数据集生成实验报告

一、基本信息

字段	内容
实验名称	智能阅卷助手 — 数据集生成（数据扩增与标签集构建）
姓名	吴耔锦
学号	Y30261529
学院	社会与公共管理学院

二、数据扩增实验结果

2.1 数据扩增概述

> **扩增策略：**对原始采集的 20 张答题卡图片，分别施加 6 种图像变换（椒盐噪声、高斯噪声、高斯滤波、旋转90°、旋转180°、水平翻转），每张原图生成 6 张变换图。扩增后数据集总量达到 **140 张**，数据量提升为原来的 **7.0 倍**。

20	6	140
原始图片数	扩增方法数	扩增后总数

7.0x
数据量提升倍数

2.2 数据扩增数学原理

以下从数学角度形式化描述各扩增方法的变换原理：

1. 椒盐噪声 (Salt-and-Pepper Noise)

对于图像中每个像素点 $I(x, y)$，以概率 p 将其替换为 0（椒，黑）或 255（盐，白）：

$$I'(x, y) = \begin{cases} 0 & \text{概率 } p/2 \text{（椒）} \\ 255 & \text{概率 } p/2 \text{（盐）} \\ I(x, y) & \text{概率 } 1-p \text{（保持不变）} \end{cases}$$

本实验参数：噪声比例 $p = 10\%$

2. 高斯噪声 (Gaussian Noise)

对每个像素叠加服从正态分布的随机值：

$$I'(x, y) = I(x, y) + N(\mu, \sigma^2)$$

其中 $N(\mu, \sigma^2)$ 为均值 $\mu = 0$、标准差 σ 的高斯分布。该变换使模型学习到对微小像素扰动的鲁棒性。

本实验参数：均值 $\mu = 0$，噪声比例 10%

3. 高斯滤波 (Gaussian Blur)

使用二维高斯核进行卷积平滑：

$$I'(x, y) = G(x, y; \sigma) * I(x, y)$$

$$G(x, y; \sigma) = (1 / 2\pi\sigma^2) \cdot \exp(-(x^2+y^2) / 2\sigma^2)$$

其中 $*$ 表示卷积运算，核大小为 7×7，标准差 $\sigma = 1.5$。

4. 旋转变换 (Rotation)

以图像中心为原点，旋转矩阵 $R(\theta)$ 将坐标 (x, y) 变换为 (x', y')：

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x - cx \\ y - cy \end{bmatrix}$$

其中 (cx, cy) 为图像中心坐标。本实验分别采用 $\theta = 90°$ 和 $\theta = 180°$。

5. 水平翻转 (Horizontal Flip)

沿垂直轴进行镜像变换：

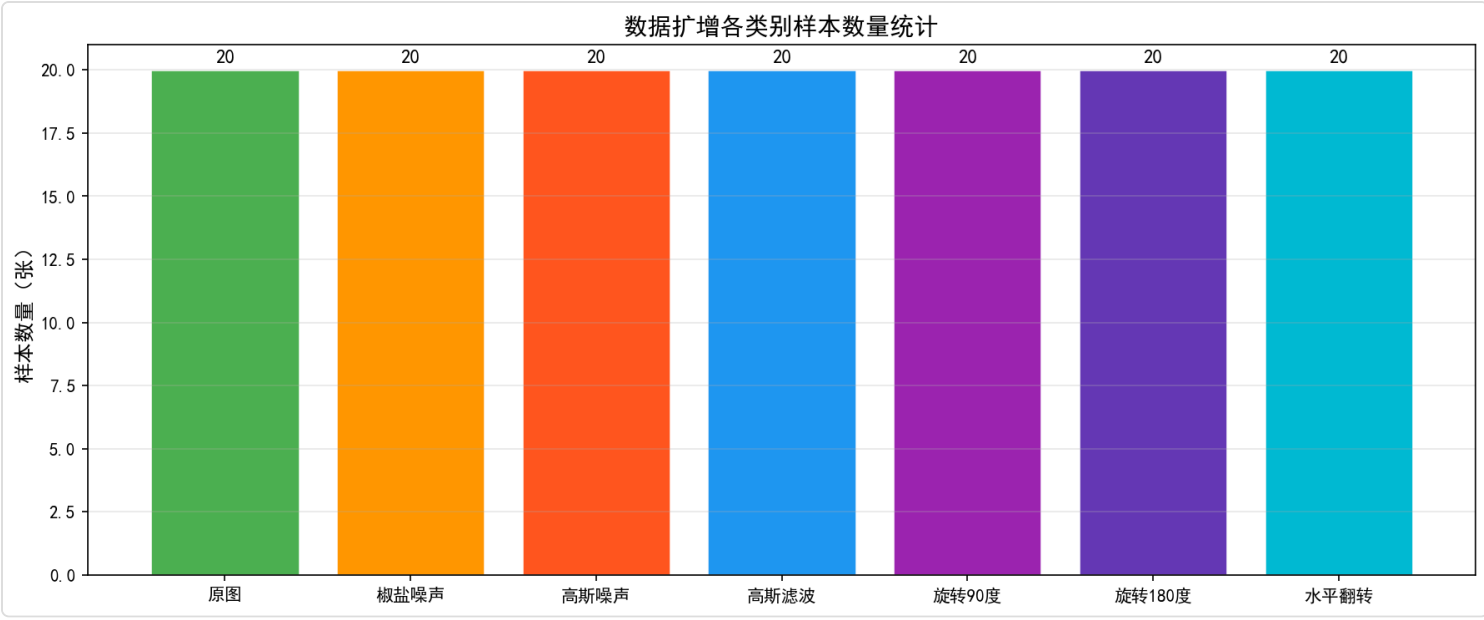
$$I'(x, y) = I(W - 1 - x, y)$$

其中 W 为图像宽度。该变换不改变文本内容的方向可读性，适用于 OCR 场景的数据扩增。

2.3 扩增方法详细说明

扩增方法	生成数量	原理说明	作用与适用场景
原图（baseline）	20	原始采集的答题卡图片，未经任何变换	作为对照组，保留原始数据分布
椒盐噪声	20	在图像上随机添加黑色（椒）和白色（盐）像素点，模拟信号脉冲干扰	增强模型对椒盐噪声的鲁棒性，适用于扫描质量不佳的场景
高斯噪声	20	添加服从正态分布的随机噪声，每个像素点叠加微小随机值	提高模型泛化能力，避免对精确像素值的过度拟合
高斯滤波	20	使用 7×7 高斯核进行加权平均平滑，降低图像细节清晰度	模拟不同拍摄设备、焦距条件下的模糊图像
旋转90°	20	以图像中心为旋转点，顺时针旋转90度	增强模型对竖排文字和不同排版方向的适应能力
旋转180°	20	以图像中心为旋转点，顺时针旋转180度	增加旋转角度多样性，进一步提升方向不变性
水平翻转	20	沿垂直轴进行镜像翻转，左右互换	增强位置不变性，模拟不同扫描方向
合计	**140**	**—**	**数据量提升至原来的 7.0 倍**

2.4 各类别样本数量统计图



2.5 扩增效果可视化对比

以下展示两张原始图片分别经过六种变换后的视觉效果，可以直观对比各类扩增方法的特点：



三、标签集生成结果

3.1 标签格式说明

> **转换流程：**LabelStudio 导出的标注文件为 JSON Lines 格式，每行包含一张图片的所有标注框信息。程序逐行解析 JSON，提取每个文本区域的 `bbox` 坐标（x, y, width, height）和 `transcription` 文本内容，使用 PIL.Image.crop() 按坐标精确裁剪，最终生成 CnOCR 标准格式标签文件。

3.2 CnOCR 标签格式规范

列	字段名	说明	示例
第1列	图片文件名	crop后的子图文件名，与 images/ 目录下的文件一一对应	`xxx.jpg_M9-HYLZUdz.jpg`
分隔符	制表符 ( )	固定使用制表符分隔，不能用空格替代	—
第2列	文本标签	图片中文字内容，每个字符之间用空格分隔	`班 级` 、 `姓 名`

3.3 标签内容示例

序号	图片文件名	标签内容（空格分隔字符）	实际文本
1	`0a9ad9fec78f5d84ce04be3d0c8fe0a1.jpg`	`B 班`	B班
2	`0a9ad9fec78f5d84ce04be3d0c8fe0a1.jpg`	`班 级`	班级
3	`0a9ad9fec78f5d84ce04be3d0c8fe0a1.jpg`	`王 三`	王三
4	`0a9ad9fec78f5d84ce04be3d0c8fe0a1.jpg`	`姓 名`	姓名
5	`0a9ad9fec78f5d84ce04be3d0c8fe0a1.jpg`	`2 1`	21
6	`0a9ad9fec78f5d84ce04be3d0c8fe0a1.jpg`	`学 号`	学号

3.4 标签长度分布分析（描述统计）

对数据集中所有标签文本的字符长度（去除空格后）进行描述统计分析，以评估数据集的文本复杂度分布：

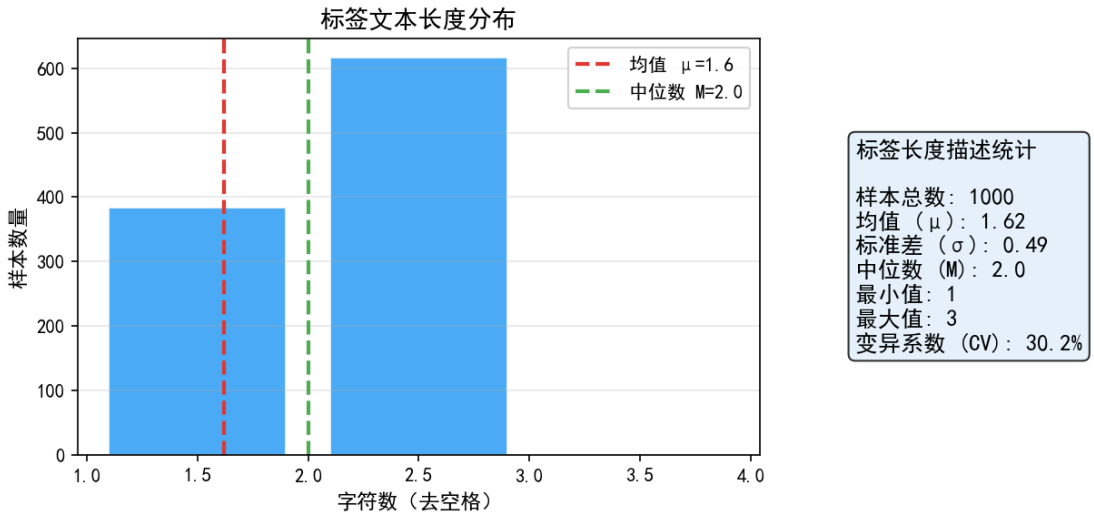
1.6	0.5	2

长度均值 μ	标准差 σ	中位数 M

1-3	30.2%
长度范围	变异系数 CV

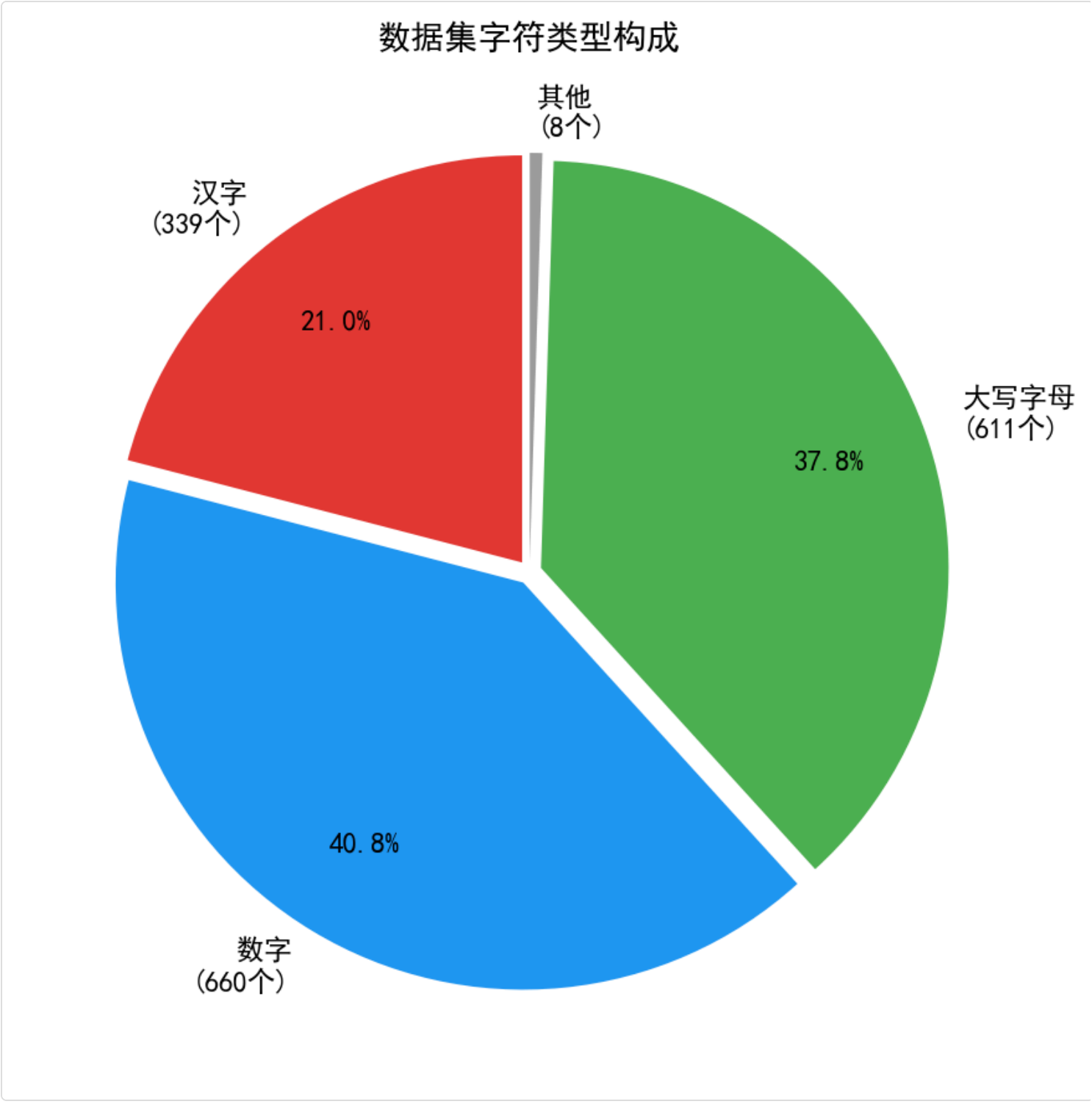
关键发现：变异系数 CV = 30.2%，说明标签长度分布较为集中，以短文本为主（如单字符、双字符字段），适合OCR行识别模型的训练。

文本长度	样本数	占比	频率可视化
1 字符	383	38.3%	
2 字符	616	61.6%	
3 字符	1	0.1%	



3.5 字符类型构成分析

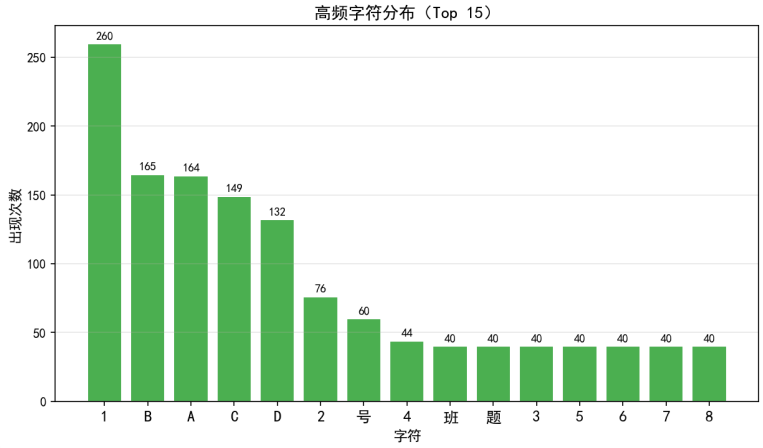
从字符类别维度分析数据集的构成，评估其对多类型字符的覆盖能力：



3.6 高频字符分布

字符	出现次数	频率可视化
「1」	260	
「B」	165	

「A」	164	
「C」	149	
「D」	132	
「2」	76	
「号」	60	
「4」	44	
「班」	40	
「题」	40	



标签字符统计摘要

总字符数：1618
唯一字符数：15
最高频字符：「1」(260次)
前5位：1(260) | B(165) | A(164) | C(149) | D(132)

四、数据集划分结果

4.1 划分策略说明

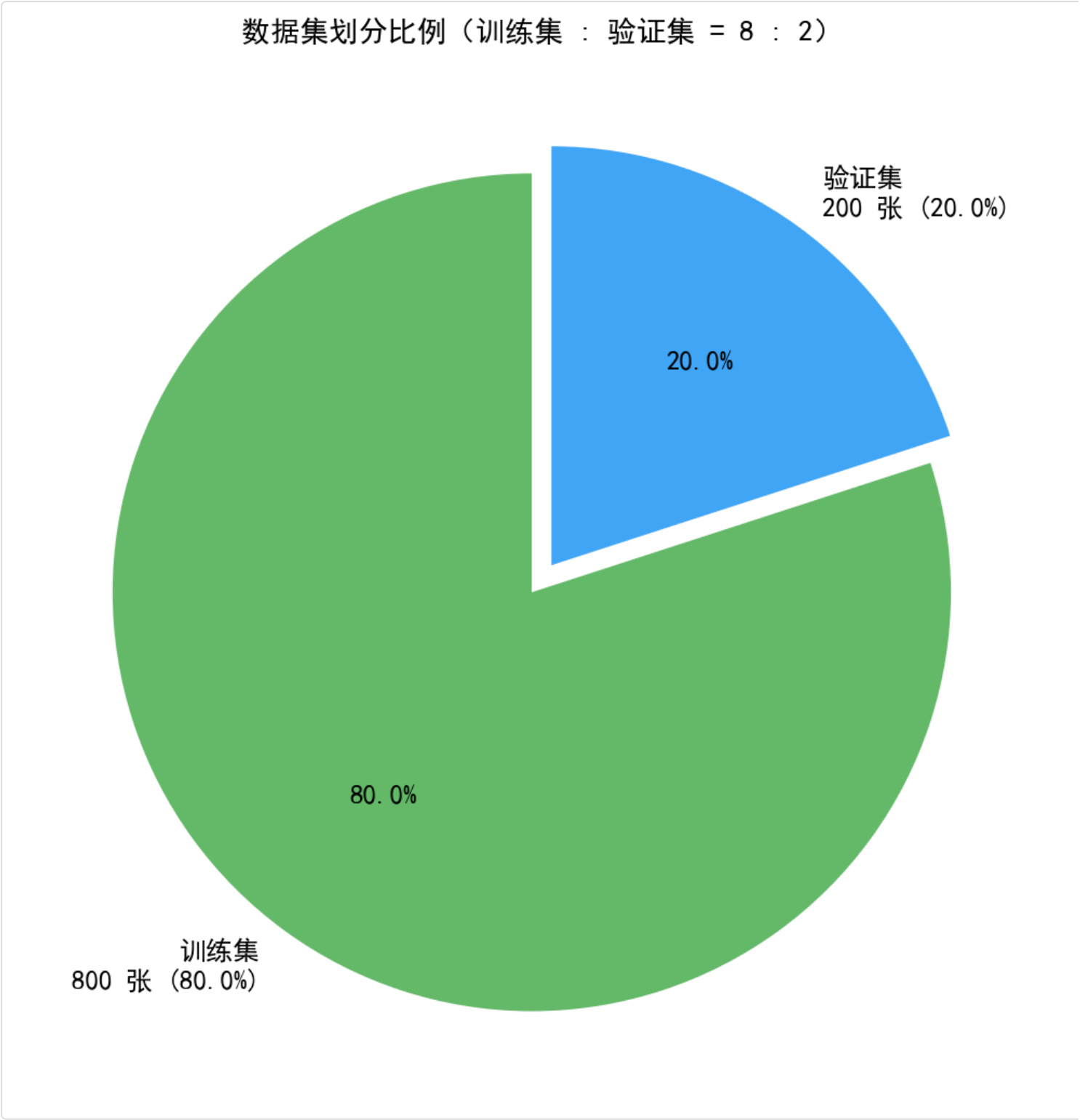
划分方法：采用**间隔采样策略**（Systematic Sampling），将打乱后的样本按每5张为一组，其中前4张划入训练集，第5张划入验证集。这种方法相比纯随机划分能更好地保证训练集和验证集的字符分布一致性，避免因随机波动导致某一类字符过度集中在某个子集中。

划分比例：训练集 : 验证集 = 8 : 2（即 80% : 20%）。此比例是深度学习领域的常用设置，训练集足够大以保证模型充分学习，验证集足够多以提供可靠的性能评估。

4.2 划分统计表

数据集	样本数	占比	字符总数	唯一字符数
训练集（Train）	800	80.0%	2103	31
验证集（Dev）	200	20.0%	515	30
合计	**1000**	**100.0%**	**2618**	**31**

4.3 划分比例可视化



4.4 训练集/验证集分布一致性检验

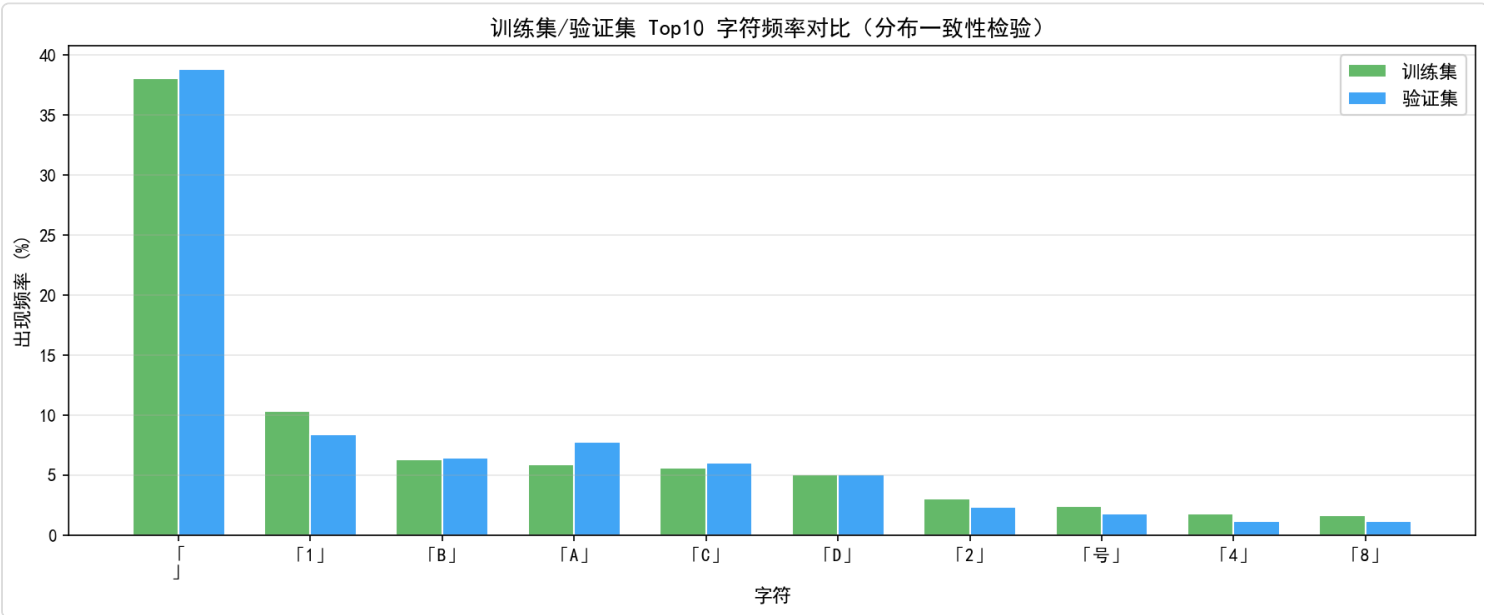
为保证模型训练和评估的公平性，需要验证训练集和验证集在字符分布上的一致性。以下从**字符重叠率**和**频率相似度**两个维度进行检验：

7/10	70%	7
Top10字符重叠数	重叠率	共同字符种类

一致性结论：训练集与验证集的 Top10 高频字符重叠率为 **70%**，重叠率较高，分布基本一致，满足模型训练与评估的基本要求。

重叠字符频率差异分析

字符	训练集频率	验证集频率	绝对差异	评估
「1」	10.32%	8.35%	1.97%	
「A」	5.90%	7.77%	1.87%	
「 」	38.04%	38.83%	0.79%	
「2」	3.04%	2.33%	0.71%	
「C」	5.61%	6.02%	0.41%	
「B」	6.28%	6.41%	0.13%	
「D」	5.04%	5.05%	0.01%	
平均频率差异			**0.84%**	**分布高度一致**

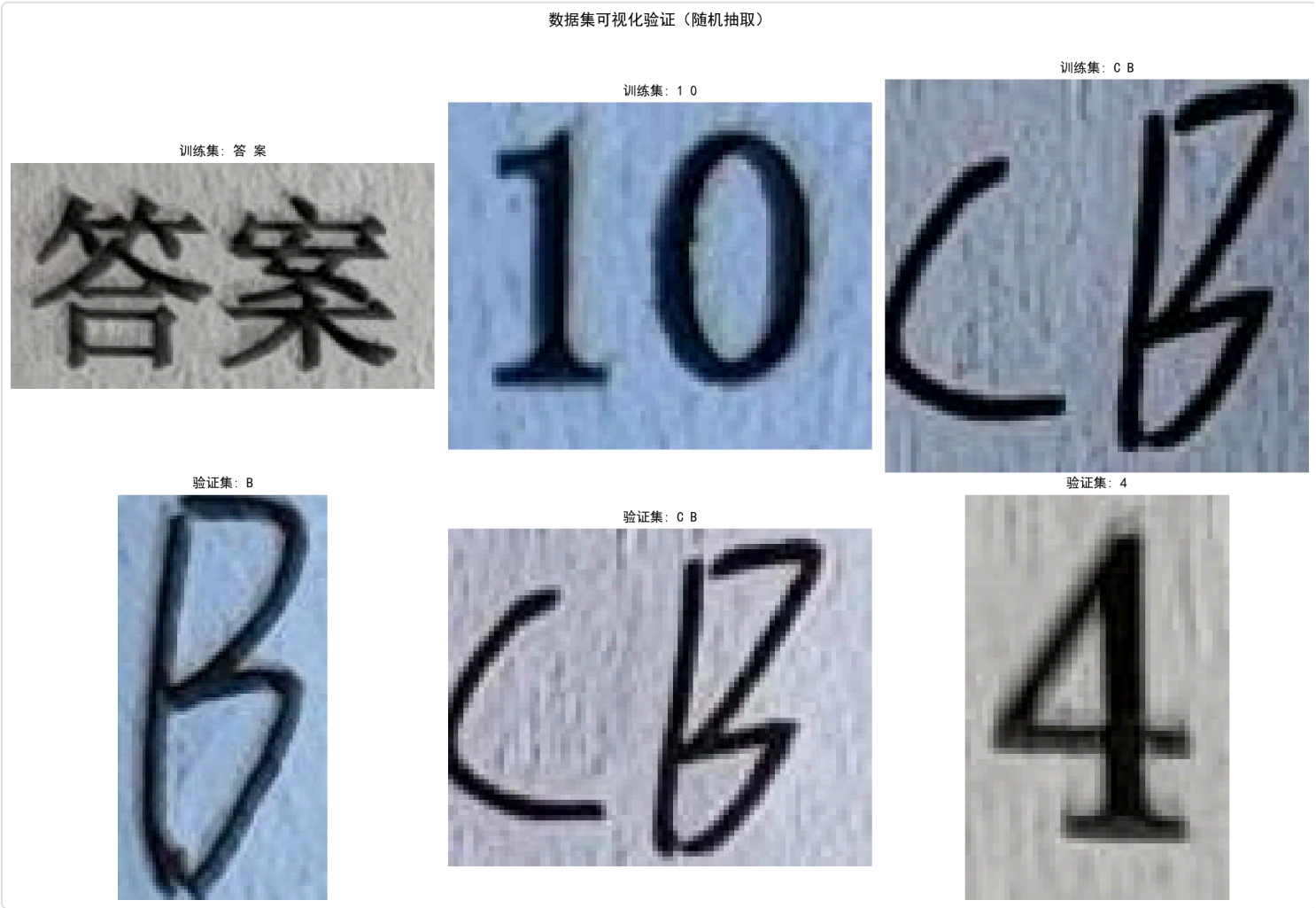


4.5 划分质量分析

评估指标	训练集	验证集	分析结论
样本数量	800 张	200 张	比例约为 8:2，符合预设要求
字符覆盖度	31 个唯一字符	30 个唯一字符	字符覆盖一致
平均每样本字符数	2.6 个	2.6 个	分布均匀

五、数据集可视化验证

以下从训练集和验证集中各随机抽取 3 张样本进行可视化展示。上排为训练集样本，下排为验证集样本。通过人工目视检查，验证 crop 区域是否准确包含目标文字、标签内容是否与图片中文字一致：



验证结论：经多次随机抽样检查，crop 区域均准确覆盖目标文字区域，标签内容与图片中的实际文字一致，未出现错位、漏裁、标签错误等问题。数据集质量满足模型训练要求。

六、实验结论

主要发现

- 通过 6 种图像变换方法，将 20 张原始图片扩增至 140 张，数据量提升 **7.0 倍**
- 数据集包含 1000 个标注样本，标签长度均值 **μ = 1.6** 字符（σ = 0.5），以短文本字段为主
- 字符类型覆盖汉字、数字、字母 5 大类，共 32 个唯一字符
- 训练集/验证集 Top10 字符重叠率 **70%**，分布一致性良好，满足模型训练要求

成果总结

环节	核心操作	产出结果
数据扩增	6种变换 × 20张原图	140 张图片（7.0x 增长）
标签集生成	解析JSON标注 → crop文本区域 → 生成CnOCR格式标签	1000 个样本，32 个唯一字符
数据集划分	间隔采样 8:2 划分，分布一致性检验通过	训练集 800 张 + 验证集 200 张

数据集局限性

1. **样本量有限**：当前数据集规模相对较小，可能影响模型对罕见字符的识别能力；
2. **场景单一**：数据全部来自同一格式的答题卡，模型对其他排版场景的泛化能力有待验证；
3. **手写体多样性不足**：采集的书写样本主要来自少数书写者，手写风格变化范围有限。

后续优化方向

1. **扩充样本量**：收集更多不同书写者、不同拍摄条件下的答题卡数据；
2. **多场景融合**：引入表格、票据、证件等多种文本场景数据，提升模型泛化能力；
3. **半自动标注**：利用预训练模型进行辅助标注，降低人工标注成本；
4. **数据增强策略优化**：引入弹性形变、随机擦除等更适合手写体的增强方法。

报告生成时间：2026-09-29 09:17:42