Leveraging Gram Matrix Representation in Shallow DNN for GI Tract MRI Image Segmentation

Fariza Siddiqua
Dept. of EEE
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
farizasiddiqua@gmail.com

Tamim Ahmed
Dept. of EEE
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
tamim.ahmed.tamim@gmail.com

Mohammed Imamul Hassan Bhuiyan
Dept. of EEE
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
imamul@eee.buet.ac.bd

***Abstract*—GI tract MRI image segmentation is important for the diagnosis and treatment of many diseases. In case of GI tract cancer or tumor treatment, radio oncologists must apply X-ray beams pointing towards the tumor cell while avoiding the other organs. The traditional segmentation process for the MRI scan is time consuming and labor intensive. A computer aided fast and accurate method is required. UNet, UNet++, Attention UNet are some of the deep learning architectures used in this purpose having remarkable performance. Most of these architectures inherit significant computational complexity. In this paper, a gram matrix oriented shallow DNN is proposed which will bypass the traditional attention mechanism. Due to the simplified computation of the proposed model, it yields similar or superior segmentation performance as compared to existing models and it has much lower number of trainable parameters in the range of 10 to 75 percent less the other models.**

***Index Terms*—U-Net, Gram Matrix, GI tract, MRI, Semantic segmentation**

I. Introduction

Deep learning algorithms have brought about revolutionary changes in medical image segmentation. Segmentation of different organs or parts is a fundamental task for the treatment and diagnosis of many diseases and abnormalities. In this paper, we focus on GI tract MRI image segmentation for better and faster treatment of gastrointestinal cancer or tumor cells. When radio oncologists need to apply X-ray beams to the tumor cells in GI tract, they must segment stomach and intestines from the tumor cells. Manual segmentation in MRI image of GI tract takes about 15 minutes to hours which prolongs the treatment. Therefore, there is a high demand of automatic image segmentation method to save time and make the segmentation process more accurate. In case of GI tract image segmentation, this is a type of multi-level semantic segmentation. Deep learning architectures take the lead of this task. U-Net [10], Attention U-Net [13], Transformer based U-Net [9], U-Net++ [16], LeViT-U-Net++ [8], U-Net with ResNet encoder [10] are some of the proposed models which can provide great segmentation result. However, there are some limitations in these models- (i) U-Net [10] performs great in medical image segmentation, but at the cost of large filter size. It requires a large number of trainable parameters. (ii) Attention U-Net [13] also performs well but it inherits quadratic computational complexity. (iii) Transformer does patchification of images before segmentation. It causes degradation in resolution of segmented images . (iv) U-Net++ [16] has too many layers and skip connections. Its bulk architecture requires expensive hardware to be embedded in. In this paper, we have proposed a Gram matrix-oriented bypass attention mechanism in the U-shaped architecture. It uses less multiplication operation and more concatenation operation. Therefore, the computational complexity becomes almost linear. Because of the implementation of Gram matrix [5], a more developed feature matrix is produced in each step, which makes it possible to reduce the number of layers and the size of filters. Therefore, the trainable parameters are drastically reduced and computation is simplified. For training and validation, we have used MRI scanned GI tract image data set provided by UW- Madison [12]. It is a three-level semantic segmentation problem where stomach, large bowel and small bowel from the MRI image. In the result analysis part of this paper, we have compared the performance of our method with U-Net [10], U-Net++ [16] and ResNet Encoder based U-Net [10], LeViT-U-Net++ [1]. The comparison reveals that the proposed method gives a better segmentation performance compared to U-Net++ [16] and LeViT U-Net++ [8] while requiring significantly reduced number of trainable parameters.

979-8-3503-2614-7/23/$31.00 ©2023 IEEE

II. Related Works

A. Medical image Segmentation Based on CNN

Encoder - decoder based U shaped CNN architectures play a pivotal role in medical image segmentation. There are many variants of U-shaped architectures. U-Net [10] is one of them which shows remarkable segmentation performance. Among the encoders of U-Net, ResNet performed best in case of GI tract MRI image segmentation. To segment small region, weighted ResUNet [11] is used. U-Net++ [16] is an elaborate version of U-Net which is has less semantic gap. Between the feature maps of encoder and decoder. For 3-D image segmentation, 3-D U-Net [3], 3-D Multi ResUNet [15] are used. A modified version of 3-D Multi ResUNet outperformed

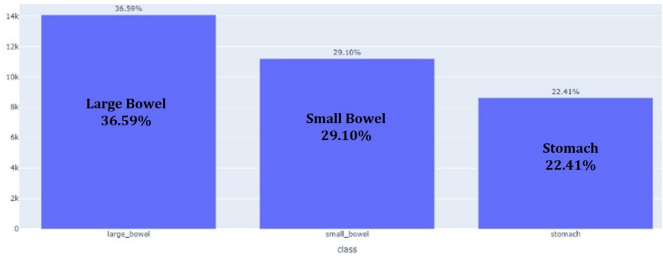

Fig. 1. Percentage distribution of training images with mask.

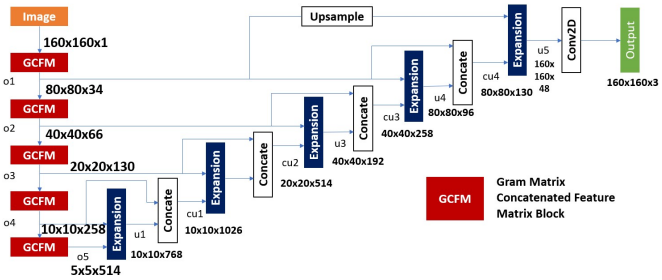

Fig. 2. Complete Block Diagram of the Architecture.

the traditional 3-D UNet and 3-D Multi ResUNet in case of lung tumor image segmentation [7].

B. Medical Image Segmentation Based on Transformers

CNN based models cannot model long range dependencies. Vision Transformer or ViT [1] is used in this task. There are many hybrid networks used in medical image segmentation. These are basically combination of transformer and U shaped architecture. TransU-Net [9] is the first hybrid structure. TransU-Net [9] is a pioneering framework for medical image segmentation that introduces self-attention mechanisms in a sequence-to-sequence prediction setup. To address the reduction in feature resolution caused by Transformers, it combines a hybrid CNN-Transformer architecture. IT-UNet [14] is used for organ at risk segmentation. For thoracic segmentation, FcTC- U-Net [2] is used. Another example of hybrid network in LeViT U-Net++ [8] used for GI tract MRI image segmentation. In this method, the usage of vision transformer is implied by combining hard inductive bias properties of CNNs. There are also some purely transformer based methods. Swin Transfomer [6] is one of them. In some case these models performs more accurate.

III. Dataset and Preprocessing

The proposed model is trained and validated with GI tract MRI scanned image data set provided by UW Madison [12]. The data set was provided for a competition in Kaggle and only training data set was available. In the train folder, there were sub folders for each patient. The scanned images of each patient were day-wise organized. Each scanned image was in PNG format. For each image 3-levels of segmentation masks were provided and the levels were stomach, large bowel, and small bowel. The masks were in RLE (Run Length Encoded) format. As each image has 3 levels (stomach, large bowel, and small bowel). A percentage distribution of images among stomach, large bowels and small bowels is shown in Figure 1. Before training the model with the dataset, the RLE masks were decoded. 70% of the training data was used for training and 30% was used for validation.

IV. Proposed Model

In Figure 2, the architecture of the proposed model is given. We have redesigned the encoder and decoder part of the U-shaped architecture. Each layer of the encoder is named as GCFM (Gram matrix concatenated feature matrix) block.

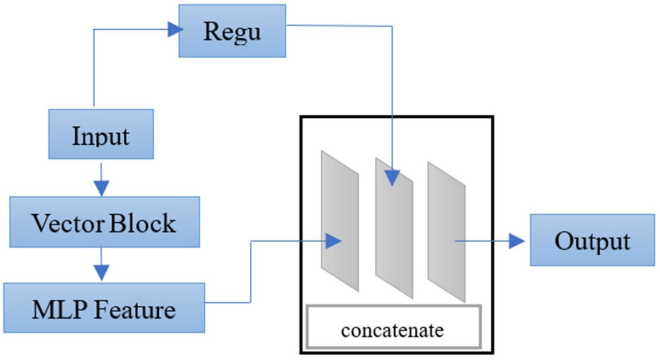

Fig. 3. Block Diagram of GCFM Block.

There is total 5 GCFM blocks. Each layer in the decoder is named as Expansion block. There are total 5 Expansion blocks.

So, what is Gram matrix and how are we leveraging it in our model?

Gram matrix is used to capture the style of an image [4]. It basically finds the correlation among the features of an image [5]. In U-Net, each convolution layer produces feature matrices which are independent of one another. To know the correlation between the components of the feature matrices, Gram matrix is used. At first, each feature matrix is converted into a 1-D vector. Then these 1-D vectors are concatenated and a 2-D matrix is generated. Gram matrix is the inner product of this 2-D matrix which represents the correlation among the components of the feature matrices.

$$GramMatrix = M \times M^t \tag{1}$$

Whereas M is generated 2-D matrix. The Gram Matrix is then concatenated in the front and back of the primary 3-D feature matrix.

A. GCFM Block

In Figure 3, GCFM block takes a 3-D tensor as an input and sends it to the vector block. The vector block converts each feature matrix into a 1-D vector. MLP feature block takes in all the 1-D vectors as inputs and produces Gram matrix. After that, the Gram Matrix is concatenated with the primary 3-D feature matrix. As Gram matrix gives an insight about which components of the feature matrix are of more importance, we can have good performance using small sized filters and less number of layers.

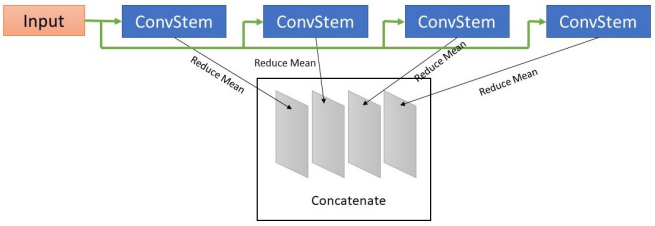

Fig. 4. Block Diagram of Vector Block.

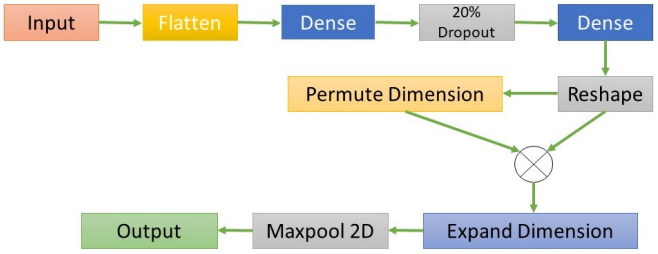

Fig. 5. Block Diagram of MLP Feature Block.

B. Vector Block

As shown in Figure 4, the vector block takes in the 3-D tensor and sends it to four parallel convstem layers. In each convstem layer, convolution, gelu activation and normalization operations take place. The output of the convstem layer is channelwise averaged. After passing through the convstem layer, a 1-D vector is generated. From 4 convstem layers, 4 1-D vectors are generated and concatenated and the final 1-D component vector is generated.

C. MLP Feature Block

In Figure 5, MLP feature block takes in the 1-D component vector as input. Then this vector is flattened. After that it is passed through two successive dense convolution layers. The inner product of the vector is determined and the Gram matrix is generated.

D. Expansion Block

According to Figure 6, the expansion block takes in two inputs. Considering the first expansion block, the first input is the output of the last GCFM block. And the 2nd input is the output of the 2nd last GCFM block. These two inputs and concatenated after passing through the convolution and transpose convolution layer. 30% dropout is used to avoid over-fitting. Then the generated output is again concatenated with the 2nd input. This is the operation of one expansion block. It repeats 5 times in our model.

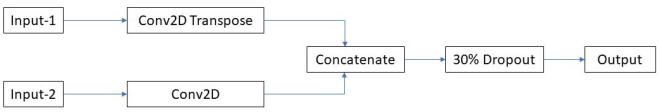

Fig. 6. Block Diagram of Expansion Block.

Loss	Train	Validation
BCE Loss	0.0975	0.1237
Tversky Loss	0.0765	0.1013

TABLE I
LOSS COMPARISON

Models	IoU	Dice Coefficient	Loss	Parameter Count
Proposed Model	0.912	0.9008	0.1073	9,314,745
U-Net	0.9107	0.9032	0.1032	36,893,154
U-Net++	0.8765	0.9012	0.1062	10,365,568
LeViT-U-Net++	0.7282	0.7954	0.1347	
ResNet Encoded UNet	0.884			

TABLE II
VALIDATION ACCURACY OF DIFFERENT MODELS

V. RESULT AND ANALYSIS

We have used 55 epochs to train the model and for the best use of GPU memory. 80% of the data were used to train the model and the rest of 20% were used to evaluate the model. For the result analysis, the performance parameter we used are:

$$DiceCoefficient = \frac{2 \times (|A \cap B|)}{|A| + |B|} \tag{2}$$

$$IoU(A, B) = \frac{|A \cap B|}{|A| + |B|} \tag{3}$$

$$DiceLoss = 1 - DiceCoefficient \tag{4}$$

$$BCELoss = BCE + DiceLoss \tag{5}$$

$$TverskyLoss_{\alpha,\beta} = \frac{TP}{TP + \alpha FP + \beta FN} \tag{6}$$

Where, BCE means Binay Cross Entropy and TP, FP and FN mean true positive, false positive and false negative respectively. In equation 6, α and β are constants. The value of each of these constants is 0.5.

Figure 7 shows the plot of BCE loss and Tversky loss with respect to number of epochs. It is observed that within 40 epochs both the losses decrease significantly.

Table I gives the minimum losses achieved for training and validation data. It is seen that Tversky loss is lower than the BCE loss.

Likewise, in Figure 8 and Figure 9, the maximum values of Dice coefficient and IoU are obtained within a mere 40 epochs. Table II provides the comparison of the performance of the proposed model with that of generic U-Net, generic U-Net++, LeViT-U-Net++ [8] and ResNet encoded U-Net [10]. It should be mentioned that for LeViT-U-Net++ [8], maximum score values are shown among those obtained for four different folds. It can be seen that the IoU of proposed model is 0.14% greater than U-Net, 4.06% greater than U-Net++, 25.24% greater than LeViT-U-Net++ [5] and 2.8% greater than ResNet encoded U-Net [6]. The dice score of the proposed method is almost same as that of U-Net and U-Net++ but 13.15% greater than LeViT-U-Net++ [5]. The validation loss is also very close to that of U-Net and U-Net++ and 20.34% less than that of LeViT-U-Net++. If we look at the parameter count, a noticeable improvement can be seen. Our model can perform almost like

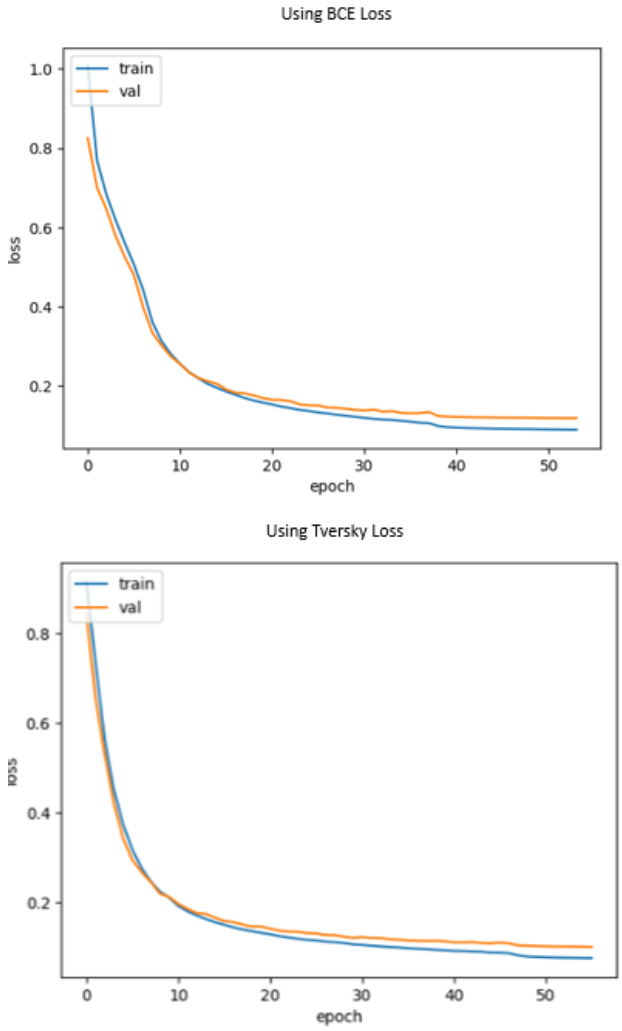

Fig. 7. Comparison between BCE loss and Tversky loss.

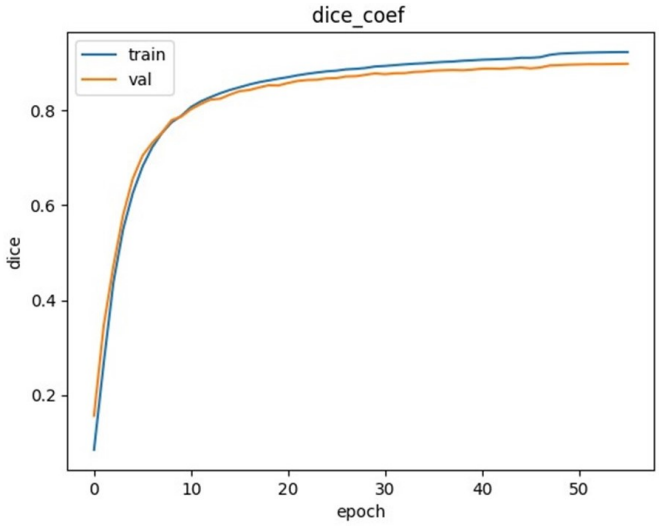

Fig. 8. Plot of Dice Coefficient.

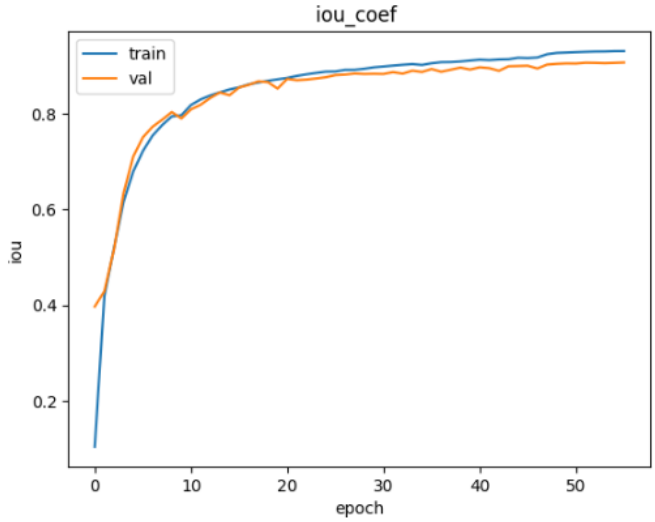

Fig. 9. Plot of IoU Coefficient.

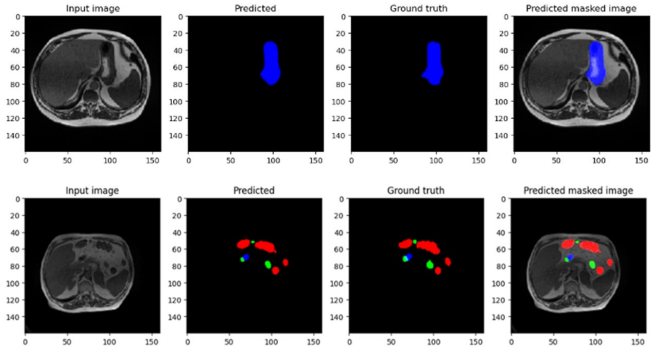

Fig. 10. Visual representation of predicted mask and comparison with ground truth (red denotes large bowels, green denotes small bowels and blue denotes stomach).

U-Net but having 75% less trainable parameters than U-Net and 10% less trainable parameters than U-Net++. It indicates that our model is much shallower than the other two without doing any sacrifice in case of segmentation performance.

Figure 10 shows the predicted masks for two randomly selected GI tract images. The masks are very similar when compared to the ground truth masks. The closeness between predicted masks and ground truth masks indicates a high level of accuracy in the model's segmentation capabilities.

Figure 11 shows the comparison of predicted masks by our proposed model and generic U-Net. The first image represents the predicted mask by the proposed model. The second image represents the predicted mask by generic U-Net. It can be seen that, in some portion, U-Net predicted better than our model, while in some portions, our model predicted better than U-Net. Therefore, it can be said that the overall segmentation performance by the proposed method is equivalent to that of

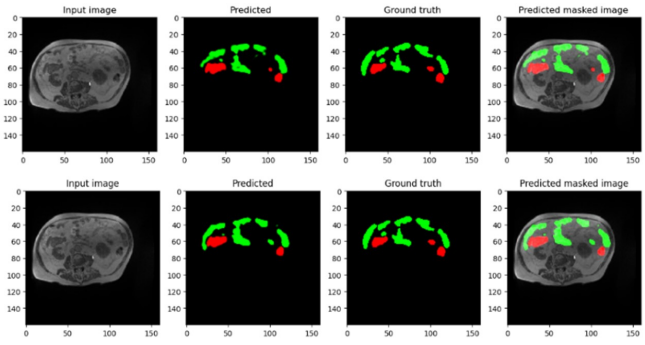

Fig. 11. Segmentation comparison between proposed model and generic U-Net (Red denotes large bowel, green denotes small bowel and blue denotes stomach).

U-Net and other existing models.

VI. Conclusion

Segmentation of MRI images of GI tract is an important issue regarding the treatment and prognosis GI tract cancer patients. This paper represents modified U-Net architecture incorporating the notion of Gram matrix. It has been shown that, using a publicly available data set, the proposed architecture can deliver a superior segmentation of stomach, large bowel and small bowel as compared to existing models. Moreover, the proposed model is rather shallow considering significantly reduced number of trainable parameters. This shallow deep neural network has a potential to be embedded in less expensive medical devices and treatment of GI tract cancer can be more feasible.

References

[1] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z Pan. Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12094–12103, 2022.

[2] Yufang He, Fan Song, Wangjiang Wu, Suqing Tian, Tianyi Zhang, Shuming Zhang, Peng Zhang, Chenbin Ma, Youdan Feng, Ruijie Yang, et al. Multitrans: Multi-scale feature fusion transformer with transfer learning strategy for multiple organs segmentation of head and neck ct images. *Medicine in Novel Technology and Devices*, 18:100235, 2023.

[3] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[4] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019.

[5] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[7] Suhail Najeeb and Mohammed Imamul Hassan Bhuiyan. Spatial feature fusion in 3d convolutional autoencoders for lung tumor segmentation from 3d ct images. *Biomedical Signal Processing and Control*, 78:103996, 2022.

[8] Praneeth Nemani and Satyanarayana Vollala. Medical image segmentation using levit-unet++: A case study on gi tract data. In *2022 26th International Computer Science and Engineering Conference (ICSEC)*, pages 7–13. IEEE, 2022.

[9] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer, 2021.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[11] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 2023.

[12] UW-Madison. [uw-madison gi tract image segmentation]. Retrieved from Kaggle Competition, July 2022.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Junyan Wu, Eric Z Chen, Ruichen Rong, Xiaoxiao Li, Dong Xu, and Hongda Jiang. Skin lesion segmentation with c-unet. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2785–2788. IEEE, 2019.

[15] Y Zhang, H Liu, and Q TransFuse Hu. Fusing transformers and cnns for medical image segmentation. arxiv 2021. *arXiv preprint arXiv:2102.08005*.

[16] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.