# SELF-SUPERVISED REPRESENTATION LEARNING FOR MEDICAL IMAGE PROCESSING

**Manish Bhardwaj**
Newcastle University
UK
m.bhardwaj2@newcastle.ac.uk

**Huizhi Liang**
Newcastle University
UK
huizhi.liang@newcastle.ac.uk

**Varun Ojha**
Newcastle University
UK
varun.ojha@newcastle.ac.uk

**ABSTRACT**

Self-Supervised learning is a promising paradigm of machine learning that assures to resolve the issue of huge "labelled data-set" requirement by the supervised learning algorithms. In order to explore this frame of reference, we evaluated various self-supervised learning and supervised learning networks that are prominent in the field of medical images segmentation. The medical image data-sets used were abdomen ct scans with 15 organs , MRI scans for spleen and Thyroid nodule ultrasound images. To improvise the data-sets, various data pre-processing techniques were implemented like re-orientation, resampling, standardization, windowing, channels addition, resizing and intensity scaling. The data was also augmented using random flips, random crops and padding. The data pre-processing of the medical imaging ct/mri scans was executed by utilizing specific libraries like Nibable or Monai so that the transformations maintain the affine matrix relationships (physical coordinates to voxel coordinates mapping) while training. The distinguished supervised learning models in the area of medical representation that we evaluated are - nnU-net 2D, nnU-net 3D, Unet. The machine learning networks picked for self-supervised learning are SMIT, SAM and BTUnet. These self-supervised learning networks were chosen because of their unique architectures and applications. After the completion of our evaluation, We have established that the self-supervised learning network requires much lesser labelled data as compared to supervised learning networks. Also, the self-supervised learning networks and can perform almost as good as, and in some cases can even surpass the performance compared to the supervised learning networks . One of the evaluations provide 81% accuracy with just 10% of labelled data in self-supervised algorithm against the 76% accuracy in supervised algorithm that used the same complete labelled data-set for training4. This substantiate there is huge potential to obtain greater success by applying the self-supervised learning in the domain of medical image segmentation. Self-Supervised learning fulfils the promise to reduce costs related to annotations for the complex medical images, and also reduces the training time thereby fostering further research in the field of medical image segmentation.

## 1 Introduction

One of the main struggle of implementing deep learning techniques in the area of medical image segmentation is the lack of labelled data for training the models/networks. In case of medical images like CT scans or MRI or Ultrasounds, a specialized skilled person is required to identify and annotate various organs in the scans and to interpret if there is any medical issue being exhibited in a particular medical image. This process of annotating the medical images consumes a lot of time of the medical professionals/experts and hence the cost of annotations is quite high specifically in the medical domain. This also leads to lack of training data that have corresponding ground truth labels, thereby inhibiting further advances of the application of deep learning in the medical domain.

"Supervised learning" is very successful technique for image segmentation of natural images like car, tree, building, animals etc. There is plenty of training data available for these natural images. However, advancement of "Supervised learning" in the area of medical field has been difficult because of the lack of training data.

Hence there has been an urgent need to recognize a deep learning methodology that utilizes very less labelled data but still can produce excellent results in image segmentation. The deep learning technique that can fill this gap is "Self-Supervised Learning".[1]

The concept of the self-supervised learning emerged by comparing human learning to machine learning. As mentioned by Yann LeCun [2] – “How is it possible for an adolescent to learn to drive a car in about 20 hours of practice? By contrast, to be reliable, current ML systems need to be trained with very large numbers of trials." For example, human would be able to categories if a particular creature is animal or not (may not know the name of the animal, but still can categorize). Over the years, human have learnt various features that constitutes an animal i.e. humans have ability to learn concepts at higher levels of abstraction that seem to develop on top of lower-level ones. This is termed as common sense. “Common sense can be seen as a collection of models of the world that can tell an agent what is likely, what is plausible, and what is impossible. Using such world models, animals can learn new skills with very few trials.”

This is the basis of "self-supervised learning" concept.

The essential task is to discover if "self-Supervised learning" is an appropriate technique for medical image segmentation? If so, it will reduce the requirement of high amount of training data for learning of these models. This would mean reduction in the cost for producing annotations of the medical images as the corresponding time required by the domain specialists/ experts would be considerably minimized. This investigation will provide the direction in which to tread ahead further to foster more advancements to apply deep learning in the field of medical image segmentation.

In order to proceed with further analysis, we picked some established supervised learning models in the field of medical domain and compared those with the selected self-supervised models. The networks were selected for evaluation based on their current performance and applications. A lot of existing research work was studied to select the networks to evaluate [2] [1] [3] [4] [5] [6] [7] [8].

As generally the medical images will contain both 3D and 2D data, hence the data-sets selected should also contain 2D and 3D data i.e. images (2D), CT scans (3D) and MRI scans (3D). Hence we picked Abdomen CT scan data-set [9] (collection of CT scans that contain 15 organs from abdomen), Spleen MRI data-set[10] and Thyroid nodule image data-set[11]. The abdomen CT scan data[9] is more complex as it contains 15 organs from the abdomen. Selection of these data-sets, fulfills our requirement to test the most general data available in the medical domain.

The "Supervised learning" models that we selected for evaluation were Unet[3], nnU-net 2D[12] and nnU-net 3D[12] as these are very widely used in the medical domain. The abdomen CT scan data-set[9] was used to evaluate Unet and nnU-net 3D, and thyroid nodule data-set[11] was used to measure performance of nnU-net 2D.

The Self-supervised models were chosen based on specific characteristics of their architecture and their applications - BTUnet[13], Segment Anything[14], SMIT[15]. only 10% to 20% of thyroid data-set [11] was used to evaluate BTUnet model and the Segment Anything model. furthermore, only 10% of the The abdomen CT scan data[9] and Spleen MRI data-set[10] was used to evaluate SMIT network.

Our major contibution to this project include:

- Extensive study to understand and research the existing work so that we can select appropriate supervised and self-supervised networks for experiments.
- Conducted research of medical images and identify appropriate image data to use.
- Pre-processing of the 3D volume scans and 2D images suitably to be used for training.
- Updated original implementation codes of the published papers of the selected model to make these appropriate to run and produce results.
- Conducted the experiments utilizing the above researched items (Networks, papers, codes, image data-sets, methodologies)
- Comparison, correlation and analysis of the results obtained to seek limitations and further advancements areas.

## 2 Related Work

There has been considerable work done in the domain of segmentation of medical images using supervised learning [3], [12], [16], [4], [5]. The self-supervised learning can take it to the next level, where from the past conceptual understanding the artificial intelligence can be utilized to segment any organ and to identify any disease/anomaly.There is also substantial effort done on self-supervised learning [17], [6], [7], [1], [8]

Based on the study and the understanding of the existing studies, following choices were made.

The "Supervised learning" models that we adopted for evaluation were Unet[3], nnU-net 2D[12] and nnU-net 3D[12] as these are very widely used in the medical domain. Like any supervised model, these networks require lot of training data for learning.

Unet is the most widely used supervised learning network in the medical domain area for image segmentation [18] [19]. Hence it was selected to be evaluated.

Because of its success, there were further modifications done in Unet model's architecture to enhance its performance. This new supervised network is nnU-[12]. Additional cascaded Unet was added to make segmentation two staged process. The loss was modified to include both dice and cross entropy. And batch normalization was replaced by the instance normalization. it has two variants in two dimensions and 3 dimensions. As cited[12], nnU-net has significantly improved performance on the medical image segmentation. This prompted to add nnU-net in our list of evaluation.

The following Self-supervised models were chosen based on specific characteristics of their architecture and their applications - BTUnet[13], Segment Anything[14], SMIT[15].

Described below is the summary of approach for selecting these networks.

BTUnet[13] was chosen because of its unique architecture of Barlow Twin loss. It overcomes the challenge of requirement of positive and negative samples for training, because all the samples are related to each other in medical domain images. it was evaluated on the Thyroid nodule data-set.

SAM[14] is a recent effective model by Meta for segmentation of natural images. We fine-tuned it for medical images to gauge its effectiveness and further usage in medical domain. The Thyroid nodule data-set was used determine its effectiveness.

SMIT[15] used self-distillation technique of teacher and student networks being trained simultaneously. This is based on Vision transformers which are very effective in image segmentation in general.

## 3 Methodology

The aim of the project is to explore self-supervised learning in the domain of medical image processing. This would entail image segmentation and classification of medical domain images by implementing various deep learning techniques and methodologies. Various types of medical images like CT scans or MRI or Ultrasound images would be utilized for this purpose with the pre-processing of these images in 2-dimensions and 3-dimensional space [20].

The methodology followed was as below:

- Medical Image Processing
  - Identify how to process CT scans/ MRI/ Ultrasound images to utilize for deep learning.
  - Processing of these images in 2D and 3D (to enable distinguish deep learning techniques efficacy in 2D and 3D)
  - Enable easy visualization of the input and output medical images.
- Deep Learning Techniques
  - Implement various deep learning methodologies and algorithms for supervised image segmentation and classification on medical images.
  - Compare the different predictions from these models.
  - Analyse various implemented approaches.
- Self-Supervised Learning
  - Implement algorithms for self-supervised learning.
  - Analyse the results.

The methodology is as depicted in our machine learning pipeline figure below (Figure1).

Medical images would generally also contain third dimensions; hence it is important to consider both 2d and 3d images for this project i.e., inclusion of CT / MRI scans and ultrasound images. Each CT/MRI scan is series of cross-sectional single channel images of shapes inside the body i.e. organs, blood vessels, bones.

The medical images need to be processed very differently from the other images. The CT scans and MRI are very specific format of images (e.g., nifty format) that contain various slices/images that are collated together to form the three-dimensional structure of the area of the scan. This provides a view that enables visualization of the organs from

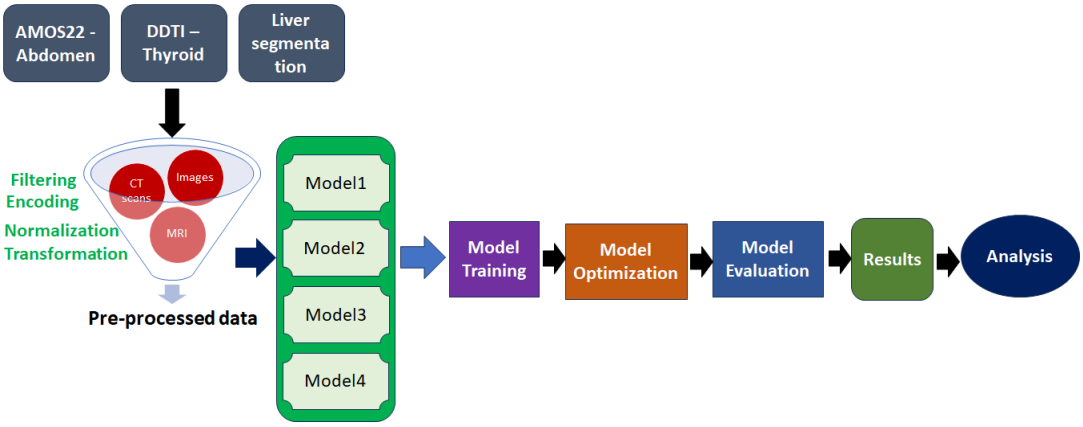



Figure 1: Machine Learning Pipeline

all angles and hence any issue can be pinpointed. There are libraries like "Nibabel" to process these files. As these images can be produced in 3 dimensions, hence there are methods to generate animation out of the input and output images along with masks.

One of the challenges with medical images is that any data-set is available in limited quantities, and these are in different formats, hence pre-processing of the medical images is a very important task.

### 3.1 Pre-Processing

Each CT scan / MRI scan or ultrasound may have different properties, which necessitates that data is pre-processed before it is used for the training of models/Networks. The most important pre-processing used are listed below:

- Orientation:

  For CT scans and MRI, this is the orientation of the scans. The data array axis of all scans should always correspond to the same physical dimensions. Otherwise, the machine learning networks will not be able to learn appropriately and will output bad results. There are three planes of orientation as below. The indices change depending on the orientation of the scan. Sometimes the first axis slices are coronal and sometimes it could be axial or sagittal.

  - Coronal – from anterior to posterior or vice versa.
  - Axial – from inferior to superior or vice versa.
  - Sagittal – from left to right or vice versa.

- Resampling:

  Reduce the size of the volume. Standardize physical resolution and size. The physical coordinates are mapped to voxel and are related by affine matrix. Hence any resizing or resampling has to be done in a way that this relationship stays intact. It allows scaling, rotation, fliping, padding and shearing/cropping.

- Standardization:

  The values in CT scan are based on the Hounsfield units which are practically between -1024 to 3071, depending upon density of region of interest. For example, fat in the liver of density of the bone. Hence for ct scans Standardization is done by dividing by 3071.

- Windowing:

  This is done to get different contrasts in the medical image depending on the task to be performed. The values of the matrix data between certain ranges can be clipped to highlight other structures within the body.

### 3.2 Supervised learning Networks

We have assessed below mentioned supervised learning models, that are prominent in the area of medical image segmentation. We have used both two dimensional and three dimensional data for evaluation on these networks. The publicly available data-sets were utilized as referred in the Introduction section 1.

### 3.2.1 U-net:

U-net is the most prominently used network for the medical image segmentation. [16], In this work by Ronneberger et al, it was showcased that unet has been able to achieve good results on 2D microscopic images. Hence as next step U-net3D model was tried in this dissertation.

This network uses encoder-decoder design where encoder (the contracting path at left side) is used to extract features, and the decoder (the expansive path at right side) consists of up-sampling of the feature-maps. The contracting path layers are encoder layers are concatenated with the corresponding decoder block using the skip connections for feature up-sampling. The architecture of the Unet network is provided in below figure 2

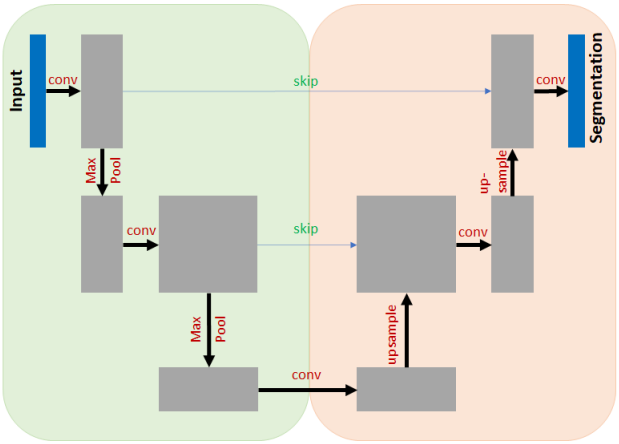



Figure 2: UNet architecture diagram [16]

U-net was executed on amos22 data-set, that consist of 3D scans including 15 organs and corresponding masks. The entire dataset of 600 CT scan was utilized. The results were not very encouraging.

### 3.2.2 3D nnU-net:

nnU-net [12] refers to no-new-net, which is a framework based on 3D U-net [16]. This includes many non-architectural changes that showcase better results. nnU-net consists of adaptive pre-processing and training scheme. nnU-Net automatically runs a five-fold cross-validation for three different U-Net networks and the model with the highest mean foreground dice score is finally chosen automatically. The three models are (i) 2D unet and conversion of 3D volume in to 2D slices, (ii) 3D unet whereby 3D convolution is used directly, (iii) cascaded Unet networks as explained below.

For 3D CT scan large data-sets, the problem is that the training is done on image patches because of limitation of infrastructure capabilities. This inhibits collection of sufficient contextual knowledge like correctly differentiate the part of one organ from another organ. Hence nnU-net contains an additional cascaded 3D Unet network.

Unet is first trained on a down sampled images (1st stage), the segmentation results are then up-sampled to original voxel space and forwarded as additional input channel to the to second 3D Unet network (2nd stage) that is trained on patches at full resolution.

Some of the differences to Unet are as below:

- Network is trained on combination of dice and cross entropy loss.
- Leaky Relu has been used instead of Relu.
- Use of instance normalization instead of batch normalization
- To enable network to appropriately learn spatial semantics, resampling of scans was done to the median voxel spacing of their respective data-set.

The figure3 below describes the nnU-net architecture.

The entire data-set of 600 CT scan was utilized. Satisfactory results were obtained when nnU-net was run on the 3D amoss ct scan data-set. The inspiration of code is from original implementation of paper[12] by Helmholtz Imaging and German Cancer Research Center (DKFZ).

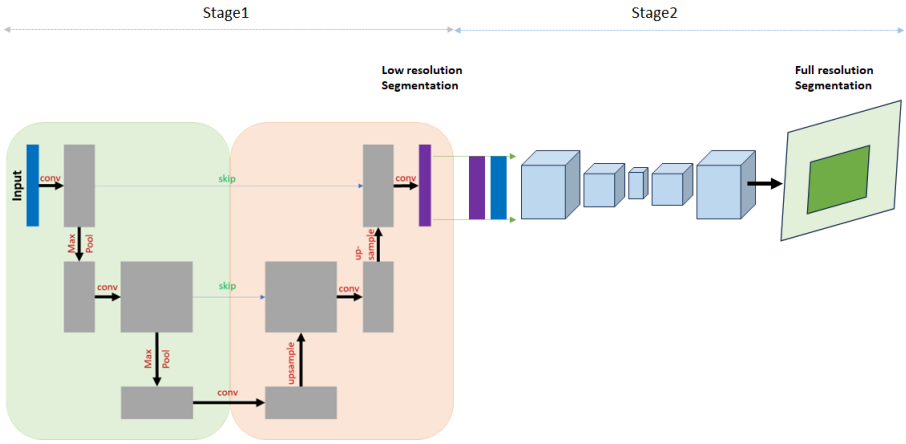



Figure 3: nnU-net architecture diagram [12]

### 3.2.3 2D nnU-net:

The 2D nnU-net also uses the stage1 and stage2 as in 3D nnU-net 3.2.2. Additionally, pre-processing is done as input patch size of 256x256, a batch size of 42 and 30 feature maps in the highest layers (number of feature maps doubles with each down sampling). This network was executed on the DDTI data-set ( Thyroid nodule ultrasound images). The entire dataset of 420 images utilized, and the results obtained were also competitive.

## 3.3 Self-Supervised learning Networks

We have assessed the following self-supervised learning models. We have used both two dimensional and three dimensional data for evaluation on these networks. We have used the publicly available data-sets as referred in the Introduction section 1. However the amount of data used to train the self-supervised network varies between 10% to 15% of what was used to train the supervised networks.

### 3.3.1 BTUnet:

BT-Unet [13] utilizes the Barlow Twins approach to pre-train the encoder of a U-Net [16] model via redundancy reduction in an unsupervised manner to learn the features. The encoder uses unlabelled data-set to learn the data representation. And then the complete network is fine-tuned to provide segmentation. Only limited number of annotated dataset is used for the fine-tuning.

BTUnet puts forward solution to the challenge that the contrastive learning, clustering and distillations perform basis the similarity maximization by efficient generation of the positive (related images) and negative (unrelated images) samples for pre-training. However, in biomedical image analysis all the input samples are related to each other [21]. In Barlow Twins there is no such requirement of positive and negative samples, therefore it seems to be more suitable for biomedical image segmentation. The encoders are based on siamese network [22]. The architecture of BTunet is depicted in Figure4 below.

BTUnet is characterized by its unique loss function [23] , that is combination of invariance and redundancy reduction of the empirical cross correlation matrix of representations' matrix.

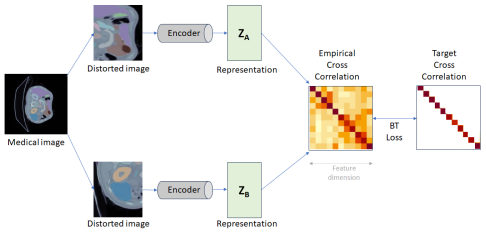



Figure 4: BTUnet Architecture diagram [13]

BTUnet network was executed on DDTI (Thyroid nodules ultrasound image data-set)[11] with inspiration from original implementation[24] . All the 420 images were used as unlabelled data for the Barlow Twin Pre-training. This is not enough number of unlabelled data for any pre-training/ self-training, hence the performance of this model may not be good. Only 10% of labelled data-set i.e. 42 images and corresponding masks were used for the fine-tuning of the network. the results obtained were just descent.

Note:

The strategy to utilize transfer learning works excellent with natural images but had been not been very effective in biomedical image analysis [25] due to large variation in the associated complex patterns of biomedical imaging as compared to natural images. However there have been many excellent developments in transfer learning after mid 2020, hence this area needs to be explored again and hence has been emphasized upon in the following networks.

### 3.3.2 Segment Anything by Meta:

"Segment Anything" [14] is an attempt to make a foundation model for image segmentation to take it to next level. While in the past there were pre-trained models which were further adapted to the downstream tasks, However recently these are being termed as "foundational model" if done on a large scale. It has taken inspiration from the large scale NLP foundation models that have strong capabilities of zero shot or few-shot generalization.

"Segment Anything" has three components: (i) Promptable segmentation task, (ii) model (SAM), and (iii) data-set (SA-1B).

Prompt is a set of points or box or mask that provides information around what to return in the segmentation task. Even when the mask is ambiguous or contains multiple objects, the output must contain mask of at least one of the objects.

The SAM (segment anything model) is based on vision transformers [26]. It uses combination of focal loss and dice loss.

The segment anything data-set consists of 11 million diverse images with 1.1 billion masks. As the number of images grew, the image encoder was also scaled from ViT-B (basic) to ViT-L (large) to ViT-h (huge). It thus works as a model for self-supervised learning.

"Segment Anything" has limitations as cited in its published paper [14] - "It can miss fine structures, hallucinates small disconnected components at times, and does not produce boundaries as crisply as more computationally intensive methods". Also "SAM is designed for generality and breadth of use rather than high IoU interactive segmentation."

Hence we chose DDTI dataset (Thyroid nodule ultrasound image data-set) to be evaluated with SAM. This dataset is not fine distinguishable features, and sometimes poses difficulties to extract the nodules. Secondly, We received just decent results with BTUnet. Thus it was worth to try this data-set to evaluate the performance of SAM.

We used ViT-H pre-trained weights for our evaluation. the code inspiration was derived from Meta [27] and [28]. Only 20% of the "thyroid nodule image" labelled data-set i.e. only 84 images were used for further fine tuning of the SAM.

The results obtained were quite competitive.

Arbitrarily, SAM zeroshot was executed on slice of a ct scan. And the results obtained were quite good visually and we could see various organs, vessels and tissues that are all segmented. The figure5 below showcases the results.

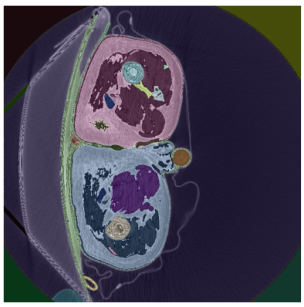


Figure 5: SAM Zeroshot on abdomen CT scan slice

#### 3.3.3 SMIT:

SMIT [15] is "Self-distillation learning with Masked Image modeling method to perform SSL for vision Transformers (SMIT)". This network is executed on 3D multi-organ segmentation from CT and MRI.

SMIT is based on Vision Transformers (ViT) [26]. ViT are known to perform well despite image noise or contrast differences. Hence these are fit for use in medical segmentation. However, training of ViT requires a large number of labelled dataset and this limitation is overcome by self-supervised learning. the code inspiration has been taken from original implementation[29]

This network combines ViT with self-supervised learning using masked image modelling and self-distillation of teacher and student network being trained concurrently. The student teacher network is pre-trained for self-supervised learning using more than 3500 CT scans from head, neck, chest, abdomen, lung, kidney etc. After pre-training, only the student network is retained for further fine-tuning and testing. The pre-trained model training weights for the above referred 3500 CT scans are available. Hence taking that as base, the student network was fine-tuned with following data-set.

- Spleen segmentation (one organ segmentation) [10]
  This is MRI data-set for spleen segmentation. Only 21 labelled MRI scans were used for training, with 9 scans for validation. The test accuracy thereafter was quite competitive despite the fact that the pretrained model was based on CT scan, but the spleen segmentation data-set was MRI scan.
- Amos22 (15 organs segmentation) [9] The student network was fine-tuned to segment for 15 organs in a CT scan. Only 40 labelled CT scans were used and 10 ct scans were used for validation. The test accuracy achieved was excellent, better than the supervised models.

## 4 Experiments and Results

### 4.1 Data-set for Experiments

In the medical domain, most commonly the scans comprise of both 2D as well as 3D properties, hence the data-sets selected for experiments also contain 2D and 3D data with scan types as images (2D), CT scans (3D) and MRI scans (3D). We picked Abdomen CT scan data-set [9] (collection of CT scans that contain 15 organs from abdomen), Spleen MRI data-set[10] and Thyroid nodule image data-set[11]. The abdomen CT scan data[9] is more complex as it contains 15 organs from the abdomen. Selection of these data-sets, fulfills our requirement to test the most general data available in the medical domain.

Described below are these data-sets chosen for the experiments on the selected machine learning networks.

Depicted below is an example of "A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation" [9]. The CT scan and corresponding masks are embedded together to showcase different organs in the abdominal CT scan. This is abdominal organ segmentation data-set with annotations of 15 organs. There are 600 CT scans / MRI with 74k annotated slices. The data-set has been collected from 8 different scanners.

Scans include – Spleen, Right Kidney, Left Kidney, Gall bladder, Esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus.

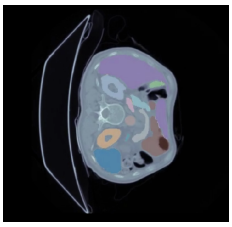


Figure 6: AMOS ct scan with ground truth

Another data-set is "A large annotated medical image dataset for the development and evaluation of segmentation algorithms" [10]. Below figure7 is the left atrium segmentation sample from the data-set.

DDTI [11]is the Ultrasound data-set for Thyroid nodule containing the images of the ultrasounds and the corresponding coordinates for the labels. It is annotated data-set for 99 cases. The figure8a 8b below is the sample where the labels/mask is also converted into an image for training.

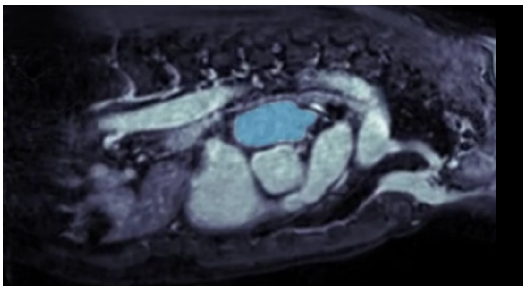
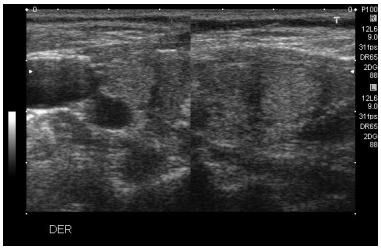


Figure 7: AMOS ct scan with ground truth



(a) Thyroid Nodule ultrasound

(b) position of nodules

Figure 8: Thyroid Nodule Dataset

### 4.2 Experimental settings

Our experiments for evaluation required processing of image data that needs access to high amount of RAM and powerful GPU. Hence the data pre-processing and the models training were executed on google Colab. The models were still taking a lot of time to get trained (some times even two days). The results were heavily dependent on data pre-processing, hence many permutations of data transformations were carried out as referred in section for "data pre-processing" 3.1. And we needed to re-run the models for evaluation. The configuration containing A100 GPU with high ram was able execute the models just satisfactorily.

The requirement for storage also expanded very quickly. Firstly, the images data required a lot of storage space. Secondly, the models being saved at the checkpoints were also consuming a lot of storage during experimentation. overall 1TB is storage seemed sufficient enough for our experiments.

Most of the models were written using pytorch in jupyter notebook, and in some cases Tensorflow / keras was used. As per complexity of the models, we separated the data pre-processing, models, training and evaluation in to different python files or the jupyter notebook. Nibable and Monai are the data pre-processing libraries utilized for 3D image transformations.

### 4.3 Evaluation Metrics

Various significant supervised and self-supervised models were attempted and the accuracy results were recorded in order to study their efficacy. The performance accuracy of the evaluated models were measured via dice loss/ dice coefficient and the intersection over union. One of the other measure for success of the network/ model is the amount of training data used i.e. number of images and corresponding ground truths used for the training / fine tuning of the network/model.

### 4.4 Results

Below is the table1 providing a quick snapshot of the results of all the models/ networks that have been evaluated as mentioned in the "Methods" section 3 and have been discussed further thereafter.

It can be immediately inferred from this table 1that the Self-Supervised networks have performed well and have produced as good results as the supervised models and sometimes surpassing, that too with very few number of annotated images.

Below is further description of the Network-wise results.

Table 1: Accuracy comparison of various Machine Learning Networks

| Type | Model | Accuracy | No. images used | Dataset |
|---|---|---|---|---|
| Self-Supervised | SMIT | 0.81 | 50 | Abdomen CT Scans |
| Self-Supervised | SAM | 0.81 | 100 | Thyroid Nodule images |
| Supervised | nnU-net 2D | 0.81 | 420 | Thyroid Nodule images |
| Self-Supervised | SMIT | 0.79 | 30 | Spleen CT Scans |
| Supervised | nnU-net 3D | 0.76 | 600 | Abdomen CT Scans |
| Self-Supervised | BTunet | 0.68 | 58 | Thyroid Nodule images |
| supervised | 3D U-net | 0.38 | 600 | Abdomen CT Scans |

- BTUnet (self-Supervised model):

  The encoder was trained with 420 unlabelled images of thyroid nodules. Therafter The fine-tuning was performed using 42 labelled images. We obtained the test accuracy of 0.68. The primary reason for a low accuracy seems to be the fact that there were low number of images for pre-trained model hence the learning of features by model was not adequate.

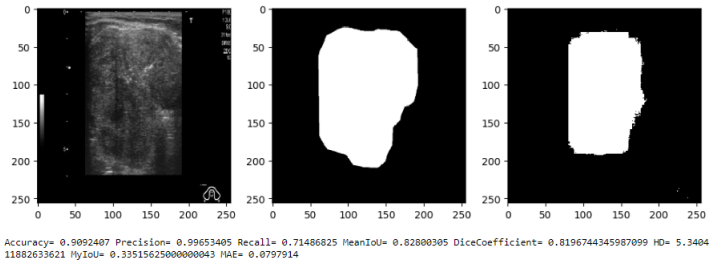



Figure 9: BTUnet self-supervised segmentation

- nnU-net 3D (Supervised model):

  The network was trained with 600 CT scans and corresponding masks of abdomen. the task was to segment 15 organs in the abdomen scans. we obtained a competitive test accuracy of 0.76. we could increase the accuracy if we had more labelled samples, it would have increased the overall learning.

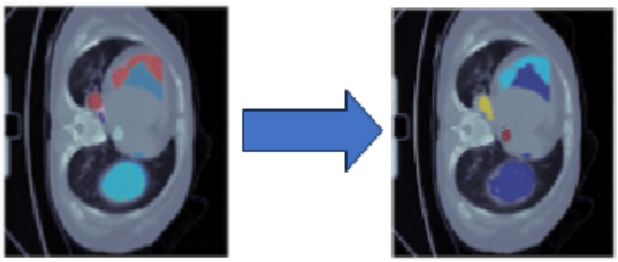


Figure 10: nnU-net 3D supervised segmentation

- SMIT (Self-Supervised model) - with Spleen segmentation:

  The SMIT network was run with two sets of data, first with Spleen Scans (one organ segmentation) and second with abdomen scans (15 organ segmentation). These are the results with Spleen segmentation (one organ). This data is MRI data, and is lesser clear than the corresponding other CT scans. This would mean that there may be a dip in accuracy.

  Only 30 images were used for fine-tuning of this model. The accuracy achieved was 79% which is quite encouraging.

- nnUnet 2D (Supervised):

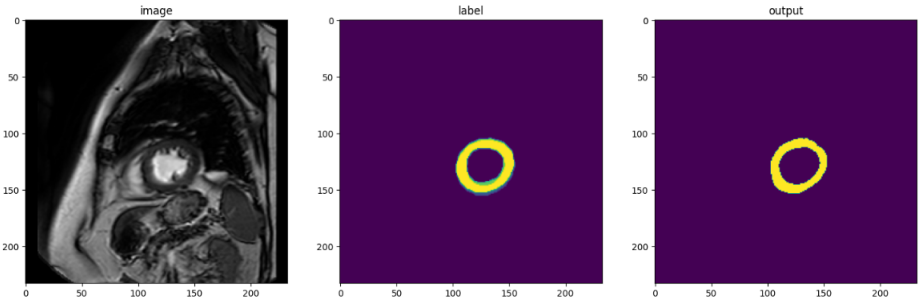



Figure 11: SMIT self-supervised spleen segmentation

This network was evaluated with Thyroid nodule image data-set. The entire data-set of 420 images with corresponding masks was used to train the network. The test accuracy obtained was 0.81, highest in the supervised network segment.

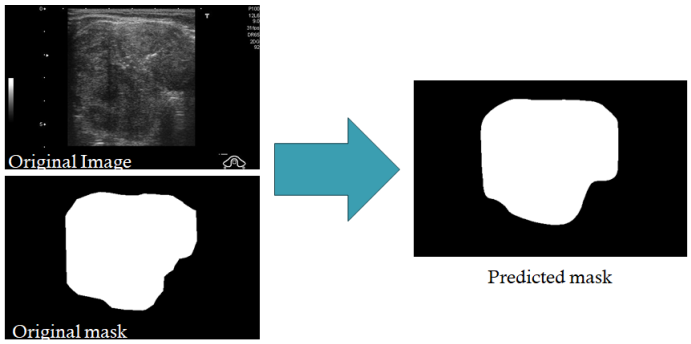



Figure 12: nnUnet 2D Supervised segmentation

- Segment Anything (Self-Supervised):
  The Segment Anything model (SAM) by Meta performs really well for the natural images that exist in the physical world, and that's because it is trained on such images. The SAM was fine-tuned to the medical images of thyroid nodules. only 84 images with corresponding masks were used for the fine-tuning. we achieved good result of accuracy of 0.81. This is as good as what we obtained for the nnUnet 2D Supervised model as in section above 1

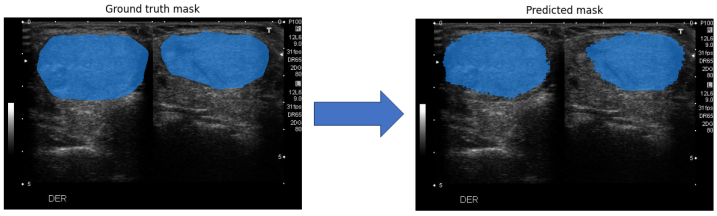



Figure 13: SAM Self-Supervised segmentation

- SMIT (Self-Supervised):
  The SMIT self-supervised network was also evaluated on abdomen CT scan dataset to segment 15 organs. While the task was difficult but the model outshined as it provided a test accuracy of 0.81. Only 50 labelled images were used as against 600 labelled images for the nnU-net 3D supervised model (nnUnet 3D accuracy 0.76 1). The network was able to learn segmentation of 15 organs and was still able to produce the excellent results.

From the results above we can easily deduce that the self-supervised SMIT model has provided excellent test accuracy despite using a complex 3D data-set to segment 15 organs, with only 50 labelled images. Similarly the SAM model

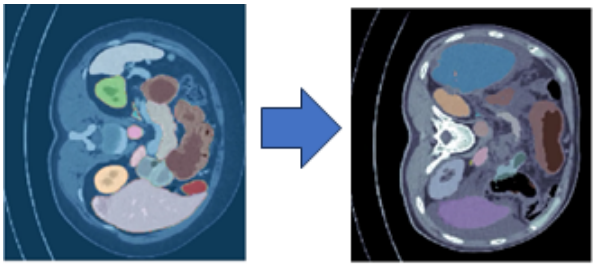


Figure 14: SMIT Self-Supervised Abdomen scan

(self-supervised) has provided almost similar test accuracy as the nnU-net supervised learning model, despite using just 84 labelled images for training. We can clearly see here that the self-supervised models have been performing really great here.

Also, while the supervised model needed to be trained on around 500 to 600 training data images, but self-supervised produced the similar excellent performance with just 50 to 80 training data images.

## 5 Conclusions

It was observed that Self-Supervised models need very less labelled data-set for training. Collecting the labelled data is the major problem in the medical domain, hence this issue is resolved by using self-supervised models.

The accuracy provided by self-supervised models are as good as against the accuracy produced by the supervised models. another critical point to emphasize is that self supervised networks used only 10% to 15% of the data used for the supervised networks.

We are presenting below the points that are the essence of the entire work:

- Self-Supervised networks need less amount of labelled data for further training. This means less number of annotations are required. Annotations in the medical imaging area are very costly and time consuming. Hence a lot of cost and delay is avoided by using the self-Supervised networks.
- Self-Supervised networks provide accuracy as good as the supervised networks and surpasses the supervised networks in some cases. This emphasizes on fact that more and more research should be focused on self-supervised learning in the area of medical domain.
- The foundation models should be built specifically for medical image segmentation. Such foundation models will make the further research very easy and quick.

We had set our investigations to seek resolution to the struggles and questions raised earlier in the Introduction section1 specific to medical domain regarding the lack of labelled training data that inhibits the advances of application of deep learning. With our experiments 3 to evaluate the selected models4.4 with chosen data 4.1, We can conclude that the self-supervised networks are the way forward for any future work in the area of medical segmentation. With our evaluations, we have been able to conclude that While self-supervised networks not only saves cost and time but also provides as good performance/accuracy results as the expansive labelled data hungry supervised models.

## 6 Future Work

We were presented with many thoughts while working on evaluating supervised and self-supervised models. Below are some important areas highlighted that should be continued further to enhance future research.

We could not find appropriate numbers of unlabelled images for pre-training or self-training of the self-supervised models. The more the number of unlabelled data, the better but at least 5000 unlabelled scans should be collected from various sources for effective pre-training/ self-training. A universal data bank should be built for the unlabelled anonymized medical image data and should be accessible publicly. As this is unlabelled data, there should not be any issues of public health information. It should be easy to collect huge amount of unlabelled data by connecting with various institutes, as there are no annotations required.

There is potential to explore SAM with 3D medical image data. This needed more complex processing and more time for further fine-tuning on 3D medical image data (to be executed on the similar lines as we have done for 2d data, by converting the 3D volumes to 2d slices).

Generally the medical images are very heavy (as most general are 3D), hence It takes many days for training and fine-tuning of networks. Hence the results and in turn any troubleshooting is delayed by several days, making it a time consuming process. The availability of better hardware with multiple advanced GPUs will make any future research faster.

There should be consistent efforts in building foundational models for the medical image segmentation specifically. This will enhance the path to future developments in AI field in medical domain. Once there is availability of good amount of unlabelled data, and appropriate hardware, the work of creating foundation models specifically for medical domain will be much more smooth.

## References


[1] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.

[2] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022.

[3] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[5] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[8] Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: A path towards autonomous machine intelligence, 2023.

[9] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation, 2022.

[10] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019.

[11] Pedraza L. and Vargas C.and Narváez F.and Durán O.and Muñoz E.and Romero E. An open access thyroid ultrasound image database, 2015.

[12] Fabian Isensee, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:1–9, 02 2021.

[13] Narinder Singh Punn and Sonali Agarwal. Bt-unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models, 2022.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[15] Jue Jiang, Neelam Tyagi, Kathryn Tringale, Christopher Crane, and Harini Veeraraghavan. Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (SMIT). In *Lecture Notes in Computer Science*, pages 556–566. Springer Nature Switzerland, 2022.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[17] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model, 2023.

[18] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net, 2022.

[19] Yin XX, Sun L, Fu Y, Lu R, and Zhang Y. U-net-based medical image segmentation., 2022.

[20] Abhishek Shivdeo, Rohit Lokwani, Viraj Kulkarni, Amit Kharat, and Aniruddha Pant. Comparative evaluation of 3d and 2d deep learning techniques for semantic segmentation in ct scans, 2021.

[21] Hao Zheng, Jun Han, Hongxiao Wang, Lin Yang, Zhuo Zhao, Chaoli Wang, and Danny Z. Chen. Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation, 2021.

[22] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020.

[23] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[24] Narinder Singh Punn and Sonali Agarwal. Bt-unet: A self-supervised learning framework for biomedical image segmentation using barlow twins with u-net models. *Machine Learning*, pages 1–16, 2022.

[25] Laith Alzubaidi, Mohammed Fadhel, Omran Al-Shamma, Jinglan Zhang, Jorge Santamaría, Ye Duan, and Sameer Oleiwi. Towards a better understanding of transfer learning for medical imaging: A case study. *Applied Sciences*, 10:4523, 06 2020.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[28] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, , and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[29] Jue Jiang, Neelam Tyagi, Kathryn Tringale, Christopher Crane, and Harini Veeraraghavan. Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit). pages 556–566, 2022.