

Capstone Project Phase A

**MRI image segmentation in cancer of the GI tract**

**Instructor name:** Miri Weiss-Cohen

**Team members:**  
 Rotem Levi

Ofek Malka

[Project GitHub link](https://github.com/rotemlv/CapstoneProject)

**Table Of Content**

[Abstract 1](#_Toc167366839)

[Introduction 1](#_Toc167366840)

[Literature Review 3](#_Toc167366841)

[Background 10](#_Toc167366842)

[U-Net 10](#_Toc167366843)

[The U-Net architecture 11](#_Toc167366844)

[Advantages and limitations 12](#_Toc167366845)

[Attention 13](#_Toc167366846)

[From language models to images 15](#_Toc167366847)

[Attention U-Net 16](#_Toc167366848)

[Transformers in U-Net 17](#_Toc167366849)

[Expected Achievements 18](#_Toc167366850)

[Research Challenges 18](#_Toc167366851)

[Research Process 18](#_Toc167366852)

[Our proposed Process 18](#_Toc167366853)

[Challenges and achievements 20](#_Toc167366854)

[Dataset 20](#_Toc167366855)

[Hyperparameters 20](#_Toc167366856)

[Product 21](#_Toc167366857)

[Use case 21](#_Toc167366858)

[Flow Chart 21](#_Toc167366859)

[GUI 22](#_Toc167366860)

[Evaluation/Verification Plan 24](#_Toc167366861)

[References 25](#_Toc167366862)

# Abstract

This research project focuses on enhancing the segmentation accuracy of MRI images for these cancers, which is critical for effective radiation therapy planning. Employing advanced deep learning techniques, particularly Transformer-based architectures integrated with U-Net models, this study aims to develop a segmentation model that surpasses the current state-of-the-art in accuracy. The proposed model leverages the inherent strengths of Transformers to manage the complex spatial relationships in MRI scans, facilitating more precise targeting of radiation doses and potentially improving patient outcomes.

By refining image segmentation processes, this project not only seeks to increase the precision of radiation therapy but also to expedite treatment workflows, addressing both the rising prevalence of GI cancers and the urgent need for more effective treatment methodologies. The ultimate goal is to provide a robust tool for oncologists and radiologists that enhances both the accuracy and speed of cancer treatment planning.

# Introduction

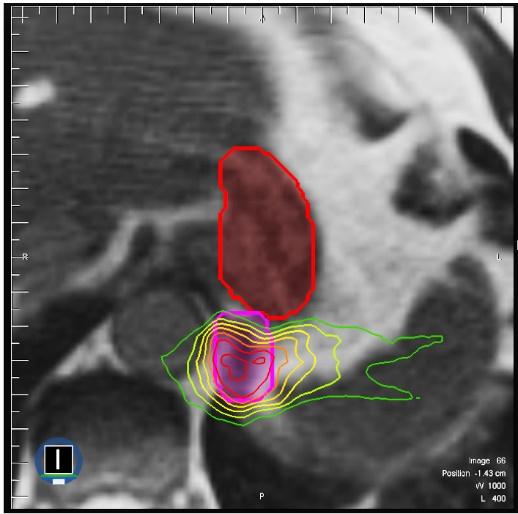
According to **[‎1, ‎2]**, cancers of the gastro-intestinal tract affect an estimated 5 million each year worldwide, representing more than a quarter of all cancers. Worryingly, the prevalence of such cancers is increasing, causing an estimated 3 million deaths per year.

Gastrointestinal (GI) tract cancers typically begin with genetic mutations in the cells lining the digestive tract. These mutations can lead to abnormal cell growth and eventually cancer. These cancers comprise a significant portion of global cancer incidences and mortalities, pose substantial challenges in medical imaging and treatment.

Gastro-intestinal cancers manifest as gastric carcinomas arise through successive changes and not de novo from normal epithelium. For the intestinal type of gastric cancer, this includes the transformation of normal mucosa into a mucosa that resembles intestinal epithelium. Treatment strategy mainly consists of surgical resection of the primary tumor and regional lymph nodes along with chemotherapy and radiotherapy **[‎3]**.

About half of patients are eligible for radiation therapy during treatment [‎**18**]. During such treatments, radiation oncologists try to deliver therapeutic doses of radiation, using X-ray beams pointed to tumors, while avoiding the healthy stomach and intestines.

With newer technology such as integrated magnetic resonance imaging (MRI) and linear accelerator systems (MR-Linacs), oncologists are able to visualize the daily position of the tumor and intestines, which can vary day to day. In these scans, radiation oncologists must manually segment (outline) the position of the stomach and intestines in order to adjust the direction of the x-ray beams to increase the delivered radiation dose to the tumor while avoiding the stomach and intestines (see **Fig 1** below).

  
**Figure 1**: segmentation of the stomach and intestines on MRI scan. The tumor (pink) is close to the stomach (red). High doses of radiation are directed to the tumor while avoiding the stomach. Dose levels represented by the rainbow of outlines, with higher doses represented by red and lower doses represented by green [‎**2**].

However, manual segmentation of the tumor and the GI organs is a time-consuming and labor-intensive process that can prolong treatments from 15 minutes a day to an hour a day, which can be difficult for patients to tolerate. Automation of this process could shorten treatment duration, ease the burden on radiologists and could provide faster treatment and quicker recovery in radiation patients.

Deep learning methods in general and segmentation U-Nets can assist in reducing manual work and allowing more patients to receive appropriate treatment by automating the segmentation process. Indeed, deep neural network-based methods have been employed in recent years for the automated diagnosis of medical illnesses. It has outperformed traditional algorithms in terms of accuracy since the characteristics are learned from data using a general-purpose learning technique rather than being built by human engineers.

Our goal is to develop an enhanced GI tract segmentation method for MRI images that rivals current SOTA in accuracy, by incorporating Transformer based architectures in our model.

The field of medical segmentation networks (and segmentation in general), even with the recent advancements (incorporating transformers and similar attention mechanisms in different forms), has yet to obtain sufficiently accurate values (the problem of semantic segmentation, unlike classification, is not yet fully solved), but strides towards that goal are constantly being made.

Since segmentation of the GI tract has seen less attention (compared to brain or liver segmentation networks research), our aim is to compare and analyze the most recent architectures for GI segmentation and improve their results, thereby allowing future research an easier path towards solving the problem of automated medical segmentation.

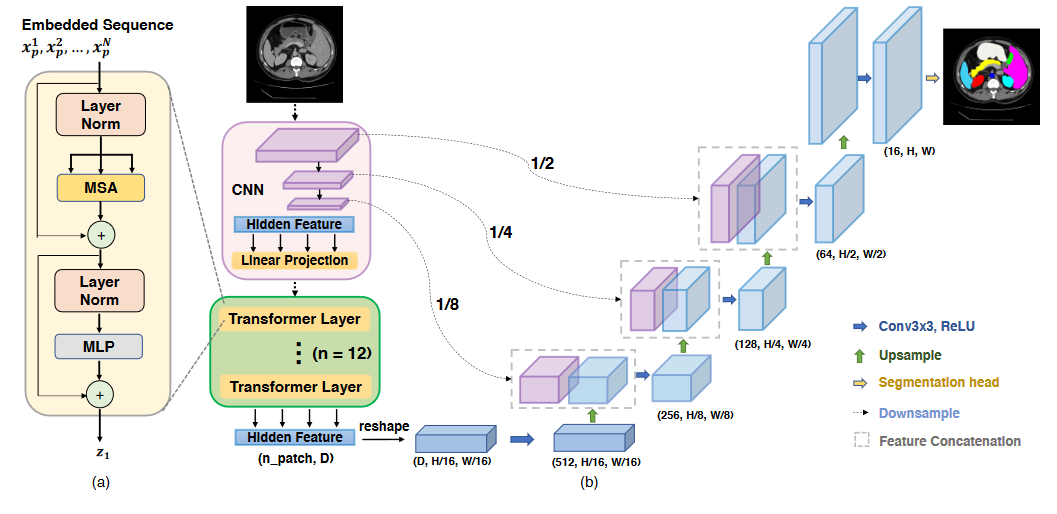
Segmentation in the context of medical imaging, such as MRI scans, refers to the process of dividing an image into multiple segments or regions, each representing a different part of the body or anatomical structure. This technique is crucial for identifying and isolating specific areas of interest, such as tumors or organs, from the rest of the image. In the field of gastrointestinal (GI) tract segmentation, this process is particularly important for diagnosing and treating various conditions.

The significance of segmentation in GI tract imaging lies in its potential to revolutionize the diagnosis and treatment of gastrointestinal disorders. Traditional manual segmentation methods are labor-intensive and time-consuming, often extending treatment durations and potentially affecting patient recovery times. By automating this process, we can significantly reduce the burden on radiologists and improve patient outcomes.

# Literature Review

Previous studies have investigated automated segmentation within the gastrointestinal (GI) tract. In this section, we will review some of the most significant research contributions to date.

TransUNet **[‎4]** is the first to explore the potential of transformers in medical image segmentation. Interestingly, the authors found that using plain transformer on tokenized image patches fails to provide sufficiently detailed segmentation, this is since the transformer treats the input as 1-dimensional sequences, focusing exclusively on global context. The combination of the transformer layers with a CNN (U-Net) mitigates these issues, providing a strong segmentation framework.



**Figure 2**: A block diagram of the TransUNet segmentation model. TransUNet uses a transformer block as a bridge layer between the encoder and the decoder.

Architecturally, A U-Net encoder-decoder topology is used, with a transformer serving as the bridge between the two parts. Between the encoder and the transformer layers, a learned linear projection is used in order to “tokenize” the input into patch embeddings before entering the transformer block, followed by the addition of positional encodings. The patches now enter the transformer component, which consists of 12 layers of MSA + MLP (Multihead self-attention and multi-layer perceptron, preceded by layer norm). The output is reshaped into the 3D tensor, a 1x1 convolution is used to reduce the feature depth before entering the decoder (vanilla U-Net decoder with bilinear upsampling). The loss function used is not specified in the paper, though a look at the code reveals it was set to

Dataset used for evaluation are:

Synapse - MICCAI 2015 MA Abdomen Labeling challenge dataset: Abdominal CT images. A random split of 18 training cases and 12 cases for validation is used. No test subsample was reserved for this (small) dataset.

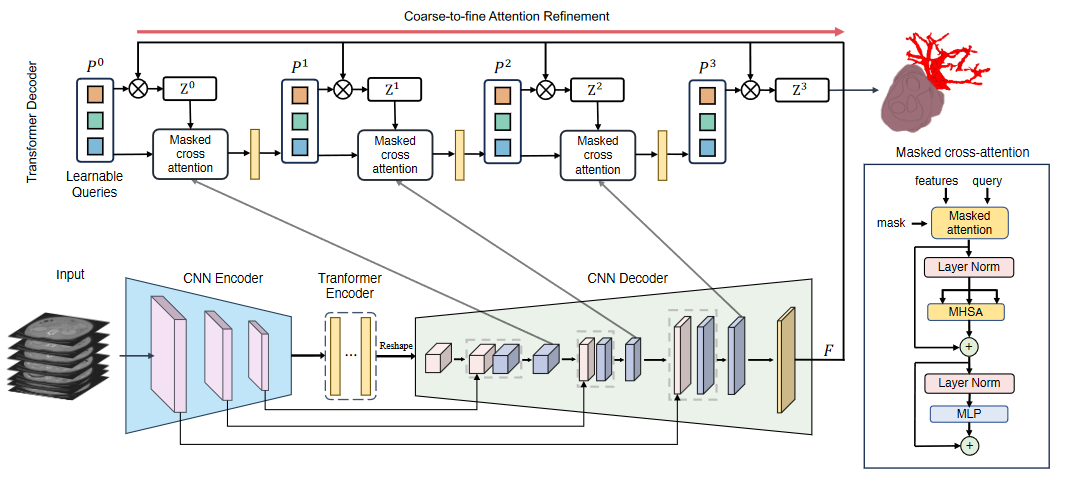
ACDC - Automated cardiac diagnosis challenge dataset: Cardiac MRI images. A random split of 70 training cases, 10 validation cases and 20 for testing is used.

Only simple data augmentation methods are used for all experiments: random rotations and flipping.

The transformer layers, as previously mentioned, are built upon the ViT (vision transformer) architecture. Combined with the transformer is a CNN, specifically ResNet-50, as the transformer backbone. Importantly, the encoder backbone (transformer and ResNet-50) are both pre-trained on ImageNet. Input resolution is set to 224 x 224, with patch size (for the transformer embedding) of 16. The model is trained by SGD with learning rate of 0.01, momentum of 0.9 and weight decay of 1e-4. Batch size is 24 and the training iterations are 14k and 20k for the Synapse and ACDC datasets respectively ().

The results show a measurable improvement from the non-transformer based U-Net architectures (V-Net, R50-Unet, AttentionUnet…), though the numbers show that transformer utilization at the time of TransUnet is still far from the technology’s potential. Indeed, at present, the following are already being explored: different transformer implementation, multi-scale skip connection enhancements, enhanced feature fusion (between the CNN and the transformer’s outputs) and other methods.

In 3D TransUNet **[‎5]** the authors of **[‎4]** consist of all the authors of the original TransUNet paper (and others), here they explore segmentation 3D medical images (CT and MRI scans, as opposed to individual slices in a regular U-Net and TransUNet).



**Figure 3**: A block diagram of the 3D TransUNet segmentation model. This method appends a transformer-decoder to the TransUNet topology.

As backbone, the authors specify the nn-Unet framework (unlike the original TransUNet, which used a ResNet-50 encoder). The image sequentialization process occurs in a similar fashion to TransUNet, though here the sequence consists of flattened 3D patches, and the equations are amended to accommodate that.

The encoder and the transformer bridge are nearly identical to the TransUNet variant, though using different projection matrices and convolution operations in order to process 3D input.

The decoder now contains an additional attention component (transformer decoder), newly added in 3DTransUNet. The specific operation is:

Here, are the previous layer’s features, are the previous layer’s query features, is the final decoder output and are the embedding matrices used to transformer the decoder features into the key and value matrices.

The transformer decoder helps extract and refine spatial information obtained within the transformer bridge of the network (and from the encoder’s feature space). This essentially turns the original U shape into a “W-Net”, with additional skip-connection stage from the decoder to the transformer decoder.

The loss function is also slightly modified from the original TransUNet,

CLS loss is defined as the cross-entropy loss for each candidate region. The loss is named “Hungarian matching loss”. are hyperparameters, set to 0.7 and 0.3 respectively.

Unlike in TransUNet, where each MRI slice was segmented separately and combined in post-processing, here the MRI scan is processed as a whole, though it is worth noting that the authors’ source code provides an option to use 3D TransUNet on 2D inputs.

4 datasets are specified for the evaluation section: BraTS2021, MSD vessel, BraTS and Pancreas.

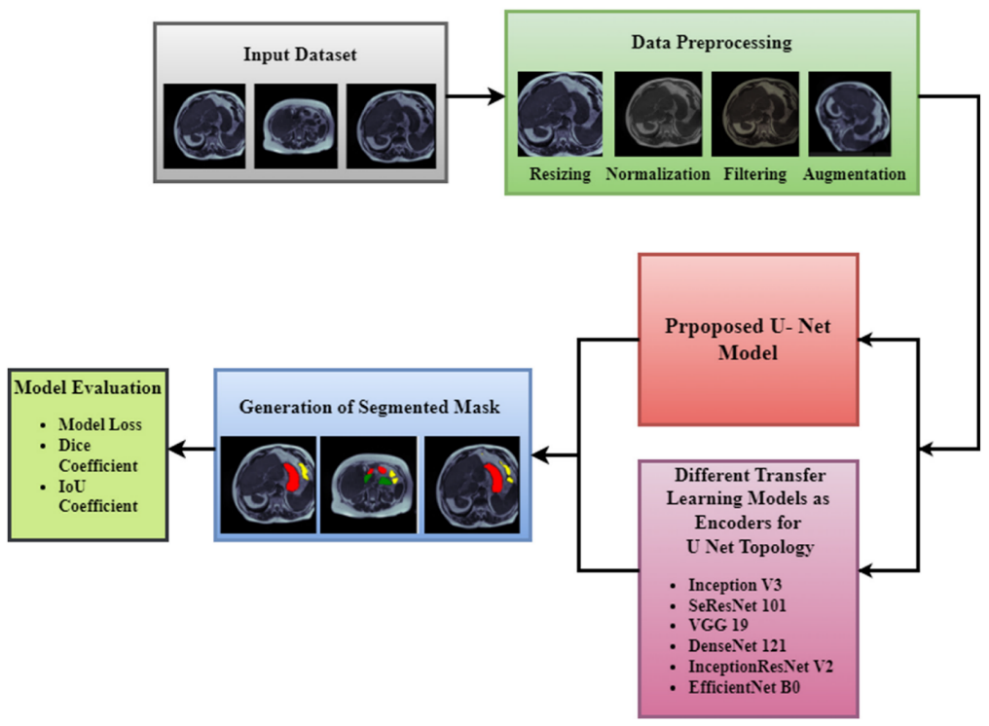
The settings used by the authors for implementing the model include a batch size of 2 (presumably limited by the large size of 3D inputs and by the GPU used), with various input sizes according to each dataset.

Extensive data augmentation is performed: random rotation, scaling, flipping, white Gaussian noise, Gaussian blurring, adjusting brightness and contrast, simulation of low resolution and Gamma transformations.

The results show that 3D TransUNet consistently leads against other 3D (and 2D) architectures, however, the improvements are in the 2-3% range. Further augmenting the nnU-Net framework with transformer component(s) seems to yield decent improvements. This idea seems to have been further explored [here](https://iopscience.iop.org/article/10.1088/1361-6560/ad0c8d) (abstract only, pay-walled article). Interestingly, the Pancreas dataset’s specificity seems to have significantly dropped without the decoder part during the ablation study (91 to 85 percent), implying that nn-Unet benefits from the modified transformer-based approach presented in the paper.

In **[‎6]** *Neha Sharma, Sheifali Gupta et al* first describe and review the issue of GI tract tumors, and the need for radiation oncologists to have accurate methods to guide X-Ray beams towards tumor tissue while avoiding healthy organs. The segmentation is performed on MRI slices of the GI tract.

The aim is to compare between a vanilla U-Net architecture and six different architectures for transfer learning: Inception V3, SeResNet50, VGG19, DenseNet121, InceptionResNetV2, and EfficientNet B0. These pre-trained models were implanted as the topology of the U-Net model (as the encoder part), while the decoder served as the trainable part of the network. The decoder thus uses the extracted features from each of these architectures’ different layers (via skip connections) in order to obtain a semantic context of the input image, while preserving spatial information.



**Figure 4**: Proposed Methodology for GI Tract Segmentation

The dataset used is the WM-Madison dataset ([link](https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation)) MRI scan dataset.

As part of the pre-processing scheme, four main operations were performed on the images:

Resizing: Images were resized to 160x160x1.

Gaussian filter: Used to blur certain portions of the image and to lessen noise (high-frequency components).

Normalization: Pixel values were divided by 255.

Augmentation: Used in order to increase the diversity of data, and increase the dataset’s size. The augmentation methods used are basic - only rotation and zoom.

The architecture used is a vanilla U-Net: 4 blocks down and 4 blocks up, with a middle block of two convolutional layers. The only difference between vanilla U-Net and the proposed method is the input (and subsequent layers) size.

Using this (vanilla U-Net built from scratch) model as a skeleton, the 6 different transfer learning architectures were then used as U-Net’s topology, and compared with the vanilla U-Net.

The different networks were implemented using the Keras framework. Among the hyperparameters described in the article, a batch size of 32 was used for all methods, with 20 epochs and a learning rate of 1e-4. The optimizer used is Adam.

In this article the vanilla U-Net somewhat drastically outperformed all transfer-learning methods, with validation loss of 0.122 compared to 0.418 for the best transfer-learning result (Inception V3). Dice score and IoU coefficient followed the same pattern (with Inception V3 in 2nd place).

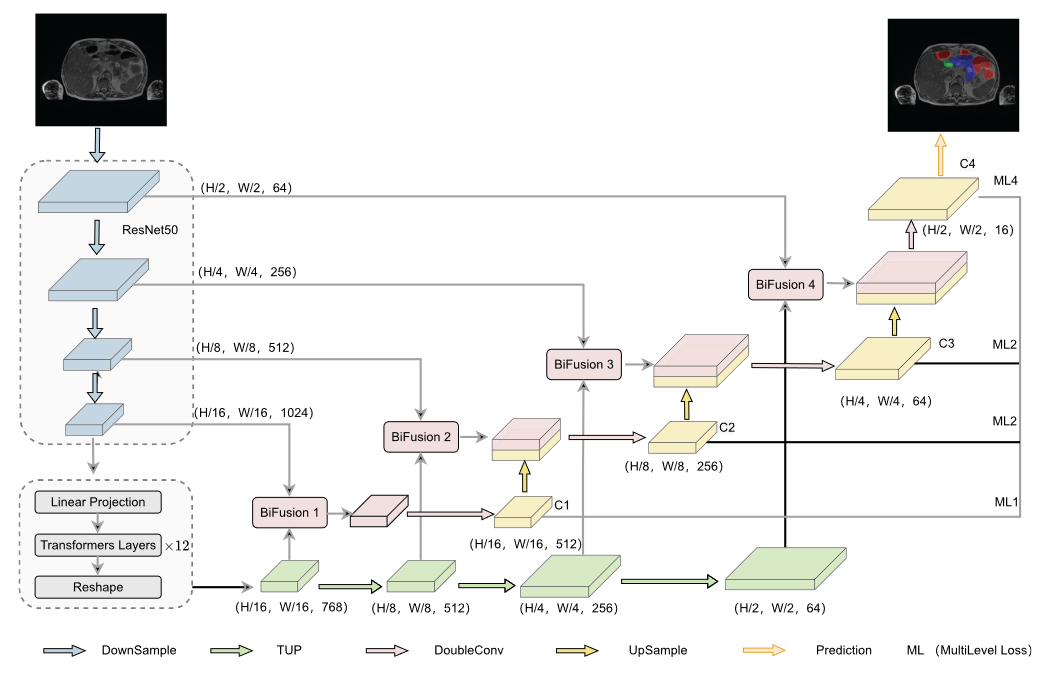
The results obtained in this article seem a bit odd, the [BiFTransNet](https://doi.org/10.1016/j.compbiomed.2023.107326) [**‎4**] article reported a lower IoU score with a much newer Transformer model on the exact same dataset, which seems unlikely given the measurable improvement even the simpler Attention U-Net provides over the vanilla architecture [‎**15**].

*Xin Jiang, Yizhou Ding et al* present **[‎7]** a method to perform GI tract segmentation using a combination of a transformer with a CNN, in order to improve upon convolutional networks limitations in learning global features. This is achieved by defining a fusion component, which effectively combines information from the transformer and the convolutional network. The proposed method improves upon existing transformer segmentation methods (specifically TransUnet) by the introduction of such “BiFusion block” – which combines information from both parts of the architecture, in addition to a multilevel loss strategy which can effectively guide the network’s training process.

The architecture is a U-Net skeleton with ResNet50 as a backbone for the convolutional part, which is then fed both into a 12-layer transformer block and to the fusion blocks (via U-Net’s skip connections). The BiFusion blocks’ output is then used as input for the multilevel loss function and up-sampled via the decoder path.

The BiFusion block consists of three sub-components: channel-attention, spatial-attention and a “multimodal fusion”, which multiplies the transformer and CNN outputs (Hadamard) prior to a convolutional layer. These three outputs are concatenated and pass a residual operation in order to fit the required decoder’s input.

The multilevel loss strategy acts as a weighted sum of the loss function, which is calculated over the final layer of each of the decoder blocks. The sum gives a large weight to the final layer, while not ignoring the previous ones. The loss function itself is an unbiased sum of BCE and Tversky loss functions. (Tversky loss is derived from the Tversky index, used as a similarity measure on sets).



**Figure 5**: A block diagram of the BiFTransNet segmentation model. The diagram shows the additional Bi-Fusion blocks, used by the model to combine transformer and CNN features.

As hyperparameters, the authors chose weights of 0.1 as the ML1 to ML3 multi-loss components’ coefficient, and 0.7 for the final layer (ML4). In addition, 0.5 was used as both Tversky constants. Batch size of 16 was chosen, with initial learning rate of 1e-2, weight decay of 1e-4, and minimum learning rate of 1e-6. The optimizer used was SGD, for 120 epochs.

The article specifies two datasets used to train and evaluate the proposed method: the UW-Madison GI segmentation dataset and the Synapse Multi-organ segmentation dataset. For data preprocessing, the images were resized to 224 by 224 (to accommodate input layer), and the following augmentation operations were used: flip, rotation, grid distortion, elastic transform and coarse dropout.

The model was then trained (on a single NVIDIA 3090 GPU).

Evaluation was performed using Dice coefficient, IoU, and Hausdorff distance. These metrics were then used to compare BiFTransNet with U-Net, FPN, Deeplabv3, Deeplabv3+ and TransUnet (the first transformer network for medical image segmentation, according to the authors). Also, the BiFTransNet is measured with and without transformer upsample process, multilevel loss and BiFusion blocks in an attempt to measure which was most effective. Finally, different multilevel loss hyperparameters were compared in order to reach the best choice.

The results show improvement upon existing methods, though the measured difference seems minor. Of note, the Dice coefficient for Gallbladder and Pancreas showed marked improvement over existing methods.

These models give us a good understanding about the landscape in our project’s field. Transformer-based U-Net seem to have provided the best models in GI tract segmentation [‎**4**], with a subsequent focus on combining the encoder features with the transformer’s output, and incorporating additional attention mechanisms to these features during the decoder stage. Additionally, fine tuning and data augmentation play a significant role in current advancements.

In conclusion, multi-level loss strategy and BiFusion can help push the medical segmentation area forwards.

# Background

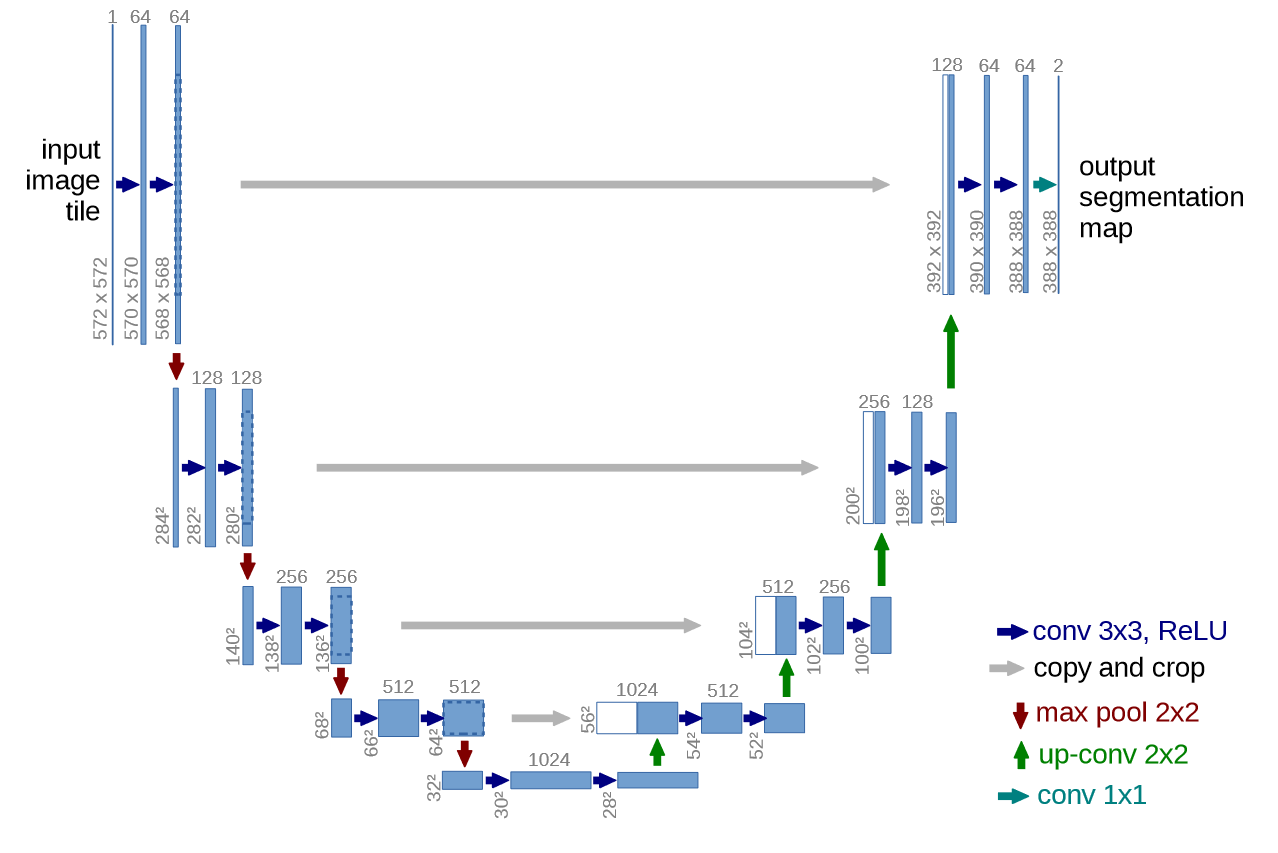
## U-Net

First proposed in a 2015 paper as a semantic segmentation model for medical images [‎**8**], U-Net has proven to be a strong performer in the field. Since then, U-Net has gained vast success as a segmentation model (or as an underlying topology) in many different fields, and even in unrelated areas (generative neural networks) [‎**9**, ‎**10**].

The U-Net architecture uses an encoder-decoder structure, using this concept to extract high-level semantic features from the image in the encoder part, and learning the spatial structure and upsampling these features towards the final segmentation mask in the decoder part of the network. Between these two paths U-Net utilizes several concatenation operations, known as the skip-connections. These operations allow the network access to spatial information from shallower layers, which is essential for accurate representation of semantic information while retaining global contexts.

As seen in the figure below, the network’s topology and its symmetric shape create a U-shaped path, which provided its namesake.

The vast improvements U-Net provided in image segmentation (medical imaging in particular) placed the network’s topology as the go-to segmentation framework in the medical field [‎**9**].



**Figure 6**: The original U-Net's topology

## The U-Net architecture

As previously discussed, U-Net utilizes the encoder-decoder architecture, with supplemental skip-connections. These are implemented by using the following blocks:

*Encoder*

The encoder path shrinks the input image in its spatial dimensions, allowing the model to extract relevant features using the convolutional layers. The encoder path is built in blocks: these blocks perform two main operations: down-sampling and a double-convolutional layer. The input is first downsampled through max-pooling and then passed through two convolutional layers with 3 by 3 kernels. Vanilla U-Net consists of four such blocks, which make up the encoder part of the network. At the end of each block, a skip connection is added which feeds its output to the respective decoder block. Each block increases the depth (features dimension) by a factor of two.

*Decoder*

The decoder’s goal is to build up the semantic information gained from the downsampling convolutional path (encoder) in size back into the original input’s dimensions. Likewise, the decoder is also built in blocks. This is done through a combination of upsampling and convolutional operations. Upsampling in vanilla U-Net is performed via transposed convolution (a convolution with fractional step, sometimes named de-convolution or up-convolution) – a method that allows the model to learn upsampling parameters (weights) during training. However, it is worth noting that there are many different techniques for CNN upsampling, examples include bilinear, trilinear, and “un-pooling”. Here the skip connections are implemented, in the form of a concatenation operation: the respective encoder features are concatenated to the previous decoder block’s output and then fed-forward to the current block. Each decoder block (symmetrically) decreases the features dimension by a factor of two.

Finally, a 1x1 convolution is performed, in order to fit the amount of segmentation masks to the number of classes in the dataset.

## Advantages and limitations

The U-Net framework provided plenty of advantages to the image segmentation field:

Efficiency: U-Net uses a constricting path, which greatly mitigates thee cost of convolutional operations over high spatially dimensioned inputs. Moreover, the skip-connections help provide a more detailed spatial information from earlier blocks, sparing the model additional work to re-capture said information.

Simplicity: U-Net consists of a small amount of relatively simple operations (as seen in the figure above), thus providing an easy skeleton to build around and append additional components to.

Accuracy: U-Net’s accuracy far surpassed the previous medical segmentation models, thanks to its design and its skip connections.

However, U-Net also faced many limitations, as it was still far away from ideal performance in semantic segmentation (in terms of accuracy).

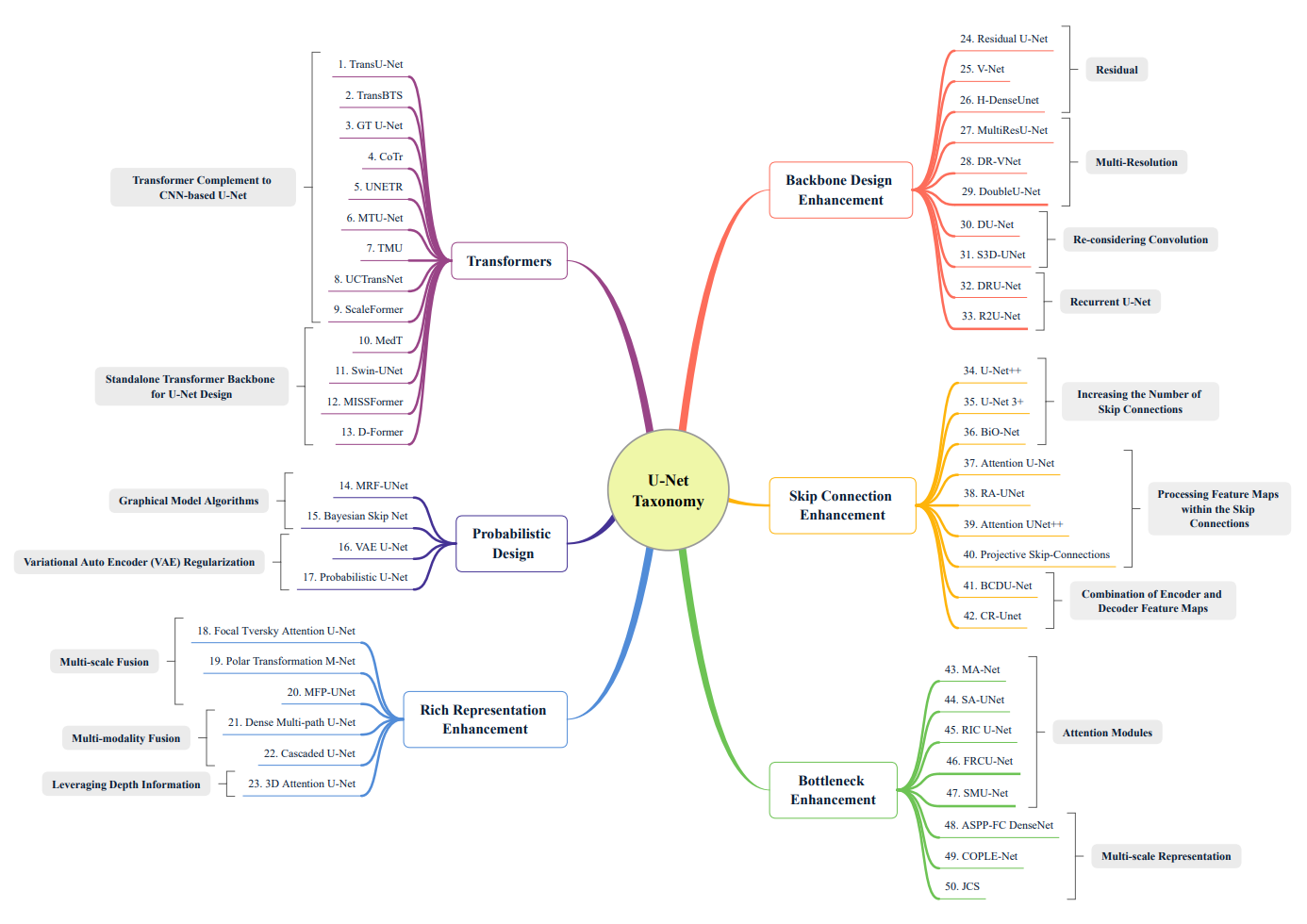
Due to the nature of the U-Net topology (shrinking path), some degree of information loss is guaranteed. This is somewhat mitigated by skip-connections, but these are inadequate to fully overcome the issue.

U-Net's design limits the information flow in the network. Due to the convolutional operation’s limited receptive field (and their inductive bias [**‎7**], U-Net lacks the ability to model long-range feature dependencies.

The pooling and convolutional operations can prevent low-level features from being transmitted to subsequent convolutional layers, causing loss of local information quality.

The skip connections’ placements are limited to same-scale layers, and thus fail to explore the relationship between feature maps from different decoding stages. Hence, U-Net cannot guarantee the consistency of the feature representations and semantic embeddings.

Many different improvements and modifications were introduced to the U-Net architecture over the years, with the incorporation of the Attention block, and its future variant (the transformer), being among the most influential [‎**11**]. Attention has since revolutionized the field of ML, computer vision and NLP, providing some of the best results seen in recent years, in many different tasks, including semantic segmentation [‎**11**, ‎**13**].



**Figure 7**: A partial graph of influential U-Net variants.

## Attention

Attention is a mechanism which attempts to mimic cognitive attention. The goal here is to obtain a weights map over the input which contains additional context based upon which regions (or words, in language models) are more relevant, in addition to information about relationships between two patches in the input.[‎19]

Historically, recurrent neural networks (RNNs) were used to capture spatial contextual information. These, however, faced issues in both accuracy (hidden outputs) and performance (owing to their recursive nature).

In 2017, the now famous article “Attention is all you need” [**‎11**]

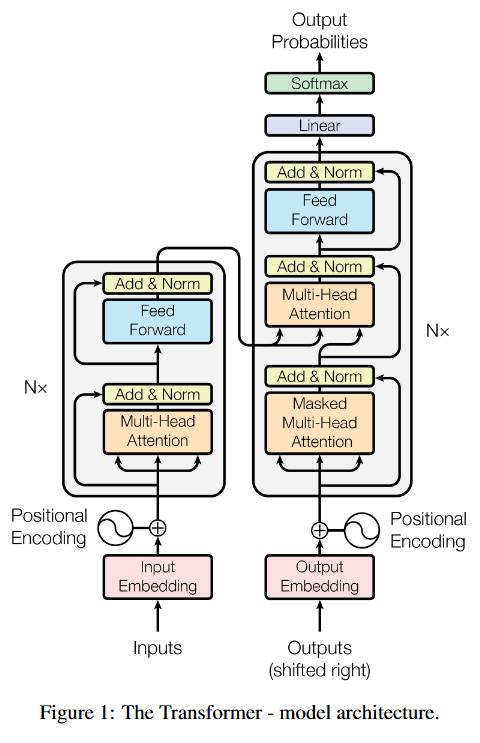
provided the attention concept in full, with an efficient and ground breaking implementation – the transformer. The Attention function is the transformer's main operation - the transformer calculates it in parallel across the tokenized input space and uses additional blocks to provide the versatility of CNNs with excellent results. Before looking at the transformer, we will discuss this operation.

The Attention map is obtained by calculating an attention function, the most common being the “Scaled-Dot-Product-Attention”:

The input is tokenized via embedding matrices to convert it to three different inputs for the attention function – queries, keys and values. Intuitively, the attention function treats the transformed patches as a database query: first the query is associated with the correct keys, and then the values are accessed according to the relevant keys. The query-key similarity is algebraically calculated via the dot-product operation.

An important note – in conventional CNNs we often see the softmax function in the end, as the loss function (or part of it), here we have it in the middle of a block inside the network, as part of the attention function. Hence, normalization (dividing by square root of the query-key dimensionality) is performed in order to keep the mean and variance of the original input, thus avoiding vanishing gradients for this function.

The attention function is then used to guide the network’s weights through an effective heat-map (attention map) which helps the rest of the model focus on important regions.



**Figure 8**: The transformer diagram. The transformer uses positional encodings to extract inter-dependencies and improve the input embedding, using Attention blocks.[‎11]

As seen in the figure above, the vanilla transformer computes the multi-head attention in each layer, in both the encoder and decoder. This operation consists of many scaled-dot-product-attention operations running in parallel. Vision transformers (ViTs) discard the decoder part, using the encoder alone in order to capture spatial information from the tokenized image.

## From language models to images

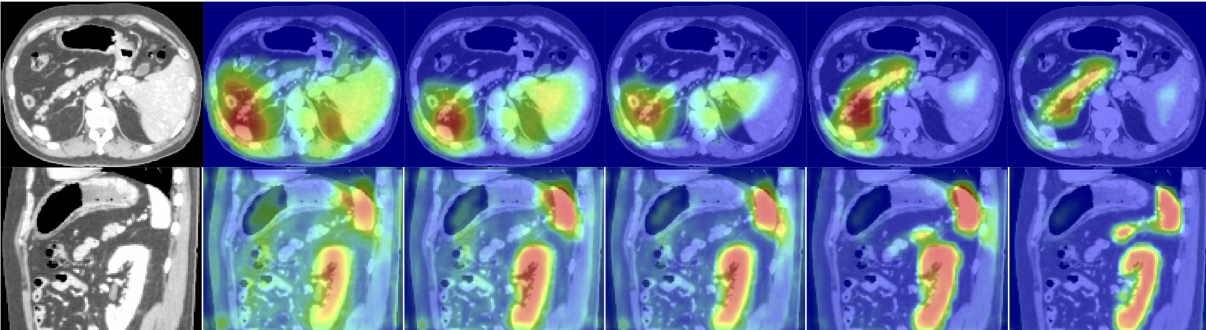
Though we discuss attention in computer vision, the concept first appeared in natural language models. The concept of attention, first developed in natural language processing (NLP), has significantly transformed the way models understand and interpret visual data in computer vision tasks. Attention mechanisms in NLP models allow the model to focus on specific parts of the input sequence when generating the output. This mechanism has been crucial in improving the interpretability and performance of NLP models by enabling them to capture long-range dependencies and relationships within the data [‎**11**, **‎13**]

The success of attention mechanisms in NLP led to their adoption in computer vision tasks, where they have been instrumental in addressing the challenges of understanding complex visual scenes and objects. The introduction of attention gates and transformers in computer vision as early as 2018 marked a significant shift in the field, enabling models to focus on relevant parts of an image or video frame, thereby improving the accuracy and efficiency of tasks such as object detection, image segmentation, and scene understanding [‎**14**].

The transformer architecture, which relies heavily on self-attention mechanisms, has been particularly influential in computer vision. Unlike convolutional neural networks (CNNs) that process images in a fixed, localized manner, transformers can model global dependencies in data, making them well-suited for tasks that require understanding the entire context of an image. This has led to the development of models like Vision Transformers (ViT), which apply the transformer architecture directly to images, treating them as a sequence of patches and leveraging self-attention to capture global contextual information [‎**14**].

ViTs recently emerged as a competitive alternative to Convolutional Neural Networks (CNNs) that are currently state-of-the-art in different image recognition computer vision tasks. The ViT model architecture was introduced in a research paper published at ICLR 2021 titled “An Image is Worth 16\*16 Words: Transformers for Image Recognition at Scale” [‎**14**].

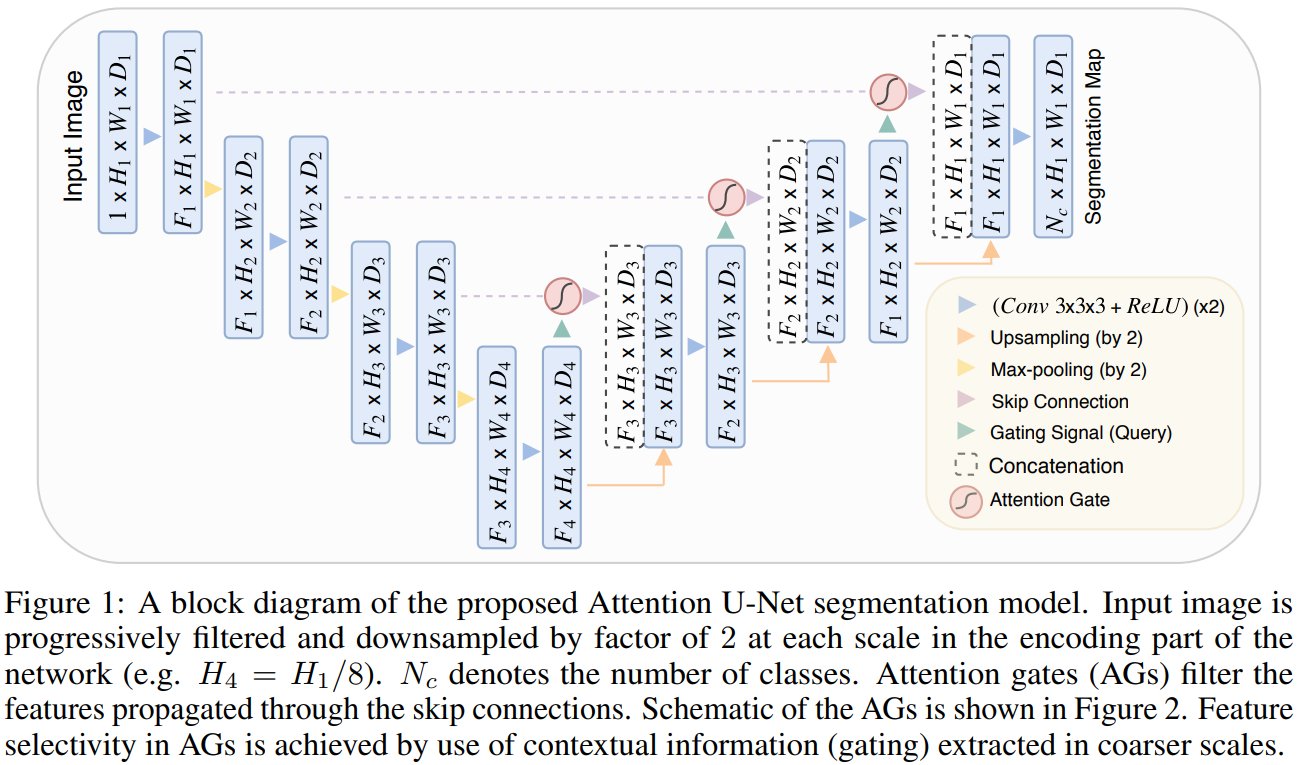
The adoption of attention mechanisms in computer vision has not only improved the performance of models on various tasks but has also opened new avenues for research and development in the field. As researchers continue to explore the capabilities and limitations of these mechanisms, it is expected that attention-based models will continue to evolve, offering new solutions to complex visual problems and pushing the boundaries of what is possible in computer vision [‎**14**]



**Figure 9**: The attention coefficients in the Attention U-Net model during different training epochs. The attention mechanism guides the model to focus on the relevant organs (pancreas, kidney, and spleen) [**‎15**].

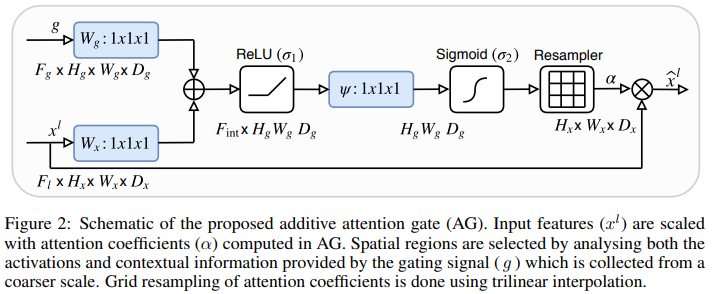
## Attention U-Net

Attention-U-Net [‎15], presented in 2018, uses a relatively simple attention mechanism: the Attention gate.



**Figure 10**: A block diagram of the U-Net segmentation model. Input image is progressively filters and downsampled by factor of 2 at each scale in the encoding part of the network. Attention gates filter the features propagated through the skip connections. Schematics of the attention gates is shown in Fig. 11. Feature selectivity in attention gates is achieved by use of contextual information extracted in coarser scales.

In the Attention U-Net model, Attention Gates (AGs) play a pivotal role in enhancing the accuracy of image segmentation tasks, particularly in medical imaging. These AGs are designed to emphasize significant features that are transmitted through skip connections, as shown in Figure 1. By utilizing data extracted from a lower resolution, AGs effectively distinguish between pertinent and irrelevant or noisy signals in the skip connections. This process occurs just before the concatenation step, ensuring that only relevant activations are combined.

  
**Figure 11**: Schematics of the additive attention gate

AGs not only filter neuron activations during the forward pass but also during the backward pass. During the backward pass, gradients from background areas are reduced in weight. This mechanism allows for the adjustment of model parameters in deeper layers, based on spatial areas relevant to the task at hand.

The integration of AGs into the U-Net model significantly enhances the model's ability to focus on target structures while suppressing irrelevant regions in input images. This is particularly beneficial in medical imaging, where identifying specific features is crucial for accurate diagnosis and treatment planning. By highlighting salient features only, AGs eliminate the need for explicit external modules for tissue or organ localization in cascaded convolutional neural networks (CNNs). This approach not only increases the sensitivity and precision of the model but also reduces the need for overt exterior tissue/organ localization units of cascading convolutional neural networks (CNNs), making the model more efficient and reliable for medical applications [‎**16**].

## Transformers in U-Net

In recent years, the integration of transformers and other sophisticated attention components into medical segmentation models has significantly advanced the field. These advancements have been driven by the success of transformers in various domains, including natural language processing, and their potential to capture long-range dependencies in data, which is crucial for medical image segmentation tasks.

The U-Net architecture has been a cornerstone in medical image segmentation. However, its convolution-based operations inherently limit its ability to model long-range dependencies effectively. To address these limitations, researchers have turned to Transformers, renowned for their global self-attention mechanisms, as alternative architectures.

The integration of a Transformer-based encoder and decoder into the u-shaped medical image segmentation architecture has shown significant potential. The Transformer encoder excels in multi-organ segmentation, where the relationship among organs is crucial. On the other hand, the Transformer decoder proves more beneficial for dealing with small and challenging segmented targets such as tumor segmentation. Extensive experiments have demonstrated that transformer U-Nets outperform competitors in various medical applications, highlighting the effectiveness of transformers and other sophisticated attention components in medical segmentation [**‎4**].

# Expected Achievements

After performing a literature review and gaining an understanding regarding GI tract MRI segmentation, our aim is to build, train and validate a segmentation system using the U-Net architecture, utilizing a transformer-bridge based approach (i.e. TransUNet, BiFTransNet) that surpasses current SOTA methods in this field.

Using the UW-Madison dataset, we will train the model to accurately segment tumor tissue, the stomach and the intestines, providing safe margins for radiology treatments.

We intend to reach this goal through a combination of the best architectural choices and hyperparameters fine-tuning. Architecturally, current literature indicates that SOTA methods focus on fine-tuning the fusion step in the network, where CNN and transformer features are combined. We will investigate and use the most effective methods for this goal. Additionally, fine tuning will entail using multi-level loss function, and exploring transformer-specific hyperparameters.

A successful project will entail:

* Improvement in accuracy measurements: DICE and IoU scores which surpass current SOTA (BiFTransNet).
* Choosing the best fusion operations and hyperparameters in our model, in order to obtain a substantial improvement over current methods.
* Provide a working system which not only accurately segments MRI images, but performs said operation in a timely manner (acceptable inference time).

## Research Challenges

The GI tract is in a dense region in the human body, with many similar-looking and overlapping organs, requiring precise analysis to decide a proper region for radiation treatment. Accurate segmentation during radiation treatment is subject to slice-to-slice image variability (noise and artifacts), and a constantly changing landscape – as the therapy affects both tumor and healthy tissue. Additionally, accurate segmentation can be affected by MR images quality, which varies depending on system models, manufacturers, sequence parameter settings, or field shimming conditions.

# Research Process

## Our proposed Process

The project’s research began with reviewing GI tract cancer – we learned about the nature of the disease and the different diagnostic and treatment methods. Our research focused on part of the treatment termed “Image-guided radiation therapy”, wherein medical imaging techniques are used during, or just prior to administering radiation treatment.

MR-Linac is a magnetic resonance-guided linear accelerator that combines MRI with radiation therapy to target and treat cancers. MRI guidance allows doctors to adjust the radiation therapy in real time with better soft tissue resolution and deliver it more accurately and effectively than ever before.

In order to create a segmentation model capable of sufficient accuracy in this task, we intend to use the U-Net architecture as the basic topology, with a transformer component as the encoder-decoder bridging block. Following recent advancements, we intend to employ a fusion strategy to capture and effectively combine the semantic information from the U–Net encoder output and the global spatial relationships embedded in the transformer outputs.

As previously discussed, U-Net first extracts semantic information and then up-samples it into a segmentation mask. The transformer (when used as encoder-to-decoder bridge component) expands the local-global context in the original encoder feature maps. This process, however, does not manage to fully mitigate the architectural information loss of U-Net – hence, additional paths (multi-scale skip connections) and attention mechanisms (feature fusion blocks) will be introduced to the architecture to further alleviate this issue.

Reviewing the medical image segmentation field, we learned that transformer-based models have already pushed results ahead. Using a transformer U-Net as our base model, current research provides the following avenues for improvements:

* Expanding the network architecture to include multi-scale context, utilizing additional skip connections.
* Alternatively, appending a coarse-to-fine attention refinement component to the model’s decoder segment, as seen in 3D-TransUNet.
* Multi-level loss function which will help guide the training process, as explored in BiFTransNet. Our aim here is to study the effect of a dynamic multi-level loss function, with hyperparameter adjustment during the training process.
* We will incorporate a combination of the Dice loss function with BCE, as seen in TransUNet, and compare its performance to a combination of Tversky and BCE loss function (from BiFTransNet).

As success criteria we will use the IoU (intersection over union ratio), the Dice coefficient (as a measure of similarity) and the HD (Hausdorff Distance) measures. We will follow both the per-organ values and the averages.

Additionally, we intend to tune available hyperparameters to provide the best evaluation metrics. The batch sizes used in recent literature were, at most, 16 and 24. Considering current video memory and time limitations, we intend to explore batch size values of 16, 32 and 64. We intend to use a learning rate decay strategy, as seen in recent work in the field. We will explore the following initial learning rates: 1e-2, 8e-2, 3e-4 (seen in TransUNet and BiFTransNet), using a cosine annealing function to gradually decrease it (lr schedule = cosine). Here we will also explore finer minimum learning rate values.

Our dataset, UW Madison GI tract dataset, is rather limited, with 85 cases in total – hence we will use data augmentation techniques to adequately expand it (as seen in recent work). For data augmentation techniques, we intend to leverage the following (inspired by the authors of BiFTransNet and 3DTransUnet): flip, rotation, grid distortion, elastic transform, coarse dropout, white Gaussian noise, Gaussian blurring, adjusting brightness and contrast, simulation of low resolution and Gamma transformations.

## Challenges and achievements

Our challenges working on this project can be separated into two types – skill related and field related challenges.

Regarding our knowledge level: both of us learned a lot in the recent seminar course we took together, where we researched and presented a transformer U-Net based network for medical segmentation. This achievement tempered our lack of expertise and familiarized us with the metrics, terms and tools used in medical image segmentation, and provided us with a more-than-adequate baseline for this work. Moreover, the academic course “Data mining and machine learning” further prepared us for the tasks at hand.

An additional challenge we face is in the project field’s properties. Firstly, the medical discipline is among the most complex and unforgiving in the modern world, we face many difficulties, despite the seminar preparation, in efficiently and fully understanding the barrage of medical terms and their interconnectedness. Secondly, the computer vision field is ever-changing, with recent advancements which require both a strong mathematical background and a capable abstract thought process for sufficient understanding.

During the literature review and the rest of our research process, we managed to gain an understanding and familiarize ourselves with many of the various idiosyncrasies in the medical image segmentation field, and in the world of transformer-based segmentation networks.

## Dataset

Our project uses a single dataset – the UW – Madison GI Tract segmentation dataset. Each case is represented by multiple sets of scan slices, identified by the day the scan took place. A case consists of scan performed over multiple days. The dataset was derived from multiple abdominal MRI scans from 85 different cases, for a total of 38,496 slices. It includes annotations for three organs: the large bowel, small bowel, and stomach. Out of the total slices, 21,906 have annotations for these organs, while the rest are labeled as background, owing to the absence of visible annotations. Although the number of slices is large enough, the case-to-case diversity is rather low, with 85 cases in total. Hence, we will use various data augmentation techniques to increase the training data’s size.

## Hyperparameters

During this stage we will find the best hyperparameter choices for our segmentation model. Here we will train our model using different combinations of our hyperparameter sets, comparing convergence speed and model accuracy. The following hyperparameters will be considered:

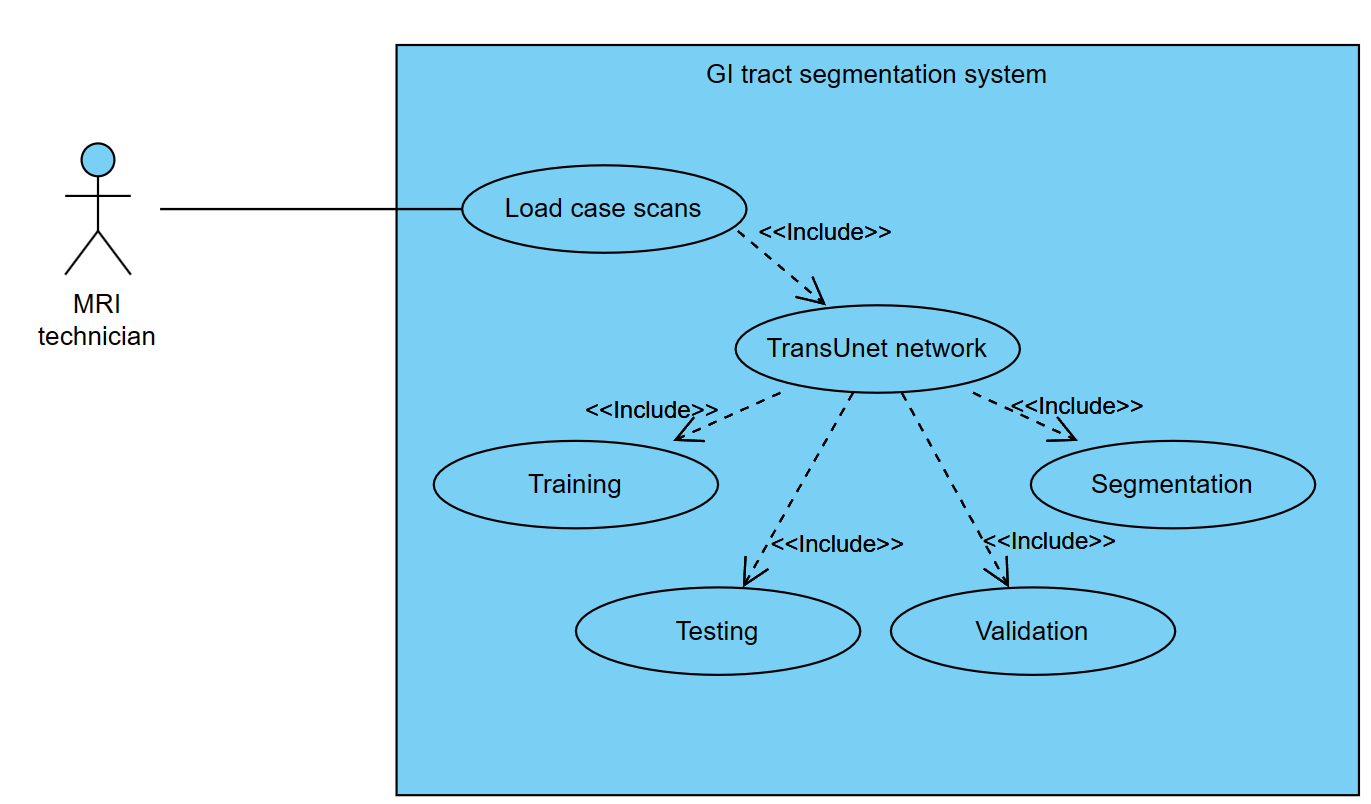
Initial learning rate: {1e-2, 8e-2, 3e-4}

Batch size: {16, 32, 64}

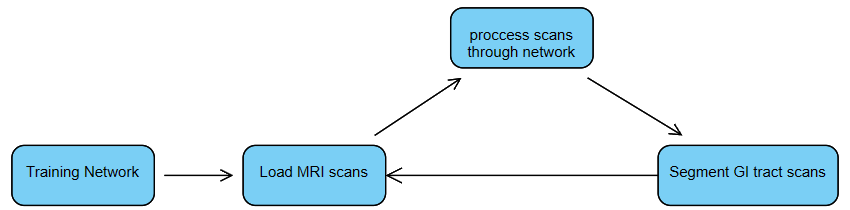
Minimum learning rate: {1e-5, 1e-6, 1e-7}

# Product

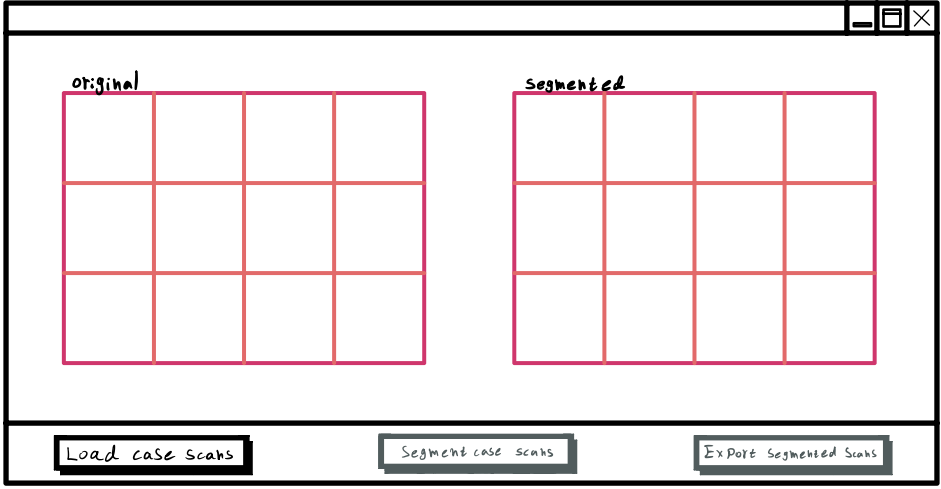
## Use case

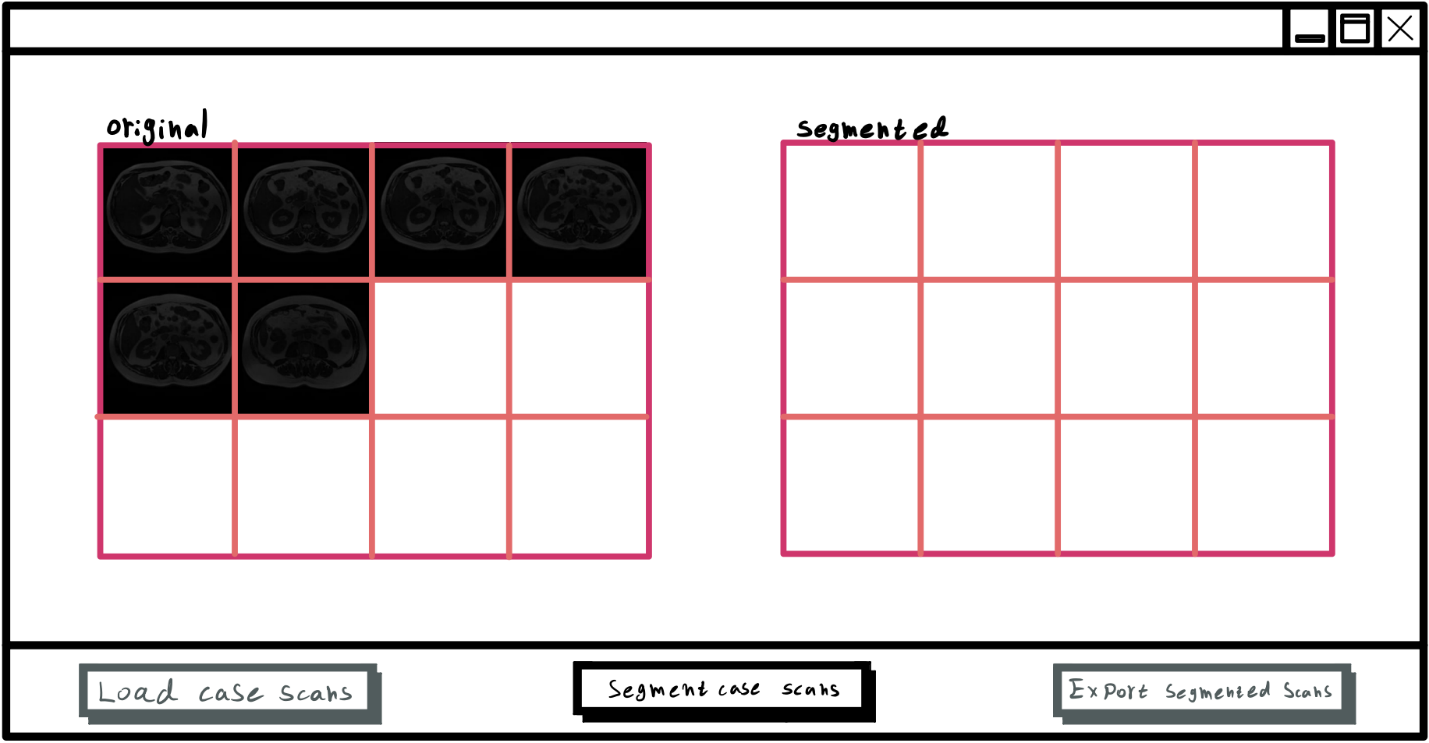
**Figure 12:** Use case diagram for GI tract segmentation system

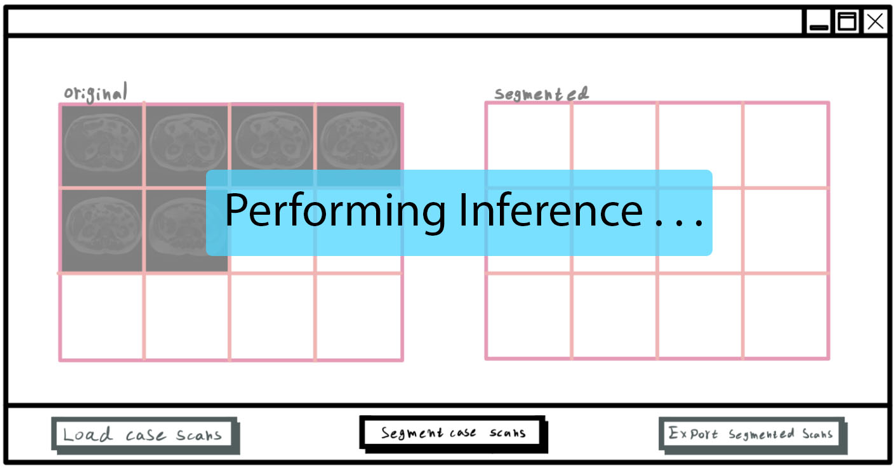
## Flow Chart

**Figure 13:** Training stage flow chart diagram

## GUI

**Figure 14:** GUI at initial stage

**Figure 15:** GUI at 2nd stage – MRI case scan selection

**Figure 16:** GUI paused during inference

**Figure 17:** GUI at final stage – MRI case scans segmented

# Evaluation/Verification Plan

In order to test our program, we have come up with some test cases detailed in Table 1 below.

|  |  |  |
| --- | --- | --- |
| # | Test Case | Expected Result |
| 1 | Click ‘Load Case Scans’ button | A file browser will pop-up, after files were chosen, GUI will advance to 2nd stage. |
| 2 | Click ‘Segment Case Scans’ button | The GUI will cease to respond to any input, a sort of spinning wheel will appear indicating inference is taking place, once done inferencing the spinning wheel disappears and GUI advance to 3rd stage. |
| 3 | Click ‘Export Segmented Scans’ button | A file browser will pop-up, once a folder is chosen the segmented files will be saved in that folder. |
| 4 | Clicking on a disabled button | No action. |
| 5 | Attempt to load unsupported image format (or non-matching dimensions) | Error message will pop-up and GUI will not advance to the next stage. |

Table 1: Suggested use cases for program validation

# References

1. Jardim, S.R.; de Souza, L.M.P.; de Souza, H.S.P. “The Rise of Gastrointestinal Cancers as a Global Phenomenon: Unhealthy Behavior or Progress?” Int. J. Environ. Res. Public Health 2023, 20, 3640. <https://doi.org/10.3390/ijerph20043640>
2. happyharrycn, Maggie, Phil Culliton, Poonam Yadav, Sangjune Laurence Lee. (2022). UW-Madison GI Tract Image Segmentation. Kaggle. <https://kaggle.com/competitions/uw-madison-gi-tract-image-segmentation>
3. Vincenzo Catalano, Roberto Labianca et al. ELSEVIER Critical Reviews in Oncology/Hematology Volume 71, Issue 2, August 2009, Pages 127-164. <https://www.sciencedirect.com/science/article/abs/pii/S1040842809000043>
4. Jieneng Chen, Yongyi Lu, Qihang Yu et al. “Transformers Make Strong Encoders for Medical Image Segmentation” arXiv:2102.04306v1 [cs.CV] 8 Feb 2021. <https://arxiv.org/pdf/2102.04306>
5. Jieneng Chen, Yongyi Lu, Qihang Yu et al. “3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers” arXiv:2310.07781v1 [cs.CV] 11 Oct 2023. <https://arxiv.org/pdf/2310.07781>
6. Neha Sharma, Sheifali Gupta et al. “U-Net Model with Transfer Learning Model as a Backbone for Segmentation of Gastrointestinal Tract” Bioengineering 2023. <https://www.mdpi.com/2306-5354/10c>
7. Xin Jiang, Yizhou Ding et al. “unified and simultaneous segmentation network for gastrointestinal images of CT & MRI”. ELSEVIER Computers in Biology and Medicine Volume 165, October 2023, 107326. <https://doi.org/10.1016/j.compbiomed.2023.107326>
8. Olaf Ronneberger, Philipp Fischer, Thomas Brox. “U-Net: Convolutional Networks for Biomedical

Image Segmentation”. arXiv:1505.04597v1 [cs.CV] 18 May 2015. <https://arxiv.org/pdf/1505.04597>

1. Reza Azad , Ehsan Khodapanah Aghdam , Amelie Rauland, Yiwei Jia et al. “Medical Image Segmentation Review: The Success of U-Net”. arXiv:2211.14830v1 [eess.IV] 27 Nov 2022. <https://arxiv.org/pdf/2211.14830>
2. Krithika Devi, K. Suganthi. “Review of Medical Image Synthesis using GAN Techniques”. ITM Web Conf. Volume 37, 2021. <https://www.itm-conferences.org/articles/itmconf/abs/2021/02/itmconf_icitsd2021_01005/itmconf_icitsd2021_01005.html>
3. Ashish Vaswani, Noam Shazeer et al. “Attention Is All You Need”. arXiv:1706.03762v7 [cs.CL] 2 Aug 2023. <https://arxiv.org/pdf/1706.03762>
4. Soufyane Ayanouz, Anouar abdelhakim Boudhir, Mohamed Ben Ahmed. “Unleashing the Power of NLP and Transformers: A Game-Changer in Medical Research and Clinical Practice and a revolution of Medical Text Analysis.: Case Study: Cancer report classification by priority”. NISS 2023. <https://dl.acm.org/doi/abs/10.1145/3607720.3607787>
5. Firuz Kamalov, David Santandreu Calonge, Ikhlaas Gurrib. “New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution”. arXiv:2305.18303v2 [cs.CY] 25 Oct 2023. <https://arxiv.org/pdf/2305.18303>
6. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. arXiv:2010.11929v2 [cs.CV] 3 Jun 2021. <https://arxiv.org/pdf/2010.11929>
7. Ozan Oktay, Jo Schlemper, Loic Le Folgoc et al. “Attention U-Net: Learning Where to Look for the Pancreas”. arXiv:1804.03999v3 [cs.CV] 20 May 2018. <https://arxiv.org/pdf/1804.03999>
8. Ahmed AL Qurri, Mohamed Almekkawy. “Improved UNet with Attention for Medical Image Segmentation”. MDPI. <https://www.mdpi.com/1424-8220/23/20/8589>
9. Salman Khan, Muzammal Naseer et al. “Transformers in Vision: A Survey”. arXiv:2101.01169v5 [cs.CV] 19 Jan 2022. <https://arxiv.org/pdf/2101.01169>
10. Noel E. Donlon, Robert Power et al. “Radiation and Immunotherapy in Upper Gastrointestinal Cancers: The Current State of Play”. MPDI. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7865314>
11. Grace W. Lindsay et al. “Attention in Psychology, Neuroscience, and Machine Learning”. PubMed. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7177153/>