Dicom数据处理

# 一、文件目录

数据集文件以Data为根目录，Data的子文件夹是医生，医生子文件夹是病人，例如patient1的文件夹路径为：Data/doctor1/patient1

一个病人文件内，包含多张CT切片，以”CT\_XXX”命名；切片目录，以”CT\_XXX.dir”命名；包含标签文件，以”RTSTRUCT\_XXX”命名；还包含生成的勾画图片，存储在此路径下的各个以勾画部位为名的文件夹内。

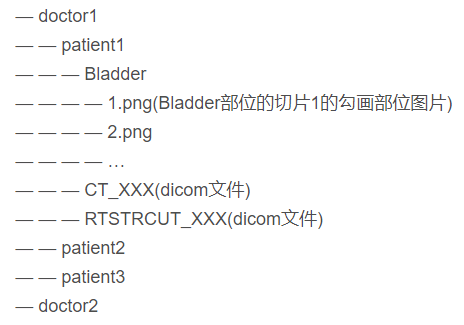
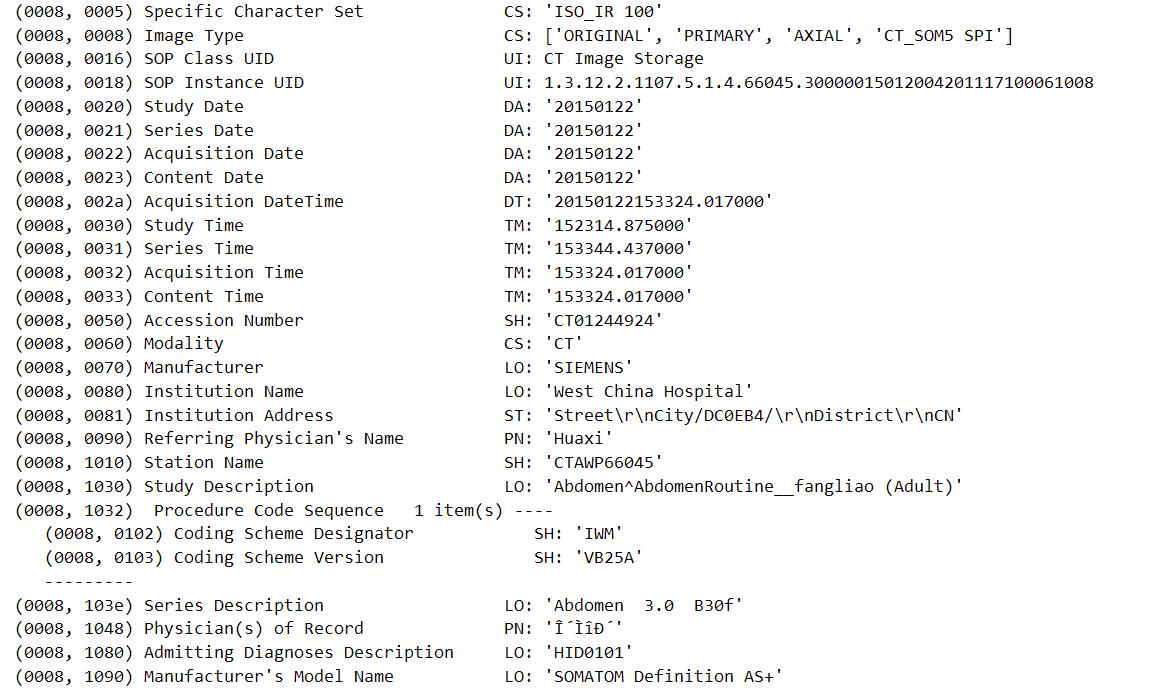
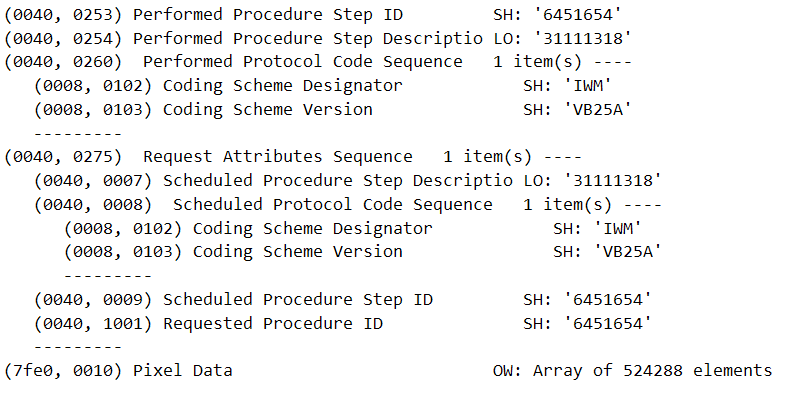


图1 文件目录示意图

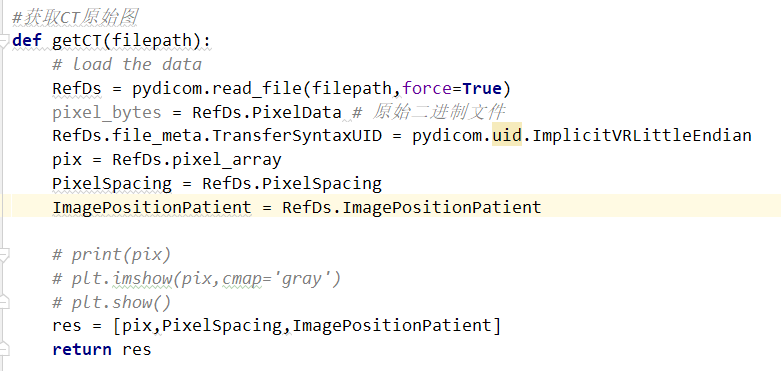
# 二、CT文件

CT文件以”CT\_XXX”命名，通过pydicom读取后，部分信息如下：





Pixel Data存储的就是CT图像，表示一个切片的CT底图，以二进制存储，需要将二进制转化成图像，用如下代码：

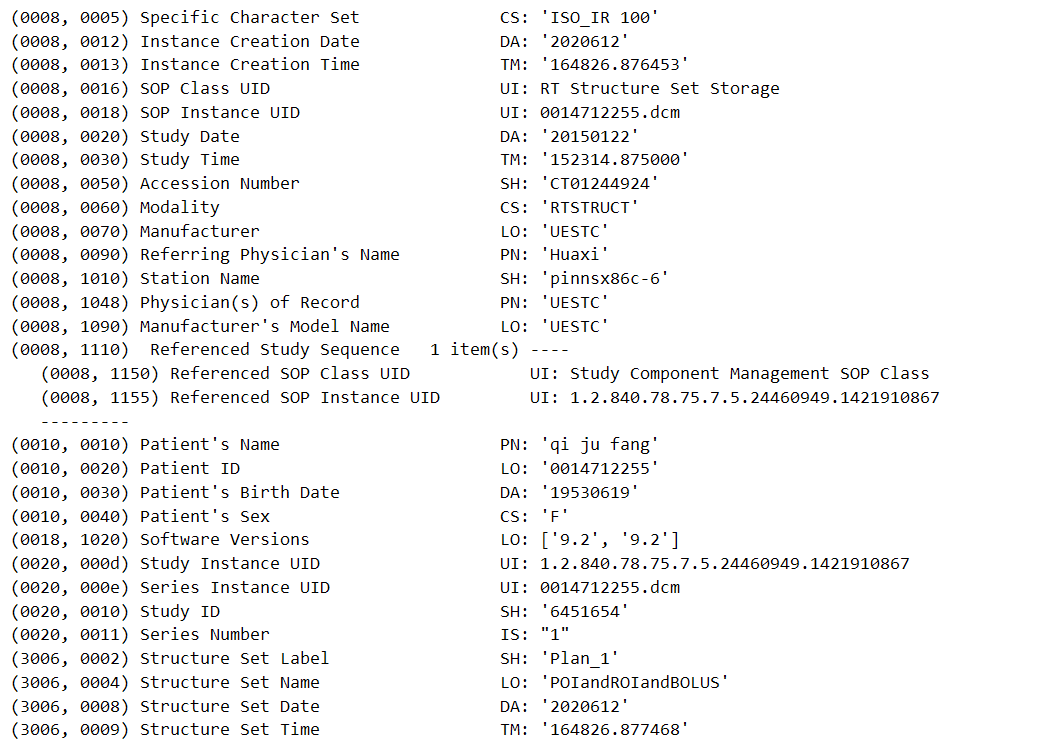


返回的res中的pix是十进制表示的图像信息，需要放在512\*512的.png图像中。

# 三、RTSTRUCT数据的读取

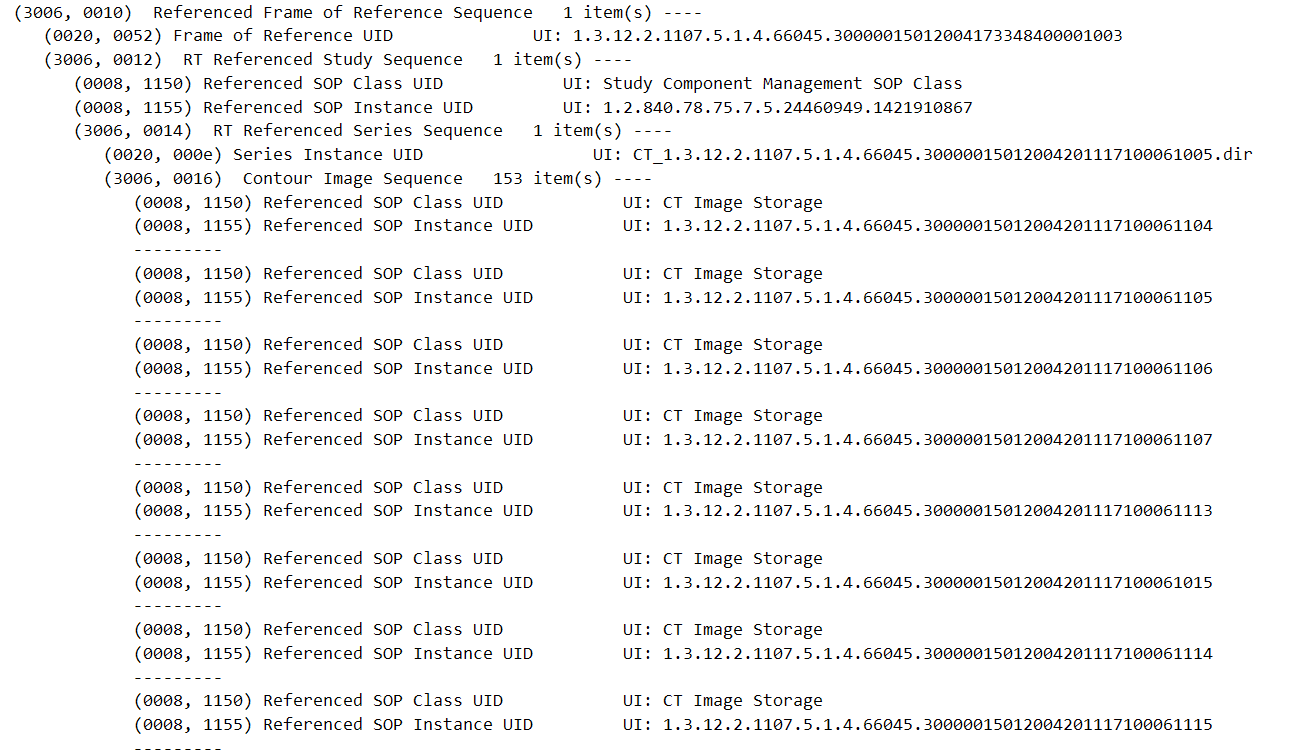
RTSTRUCT文件包含勾画信息、Referenced Frame of Reference Sequence、Structure Set ROI Sequence和ROI Contour Sequence四个部分。

1.勾画信息：勾画信息包含病人的姓名、勾画机器，对应切片名称等



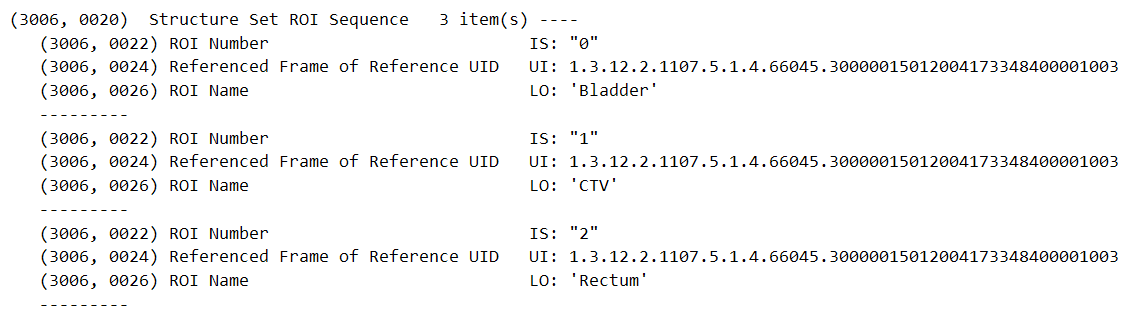
2. Referenced Frame of Reference Sequence

这部分包含所有切片信息，即对应的CT底图名称列表



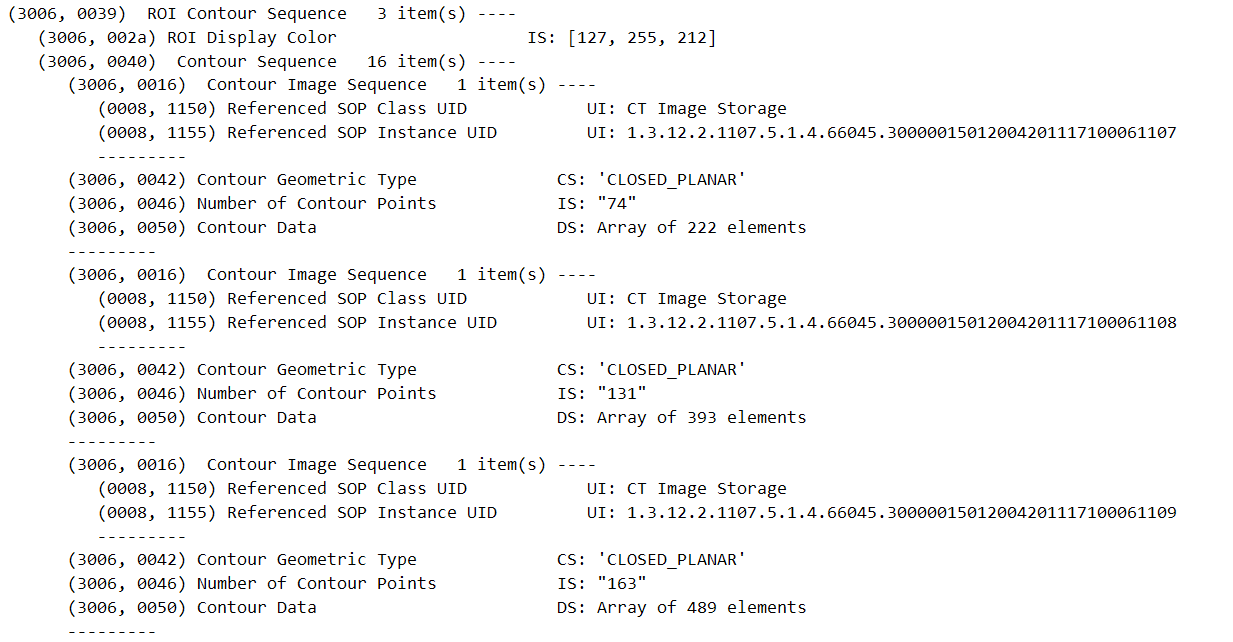
3.Structure Set ROI Sequence

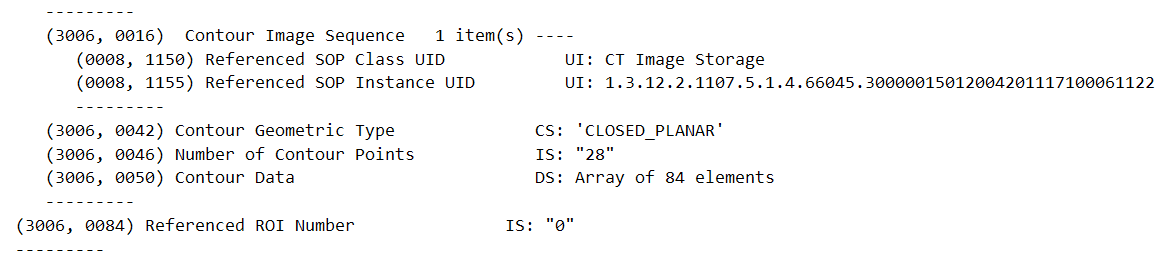
这个部分将部位名称进行编号，与后面ROI Contour Sequence中的Referenced ROI Number对应



4. ROI Contour Sequence

这部分包含勾画的图像信息，存储在Contour Data中。

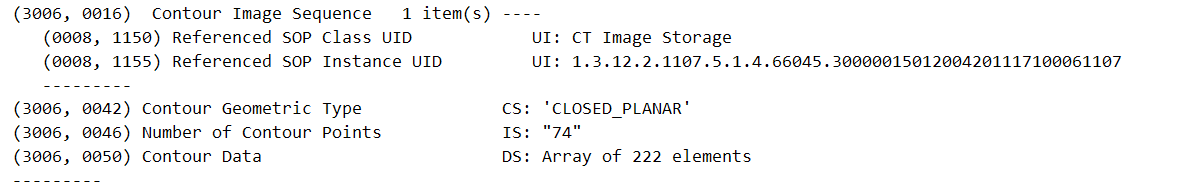




**ROI Contour Sequence 3 item(s)**：表示有3个部位的标注数据

**Contour Sequence 16 item(s)**：表示当前部位有16张切片有标注数据

下图表示一张切片的标注数据：



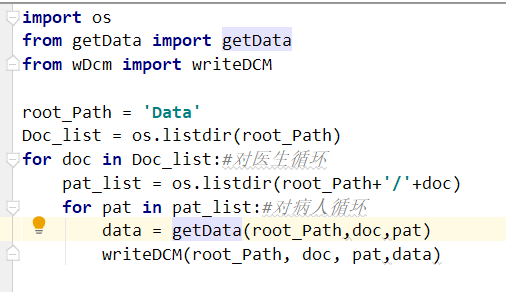
**Contour Data**：存储形式为勾画图像分为多个连通区域，每一个连通区域存储为一个Contour Data，存储的是当前连通区域的轮廓的顺时针世界坐标顺序，每个坐标点包含三个方向的值。

# 四、勾画模型的训练

输入CT图像以及专家勾画的标签图，训练UNet等模型，输出CT图像对应的ROI的预测勾画图像，存为.png。

# 五、RTSTRUCT数据的生成

经过模型预测的数据将会以.png形式存储在每个病人路径下，首先读取png的切片名称、png的勾画数据，勾画部位，医生名称，病人名称信息，并存储为二维形式，然后新建RTSTRUCT\_XXX.dcm文件，根据每一条信息写入RTSTRUCT中。



上图getData()是获得二维形式的数据，writeDCM()是对每个RTSTRUCT进行写入