BTCV

Modality	CT
Organ	
Type	3D
URL	https://www.synapse.org/#!Synapse:syn3193805/wiki/217753
분야	Medical
자료	

Vanderbilt University Medical Center (VUMC)의 CT Scanners로부터 습득한 **Medical Segmenation Dataset**이다.

데이터셋은 Abdomen(복부)과 Cervix(자궁경부)로 나뉘며,
Medical Image Segmentation은 주로 전자를 다룬다.

따라서 **Abdomen Dataset**에 대해 설명하겠다.
참고자료는 **[Link]**를 이용하였고, 데이터 분석 코드는 **[GitHub]**에 올려놓았다.

= Category =

Abdomen은 총 13개의 organ을 지닌다.
(1) spleen

(2) right kidney

(3) left kidney

(4) gallbladder

(5) esophagus

(6) liver

(7) stomach

(8) aorta

(9) inferior vena cava

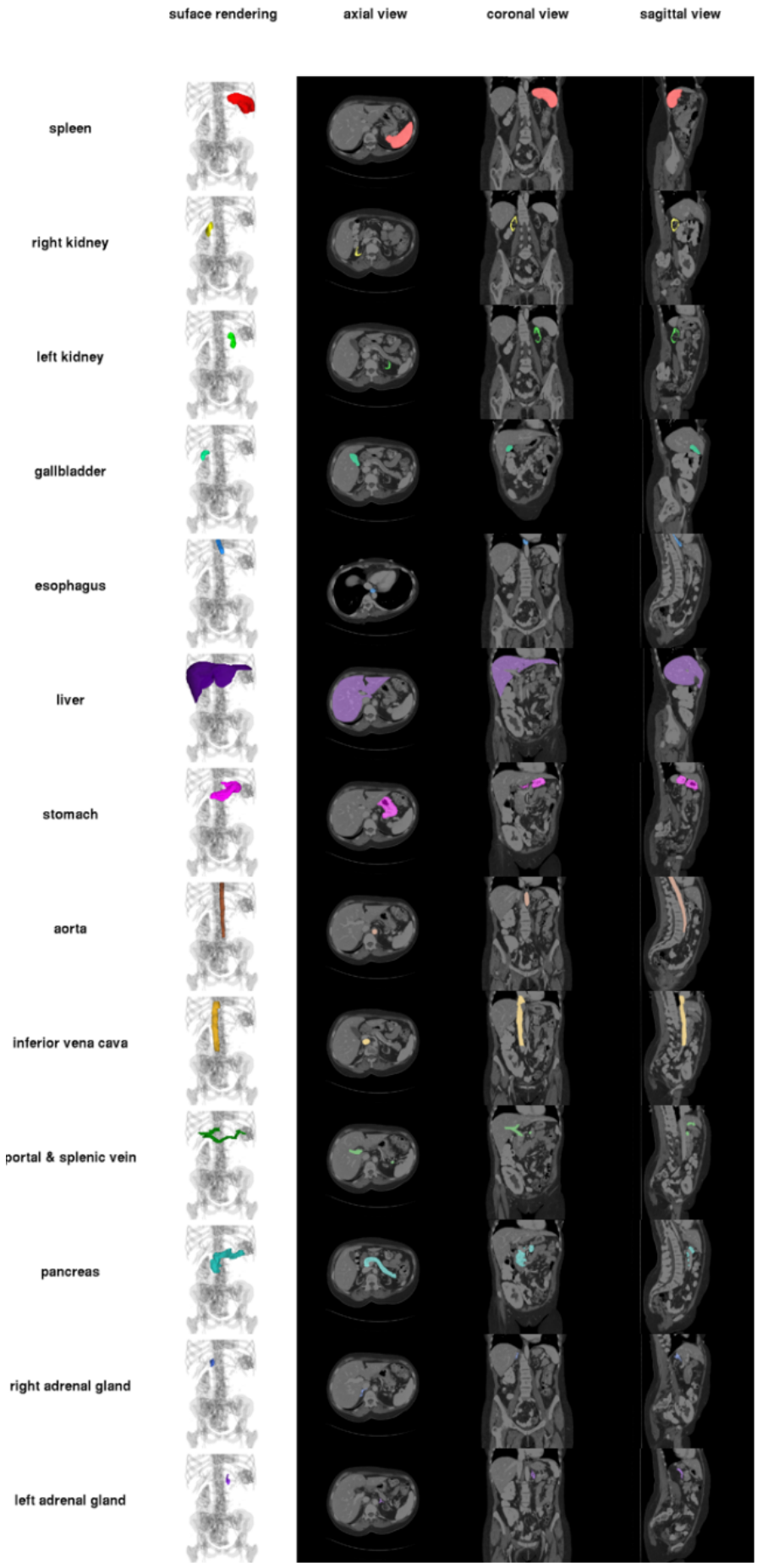
(10) portal vein and splenic vein

(11) pancreas

(12) right adrenal gland

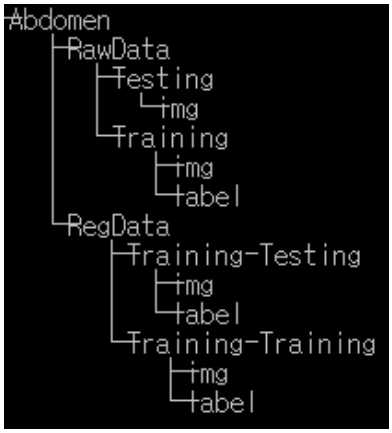
(13) left adrenal gland

다음은 각 organ의 Color Segmentation Map이다.


suface rendering
axial view
coronal view
sagittal view
spleen
right kidney
left kidney
gallbladder
esophagus
liver
stomach
aorta
inferior vena cava
portal & splenic vein
pancreas
right adrenal gland
left adrenal gland

axial view는 몸을 위 or 아래 방향으로 확인한 것이고, **coronal view**는 몸을 앞 or 뒤 방향으로 확인한 것이다. **sagittal view**는 몸을 오른쪽 or 왼쪽 방향으로 확인한 것이다. 데이터셋에는 axial view만 저장되어 있기 때문에 나머지는 고려하지 않아도 된다.

= Dataset Structure =



RawData(1.53 GB)는 Trainining과 Testing으로 이루어진 데이터셋이다.
Trainining/img에는 30명의 환자에 대한 CT 영상이 있고, Testing/img에는 20명의 환자에 대한 CT 영상이 있다.
Training/label에는 30명에 대한 Segmentation Map이 있다.
참고로 Testing은 label이 없기 때문에 **Synapse** 사이트를 이용해서 모델 평가를 해야한다.

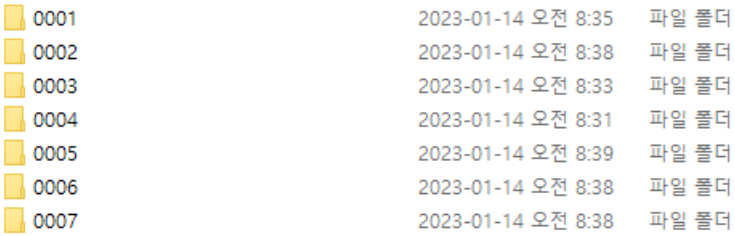
다음은 Trainining/img의 파일 일부를 나타낸 것이다.

img0001.nii	2015-03-25 오전 12:38	ALZip GZ File	43,377KB
img0002.nii	2015-03-25 오전 12:38	ALZip GZ File	38,541KB
img0003.nii	2015-03-25 오전 12:38	ALZip GZ File	50,456KB
img0004.nii	2015-03-25 오전 12:38	ALZip GZ File	39,978KB
img0005.nii	2015-03-25 오전 12:38	ALZip GZ File	30,925KB
img0006.nii	2015-03-25 오전 12:38	ALZip GZ File	28,700KB
img0007.nii	2015-03-25 오전 12:38	ALZip GZ File	48,925KB
img0008.nii	2015-03-25 오전 12:38	ALZip GZ File	35,019KB

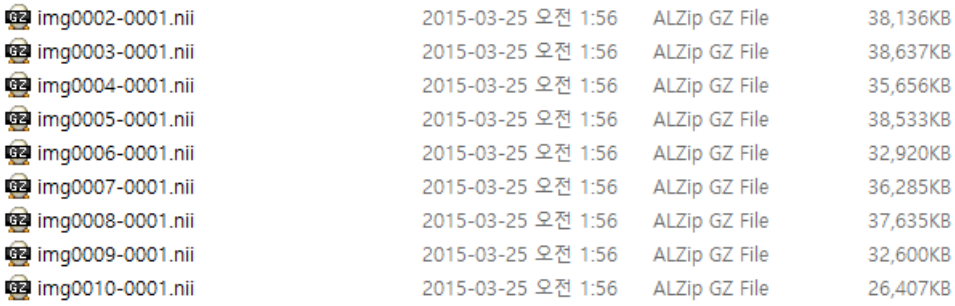
RegData(37.3 GB)는 Training-Testing과 Training-Training으로 이루어진 데이터셋이다.
Training-Testing은 test data에 train data를 등록시킨 것이다. 각 테스트 환자마다 30개(훈련 환자의 수)의 CT 영상을 지니므로, 총 30x20(600)개의 파일이 존재한다.
Training-Training은 train data에 train data를 등록시킨 것이다. 각 훈련 환자마다 29개(자신을 제외한 훈련 환자의 수)의 CT 영상을 지니므로, 총 29x30(870)개의 파일이 존재한다.

(Registered data are categorized into two parts, (1) the pair-wise registrations between training datasets (30 x 29 = 870), and (2) the registrations from the training data to the testing data (30 x 20 = 600))

다음은 Training-Training/img의 디렉토리 일부를 나타낸 것이다.



다음은 Training-Training/img/0001의 파일 일부를 나타낸 것이다.



img0002-0001.nii.gz는 Training에서 두 번째 환자를 Training에서 첫 번째 환자에 등록시킨 것이다.
(In the example case above, subject 0002 in the training datasets was registered to subject 0001 in the training datasets)

어떻게 등록시켰는지는 아직까지 잘 모르겠다. **Synapse** 사이트에도 설명이 미약하다.
다만, RegData는 학습 및 테스트 시 이용하지 않기 때문에 몰라도 된다고 생각한다.

= File Type =

CT 영상은 기본적으로 MRI와 같이 nii.gz이다.
nii.gz는 `nibabel` 모듈을 통해 얻고 출력할 수 있다.

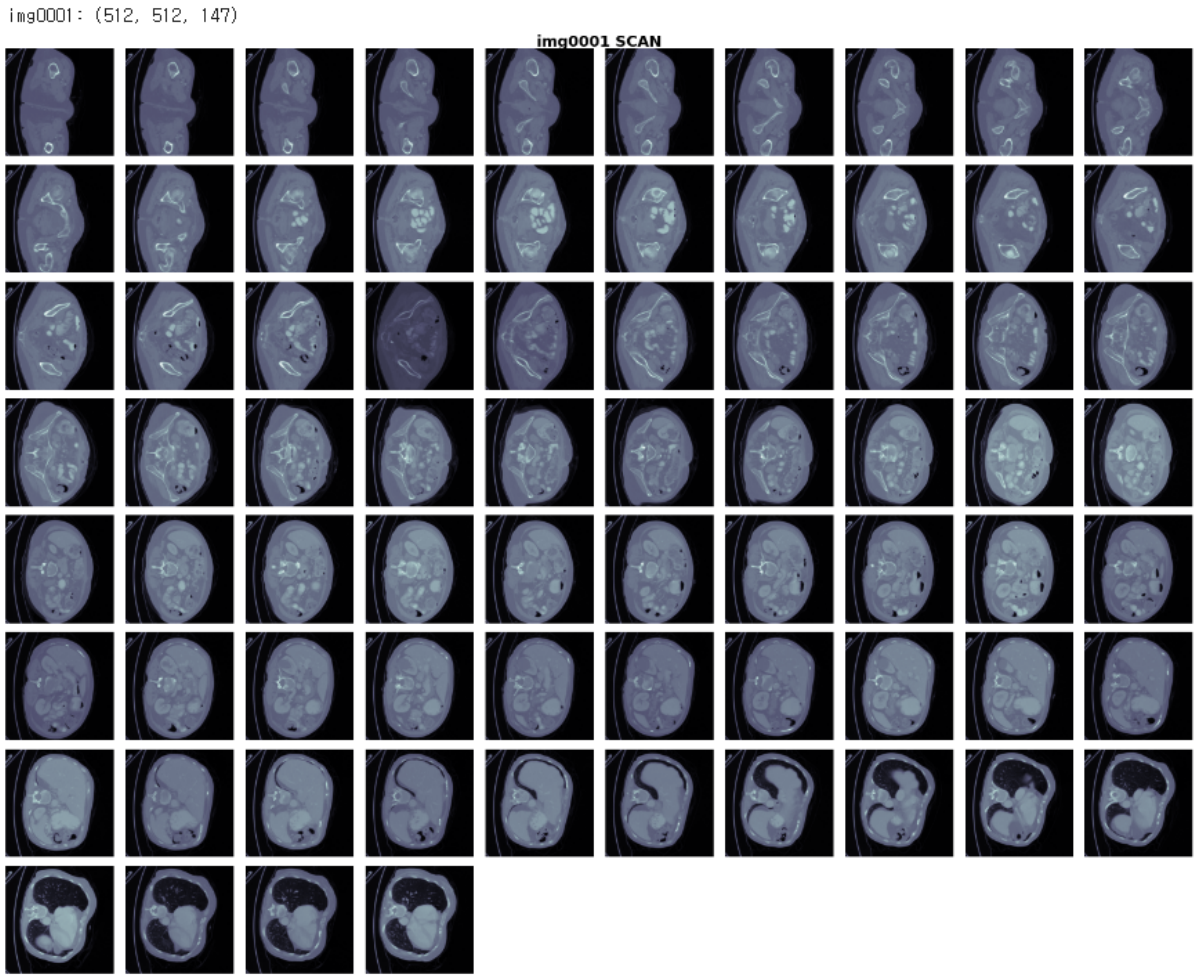
```
path='/media/ahnsunghyun/HDD/dataset/BTCV/Abdomen/Abdomen'
id='0001'
data_path=os.path.join(path,'RawData','Training','img','img'+id+'.nii.gz')
image=nib.load(data_path).get_fdata() # get
ct=image[:,:,118]
plt.imshow(ct) # print
```

= Data Shape =

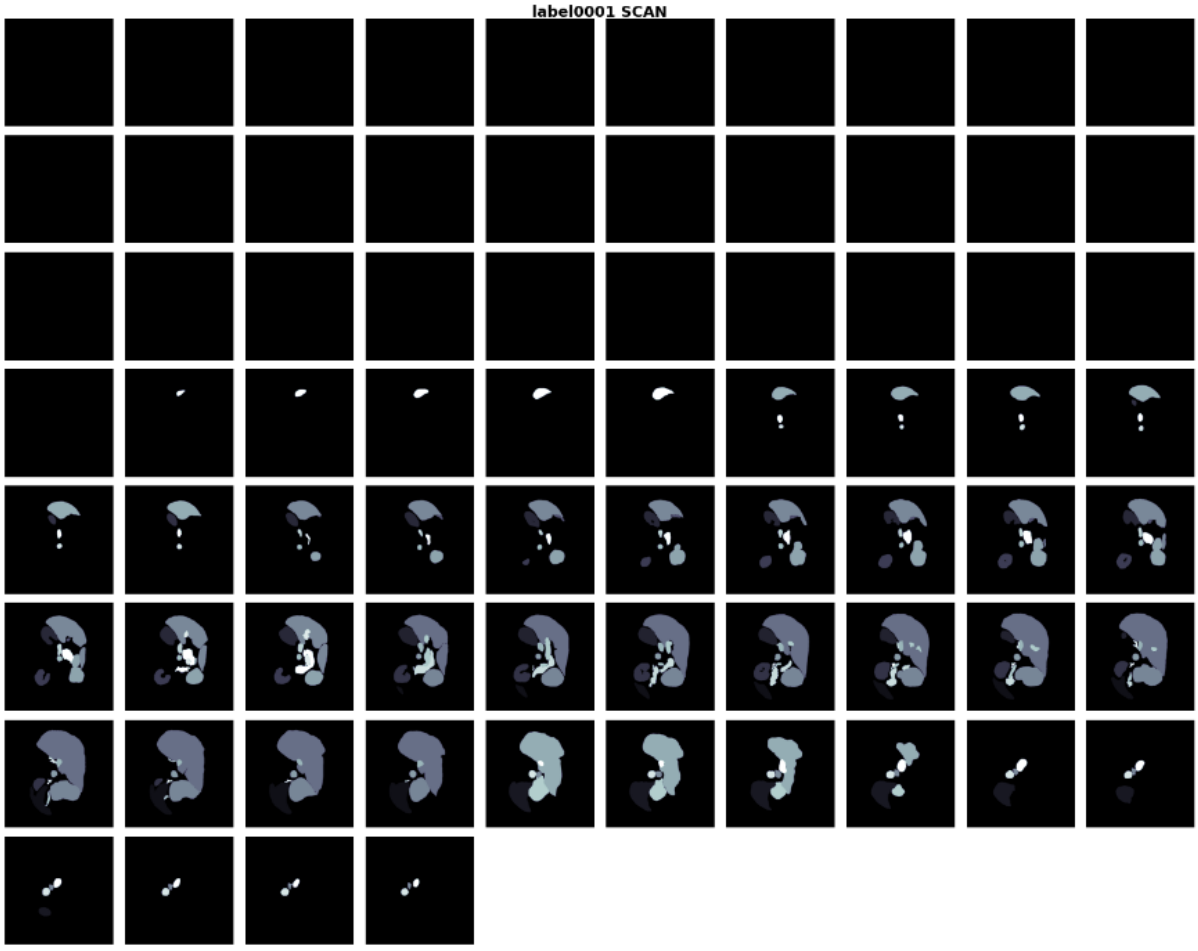
3D Data이기 때문에 data의 shape은 (height x width x slice)이다. 이 때, slice는 환자마다 다를 수 있다고 한다. (85~198)
(The 50 scans were captured during portal venous contrast phase with variable volume sizes (512 x 512 x 85 - 512 x 512 x 198).)

또한 slice에는 모든 organ이 포함될 수도 있고 포함되지 않을 수도 있다.

다음은 1번 환자의 CT 영상(img0001.nil.gz)을 2D로 출력한 것이다. 1번 환자 데이터의 Shape은 (512,512,147)이다.
출력할 때는 slice를 한 개씩 건너뛰면서 출력했다. —> 74장 출력

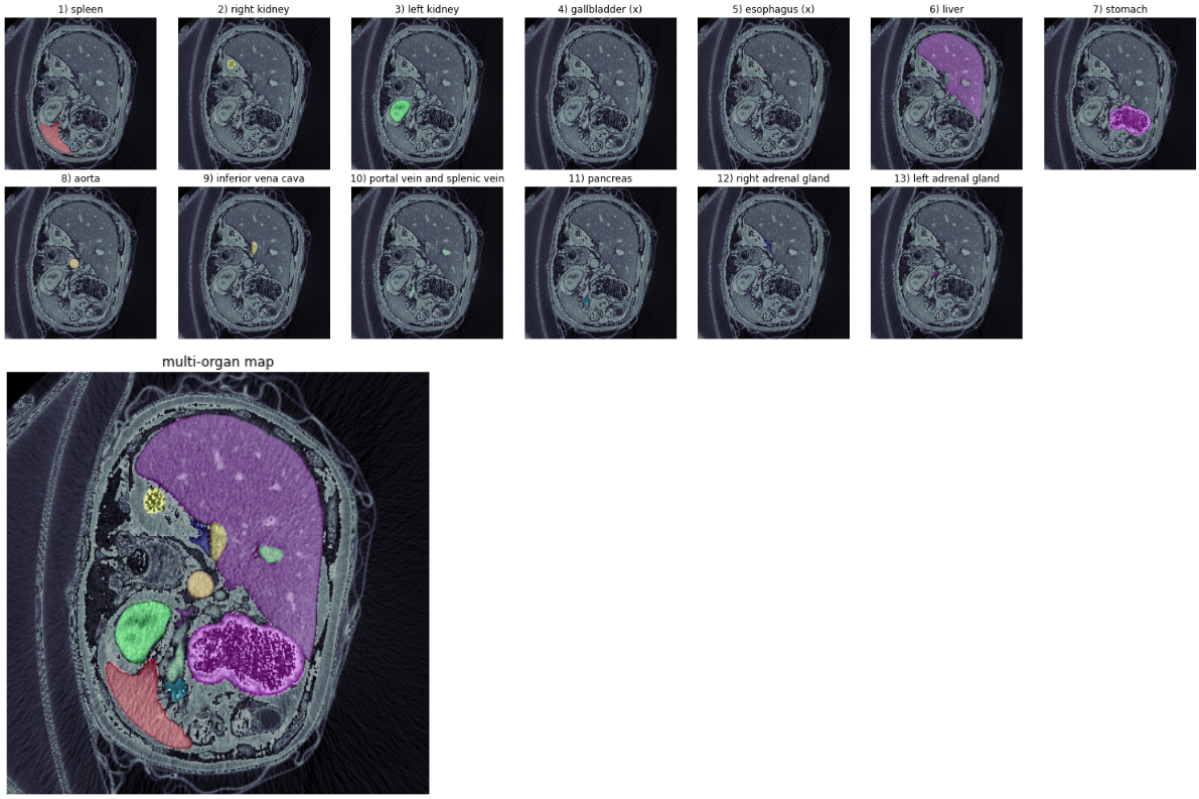


다음은 1번 환자에 대한 정답 영상(label0001.nil.gz)을 2D로 출력한 것이다. label의 shape은 img의 shape과 같다.
참고로 2D 출력에서 각 픽셀은 0과 13 사이의 수(카테고리 수)로 저장된다.



= Color Segmentation Map =

label 영상에 적힌 숫자를 보고 sigle-organ과 multi-organ에 대해 직접 Color Segmentation Map을 만들어봤다.
1번 환자의 118번째 slice를 이용하였다.
참고로 (x) 표시는 해당 organ이 존재하지 않는다는 의미이다.
코드는 **[GitHub]**를 확인하면 된다.



= Leaderboard =

Synapse 사이트의 **leaderboard**에는 Testing에 대한 모델 평가 결과가 적혀 있다.
모델 평가 지표는 DICE, Mean Surface Distance(MSD), Hausdorff Distance(HD)를 이용한다.

leaderboard에 파일을 올리면 single-organ에 대한 평가도 수행할 수 있다고 한다.
이 내용은 **How to Participate**를 참고하면 된다.

(**PLEASE NOTE** We are also scoring leaderboards by label as well (rather than just by mean of all labels). Please make sure to select all the appropriate challenges. If you use the python code, this is all handled for you automatically.)

다음은 laderboard 페이지의 일부이다.

Abdomen Leaderboard

Standard Registration

ID	name	entity	status	team	user+ID	Dice	Mean+Surface+Distance	Hausdorff+Distance
9729220	DAE	syn45208328	SCORED	lowLevel	@sallyY	0.9213	0.65477	16.8232
9728831	disrupted_autoencoder	syn44009192	SCORED		@sallyY	0.92123	1.493	28.2007
9728743	Disrupted_AutoEncoders	syn44009192	SCORED	lowLevel	@sallyY	0.92123	1.493	28.2007
9727802	SHA_TIME	syn34344204	SCORED		@20020135	0.91874	0.8104	20.5334

Spleen - Standard Competition

ID	name	entity	status	team	user+ID	Dice	Mean+Surface+Distance	Hausdorff+Distance
9728744	Disrupted_AutoEncoders	syn44009192	SCORED	lowLevel	@sallyY	0.98021	0.21688	18.8651
9727081	submit_test2	syn34344204	SCORED		@cjhsmile	0.97568	0.26151	19.1774
9727061		syn34344204	SCORED		@cjhsmile	0.97568	0.26151	19.1774
9723809	'first_test'	syn34344204	SCORED	Bai-Team	@baiqqq	0.97568	0.26151	19.1774