队伍编号	MC2400990
题号	B

基于 Yolov5s 和 ResNet50 的甲骨文智能识别中原始拓片单字自动分割与识别研究

摘 要

本文旨在研究甲骨文智能识别中原始拓片单字自动分割与识别问题，甲骨文作为中国古代文字的重要形式，其智能化识别对于古文字学研究具有里程碑意义。

对于第一问，本文对图像的噪点进行处理。本文进行了六层的预处理建模。依次为双边滤波算法、图像灰度化、局部对比度增强算法、锐化、自适应阈值二值化方法以及灰度反转，六层预处理模型的串联应用，可以有效地提高图像的质量和清晰度。

在第二问和第三问中，我们以 YOLOv5s 为基础模型，并在其中融入了注意力模块，能强化对甲骨文图像中的深层特征和全局信息的捕捉,极大提高了字符检测的准确率和系统的整体效率,确保了在各种背景条件下的高性能的模型应用。

在第四问中，我们利用从前一步骤中训练的模型，从混合甲骨文数据集中框选出单个甲骨文，并应用 ResNet 深度学习架构来进行文字识别。ResNet 通过引入一个创新的分割注意力机制，有效地提升了网络对图像中不同区域的注意力分配，从而提高了模型对于复杂字符形态的识别能力。这一技术适用于处理甲骨文这类历史文本，能够处理文本的多样性和变异性。通过这种方法，我们实现了高精度的字符识别，并将结果进行了可视化展示。

关键字:图像增强，YOLOV5S，ResNet50

目录

一、问题重述

1.1 问题背景

甲骨文拓片图像分割是甲骨文数字化工程中的关键问题，旨在从复杂背景中准确提取出独立文字区域。然而，现有的图像分割方法在面对点状噪声、人工纹理和固有纹理等干扰元素时表现不佳，误分割率较高，且受到图像来源多样性的影响。如何有效应对甲骨文拓片图像中的干扰元素，提高文字区域的准确性和稳健性，仍然是一个具有挑战性的问题。

1.2 问题提出

1.2.1 问题 1

对三张甲骨文原始拓片图片，进行图像预处理，建立甲骨文图像预处理模型。该模型旨在实现对甲骨文图像中干扰元素的初步判别和处理。

1.2.2 问题 2

针对甲骨文原始拓片图像，需要建立一个快速准确的甲骨文图像分割模型，以实现对不同的甲骨文原始拓片图像进行自动单字分割。

1.2.3 问题 3

使用问题 2 建立的甲骨文图像分割模型，对 200 张甲骨文原始拓片图像进行自动分割。确保模型能够准确地识别并分割出图像中的单个文字区域。

1.2.4 问题 4

在前三个问题的基础上，针对甲骨文原始拓片图像的单字分割研究，需要采用适当的方法进行甲骨文原始拓片的文字识别。

二、问题分析

2.1 问题一的分析

首先，考虑到拓片的年代久远和存储条件的多样性，图像中可能存在大量的噪声和损伤，问题一的主要目的则是去除噪声并增强文本特征。甲骨文图像的集合了点状噪声

、人工纹理和固有纹理。通常一个好的预处理需要对固有纹理进行放大与增强来突出文字特征，并且还需要实现淡化点状噪声，综合以上两点来增强图像清晰度。

2.2 问题二的分析

问题二的核心在于选择合适的机器学习模型进行对甲骨文图像的有效且快速的分割，然而，对于本任务来说，由于图像内存在大量噪点因素，由于甲骨文文字整体复杂度并不高，这些噪点相较于实际的甲骨文文字并没有极大的特征差别，容易造成较大干扰，二常规的图像分割算法并不能有效进行建模与分割，选择较为复杂的分割模型是本题的关键。

2.3 问题三的分析

问题三是问题二的实际应用，问题三的测试文件具有随机性与多样性，需要对问题二建立的模型进行多次调参。

2.4 问题四的分析

问题四的难点在于给定的数据集体量较小，无法训练至较好的效果来应对实际甲骨文图片的多样性。因此关键在于对数据集进行扩充，例如对给定的数据集进行数据增强或者数据反转微调。接着是选择图像识别模型，对问题二三分割模型相结合进行模型的组合，实现识别任务。

三、模型假设与符号说明

3.1 模型假设

（1）预处理假设

甲骨文拓片图像中存在各种干扰元素，如点状噪声、人工纹理和固有纹理等，它们会对图像分割产生影响，降低图像的质量和清晰度。然而，通过合适的图像预处理方法，如去噪、灰度化和二值化等操作，可以有效地减少这些干扰元素的影响，提高图像的质量和清晰度。同时，采用特征提取和判别方法，能够有效地区分文字区域与干扰元素，为后续的图像分割工作做好准备。

（2）图像分割模型假设:

假设甲骨文图像中的文字区域具有特定的形态特征，如连续性和紧密性。若图像分割模型能够精确识别并分割出文字区域，并且具备处理各种来源和类型的甲骨文图像的能力。此外，假设建立的图像分割模型具备一定的泛化能力，即能在未曾接触的数据集上产生良好的分割效果。通过适当的预处理方法，如去噪、灰度化和二值化等预处理步

骤，可以提高图像的质量和清晰度，为图像分割模型提供更好的输入。

（3）文字识别模型假设:

假设甲骨文图像中的文字具有独特的形态和结构特征，如笔画组成和笔画顺序。若所建立的文字识别模型能够精确识别甲骨文图像中的文字，并输出正确的文字序列。此外，假设该文字识别模型具备处理各种字体、字形和不同风格的能力，并且具备一定的鲁棒性和泛化能力，能够在不同情况下产生可靠的识别结果。通过适当的模块，如特征提取和判别方法，可以进一步提高识别模型的性能和鲁棒性，使其在复杂环境下也能表现出良好的识别能力。

3.2 符号说明

I：原始甲骨文拓片图像。
Ip：预处理后的甲骨文拓片图像。
D：图像的干扰元素，可能包括噪声、人工纹理和固有纹理。
S：图像分割算法或模型。
R：字符识别算法或模型。
C：甲骨文字符集。
c：单个甲骨文字符。
V：特征向量空间。
f (.)：特征提取函数，将图像映射到特征空间。
B：字符的边界框。
T：测试集中的甲骨文图像。
Y：识别结果集。
θ：模型参数。
η：学习率或其他优化参数。
L：损失函数。
A：模型评估指标集，如准确度、召回率等。
G：增强后的甲骨文图像数据集，用于训练时的数据增强。
N ：噪声模型，用于生成合成的干扰元素以训练模型。
M：甲骨文的异体字集。
α：某一特定的异体字。
E：图像的增强算法或操作，用于改善图像质量，例如对比度调整。
Z：图像的 z 分数，用于标准化图像亮度和对比度。
P ：预测函数，用于基于特征向量 V 输出字符 c 的概率。
H：信息熵，用于评估模型的不确定性。
ϵ：模型容忍的误差阈值，用于决定是否接受分割或识别结果。

四、问题一的建立与求解

4.1 模型的建立

对于给定的三张甲骨文图片，建立模型的核心目标就是找到效果较好的去噪组合。

4.1.1 模型的选择

预处理模型主要分为六层。第一层选择双边滤波算法进行进行图像的去噪，对图像进行效果减弱，淡去纹理噪声[1]，第二层择使用图像灰度化，目的在于强化局部白色字体的清晰度。接着通过第三层局部对比度增强（CLAHE）算法进一步增强图像的对比度，使细节更加清晰。第四层通过增强锐化白色部分，突出图像中的亮部。第五层采用自适应阈值二值化方法，将图像转换为黑白二值图像，以便更好地提取图像中的目标特征。最后一层则是灰度反转，将二值化后的图像进行反转处理，以白色为字体的主要颜色，以便后续识别。六层预处理模型的串联应用，可以有效地提高图像的质量和清晰度。预处理结构如下图所示：

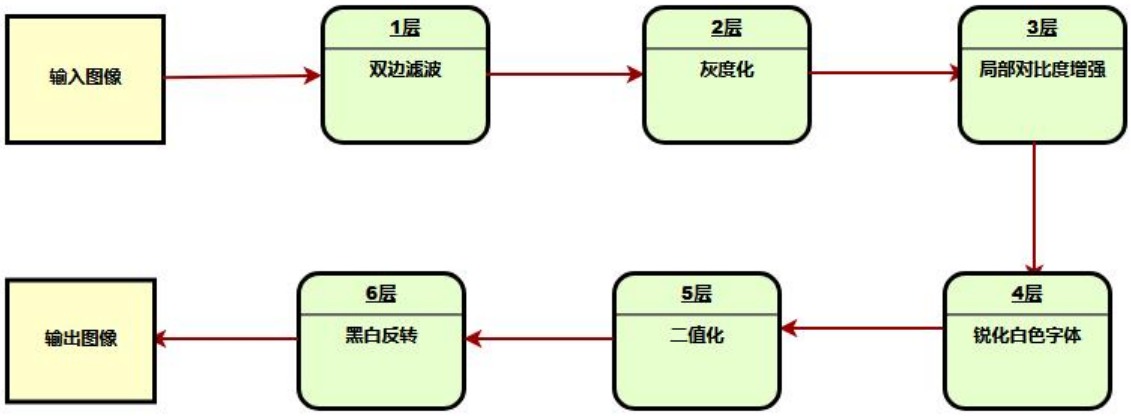


图 1图像预处理模型结构

4.1.2 数学模型和算法流程

图像预处理 $I_p = E(I)$

对于每张原始拓片图像 I，我们定义预处理函数 E 来得到预处理后的图像 I_p。

噪声去除：$I_{smooth} = f_{smooth}(I, K)$

其中 f_{smooth} 是一个滤波函数，K 是滤波器核。对于中值滤波器：K = *Median Kernel*

光照校正：$I_{adj} = f_{hist}(I_{smooth})$

二值化：$I_{bin} = f_{thresh}(I_{adj}, T)$

其中 f_{thresh} 是一个阈值函数，T 是由大津法计算得到的阈值。

反转：$I_{clean} = f_{morph}(I_{bin}, M)$

其中 f_{morph} 表示形态学操作函数， M 是形态学结构元素。

最终的预处理图像 $I_p = I_{clean}$

4.2 结果检验与分析

中间层的局部对比度增强是该模型的关键，能非常好的强化文字的特征，后续的操作层将进一步对此进行数据增强。

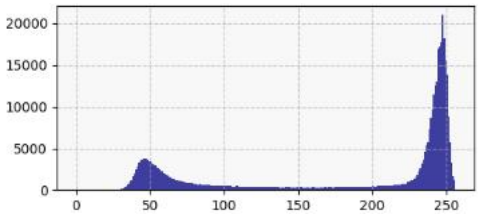


图 2 降噪前

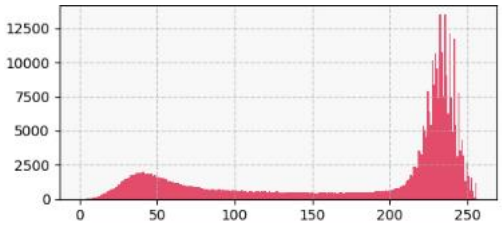


图 3 降噪后

图 4 为原始图像，图 5 为经过预处理后的效果，可以看出本模型将甲骨文图像的各部分字体进行了更为明显的突出，使得后序对分割模型的训练提高了精准度。

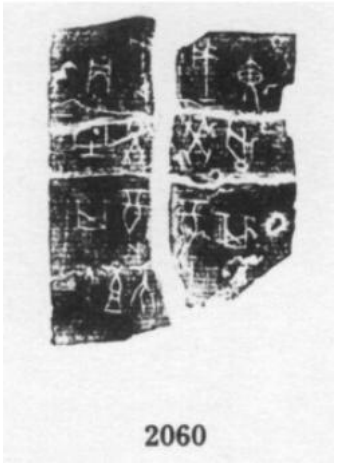


图 4

图 5

五、问题二、三模型的建立与求解

5.1 模型的建立

5.1.1 模型的选择

随着深度学习技术发展，来越多的目标检测技术应用到文字检测中。Redmon 等在 2016 年提出的 YOLO[2]系列性能更高,Redmon 在 2017 年提出 YOLOv2,使用 darknet-19 网络作为主干网络，在简化网络续构的同时提高了目标检测的准确率。Alexey 在 2020 年提出 YOLOv4[3],使用 CSPDarknet53 作为主干网络并采用 FPN 和 PAN 融合特征图，进一步提高了检测准确率。同年，Ultralytics 提出了 YOLOv5，模型检测准确率高于以往的目标检测模型且检测速度快,YOLOv5 也因此成为目前目标检测表现最好的网络模型之一。甲骨文存在大量与文字特征相似的裂痕，使用 YOLOv5 等现有技术进行文字检测会受到裂痕等复杂纹理背景信息的干扰以及甲骨文文字具有密集、粘连的特点，造成识别失误。在本研究中，我们选择了 YOLOv5s 作为甲骨文图像识别任务的模型。YOLOv5 是一种轻量级目标检测算法，具有快速和高效的特点，适用于处理大规模数据集。我们选择 YOLOv5s 版本是因为其在保持较高检测性能的同时具有较小的模型体积，可以在资源受限的环境下进行部署和应用。此外，本文在 YOLOv5s 的基础上引入注意力机制，进一步提高了检测精度和速度。

5.1.2 数学模型和算法流程

5.1.2.1 yolov5s 模型

本文提出了一种基于 YOLOv5s 的甲骨文文字检测算法，主要包括 Input 输入端、Backbone 特征提取端、Neck 颈部端和 Prediction 预测端。在 Input 端，采用了 Mosaic 数据增强和自适应锚框计算方法。Backbone 部分由切片结构 Focus、4 次卷积 ConV、C3 模块、空间金塔池化 (Spatial pyramid pooling, SPP) 模块和自注意力机制模块组成。该模块能够更好地关注甲骨文文字的深层特征，同时捕捉全局信息，减弱甲骨文图像上裂痕和粘连对文字检测的干扰。Neck 端采用 FPN 和 PAN 结构，融合特征过下采样输出端生成 3 个特征图，用于检测不同尺寸的目标。该甲骨文文字检测算法的网络结构如图 6 所示。

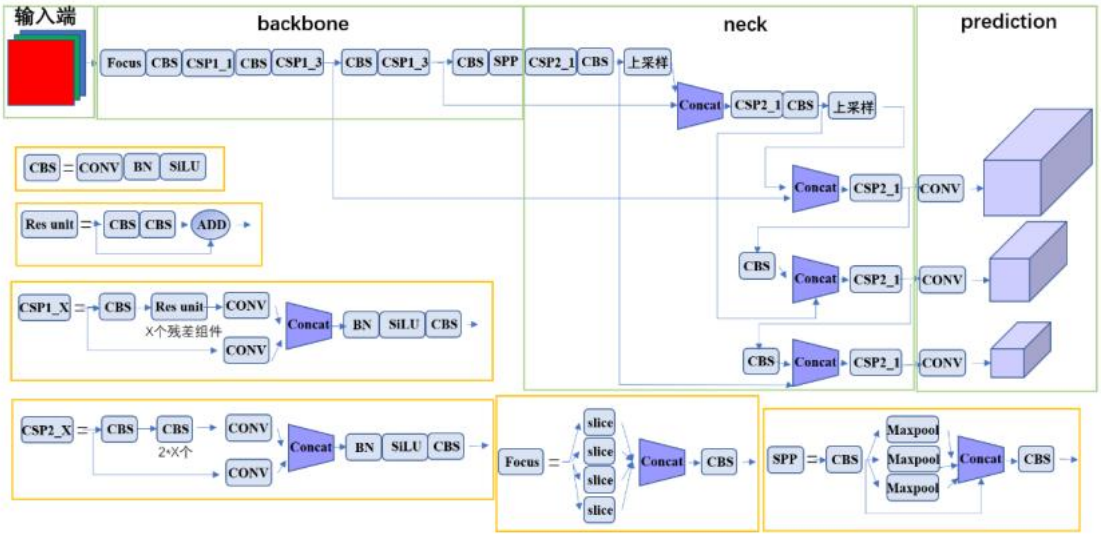


图 6

5.1.2.2 标注文件转换

完成建立模型后，将要对数据集进行格式转换以适应yolov5s的训练要求，从提供的训练集数据格式中可以看到，Train 文件夹中的每张图像都有一个对应的JSON文件，包含字符的位置信息。这个信息是用于训练和验证分割模型的标注数据。这些JSON 文件包含了图像名称和一系列标注（ ann ），每个标注代表一个甲骨文字符的边界框，格 式为 [x_min, y_min, x_max, y_max, label]。首先思路则是将json标注集文件转化为txt格式的标注集文件。

对于yolov5s数据及要求：

假设图像的高和宽分别为h,w,bbox的左上角坐标为(x1,y2)，右下角坐标为(x2,y2)，则可求得bbox中心坐标(x_cy_c)为

xc=x1+(x2-x1)/2 =(x1+x2)/2

yc=y1+(y2-y1)12=(y1+y2)12

假设yolo的5个数据分别为:label,x,y,w,h，则有对应关系:

x_=(x1+x2)/2wy_=(y1+y2)/2hw_=(x2-x1)/wh_=(y2-y1)/h

反过来，则有:

*x1=w*x-05*w*WX2=w*x +05*w*Wy1=h*y_-05*h*hy2=h*y_+0.5*h*h*

因此需要进行标注文件转换建模，由于是分割任务，类别为0即可。给出的json文件包含了对应的文件名、图片左上角坐标以及图片右下角坐标(ann)，进行对应的转换即可，以下为数学公式

-> 获取对象边界框的坐标

x_min, y_min, x_max, y_max, class_index = ann

->计算边界框的中心点坐标和宽度、高度

*x_center = (x_min + x_max) / (2 * img_width)*

*y_center = (y_min + y_max) / (2 * img_height)*

width = (x_max - x_min) / img_width

height = (y_max - y_min) / img_height

->将对象信息转换为YOLOv5格式

5.2 结果检验与分析

5.2.1 评估指标

本文选取训练损失（Loss）、准确率(Precision，P)、召回率(Recall），平均精度均值(mAP0.5)作为评估指标。

（1）训练损失（Loss）：表示模型预测结果与真实结果差距。

（2）准确率（Precision）：表示正确检测的物体的个数占总检测物体个数的百分比。

（3）召回率（Recall）：表示正确检测的物体的个数占测试集中物体的总数的百分比。

（4）平均精确度（AP）：以召回率（Recall）为横坐标，准确率（Precision）为纵坐标，在一定阈值的基础之上形成的曲线称为 P-R 曲线。P-R 曲线下方围成的面积即为平均精确度（AP）。

（5）平均精度均值(mAP0.5)：基于 0.5IoU 的平均精度均值(mean AP, mAP)。

5.2.2 模型训练

本文使用 Windows10 64 位系统,实验环境为 python3.10、pytorch1.8.0。模型需在 NVIDIA RTX3060GPU 运行，在相同超参数下进行训练、验证和测试。图片设置为 640X640JPG 格式，Batchsize 设置为 3，训练 40 个 epoch，以下是训练过程中一个过程数据，我们对其观察其模型的在训练上的效果

5.2.3 结果检验与分析

（1）train 损失函数图像

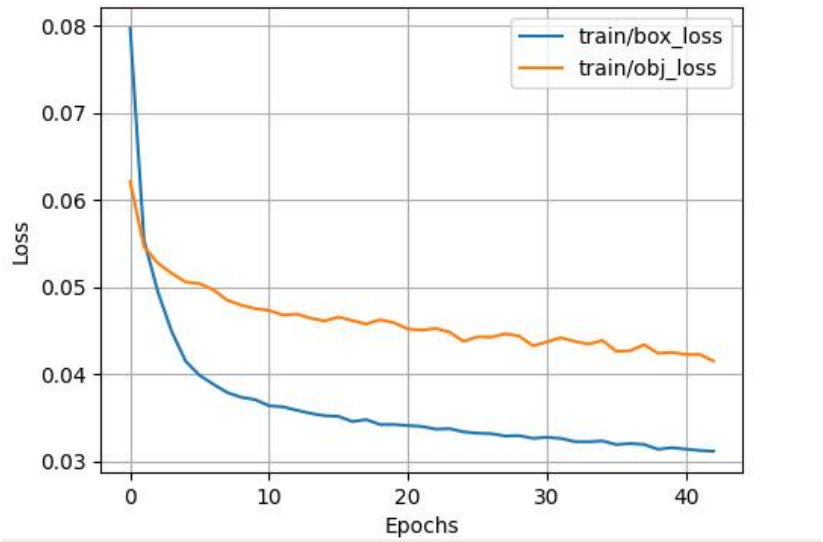


图 7

（2）test 损失函数图像

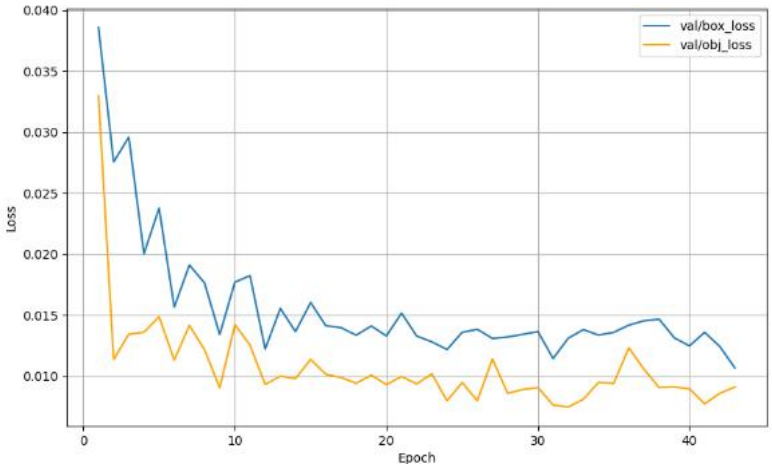


图 8

（3）precision 结果图像

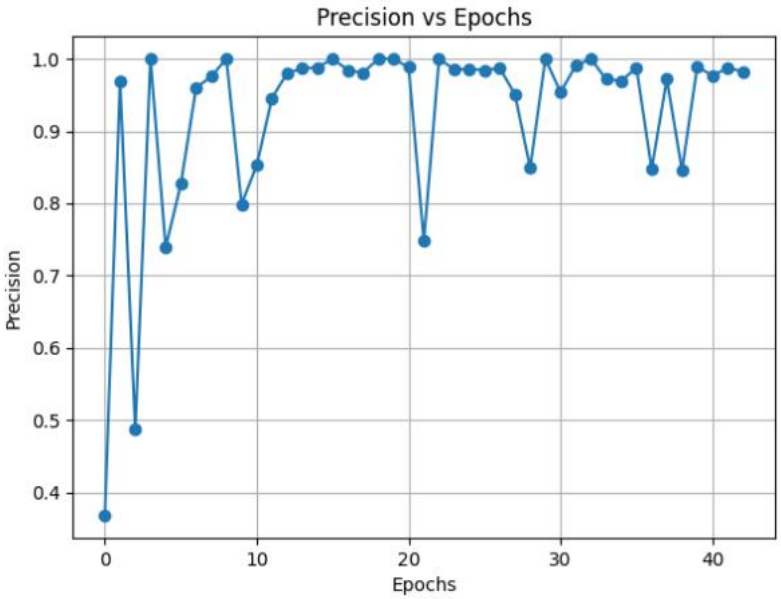


图 9

（4）recall 结果图像

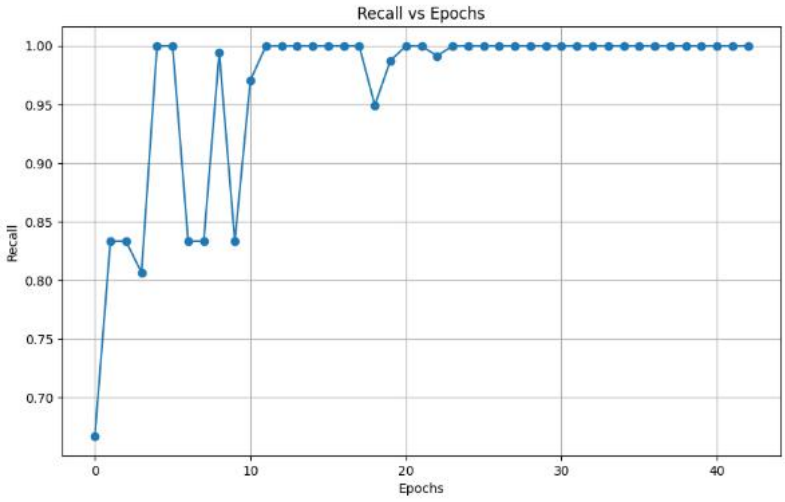


图 10

（5）mAP0.5结果图像

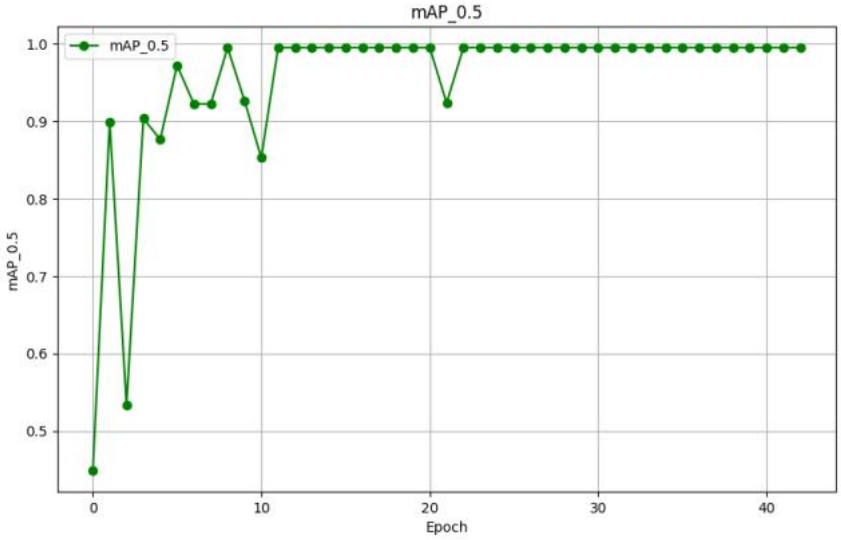
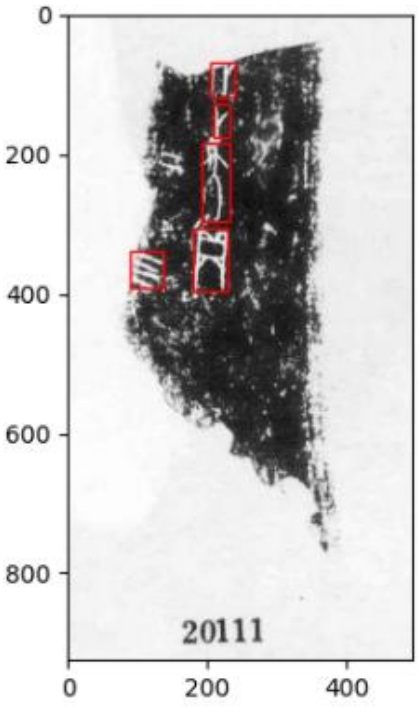


图 11

以图像‘020111.jpg’为例,该分割的矩阵坐标分别为:[204.0, 67.0, 239.0, 118.0, 1.0], [209.0, 127.0, 233.0, 175.0, 1.0], [90.0, 339.0, 137.0, 392.0, 1.0], [191.0, 183.0, 231.0, 297.0, 1.0], [180.0, 306.0, 230.0, 397.0, 1.0]，如下图所示，能精准的对该甲骨文文字进行识别。



模型训练的结果参数保存至指定的 excel 文件中,表格内部分结果如图 12 所示：

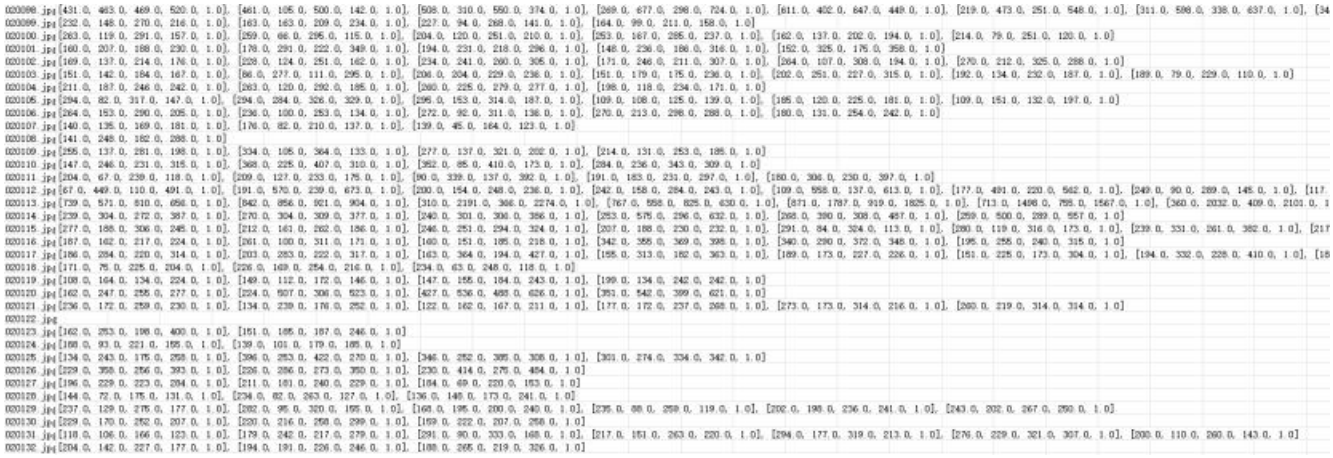


图 12

六、问题四模型的建立与求解

6.1 模型的建立

问题四在前两问字体分割的基础上，进行甲骨文的识别。本题选择 ResNet 作为主要的网络结构，对附件四的训练集和测试集进行调试。

6.1.1 ResNet 分类识别模型

ResNet 是一种深度残差网络[4]，用于图像分类任务。其核心思想是通过引入残差块（Residual Block）来解决深层神经网络训练中的梯度消失和梯度爆炸问题。ResNet 的基本结构是由多个残差块组成的深层网络。每个残差块包含两个主要分支：一个是跨层的快捷连接（Shortcut Connection），另一个是主路径（Main Path）。主路径由一系列卷积层组成，用于学习特征表示；快捷连接直接将输入添加到主路径的输出上，从而保留了原始输入的信息。这种设计使得网络能够更轻松地学习恒等映射，从而更有效地训练深层网络。ResNet 在 ILSVRC 和 COCO 等多个视觉任务上取得了优秀的性能，成为了深度学习中的经典模型之一。

6.1.1.1 残差结构

残差结构 1： 原论文将输入 X(即 input)经过一系列处理之后得到残差 F(X)，若是在 shortcut 分支上不经过 downsample 处理(即不经过 conv1x1 卷积+BN 处理)，则在最终得到的残差映射函数为 F(X) + X，此种结构一般用在 conv_x 组块中的非第一层之后的层，即用在不需要改变前后输入输出维度的层。

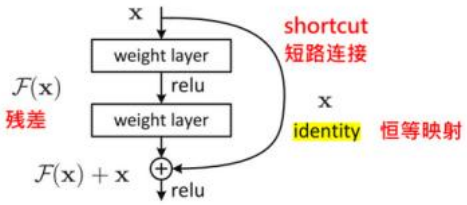


图 13

残差结构 2： 残差结构 2 提到的类似这种需要在 shortcut 分支上进行 downsample 处理的结构，一般用在每个 conv_x 组块的第一层中，即上一层的输出 out_channel 不符合此层所需要的 in_channel，此时需要用 conv1x1 卷积进行升维操作，此时得到的残差映射函数为 F(X) + G(X)，G(X)为 shortcut 分支上对输入 X 进行处理后得到的恒等映射。

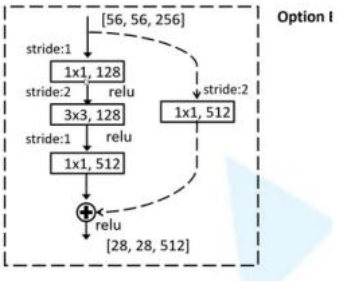


图 14

6.1.1.2 模型架构

具体来说，本任务选择了 ResNet50 模型，是 ResNet 系列中的一个具体模型，它由 50 层深度组成，包括堆叠的残差块。BottleNeck 应用于该网络的 Residual 结构中。整个网络结构由卷积层、批量归一化层、激活函数和残差块组成。ResNet50 首先包含一个 7x7 的卷积层用于图像的初步特征提取，然后是一系列堆叠的残差块。每个残差块内部包含了多个卷积层，以及跨层连接的快捷路径。这种深度残差网络结构使得 ResNet50 在处理大规模图像分类任务时表现出色，并且在训练过程中可以避免梯度消失和梯度爆炸问题。具体结构如图 15 所示。

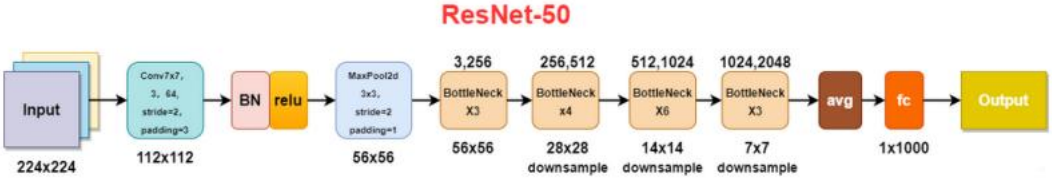


图 15

6.1.2 数学模型和算法流程

6.1.2.1 图片分割

在第二三问进行准确分割的基础上，将单独的甲骨文字分割出来，方便后续识别任务的进行如图

图 16

6.1.2.2 算法流程

1.输入图像：ResNet-50 的输入是一张大小为 224x224 的 RGB 图像。

2.卷积和池化：输入图像首先经过一层 7x7 的卷积层，然后通过一个 2x2 的最大池化层进行下采样，将图像尺寸缩小为原始的一半。

3.残差块：ResNet-50 由多个残差块组成，每个残差块内部包含多个卷积层。这些残差块有助于解决梯度消失和梯度爆炸等问题，使得网络可以训练得更深。具体来说，ResNet-50 包含 4 个阶段（stage），每个阶段包含多个残差块。

4.全局平均池化：在最后一个残差块的输出上应用全局平均池化，将每个特征图的尺寸降为 1x1，这样得到的特征图可以直接输入到全连接层中。

5.全连接层：全局平均池化后的特征图被展平为一个向量，并传递给全连接层。全连接层负责将这些特征映射到对应的类别上，输出图像的分类概率。

6.Softmax 层：最后一层是 softmax 层，将全连接层的输出转换为概率分布，表示输入图像属于每个类别的概率。

6.2 结果检验与分析

如图 17 所示，模型能准确识别出对应的甲骨文字。

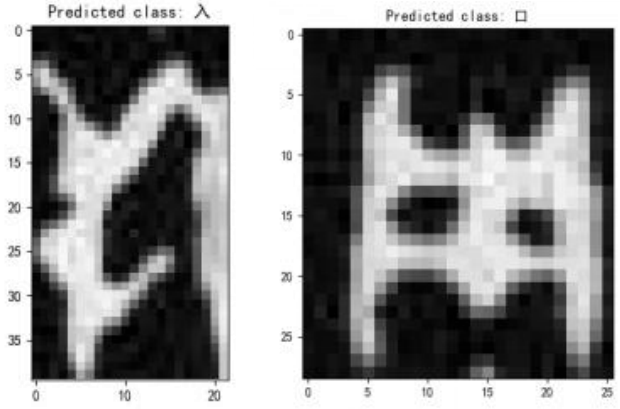


图 17

七、评价与改进

7.1 模型的评价

该工作结合了 改进的 YOLOv5s 和 ResNet50 两种模型，分别对甲骨文图像进行分割与分类，充分利用它们各自的优势来完成甲骨文图像的分割和识别分类任务。其中，YOLOv5s 能够快速地对图像进行目标分割，而 ResNet50 则能够对图像中的对象进行更准确的分类识别。整体而言，本模型在图像分析领域取得了一定的成果，但仍需要针对具体任务和数据集进行进一步的优化和调整，以提高模型的性能和泛化能力。

7.1.1 模型优点

1.对于改进的 yolov5s 模型而言，能非常好的处理图片的特征，以至于较为模糊的图像也能进行识别。在 40 个训练轮次内即可到达一个较为理想的训练效果，并且对计算机设备的要求并不需要很高。

2.对于改进的 resnet 模型而言，模型可以学习到更加抽象和复杂的特征，从而提高了模型的性能和泛化能力。除此之外，通过使用残差连接使得模型能够更有效地传播梯度。该模型还具备更快的收敛速度，在相同的训练迭代次数下，模型可以达到更好的性能。

7.1.2 模型缺点

1.对于复杂或模糊的图像处理能力不足：在处理复杂、模糊或低分辨率的图像时，可能存在识别错误或漏识别的情况。特别是对于一些细微或不规则的图像特征，模型可能无法进行准确识别。

2.容易受到噪声干扰：对于输入数据中的噪声或干扰比较敏感。如果输入数据中存在大量噪声或不相关的信息，模型可能会产生误导性的结果，从而降低了整体性能。

3.需要大量的训练数据和计算资源：尽管本文改进的模型在该任务上表现出色，但要达到最佳性能需要大量的训练数据和计算资源。对于一些资源受限或数据稀缺的情况，可能难以训练出高质量的模型。

4.存在过拟合的风险：由于网络深度较大，存在过拟合的风险，特别是当训练数据不足或训练过程未充分调优时。过拟合会导致模型在测试数据上的泛化能力下降，从而影响了模型在实际应用中的表现。

7.2 模型的改进

（1）对于分割模型而言，需要在原本的模型基础上填充一些功能型模块，例如注意力机制

（2）对于识别模型而言，首先需要扩充数据集，或者对数据集进行各种增强处理。其次，该模型为组合模型架构，对照实验数量不够，可能有另外的组合架构会实现更优的效果。

（3）进行超参数的调整，需要后续进行更多的对照实验，得到更佳的参数组合。

八、参考文献

[1]王玉灵.基于双边滤波的图像处理算法研究[D].西安电子科技大学[2024-04-14].DOI:10.7666/d.y1668012.

[2]Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//Computer Vision & Pattern Recognition. IEEE, 2016. DOI:10.1109/CVPR.2016.91.

[3]Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020. DOI:10.48550/arXiv.2004.10934.

[4] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016. DOI:10.1109/CVPR.2016.90.