Deep Learning and its applications to Signal and Image Processing and Analysis
Final Project – Medical Segmentation Decathlon

Daniel Duenias and Reut Moshe

Abstract

Segmentation is known to be the most studied medical image processing task, but the various segmentation challenges were usually organized separately, so the development of the algorithms was driven by the need to deal with one specific clinical problem. The Medical Segmentation Decathlon (MSD)—is a biomedical image analysis challenge, in which algorithms compete in a multitude of both tasks and modalities. The hypothesis is that a method capable of performing well on multiple tasks would generalize well to a previously unseen task and might outperform a custom solution. In this work, we present a generic approach based on a 3D U-Net architecture that can perform the segmentation task accurately when trained on various medical segmentation tasks. In addition, we show an inclusive ablation study that contains a comparison between several hyperparameters, models, and augmentation techniques. We will present the advantages of using augmentations, the benefit of ensemble models, the importance of the receptive field for each specific type of dataset, and other findings and conclusions that emerged from the study.

Introduction and Objective

Segmentation is the process of dividing an image into regions with similar properties such as gray level, color, texture, brightness, and contrast. The role of segmentation is to subdivide the objects in an image; in the case of medical image segmentation, the aim is to study anatomical structure [1]. Thus, nowadays, medical images play a crucial role in assisting healthcare providers to treat and diagnose patients. The understanding of medical images depends mainly on the visual interpretation of the radiologists. This is time-consuming and may be subjective, depending on the

radiologist [2]. Still, image segmentation is often the first and the most critical step in the analysis of medical images for computer-aided diagnosis and therapy. Medical image segmentation is a challenging and complex task due to the intrinsic nature of images [3].

One of the most challenging tasks in medical image analysis is to deliver critical information about the shapes and volumes of organs. Using medical image segmentation allows identify the pixels of organs or lesions from background medical images such as CT or MRI images [4]. Segmentation is so far the most widely researched medical image processing task and it faces many challenges. One of the big barriers is the unavailability of an annotated dataset. Collecting annotations is often very tough and very tedious and expensive. Another challenge to deal with is unbalanced labels. It can be overcome by collecting a suitable set of negative samples. The negative set must contain cases that have similar properties but are not positive. Administrative challenges include unavailability of datasets, and privacy and legal issues [4, 5].

As a result of these challenges, many recent medical segmentation systems rely on powerful deep learning models to solve highly specific tasks [6]. In other words, although segmentation is so far the most widely investigated medical image processing task, the various segmentation challenges have typically been organized in isolation, such that algorithm development was driven by the need to tackle a single specific clinical problem. This limits our understanding of the generalizability of the proposed contributions. The Medical Segmentation Decathlon was created to develop a generic model capable of performing well on multiple tasks that needs to implement [7].

To address this challenge, we used the U-NET architecture. U-NET was developed for biomedical image segmentation – datasets with very few images. Our model is trained on a training set that includes various types of augmentation. The evaluation of the model results was with the Dice Score function. We also optimized our results by hyper-parameters search.

Data Description

The Medical Segmentation Decathlon is a collection of medical image segmentation datasets. It contains a total of 2,633 three-dimensional images collected across multiple anatomies of interest, multiple modalities, and multiple sources. Specifically, it contains data for the following body organs or parts: Brain, Heart, Liver, Hippocampus, Prostate, Lung, Pancreas, Hepatic Vessel, Spleen, and Colon [7].

Medical Segmentation Decathlon aims to create a segmentation algorithm that generalizes across 10 datasets corresponding to different entities of the human body. These algorithms may dynamically adapt to the specifics of a particular dataset, but are only allowed to do so in a fully automatic manner [8].

The MSD data set is publicly available under a Creative Commons license CC-BY-SA4.0, allowing broad (including commercial) use. The training data is available at http://medicaldecathlon.com/.

In the attached we focused on only 2 of them: Heart and Hippocampus.

Heart

The heart is a muscular organ in most animals. It pumps blood through the blood vessels of the circulatory system [9]. The pumped blood carries oxygen and nutrients to the body while carrying metabolic waste such as carbon dioxide to the lungs. In humans, the heart is approximately the size of a closed fist and is located between the lungs, in the middle compartment of the chest [10].

The heart is located between the lungs in the middle of the chest, behind, and slightly to the left of the breastbone (sternum). A double-layered membrane called the pericardium surrounds the heart like a sac. The outer layer of the pericardium surrounds the roots of the heart's major blood vessels and is attached by ligaments to your spinal column, diaphragm, and other parts of the body. The heart weighs between 7 and 15 ounces (200 to 425 grams) and is a little larger than the size of your fist. By the end of a long life, a person’s heart may have beat (expanded and contracted) more than 3.5 billion times. In fact, each day, the average heart beats 100,000 times, pumping about 2,000 gallons (7,571 liters) of blood [29].

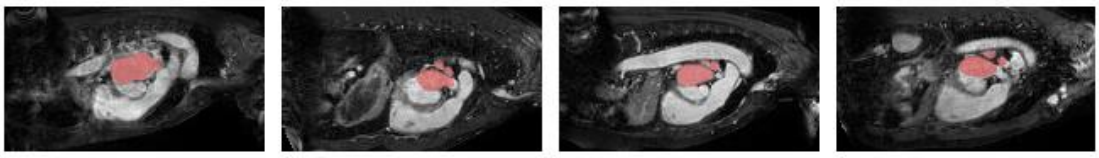
In addition, the human heart contains four chambers that are situated just to the left of the midline of the thoracic cavity. The upper two chambers (atria) are divided by a wall-like structure called the interatrial septum. The lower two chambers (ventricles) are divided by a similar structure called the interventricular septum. Between each atrium and ventricle, valves allow blood to flow in one direction, preventing backflow [11]. One of the four chambers of the heart is the left atrium which is located on the left posterior side. The left atrium's primary roles are to act as a holding chamber for blood returning from the lungs and to act as a pump to transport blood to other areas of the heart. The walls of the left atrium are slightly thicker than the walls of the right atrium. Oxygen-rich blood from the lungs enters the left atrium through the pulmonary vein. The blood is then pumped into the left ventricle chamber of the heart through the mitral valve. From there, the blood is ready to be pumped into the body to deliver oxygen-rich blood to all bodily tissues. Mitral valve prolapse is a common affliction in which the mitral valve between the left atrium and left ventricle does not close properly. This condition does not typically require treatment; however, some patients with mitral valve prolapse can develop more serious conditions that require treatment. One such condition is mitral valve regurgitation, in which blood leaks back into the left atrium through the mitral valve [30].

Furthermore, the heart which evolved hundreds of millions of years ago developed a cardiovascular system with complete separation between pulmonary and systemic circulations incorporated into a single pump with chambers dedicated to each circulation. A lower pressure right heart chamber supplies deoxygenated blood to the lungs, while a high pressure left heart chamber supplies oxygenated blood to the rest of the body. Due to the complexity of morphogenic cardiac looping and septation required to form these two chambers, congenital heart diseases often involve maldevelopment of the evolutionarily recent right heart chamber. Consequently, some diseases predominantly affect structures of the right heart, including arrhythmogenic right ventricular cardiomyopathy (ARVC) and pulmonary hypertension [12].

To sum up, structure and function in any organ are inseparable categories, both in health and disease. Whether we are ready to accept it or not, many questions in cardiovascular medicine are still pending, due to our insufficient insight into the basic science [13].

The heart dataset was provided by King's College London (London, United Kingdom), originally released through the Left Atrial Segmentation Challenge (LASC) [14]. This dataset includes 30 MRI datasets covering the entire heart acquired during a single cardiac phase (free breathing with respiratory and ECG gating). Images were obtained on a 1.5T Achieva scanner (Philips Healthcare, Best, The Netherlands) with a voxel resolution of 1.25 x 1.25 x 2.7 mm^3. The left atrium appendage, mitral plane, and portal vein endpoints were segmented by an expert using an automated tool followed by manual correction [15].

The dataset contains 30 3D volumes and is split into 20 training images and 10 testing images with variable volumes sizes. The training set also contains ground truth binary volumes that are used as labels. Voxel with value 1 represents the left atrium, and 0 represents the background [7, 16]. The data contains decent resolution (around 320x320x130) images while some of the images are noisier and less clear than the others. Combining this with the fact that the dataset is small, we get high variance data set. The whole area of the heart and its surroundings has roughly the same gray level, so the segmentation has to consider mainly shapes, edges, and special features.



Single z slice with its ground truth in red

Hippocampus

Hippocampus is a complex brain structure embedded deep into the temporal lobe. It has a major role in learning and memory. It is a plastic and vulnerable structure that gets damaged by a variety of stimuli. Studies have shown that it also gets affected in a variety of neurological and psychiatric disorders. In the last decade or so, a lot has been learned about conditions that affect the hippocampus and produce changes ranging from molecules to morphology. Progresses in radiological delineation,

electrophysiology, and histochemical characterization have made it possible to study this archicerebral structure in greater detail [17].

The hippocampus plays a fundamental role in long-term memory, response to stress, and contextualization of emotional experience. The entorhinal cortex is part of the parahippocampal gyrus but is the main hippocampal input structure, a pathway along which information converges to be memorized [18]. Numerous studies in animal models imply that hippocampal neurogenesis is important for functions such as learning, memory, and mood. Interestingly, hippocampal neurogenesis is very sensitive to physiological and pathological stimuli [19].

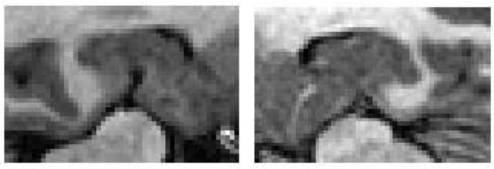
The hippocampal formation is a prominent region of the adult cerebral cortical mantle that is usually described as an infolding of the medial temporal lobe. Based on a modern understanding of how the trisynaptic or intrahippocampal circuit is organized structurally and functionally, it is convenient to define the hippocampal formation as consisting of two major divisions: hippocampal region, and retrohippocampal region (with subiculum, presubiculum/postsubiculum, parasubiculum, and finally entorhinal area—the traditional starting point of the trisynaptic circuit). Topologically, the hippocampal formation displays a unique cortical architecture: each of its areas has a longitudinal axis, overall forming an 8-membered palisade beginning at the mantle's embryologically medial edge with the dentate gyrus, followed sequentially by Ammon's horn (with its successive fields CA3, CA2, and CA1), subiculum, presubiculum and postsubiculum, parasubiculum, and entorhinal area. Superimposed on this macrostructure, the intrahippocampal circuit—which interconnects all 8 cortical areas—is arranged along the transverse axis of each area, and thus the hippocampal formation as a whole [20].

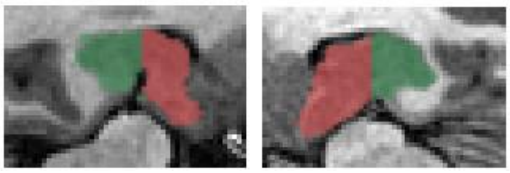
Damage to the hippocampus typically produces temporally graded retrograde amnesia, whereby memories acquired recently are impaired more than memories acquired remotely. This phenomenon has been demonstrated repeatedly in a variety of species and tasks. It has also figured prominently in theoretical treatments of memory and hippocampal function [21].

To sum up, the hippocampus is one of a group of remarkable structures embedded within the brain's medial temporal lobe. Long known to be important for memory, it has been a prime focus of neuroscience research for many years [22].

The dataset consisted of MRIs acquired in healthy adults and adults with a non-affective psychotic disorder taken from the Psychiatric Genotype/Phenotype Project data repository at Vanderbilt University Medical Center (Nashville, TN, USA). Patients were recruited from the Vanderbilt Psychotic Disorders Program and controls were recruited from the surrounding community. All participants were assessed with the Structured Clinical Interview for DSM-IV [23]. Structural images were acquired with a 3D T1-weighted MPRAGE sequence (TI/TR/TE, 860/8.0/3.7 ms; 170 sagittal slices; voxel size, 1.0 mm^3). All images were collected on a Philips Achieva scanner (Philips Healthcare, Inc., Best, The Netherlands). Manual tracing of the head, body, and tail of the hippocampus on images was completed following a previously published protocol [24, 25]. For this dataset, the term hippocampus includes the hippocampus proper (CA1-4 and dentate gyrus) and parts of the subiculum, which together are more often termed the hippocampal formation [16, 26].

The dataset contains 394 3D volumes which 263 images for training and 131 for the test. Here also, the sizes of the volumes are not constant. The training set also contains ground truth volumes that are used as labels. The segmentation in the hippocampus is into three segments, and each voxel in the ground truth label is 0,1, or 2. Voxel with value 2 represents the anterior, 1 represents the posterior, and 0 represents the background. This data has a low resolution of around 40x60x40. The hippocampus has a unique gray level that is a bit different from its close environment. However, the borderline between both of its parts seems to not have many conclusive features, a fact that might make the training difficult.



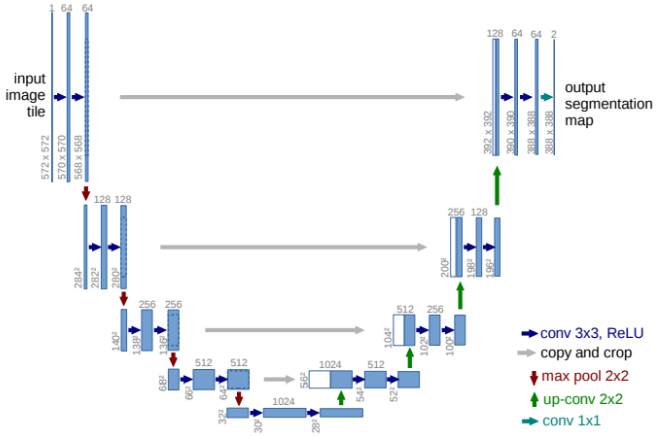


Single z slice with its ground truth (red – posterior, green – anterior)

Methods and Algorithms

1. Model architecture

To address this challenge, we used the U-NET architecture. U-NET was developed for biomedical image segmentation – datasets with very few images. The architecture consists of a contracting path and a symmetric expanding path. The contracting path's role is to capture context [27].



In our model, we used 3D U-NET architecture with padding to keep the output image in the same dimensions (no cropping is needed in the skip connections). To reduce overfitting, we used dropout (we decided where to add the dropout layers according to Alex Kendal et al.[28] and L2 batch normalization.

2. Loss functions

As loss functions, we use a combination of the cross entropy and the dice score as a loss function. The cross-entropy for the ground truth label y_i and the prediction $\widehat{y_i}$ is defined as follows:

$$cross\ entropy\ loss = -\sum_{i=1}^{n} y_i log\widehat{y_i}$$

The dice score is defined as follows:

$$dice\ score = \frac{2 * TP}{(TP + FP) + (TP + FN)}$$

The dice loss is defined as follows:

$$dice\ loss = 1 - dice\ score$$

Due to the class imbalance, we used weighted dice and weighted cross entropy.

The combination of the two losses was calculated as follows:

$$combined\ loss = \alpha \cdot dice\ loss + (1 - \alpha) \cdot cross\ entropy\ loss\ ,\ 0 \leq \alpha \leq 1$$

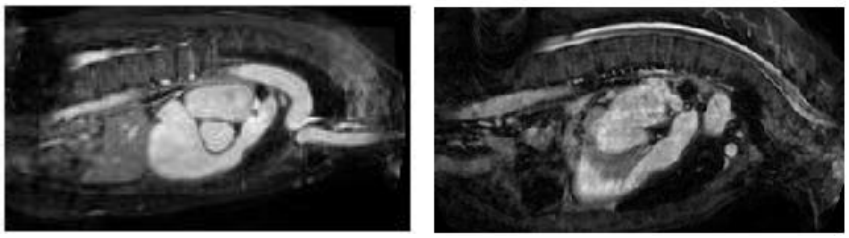
Our U-Net model needed the input to be in the size of $2^{UNet\ depth}$ in each axis to work properly and the images in the dataset were in various sizes so we padded the images with 0 to a specific size and the label with -100 in those places. Thus, the loss calculation excluded the -100 voxels and enabled a faster and better convergence.

3. Augmentations

Since the number of images in the training dataset is limited and for better generalization, we used several augmentations:

Noise

Adding random noise is one of the approaches to improving generalization error and expands the size of the training dataset. The addition of noise during the training of a model has a regularization effect. random gaussian noise with different σ while the σ is also chosen randomly. That way some images have different noise levels (and some don't have any noise at all). Looking at the data we saw that some images are noisier than the others, so we decided to apply the noise augmentation this way. It also helps the model to focus on the shapes and details of the image and not rely hardly on the gray levels.

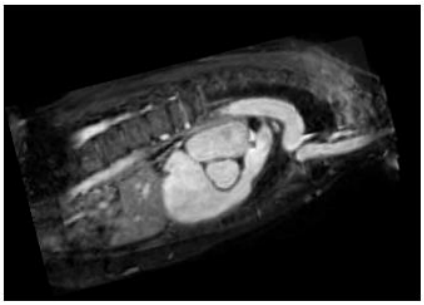


Left - without the augmentation, right – noise augmentation

Rotation

Another common data augmentation technique is random rotation. Random rotation can improve the model performance without collecting and labeling more data. We

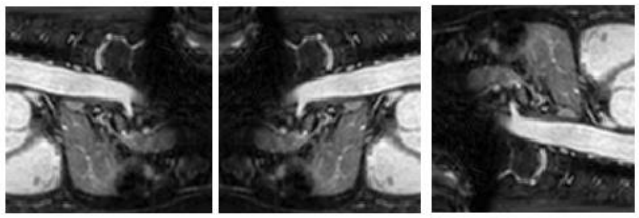
performed a random small rotation (-2 to 2 degrees) \ big rotation (-10 to 10 degrees) angle rotation. The rotation is random in each axis separately.



Z axis rotation

3D X, Y and Z axes flips

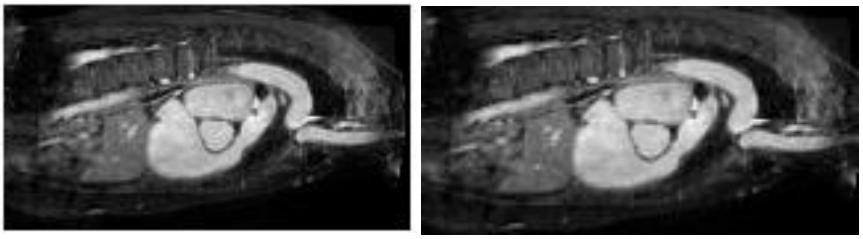
Adding more information and images to learn from without having to go through the time-consuming process of collecting and labeling more training data can be done by creating several versions of the images in various orientations. We performed a random flip with p=0.5 on each axis (called "basic aug" in the Hippocampus parameters sweeps).



X and y axes flips

Rescale

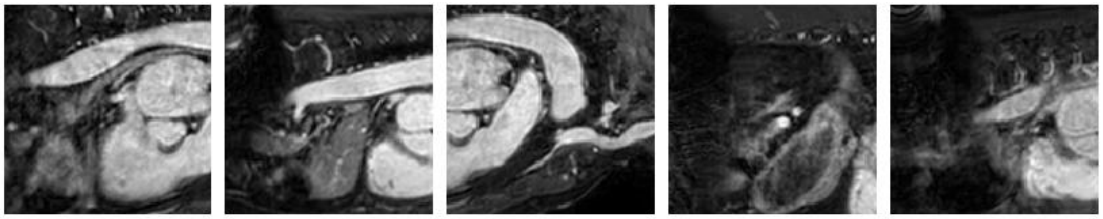
Image rescale augmentation is an augmentation technique where we randomly pick the short size of an image within a dimension range. Understanding that rescaling creates interpolation that changes the pixel's values but not the essence of the image, we used small (up to $\pm$5%) and big (up to $\pm$10%) random rescaling for better generalization and enriching our dataset.



Left – regular image, right – cropped to the same size, +10% scaled image

Random cropping

Random cropping was used in the heart dataset. Random cropping of 96 pixels cube for both creating more samples and enabling the data to pass through the network (smaller images - memory-wise). we tried cropping around the segmentation target and random cropping and found that it is important for the generalization (validation set) to crop around places that have only a background.



We decided not to add more augmentations because of the high number of hyperparameters that we wanted to test and the training times.

4. Ensemble, 5 folds and test time augmentations

We wanted the models to learn from all the training data and still be able to monitor their progress with the validation set so we divided the training data into 5 folds of train-validation sets (80% train, 20% validation) without validation overlap. In addition, to reduce the variability of the random initializations of the models, we created an ensemble model by averaging the soft predictions of a few of the best models in each fold. Therefore, our final model was an ensemble model of several models in each data fold (for example, 3 models in each fold – 15 models). Eventually, we added test time augmentations (the best augmentations according to the ablation study presented below) to get even better results.

5. Transfer learning

Trying to face the "small Heart dataset" problem, we conducted several experiments harnessing the power of transfer learning. Due to our 3D U-Net architecture, we were unable to use image-net pre-trained encoders such as VGG, so we tried to train a Heart model using the encoder of a trained Hippocampus model (and vice versa, just to see the effect). Unfortunately, the results were unsuccessful. Analyzing the results, we concluded that due to the big difference in the datasets, both resolution and shapes (internal organs and part of a brain), the features that each model's encoder learns are not similar enough to be transferred.

Challenges and Difficulties

Small training dataset with large variability

The heart dataset was a more challenging one due to the small number of images in the training dataset - only 20 3D volumes for the Heart. We used a variety number of augmentations to produce additional information the model can learn from.

Segmenting two neighboring small structures with high precision

The hippocampus presented a low resolution with a high demand for distinguishing between 2 close structures of the same tissue.

Preprocessing the data

Preparing the images to enter the pipeline while each image has different dimensions and adding augmentations was one of the main challenges. For example, creating some 3D augmentation (such as rotation) which we had to implement ourselves, or finding the best way/size to pad the images for the same dimensions.

Imbalance data

The voxel distribution per class in the images was sharply unbalanced. The voxels labeled with the desired class were much smaller than the background of the image. To solve this challenge, we used weighted loss and give more weight to the less common labels.

Long running time

Given the large size of the images (especially those of the heart section), the running time of each model was very long. Mainly for the hyper-parameter searching step, we

had to plan the experiments in the best possible way, while combining running on several GPUs at the same time, to get as much information as possible.

Memory

The memory consumption bottleneck created by the size of the data, the depth of the network, the batch size, the initial number of features, and the 3D convolution that we chose to use forced us to look for creative solutions and make some compromises during the process.

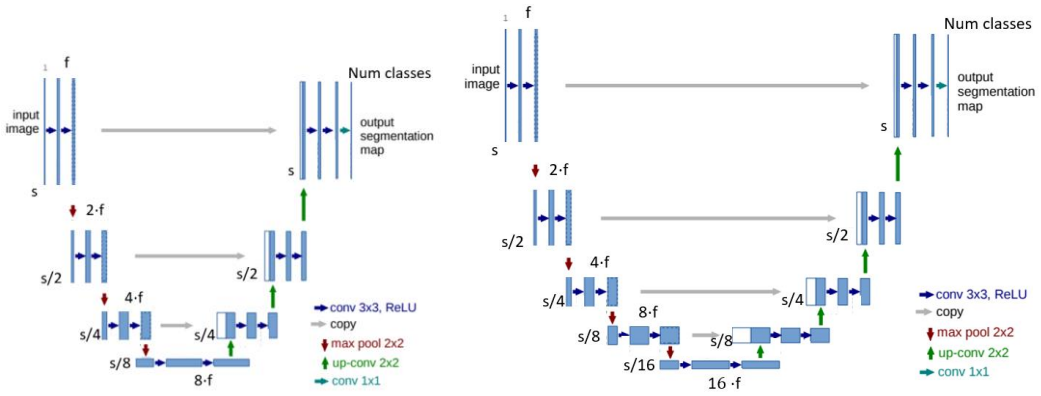
Experiments

We used the 3D U-Net model for its well-known performance in the segmentation task. However, we wanted to optimize it so that the same architecture will give us the best possible results on both segmentation tasks (as the challenge demands). To do that, we decided about a set of changeable parameters and initiated several "parameters sweeps" looking for the optimal set. Each dataset had 3 parameters sweeps. The first one was coarse and had a low number of epochs (to train a lot of different models). The goal was to roughly filter irrelevant values of parameters. The second sweep's goal was to better focus on the relevant values while training for a longer time to get a higher result and understand each parameter effect (in the Hippocampus data this sweep started to check the effect augmentations because of the computation time allowed it – smaller data resolution). The third and final sweep was intended to find the best set of parameters, including augmentations, using a high number of epochs for training (without overfitting). All the sweeps were done on data fold 1 (out of 5).

The set of parameters we tested:

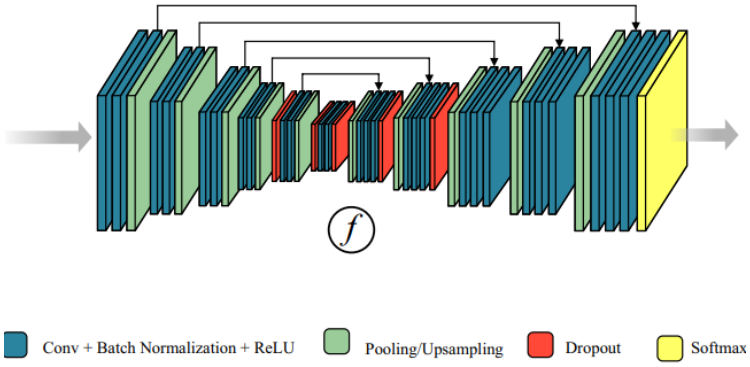
1. Model depth – understanding the effect of the depth on the receptive field and its relevance for each dataset we tried 2 different depths, 4 layers deep and 3 layers deep (UNet and Short_UNet correlatively). The receptive fields were 140 (UNet) and 68 (Short_UNet). We hypothesized that because of the sizes of the images, both the Short_UNet and the UNet should perform about the same in the Hippocampus dataset. However, cropping the Heart dataset to both sizes of receptive fields and observing the details we thought that the bigger receptive field is necessary for understanding the context of the

segmentation target and its difference from other similar, close by shapes and colors in the image.



Our architectures (left- UNet_Short, right- UNet). While 'f' is the initial number of features and 's' is the image's shape (in each axis s can be different). We pad with zeros before the convolution to preserve sizes.

2. Number of initial features – in our model, each depth layer doubles the features in the layer above it. More initial features results in more features map in each depth layer. This parameter was limited due to the high memory consumption of the 3D convolutions.
3. Class weights – both datasets are imbalanced (a lot of background). The weights were calculated using different methods.
4. Loss function mixture factor – we used a weighted mixture of cross entropy and dice loss.
5. Learning rate
6. Batch size – the batches size was limited due to the 3D convolutions (because of high memory consumption) and was entangled with the number of initial features (the maximum batch size in the heart dataset was 4).
7. Dropout probability – reading several works on the subject [28] we decided to add the dropout layers as follows:



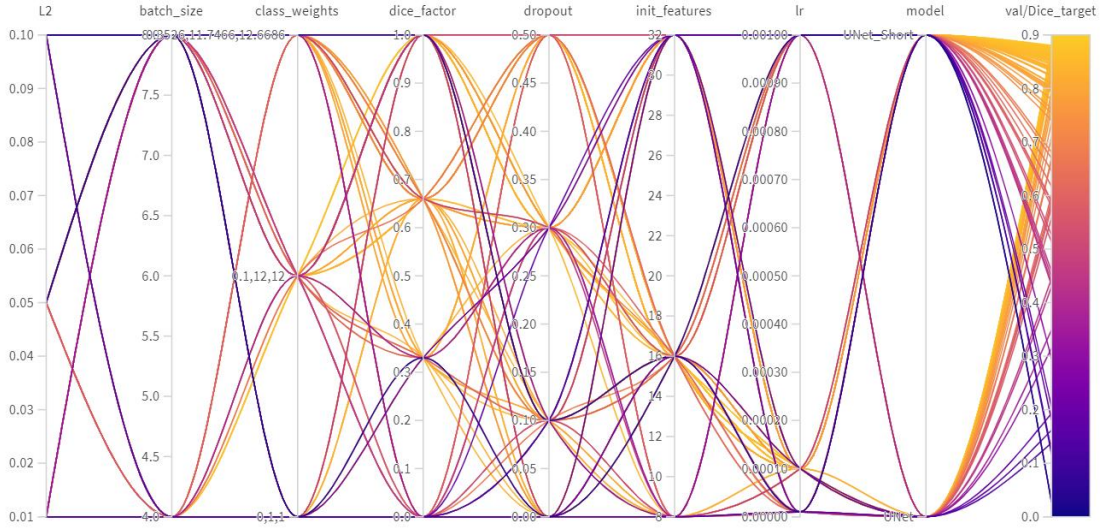
8. L2 factor
9. Augmentations (listed and elaborated above)

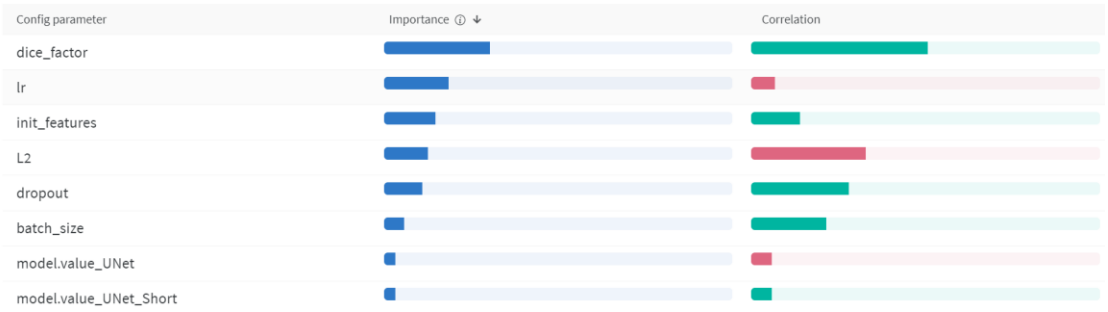
For evaluating our results, we used the Dice coefficient as our main metric. averaging the Dice coefficient between all the target classes (1 in the heart and 2 in the hippocampus). Another metric we used for training was IOU. Those two metrics were chosen for their relevance in the segmentation task. We used Dice as our main metric because that was the challenge's main metric. As a subjective measure, we used visualizations of the predictions (which will be presented later).

All the sweeps were maximizing the Dice coefficient of the target in the validation.

<u>Hippocampus:</u>

1. Coarse initiation sweep – random sweep with 168 runs, 60 epochs

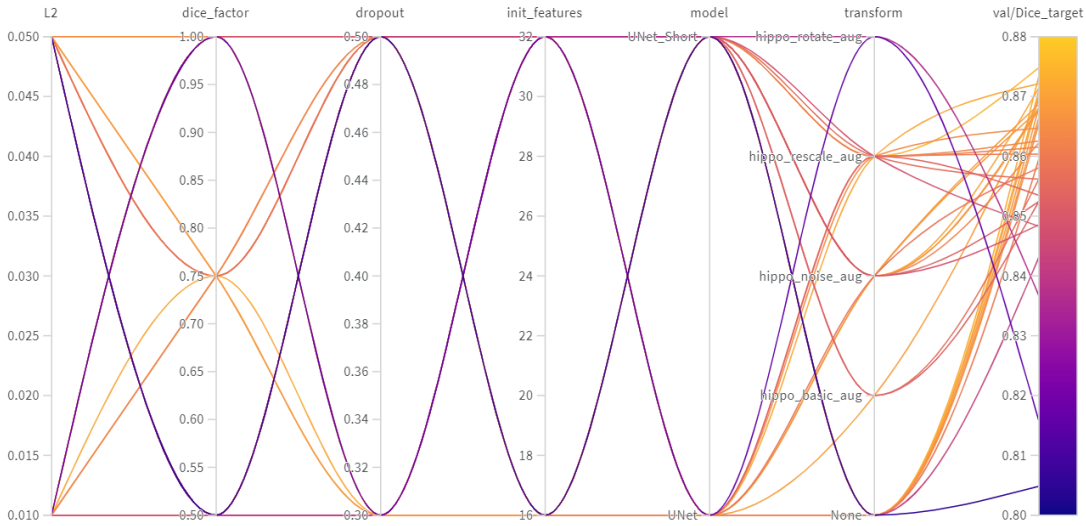




The correlation bar is green if the higher value gives higher results and red if higher value gives lower results. For categorical parameters (such as a model type) the bar is green if it brings higher results. The bigger the bar, the higher the correlation of the parameter.

As we can see here, the Dice factor is important and positively strongly correlated (which is not a surprise considering we are optimizing the Dice coefficient). The learning rate is negatively correlated (even though there is a low number of epochs) and has high importance. However, the model type has no high significance, as we predicted in this data. A high number of features also results in a better score (our maximal number was 32 due to memory limitations) because it increases the chances of finding good features (in the basic 2D U-Net it was 64). Narrowing the ranges and adding augmentations, we initiated the second sweep.

2. Second sweep - random sweep with 53 runs, 400 epochs

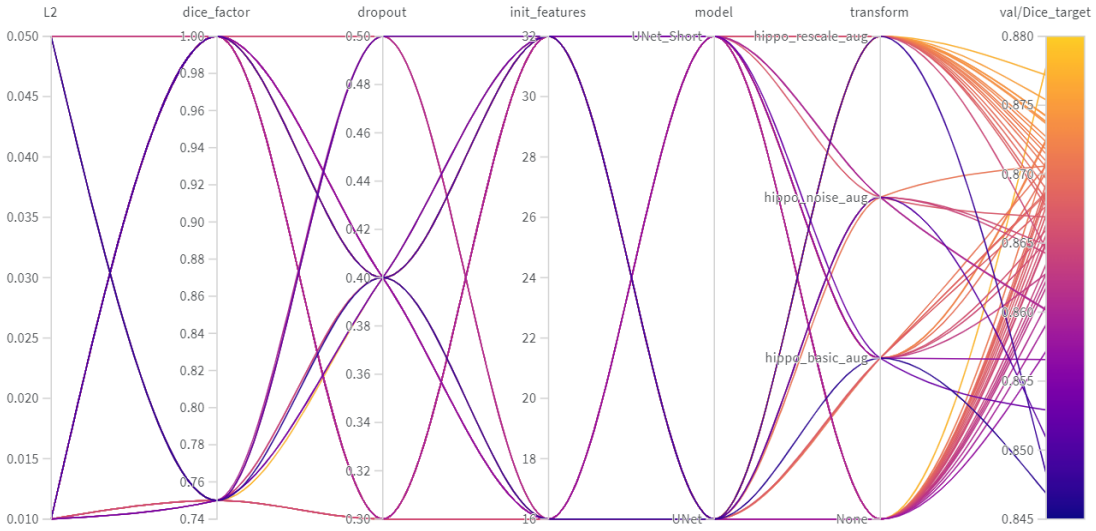


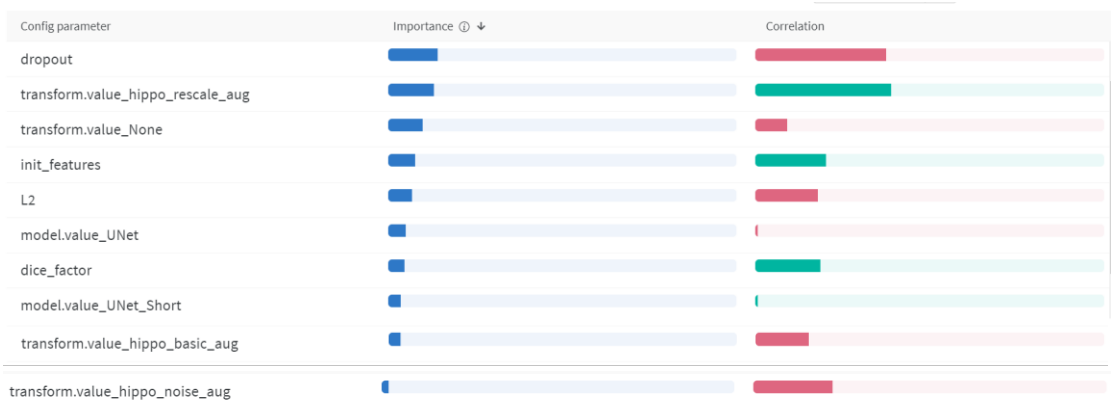
First, we can see that instead of spreading all over the scale starting with an almost 0 score, we are generally doing much better and the scores are between 0.8 to 0.88 while most are closer to the top.

We assume that because all the images (training, validation, and test) have the same orientation, the rotation was harmful, doesn't matter how small it was. On the other hand, small rescaling did perform better as we predicted. The flipping (basic_aug) worked had a small correlation and importance.

We saw that 400 epochs cause slightly overfitting so for the final sweep we changed it to 200 (where no model overfitted – with correlation to the dropout and L2 regularization we chose).

3. Final sweep - random sweep with 57 runs, 200 epochs





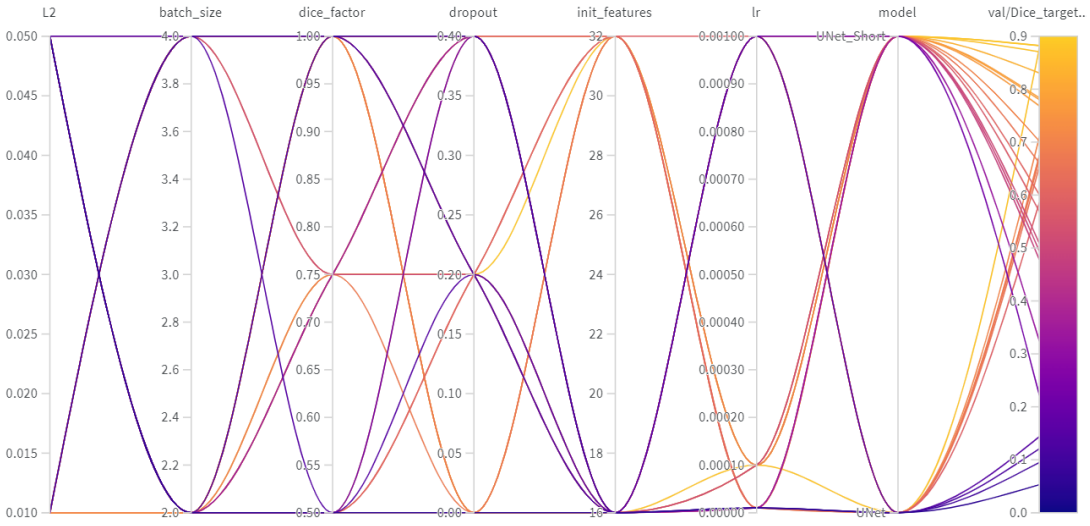
Here we used noise with bigger σ to see the effects and found that it was too noisy for this dataset (as it is a low resolution dataset).

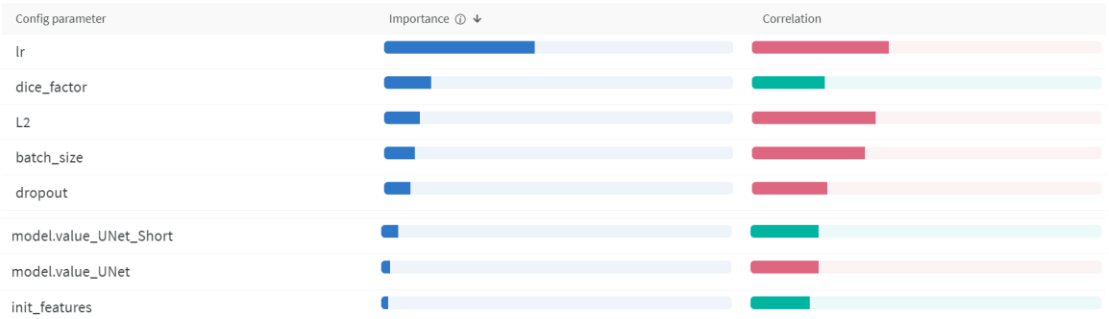
As we can also see, the dropout should not be so high and there are not many differences between the depths of the networks.

Heart:

In the beginning, we tried random cropping around the target. That was not so successful on the validation data because the model was missing a lot of places that are only background and are a bit similar to the segmentation target. That is why we decided to use random cropping in the whole image and decrease the learning rate for more stable training. In all these trainings we used the same random cropping.

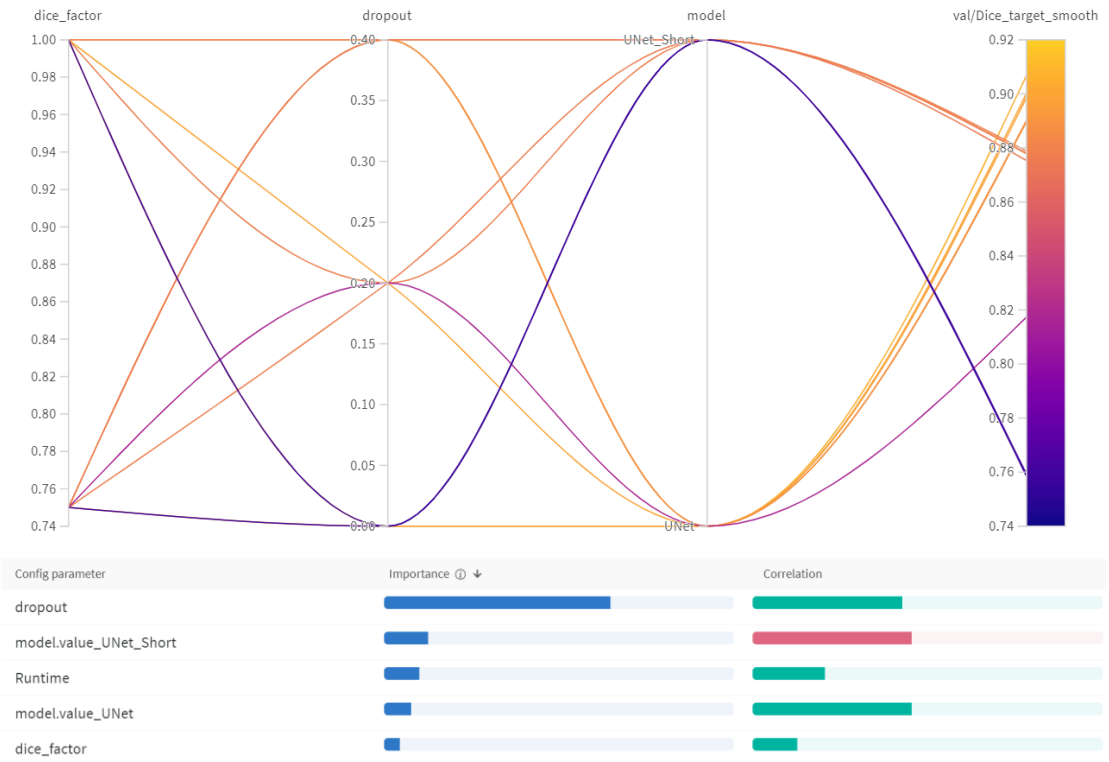
1. Coarse initiation sweep - random sweep with 32 runs, 300 epochs





We can see here that the learning rate is negatively correlated (even though there is a low number of epochs) and have the highest importance. However, the model depth has no high significance, not as we predicted in this data. Nevertheless, focusing only on the models with the high results our hypothesis holds (as shown in the next sweep). Here we had memory limitation and for 32 features and batch size 4, the model crashed so that is the reason for the misleading correlation between the batch size and the score.

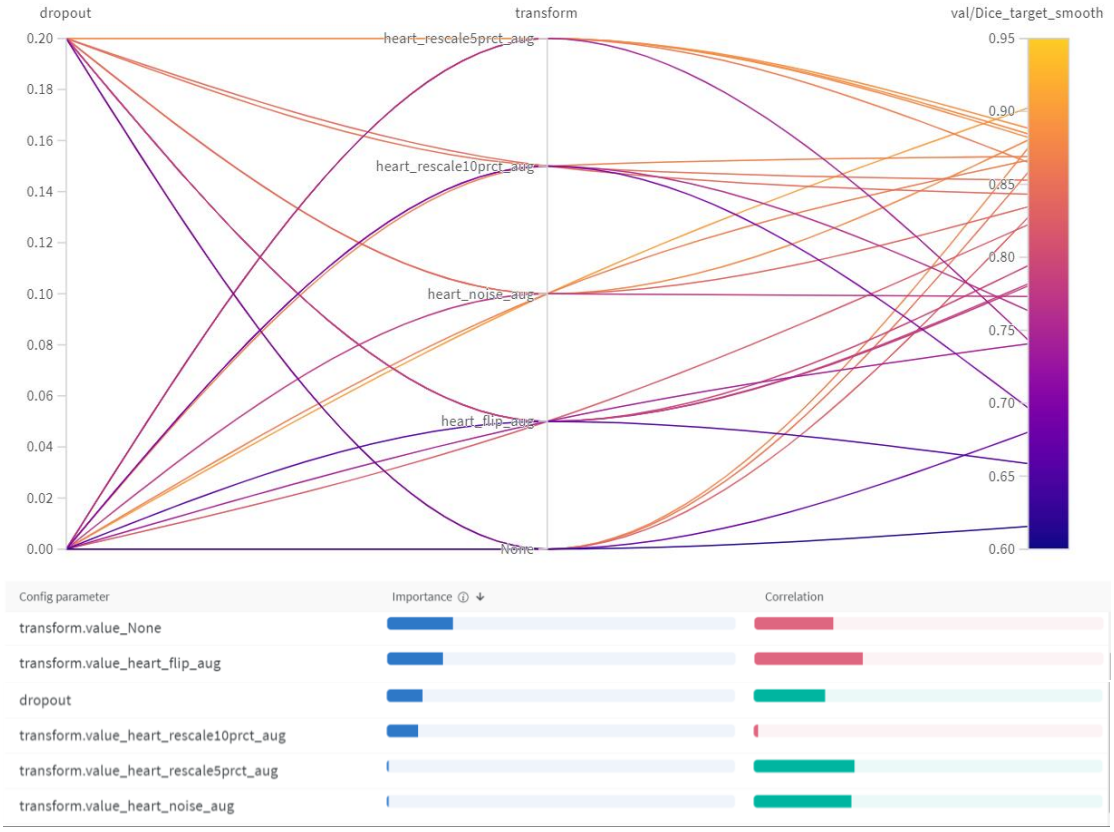
2. Second sweep (filtering some parameter ranges) - grid sweep with 12 runs, 500 epochs (grid checking all the parameters).



As we can see here, for that data, the receptive field needs to be bigger for the results to be better. As explained before, there are a lot of similar shapes and colors around

the heart and to segment a specific part of the heart the model must be able to "see" all of those together. Here, again, adjusting the parameters range we zoom into the 0.74 - 0.91 part of the score scale.

3. Final sweep (augmentation) - grid sweep with 30 runs (multiple models per parameters set for more stability), 500 epochs

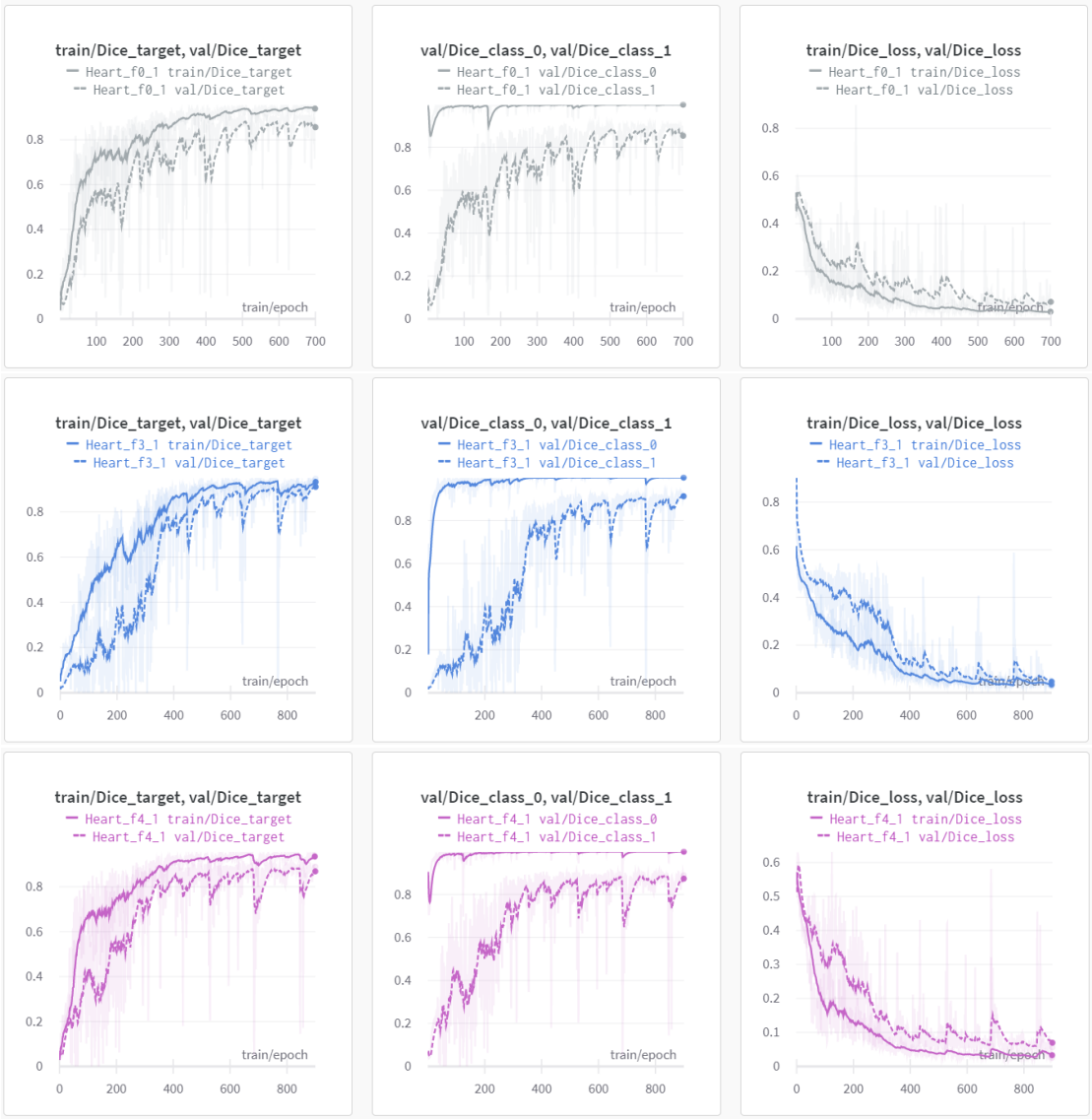


We intentionally removed the rotation augmentation after we saw the results on the hippocampus data and for faster training runtime. Here the flipping decrees the score, but the small rescaling (5%) and the noise have a positive correlation (although their importance is rather low, they are not harmful). We can also observe that without augmentation the scores are worse, and that parameter has high importance. The dropout should be slightly higher to get the best results.

Taking into consideration all the results above, we decided that the architecture should be the U-Net (with the 140 receptive field) and the number of initial features should be 32. Also, to get the best results for both datasets, the dropout probability should be 0.3 and the learning rate 0.0001. the L2 regularization factor is 1% and 2.5% in the hippocampus and the heart datasets correlatively. We also composed an

augmentation that fits each dataset (which will be also used for the test time augmentations) and trained several models on every fold of the 5 data folds. All the final models' train were stopped before overfitting.

Presented below, the training curves of three models, in three different data folds of the heart dataset.



From the top to the bottom are folds 1, 4 and 5 in correlation. To the left, Dice target is the Dice coefficient target class. The middle graph shows the validation Dice per class. The right graph shows the weighted dice loss of all 2 classes (background included – class 0). Both left and right - stripped line for validation and full line for training.

Because most of the pixels are background, even with the weighted loss we get that the background receives higher scores. The training is very noisy because each epoch here contains a small number of images, and it is being randomly cropped only once. Therefore, some epochs may receive only easy background and get high scores while

others can get more complex parts. We can also see that some folds get better results than others. That should be an advantage for the ensemble model's generalization abilities.

Presented below, the training curves of three models, in three different data folds of the hippocampus dataset.

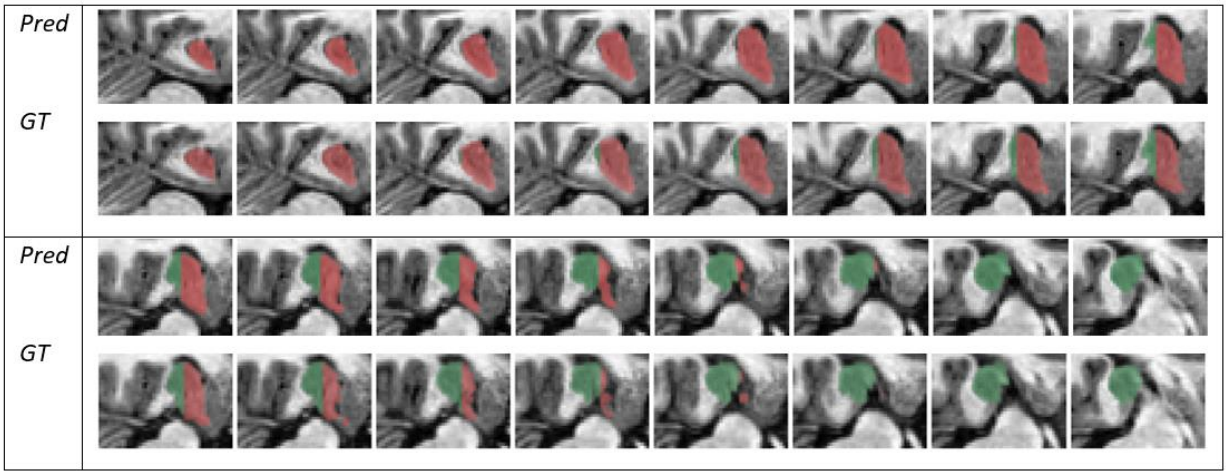
From the top to the bottom are folds 1 to 3 in correlation. To the left, Dice target is the average Dice coefficient of the two target classes. The middle graph shows the validation Dice per class. The right graph shows the weighted dice loss of all 3 classes (background included – class 0). Both left and right - stripped line for validation and full line for training.

Here also, the background receives a higher score than both other classes. Another observation is that on most of the folds, the model scores higher in class 2 rather than class 1 at the beginning but then it changes, and class 1 gets a higher score eventually.

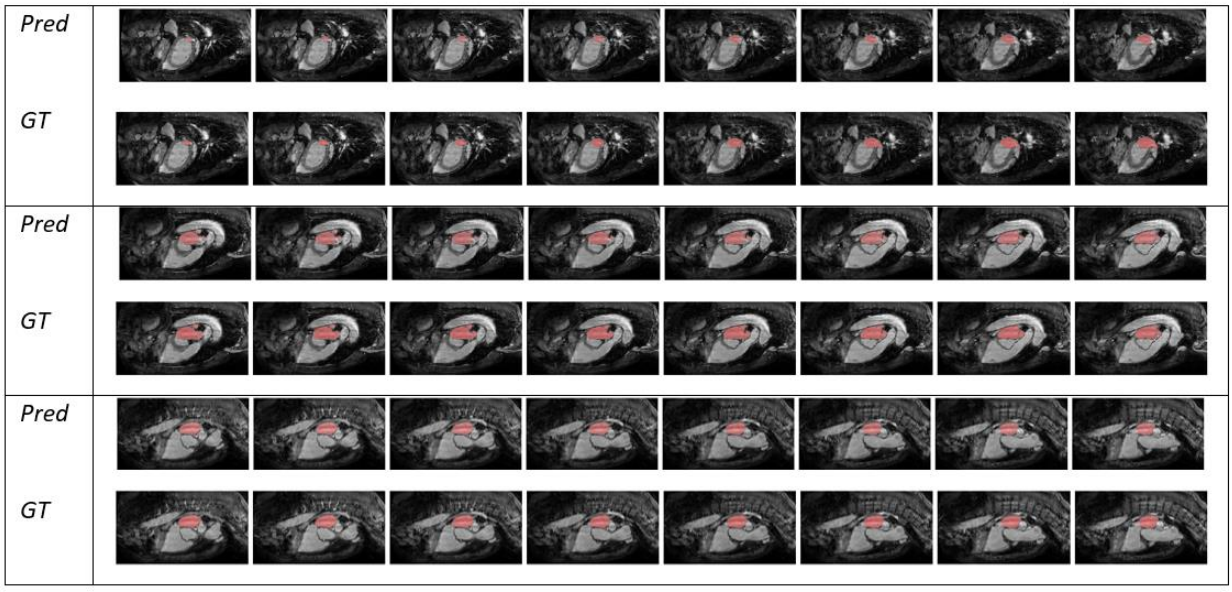
The hippocampus has smaller resolution and bigger dataset so here the convergence is much faster and cleaner than the heart dataset. The same as in the heart dataset, we can also see that some folds get better results than others. The ensemble model's generalization abilities should benefit from this here as well.

IOU and other graphs were excluded from this report for the reader's convenience.

Here we present the prediction over a random validation image and the ground truth.



Hippocampus – prediction and ground truth spread on several z slices



Heart – prediction and ground truth spread on several z slices

Visually, on both datasets, the models seem to achieve good results. In the hippocampus dataset, it seems that the model's prediction favors the red class a little bit which is in correlation to what we saw in the graphs earlier.

Results

In the MSD challenge we didn't receive our leaderboard results yet (and probably never will):

Created	Phase	User	Comment	Evaluations
July 28, 2022, 8:33 a.m.	Challenge	duenias		Succeeded Evaluation is under review by the challenge admins.

That is why we eventually are using our 5 folds models to get average cross-validation results and hope that it will be close enough to the real test results. Another metric that was used in the challenge is NSD. The NSD metric is subjected to a parameter τ which we couldn't find on the challenge's website, and it should also vary between different image resolutions. That is why we will not present the NSD in our report.

First, the results of the ensemble ablation study are presented in the table below. These were obtained using a different number of models that have been trained on the same data fold (showing only data fold 1)

dataset	ensemble	fold 1	fold 2	fold 3	fold 4	fold 5	cross val
Heart	1 model	91.2	91.6	89.3	91.7	90.1	90.78
	2 models	91.5	92.1	90	91.8	90.9	91.26
	4 models	91.4	92.3	89.8	92	91.2	91.34
Hippocampus average dice	1 model	88	89.6	88.3	88.5	88.4	88.56
	2 models	88.3	89.8	88.5	88.8	88.5	88.78
	4 models	88.5	90.1	88.4	88.9	88.7	88.92

There is a significant increment of the scores averaging the soft prediction of 2 models and another, a bit less powerful, using 4 models and not 2.

In the table below we can see the effect of the test time augmentations

dataset	apply test aug	fold 1	fold 2	fold 3	fold 4	fold 5	cross val
Heart	with aug	91.2	91.6	89.3	91.7	90.1	90.78
	without aug	91	91.8	89.6	91.5	90.3	90.84
Hippocampus average dice	with aug	88	89.6	88.3	88.5	88.4	88.56
	without aug	88.2	89.6	88.1	88.7	88.5	88.62

It is less significant than the ensemble effect but still results in a slightly better score.

In both ensemble models and test time augmentation, even though each fold is affected differently, the total cross validation is higher. Combining both ensembles of

4 models and the test time augmentations our final results are presented in the next table.

	fold 1	fold 2	fold 3	fold 4	fold 5	cross val
Heart	91.6	92.2	90.1	92.1	91.1	91.42
Hippocampus class 1	88.7	90.6	88.8	90	89.3	89.48
Hippocampus class 2	88.3	89.4	88.6	90.2	88.4	88.98

A link to the challenge's leaderboard (for scores and ranking observation):

https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/

Conclusions and summary

In this work, we presented the power of an organized and inclusive ablation study. We showed the advantages of using relevant augmentations in training and evaluation. In addition, the benefit of ensemble models was discussed and how a simple averaging between models' predictions can easily improve the results. Another major conclusion was the importance of the receptive field for each specific dataset and the intuition of how big it should be.

Dealing with this challenge, we better understand the importance of using cross validation to assess the results in cases when there is no test data to check on. Using cross validation may take longer training time and more work, however, the results of it are more accurate.

Having the memory limitation difficulty, we also learned a lot about the balances and compromises between model's capacity and memory limitations, especially in datasets of 3D images.

References

.1 Sharma, N. and L.M. Aggarwal, *Automated medical image segmentation techniques.* J Med Phys, 2010. **35**(1): p. 3-14.

.2 Patil, D.D. and S.G. Deore, *Medical image segmentation: a review.* International Journal of Computer Science and Mobile Computing, 2013. :(1)2p. 22-27.

.3 Zhang, P., et al., *Domain Adaptation for Medical Image Segmentation: A Meta-Learning Method.* J Imaging, 2021. **7**(2).

.4 Hesamian, M.H., et al., *Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges.* J Digit Imaging, 2019. **32**(4): p. 582-596.

.5 Razzak, M.I., S. Naz, and A. Zaib, *Deep learning for medical image processing: Overview, challenges and the future.* Classification in BioApps, 2018: p. 323-350.

.6 Perslev, M., et al. *One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019. Springer.

.7 Antonelli, M., et al., *The Medical Segmentation Decathlon.* Nat Commun, .2022 :(1)13p. 4128.

.8 Isensee, F., et al., *nnu-net: Self-adapting framework for u-net-based medical image segmentation.* arXiv preprint arXiv:1809.10486, 2018.

.9 Venes, D., *Taber's cyclopedic medical dictionary*. 2017: FA Davis.

.10 Kresh, J.Y. and J.A. Armour, *The heart as a self-regulating system: integration of homeodynamic mechanisms.* Technol Health Care, 1997. **5**(1-2): p. 159-69.

.11 Mishra, S. and V. Upadhyay, *A Mathematical Study Of Two Phase (One Phase Is Newtonian And Other Is Non-Newtonian) Coronary Blood Flow In Venules Using Herschel–Bulkley Model During Angina.*

.12 Pirruccello, J.P., et al., *Genetic analysis of right heart structure and function in 40,000 people.* Nat Genet, 2022. **54**(6): p. 792-803.

.13 Torrent-Guasp, F., et al., *Towards new understanding of the heart structure and function.* Eur J Cardiothorac Surg, 2005. **27**(2): p. 191-201.

.14 Tobon-Gomez, C., et al., *Benchmark for Algorithms Segmenting the Left Atrium From 3D CT and MRI Datasets.* IEEE Trans Med Imaging, 2015. **34**(7): p. 1460.1473-

.15 Ecabert, O., et al., *Segmentation of the heart and great vessels in CT images using a model-based adaptation framework.* Med Image Anal, 2011. **15**(6): p. 863-76.

.16 Simpson, A.L., et al., *A large annotated medical image dataset for the development and evaluation of segmentation algorithms.* arXiv preprint arXiv:1902.09063, 2019.

.17 Anand, K.S. and V. Dhikav, *Hippocampus in health and disease: An overview.* Ann Indian Acad Neurol, 2012. **15**(4): p. 239-46.

.18 Soudry, Y., et al., *Olfactory system and emotion: common substrates.* Eur Ann Otorhinolaryngol Head Neck Dis, 2011. **128**(1): p. 18-23.

.19 Kuruba, R., B. Hattiangady, and A.K. Shetty, *Hippocampal neurogenesis and neural stem cells in temporal lobe epilepsy.* Epilepsy Behav, 2009. **14 Suppl 1**: p. 6.5-73

.20 Cenquizca, L.A. and L.W. Swanson, *Spatial organization of direct hippocampal field CA1 axonal projections to the rest of the cerebral cortex.* Brain Res Rev, 2007. **56**(1): p. 1-26.

.21 Clark, R.E., N.J. Broadbent, and L.R. Squire, *Hippocampus and remote spatial memory in rats.* Hippocampus, 2005. **15**(2): p. 260-72.

.22 Andersen, P., et al., *The hippocampus book*. 2006: Oxford university press.

.23 First, M.B., et al., *Structured clinical interview for DSM-IV-TR axis I disorders, research version, patient edition*. 2002, SCID-I/P New York, NY, USA:.

.24 Woolard, A.A. and S. Heckers, *Anatomical and functional correlates of human hippocampal volume asymmetry.* Psychiatry Res, 2012. **201**(1): p. 48-53.

.25 Pruessner, J.C., et al., *Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories.* Cereb Cortex, 2000. **10**(4): p. 433-42.
.26 Amaral, D.G. and M.P. Witter, *The three-dimensional organization of the hippocampal formation: a review of anatomical data.* Neuroscience, 1989. **31**(3): p. 571-91.
.27 Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
.28 Kendall, A., V. Badrinarayanan, and R. Cipolla, *Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding.* arXiv preprint arXiv:1511.02680, 2015.

29. Texas Heart Institute. (n.d). *Heart* Anatomy. Retrieved July 30, 2022, from https://www.texasheart.org/heart-health/heart-information-center/topics/heart-anatomy/

30. The Healthline Editorial Team (2018, January 22). *Left Atrium*. Healthline. https://www.healthline.com/human-body-maps/left-atrium#1