2024-2025 学年第 2 学期

图像数据挖掘技术与应用课程

（非试卷）

课程代码 05b00081

适用年级 2022 级

适用专业 数据科学与大数据技术

考核要求：

甲骨文是我国目前已知的最早成熟的文字，现希望通过对已标记的甲骨文图像进行特征提取和建模，从而在一定范围能够准确识别出甲骨文图像对应的汉字。以下为具体任务。

任务一：附件 1（train 文件夹）为单字图像标注数据集，各子文件夹的名称即为子文件夹内图像对应的汉字标记（共 75 个汉字，非均匀分布的图像样本 36441 张）。train 文件夹中的文件结构示例如下：

```
+train
  +宾
    -0000.bmp
    -0001.bmp
    ...
  +丙
    -0000.bmp
    -0001.bmp
    ...
```

设计单字图像识别的合理模型（包括预处理算法），从

train 文件夹中自行划分出训练样本和验证样本对模型进行实验，得到实验情境下的最优模型，将实验设计、过程及结果整理到报告中；

任务二：使用最优模型对 test 文件夹中的图像进行识别，将结果填入“预测结果.xlsx”的`label`列中。注意：`label`列要填入的是与测试图像名称一一对应的中文汉字，应使用 pandas 批量填入。

注意事项：

1.项目完成后，提交“预测结果.xlsx”文件（名称需改为学号姓名.xlsx）、报告模板.docx、以及以 zip 格式压缩的程序源代码（不得包含任何数据集文件，模型文件仅保存用于任务二的最终模型参数 pth 文件）；

2.严禁抄袭，若与他人的程序代码或报告高度雷同，一经核实，抄袭者与被抄袭者同为 0 分处理；提交的程序代码运行结果若与报告所反馈的数据严重不符，一经核实，以 0 分处理。