A HIERARCHICAL TRANSFORMER ENCODER TO IMPROVE ENTIRE NEOPLASM SEGMENTATION ON WHOLE SLIDE IMAGES OF HEPATOCELLULAR CARCINOMA

Zhuxian Guo[⋆†] *Qitong Wang*[⋆] *Henning Müller*[†]
Themis Palpanas[⋆‡] *Nicolas Loménie*[⋆] *Camille Kurtz*[⋆]

[⋆] Laboratory of Informatics Paris Descartes (LIPADE), Université Paris Cité, Paris, France
[†] University of Applied Sciences of Western Switzerland (HES-SO Valais), Sierre, Switzerland
[‡] French University Institute (IUF), Paris, France

ABSTRACT

In digital histopathology, entire neoplasm segmentation on Whole Slide Image (WSI) of Hepatocellular Carcinoma (HCC) plays an important role, especially as a preprocessing filter to automatically exclude healthy tissue, in histological molecular correlations mining and other downstream histopathological tasks. The segmentation task remains challenging due to HCC's inherent high-heterogeneity and the lack of dependency learning in large field of view. In this article, we propose a novel deep learning architecture with a hierarchical Transformer encoder, HiTrans, to learn the global dependencies within expanded 4096×4096 WSI patches. HiTrans is designed to encode and decode the patches with larger reception fields and the learned global dependencies, compared to the state-of-the-art Fully Convolutional Neural networks (FCNN). Empirical evaluations verified that HiTrans leads to better segmentation performance by taking into account regional and global dependency information.

***Index Terms*—** Digital histopathology, HCC, Neoplasm segmentation, Transformer architecture, Semantic segmentation, Deep learning.

1. INTRODUCTION

Hepatocellular Carcinoma (HCC) is a primary tumor of the liver and is now the fifth most common cancer worldwide [1]. HCC is a highly heterogeneous molecularly and histologically as a cancer. A series of ongoing studies have shown that the HCC phenotype appears to be closely related to particular gene mutations [2].

Clustering the WSI representations of certain phenotypes to particular gene mutations by mining the relationships of the representations and their corresponding transcriptomic data is clinically meaningful. Imaging-based multi-omics can help physicians understand the morphology and micro-environmental cell population changes related to certain mutations. Multiple Instance Learning (MIL) such as the framework Clustering-constrained Attention Multiple Instance Learning (CLAM) [3] has been applied in [4] to predict the activation of 6 common HCC immune gene signatures within a roughly selected neoplasm area. A robust automatic neoplasm segmentation can then act as a preprocessing filter, not only to produce the annotation in lieu of pathologists but also to exclude the potential annotation bias to specific highly predictive regions when using the annotations provided by experienced pathologists.

Automatic neoplasm segmentation remains an important challenge today mainly due to: (1) the inherently high tissue heterogeneity, and (2) the lack of consideration of effectively aggregated large-field relational features. For instance, HCC is a highly heterogeneous cancer and it has distinct morphological phenotypes. The morphological appearance of different phenotypes is very different across cases, which makes the tumor segmentation model hard to generalize.

The *de facto* WSI segmentation models are patch-based convolutional neural networks (CNN) like [5, 6], and the patch size usually ranges from 128×128 to 512×512 due to GPU memory limitations. Such a patch size is very small compared to the gigapixel size of the WSI, which limits the receptive field of the model and leads to a limited ability to capture large-scale aberrant tissue structures in HCC such as macro-trabeculae[1], pseudoglandular and necrotic foci architectural patterns, etc.

Previous work [7] has studied how the proposed multi-scale CNN models take into account the histological features at different scales, ranging from nuclear aberrations through cellular structures to the global tissue architecture, by using patches at different spatial scales as input. For the multi-scale CNN model, larger-scale patches, which are centered aligned to the smallest scale patch at full resolution (detail patch), are downsampled to similar size to the detail patch in order to fit the segmentation framework. A general image downsampler is usually applied while the WSI complexity is quite different from other types of images. In this context, we aim to develop an entire HCC neoplasm segmentation framework

[1] Neoplastic cells of macrotrabecular-massive HCC are arranged in thick trabeculae surrounded by vascular spaces.

by using state-of-the-art approaches to mimic the WSI exploration by pathologists in a hierarchical fashion and thus to serve for downstream tasks in digital histopathology.

Transformers [8] rely on global self-attention mechanisms and have achieved excellent performance in many tasks with global dependency requirements such as sequence modeling and language modeling. They were modified for computer vision tasks, called Vision Transformers (ViT) to serve as an alternative to CNNs in feature extraction [9]. In digital pathology, as a follow up of CLAM, ViT was shown to be stronger in feature aggregation than simpler attention-weighted average mechanisms. They can also be stacked as a hierarchical architecture to effectively aggregate the WSI features to a slide-level representation [10].

Inspired by the success of the application of ViT on representation learning of gigapixel images, we propose in this article a framework, called HiTrans, with hierarchical-based Transformer encoder to enlarge the field of view and to enhance the entire HCC neoplasm segmentation. Such a contribution allows to dramatically increase the segmentation field of view to 4096×4096. The WSI patches are encoded with larger fields of view compared to conventional Fully Convolutional Neural networks (FCN), and are decoded by taking into account regional and global dependency information. The experimental results with a large real dataset demonstrate that the proposed HiTrans framework can lead to better entire HCC neoplasm segmentation, quantitatively and qualitatively.

The dataset used in this study is introduced in Sec. 2. Sec. 3 presents the data preprocessing pipeline, the baseline architecture, and the proposed network training protocol. Experimental results are provided in Sec. 4.

2. DATASET

The PAIP liver cancer segmentation challenge was held in 2019 (PAIP 2019) [11] as part of the MICCAI 2019 Grand Challenge for Pathology. The PAIP 2019 training cohort consists of 50 anonymized WSIs at the 20× magnification in ScanScop Virtual Slide (SVS) format. Each WSI was selected from the HCC resection slides from one patient, which means the 50 WSIs in the training cohort belong to 50 different individuals. The Edmonson-Steiner tumor grade distribution is 7, 23, 20 for Grade I, II, III, respectively. The slides were all stained with conventional hematoxylin and eosin (H&E) staining and were digitized with an Aperio AT2 whole-slide scanner. The WSI size ranges from 35855×39407 to 64768×47009. The training cohort WSIs come with two-layers of annotation for whole tumor areas and viable tumor areas. Only the first annotation layer (i.e., the whole tumor area) was used in this study. The whole tumor area means that the entire neoplasm that can be observed on the WSI, including all dispersed viable tumor cell nests, tumor necrosis and tumor capsules.

3. METHODS

The proposed HiTrans framework (Fig. 1) takes 4096×4096 WSI patches as input. A hierarchical Transformer encoder add-on module is added between a ResNet [12] encoder backbone and a modified U-Net decoder to learn the global dependencies (red dashed box). Sec. 3.1 introduces the data preprocessing pipeline. The proposed architecture details are illustrated in Sec. 3.2 and the training protocol is described in Sec. 3.3.

3.1. Data preprocessing

Since the WSI tissue mask is not provided, we followed a conventional pipeline to patchify WSIs to create the pairs of high tissue percentage 4096×4096 patches and their corresponding neoplasm masks. The 50 WSIs were split into 30, 10, and 10 for training, validation, and test, respectively.

3.2. Proposed architecture

A hierarchical add-on Transformer encoder module that contains two Transformers is added between the CNN feature extractor and the decoder to learn subtle global dependencies, as shown in Fig. 1 the red dashed box.

[CNN feature extractor] The intermediate layers of a pretrained 18-layer ResNet were used as an encoder for feature extraction. The aforementioned ResNet was pretrained on 57 histopathology datasets in a self-supervised learning fashion [13] following the SimCLR [14] contrastive learning setting. The adaptive average pooling layer and the dense layer at the end of the ResNet were removed to make it as a 2D feature map extractor. Five feature maps are generated (Fig. 1, map1 to map5).

[Transformer encoder I] Transformer encoder I is a 12-layers standard Transformer encoder with six attention heads and 384-length hidden dimension. Map5 is unfolded into 64 (8×8) seamless 16×16 sub-patches, and each sub-patch contains 256 mini patch embeddings. The sub-patches are linearly transformed (Fig. 1, linear1) to fit the hidden dimension of Transformer encoder I. An extra segmentation embedding SEG is added. All output SEG embeddings of each sub-patch that contain the regional features of each sub-patch are kept to form the inputs of Transformer encoder II. All output mini patch embeddings are reshaped to a 128×128 feature map to act as a skip connection to provide finer regional features for decoding.

[Transformer encoder II] Transformer encoder II has 12 layers with three attention heads, and the hidden dimension is 192. It takes the output SEG embeddings from Transformer encoder I to learn the global dependencies among the sub-patch features, within a 4096×4096 patch. The output embeddings of the sub-patches are reshaped to a 8×8 feature map. Thanks to the self attention mechanism, each element

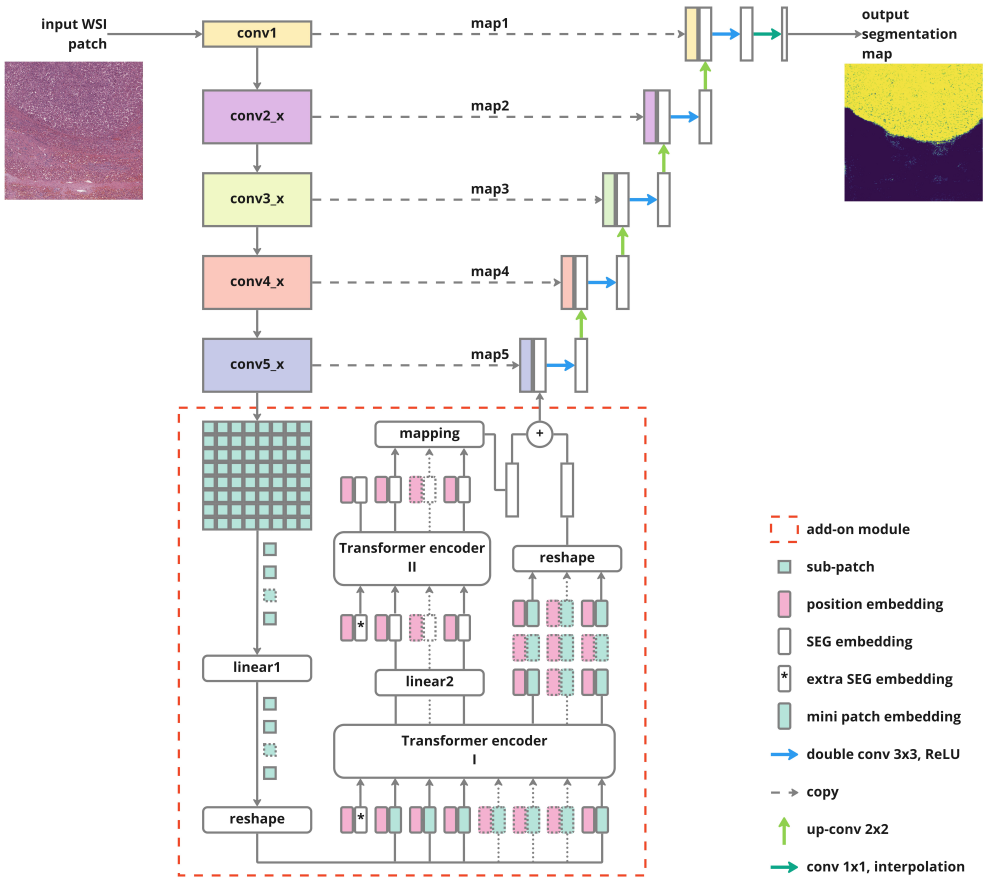


Fig. 1. Proposed hierarchy-based Transformer encoder architecture (HiTrans) for entire HCC neoplasm segmentation.

in the feature map contains its global dependency information. Each element is then mapped and added up to its spatial corresponding elements on the 128×128 feature map from Transformer encoder I. We added up the feature maps from the two Transformer encoders instead of doing concatenation in order to alleviate the block-biased prediction.

[Global dependency learning] The feature maps from the two Transformer encoders are fused to merge regional and global features. The new feature map is then concatenated with map5 to maintain localization accuracy. Performing a global dependency learning through this hierarchical Transformer encoder architecture, the WSI patches are encoded with larger fields of view compared to CNN. The proposed architecture also provides the decoder with regional and global dependency information for a finer segmentation.

[Convolutional decoder with shortcuts] Transposed convolutional layers (Fig. 1, up-conv 2×2) expand the feature map size. Map1 to map4 are concatenated with the expanded feature maps of the $l-1$ layers, and then pass the double convolutional layers that halve the number of channels. At the end, the 2048×2048 feature maps will be decoded to a 4096×4096 segmentation map by a 1×1 convolutional layer following a bilinear interpolation with corner pixels alignment.

3.3. Network training

[General training protocol] The model was trained on a Nvidia A100 SXM4 80GB GPU for 100 epochs using the AdamW optimizer [15] with a batch size of 2 and a early stopping patience of 10. Base learning rate was set to 5e-4, with the first 10 epochs used to warm up followed by decay using a cosine schedule to reduce the base learning rate to the minimum learning rate 1e-6. Weight decay rate was increased gradually from 1e-2 to 1e-4 following a cosine schedule.
[Alternate training] Alternate training strategy was adopted to overcome the convergence difficulty in training this hierarchical semantic segmentation architecture with two stacked Transformer encoders. The ResNet feature extractor, Transformer encoder I and II were trained Alternately to maximize the framework ability.

4. RESULTS

4.1. Evaluation metric and results

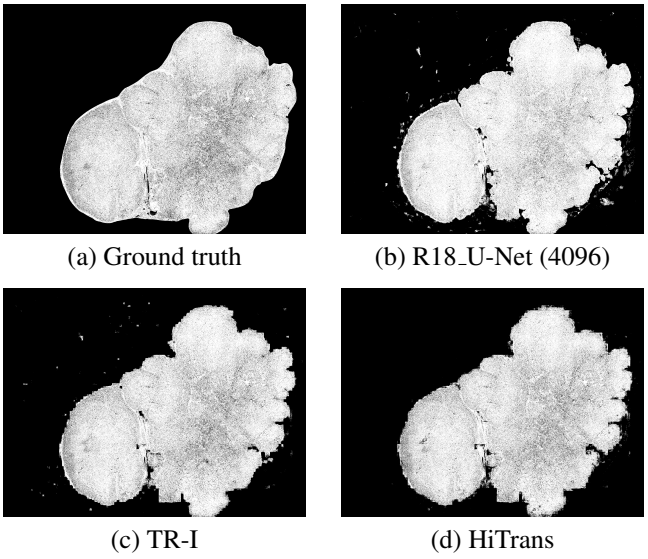
The WSI segmentation for all models was performed in seamless patch-wised inference units. The average Jaccard index of the 10 WSIs in the test set is used as a quantitative score to

Exp.	Method	Patch size	Avg. Jaccard
1	U-Net	512	0.6659
2	R18_DLv3	512	0.6272
3	R50_DLv3	512	0.6803
4	R18_PSPNet	512	0.6252
5	R50_PSPNet	512	0.6695
6	SegFormer-B0	512	0.6352
7	SegFormer-B2	512	0.6027
8	HiTrans	4096	**0.7513**

Table 1. Experimental results and transversal comparison.

Exp.	Method	Patch size	Avg. Jaccard
1	R18_U-Net	512	0.6609
2	R18_U-Net	4096	0.7202
3	TR-I	4096	0.7172
4	HiTrans	4096	**0.7513**

Table 2. Quantitative results of the ablation study.



(a) Ground truth (b) R18_U-Net (4096)

(c) TR-I (d) HiTrans

Fig. 2. Qualitative segmentation results of the ablation study.

evaluate the entire neoplasm segmentation performance. Notably, the results we presented cannot be directly compared with PAIP 2019 leaderboard results, because we focused on evaluating the model performance and avoided the usage of postprecessing steps, including manually-designed WSI postprocesing strategies, ensemble learning, and overlapped inference. Moreover, the results in this study is only trained and tested on the training cohort of PAIP 2019 and only the whole tumor area annotation is used for training.

Compared with three state-of-the-art (SOTA) semantic segmentation frameworks, U-Net [16], DeepLabV3 [17], PSPNet [18], and one SOTA Transformer-based framework, SegFormer [19], the proposed HiTrans framework has better performance on HCC segmentation task (Tab. 1, Exp. 8). Among the FCNN segmentation frameworks, DeepLabV3 with a ResNet-50 backbone has the best performance (Tab. 1, Exp. 3) thanks to the stronger feature extractor and the Atrous Spatial Pyramid Pooling (ASPP) modules, which probe convolutional features at multiple scales while avoiding increasing network size too much. For SegFormer, the smaller variant SegFormer-B0 (Tab. 1, Exp. 6) has better performance than the larger variant SegFormer-B2 (Tab. 1, Exp. 7), possibly due to the convergence difficulty in training larger Transformer-based model.

4.2. Ablation study

The proposed hierarchical Tansformer encoder framework can learn global dependencies and bring muti-scale cues for the decoder during segmentation inference. In the ablation study, we removed the Transformer hierarchical add-on module, in order to evaluate the above argument, namely, that knowledge of global dependencies can enhance segmentation performance. Experiments without the add-on module using 512×512 (Tab. 2, Exp. 1) and 4096×4096 (Tab. 2, Exp. 2) patches were conducted separately. Besides, an experiment using an architecture without Transformer encoder II was conducted (Tab. 2, Exp. 3). By comparing the experimental results, taking larger patches as input and using HiTrans to learn the global dependencies can lead to better segmentation results. Comparing with the results from add-on module dropped (Fig. 2, b) and Transformer encoder II dropped (Fig. 2, c) architecture, HiTrans can further improve the precision (Fig. 2, d) thanks to this regional and global dependency-aware architecture.

5. CONCLUSIONS

In this article, we introduce a hierarchical Transformer-based segmentation architecture, HiTrans, for HCC entire neoplasm segmentation. HiTrans can efficiently learn the regional and global dependencies within 4096×4096 WSI patches by encoding and decoding the WSI in a hierarchical fashion. The experimental results with a large real dataset demonstrate that HiTrans can lead to quantitatively and qualitatively better entire HCC neoplasm segmentation. In our future studies, we aim at developing a robust slide-wise context aware framework by leveraging different strategies in global dependency learning like graph-based neural networks. We will also explore the application on other tasks.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by PAIP 2019 Challenge. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

This work was supported by Data Intelligence Institute of Paris (diiP), IdEx Université Paris Cité (ANR-18-IDEX-0001), and Translational Research Program in Cancerology INCa-DGOS - PRTK-2020, and was performed using HPC resources from GENCI-IDRIS (2022-AD011012825R1) made by GENCI. Qitong Wang is funded by China Scholarship Council.

8. REFERENCES

[1] Jacques Ferlay et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–386, 2015.

[2] Julien Calderaro et al., "Molecular and histological correlations in liver cancer," *Journal of Hepatology*, vol. 71, no. 3, pp. 616–630, 2019.

[3] Ming Y. Lu et al., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[4] Qinghe Zeng et al., "Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology," *Journal of Hepatology*, vol. 77, no. 1, pp. 116–127, 2022.

[5] Blanca Maria Priego Torres et al., "Automatic segmentation of whole-slide h&e stained breast histopathology images using a deep convolutional neural network architecture," *Expert Syst. Appl.*, vol. 151, pp. 113387, 2020.

[6] Mousumi Roy et al., "Convolutional autoencoder based model HistoCAE for segmentation of viable tumor regions in liver whole-slide images," *Scientific Reports*, vol. 11, no. 1, pp. 139, 2021.

[7] Rüdiger Schmitz et al., "Multi-scale fully convolutional neural networks for histopathology image segmentation: From nuclear aberrations to the global tissue architecture," *Medical Image Anal.*, vol. 70, pp. 101996, 2021.

[8] Ashish Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[9] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[10] Richard J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *CVPR*, 2022, pp. 16123–16134.

[11] Yoo Jung Kim et al., "PAIP 2019: Liver cancer segmentation challenge," *Medical Image Anal.*, vol. 67, pp. 101854, 2021.

[12] Kaiming He et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[13] Ozan Ciga et al., "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, pp. 100198, 2022.

[14] Ting Chen et al., "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[15] Ilya Loshchilov et al., "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.

[16] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[17] Liang-Chieh Chen et al., "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[18] Hengshuang Zhao et al., "Pyramid scene parsing network," in *CVPR*, 2017, pp. 6230–6239.

[19] Enze Xie et al., "Segformer: Simple and efficient design for semantic segmentation with transformers," in *NeurIPS*, 2021, pp. 12077–12090.