可以学习的注意力头，通过加入交叉注意力模块，缩放点击注意力函数可以建模每个形状先验特征与不同位置的卷积局部特征之间的相似度权重。FFN 为多层感知器结构，将注意力打分结果映射到隐特征当中。接下来，通过使用跳跃连接将编码器得到的中间层特征图和 Transformer 解

码器得到的融合特征图连接在一起得到了解码器的输入 $\boldsymbol{F^d = F^e \oplus F^{fus}}$，其中的“ $\oplus$ ”可以是拼接操作，按位加法操作或是按位操作乘法操作等等。最后融合后的特征被输入到由多个上采样层组成的解码器中，逐步恢复为尺寸与原图大小相同的蒙板，即$\boldsymbol{M = D_\theta(F^d) \in \mathbb{R}^{H \times W}}$，蒙板中的像素值即为模型预测图片中像素属于文字区域的概率，从而实现了对文字区域的检测。

3.1.3 监督标签生成过程

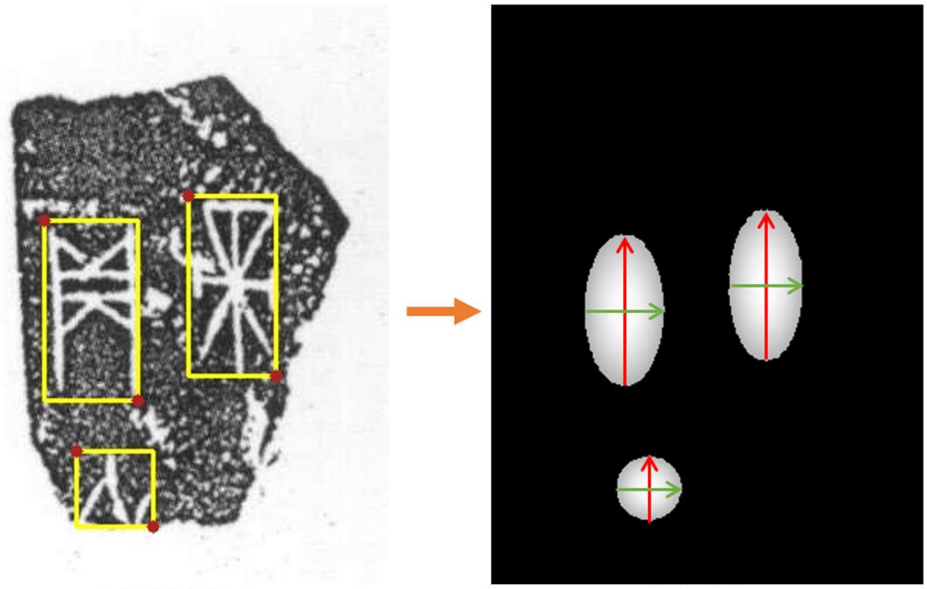


图 3-3 分割标签生成过程示意图

如前文所述，本方法采用分割的方法预先对文字区域进行检测，所以需要生成用于监督蒙板的真实标签。对于每一张训练图片，在预处理阶段仅为其生成字符级标注的区域图 (Region Map)。生成区域图的过程涉及将每张训练图片中的甲骨文字标注的包围框坐标转换为像素标签信息，其生成过程示意图如3-3所示。具体讲，对于每一张拓片，其标注形式为每个文字包围框的左上角坐标和右下角坐标表示，它们形成了一个集合，$\mathbf{B} = \{\mathbf{b_i} \in \mathbb{R}^4 | \boldsymbol{i} = \mathbf{1}, \mathbf{2} \ldots \boldsymbol{N}\}$，其中的 $\mathbf{b_i}$ 为实数域的四维向量即左上角和右下角顶点的横纵坐标，$\boldsymbol{N}$ 为一张拓片上被标注文字的个数。在生成标签时，通过采用近似的二维高斯分布概率密度函数来计算包围框内每个像素的值，其计算公式如下：

$$P(x,y) = \frac{A}{2\pi\sigma^2} \cdot e^{-\frac{(x-x_c)^2+(y-y_c)^2}{2\sigma^2}} \quad \cdots\cdots\cdots\cdots\cdots (3.5)$$

其中的 (x, y) 是标注包围框内各个像素的坐标索引，(x_c, y_c) 为标注包围框的中心坐标，可以由标注信息中的左上角和右下角坐标计算得到。A 是一个系数，用于将计算得到的每一个像素的概率密度值放大，使得中心像素的值接近于 1。根据公式可以看出，对于每一个标注包围框，将包围框的中心指定为二维高斯分布的中心后，其余框内像素的值便均可由上述公式计算得出，框内的像素值呈现由中心向两侧逐渐减小的趋势。通过使用概率密度函数的计算，高斯分布的特性可以让靠近文字中心区域的像素具有更高的像素值，而靠近文字区域边缘的像素具有相对较低的像素值。这种软标签生成策略反映出越靠近文字中心的区域包含了丰富的字形信息和文字的主体结构而边缘区域则可能包含噪声或无用的信息，需要相对较低的关注。

3.1.4 优化过程

本方法在像素级别使用带权重的二值交叉熵损失训练模型。二值交叉熵损失函数公式如下：

$$\mathcal{L}_b = \frac{1}{N}\sum_{i=1}^{N} y \cdot \log y' + (1-y)\cdot \log(1-y') \cdots\cdots\cdots (3.6)$$

其中的 y 是每一个像素的标签值，y' 是同一位置模型输出的像素预测值，N 是像素的总个数，与图像的尺寸有关。在本工作中，形态先验特征也在模型的优化过程中被使用，但其参数会被固定，不参与梯度回传。具体而言在每轮训练过程中得到预测结果后，再将具有高置信度的区域裁剪出来并送入前文提到的用于获得结构先验信息的分类网络以得到每个预测区域的结构隐特征 $\boldsymbol{z}^c = f_\theta(\boldsymbol{I}^c)$，其中 $\boldsymbol{I}^c$ 是从输入图片裁剪后得到的区域图片。接下来，可以反映一次预测结果质量的动态权重通过计算预测区域的隐特征和先验特征字典的欧氏距离的最小值获得，其计算公式如下：

$$\boldsymbol{w} = \min(\|\boldsymbol{z}^1, \boldsymbol{z}^c\|_2, \|\boldsymbol{z}^2, \boldsymbol{z}^c\|_2, \ldots, \|\boldsymbol{z}^N, \boldsymbol{z}^c\|_2) \cdots\cdots\cdots (3.7)$$

如果模型可以给出的准确的文字预测区域，则该区域内的像素所包含的结构特征已经被前文提到的分类模型学习到，因此，最小化的操作可以看作是一种动态地判别策略，即一个被正确检测的文字的隐特征应该与先验字典中的某个向量特征相似，具有足够小的欧氏距离，被赋予了较小的权重，而错误检测的区域与每个

向量都不相似，具有较大的欧式距离，则拥有较大的权重。

3.2 实验设计与结果

3.2.1 实验数据集

本工作使用由安阳师范学院提供的公开甲骨文检测数据集[42]进行所有的实验。该检测数据集包括 9154 张图片和对应的文字包围框标注。参照近几年光学文字识别领域对于数据集划分的设置[43],[44]，本方法按照 4：1：5 的比例将甲骨文字检测数据集划分为训练集，验证集和测试集，以达到相对严格的测试情况。最终训练集包括 3688 张图片，验证集包括 900 张图片，测试集包括 4566 张图片。除此之外，为了获得形状先验，本工作还使用了一个由 125 类甲骨文单字图片所组成的数据集[45]，该数据集中的图片均为从拓片图像中裁剪得到。

3.2.2 评估指标

为了综合地衡量预测框和真实框的匹配程度，本工作选择使用准确率 (P)，召回率 (R) 和 F_1 分数评估检测结果。在文字检测的评估标准中，准确率指检测正确的预测框的个数 (B_P) 占全部预测框 (B) 总数的比例，主要是判断检测效果的指标，其计算公式为：

$$P = \frac{B_P}{B} \quad (3.8)$$

召回率指被正确检测的框的个数 (G_P) 占全部真实标注框 (G) 数量的比例，主要是判断漏检的指标，其计算公式为：

$$R = \frac{G_p}{G} \quad (3.9)$$

F_1 分数通过召回率和准确率计算得到，是二者的调和平均，取值范围在 0 到 1 之间，其计算结果越高代表检测效果越好，其计算公式为：

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \quad (3.10)$$

为了判断一个预测框是否正确以及一个真实标注框是否被检测到方法是计算预测框和真实标注框的交并比。交并比的计算方法如图 3-4 所示。当一个预测框与某个真实框的交并比大于所设阈值时，该预测框为一个检测正确的预测框，该真

实框为一个被正确检测的真实框。当一个预测框与所有真实框的交并比都小于所设阈值时，即为一个误检测框；当一个真实框与所有预测框的交并比都小于所设阈值时，即为一个漏检测框。通常情况下，阈值的选择可以为 0.25，0.5 和 0.75，其中 0.5 是常用且较为公平的设置，0.75 则是更严格的比较。

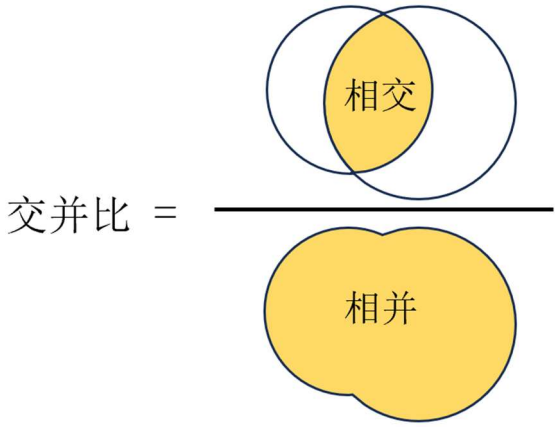


图 3-4 交并比计算示意图

3.2.3 实验设置

本工作模型代码基于 Python 实现，采用 Pytorch 库进行模型搭建。本工作选择了三个应用广泛的编码器-解码器架构网络作为基线模型，分别是 UNet[47]，U^{2}Net[46]和 HourglassNet[32]。对于基线模型，本工作通过修改了模型的通道数使得三个模型的可学习参数在相同的数量级下。在交叉注意力特征融合分支，本工作使用 3 个层堆叠成交叉注意力解码器，每个层内的多头数量设置为 4，特征向量维度设置为 64 以便于残差连接特征融合。为了获得代表甲骨文单字的先验结构特征，本工作选择使用 34 层的 ResNet 骨干网络作为编码结构特征先验的模型。本工作选择使用 Adam 作为网络训练的优化器，所有模型的参数均为随机初始化并从头训练，本工作使用指数学习率调度器在每次迭代中调整学习率，学习率在前几轮训练中会逐步增大，在后续的训练轮次中逐步变小。在训练过程中，本工作采用了颜色扰动和随机裁剪策略进行数据增强，并且将图片的分辨率均设置为 512×512。在实验结果评估时，计算 IoU 的阈值被指定为 0.5。具体超参数设置见表 3-1。

表 3-1 实验超参数设置

超参数	值
训练轮数 (epoch)	90
批量大小 (batch size)	8
初始学习率	0.001
图像尺寸	512×512
IoU 阈值	0.5

3.2.4 实验结果对比

表 3-2 在测试集上的定量比较结果

Backbone	Additional	Precision	Recall	F1-Score	Size
HgNet[32]	w/o	0.668	0.635	0.651	13.7MB
HgNet(Ours)	w/	**0.694**	**0.701**	**0.696**	11.4MB
UNet[47]	w/o	0.672	0.684	0.678	7.0MB
UNet(Ours)	w/	**0.693**	**0.712**	**0.702**	7.6MB
U^2Net[46]	w/o	0.722	0.718	0.716	4.5MB
U^2Net(Ours)	w/	**0.757**	**0.739**	**0.748**	5.5MB

表 3-2 展示了改进的方法和对应的基线模型在新划分的测试集上的评估指标对比结果以及各个模型的参数大小。改进的方法在没有引入过多参数的情况下，提升了基线模型的检测表现。具体而言，对于 HourglassNet，改进的方法显著提升了基线网络的表现：准确率提升了 2.6%，召回率提升了 6.6%。在加入了本工作提出的交叉注意力特征融合模块后，UNet 基线模型的准确率提升了 2.1%，召回率提升了 2.8%。与 U^2Net 基线模型相比，改进的方法在准确率上提升了 3.2%，

召回率提升了 2.1%。从这些结果来看，U^{2}Net 是更适合处理甲骨文字检测的分割模型，因为它拥有相对最少的模型参数，但却表现出最佳的检测效果。这表明改进的方法不仅在提高检测性能方面具有竞争力，而且能够以较小的额外参数量来实现这一目标。

图 3-5 拓片图像与标注边界框

为了更直观的展示甲骨文字检测的效果，图 3-5 中展示了带有标注框的甲骨文拓片图像并且在图 3-6 至图 3-8 中展示了基线模型和加入了交叉注意力特征融合后的模型的检测结果对比。与基线模型相比，融入了先验信息的模型能够识别

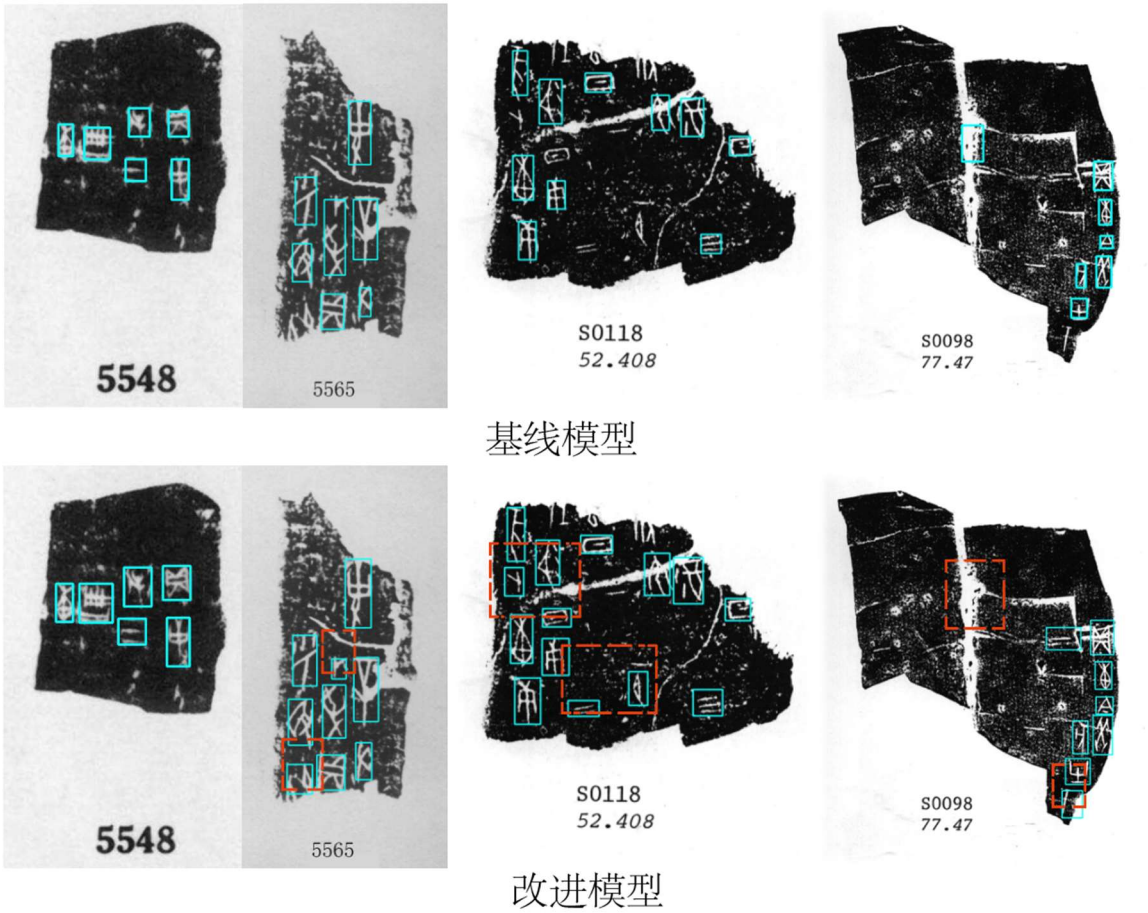


图 3-6 HourglassNet 模型检测结果对比

相对较小文字区域，而且对于位置较为临近的文字也可以精准地区分它们的边界。从图 3-6 可以看出，在第二列的图片中，基线模型将尺寸较小的文字与周围较大的文字预测成了同一给文字，而改进模型可以良好地区分它们的边界，同时残缺的文字也被检测到；在第三列和第四列的图片中，基线模型漏检或错检了一部分文字，但加入了先验知识模型避免了这种情况。在图 3-7 的第二列中，基线模型漏检了小尺寸的古文字，但改进模型将其检测到；在第三列的图片中，基线模型对于裂痕穿过的文字产生了部分的预测，而改进模型很好地将右上部分区域的文字检测出来，在第四列的图片中，基线模型将位置临近的不同文字一并框出，改进模型则对于边界有了准确地预测，避免了这种情况。在图 3-8 中，改进模型同样可以对尺寸较小的文字产生精确的预测。但是从 3-6 和 3-8 的第四列预测结果可以看出，改进模型将部分划痕预测为了文字，产生了错误的结果。

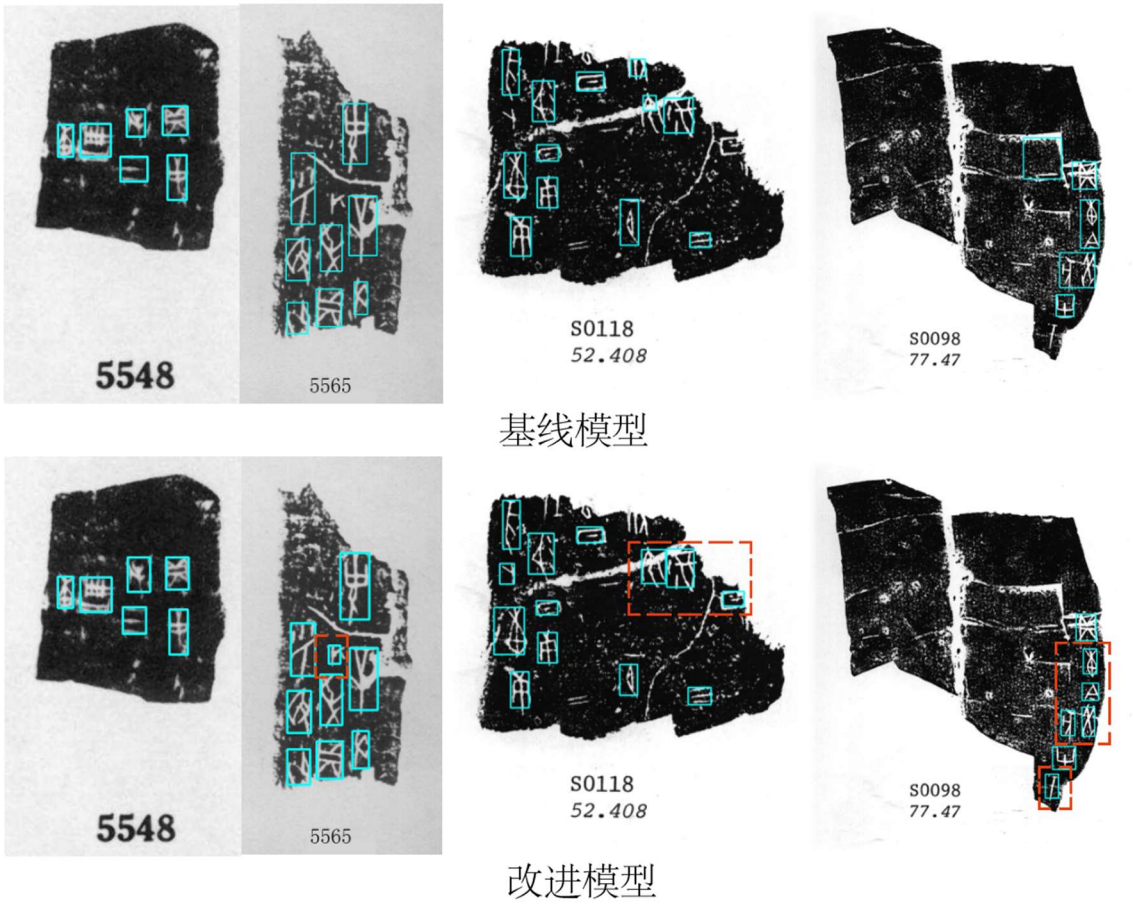


图 3-5 UNet 模型检测结果对比

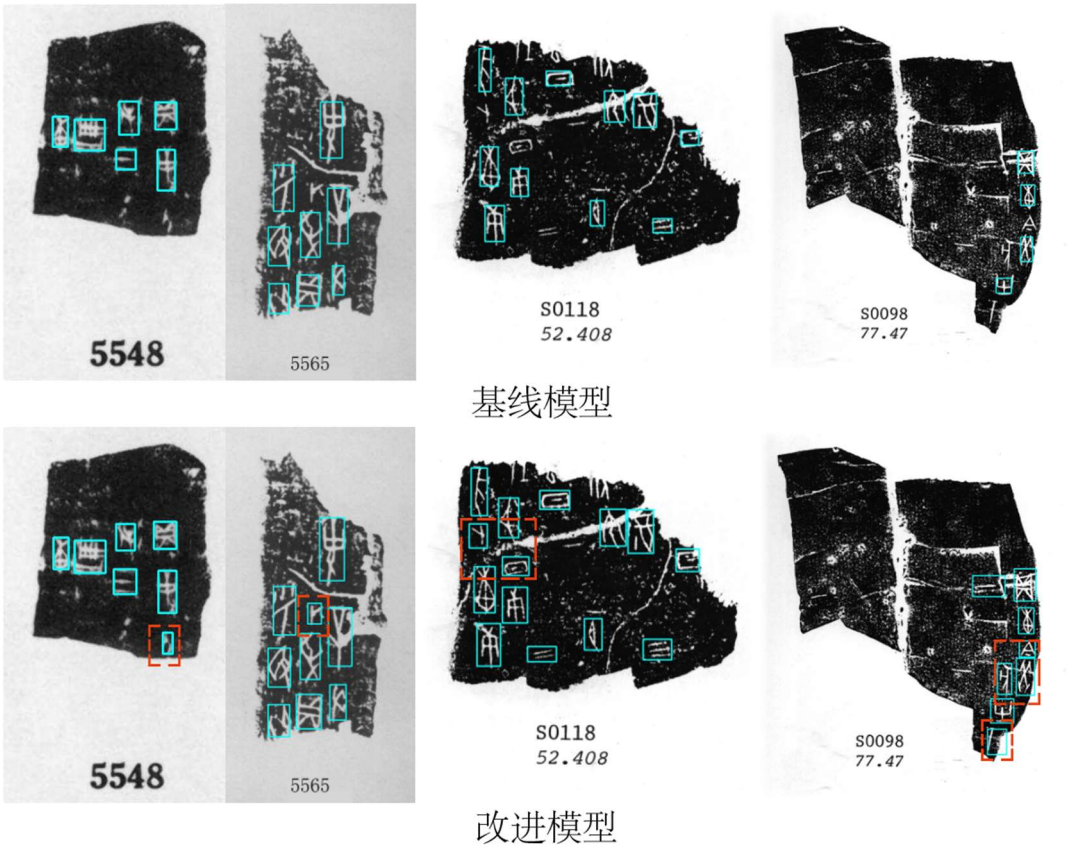


图 3-6 U²Net 模型检测结果对比

3.2.5 消融实验

本工作在三个方面进行了消融实验。为了证明加入的交叉注意力机制的有效性，而非引入额外参数的影响，本工作选择对比更大参数量的骨干网络与加入交叉注意力的网络模型之间的效果。如表 3-3 所示，提高网络中可学习的参数量可以提高网络的能力，但加入交叉注意力的网络模型在拥有较少的参数量的情况下，可以取得更好的结果。

表 3-3 模型参数消融实验结果对比

	Precision	Recall	Size
U²Net(Large)	0.749	0.736	75.5MB
U²Net(Ours)	**0.757**	**0.739**	5.5MB

为了探究文字先验对检测效果的影响，本工作更换了不同的嵌入表示作为注意力机制解码器输入，如表 3-4 所示，把特征融合的交叉注意力模块换成自注意力模块后，效果只有微弱的提升，说明注意力模块的权重计算可以对隐特征进行筛选；当采用随机初始化输入时，模型的预测效果有所下降，不如采用文字先验作为输入，产生这种现象的原因是骨干网络模型和交叉注意力模块的参数均为不含有任何预训练信息的随机初始化参数，当输入无法提供有效的先验信息时，会使模型的优化过程更加困难。

表 3-4 输入形式消融实验结果对比

	Precision	Recall	F1-Score
Random Init	0.719	0.705	0.712
Self-Attention	0.727	0.719	0.723
Shape Prior	**0.757**	**0.739**	**0.748**

本工作还对送入解码器前的特征图融合方式进行了消融实验，本工作分别采用按位加，按位乘和直接拼接的方式连接编码器的输出和交叉注意力层的输出。如表 3-5 所示，按位加方式取得了明显的效果提升，而且，与拼接方式相比，按位加的融合方式可以减少模型的参数量。

表 3-5 融合方式消融实验结果对比

	Precision	Recall	F1-Score
Add	**0.757**	**0.739**	**0.748**
Concatenate	0.729	0.721	0.725
Multiply	0.725	0.718	0.721

3.2.6 迁移预测

本章提出了一个基于交叉注意力特征融合的甲骨文字检测模型，并在甲骨文拓片数据集上进行训练与评估，实验证明改进的模型可以完成甲骨文字检测的任务。此外，将在甲骨文拓片数据集上训练好的模型迁移到金文数据集检测金文拓片中的古文字时，依然可以取得不错的效果，其预测结果见图 3-9，在每张图中，第一列展示了在相对规范工整的拓印文字情况下的预测结果，第二列展示了存在倾斜旋转的拓印文字情况下的预测结果，第三列展示了文字密集情况下的预测结果。从图中的预测结果可以看出，当处理方向平直的文字时，模型能够准确地检测金文拓片中的文字，并输出正确的预测框。这种情况下，文字的排列方式通常比较规整，文字之间的间距和对齐比较明显，有利于模型进行准确的定位和检测；当文字方向有所倾斜，虽然模型依然可以对文字区域输出预测框，但出现了无法

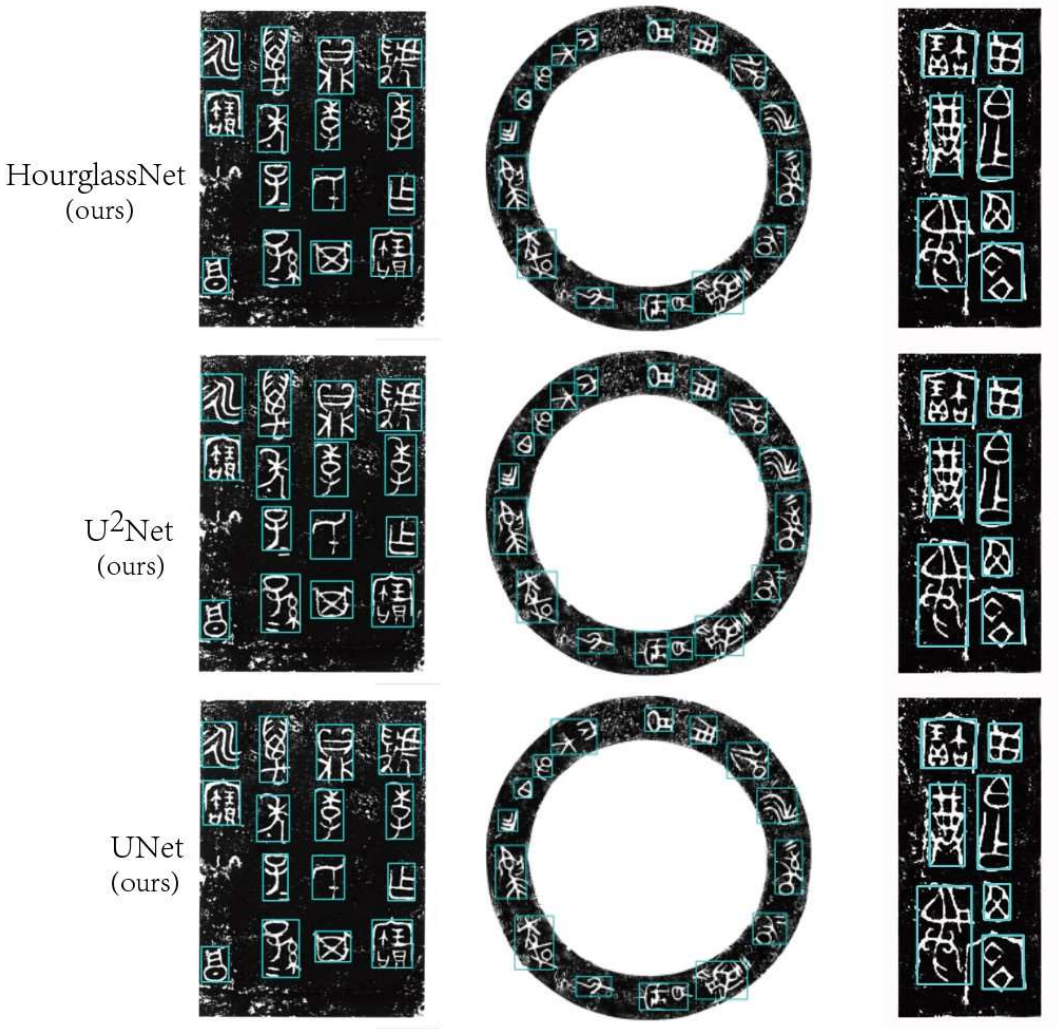


图 3-7 金文数据集预测结果

精确区分不同文字的情况，预测效果有所下降；在文字布局紧密的情况下，模型对于文字区域依然有鲁棒的预测结果。金文作为刻在青铜器物上的文字，已经被精心设计修正，所以拓印后的结果与随意刻印在甲骨上的甲骨文相比，文字部分虽然已经有很大的变化，但也更加工整清晰，同时因其以拓片为存储形式的方式与甲骨存储形式相同，而训练过的模型已经学习到了拓片中代表文字区域的笔划结构信息，能够区分拓片中的文字笔划与背景区域，故将模型迁移至金文检测也可以产生合理的预测结果。

3.3 本章小结

本章节首先详细介绍了本工作提出的基于交叉注意力机制的甲骨文字检测模型的整体结构，接下来介绍了完成甲骨文字检测的实验流程和评估方法，最后通过实验展示本文所提出的模型在甲骨文字检测效果上的优势。本章的工作展示了基于先验知识的甲骨文字检测模型的潜力，利用文字先验知识指导模型并提升模型检测能力具有重要研究价值。此外，将本工作提出的模型迁移至其他古文字数据集依旧可以取得良好的检测结果，证明了模型迁移预测的能力和在古文字检测任务中迁移学习的可能。

第 4 章 基于伪类别标签的甲骨文字检测模型

4.1 问题描述

本章的工作旨在通过对多类别标签的监督，显式地利用不同文字的多类别标签的先验信息为模型提供区分文字区域和背景区域的能力。多类别标签是指将相同文字看作同一个类别，为不同的文字赋予不同的类别标签，由于多类别标签的获取方式不是由专家手工标注得到，而是通过预训练的分类网络并预测得到，故称为伪类别标签，同时该多类别标签并不需要标识文字的具体释义，只作为区分不同文字的标志。通过对多类别伪标签的监督，使其包含了可以区分文字区域和背景区域的局部信号；最后通过特征融合，将该局部信号传递到特征图中用于文字区域的预测。

4.2 模型总览

本章提出的基于伪类别标签的检测模型整体结构如图 4-1 所示，该模型包括三个部分：图像特征提取模块、伪类别标签预测模块和区域预测模块。首先，图像特征提取模块作为骨干网络，负责提取输入图像的特征图，这些特征图包含了图像的高级语义信息和低级视觉特征。其次，伪类别标签预测模块通过有监督的

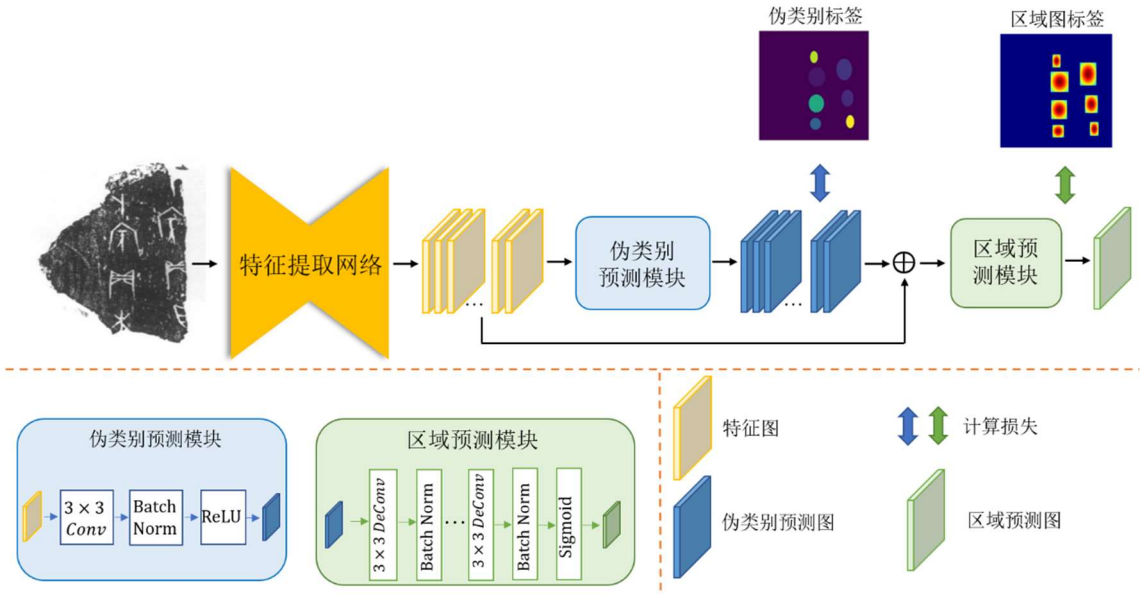


图 4-1 网络整体结构图

方式对每个文字区域的类别进行预测，并将这些预测结果作为先验信息，与骨干网络提取的特征图相加。这种方式能够有效地引导模型关注文字区域。区域预测模块接收融合了多类别先验信息的特征图，输出最终的预测结果。

除了模型结构外，本章还介绍了监督信息的生成方式，即如何生成用于训练的蒙板。通过将标注数据转换为伪类别标签，结合像素级别的信息，还可以生成准确且高质量的监督信号，为模型的训练提供有效的指导。此外，还设计了整体的损失函数，综合考虑了定位精度、类别准确性和区域完整性等方面的因素，以促进模型的全面优化和训练收敛。

4.2.1 标签生成过程

本工作分别生成代表了文字区域的区域图标签和代表不同文字信息的伪类别图标签。它们的形式均是二维数组。具体生成过程如图 4-2 所示。对于区域图标签的生成，本工作采用与第三章相同的方式，借助高斯分布概率密度函数对每一个包围框分别生成其内部像素的标注值。如 4-2 的热力图所示，在每一个包围框内部，越接近包围框中心的像素颜色越鲜艳，代表拥有较高的像素值即较高的

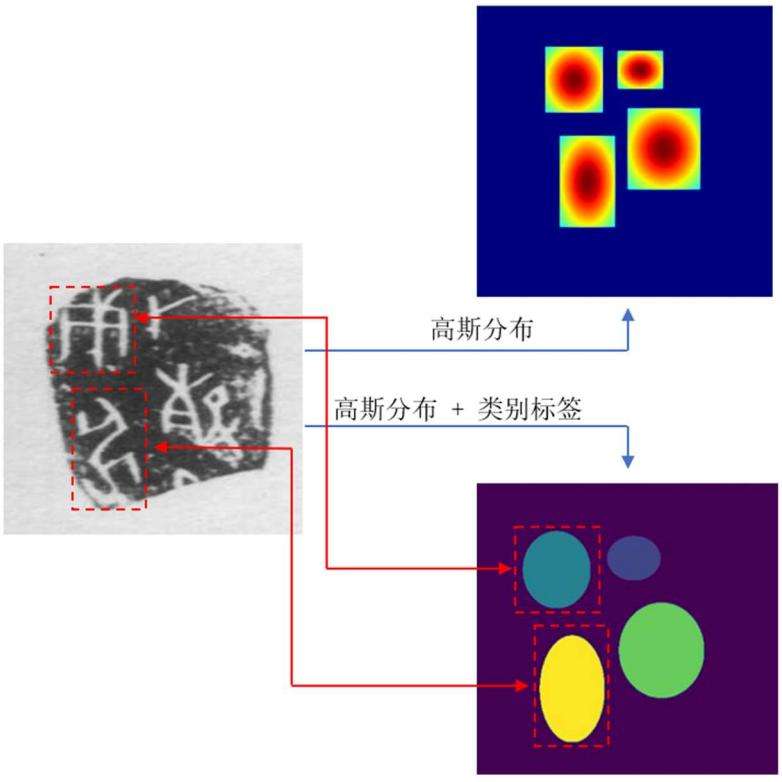


图 4-2 标签生成过程示意图

概率值；越接近边缘的像素则越趋于冷色，代表像素值较低。本工作还为每张拓片图片生成伪类别标签图。首先将标注包围框内的文字图像裁剪出来，并送入第三章中提到的分类 ResNet 中得到每个文字的具体类别索引，接下来将区域图中像素值大于阈值的像素替换为该文字的类别索引值。伪类别预测图的标注形式类似于实例分割的标注，如 4-2 所示，不同的文字区域拥有不同的颜色，即代表了不同的类别。

4.2.2 网络结构

本工作选择常用的语义分割网络作为特征提取骨干网络 $E(\cdot)$，对于输入图片 $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$，模型将其编码为与图像等大的高维特征图$\boldsymbol{F} = E(\boldsymbol{I}) \in \mathbb{R}^{H\times W\times d}$，获得特征图后，首先将特征图送入到伪类别标签预测模块中，该模块由堆叠的卷积层，批量归一化层和非线性激活函数组成，伪类别预测模块根据输入的特征图预测类别蒙板$\boldsymbol{M}_{pc} = f_p(\boldsymbol{F}) \in \mathbb{R}^{H\times W\times d}$，其中每个位置的向量代表了该像素属于每个字的预测概率。对类别模板的监督可以让模型拥有理解文字特定结构的能力，同时，受到 Liu 等人工作的启发[51]，本工作还将类别蒙板与特征提取网络输出的特征图按位置相加得到融合特征图 $\boldsymbol{F}' = \boldsymbol{F} \oplus \boldsymbol{M}_{pc}$ ，显式地将类别特征信息融入到原始特征图中，提供代表文字区域的独特信息。接下来，区域预测模块根据融合特征图输出最终的区域预测图 $\boldsymbol{M}_r = f_r(\boldsymbol{F}') \in \mathbb{R}^{H\times W}$，其中每个位置的值代表了其属于文字区域的概率。在训练阶段，模型会输出伪类别预测图和区域预测图，他们同时被预先生成的标签所监督；在预测阶段，仅输出区域预测图用于生成最后的包围框表示，作为最终的预测结果。

4.2.3 损失函数设计

在训练过程中，区域预测图和伪类别预测图同时被监督，并采用参数 λ_1和 λ_2 平衡两个监督信息的权重，具体计算公式如下：

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_r + \lambda_2 \cdot \mathcal{L}_{pc} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4.1)$$

二元交叉熵损失 $\mathcal{L}_r$ 作为监督区域图的损失函数，因为其本质上是一个像素级别的二分类的问题；交叉熵损失 $\mathcal{L}_{pc}$ 作为监督伪类别图的损失函数。

4.3 实验设计与结果

4.3.1 实验设置

本工作模型代码基于 Python 实现，采用 Pytorch 库进行模型搭建。本工作选择了三个在分割领域应用广泛模型作为骨干网络，分别是 UNet[47], U^{2}Net[46]和 HourglassNet[32]。对于选取的骨干网络，通过修改了模型的通道数使得三个模型的可学习参数在相同的数量级下。Adam 作为本工作的优化算法，所有模型参数均为随机初始化从头训练，并使用指数学习率调度器在每次迭代中调整学习率。在训练过程中，输入图片通过颜色扰动进行数据增强，以获得更多样化的数据，并且每张图片的分辨率均设置为 512×512。实验数据集和评估指标均均采用与第三章相同的设置。其他具体超参数设置见表 4-1。

表 4-1 实验超参数设置

超参数	值
训练轮数 (epoch)	90
批量大小 (batch size)	8
初始学习率	0.003
图像尺寸	512×512
平衡系数 λ_1	1
平衡系数 λ_2	0.1

4.3.2 实验结果对比

表 4-2 分别展示了在验证集和测试集上骨干网络与改进方法的评估结果。添加了类别预测分支之后，HourglassNet 模型的准确率值和召回率值在测试集上均取得较大的提升，准确率从 66.8%提升到了 70.9%，召回率从 61.5%提升到了 68.4%；UNet 模型的准确率提升较为明显，有 4.3%的提升，召回率有 1.4%的提

升；U²Net 模型的准确率也有一定的提升，达到了 1.9%，召回率仅有 0.7%的提升，提升较小。通过实验结果可以看出，新引入类别预测分支以及多类别预测子任务后，多类别标签提供了文字区域独特的信号，使模型能够对文字区域拥有更高的关注，从而提升了检测效果。

表 4-2 甲骨文验证集和测试集对比结果

	Precision		Recall		F1-Score	
	验证集	测试集	验证集	测试集	验证集	测试集
HgNet	71.4	66.8	67.9	61.5	69.6	64.1
HNet (ours)	71.3	**70.9**	**68.3**	**68.4**	**69.9**	**69.6**
UNet	69.0	67.2	69.5	68.4	69.2	67.8
UNet (ours)	**71.9**	**71.5**	**70.3**	**69.8**	**71.1**	**70.6**
U²Net	73.1	72.2	**72.5**	71.8	72.8	71.6
U²Net (ours)	**73.8**	**74.1**	72.1	**72.5**	**72.9**	**73.4**

本工作还比较了改进的方法与其它回归模型和文字检测模型在验证集和测试集上的检测效果，以及不同网络模型的参数大小，其结果见表 4-3。与文字检测模型 DBNet 相比，改进的模型能够取得更好的结果。与随机初始化参数从头训练的 RetinaNet 模型[57]和 YOLO 模型[59]相比，改进的方法在准确率和召回率指标上都取得了更好的结果。与使用了在 ImageNet 上预训练[58]过的权重初始化的 YOLO 模型 (见表中 YOLO*)相比，改进方法取得了与之相近的结果，然而在网络结构上，改进模型不仅显著降低了参数量，而且避免了回归模型中必不可少的先验锚框生成和非极大值抑制操作，实现了对模型的简化。

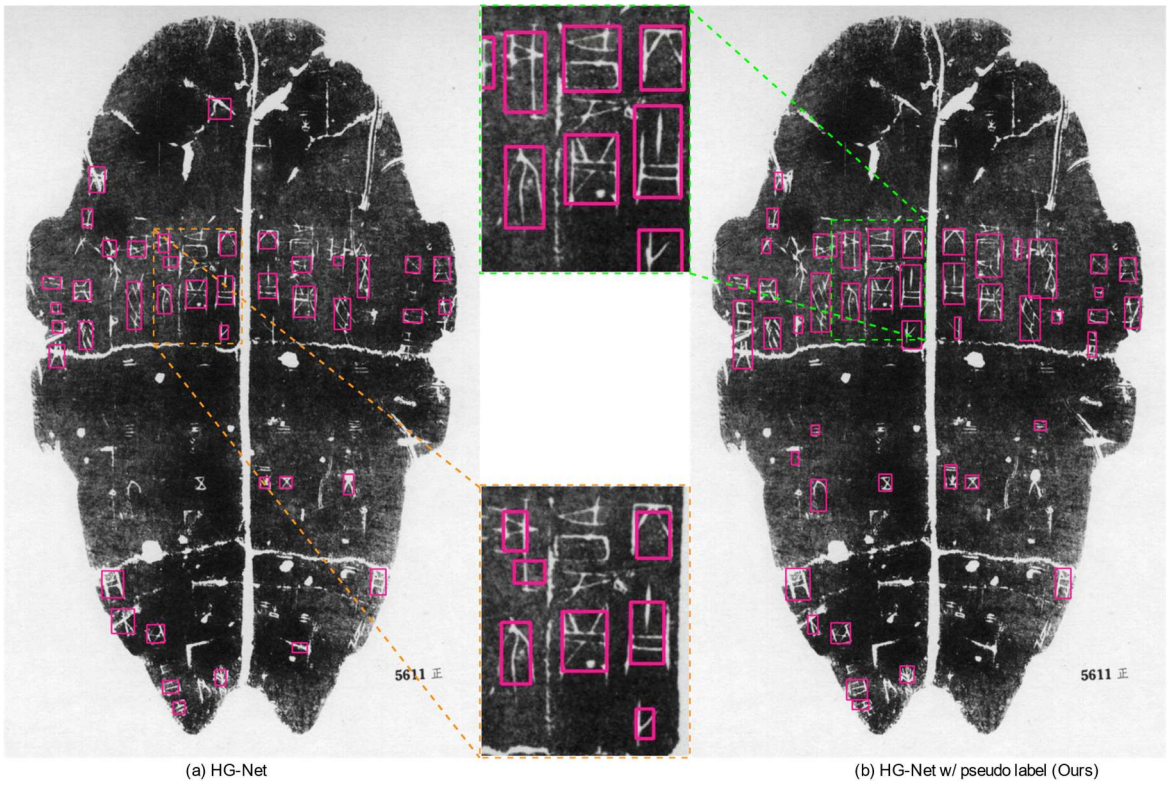
表 4-3 与回归模型的定量比较结果

	Precision		Recall		F1-Score		Size
	验证集	测试集	验证集	测试集	验证集	测试集	
RetinaNet	57.3	56.1	56.4	55.4	56.9	55.4	60.8M
YOLO-v3	44.1	46.3	45.6	46.9	44.8	46.6	125.2M
YOLO-v3*	73.2	73.7	**72.8**	**74.1**	72.7	**73.9**	
DBNet	68.8	69.4	64.5	65.2	66.5	67.5	104.1M
HgNet (ours)	71.3	70.9	68.6	68.4	69.9	69.6	6.6M
UNet (ours)	71.9	71.5	70.3	69.8	71.1	70.6	6.9M
U^2Net (ours)	**73.8**	**74.1**	72.1	72.5	**72.9**	73.4	5.9M

4.3.3 预测结果可视化对比

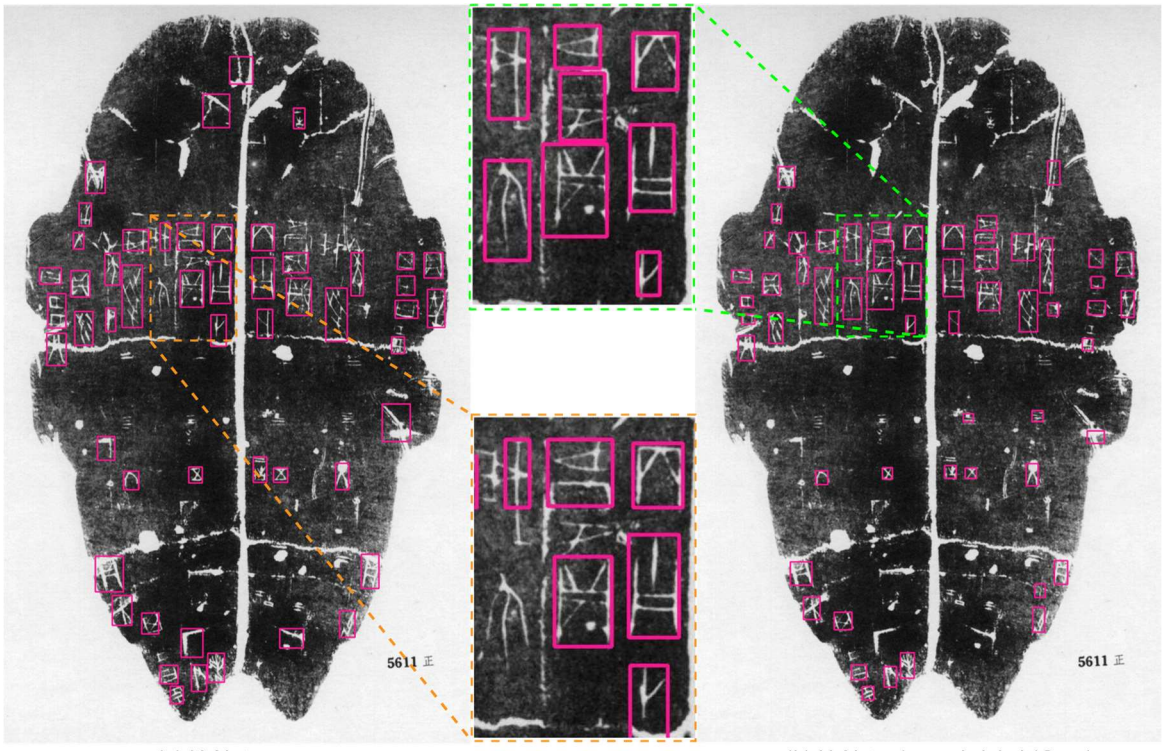
图 4-3 至图 4-6 分别展示了基线模型与本工作提出的改进方法在相同的尺寸较大的图片中预测结果的对比，每张图片的左侧为基线模型的预测结果，右侧为改进模型的预测结果，本工作还放大了某些区域的预测结果。从图 4-3 的放大图中可以看出，右侧改进模型的预测结果图片中检测到了笔划相对较细的文字，这表明改进模型在识别细微文字方面的能力得到了显著提升，从而产生了更精准的预测结果。而从图片上方区域的观察可以发现，改进模型减少了对裂痕区域的预测，这进一步减少了假阳性的预测结果。这意味着改进后的模型由于获得了潜在的先验知识在辨别裂痕和其他无关区域方面表现更加准确。图 4-4 展示了 HourglassNet 模型的对比结果，从局部放大的图片可以看出，与基线模型相比，改进模型在文字较为密集的区域展现出了更为精准的边界预测能力。这说明改进模型能够更准确地识别文字之间的边界，从而完整地框住每个文字；改进模型在

处理笔划较细的文字时也表现出了显著的进步，不再漏掉这些细微的文字信息，而是能够产生相应的检测结果。图 4-5 展示了 U^2Net 模型的对比结果，从放大图



(a) HG-Net (b) HG-Net w/ pseudo label (Ours)

图 4-4 U-Net 网络预测结果对比



(a) U-Net (b) U-Net w/ pseudo label (Ours)

图 4-3 HourglassNet 网络预测结果对比

可以看出，改进模型对小尺寸的文字区域拥有更好的预测效果，对比其他未放大部分的预测结果也可以发现，U^{2}Net 在文字密集区域也拥有精准的预测结果，可以准确的区分文字的边界，对于内容复杂的拓片图像更加鲁棒。

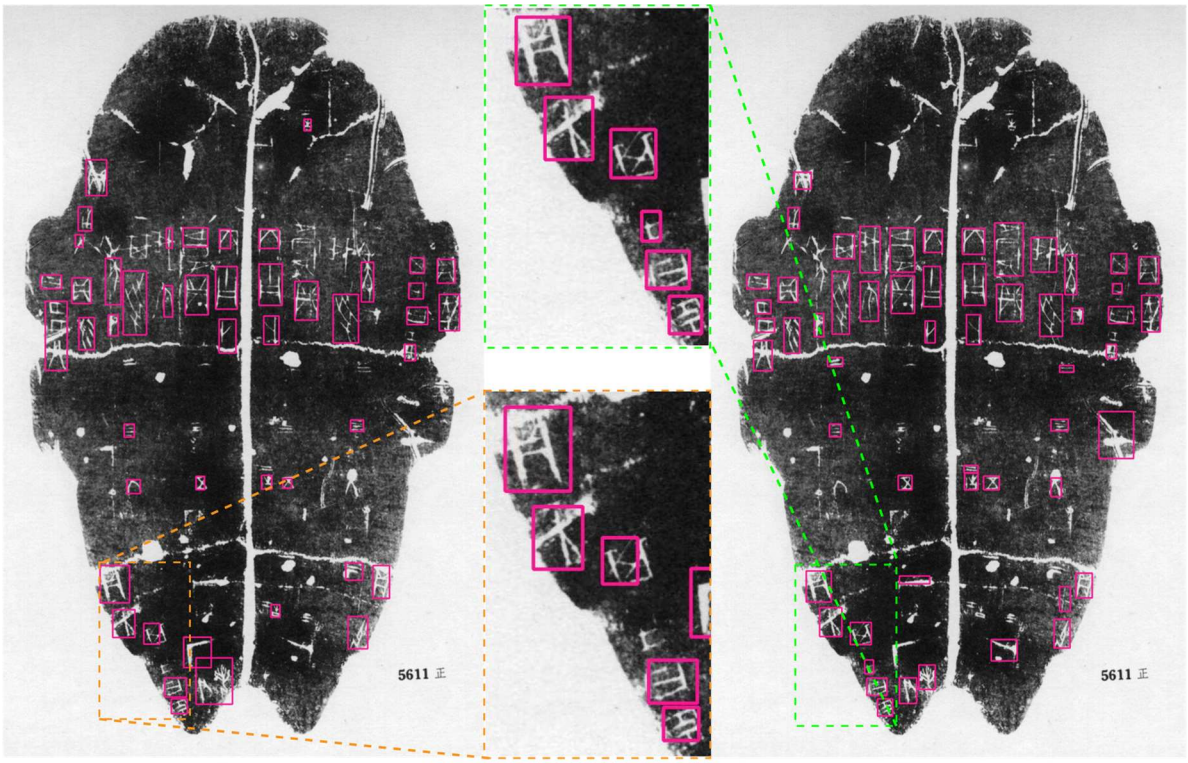
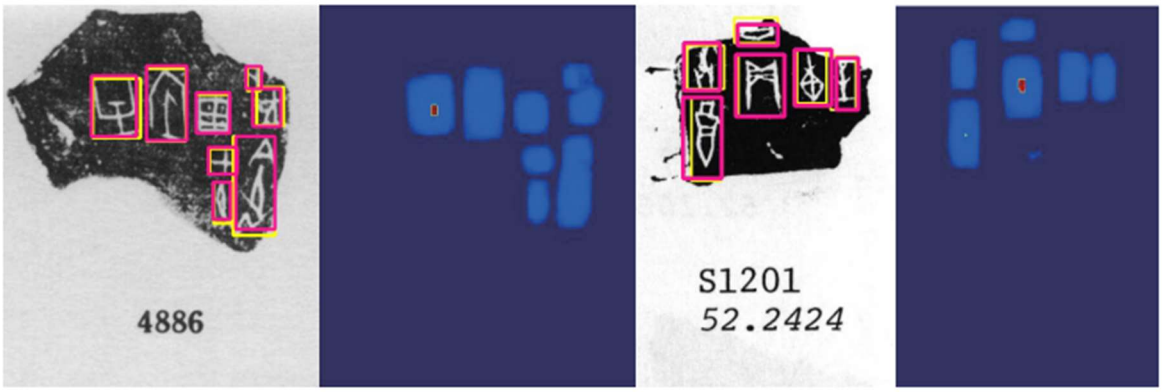


图 4-5 U^{2}Net 网络预测结果对比

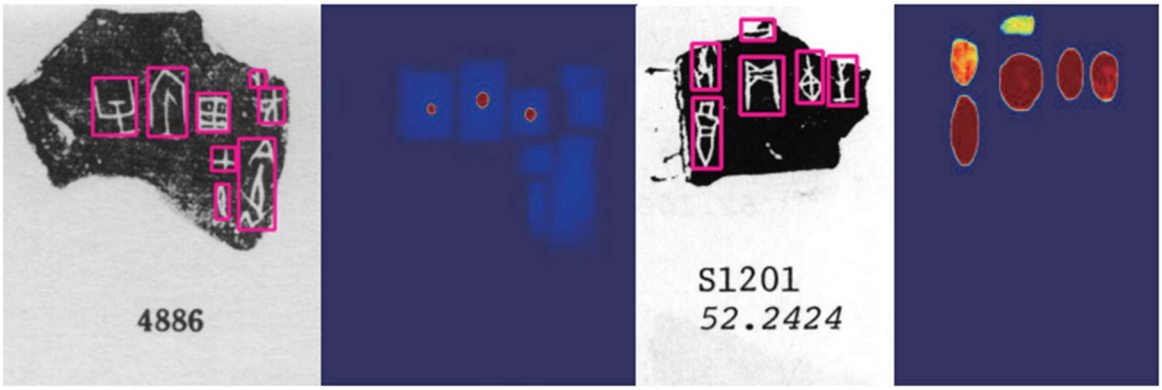
4.3.4 模块可视化分析

图 4-6 中展示了一些文字分布较为密集的拓片图像的预测结果并且可视化了对应的伪类别预测分支的输出以探究伪类别标签的作用。为方便比较，拓片的标注信息也在第一行的预测结果图片中同时展示出来，每张图片中粉红色的包围框为预测结果，第一行图片中黄色的包围框为标注信息。在第一列和第三列所展示的预测结果中，每个预测包围框的位置与每个标注框的位置都很好地重叠在一起，说明改进模型可以在文字分布较为密集的情况下产生准确的预测结果。在第二列和第四列的可视化结果中，每个改进模型的伪类别分支的输出对于文字区域所计算产生的值和背景区域计算所产生的值不仅有显著的区别，同时在边界处还具有明显的区分，而且在其他划痕的区域的值也与文字区域有所不同，说明该模块可以对于文字区域产生区别于其他区域的检测特征。在模型前向计算的过程中，伪

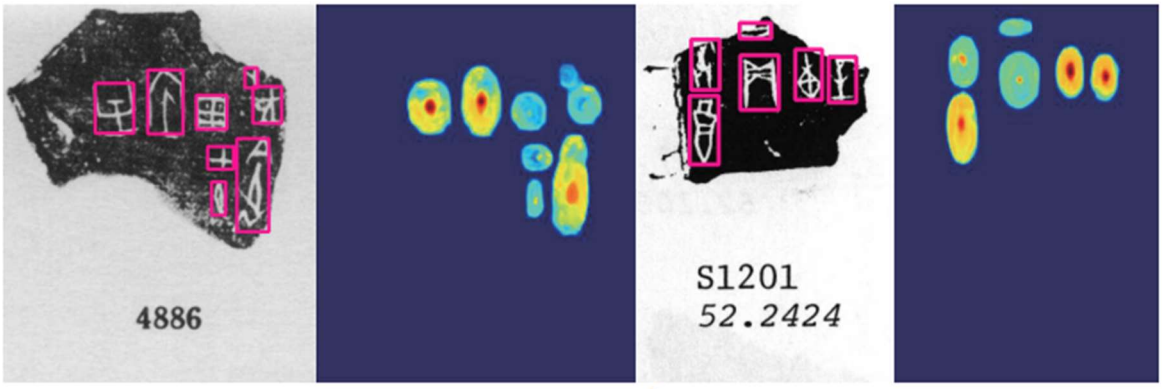
类别预测图作为一种先验信息与初始特征图相融合，提供了文字区域与其他区域不同的信号，从而使得接下来的区域预测模块能够更轻易的区分不同的位置，产生更精确的结果。



(a) HourglassNet



(b) UNet



(c) U^{2}Net

图 4-6 伪类别预测图可视化

4.3.5 迁移预测

本章提出了一个多类别伪标签特征融合的甲骨文字检测模型，同第三章，我们也在金文数据集中观察训练好的模型的检测效果，其结果见图 4-7，每一列的布局与第三章相同。从第一列的预测结果可以看出，在较为简单的情境下，模型能够准确地识别平整文字的整体位置和文字之间的边界；在第二列的预测结果中，当文字方向有所倾斜时，虽然模型依然可以对水平文字输出准确地预测，但对于倾斜的文字预测效果有所下降；在第三列的预测结果中，模型可以对密集的文字区域产生准确的预测结果。

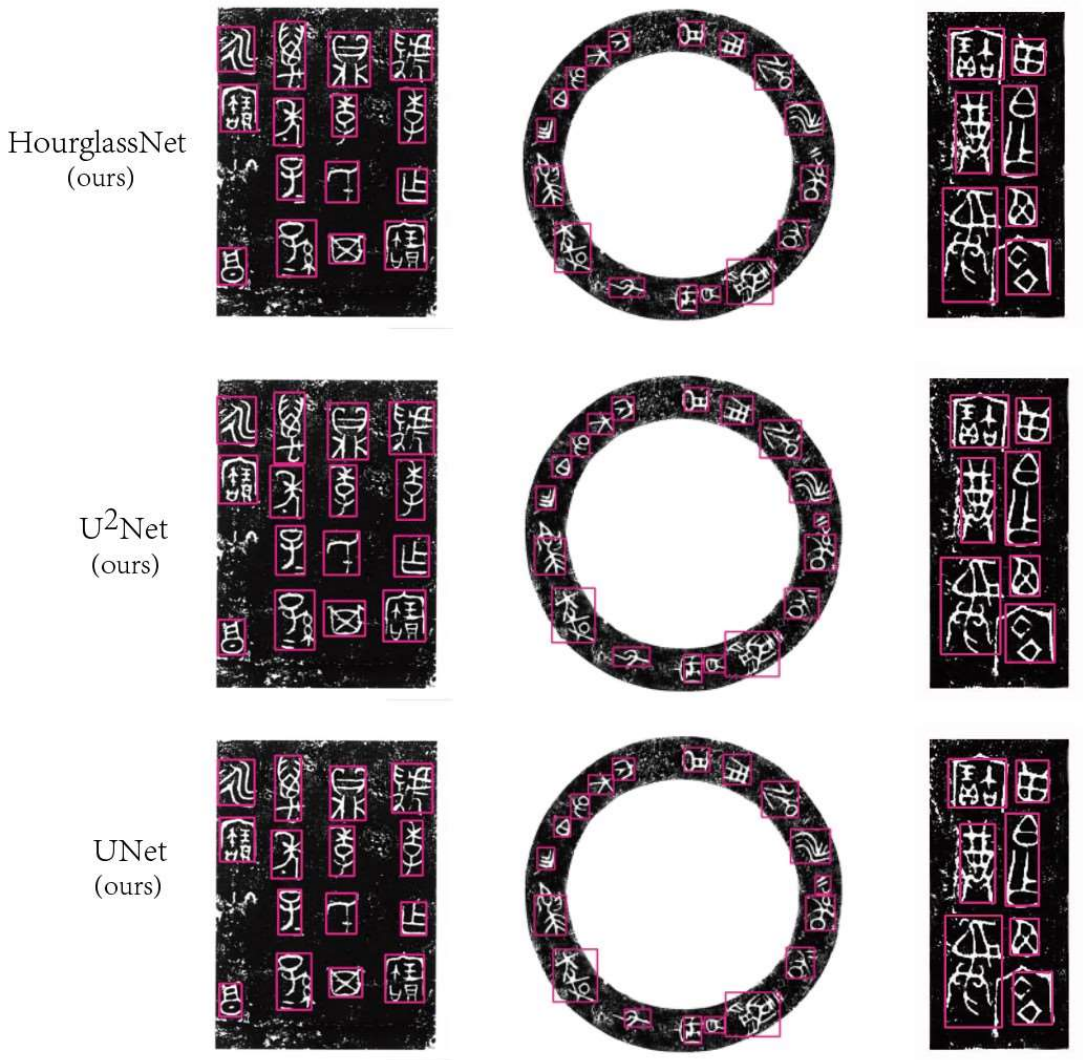


图 4-7 金文拓片预测结果

目前还没有一个规范化标注的金文拓片数据集，但通过本文的实验证明，利用预训练好的模型进行迁移预测是一种有效的方法。这一方法不仅可以产生较好

的预测结果，还能够为构建新的数据集提供重要的辅助，同时，也证明了通过迁移学习利用预训练模型完成与考古交叉的视觉任务的可行性，通过迁移学习，利用预训练好的模型在其他领域学到的知识，对其他古文字数据进行特征提取，为后续的标注工作提供支持，对古文字研究和人工智能模型发展都具有重要意义。

4.4 本章小结

本章节提出了基于伪类别标签辅助的甲骨文字检测模型，通过简单且有效的方式利用了文字先验信息。本章首先介绍了模型的总体结构，然后对于监督数据的生成方式进行了描述，最后介绍了整个模型从输入到输出的工作流程。接下来利用甲骨文数据集对模型的性能进行验证，并通过对比实验和可视化分析展示了提出方法的有效性。实验结果展示本方法可以产生聚焦于文字区域的特征，在不引入过多数量的情况下，提升轻量级模型的性能。在金文数据集中准确的预测结果也说明了本章提出的模型在优化过程中学习到了代表拓片古文字的特征，因此在其他数据集中依旧可以使用，证明了模型在古文字检测任务上迁移学习的能力。

第 5 章 总结与展望

5.1 总结

本文主要对拓片甲骨文字检测任务进行探究，通过探究如何利用甲骨文字形先验知识增强模型对甲骨文字区域的关注能力，提高计算机对于甲骨文字的自动化检测能力。

本文提出了基于交叉注意力机制的特征融合模型，该方法同时利用输入图像的高维特征与甲骨文单字的结构先验特征向量，生成最终的预测结果。通过交叉注意力的查询机制，代表文字结构的特征向量被引入到原始特征图中，使得文字特征与背景特征可以被明显地区分，产生精确的检测结果。实验结果表明，基于交叉注意力将先验知识融合的方法应用于多个常用的分割模型中都获得了指标的提升，并产生了较好的检测效果。

本文还提出了基于多类别伪标签的检测模型，该方法通过多类别预测子任务使模型理解文字区域的结构提示，从而产生可以明确分辨文字区域的特征，最终得到良好的检测结果。实验结果表明，多类别伪标签特征可以作为区别文字与其他部分的有效信号；基于伪类别标签的检测模型可以在满足轻量级要求的同时，达到较好的检测效果。

本文还展示了将训练好的模型迁移到其他古文字数据集的预测结果，证明了在甲骨文数据集中学习到的特征，可以迁移到其他古文字预测任务中。预测结果表明了检测模型在古文字检测任务中的迁移学习能力也证明了将人工智能方法应用于古文字数字化研究是有效的办法。

5.2 展望

针对模型设计方面，本文设计的两个方法在检测准确性上还有很多的进步空间，对于模型本身的设计以及对于参数的调优等办法都存在提升模型性能的可能性。同时，本文的出发点是利用古文字独有的先验信息来辅助模型获得更准确的预测，对于先验信息的获取方式以及表示形式都是值得进一步探寻的方向。

针对甲骨文及古文字数字化的辅助工作方面，通过结合人工智能的方法，将

更多人工智能领域的方法应用于考古研究领域，可以加速考古研究的进程；将更多专业领域的数据应用到人工智能模型中，可以探索人工智能算法在各种任务上的适用性，本工作表明了迁移学习在古文字检测任务上的适用性，通过迁移学习，利用预训练好的模型在其他领域学到的知识，对其他古文字数据进行特征提取，从而为后续的标注工作提供支持，对于学科交叉研究有正向积极的作用，可以促进多个领域的协同发展。

参考文献

[1] 张更明. 甲骨文: 解读中华文明的古老密码[J]. 协商论坛, 2023, (11): 56-58.

[2] 吴琴霞, 高峰, 刘永革. 基于本体的甲骨文专业文档语义标注方法[J]. 计算机应用与软件, 2013,30(10): 60-63.

[3] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2020,9726-9735.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, et al. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML),2020.

[5] Simonyan, K., & Zisserman. A. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR,2014,abs/1409.1556.

[6] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition,2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:770-778.

[7] Goodfellow, Ian J. et al. Generative Adversarial Nets. Neural Information Processing Systems (NIPS),2014.

[8] Zhang, Lvmin et al. Adding Conditional Control to Text-to-Image Diffusion Models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV),2023: 3813-3824.

[9] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., & others. Gradient-based learning applied to document recognition. Proceedings of the IEEE,1998, 2278–2324.

[10] Krizhevsky, A., Sutskever, I., & Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NIPS),2012.

[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke,V. Rabinovich, A. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015:1-9.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. CoRR,2014, abs/1409.0473.

[13] Parikh, A.P., Täckström, O., Das, D., & Uszkoreit, J. A Decomposable Attention Model for Natural Language Inference. ArXiv,2016, abs/1606.01933.

[14] Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez. Attention is All you Need. Neural Information Processing Systems (NIPS),2017.

[15] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. End-to-End Object Detection with Transformers. ArXiv,2020,abs/2005.12872.

[16] Wang, Yuxin et al. ContourNet: Taking a Further Step Toward Accurate Arbitrary-Shaped Scene Text Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2020: 11750-11759.

[17] Kingma, D.P., & Ba, J. Adam: A Method for Stochastic Optimization. CoRR, 2014,abs/1412.6980.

[18] 王慧慧. 大规模甲骨文数据集构建及算法研究[D].河南大学,2019.

[19] Li Q, Yang Y and Wang A. Recognition of inscriptions on bones or tortoise shells based on graph isomorphism. Computer Engineering and Applications,2011,112-114.

[20] Xing Jici et al. Oracle bone inscription detection: a survey of Oracle bone inscription detection based on deep learning algorithm. AIIPCC '19,2019.

[21] Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. AAAI Conference on Artificial Intelligence,2016.

[22] Tang, Jing Rui et al. Few Could Be Better Than All: Feature Sampling and Grouping for Scene Text Detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2022: 4553-4562.

[23] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. EAST: An Efficient and Accurate Scene Text Detector. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017:2642-2651.

[24] Liao, Minghui et al. Real-time Scene Text Detection with Differentiable

Binarization. ArXiv, 2019,abs/1911.08947.

[25] Baek, Youngmin et al. Character Region Awareness for Text Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ,2019: 9357-9366.

[26] Li, Xiang et al. Shape Robust Text Detection With Progressive Scale Expansion Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2018:9328-9337.

[27] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network. ArXiv,2020,abs/2002.10200.

[28] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2021:3122-3130.

[29] Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., & Bai, X. Multi-oriented Text Detection with Fully Convolutional Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016,4159-4167.

[30] Jun Guo, Changhu Wang, Edgar Roman-Rangel. Building hierarchical representations for oracle character and sketch recognition. IEEE Transactions on Image Processing,2015,104–118.

[31] Lin Meng and Tomonori Izumi. A combined recognition system for oracle bone inscriptions. International Journal of Advanced Mechatronic Systems(IJAMS) ,2017

[32] Newell, Alejandro et al. Stacked Hourglass Networks for Human Pose Estimation. European Conference on Computer Vision (ECCV),2016.

[33] Kittenplon, Yair et al. Towards Weakly-Supervised Text Spotting using a Multi-Task Transformer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2022: 4594-4603.

[34] Ye, Maoyuan et al. DeepSolo: Let Transformer Decoder with Explicit Points Solo for Text Spotting. 2023 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR) ,2022:19348-19357.

[35] Xie, Xudong et al. Toward Understanding WordArt: Corner-Guided Transformer for Scene Text Recognition.European Conference on Computer Vision (ECCV),2022.

[36] 刘海强. 基于深度学习的火星地形分割技术研究[D].吉林大学, 2023. DOI:10.27162/d.cnki.gjlin.2023.007607.

[37] Shelhamer, Evan et al. Fully convolutional networks for semantic segmentation. 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ,2015: 3431-3440.

[38] Zeiler, Matthew D. et al. Adaptive deconvolutional networks for mid and high level feature learning. 2011 International Conference on Computer Vision (ICCV) (2011): 2018-2025.

[39] Dumoulin, Vincent and Francesco Visin. A guide to convolution arithmetic for deep learning, 2016, ArXiv abs/1603.07285 .

[40] Melkman A., On-line Construction of the Convex Hull of a Simple Polygon, Information Processing Letters,1987.

[41] Gehring, Jonas et al. Convolutional Sequence to Sequence Learning. ArXiv, 2017,abs/1705.03122 .

[42] Jing Xiong,Qingju Jiao,Guoying Liu,Yongge Liu. Oracle Bone Inscriptions Big Knowledge Management and Service Platform. ICPCSEE Steering Committee. Abstracts of the 5th International Conference of Pioneering Computer Scientists, Engineers and Educators,2019.

[43] He, Mengchao et al. Contest on Robust Reading for Multi-Type Web Images. 2018 24th International Conference on Pattern Recognition (ICPR),2018.

[44] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng. Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT.2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),2017, Vol. 01. 1454–1459.

[45] Meng Lin. 2017. Recognition of Oracle Bone Inscriptions by Extracting Line Features on Image Processing. In International Conference on Pattern Recognition

Applications and Methods

[46] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., & Jägersand, M. U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. Pattern Recognition,2020,106,107404.

[47] Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv,2015, abs/1505.04597

[48] 门艺.甲骨文献的信息化与利用[J].兰台世界,2008.

[49] Flad, R.K. Divination and Power: A Multiregional View of the Development of Oracle Bone Divination in Early China. Curr. Anthropol,2008, 49, 403–437

[50] Fu X, Yang Z, Zeng Z, Zhang Y, Zhou Q. Improvement of Oracle Bone Inscription Recognition Accuracy: A Deep Learning Perspective. ISPRS International Journal of Geo-Information,2022; 11(1):45.

[51] Liu, Rosanne et al. “An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution.” ArXiv,2018, abs/1807.03247 .

[52] 郭昱宇. 基于场景理解的视觉关系生成研究[D]. 电子科技大学,2022.

[53] 刘志基. 谈古文字信息化处理中“字”的处理问题,2002.

[54] Xin-Lun Zhou, Xing-Cheng Hua and Feng Li, "A method of Jia Gu Wen recognition based on a two-level classification," Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR), Montreal, QC, Canada, 1995.

[55] Jun Guo, Changhu Wang, Roman-Rangel E, Hongyang Chao, Yong Rui. Building Hierarchical Representations for Oracle Character and Sketch Recognition. IEEE Trans Image Process. 2016.

[56] Feng Gefei, Gu Shaotong. Feature Extraction Method of Oracle Bone Inscriptions Based on Mathematical Morphology. Journal of Chinese Information Processing. 2013, 27(2): 79-86.

[57] Lin, Tsung-Yi et al. “Focal Loss for Dense Object Detection.” IEEE Transactions on Pattern Analysis and Machine Intelligence 42(TPAMI), 2017.

[58] J. Deng, W. Dong, R. Socher, L. Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database, 2009 IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR),2009:248-255

[59] Redmon, Joseph and Ali Farhadi. YOLOv3: An Incremental Improvement. ArXiv,2018, abs/1804.02767 2018.

[60] Neubeck, Alexander and Luc Van Gool. Efficient Non-Maximum Suppression. 18th International Conference on Pattern Recognition (ICPR),2006: 850-855.

[61] Han, Xu et al. IsOBS: An Information System for Oracle Bone Script. Conference on Empirical Methods in Natural Language Processing (EMNLP) ,2020.

[62] Lin, Xiaoyu et al. Radical-based extract and recognition networks for Oracle character recognition. International Journal on Document Analysis and Recognition (IJDAR), 2022.

[63] Shi, D. et al. RCRN: Real-world Character Image Restoration Network via Skeleton Extraction. Proceedings of the 30th ACM International Conference on Multimedia,2022.

[64] Long, Shangbang et al. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. European Conference on Computer Vision (ECCV) ,2018.

[65] 邱锡鹏神经网络与深度学习，机械工业出版社，https://nndl.github.io/, 2020.

[66] Liu, Guoying et al. An Oracle Bone Inscription Detector Based on Multi-Scale Gaussian Kernels. Applied Mathematics, 2021: 224-239.

[67] 吴振武.吉林大学藏甲骨集[M].上海:上海古籍出版社,2021.