جامعة محمد بن زايد للذكاء الاصطناعي
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

BMVC 2022
Nov. 21-24, 2022, London

TransResNet: Integrating the Strengths of ViTs and CNNs for High-Resolution Medical Image Segmentation via Feature Grafting

Muhammad Hamza Sharif, Dmitry Demidov, Asif Hanif, Mohammad Yaqub, Min Xu

CODE

Introduction

Background

- High-resolution images have rich semantic information that can improve the diagnostic capability of the underlying method.

Challenges

- Existing DL methods for medical image segmentation (IS) are designed for input images having small spatial dimensions and perform poorly on high-resolution images due to discrenpiences between sampling depth and receptive field size.
- CNN-based (IS) methods cannot capture global context details due to narrow and fixed receptive fields, while transformers-based (IS) methods are computationally prohibitive and often downsample the input image before processing.

Solution

- We need a combined approach that captures rich local and global information without increasing the computational complexity associated with high-resolution images and gives accurate segmentation results.

Method

- We introduce **TransResNet** using two encoder modules: One is ***CNN*** based for extracting local feature details, other is ***transformer*** based for extracting global features.
- We introduce ***Cross Grafting Module (CGM)*** to combine the feature maps from both encoder branches. *CGM* generates ***grafted features*** enriched in both local and global semantic cues.
- We design our decoder in a ***staggered manner***, which first receives the flow of features from the transformer branch, followed by a cross grafting module, and finally CNN branch.

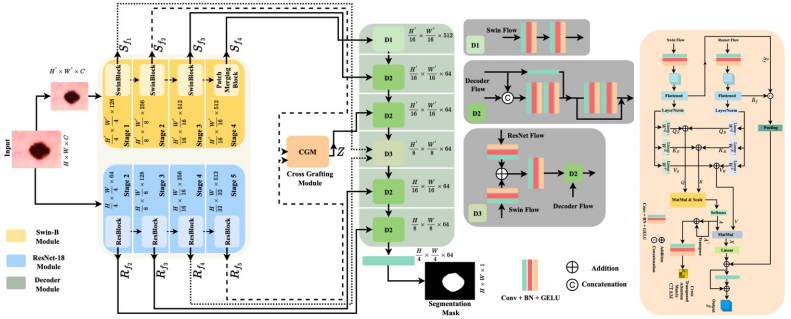

Figure 1: An overview of the architecture of **TransResNet** for high-resolution medical image segmentation. Our **TransResNet** uses the parallel branches from Swin-transformer and Resnet-18 backbones as encoders. The core module of our architecture is the **Cross Grafting Module (CGM)**, which is used to learn enriched features The decoder module aggregates the flow of feature input maps from the swin block, CGM block, and ResNet block. D1, D2, and D3 are subblocks of the decoder with their structure on the right side.

Experimental Results

We evaluate our method on ten datasets for three segmentation tasks

- Skin Lesion Segmentation (2 datasets)
- Retinal Vessel Segmentation (3 datasets)
- Polyp Segmentation (5 datasets)

Methods	Kvasir		ClinicDB		ColonDB		EndoScene		ETIS	
	mDice↑	*mIoU*↑	*mDice*↑	*mIoU*↑	*mDice*↑	*mIoU*↑	*mDice*↑	*mIoU*↑	*mDice*↑	*mIoU*↑
U-Net [30]	0.818	0.746	0.823	0.750	0.512	0.444	0.710	0.627	0.398	0.335
U-Net++ [47]	0.821	0.743	0.794	0.729	0.483	0.410	0.707	0.624	0.401	0.344
ResUNet++ [1]	0.813	0.793	0.796	0.796	-	-	-	-	-	-
SFA [8]	0.723	0.611	0.700	0.607	0.469	0.347	0.467	0.329	0.297	0.217
PraNet [7]	0.898	0.840	0.899	0.849	0.712	0.640	0.871	0.797	0.628	0.567
SANet [41]	**0.904**	**0.847**	0.916	0.859	**0.753**	**0.670**	**0.888**	**0.815**	**0.750**	**0.654**
TransFuse [46]	**0.918**	**0.868**	**0.918**	**0.868**	**0.773**	**0.696**	**0.902**	**0.833**	**0.733**	**0.659**
TransResNet	0.881	0.824	**0.917**	**0.861**	0.685	0.604	0.874	0.804	0.564	0.493

Table 1: Quantitative results on **polyp segmentation datasets** compared with seven SOTA methods. The red and green color cells represent the highest and the second highest scores respectively. Performance is measured by **mean Dice** and **mean IoU** scores. "-" indicates results are not available.

Methods	ISIC-2016		test-PH2	
	mIoU↑	*mDice*↑	*mIoU*↑	*mDice*↑
U-Net [30]	0.825	0.878	0.739	0.836
U-Net++ [47]	0.818	0.889	0.812	0.889
Attn U-Net [27]	0.797	0.874	0.695	0.805
CE-Net [13]	**0.842**	**0.905**	**0.824**	0.894
CA-Net [12]	0.807	0.881	0.751	0.846
TransFuse [46]	0.840	0.900	0.823	**0.897**
TransResNet	**0.843**	**0.907**	**0.831**	**0.905**

Table 2: Quantitative results on **skin lesion segmentation** datasets compared with six SOTA methods. Performance is measured by **mean Dice** and **mean IoU** scores.

Methods	HRF	IOSTAR	CHASE
	mF1↑	*mF1*↑	*mF1*↑
DRIU [23]	0.783	**0.825**	0.810
HED [43]	0.783	**0.825**	0.810
M2U-Net [20]	0.780	0.817	0.802
U-Net [30]	**0.788**	0.812	**0.812**
TransResNet	**0.802**	**0.835**	**0.821**

Table 3: Quantitative results on **retinal vessel segmentation** datasets compared with four SOTA methods. Performance is measured by the **mean F1** score.

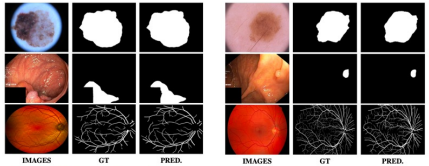

Figure 2: Qualitative results on all three segmentation tasks. The figure shows an example image, ground truth (**GT**), and predicted (**PRED**) segmentation mask for the skin lesion segmentation task (row 1), the polyp segmentation task (row 2), and the retinal vessel segmentation task (row 3).

Analysis

Train Image Resolution	Test Image Resolution	Model Performance
Higher	Higher	Increases
Lower & upscaled	Higher	Decreases
Higher	Lower & upscaled	Increases
Lower & upscaled	Lower & upscaled	Slightly decreases

Table 4: Analysis of model performance based on image resolution during training and inference. Lower-resolution images are upscaled to **1024 x 1024.**

Acknowledgment

This research work is supported by the **Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)**. For more details, please contact: muhammad.sharif@mbzuai.ac.ae or friend me via Linkedin.