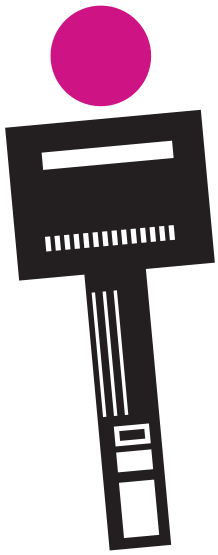

RAM ROBOTICS AND MECHATRONICS

DEEP LEARNING-BASED CATHETER SEGMENTATION IN FLUOROSCOPIC VIDEOS FROM HEAD AND NECK CANCER PATIENTS

A.H. (Afzal) Manik

BSC ASSIGNMENT

Committee:
dr. F.J. Siepel
M.M. Rocha
A. Briassouli, Ph.D

June, 2025

045RaM2025
Robotics and Mechatronics
EEMCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | TECHMED CENTRE UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Contents

Chapter 1

Introduction

1.1 Clinical background

1.1.1 Disease context

Head and neck cancer (HNC) is a blanket term used to refer to various malignancies that occur in the head and neck region, 90% of which are squamous cell carcinomas [1]. This type of cancer is commonly caused by tobacco and alcohol usage [2]. Human papillomavirus has also recently been identified as a risk factor, but to a lesser extent [3]. With the decrease in tobacco usage globally, the rate of HNC cases has slowly started declining, showing a decrease of 0.22% per year from 2002 to 2012 for instance in the US [4]. Nevertheless HNC is still the sixth most common type of cancer worldwide [5], reaffirming that research into diagnostics and treatment around this illness is still of great importance.

A common long term complication caused by HNC and its treatment is swallowing/deglutition difficulties from the mouth to the esophagus. A study by Garcia-Peris et al. showed that oropharyngeal dysphagia (OD), as it's called, was present in 50.6% of the patients studied [6]. This ailment is not only detrimental for the nutrition and hydration of the patient [7] [8], but it also brings about other issues, such as aspiration pneumonia and chest infections [9]. These factors decrease quality of life at best and increase risk of mortality at worst, making it valuable to examine the mechanisms and aetiology of the complication.

1.1.2 Current diagnostics

Videofluoroscopic Swallow Study (VFSS) is one of the main ways to medically examine OD and has generally been regarded as one the golden standards [10]. In VFSS a patient is given a bolus loaded with a contrast agent. During swallowing, an x-ray device is used to capture a fluoroscopic image of the swallowing act [11] [12]. This footage is then used by the examiner to characterize deglutition pathology [13]. An example of a VFSS frame can be seen in Figure 1.1.

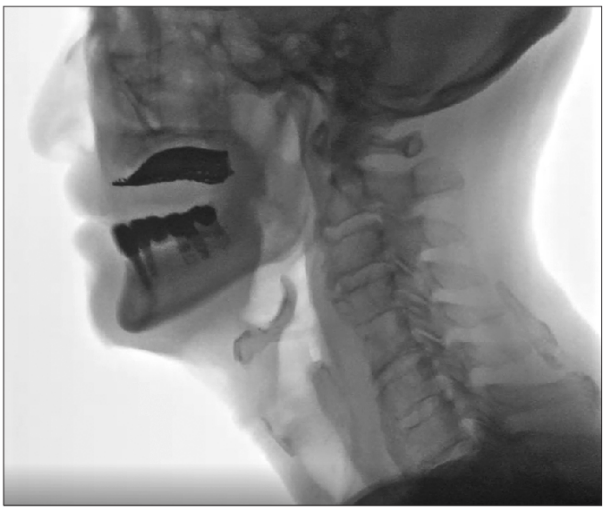

Figure 1.1: Example of an image taken in a videofluoroscopic swallow study (VFSS) [14].

Another common method of medical imaging is Fiberoptic Endoscopic Evaluation of Swallowing (FEES) [15]. FEES evaluates the process by inserting an endoscope through the nose into the throat to examine the movement and anatomy of the pharynx and larynx [16]. FEES is cheap, can be performed at bedside and does not contribute to radiation exposure [15] [17]. However, it can only visualize the surface structures of the throat. It has also been shown that it is not as good at detecting aspiration events as VFSS, suggesting that supplementary VFSS after FEES might be useful [18] [10].

Instrumental techniques using pressure data have become increasingly popular to investigate swallowing in the last decade [9]. One such technique is High Resolution Impedance Manometry (HRIM). In this method a catheter with pressure sensors and electrodes is inserted into the throat. The technique combines manometry to measure contractile activity with electrical impedance measurements to track bolus movement during swallowing [19] [20]. Although no visual information is collected, the acquired data can be used to thoroughly investigate and analyze deglutition behavior and possible pathophysiology in an objective manner [21].

To get a more accurate assessment of OD, HRIM can be used in conjunction with VFSS, especially in the HNC patient population [22]. VFSS relies on judgment made by clinicians, which can be subjective. HRIM on the other hand is usually more objective, but due to reduced pressure observed in measurements made on HNC patients, it is more challenging to annotate the manometric regions. This reduces its objectivity [23]. Thus, combining both techniques overcomes their individual shortcomings [24].

To integrate HRIM with VFSS one must find a method to match the data from both sources. Manual annotation of the catheter and sensors is an option, but time consuming and subject to inter-rater variability. Hence, an automatic process is more preferable. A possible way to automate the task is through a knowledge-based approach using computer vision techniques. Deep learning (DL) is another approach to this problem. Advances in hardware, improvements in the methodology and a general increase in available data have driven improvement in the technique, making it an increasingly popular tool for medical imaging [25] [26]. This development suggests that the use of neural networks is a promising alternative to the conventional methods.

1.2 Research goal

To enable the objective characterization of OD in the HNC population. the first obstacle to address is overcoming the difficulties of integrating HRIM and VFSS in this patient population . A way to achieve this is to automatically detect the manometric regions in the VFSS and translate them into the HRIM data. To do this, we first need to accurately detect the HRIM catheter in VFSS videos. Recognizing what part of the image is the catheter is the fundamental basis of the task. The goal of this project is to tackle this problem through a deep learning approach, with the main aim being to create a model

that is proficient at accurately detecting a catheter in fluoroscopy footage. Furthermore, the project will endeavor to analyze what model characteristics (such as hyperparameters) are most influential and beneficial for model creation for this task. The hypothesis is that the deep learning approach will indeed prove fruitful and result in an effective methodology to detect a catheter in fluoroscopy footage.

1.3 Technical background and related works

Deep learning (DL) is a branch of machine learning inspired by biological neural networks. DL models consist of mathematical functions called artificial neurons interconnected and layered to form an artificial neural network (ANN). These networks process data in a way determined by three key components. Their so-called architecture, which includes things like how the different layers are connected, what kind of functions the neurons apply and how the different layers of the network are connected.The parameters learned after training, which will be elaborated on shortly. And finally obviously the input data. The parameters of a model determine the strength of the connections between the neurons, thus influencing the way data gets transformed as it passes through the network. The model acquires its parameters by training, which is where the learning in deep learning comes into play. The networks are given large amounts of data. This data is passed through the network and used to iteratively adjust the parameters toward values that improve performance on the task [27] [28].

As mentioned before, the use of deep learning for medical imaging has risen considerably in modern times, with the most popular models at the moment being Convolutional Neural Networks (CNN). These models use convolution kernels, an already well known concept in computer vision techniques, to detect patterns. Traditional image processing techniques based on convolution kernels were often hand-crafted and tuned, which could lead to shallow and suboptimal models. CNN's on the other hand are able to tune and refine themselves due to being a DL network, permitting models with higher depth and better tuned parameters [29].
Networks of the transformer architecture type are a relatively new innovation in the DL field and are seeing a rapid rise in popularity in the medical imaging field in recent times. Models of this type were originally designed for natural language processing, but their inherent adaptability turned out to also make them greatly effective in various other tasks, including ones in medical imaging [30].
Nevertheless, the architecture that dominates the field is U-net. U-net is a type of CNN showing outstanding performance in medical image segmentation. The network consists of an encoder, responsible for feature extraction through a series of convolutional layers and downsampling, a decoder for rebuilding the image to its original resolution, and finally, skip connections for retention of image details lost in downsampling. Despite the advancements made in other architecture types, U-net and its variants remain the most popular for medical imaging tasks [31] [32].

DL techniques have already been proposed previously for VFSS analysis tasks. In 2020 Caliskan et al. for instance used Mask-RCNN to detect and track the bolus during a swallowing event in VFSS footage [33]. Another use case found for DL in VFSS analysis is the automatic detection of aspiration events. These events are challenging to identify due to often being brief and only occurring over a few frames. By using DL to detect the aspiration events, as done with CNN's in [34] and [35], clinicians are able to shift their focus on diagnosis and interpretation of the detected events.

Catheter detection with DL has been explored in various other research papers, establishing a strong proof of concept. Models proposed include ones that detect thin intravenous catheters, several catheter types in neonates and even catheters used in cerebral angiography [36] [37] [38].

Chapter 2

Methods

This chapter will detail how the research goal of this project was achieved.
The project is divided in two phases. In the first phase a base model was constructed. Its architecture was chosen based on a review of the literature, whilst the optimal combination of key hyperparameters was determined through a grid search. In the second phase two main extensions were made: augmentations were added to the data preparation techniques; and, the base model was modified by the addition of attention mechanisms in an attempt to further improve the segmentation results.

2.1 Base model

2.1.1 Dataset and data preparation

The available dataset consisted of approximately 1800 frames from VFSS footage of 16 patients at the Netherlands Cancer Institute (NKI), along with their ground-truth (GT) masks. The images had varying resolutions in the order of 1000x1000. All images were taken from a lateral perspective with patients positioned upright and facing the viewer's left. Some sample images of the dataset can be seen in figure 2.1.

The applied data split method was a so-called single train-val-test split. "Single" refers to the fact that the subsets were explicitly not mixed, meaning that for instance the images in the validation subset were never used for training. This method divides the dataset in three subsets:

1. Training set: used to train the model. This subset contained ≈800 images that were randomly sampled from the videos.

2. Validation set: used to evaluate the performance of the model at the end of each epoch, and the performance of the best model at the end of a run. This subset contained ≈200 images that were also randomly sampled from the videos.

3. Test set: Used to evaluate the performance of the best model at the end of a run. This subset contained all the videoframes of 4 complete swallow videos, adding up to ≈850 images.

Since there are only two classes in this segmentation task (catheter and background), with the catheter size being small compared to the background, this specific task is referred to as a binary semantic segmentation task with a large class imbalance. For these types of tasks retaining the finer details of the smaller class is crucial, meaning that it is beneficial to use the highest resolution possible for the images. Considering this and among other things the memory and speed constraints of the available hardware, a resolution of 640x640 was chosen. Each image was transformed to this square resolution through resizing. The interpolation utilized for resizing however introduced intermediate pixel values for the binary masks. To restore their binary nature, the masks were rebinarized with simple thresholding, rounding the intermediate values to their nearest binary value.
Shuffling was employed to achieve better generalization of the model. The training set was shuffled at the end of each epoch, the validation and testing data however remained unshuffled, ensuring accurate metric scores.

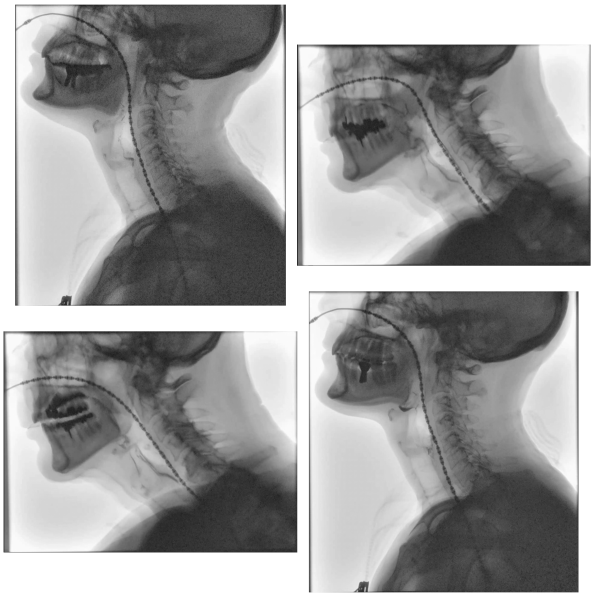

Figure 2.1: Randomly sampled raw images from the dataset. The images have varying resolutions and the patients always face the viewer's left.

2.1.2 Model architecture

The base architecture used was ResUNet, a U-net variant based on ResNet.
As mentioned before in section 1.3, U-net has been specifically designed to be used for biomedical imaging applications and is widely popular in the field [39], making it a reliable architecture to expand on.
Higher layer depth has been known to facilitate faster convergence, however the vanishing gradient problem arises as network depth increases caused by the loss of feature identities [40]. To overcome this issue a new architecture was proposed utilizing residual learning: the ResNet. This network passes the feature maps to deeper layers through feedforward connections, helping to preserve the feature maps better. ResUNet combines the strength of both architectures. For the network, the basic building blocks of U-net are transformed by adding batch normalization and identity mapping (see Figure 2.2), but the basic U-shape is left unchanged as seen in Figure 2.3. Due to the convolution layers and downsampling, this architecture demands images with sides divisible by 16 as input.

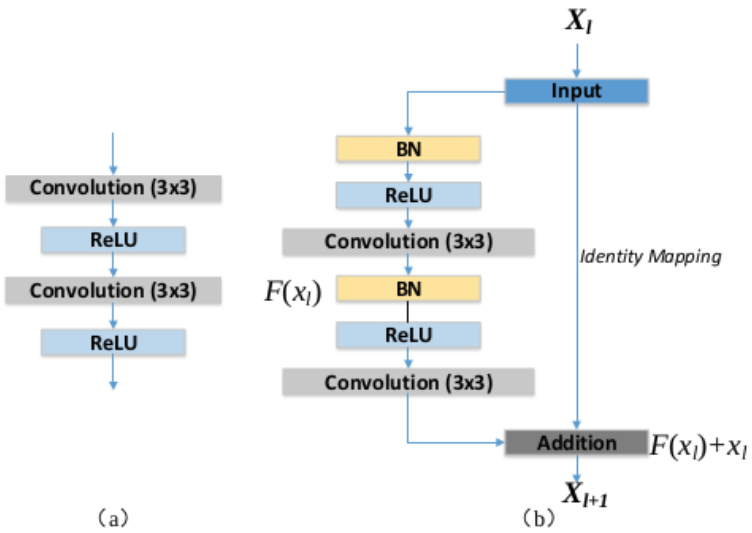

Figure 2.2: Basic building blocks: (a) base unit of U-net, (b) base unit of ResUNet [41].

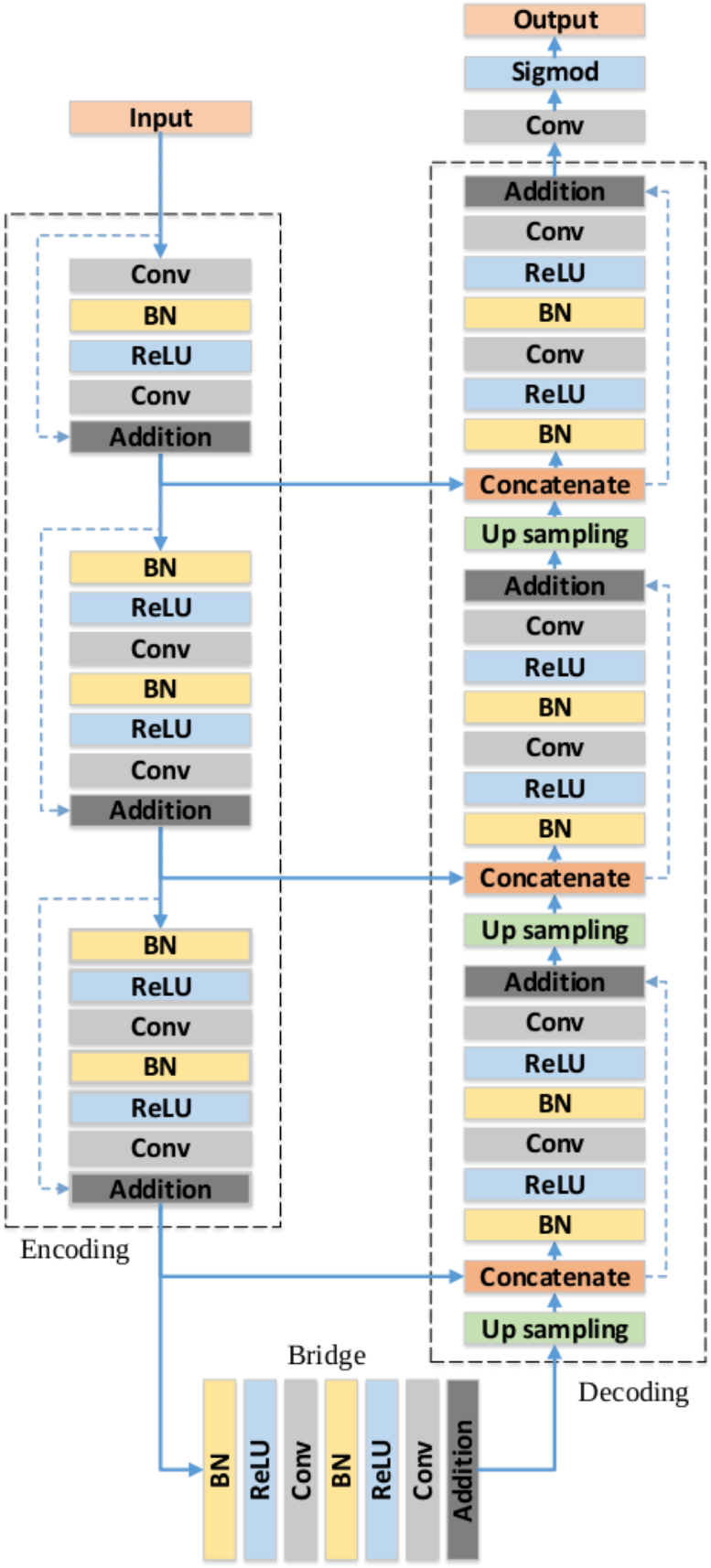

Figure 2.3: Schematic structure of the ResUNet architecture [41].

2.1.3 Training procedure and hyperparameter tuning

A grid search was conducted to analyze the effects of different hyperparameters and find the optimal combination. The three hyperparameters that were tuned were: optimizer, loss function and learning rate. Each had three options described in the subsections below. Giving a total of $3^3 = 27$ combinations of hyperparameters run. Hyperparameters that were kept constant were:

1. Batch size = 4. A small batch size is good for better generalization, but slows down training. 4 was chosen to strike a balance between generalization and training time.
2. Image size = 640x640. As mentioned before, the chosen architecture requires input images to have specific shapes. Along with the considerations about high resolution and available hardware, a resolution of 640x640 was chosen.
3. Epochs=50. 50 epochs gives the training curve enough room to plateau whilst still being small enough for minimal risk of overfitting. It should be noted that the model used for performance evaluation is not the model of the last epoch after training, but instead the model of the epoch with the best performance based on the validation set.

Optimizers

The optimizers used were: Stochastic Gradient Descent (SGD), Adam and AdamW. SGD is a simple gradient-based optimizer known for good generalization but relatively slow convergence [42]. Adam is an adaptive optimizer designed to converge quicker by adapting its learning rate for each parameter. This optimizer is widely used for medical image segmentation and is known to produce good results [43]. Finally, AdamW is a modification of Adam introduced relatively recently. This algorithm aimed to be a refinement of Adam, with the authors showing that the optimizer outperforms standard Adam [44].

Loss functions

The loss functions used were: Dice loss, Tversky loss and Weighted Binary Cross Entropy loss (WBCE). Dice loss (see equation 2.1) is based around the Dice Similarity coefficient. It is the most commonly used loss function for medical image segmentation and can serve as a strong baseline. This loss belongs to a family of loss functions called overlap-based losses that have shown to be useful in tasks with large class imbalances [45] [46].
Tversky loss (see equation 2.2) is another overlap based loss. Its a generalization of Dice loss, allowing more flexibility by allowing you to tune how heavily false negatives and false positives in your prediction image get penalized. The version of Tversky loss used had a weight of 0.1 for the false positives and 0.9 for the false negatives.
Finally, Weighted Binary Cross Entropy (WBCE) (see equation 2.3) loss was also selected to heavily target the large class imbalance in the task. This loss is a pixel-wise classification loss with weighting of the two classes. The weights used in the WBCE loss of this project were a catheter weight of 1 and a background weight of 0.001.

$$\text{Dice loss} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \tag{2.1}$$

$$\text{Tversky loss} = 1 - \frac{\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i \hat{y}_i + \alpha \sum_{i=1}^{N} (1 - y_i)\hat{y}_i + \beta \sum_{i=1}^{N} y_i (1 - \hat{y}_i) + \epsilon} \tag{2.2}$$

$$\text{WBCE loss} = \frac{1}{N} \sum_{i=1}^{N} \left[w_i \cdot (-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)) \right] \tag{2.3}$$

	$y_i \in \{0, 1\}$	Ground truth label (binary)
	$\hat{y}_i \in [0, 1]$	Predicted probability
	$\alpha = 0.1$, $\beta = 0.9$	Tversky coefficients
where:	$w_i = y_i \cdot w_{\text{fg}} + (1 - y_i) \cdot w_{\text{bg}}$	Weighting for WBCE
	$w_{\text{fg}} = 1.0$, $w_{\text{bg}} = 0.001$	Foreground and background weights
	$\epsilon = 10^{-6}$	Smoothing constant to avoid division by zero
	N	Total number of pixels

Learning rates

The learning rates that were tested were: 1e-2, 1e-3 and 1e-4. 1e-3 is a default learning rate used for Adam, and served as a baseline. Since ideal learning rates can vary per optimizer, learning rates an order of magnitude 10 higher and lower were also used. 1e-2 is a more aggressive learning rate, which risks instability, but reduces convergence time and the chance of getting caught in local minima. This higher learning rate is theorized to be especially compatible with SGD. Finally, a learning rate of 1e-4 is also included to analyze the effect of reducing the learning rate instead.

2.1.4 Performance evaluation

Due to the large class imbalance in this segmentation task, straightforward general per pixel prediction accuracy will be heavily biased. The model could for instance classify all the pixels as background, but since the catheter pixels are so few number, the resulting accuracy will still be high. Only considering how much of the catheter pixels are detected is a better criterion called recall (see equation 2.4). This metric however fails to consider False Positives (FP). The metric opted for instead is the Dice Similarity Coefficient (DSC) also called the Dice score . With the incorporation of false positives into the formula for DSC (see equation 2.5), this coefficient gives a better measure of how good the model is segmenting overall. Intersection over Union (IoU) (see equation 2.6) is another commonly used metric that is similar to DSC, but slightly more stringent. It is included to facilitate easier comparison with results reported in the literature, where IoU is sometimes preferred.

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$DSC = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.5}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{2.6}$$

The final performance evaluation of each model inference was performed over each individual image in the validation set, later referred to as *Random Frames.*, and the test set, later referred to as *Full Videos.* For the performance evaluation the images were not resized. Instead, the images were zero-padded to a square shape having sides equal to a multiple of 16, matching the network's required input size

2.2 Extensions

2.2.1 Augmentations

Due to the limited variation in the current dataset, namely the consistent orientation of the catheter from left-to-right, as seen in figure 2.1, and the similar nature of radiation dosages, we used data augmentation techniques. Augmentations artificially increase the diversity of the dataset, facilitating generalization, through transformations of the dataset. The augmentations however must remain anatomically and clinically realistic to be useful to model training.

Horizontal flipping is a straightforward augmentation which simply flips the image across the y-axis. VFSS fluoroscopy images are usually captured in a lateral view, but may be taken from either the left side or the right side, thus horizontal flipping preserves realism. This augmentation was applied to the training, validation and the evaluation sets.

Small angle rotations ($< |10°|$) emulate the slight variations of patient positioning, such as minor head tilts to the front or back. This augmentation was only applied to the training set.

Finally, Gaussian noise can be added to mimic lower image quality in lower-dose X-rays. This augmentation will result in a more grainy texture in the images. Like small angle rotations, this augmentation was also only applied to the training set.

The augmentations were randomly applied to the datasets with a probability of 50%. The training set augmentations were reapplied randomly each epoch, whilst the augmentations for the validation set and evaluation sets were non-randomized per model run. An example of these augmentation can be seen in Figure 2.4

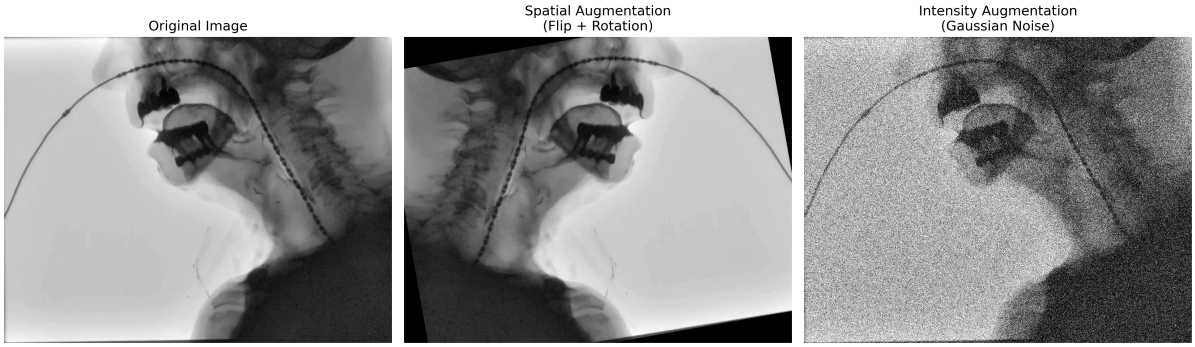

Figure 2.4: Example of the applied image augmentations.

2.2.2 Attention mechanisms

Incorporating attention mechanisms into models has shown to be greatly beneficial for image classification tasks. In the case of ResUNet, a version with added Attention gates has already been proposed before by Ehab et al. and has proven to be more effective than U-net and ResUNet at tackling tasks with large class imbalances [47]. The architecture introduces attention blocks in the skip connections of the U-net structure. These attention blocks use gating signals originating from a lower layer and the identity map from the encoder to compute attention weights, which in turn are used for determining what part of the image is most relevant to focus on. This mechanism is called spatial attention. Figure 2.5 shows the structure of the architecture.

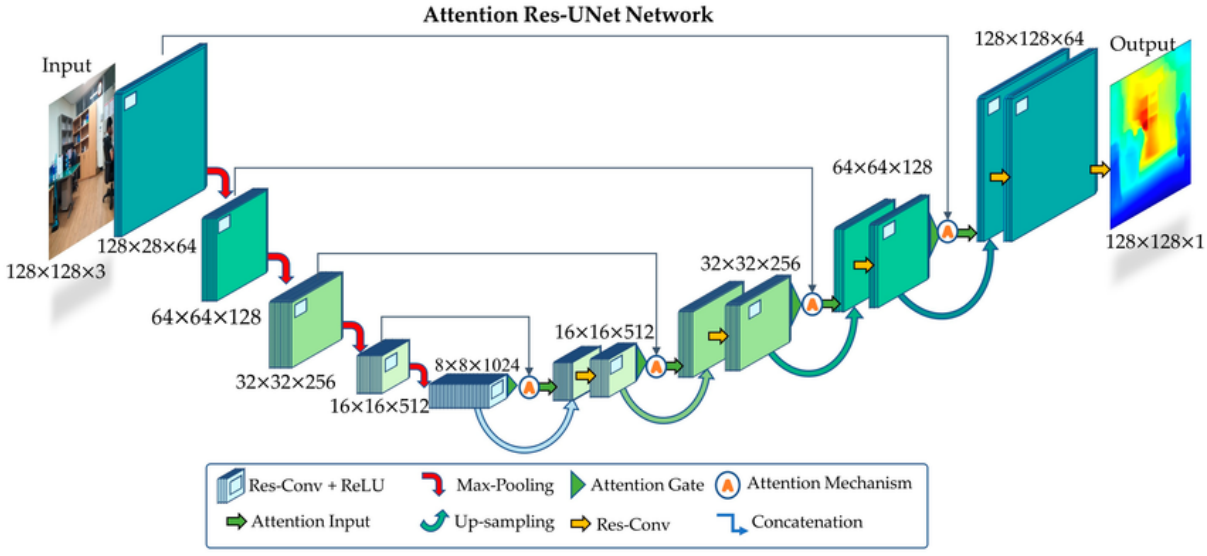

Figure 2.5: Schematic structure of the attention ResUNet architecture [48]

Chapter 3

Results

3.1 Base model

In the section the results of the grid search will be presented, as well as performance of the best model.

3.1.1 Grid search

The training and validation scores of every model was recorded. For readability's sake, the training and validation curves of only the top 12 models are displayed in figures 3.1 and 3.2. The curves all exhibit a steep rise at the start, followed by a steady plateauing.

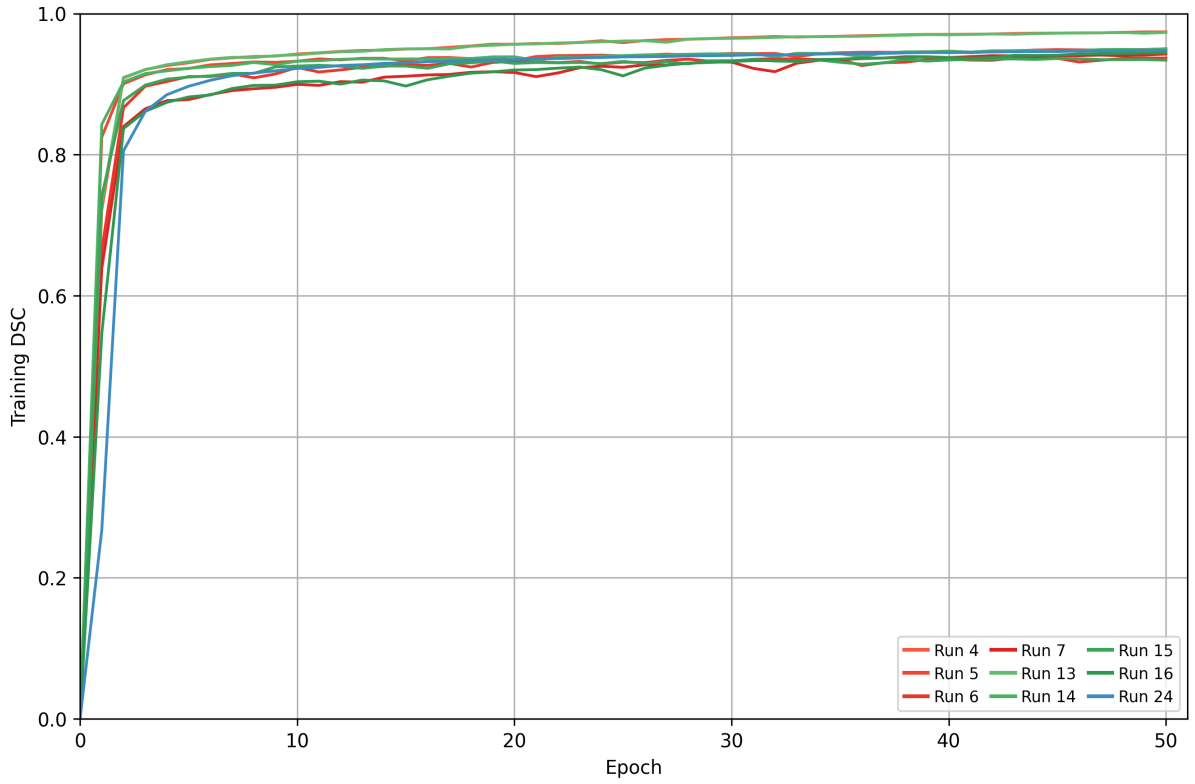

Figure 3.1: Train Dice scores over the course of training for the top 9 runs. The models with adam, adamW and SGD as optimizer are shades of red, green and blue respectively.

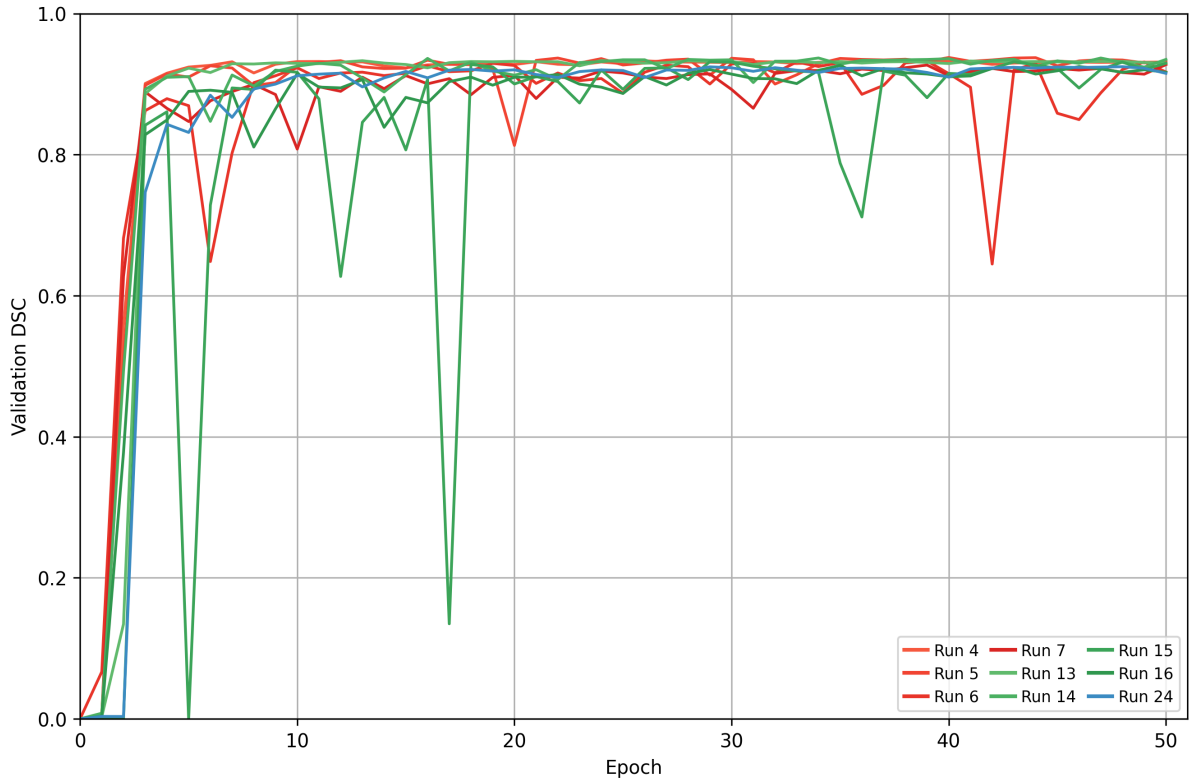

Figure 3.2: Validation Dice scores over the course of training for the top 9 runs. The models with adam, adamW and SGD as optimizer are shades of red, green and blue respectively.

Table I displays the results of the performed grid search, showing the metric scores for each hyperparameter combination. Since The *Full Videos* dataset has low variety, more value was given to the scores on the *Random Frames* dataset.

Table I: Evaluation results of the 27 models run in the grid search, with the scores rounded to 2 decimal points. The best scores for both evaluation sets (*Random* Frames and *Full Videos*) are highlighted in bold. For this grid search the batch size, image size and number of epochs were set to 4, 640x640 and 50 respectively

Run #	**Hyperparameters**			**Results**			
	Optimizer	**Loss**	**LR**	**Random Frames**		**Full Videos**	
				Dice	**IoU**	**Dice**	**IoU**
1	Adam	Dice	0.0001	0.82	0.71	0.85	0.74
2	Adam	Dice	0.001	0.82	0.72	0.88	0.78
3	Adam	Dice	0.01	0.80	0.68	0.81	0.69
4	Adam	Tversky	0.0001	0.77	0.65	0.75	0.61
5	Adam	Tversky	0.001	0.72	0.59	0.79	0.67
6	Adam	Tversky	0.01	0.68	0.56	0.73	0.60
7	Adam	WBCE	0.0001	0.63	0.47	0.69	0.54
8	Adam	WBCE	0.001	0.51	0.35	0.56	0.40
9	Adam	WBCE	0.01	0.46	0.30	0.52	0.35
10	AdamW	Dice	0.0001	0.86	0.76	0.87	0.78
11	AdamW	Dice	0.001	**0.86**	**0.77**	0.87	0.77
12	AdamW	Dice	0.01	0.83	0.72	0.84	0.73
13	AdamW	Tversky	0.0001	0.83	0.72	**0.89**	**0.81**
14	AdamW	Tversky	0.001	0.69	0.55	0.72	0.58
15	AdamW	Tversky	0.01	0.69	0.57	0.70	0.58
16	AdamW	WBCE	0.0001	0.67	0.52	0.67	0.52
17	AdamW	WBCE	0.001	0.51	0.35	0.57	0.41
18	AdamW	WBCE	0.01	0.54	0.37	0.58	0.41
19	SGD	Dice	0.0001	0.06	0.03	0.06	0.03
20	SGD	Dice	0.001	0.81	0.69	0.82	0.70
21	SGD	Dice	0.01	0.83	0.72	0.86	0.75
22	SGD	Tversky	0.0001	0.42	0.28	0.46	0.30
23	SGD	Tversky	0.001	0.81	0.70	0.87	0.78
24	SGD	Tversky	0.01	0.83	0.72	0.88	0.79
25	SGD	WBCE	0.0001	0.04	0.02	0.03	0.02
26	SGD	WBCE	0.001	0.05	0.02	0.05	0.02
27	SGD	WBCE	0.01	0.11	0.06	0.12	0.06

3.1.2 Best model

Based on the performance scores shown in table I, the model from run 11 was chosen to proceed with. Figure 3.3 and 3.4 show histograms of the evaluation Dice scores of this model.

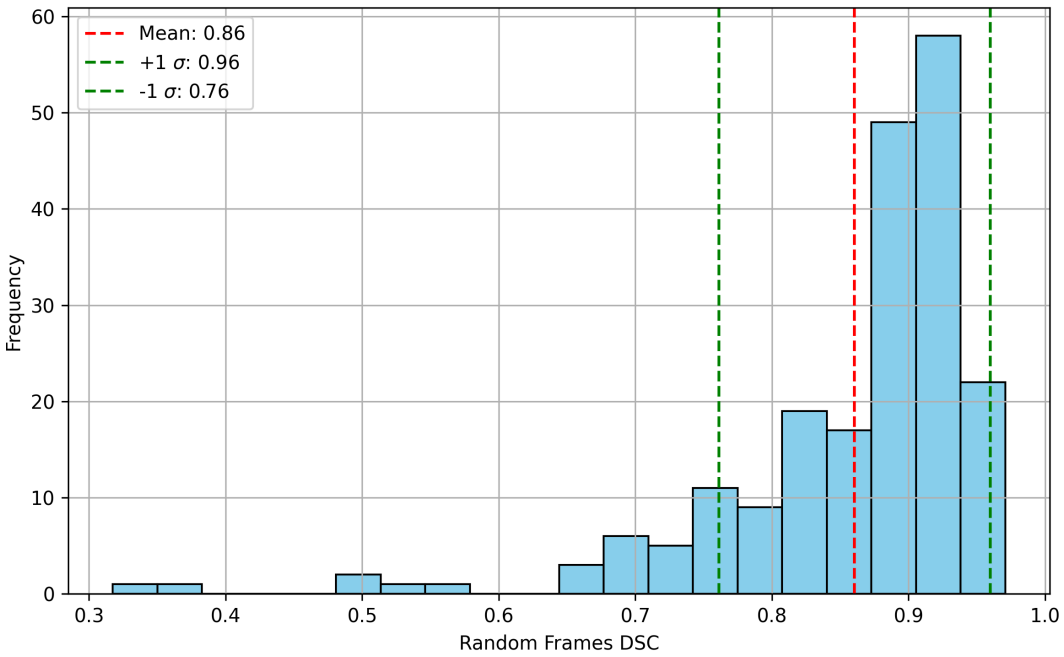

Figure 3.3: Dice score distribution of the final evaluation on the *Random Frames* dataset using the model from run 11

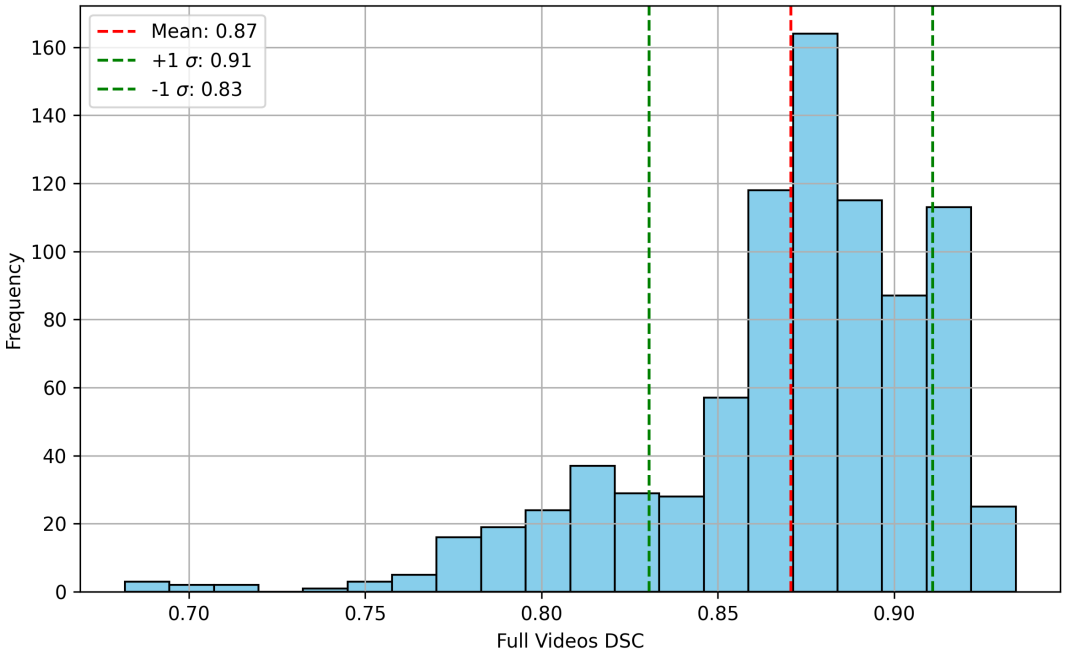

Figure 3.4: Dice score distribution of the final evaluation on the *Full Videos* dataset using the model from run 11

Figure 3.5 displays select sample segmentation masks overlaid on the original image, and compares them with the ground truth mask (also overlaid on the original) and the original image itself.

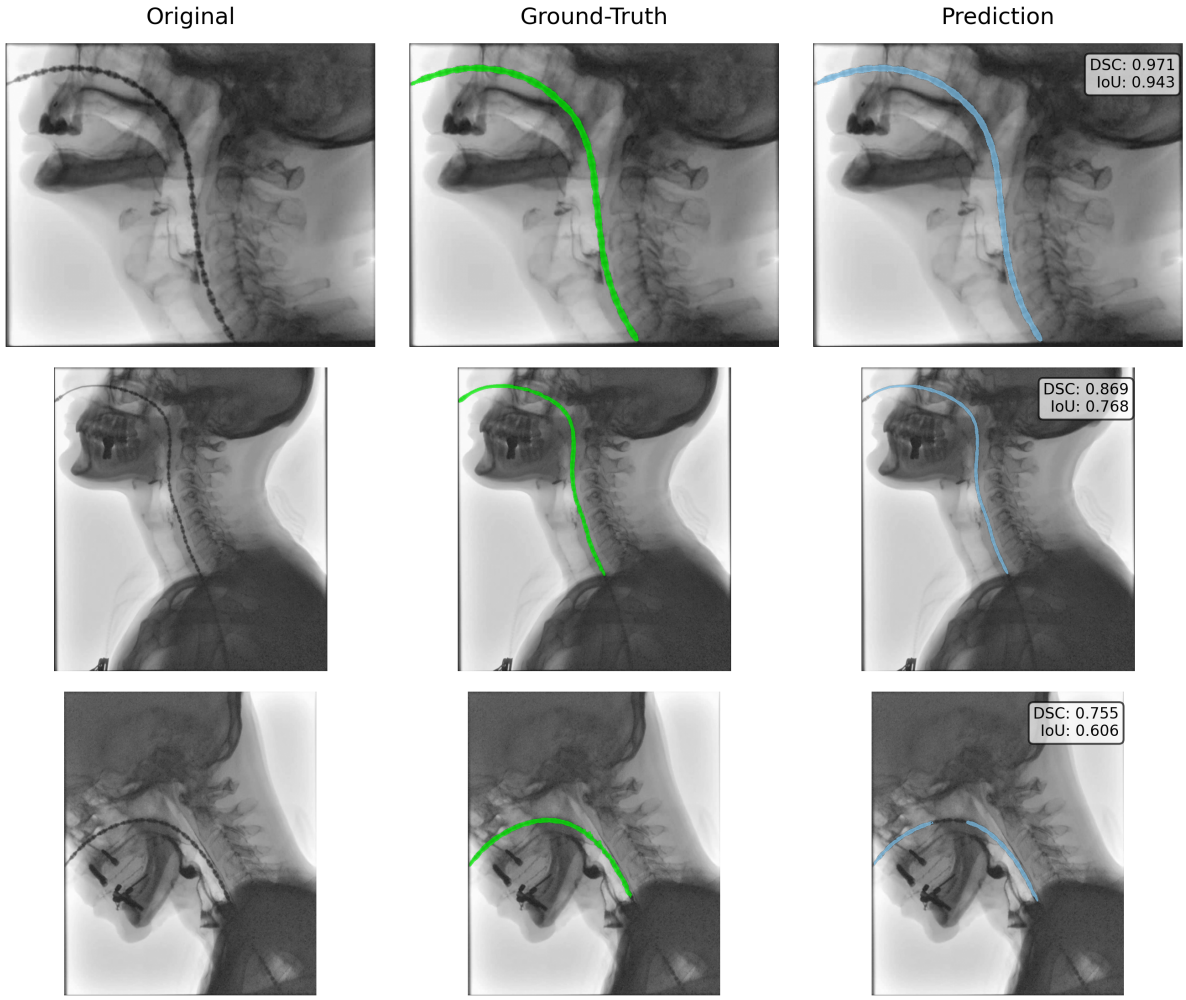

Figure 3.5: Comparison of the predicted mask generated by the best performing model. The first picture shows the original input image. In the second picture the ground truth mask is overlaid in green. The third one shows the predicted mask by the model overlaid in light-blue, along with the Dice score and IoU of the prediction. The three rows represent three levels of performance; "good", "typical" and "poor". The typical case is acquired by taking the prediction with a score closest to the average score. The good and poor case are found in a similar manner by taking the prediction with a score closest to $\pm 2\sigma$.

3.2 Extended model

For the extended model the best performing model from the grid search (run 11) was used as a foundation to build off of. All the hyperparameters were copied, except for the batch size. Due to the added layers in attention ResUNet, a batch size of 2 was used instead to stay within the memory limitations of the available hardware.

3.2.1 Augmentations

Table II shows the results of applying the augmentations to the different subsets of the data pool.

Table II: Effect of data augmentation on performance with scores rounded to three decimal points. The models were trained using the hyperparameters of run 11, with the exception of the batch size being equal to 2

Augmentations		**Results**			
Train Set	**Eval Set**	**Random Frames**		**Full Videos**	
		Dice	**IoU**	**Dice**	**IoU**
No	No	0.867	0.776	0.879	0.787
No	Yes	0.754	0.632	0.760	0.631
Yes	No	0.867	0.775	0.869	0.771
Yes	Yes	0.868	0.775	0.866	0.766

3.2.2 Attention mechanisms

In table III the performance results of 4 different scenarios are presented. In the first two scenarios base ResUNet is used as architecture, only differing from each other in whether augmentations were applied. Similarly, the last two scenarios also only differ in employment of augmentations, but use Attention ResUNet instead.

Table III: Comparison of model architectures and augmentation settings. Scores are rounded to three decimal points. All models were trained using consistent hyperparameters (based on Run 11, but batch size was 2 instead of 4) except for the model and augmentation scheme.

Model architecture	**Augmentations**	**Random Frames**		**Full Videos**	
		Dice	**IoU**	**Dice**	**IoU**
ResUNet	No	0.867	0.776	0.879	0.787
ResUNet	Yes	0.868	0.775	0.866	0.766
Attention ResUNet	No	0.860	0.766	0.874	0.780
Attention ResUNet	Yes	0.874	0.784	0.882	0.790

Similarly to the last section, figure 3.6 and 3.7 show the metric score distribution of the best performing model. Which in this case is the Attention ResUNet model based on run 11 with augmentations.

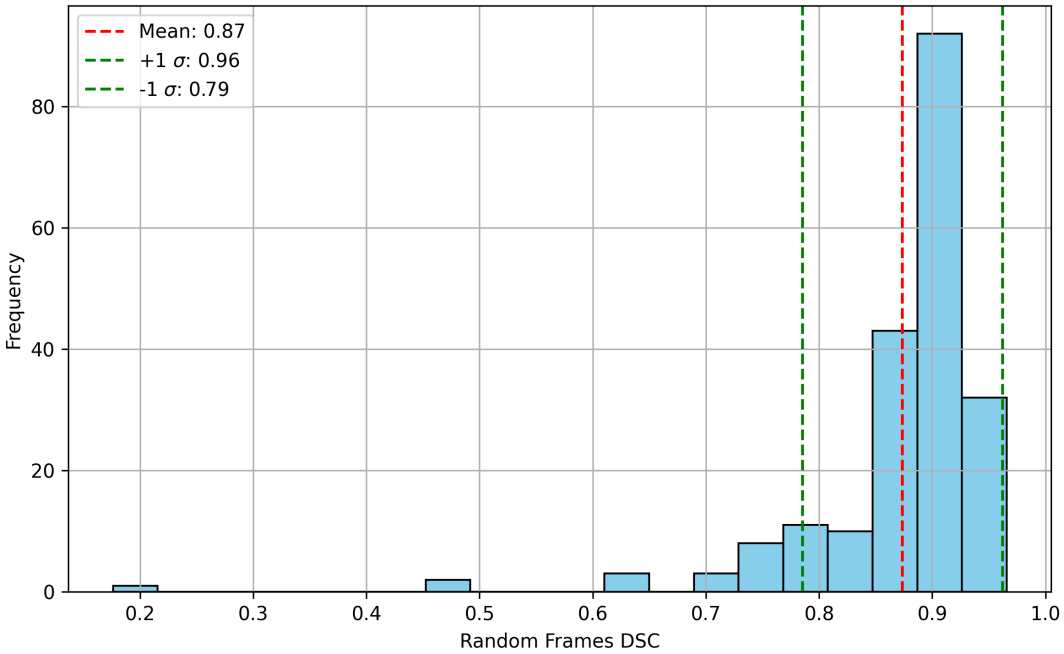

Figure 3.6: Dice score distribution of the final evaluation on the *Random Frames* dataset using augmentations and the Attention model based on run 11

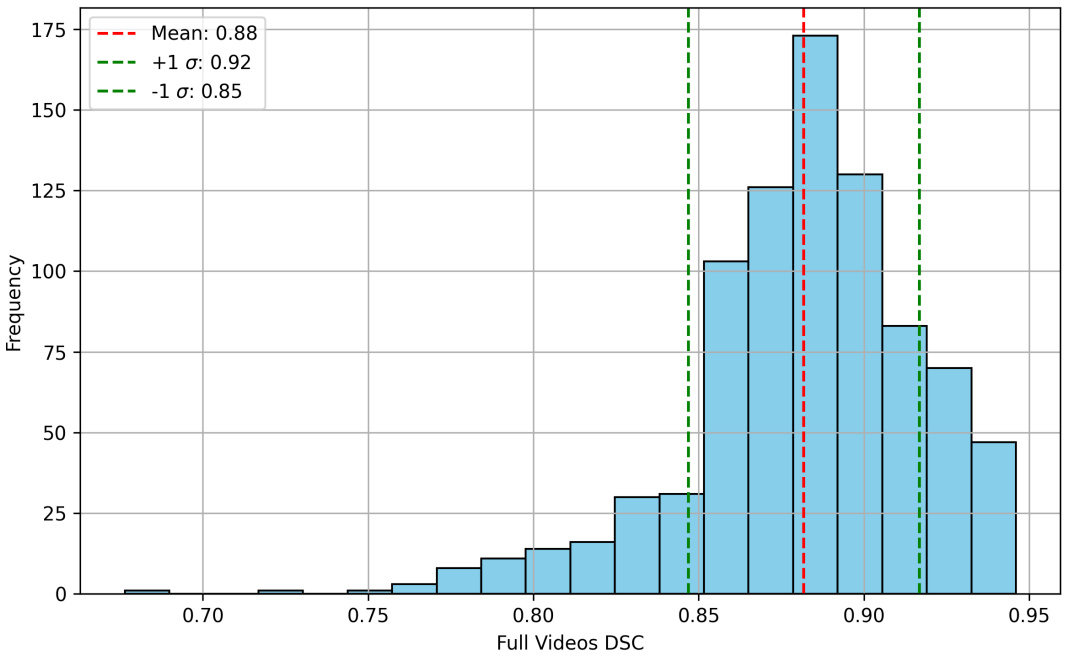

Figure 3.7: Dice score distribution of the final evaluation on the *Full Videos* dataset using augmentations and the Attention model based on run 11

Figure 3.8 shows the attention maps of some sample images used for inference. The number at the end of the map indicates how deep it is made in the architecture, with 4 being the deepest.

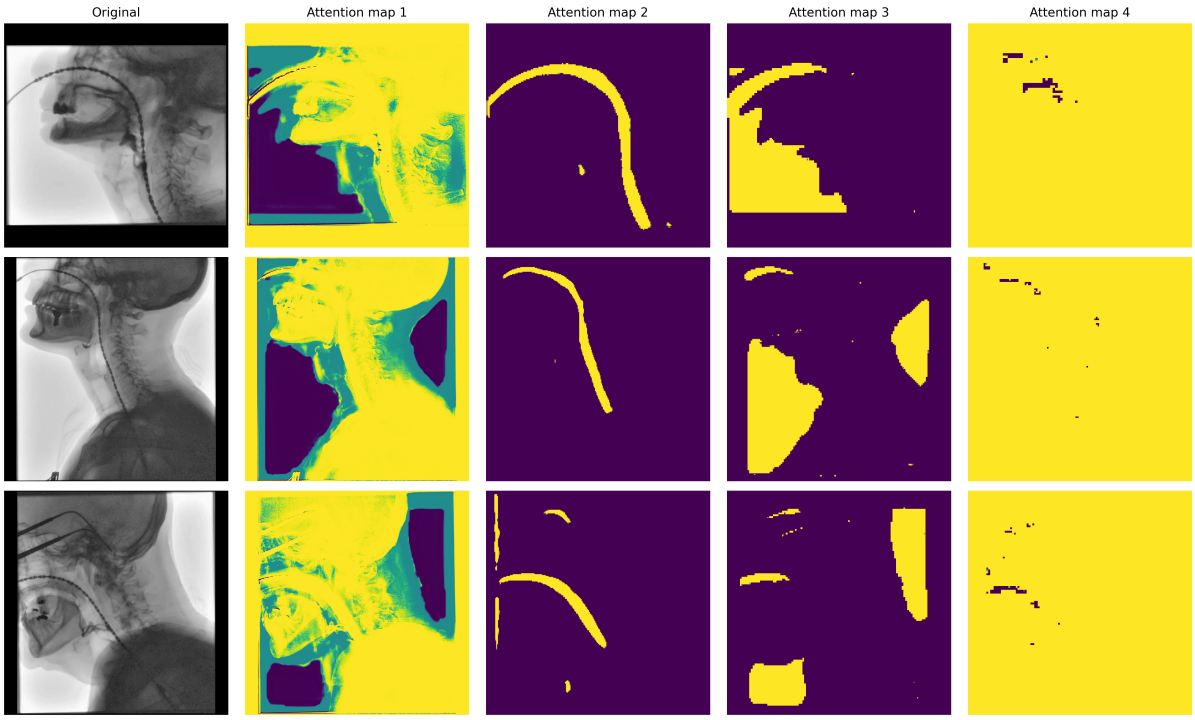

Figure 3.8: Attention maps generated by the Attention ResUNet model for a evaluation on select images from the evaluation sets. similarly to the inference images from figure 3.5, the inference quality on the images decrease going from the top row to the bottom, with the middle row showing a raw image and its maps of a typical performance.

Chapter 4

Discussion

4.1 Interpretation of findings

The grid search performed in this project resulted in some key findings. These findings are discussed below.

The training and validation curves seen in figures 3.1 and 3.2, show that the top model runs converged relatively quickly, often reaching their maximum training and validation Dice scores before the 15th epoch. Additionally, the high scores suggest that the vanishing gradient problem was not encountered during training. This outcome supports the notion that usage of ResUNet as architecture results in fast convergence whilst keeping metric scores high

From the results in table I it is apparent that the runs using adam or adamW outperforms SGD on average. It is however observed that the SGD models had high variation in their performance, with run 21 for instance having a *Random Frames* DSC of 0.83, whilst run 25 has a score of 0.04. This hints at the possibility that some SGD runs got stuck in local minima early in the training process. Additionally it is noted that the SGD runs strongly benefited from higher learning rates, suggesting that using even higher learning rates might result in better scores.
Another observation made is that Dice loss and Tversky loss consistently outperformed WBCE loss. Reestablishing the fact that overlap-based losses are most effective for segmentation tasks with large class imbalances. Regarding the scores for the two evaluation sets, it is seen that the models are slightly worse at detecting the catheter in the *Random Frames* dataset than the *Full Videos* dataset, indicating that generalization could be improved.

Figure 3.4 shows that the best model exhibited good prediction performance with high consistency on the *Full Videos* dataset, with 95% of the predictions having a score between 0.79 and 0.95. Notably, even the lower scores weren't that bad, with the lowest score being 0.65
The scores seen in figure 3.3 however are less tightly grouped together and more skewed. The majority of predictions were still adequate, but the histogram shows a considerable amount of outliers far from the average score. The worst score being as low as 0.32. This again hints at issues in generalization.

Overall, The best model of the first phase demonstrated reliable catheter segmentation, with the typical Dice score being around 0.87. In the poor predictions, like the one seen in figure 3.5, it is noted that the model often struggles with catheter detection around the jawbone. This localized degradation around an area with higher radiation attenuation suggests the issue to be caused by lack of contrast.

By exclusively augmenting the evaluation sets, it is shown in table II that the model is poorly generalized. The *Random Frames* DSC went from 0.867 to 0.754 which is a decrease of almost 15%. By also augmenting the training set the model returned to its high metric performances. The scenario where only the training set was augmented saw no significant decrease in metric scores, indicating that augmentations do indeed help with generalization.

Table III shows that in both the cases of augmented and unaugmented dataset the addition of attention mechanisms did not increase the average metric scores in any meaningful way. In the histograms of the evaluation sets prediction scores seen in figures 3.6 and 3.7 it is however observed that the number of low value outliers are reduced, suggesting a slight increase in performance due to the usage of the attention ResUNet architecture.

The attention maps seen in figure 3.8 reveal that the model is properly focusing on relevant areas. In the first map, made in the most shallow attention layer, the model identifies the patient as an important structure. The second map focuses specifically on catheter like structures, by characterizing them as thin, linear and having high radiation attenuation. The third map directs the model's attention to the outside of the patient, presumably telling it where the entry point is of the catheter. Finally, the fourth map highlights everything except small high contrast areas very close to the catheter. This map is likely used to suppress problem areas that could otherwise be included in the segmentation map.

4.2 Limitations

Despite generating important findings, the methodology used in this project had some limitations that should be acknowledged.

Resizing the input images was necessary to train the model, however making them square introduced slight geometric distortions. To maintain anatomical realism it is important to change the aspect ratio as little as possible. It would thus be better if the input images were zero-padded to a set resolution or resized dynamically per image to stay close to their original aspect ratios.

Retraining models using identical hyperparameters from the grid search yielded similar but non-identical performance metrics, demonstrating the inherent variability in neural network initialization and training dynamics. For this reason it could be argued that the results of the grid search, especially the ones that are close to each other, are not reliable for comparison.

The last point revolves around comparability of the data from the first phase and the second phase of the project. In the second phase of the project when the augmentations were applied, a batch size of 2 was used instead of 4 due to memory limitations. This change, although small, does decrease comparability between the recorded results.

4.3 Recommendations

Since it was noted that the models in some cases especially struggled with low contrast areas such as around the jawbone, adding contrast enhancing augmentations to the augmentation pipeline might further improve the average performance. Contrast Limited Adaptive Histogram Equalization (CLAHE) for instance could be a suitable candidate, since its applies local contrast enhancement based on need.

Furthermore, it could be useful to train additional models based on less effective model blueprints for the attention ResUNet analysis . Since the performance of the best model was already near the ceiling, further improvements were minimal and hard to precisely quantify. By using an ineffective base model, the improvements, if there are any, will be bigger and thus more easily ascribable to the change in architecture.

Chapter 5

Conclusion

To summarize, in this paper a deep learning model was constructed based on literature and a performed grid search for semantic segmentation of a catheter in VFSS footage. The grid search revealed that adaptive gradient based optimizers like adam and adamW are generally better for the training models for this task, as are overlap-based loss functions. Furthermore, it was observed that the addition of augmentations to the dataset and attention mechanisms to the model architecture further improved model performance, with augmentations especially having a significant impact on generalization. All of this culminated in a final model employing an Attention ResUNet architecture, achieving average Dice scores upwards of 0.87 on evaluation sets.

Bibliography

1. Mody, M. D., Rocco, J. W., Yom, S. S., Haddad, R. I. & Saba, N. F. Head and neck cancer. *The Lancet* **398,** 2289–2299 (2021).
2. Chow, L. Q. Head and Neck Cancer. *The New England Journal of Medicine* **382,** 60–72 (2020).
3. Argiris, A., Karamouzis, M. V., Raben, D. & Ferris, R. L. Head and neck cancer. *The Lancet* **371,** 1695–1709 (2008).
4. Mourad, M. *et al.* Epidemiological Trends of Head and Neck Cancer in the United States: A SEER Population Study. *Journal of Oral and Maxillofacial Surgery* **75,** 2562–2572 (2017).
5. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **74.** Epub 2024 Apr 4, 229–263 (May 2024).
6. García-Peris, P. *et al.* Long-term prevalence of oropharyngeal dysphagia in head and neck cancer patients: Impact on quality of life. *Clinical Nutrition* **26,** 710–717 (2001).
7. Espitalier, F. *et al.* International consensus (ICON) on assessment of oropharyngeal dysphagia. *European Annals of Otorhinolaryngology, Head and Neck Diseases* **135.** Supplement, S17–S21 (2018).
8. Nguyen, N. P. *et al.* Severity and Duration of Chronic Dysphagia Following Treatment for Head and Neck Cancer. *Anticancer Research* **25,** 2929–2934 (2005).
9. Rommel, N. & Hamdy, S. Oropharyngeal dysphagia: manifestations and diagnosis. *Nature Reviews Gastroenterology & Hepatology* **13,** 49–59 (2015).
10. Helliwell, K., Hughes, V., Bennion, C. & Manning-Stanley, A. The use of videofluoroscopy (VFS) and fibreoptic endoscopic evaluation of swallowing (FEES) in the investigation of oropharyngeal dysphagia in stroke patients: A narrative review. *Radiography* **29,** 284–290 (2023).
11. Specialisten, F. M. *Orofaryngeale dysfagie* Accessed May 4, 2025. `https://richtlijnendatabase.nl/richtlijn/orofaryngeale_dysfagie/startpagina_-_orofaryngeale_dysfagie.html`.
12. Rugiu, M. Role of videofluoroscopy in evaluation of neurologic dysphagia. *Acta Otorhinolaryngologica Italica* **27,** 306–316 (2007).
13. Kim, J. *et al.* Validation of the Videofluoroscopic Dysphagia Scale in Various Etiologies. *Dysphagia* **29,** 438–443 (2014).
14. Lim, C., Lee, H. J. & Park, Y. A Case of Dysphagia due to Cricopharyngeal Dysfunction and Diffuse Idiopathic Skeletal Hyperostosis. *Journal of the Korean Dysphagia Society* **12,** 74–78 (Jan. 2022).
15. Schindler, A., Baijens, L. W. J., Geneid, A. & Pizzorni, N. Phoniatricians and otorhinolaryngologists approaching oropharyngeal dysphagia: an update on FEES. *European Archives of Oto-Rhino-Laryngology* **279,** 2727–2742 (2021).
16. Langmore, S. E., Schatz, K. & Olsen, N. Fiberoptic endoscopic examination of swallowing safety: A new procedure. *Dysphagia* **2,** 216–219 (1988).
17. Baijens, L. W. J. *et al.* European white paper: oropharyngeal dysphagia in head and neck cancer. *European Archives of Oto-Rhino-Laryngology* **278,** 577–616 (2020).
18. Scharitzer, M., Roesner, I., Pokieser, P., Weber, M. & Denk-Linnert, D. M. Simultaneous Radiological and Fiberendoscopic Evaluation of Swallowing ("SIRFES") in Patients After Surgery of Oropharyngeal/Laryngeal Cancer and Postoperative Dysphagia. *Dysphagia* **34,** 852–861 (2019).
19. Lee, T. H., Lee, J. S. & Kim, W. J. High resolution impedance manometric findings in dysphagia of Huntington's disease. *World Journal of Gastroenterology* **18,** 1695–1699 (2012).

20. Omari, T. I. *et al.* A Method to Objectively Assess Swallow Function in Adults With Suspected Aspiration. *Gastroenterology* **140,** 1454–1463 (2011).

21. Omari, T. I. *et al.* Reproducibility and Agreement of Pharyngeal Automated Impedance Manometry With Videofluoroscopy. *Clinical Gastroenterology and Hepatology* **9,** 862–867 (2011).

22. Omari, T. I. *et al.* Defining Pharyngeal and Upper Esophageal Sphincter Disorders on High-Resolution Manometry-Impedance: The Leuven Consensus. *Neurogastroenterology & Motility* (2025).

23. Neijman, M. *et al.* The Use of Pharyngeal High-Resolution (Impedance) Manometry in Patients With Head and Neck Cancer: A Scoping Review. *American Journal of Speech-Language Pathology* **33,** 3100–3120 (2024).

24. Lee, T. H. *et al.* High-resolution impedance manometry facilitates assessment of pharyngeal residue and oropharyngeal dysphagic mechanisms. *Diseases of the Esophagus* **27,** 220–229 (2014).

25. Shen, D., Wu, G. & Suk, H.-I. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering* **19,** 221–248 (2017).

26. Suzuki, K. Overview of deep learning in medical imaging. *Radiological Physics and Technology* **10,** 257–273 (2017).

27. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521,** 436–444 (2015).

28. Peter, S. C. *et al.* in *Encyclopedia of Bioinformatics and Computational Biology* (eds Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C.) 661–676 (Academic Press, Oxford, 2019). ISBN: 978-0-12-811432-2. https://www.sciencedirect.com/science/article/pii/B9780128096338201970.

29. Kim, M. *et al.* Deep Learning in Medical Imaging. *Neurospine* (2019).

30. Takahashi, S. *et al.* Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *Journal of Medical Systems* **48** (2024).

31. Liu, X., Song, L., Liu, S. & Zhang, Y. A Review of Deep-Learning-Based Medical Image Segmentation Methods. *Sustainability* **13** (2021).

32. Liu, X., Song, L., Liu, S. & Zhang, Y. A Review of Deep-Learning-Based Medical Image Segmentation Methods. *Sustainability* **13** (2021).

33. Caliskan, H., Mahoney, A. S., Coyle, J. L. & Sejdić, E. *Automated Bolus Detection in Videofluoroscopic Images of Swallowing Using Mask-RCNN* in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2020), 2173–2177.

34. Peter, S. C. *et al.* in *Encyclopedia of Bioinformatics and Computational Biology* (eds Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C.) 661–676 (Academic Press, Oxford, 2019). ISBN: 978-0-12-811432-2. https://www.sciencedirect.com/science/article/pii/B9780128096338201970.

35. Iida, Y. *et al.* Detection of aspiration from images of a videofluoroscopic swallowing study adopting deep learning. *Oral Radiology* **39,** 553–562 (2023).

36. Henderson1, R. D. E., Yi, X., Adams, S. J. & Babyn, P. Automatic Detection and Classification of Multiple Catheters in Neonatal Radiographs with Deep Learning. *Journal of Digital Imaging* **34,** 888–897 (2021).

37. Ghosh, R., Wong, K., Zhang, Y. J., Britz, G. W. & Wong, S. T. C. Automated catheter segmentation and tip detection in cerebral angiography with topology-aware geometric deep learning. *Journal of NeuroInterventional Surgery* **16,** 290–295. ISSN: 1759-8478. eprint: https://jnis.bmj.com/content/16/3/290.full.pdf. https://jnis.bmj.com/content/16/3/290 (2024).

38. Lee, H., Mansouri, M., Tajmir, S., Lev, M. H. & Do, S. A Deep-Learning System for Fully-Automated Peripherally Inserted Central Catheter (PICC) Tip Detection. *Journal of Digital Imaging* **31,** 393–402 (2017).

39. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) **9351** (Springer International Publishing, 2015), 234–241. https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28.

40. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition (2015).

41. Zhang, Z., Qingjie & Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters* (2017).

42. Wilson, A. C., Roelofs, R., Stern, M., Srebro, N. & Recht, B. *The Marginal Value of Adaptive Gradient Methods in Machine Learning* 2018. arXiv: 1705.08292 [stat.ML]. https://arxiv.org/abs/1705.08292.

43. Ramachandran, P., Eswarlal, T., Lehman, M. & Colbert, Z. Assessment of Optimizers and their Performance in Autosegmenting Lung Tumors. *Journal of Medical Physics* **48.** Epub 2023 Jun 29, 129–135 (Apr. 2023).

44. Loshchilov, I. & Hutter, F. *Decoupled Weight Decay Regularization* 2019. arXiv: 1711.05101 [cs.LG]. https://arxiv.org/abs/1711.05101.

45. Kato, S. & Hotta, K. Adaptive t-vMF dice loss: An effective expansion of dice loss for medical image segmentation. *Computers in Biology and Medicine* **168,** 107695. ISSN: 0010-4825. https://www.sciencedirect.com/science/article/pii/S0010482523011605 (2024).

46. Yue, Z., Shijie, L., Chunlai, L. & Jianyu, W. Rethinking the Dice Loss for Deep Learning Lesion Segmentation in Medical Images. *Journal of Shanghai Jiaotong University (Science)* **26,** 93–102 (2021).

47. Ehab, W. & Li, Y. *Performance Analysis of UNet and Variants for Medical Image Segmentation* 2023. arXiv: 2309.13013 [eess.IV]. https://arxiv.org/abs/2309.13013.

48. Jan, A. & Seo, S. Monocular Depth Estimation Using Res-UNet with an Attention Model. *Applied Sciences* **13,** 6319 (May 2023).