# **数据预处理梳理及针对多情况大数据集模型变更方案调整**

## **一、预处理清洗流程**

为满足医院方对精度的要求，数据预处理清洗分为以下四步：

### **（一）第一步：排除无盒子照片**

计算各个照片的黑色像素占比，排除高占比照片。  
第一类 NG（worst\_product\_cleaning）：无盒子的情况。  
筛选依据：无盒子时，黑色像素或白色像素占比会异常大，可据此筛选出无盒子的照片。

### **（二）第二步：训练二分类模型识别盖子有无**

使用 OK 集中有盖子和无盖子的图片作为数据，训练一个二分类模型来识别盖子的有无。  
模型选择：采用蒸馏模型，通过调参实现较好效果。  
时间安排：预计 7.10 日期末结束后开始落实。

### **（三）第三步：边缘切割与结构相似性比对**

对新图像进行边缘切割，采用 Sobel 算子提取图像边缘，并与参考 OK 图像进行结构相似性（SSIM）比对，进行多尺度 SSIM 分析（全局 + 局部区块），同时自动生成参考模板（OK 集聚类中心）。  
目的：排除第二类 NG（semi\_product\_cleaning），包括盒子位置是否固定、盖子是否扣好、是否叠加等情况。  
依据：OK 集中所有通过的盒子都在固定位置，而 NG 中存在不固定位置但正确穿线的图片；没扣好、没底层或叠加放置多个盒子时，其边缘切割会有尖锐的地方，可据此筛除。

### **（四）第四步：图片分组与旋转处理**

经过前两步清洗，得到与之前集合类似但区分了盖子有无的图片。  
处理方式：利用第二步的模型分出盖子的有无，有盖子的图片旋转 180 度，使其与之前训练模型的默认识别位置及方向一致。将图片分为有盖子（Capok 组）和无盖子（Capout 组）两大组，由于 OK 集中图片角度偏差肉眼无法观测（<3 度），在数据增强的 transform 中可将角度调整范围适当调小为 ±5 度。

## **二、数据增强与模型训练**

### **（一）数据增强与分割**

分为两大组后，对所有数据进行增强处理。判断 NG 与 OK 的关键依据为线是否规矩、针是否卡住，这需要强调局部及细节的区分。在理想情况下，两组内图片已较为规矩，只需关注针线形状问题。因此，可对不同组的图片进行特定位置分割，将特定位置作为学习对象，以获得更好的效果（此为设想，非项目实际必要实现）。

### **（二）模型训练**

采用孟文瑜学姐的首阶段模型，对 transform 后的图片进行处理：

Capok 组：投喂经处理的图片，训练 Capok 二分类模型，重点关注盖子闭合状态与针线位置关系。

Capout 组：投喂相应图片，训练 Capout 组二分类模型，聚焦针线形态细节（如弯曲度、卡位）。

NG 判断：无针线的情况判为 NG。

具体参考照片见另一文档末尾处。