Analyzing Attention Gate Variants in UNet for Knee Recess Distension

Mason Hu[1] and Pascal Tyrrell[2]

[1]Department of Computer Science, University of Toronto
[2]Department of Medicine, University of Toronto

August 19, 2023

Abstract

Medical imaging plays an increasingly vital role in clinical diagnosis, making the automation of detection, classification, and segmentation essential. This study examines the state-of-the-art UNet for knee recess distention ultrasound segmentation, focusing on the impact of attention gates, residual blocks, and augmented attention modules. By integrating residual connections into attention-gated UNets, a novel UNet variant, AgResUNet, is proposed. This variant demonstrates improved mean IoU and produces segmentations that are semantically more aligned with the ground truth. The research also explores the interplay between attention and residual blocks within AgResUNet and uncovers a critical bug in PyTorch affecting evaluation performance for batch normalized models. The code for this research is accessible at https://github.com/masonhyz/agv-unet.

1 Introduction

In healthcare and medical studies, we rely substantially on imaging methods to diagnose lesions and illnesses. We can automate this process using image classification algorithms, but we still waste a lot of time and computation on parts of the images that are not relevant—background, irrelevant tissues, etc. As our understanding of medical sciences continues to evolve, the need for more precise, automatic, and swift diagnostic techniques is becoming increasingly apparent.

Segmentation, the technique that classifies each pixel of an image by its true relevance, is crucial for machine learning in medical imaging. Segmenting medical images shows its worth by enhancing diagnosis, time efficiency, improving treatment planning and so on. However, manual segmentation requires too much human labor, so we aim to automate this process by using by massively feeding in data into machine learning algorithms. The state-of-the-art model for segmentation tasks, giving us pixel-accurate information, is U-Nets. Since U-Nets assign each pixel a label, we will obtain more spatial context information, achieve multi-object detection, and carry out fine-grained analysis.

As images become more sophisticated, the model usually gets overwhelmed by the computation and loses performance. There is a myriad of U-Net models and variants in the current literatures, but none of them provides a comprehensive analysis on their performance on knee recess distension ultrasound segmentation. My purpose of this research is to solve this problem and fill in the gap by constructing a U-Net model for knee distension, and more importantly, improve the performance of U-Nets through introducing various modules. I will build on top of Nana's project, which analyzed the effect of attention gates on U-Net. I will test if different attention U-Nets (specifically, AgUNet, ResUNet, AgResUNet, RAUNet) boost the model's overall performance. I will explore how they achieve their performance, that is, if they improve the model by highlighting the important features and suppressing the irrelevant ones, or by optimizing the training procedure through alleviation of vanishing gradient.

1.1 Hypothesis

Attention gates, residual blocks and/or augmented attention modules significantly improves the performance (in particular, mean IoU) of U-Nets on knee recess distension ultrasound segmentation tasks.

1.2 Objectives

- Implement attention U-Net and add residual blocks to the convolutional layers; implement attention gates (AG) and augmented attention modules (AAM). Construct AgResUNet using by combining AGs with Res blocks and RAUNet through AAM and transfer learning on the ResNet34 weights.

- Fine-tune the hyperparameters of each U-Net. Assess the performance using mean intersection over union (MIoU) metric and test our hypothesis by comparing the U-Net variants to the baseline U-Net.
- Visualize the attention gates variant quantitatively. Qualitatively assess the performance of the models and seek out limitations. Similarly, identify the advantages of each model by evaluating individual segmentations. Propose potential improvements for future studies.

2 Methods

2.1 Fundamental Theory Establishment: This section will encompass a detailed literature review and exploration of attention mechanisms in neural networks. The focus will be on U-Nets, residual blocks, and attention mechanisms. It will establish a theoretical understanding of how attention cooperate with neural networks, their benefits, drawbacks, and the context in which they can be used effectively.

2.2 Data Preprocessing: This initial phase entailed acquiring and preprocessing the ultrasound images dataset used for diagnosing knee recess distension. This process involved image loading, normalization of pixel values, resizing the images to an appropriate format for model input, and converting images to the appropriate color channels. Following the preprocessing stage, the data are split into three subsets: training, validation, and testing. This division allows for model training, tuning, and a final evaluation of model performance on unseen data.

2.3 Training Environment Setup: A crucial step for an organized training process. The environment for model training was set up with remote GPU access to manage the computational requirements of deep learning models. The setup involved configuring the remote server, specifying SSH access and deployment paths, installing the necessary software packages and libraries, and logging details. Helper functions for viewing progressions and comparisons is also implemented.

2.4 Standard U-Net Construction: In Pytorch, the U-Net model was built from scratch based on the original architecture shown in papers, and was fit on the knee recess distention dataset offered by MiDATA lab. This architecture served as the foundation upon which the attention mechanisms will be added.

2.5 Implementation of attention and U-Net variants: Residual blocks are integrated into the convolutional layers of the U-Net model. These blocks serve to create shortcuts around groups of layers, allowing the gradient to be directly backpropagated to earlier layers. Attention gates will be designed and will be used both in isolation and in combination of the residual connections. Finally, AAM was implemented with the help of transfer learning on ResNet34 to form RAUNet. See U-Net inheritance hierarchy specific for my project (Figure 2.1) for a detailed dependence relationship.

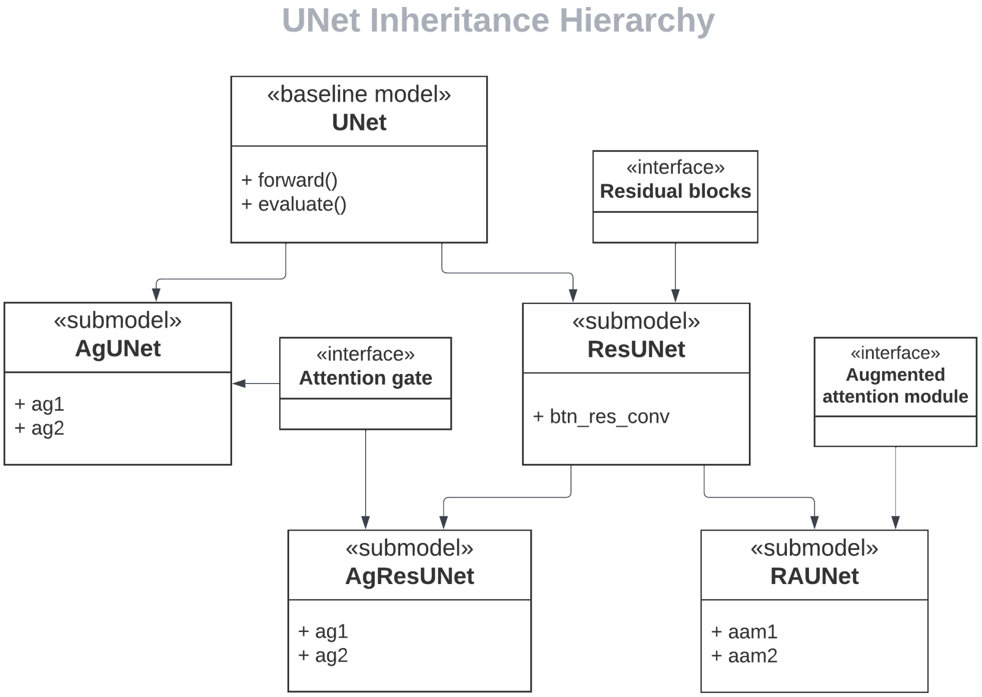


Figure 2.5

2.6 Model Assessment: Once the models have been trained, their performance were assessed using the intersection over union (IoU) metric, the most effective measure for evaluating segmentation tasks. The model's training progression, along with validation accuracy and training loss, were visualized and plotted for detailed examination. Moreover, statistics like Hausdorff distance and percentage of one-piece mask were reported.

2.7 Comparison with the Baseline Model: Through the hyperparameter tuning and after the optimal model and training configurations were found, the models were evaluated on the test set, and their accuracies were reported. The performances of the best attention-enhanced models were compared with the standard U-Net model visually. Various aspects of performance, including IoU scores and other statistics, will be presented in comparative charts.

2.8 Qualitative Assessment: A qualitative examination of the models' performance was conducted, with particular emphasis on how the models handle edge cases, their success in highlighting areas of interest, and how they suppress irrelevant regions. Attention weight maps will be visualized. I also conducted size by side comparison of model performance on typical input ultrasound images.

Limitations and Further Improvements: The final stage of the methodology involved a detailed discussion on the limitations observed in the implemented models, as well as on the dataset annotations. Based on these limitations and the results from the models, suggestions for further improvements and potential directions for future research were proposed.

3 Architecture:

In this research, a sophisticated architecture known as AgResUNet is introduced. This innovative structure is designed through the strategic integration of attention gate modules and residual blocks based on the UNet architecture. To provide clarity in understanding the interactions and functions among the models, a comprehensive UML diagram is made, offering insights into the novel AgResUNet. Other illustrations that elucidate the residual connections and the mechanisms of the attention gate are in the appendix.

3.1 Why residual blocks?

Residual block, or skip connection, is a key architectural component in deep neural networks that allow the training of much deeper models without hindering performance. Essentially, the original inputs of each layer are preserved and added algebraically to the output from the convolutional blocks (see Figure 3.1), skipping the convolution. This means instead of learning the mapping F(x), through residual blocks we learned F(x)-x, which is easier to optimize since we are pushing the residuals to zero instead of stacking non-linear layers to form an identity. This alleviates the vanishing gradient problem, where gradients become too small for the network to learn effectively as they are propagated back through many layers. (Kaiming He et al. 2015)

3.2 Why attention gates?

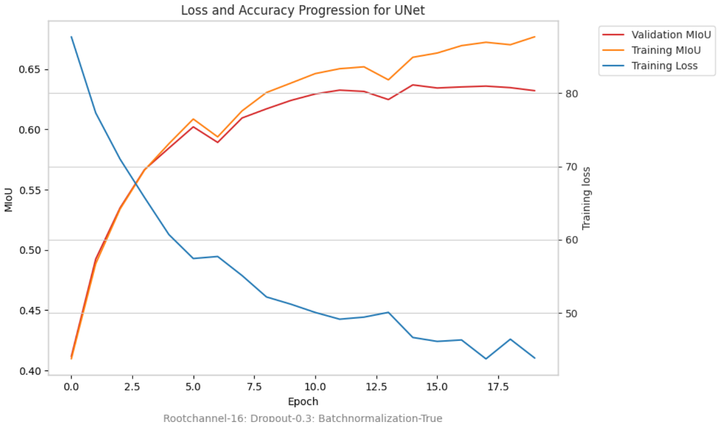
Unlike attention in transformer models or in the context of NLP, attention gates in U-Net architectures are designed to allow the network to focus on specific spatial regions that are more informative for the task at hand, much like how human visual attention operates. Within the context of image segmentation, certain areas of an image may hold more pertinent information than others, and attention gates help the network to allocate more resources to those salient regions. Essentially, an attention gate combines the clear spatial information with a coarser feature map and converts them into a map of attention weights (from 0 to 1) using a 1*1 convolution (see Figure 3.2). The mechanism significantly improves the efficiency and effectiveness of the U-Net model, especially in segmentation tasks where a detailed localization is needed. (Ozan Oktay et al. 2018)

4 Results

After tedious hyperparameter tuning, conclusion is reached that in general, UNets with root channel of 16, batch normalization, and dropouts has the highest validation accuracy. This applies universally for plain UNet, AgUNet, ResUNet, and AgResUNet. The detailed comparison of most tunable hyperparameter (root channel, dropout, batch normalization, loss) is shown in Table 4.0.

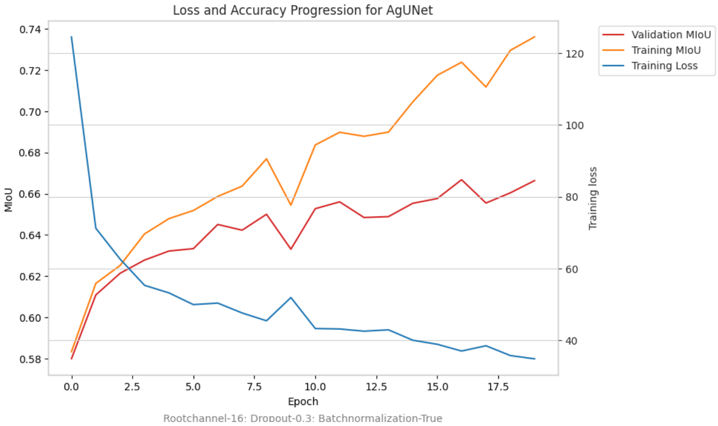
4.1 Progressions

4.1.1 Plain UNet



Test MIoU for plain UNet: 0.638

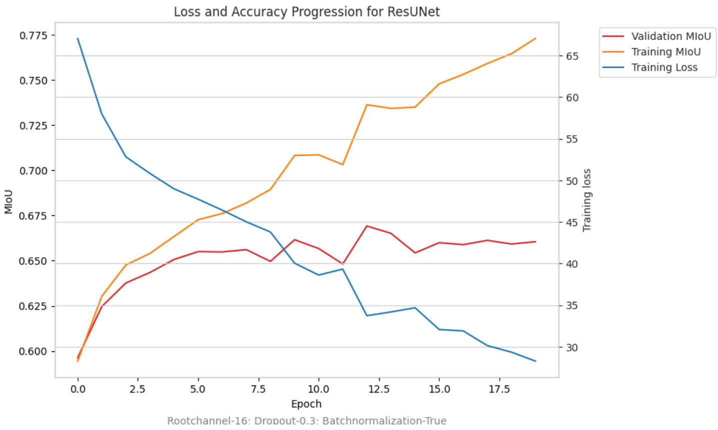
4.1.2 AgUNet



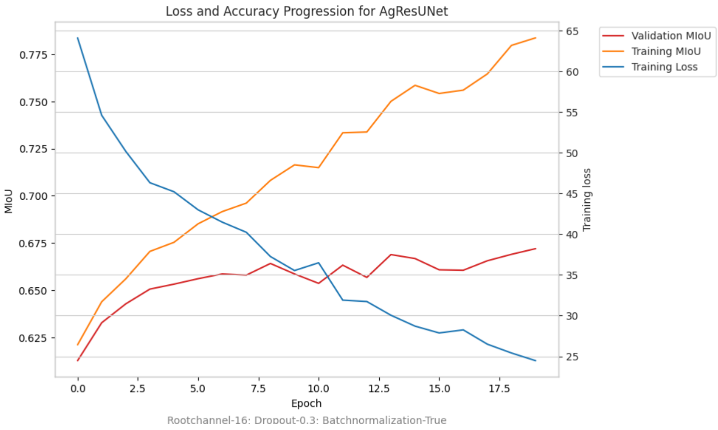
Test MIoU for AgUNet: 0.668

4.1.3 ResUNet

Test MIoU for ResUNet: 0.670



4.1.4 AgResUNet



Test MIoU for AgResUNet: 0.672

4.2 Characteristic metrics

As we can see in (Table 4.2), each of the four metrics of the five UNet variants were reported. The dice coefficients of each model are universally about 0.1 higher the MIoU. Hausdorff distances exhibit no significant insights, while AgResUNet has an obvious low in percentage of one-piece masks, revealing the complexity of the model while showing signs of overfitting at the same time. To this end, we might construct a connectivity regularizer to guide the model to produce more one-piece masks, if one-piece masks are what we

Table 4.2 Characteristic metrics

	UNet	**AgUNet**	**ResUNet**	**AgResUNet**	**RAUNet**
MIoU	0.638	0.668	0.670	0.672	0.44
Dice Coefficient	0.746	0.783	0.774	0.779	0.34
Hausdorff distance	6.220	6.073	6.244	6.196	50.6
Percentage of one-piece masks	67.6	68.0	70.53	35.6	15.6

care about. More on this in the discussions section. To my disappointment, the RAUNet models are not living up to their expectations, receiving an abysmal accuracy and failing in every non-trivial task. This might actually be owing to the model's specialty on detecting intensive light areas, which is incompatible with our ultrasound images. (Zhen-Liang Ni et al. 2019)

In a nutshell, standard UNet, attention gate UNet, residual UNet, attention gate residual UNet, are similar in knee recess distension ultrasound segmentation tasks performance, with the exception of standard UNet being considerably weaker. In other words, the incorporation of attention gates and residual blocks made significant improvement effect on the model's performance. To no one's surprise, when the two mechanisms are coexistent in a UNet, i.e. in AgResUNet, the performance of the model further improved as much as 0.02. However, this might have overcomplicated the situation, since the masks are rarely in one piece and we can see jagged borders and isolated pixels. In general, residual blocks and attention gates are evidently improving UNet.

4.3 Qualitative comparison

Figure 4.3 provides a detailed qualitative comparison between the four UNet variants, evaluated across four typical individual cases, each represented in a separate row. Key areas of interest are highlighted with red boxes for further

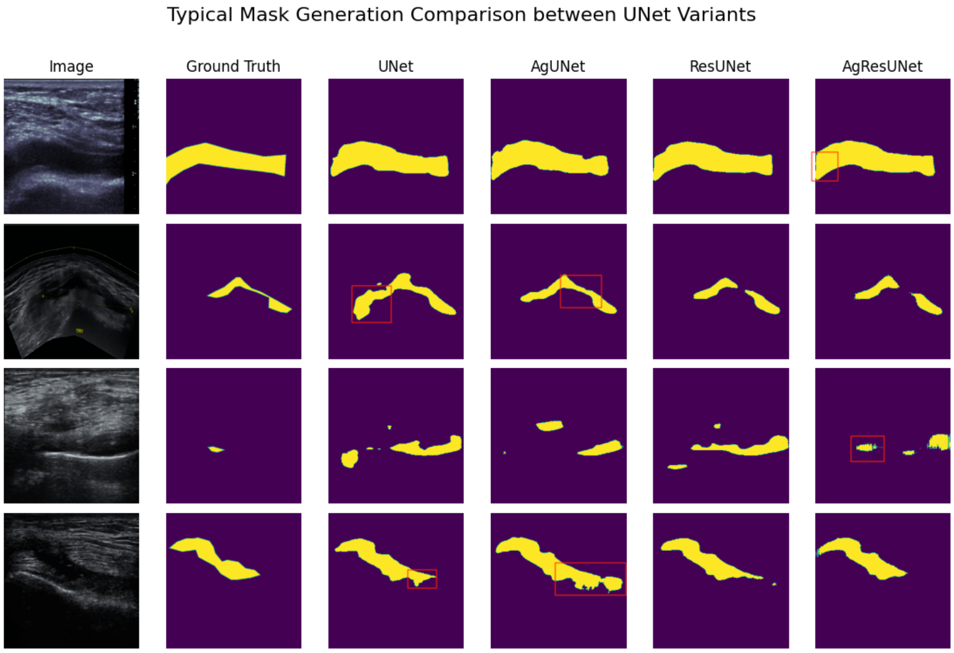


Figure 4.3

examination below.

First Row: This represents the most straightforward case, where segmentation is relatively simple due to the apparent appearance of the knee as a thick black slice in the ultrasound image. Though all UNet models performed well, AgResUNet exhibited the closest agreement with the ground truth, particularly along the leftmost margins.

Second Row: Here, the ground truth consists of a thin sliver, posing a more complex challenge. The standard UNet significantly overestimated the grey areas not part of the knee, whereas AgUNet, though excluding the grey area, surpassed both ResUNet and AgResUNet by connecting two regions, forming a slender neck area.

Third Row: This case is perplexing, as the ground truth appears to have misled all the models. All UNet variants overestimated the masks, with AgResUNet failing the least, as it correctly identified the one part of the ground truth at the very least.

Fourth Row: In this instance, both UNet and AgUNet included unnecessary protrusions in the segmentation mask. AgResUNet performed best, despite having less distinct boundaries.

This qualitative analysis reveals AgResUNet as the most robust model in segmentation tasks, generally producing the highest quality segmentation masks. While the remaining models each have their unique strengths and weaknesses, the plain UNet often misinterprets the grey areas, a mistake avoided by AgResUNet. Overall, this comparison underscores AgResUNet's superior performance and illustrates specific challenges and successes encountered by each model variant.

5 Discussion

5.1 Attention paradox

During the qualitative assessment phase of visualizing attention maps in AgUNet and AgResUNet, a remarkable observation was made—the attention weight maps for both models are nearly inverse to each other (see Figure 5.1). Specifically, for every pair of attention maps across each layer of the UNet, the one with a residual connection is predominantly red with occasional green strokes, whereas its counterpart is mostly blue with slightly red centers. This phenomenon effectively means that AgUNet focuses on the foreground, while AgResUNet emphasizes almost uniformly on the background, an unexpected behavior for a segmentation task.

The questions this observation raises are intriguing: Why would a model concentrate on the background in segmentation? Why is attention in the foreground reduced, and how does this correlate with the model's relative high accuracy and performance compared to UNet and AgUNet? Since attention maps with entirely contrasting properties produce the same accuracy, it can be inferred that it is the residual blocks that shift the model from one extreme to the other. This insight hints at the relative insignificance of attention gates on UNets compared to residual blocks, supporting the null hypothesis that attention gates may not substantially affect the overall performance of the architecture.

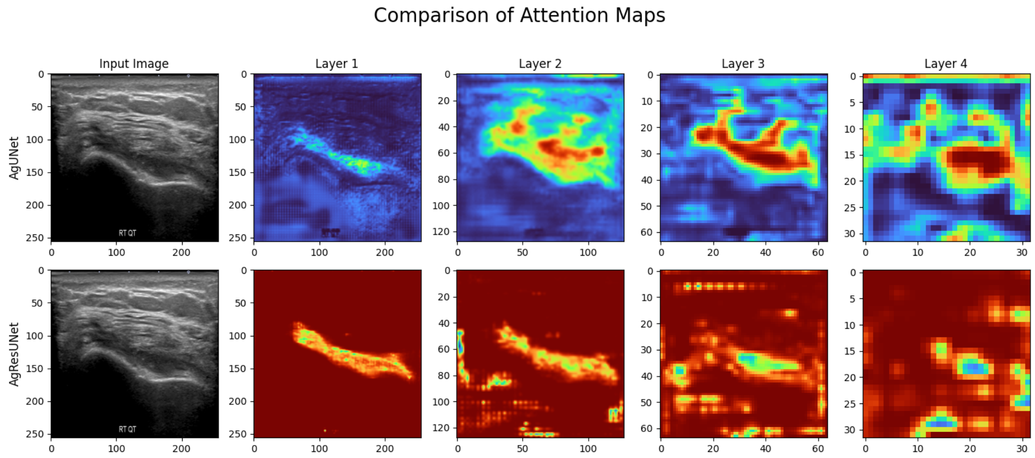


Figure 5.1

5.2 Batch norm bug

During the model evaluation process, an anomalous behavior was detected where the models performed poorly in general assessments but exhibited strong performance in generating individual prediction masks. Upon a detailed investigation of every prediction, it was observed that some predictions had virtually zero Intersection over Union (IoU) values, contrasting with the typical ones that were higher than 0.4. Even though the masks seemed to have a non-trivial intersection with the ground truth, the discrepancy persisted.

The root cause of this discrepancy was found to be the method model.eval() used on the UNet model during the evaluation phase. This behavior is documented as a common problem within the PyTorch community for users training deep neural networks on GPUs (see PyTorch Forum Thread). The issue stems from the batch normalization modules carrying training running mean and variances into the evaluation phase. Given the use of a batch size of 1 in this case, the running statistics were highly variable for each batch when calling ‘model.eval()’, leading to improper normalization of the inputs in layers if the batch size was above 32. Some of these incorrect normalizations coincided with the inputs to produce near-zero activations, resulting in zero IoU.

After four hours of careful analysis and 120 lines of debugging, the bug was addressed by adopting a solution from the community. The fix involved setting the 'running_mean' and 'running_var' of all child BatchNorm2D modules recursively after calling 'model.eval()', thus eliminating the volatile effect. This adjustment allows the model to follow the PyTorch guidelines, utilizing the essential "good practice" 'model.eval()' in the evaluation phase while still being able to achieving sensical accuracies.

This unexpected finding underscores the complexity and nuanced challenges of the model training and evaluation process. Interestingly, this thorough debugging session also revealed that some data points contained empty ground truths (such as 1984_16.png, see Figure 5.2), leading to further inquiries into label quality.

5.3 Label Quality

The label quality, bluntly, is quite rough. Specifically, the ground truth annotations are edgy and sharp, while the true edges of the knee area are smooth and round. This is a huge challenge for deep model training and the expressivity of the labels cannot keep up with the expressivity of the neural network. While this limits the model training process, it also degrades the true model performance evaluation since even a perfect mask would not produce excellent metrics due to having too smooth of an edge.

5.4 Limitation: Dataset complexity

While label quality is an obstacle to the training and evaluation process, what really limits this study from exhibiting the potential of the UNets is the lack of complexity in datasets. Close to monotone inputs are the reason for two-thirds of the data being disused—triple channel input has a single channel representation. Single big patches of knees are to some extent equivalent to minimizing the perimeter while maximizing areas—meanwhile the perimeters are what makes the model useful since boundaries are the focus of a segmentation neural network. Coupled with the fact that the labels lack precision, this dataset is more suitable for applicational purposes—constructing a mask generator—instead of researching attention mechanisms.

5.5 Limitation: Computational setback

Even though the 1080 Ti tremendously accelerated the training processes, the inability to parallel process more than 2 models at a time made hyperparameter tuning a bit tedious and unorganized. Training 2 models simultaneously, the hardware would work at a 6 hour per model rate. But training 3 at a time, the hardware significantly slows down to 17 hour per model.

5.6 Future work: Overfitting and targeted regularizers

Even though dropout and batch normalization are applied and are significantly helpful, the overfitting issue is still prominent. However, as the gap between training and validation accuracy become larger and larger, the absolute validation accuracy did not drop. They even sometimes slightly increased. I have been wondering if it is possible for this training behaviour to continue for the next 30 or 40 epochs. If so, the model will most probably change enough to produce other excellent insights. However, we still need to make sure the validation accuracy does not plummet. As efforts to alleviate overfitting, L2 regularizers and the pre-proposed metric regularizers like connectivity regularizers might play a crucial role.

5.7 Future work: Augmentation

As outline in the original UNet paper (Olaf Ronneberger et al. 2015), UNets are great models especially with data augmentation. Since I personally am working on a private dataset on a remote protected server, the augmentation process would require more tedious work in manipulating the directories through terminal. Hope to see this part in future continuations of the project.

5.8 Future work: Field knowledge incorporation

This study requires more professional knowledge than I thought. Simply put, a radiologist would think differently from I would about what the best segmentation for a specific image is. And a physician would think differently too. Due to different anatomical regions in the knee having different functions, there might also be some segmentation masks that work particularly well in one of them but not the others. We can ask for experts' opinions on this, but if I had more time, I would also study some anatomy and propose some statistics aligning with anatomical or functional landmarks. Furthermore, more acute accuracies like metrics of boundary precisions should be investigated.

6 Conclusion

The integration of attention gates and residual blocks has markedly enhanced the performance of the traditional UNet, as evidenced in this research. Particularly, the AgResUNet, as introduced in this study, has demonstrated superiority in a multifaceted assessment, showing strong capabilities in segmentation tasks. Despite its notable successes, certain aspects such as paradoxical attention mechanics and ambiguity in boundary delineations indicate areas for further refinement. These nuanced challenges not only add complexity to the model but also highlight promising avenues for future research, underscoring the potential for continued innovation and development in this field.

References

[1] Ronneberger O. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI 2015*, `https://arxiv.org/abs/1505.04597`

[2] He K. (2015). Deep Residual Learning for Image Recognition. `https://arxiv.org/abs/1512.03385`

[3] Oktay O. (2018). Attention U-Net: Learning Where to Look for the Pancreas. *MIDL'18*, `https://arxiv.org/abs/1804.03999`

[4] Ni Z. (2019). RAUNet: Residual Attention U-Net for Semantic Segmentation of Cataract Surgical Instruments. *26th International Conference on Neural Information Processing (ICONIP2019)*. `https://arxiv.org/abs/1909.10360`

Appendix

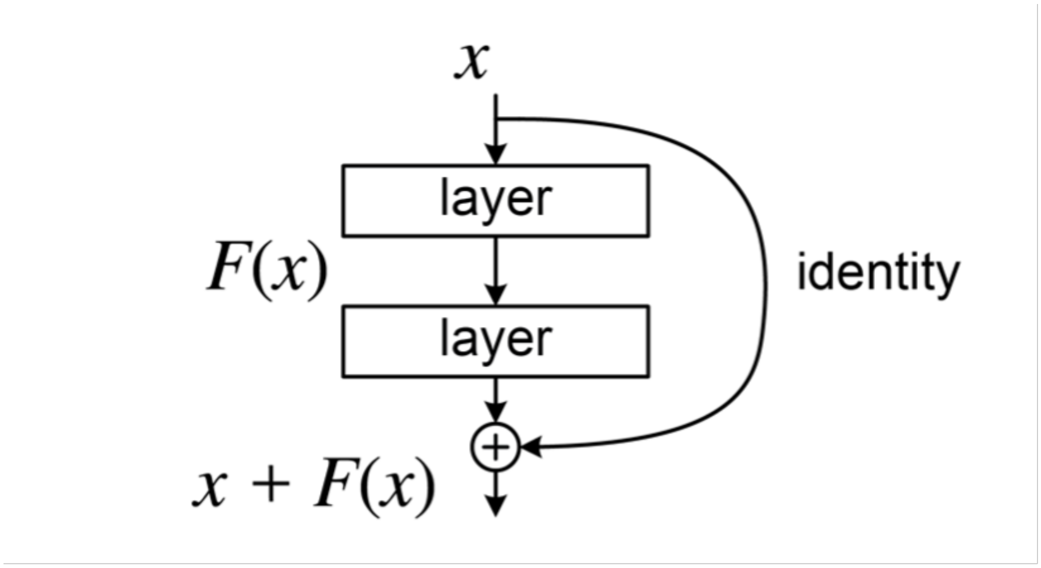


Figure 3.1 Residual block

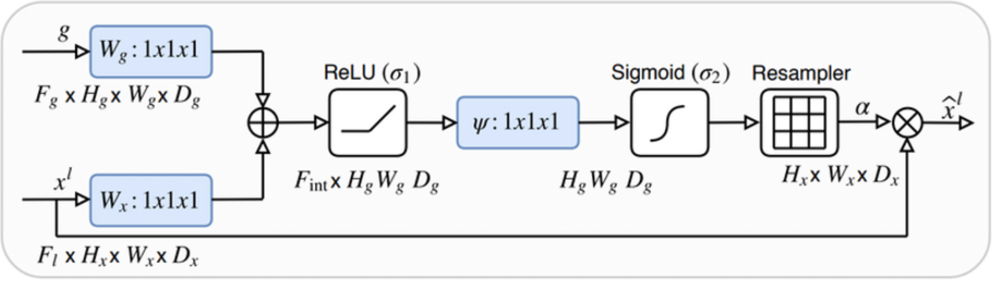


Figure 3.2 Attention gate mechanism

Table 4.0: Hyperparameter tuning with MIoU statistic

Model		**Loss Function (D=0.3)**			**Dropout (BCE)**	
Type	**RChannel**	**BCE**	**Diceloss**	**IoUloss**	**0**	**0.5**
With Batch Norm						
UNet	8	0.632			0.634	
	16	0.636	0.525	0.513	0.633	
	32	0.632			0.629	
	64	0.633	0.512	0.484	0.636	
AgUNet	8	0.668				
	16	0.668	0.511	0.513	0.667	0.657
	32	0.661				
	64	0.660	0.509	0.505	0.662	0.668
ResUNet	8	0.668			0.669	
	16	0.670			0.667	
	32	0.663			0.666	
	64	0.662			0.645	
AgResUNet	8	0.666			0.652	
	16	0.672				
	32	0.671			0.670	
	64	0.651				
Without Batch Norm						
UNet	8					
	16	0.613			0.622	
	32					
	64	0.600			0.607	
AgUNet	8					
	16	0.623			0.622	
	32					
	64	0.623			0.619	0.617
ResUNet	8					
	16	0.637			0.630	
	32					
	64	0.632			0.631	
AgResUNet	8					
	16	0.630			0.640	
	32					
	64	0.629			0.626	

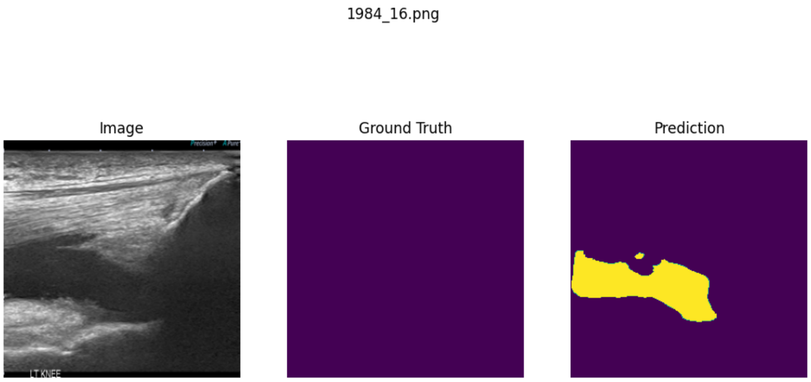


Figure 5.2