AI604 term project rebuttal / author response

1. Additional question

1.1. Question 1

- The reviewer 1 understood our main idea and contributions. **Yes**
- The reviewer 2 understood our main idea and contributions. **Yes**
- The reviewer 3 understood our main idea and contributions. **Yes**
- The reviewer 4 understood our main idea and contributions. **Yes**
- The reviewer 5 understood our main idea and contributions. **Yes**
- The reviewer 6 understood our main idea and contributions. **Yes**

1.2. Question 2

- The reviewer 1 provided constructive feedback or advice. **No**, no suggestion was given.
- The reviewer 2 provided constructive feedback or advice. **No**, no suggestion was given.
- The reviewer 3 provided constructive feedback or advice. **Yes**, we will include computation costs and GPU memory comparison experiments in our following reports. The reviewer also suggested including Transformer architecture diagram, but we shall decide according to the length of the final version of the paper. We made changes to the diagrams to make it clearer (Figure 1 2 3). We will be providing better explanations for the training procedure in the upcoming versions.
- The reviewer 4 provided constructive feedback or advice. **Yes**, the explanations on how the segmentation reconstruction head works will be provided in the later versions. Our approach will be mostly convolution-free (mathematical operation-wise), and we are still investigating various methods for reconstruction. The experiment setup will be further organized and released accordingly. Note that the experiments provided in the progress report were only preliminary, and more experiments may be conducted with different settings.
- The reviewer 5 provided constructive feedback or advice. **Yes**, the reviewer asked some questions. For the first question, pixel-by-pixel accuracy is sometimes used as evaluation metric in segmentation tasks. It is often provided as

$$\text{Acc} := \frac{TP + TN}{TP + TN + FP + FN}$$

However, intersection-over-union (IoU) or other metrics designed for segmentation are often preferred because pixel-by-pixel accuracy can provide misleading results when the class representation is small within the image. This is because the measure can be biased in reporting how well the model identify non-present class. In our particular dataset, there are some small regions in the image, which can lead to misleading metrics.

For the second question, domain adaptation and transfer learning are often used interchangeably in computer vision literature. Even though it is not entirely correct, but we follow through with the general academic consensus. In our case with medical image tasks, domain adaptation can be helpful in a case where the model needs to be adopted in different hospitals. The model is expected to perform equally well with semantically similar but distributional different images, which is often the case in real-life circumstances.

- The reviewer 6 provided constructive feedback or advice. **Neutral**. The author suggests some different variants of ViT such as DeiT and CaiT. Some of these model setups are already being investigated for the upcoming progress report. The author also says a large private dataset is needed for training ViT, but we have already observed nice performance in some of the preliminary experiments with both pretrained and non-pretrained models. We also plan to experiment with the recently proposed MLP Mixer, convolution- and attention-free model for images.

1.3. Question 3

Redrawn figures in response to Reviewer 3 are provided below:

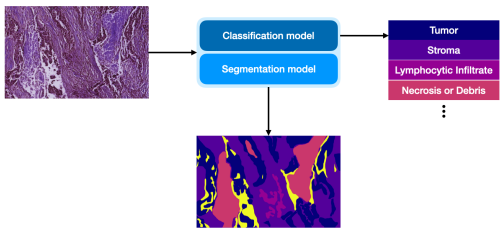


Figure 1: Block diagram of the model for WSIs segmentation and classification

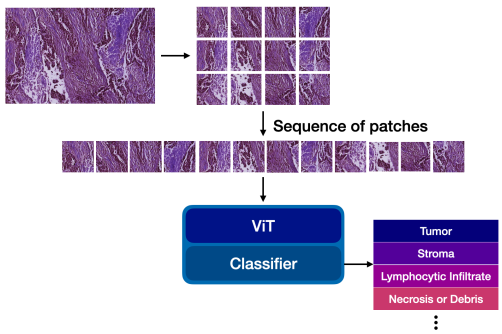
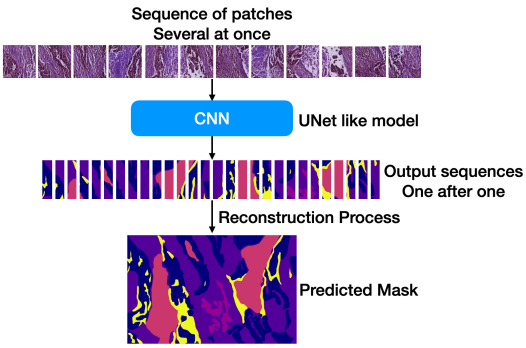
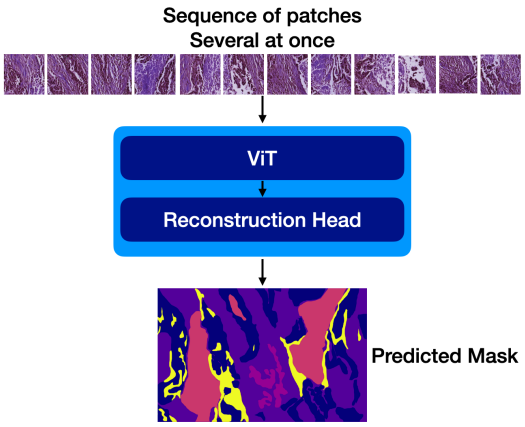


Figure 2: Block diagram of the model for WSIs classification



(a) Previous patch-based approach



(b) Proposed segmentation architecture

Figure 3: WSI segmentation model diagram