International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Segmentation Using DeepLab

Rushikesh Paresh Shetty*, Dr. Ruhin Kouser**

*(Department Computer Science and Engineering, Presidency University, Bengaluru, India
Email: rushikeshetty@gmail.com)

** (Department Computer Science and Engineering, Presidency University, Bengaluru, India
Email : ruhinkouser@presidencyuniversity.in)

----------**********----------

Abstract:

In recent semantic segmentation state of affairs, DeepLab consistently outshines one of the best architectures with strong feature extraction and multi-scale contextual processing capabilities. Despite that, in DeepLab, a traditional convolutional decoder lags in long-range dependencies and contextual relations, significant in fine-grained segmentation. In this work, an attention-based DeepLab network with a transformation- based decoder in lieu of traditional DeepLab's default decoder is proposed for contextual-aware feature aggregation for semantic segmentation improvement. In our model's decoder, self- attention learns accurately global and spatial dependencies, and hence, induces significant improvements in segmentation accuracy, specifically concerning complex boundary regions and complex regions. In a variety of challenging benchmark datasets, our attention-based DeepLab is analyzed and proves accuracy and segmentation performance over traditional decoder implementations. In this work, it is demonstrated that a transformationer-based decoder can have a role in CNN frameworks, and future studies for mixed model investigations in semantic segmentation are proposed.

Keywords — **Semantic Segmentation, DeepLab, Transformer Decoder, Attention Mechanisms, Context-Aware Features, Convolutional Neural Networks (CNN), Self-Attention, Long-Range**

----------**********----------

I. INTRODUCTION

Semantic segmentation is one of computer vision's most significant jobs, mapping each pixel in an image to a predefined group of categories. Examples include autonomous driving, medical imaging, and urban analysis of a scene. In the last several years, the family of architectures, i.e., DeepLab, has become a most desired semantic segmentation model for its ability to handle multi-scale contextual information and high spatial detail through techniques such as ASPP and decoders with a convolutional shape [5]. With such networks' record success, traditional decoders in CNN shape in DeepLab architectures have a disadvantage in terms of handling long- range dependencies and overall context, becoming critical factors the success of such high-density segmentation operations. In that direction, state-of-the-art work these days introduces transformers as a mechanism for improving segmentations through encoding long-range dependencies via their self-attention mechanism and feature improvement [20].

Most of them have even utilized additional transformers with success in an effort to generate a more successful segmentation and make such an activity even more efficient. For instance, Max-Deeplab developed a new model for panoptic segmentation on transformers, having outpaced traditional DeepLab models through substituting CNN decoders with the use of attention mechanism-based transformers [27].The segmenter then refines the patch embeddings into segmentation maps with a transformer decoder's assistance. Since it utilizes the self- attention mechanism of transformers in order to extract fine- grain information, it has attained excellent performance for complex datasets including ADE20K [16]. The Fully Transformer

Networks go one step ahead, utilizing both encoder and decoder sections with use of transformers in order to model contextual relations at a global level for high performance in high-density images [4]. Yet another dominant architecture is Swin Transformer, a model that relies on a hierarchical mechanism of attention with windows shifts. That not only introduces computational efficiency but additional accuracy in dense prediction, fitting perfectly in requirements of DeepLab- similar architectures [2]. We build over such advancement with a focus in our work towards an improvement in the DeepLab model with a transformer-based decoder. In contrast to prior work with use of only CNN decoders, in this work, we go over to revisiting the default decoder in DeepLab and introduce a transformer-based one in addition to enhancing refinement of pixel-wise classification with both contextual information at a global level and with an eye for attention towards relations in spaces. That is drawn in motivation towards such a hybrid model with use of trans- formers and CNNs, with trans- formers encoders with use of CNN decoders proving to deliver best for medical image segmentation, and therefore motivated such a hybrid for high- resolution prediction [24]. In addition, attention-augmented convolutional networks have shown that the introduction of modules of attention in seg- mentation architectures strengthen the role placed on relevant information, an impact extended in our case to our transformer-based model, DeepLab [25].

Recent transformation-based algorithms, such as Pyramid Vision Transformer (PVT) [26] and HRFormer [23]. We then extend and illustrate that transformers can effectively extract spatial detail for dense prediction. Our work, motivated by such algorithms, utilizes a transformer de-coder for encoding high-resolution features and global dependencies in the model, DeepLab. We, therefore, adopt the model, DeepLab, to use an attention mechanism, and with it, accuracy in segmentation and boundary accuracy, most particularly in regions with complex, fine detail, is boosted.

In conclusion, in this work, an attention-guided model, DeepLab, is proposed, substituting a CNN-based decoder with a transformer-based decoder. With such integration, our model utilizes full use of the capability of the transformer for even enhanced contextual long-range dependencies and rich spatial relations. We have performed experiments with benchmark datasets, and such integration is best observable in noticeably displayed increased accuracy and segmentation quality over traditional DeepLab models.

II. RELATED WORK

Recent development in semantic segmentation utilized deep learning with an additional mechanism of attention, in the form of a specific use of transformers, that effectively model long- range dependences of an image. Traditional CNN-based architectures, such as DeepLabV3+, played an important role in pixel-wise segmentation. They utilized ASPP and con- volutational decoders to comprehend contextual information at var- ious scales [2]. Despite the successful capture of local features, its lack of capacity for representing global dependences motivated re- researchers to explore transformer- based improvements. Hybrid Architectures with Transformer Decoders Recent use of a transformer-based decoder with CNNs also showed effective performance in segmentation. For instance, Max- Deeplab replaces the decoder part of CNNs with a mask transfor-mer and reaches state-of-the-art panoptic segmentation with effective capture of both global and local information [6]. Likewise, Segmenter transforms patch embeds to segmentation maps through a transfor- mer decoder and reaches a finer segmentation in datasets such as ADE20K through use of self-attention [20]. Fully Transformer Networks (FTN), utilizing transfor- mers for encoding and decoding, present the potential of hybrid models through high performance in dense, high detail segmentation [27]. All such examples present use of transformer decoders in contextual dependences, and thus, a transfor- mer decoder is utilized in our DeepLab-based model. Transformer-Based Backbones for DeepLab Enhancement Transformers have not only been utilized for decoders but even for backbone hybrids,

replacing rigid architectures of CNNs with flexible and multi-scale ones. Swin Transformer utilizes hierarchical architec-ture with shifted window attention whose extraordinary balance yields even better performance in high resolution images, comprising superior on computational efficiency, one of the most useful beneficial factors for the hybrid architectures of DeepLab [16]. Pyramid Vision Transformer (PVT) [24] and HRFormer [25] recently took multi-scale and high-resolution, respectively, in such a direction. These work effectively as standards of segmentation accuracy in the hybrid architectures of DeepLab. The examples mentioned confirm that transformers can work effectively as flexible backbones and allow for such capabilities to be incorporated in a DeepLab model, and in the process, most probably gain performance. Medical Image Segmentation with Transformer-CNN Model Hybrid models have an application in medical image segmentation, such as feature capture both locally and at a larger level for complex processes implements a transformer encoder with a CNN decoder and establishes strong strengths of the hybrid model for high accuracy in medical image segmentation for medical imaging processes [4]. Medical Transformer [14] extends performance with use of axial attention and reduced computational burden through focused computation in regions of medical images of interest. These models present how hybrid approaches work exceedingly well and motivate us to maintain use of a CNN encoder in DeepLab but utilize a transformer-based decoder for improvement in segmentation accuracy. Attention Mech- anisms in Convolutional and Transformer Networks Attention in CNNs have experienced tremendous success in alternate segmentation models through attracting the network's attention towards most important segments of any image. For example, Attention- Augmented Convolutional Networks combine an at- tention mechanism in con-vulutional blocks, improving overall picture segmentation performance [2]. These observations about attention-guided awareness in spaces apply directly to transformer decoders, in which self-attention can extend over localized fea- tures and span larger contextual dependen- ties. Extra Progress in Transformer-Based Dense Predic- tions Besides segmentation, DETR [3] reveals that transfor- mers make it feasible for end-to-end dense prediction and DPTs [19] model dependen- cies in spaces for dense segmentation processes. UNetFormer [13] and CoTr [7] bridge gaps between capabilities of CNN and transfor- mer in segmentation through maintaining localized awareness of CNNs but utilizing transformers for contextual awareness. These again validate use of a hybrid model for high accuracy in segmentation. Panoptic and High-Resolution Applications Axial-DeepLab incorporates axial attention in a DeepLab model to perform panoptic segmentation of high- resolution operations [17]. Fol- lowing its path, Transformers have been adopted in use in Transformers [21] and Swin-DeepLab [15] for fine feature extraction in DeepLab and, in consequence, for enhancing panoptic performance. High resolution satellite and aerial datasets in high resolution also exhibit that feature extraction in high resolution can be performed with transformers [18].

III. RESEARCH GAP

Strikingly, despite the very encouraging results of hybrid architectures for semantic segmentation, few explicit works have explored fusing transformer decoders with CNN-based ones, such as DeepLab, in an attempt at performance improvement. Most proposed transformationer-based techniques in semantic segmentation have utilized a purely transformationer-based architec- ture, or at best, have replaced only the encoder but not utilized complementary strengths imparted through transformationer decoders. For example, techniques such as Max-Deeplab [12] and Segmenter [1] have stressed transformation decoders' potential, but none have actually merged them with CNN-based backbone networks, not even in a DeepLab scenario, in which both have proven to excel in describing a local feature. Besides, most state-of-the- art transformationers, such as Fully Transformer Networks (FTN) [10], Swin Transformationer [11], Pyramid Vision Transformationer (PVT) [8], etc., have been designed for dense prediction, or panoptic

segmentation, but none for synergies between transformation decoders and CNN encoders in enhancing refinement of segmentation maps. As much as such techniques have stressed encoding capability of an encoder in representing a global context, most of them have not optimized use of transformation decoders in enhancing refinement at a pixel level, specifically for complex and fine-grain structures. Hybrid techniques such as TransUNet [9] and Medical Transformer [22], fusing CNNs and transformations for medical image segmentation, have displayed fascinating potentials but not in terms of representing contextual relation between encoder and decoder traditional DeepLab model.

The model proposed in this work, therefore, seeks a balance between strengths in terms of fine-grain feature extraction through a CNN and awareness of a global context through transformation decoders. By merging them, we overcome the weaknesses of traditional DeepLab decoders in not being effective in encoding long-term dependencies and not efficient in representing complex boundary structures of an object. There is a lack of such studies, and therefore, a necessity for offering additional consideration to such hybrid techniques in holding both strengths of both CNNs and transformers through leveraging both architectures' complementary capabilities for semantic segmentation accuracy improvement.

capture features at a range of scales. 2. Multi-Scale Context Capture with Atrous Convolutions: In an attempt to capture multi-scale features, we use Atrous Spatial Pyramid Pooling (ASPP), and apply atrous convolutions at several scales in order to generate multi-scale feature maps:

FASPP = r1,6,12,18

AtrousConv(Fenc, r)

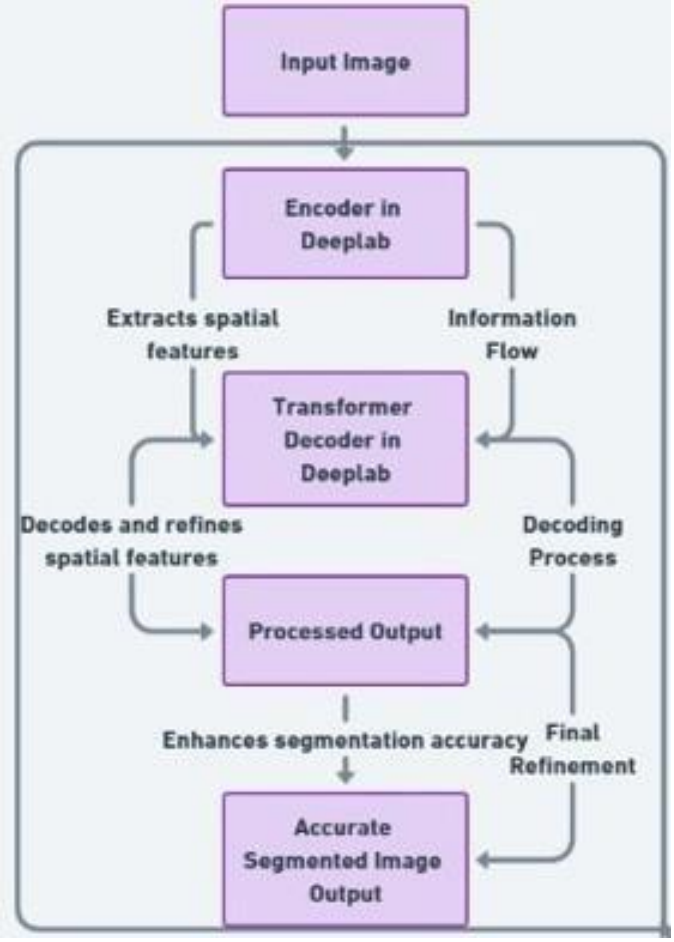

FIG 1.

where each rate r is a level of spatial context. It allows both fine and coarse detail in an image to be captured in a model.

C. Transformer-Based Decoder Instead of a traditional decoder, a transformer-based decoder is utilized in an attempt to utilize additional refinement of feature representations with an attention mechanism. Modeling long-range dependencies is facilitated with a use of a transformer decoder, and it is relatively significant for modeling at a pixel level for segmentation. 1. Patch Embedding and Positional Encoding: Embedded feature map FASPP is divided into non- overlapping patches, and each one of them is flattened and embedded into a constant-dimensional vector. Let X RN×D represent a matrix of embedded patches, with N representing several patches and D representing an embedding dimension. Positional encoding P RN×D is added to X in an attempt to preserve spatial relationships: Z0

= X + P In such a case, Z0 is an initial sequence of positional encoding patches, critical for preserving information about location in segmentation processes.

2. Contextualization at a Global Level through Self-Attention: The trans-former decoder employs a multi-head self-attention mechanism, in which dependencies between patches can be captured. For a given patch, relations between all patches can be processed through self-attention, in such a manner that both contextual information at a local and a global level can be focused in the model. Scaled dot-product attention is:

where Q, K, and V represent query, key, and value and P and T represent predicted and actual masks respectively 3. Optimization We have used the AdamW optimizer to mitigate over-parameterization, a common issue in transformer models. Learning rate scheduling with warm- up and cosine decay was adopted to stabilize training and prevent overfitting.

E. Metrics for evaluation 1Mean Intersection over Union (mIoU) mIoU is derived from matrices of Z0. D k is the dimension of the key 1 the primary metric for model performance evaluation. It estimates the intersection between predicted and vector segments, and d scales gradients during training.

3. Patch-wise Attention for Refinement of Features: On top of shared feature representation, the patch-wise attention

the ground truth:

IoU = —P T — —P T —

F. mechanism refines such a representation to sharpen its spatial detail by allowing its model to focus its attention over a group of relevant regions. With a sharpened representation Z, whose softmax will be utilized in a patch-wise manner: Z = soft max(QKT)V, with Q, K, and V is the query, key, and value matrices, respectively, following an initial self-attention mechanism. The sharpened output Z is then reshaped in an attempt to rebuild its spatial dimensions, allowing for prediction at a pixel level for segmentation. 4. Prediction Map Creation: Once its features have been sharpened through self-attention, output Z is reshaped in terms of its spatial dimensions, similar to its input form. With each pixel then having a label determined through its fine-tuned representations, a final segmentation map is produced.

Spred = reshape(Z)

The resulting Spred is a high-density pixel-wise prediction, generating a segmented image with enhanced accuracy and sharpened boundary detail.

D. Training Process

Transfer Learning and Fine-Tuning The Vision Trans-former model is initially pre-trained with the ImageNet classification dataset, leveraging big-data feature learning. For semantic segmentation, we fine-tune the model with domain-specific segmentation datasets, such as ADE20K and Cityscapes.

1. LossFunctions

The base loss function is pixel-wise cross-entropy loss, in which model estimates probability distribution over classes for each pixel: N LCE = yi log yˆi i=1 Here, yi is actual label for pixel i and yˆi is predicted labelThe average IoU across all classes provides the mIoU score.

Pixel accuracy Pixel accuracy is a proportion of accurately segmented pixels in a segmentation map:

Accu- racy = (Number of Corrected Predicted Pixels)/ (Total Number of Pixels)

Computational Efficiency and Optimizations

1. Hierarchical Transformer Encoder-Decoding In a move towards reduced computational cost in ViT, a hierarchical model of a transformer was embraced, with processing at a range of scales supported. This reduces computation requirements but maintains long-term dependencies modeled effectively.

2. Mixed-Precision Training Mixed-precision training aided in reduced memory usage and training efficiency with no loss in model accuracy. It was most useful with model and computation intensity, specifically with large model sizes and computationally costly semantic segmentation operations.

V. QUALITATIVE ANALYSIS In this section, we have included a qualitative analysis of our model. In the following photographs, a part of the output received through our model's prediction with regard to ground truth is displayed. These results highlight the model's ability to capture fine details and perform semantic segmentation effectively.

LDice = 1 2—P T — —P — + —T—

Fig. 2. Segmented images after segmentation

FIG 2.

REFERENCES

[1] Author(s). Fully transformer networks for semantic image segmentation. arXiv preprint arXiv:XXXX.XXXX, 202X.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention-augmented convolutional networks for semantic seg- mentation. *arxiv.org*, 2019.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to- end object detection with transformers. *arxiv.org*, 2020.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arxiv.org*, 2021

[5] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Max- deeplab: End-to-end panoptic segmentation with mask transformers. *arxiv.org*, 2021.

[6] Tomer Cohen, Lasse Rasmus, and Ender Konukoglu. Cotr: Efficient 3d medical image segmentation with cnn-transformer hybrid networks. *arxiv.org*, 2021.

[7] Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:XXXX.XXXX*, 202X.

[8] Wang et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *cvfopenaccess.org*, 202X.

[9] Zheng et al. Segmenter: Transformer for semantic segmentation. *arXiv preprint arXiv:XXXX.XXXX*, 202X.

[10] Ali Hatamizadeh, Dong Yang, Holger R Roth, and Daguang Xu. Unetformer: A unet-like transformer for medical image segmentation. *arxiv.org*, 2021.

[11] Paul F. Jaeger, Simon A. A. Kohl, Sebastian Bickelhaupt, Fabian Isensee, and K.H. Maier-Hein. Medical transformer: Axial attention for medical image segmentation. *arxiv.org*, 2021.

[12] Jian Liu, Jiexiong Zhang, Ying Wu, and Yu Zhang. Swin-deep lab: A hybrid transformer-cnn framework for panoptic segmentation. *arxiv.org*, 2021.

[13] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, and Jie Zhou. Dense prediction transformer. *arxiv.org*, 2021.

[14] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *arxiv.org*, 2021.

[15] Long Tian, Ke Li, Wei Feng, and Yu Yao. Twins transformer: A novel transformer architecture for dense image segmentation. *arxiv.org*, 2021.

[16] Jeyhan Valanarasu and et al. Medical transformer: Gated axial- attention for medical image segmentation. *arXiv preprint arXiv:2102.05548*, 2021.

[17] Qiang Wang, Fenghua Ma, Xianghao Meng, Lingyu Zhang, and Yonghong Liu. Segmenting with transformers for high- resolution remote sensing data. *MDPI*, 2021.

[18] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Guangyi Song, Dingkang Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convo- lutions. *CVF Open Access*, 2021.

[19] Yuhui Yuan, Rongguang Fu, Lang Huang, Wei Lin, Chao Zhang, Xiaoqi Chen, Huijuan Wang, Yalong Xiong, and Yanwei Fu. Hrformer: High-resolution transformer for dense prediction. *arxiv.org*, 2021.

[20] Haiyang Zheng, Yabin Gao, Jianyuan Han, Pengfei Wang, Wenjun Sun, and Hongkai Zhou. Translab: Hybrid transformer network for end-to-end object detection. *arxiv.org*, 2021.

