TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation

Yundong Zhang[*1], Huiye Liu[*1,2] [✉], and Qiang Hu[1]

[1] Rayicer, Suzhou, China
huiyeliu@rayicer.com
[2] Georgia Institute of Technology, Atlanta, GA, USA

Abstract. Medical image segmentation - the prerequisite of numerous clinical needs - has been significantly prospered by recent advances in convolutional neural networks (CNNs). However, it exhibits general limitations on modeling explicit long-range relation, and existing cures, resorting to building deep encoders along with aggressive downsampling operations, leads to redundant deepened networks and loss of localized details. Hence, the segmentation task awaits a better solution to improve the efficiency of modeling global contexts while maintaining a strong grasp of low-level details. In this paper, we propose a novel parallel-in-branch architecture, TransFuse, to address this challenge. TransFuse combines Transformers and CNNs in a parallel style, where both global dependency and low-level spatial details can be efficiently captured in a much shallower manner. Besides, a novel fusion technique - BiFusion module is created to efficiently fuse the multi-level features from both branches. Extensive experiments demonstrate that TransFuse achieves the newest state-of-the-art results on both 2D and 3D medical image sets including polyp, skin lesion, hip, and prostate segmentation, with significant parameter decrease and inference speed improvement.

Keywords: Medical Image Segmentation · Transformers · Convolutional Neural Networks · Fusion

1 Introduction

Convolutional neural networks (CNNs) have attained unparalleled performance in numerous medical image segmentation tasks [9,12], such as multi-organ segmentation, liver lesion segmentation, brain 3D MRI, etc., as it is proved to be powerful at building hierarchical task-specific feature representation by training the networks end-to-end. Despite the immense success of CNN-based methodologies, its lack of efficiency in capturing global context information remains a challenge. The chance of sensing global information is equaled by the risk of efficiency, because existing works obtain global information by generating very large receptive fields, which requires consecutively down-sampling and stacking

---

[*] These authors contributed equally to this work.

convolutional layers until deep enough. This brings several drawbacks: 1) training of very deep nets is affected by the diminishing feature reuse problem [23], where low-level features are washed out by consecutive multiplications; 2) local information crucial to dense prediction tasks, e.g., pixel-wise segmentation, is discarded, as the spatial resolution is reduced gradually; 3) training parameter-heavy deep nets with small medical image datasets tends to be unstable and easily overfitting. Some studies [29] use the non-local self-attention mechanism to model global context; however, the computational complexity of these modules typically grows quadratically with respect to spatial size, thus they may only be appropriately applied to low-resolution maps.

Transformer, originally used to model sequence-to-sequence predictions in NLP tasks [26], has recently attracted tremendous interests in the computer vision community. The first purely self-attention based vision transformers (ViT) for image recognition is proposed in [7], which obtained competitive results on ImageNet [6] with the prerequisite of being pretrained on a large external dataset. SETR [32] replaces the encoders with transformers in the conventional encoder-decoder based networks to successfully achieve state-of-the-art (SOTA) results on the natural image segmentation task. While Transformer is good at modeling global context, it shows limitations in capturing fine-grained details, especially for medical images. We independently find that SETR-like pure transformer-based segmentation network produces unsatisfactory performance, due to lack of spatial inductive-bias in modelling local information (also reported in [4]).

To enjoy the benefit of both, efforts have been made on combining CNNs with Transformers, e.g., TransUnet [4], which first utilizes CNNs to extract low-level features and then passed through transformers to model global interaction. With skip-connection incorporated, TransUnet sets new records in the CT multi-organ segmentation task. However, past works mainly focus on replacing convolution with transformer layers or stacking the two in a sequential manner. To further unleash the power of CNNs plus Transformers in medical image segmentation, in this paper, we propose a different architecture—*TransFuse*, which runs shallow CNN-based encoder and transformer-based segmentation network in parallel, followed by our proposed *BiFusion* module where features from the two branches are fused together to jointly make predictions. TransFuse possesses several advantages: 1) both low-level spatial features and high-level semantic context can be effectively captured; 2) it does not require very deep nets, which alleviates gradient vanishing and feature diminishing reuse problems; 3) it largely improves efficiency on model sizes and inference speed, enabling the deployment at not only cloud but also edge. To the best of our knowledge, TransFuse is the first parallel-in-branch model synthesizing CNN and Transformer. Experiments demonstrate the superior performance against other competing SOTA works.

2   Proposed Method

As shown in Fig. 1, TransFuse consists of two parallel branches processing information differently: 1) CNN branch, which gradually increases the receptive field

and encodes features from local to global; 2) Transformer branch, where it starts with global self-attention and recovers the local details at the end. Features with same resolution extracted from both branches are fed into our proposed BiFusion Module, where self-attention and bilinear Hadamard product are applied to selectively fuse the information. Then, the multi-level fused feature maps are combined to generate the segmentation using gated skip-connection [20]. There
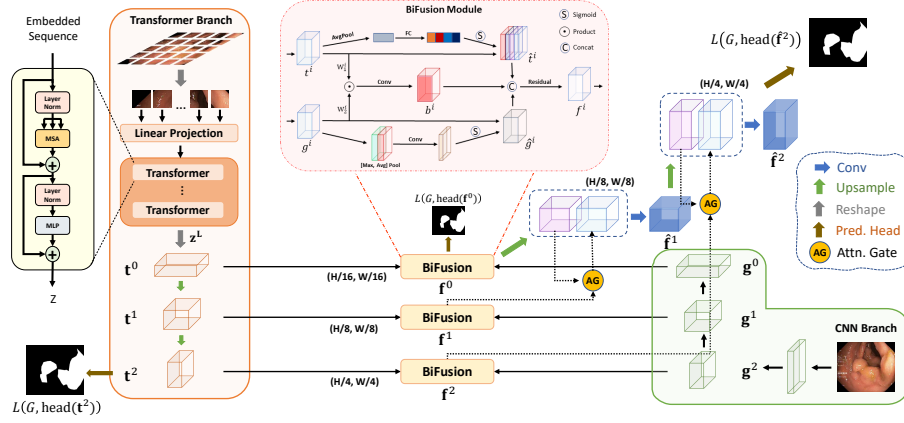

Fig. 1: Overview of TransFuse (best viewed in color): two parallel branches - CNN (bottom right) and transformer (left) fused by our proposed BiFusion module.

are two main benefits of the proposed branch-in-parallel approach: firstly, by leveraging the merits of CNNs and Transformers, we argue that TransFuse can capture global information without building very deep nets while preserving sensitivity on low-level context; secondly, our proposed BiFusion module may simultaneously exploit different characteristics of CNNs and Transformers during feature extraction, thus making the fused representation powerful and compact.

Transformer Branch. The design of Transformer branch follows the typical encoder-decoder architecture. Specifically, the input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ is first evenly divided into $N = \frac{H}{S} \times \frac{W}{S}$ patches, where S is typically set to 16. The patches are then flattened and passed into a linear embedding layer with output dimension D_0, obtaining the raw embedding sequence $\mathbf{e} \in \mathbb{R}^{N \times D_0}$. To utilize the spatial prior, a learnable positional embeddings of the same demension is added to $\mathbf{e}$. The resulting embeddings $\mathbf{z}^0 \in \mathbb{R}^{N \times D_0}$ is the input to Transformer encoder, which contains L layers of multiheaded self-attention (MSA) and Multilayer Perceptron (MLP). We highlight that the self-attention (SA) mechanism, which is the core principal of Transformer, updates the states of each embedded patch by aggregating information globally in every layer:

$$\mathrm{SA}(\mathbf{z}_i) = \mathrm{softmax}\left(\frac{\mathbf{q_i}\mathbf{k}^T}{\sqrt{D_h}}\right)\mathbf{v},\tag{1}$$

where $[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{W}_{qkv}$, $\mathbf{W}_{qkv} \in \mathbb{R}^{D_0 \times 3D_h}$ is the projection matrix and vector $\mathbf{z}_i \in \mathbb{R}^{1 \times D_0}$, $\mathbf{q_i} \in \mathbb{R}^{1 \times D_h}$ are the ith row of $\mathbf{z}$ and $\mathbf{q}$, respectively. MSA is an extension of SA that concatenates multiple SAs and projects the latent dimension back to $\mathbb{R}^{D_0}$, and MLP is a stack of dense layers (refer to [7] for details of MSA and MLP). Layer normalization is applied to the output of the last transformer layer to obtain the encoded sequence $\mathbf{z}^L \in \mathbb{R}^{N \times D_0}$. For the decoder part, we use progressive upsampling (PUP) method, as in SETR [32]. Specifically, we first reshape $\mathbf{z}^L$ back to $\mathbf{t}^0 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times D_0}$, which could be viewed as a 2D feature map with D_0 channels. We then use two consecutive standard upsampling-convolution layers to recover the spatial resolution, where we obtain $\mathbf{t}^1 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times D_1}$ and $\mathbf{t}^2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times D_2}$, respectively. The feature maps of different scales $\mathbf{t}^0$, $\mathbf{t}^1$ and $\mathbf{t}^2$ are saved for late fusion with corresponding feature maps of the CNN branch.

CNN Branch. Traditionally, features are progressively downsampled to $\frac{H}{32} \times \frac{W}{32}$ and hundreds of layers are employed in deep CNNs to obtain global context of features, which results in very deep models draining out resources. Considering the benefits brought by Transformers, we remove the last block from the original CNNs pipeline and take advantage of the Transformer branch to obtain global context information instead. This gives us not only a shallower model but also retaining richer local information. For example, ResNet-based models typically have five blocks, each of which downsamples the feature maps by a factor of two. We take the outputs from the 4th ($\mathbf{g}^0 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_0}$), 3rd ($\mathbf{g}^1 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_1}$) and 2nd ($\mathbf{g}^2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_2}$) blocks to fuse with the results from Transformer (Fig. 1). Moreover, our CNN branch is flexible that any off-the-shelf convolutional network can be applied.

BiFusion Module. To effectively combine the encoded features from CNNs and Transformers, we propose a new BiFusion module (refer to Fig. 1) that incorporates both self-attention and multi-modal fusion mechanisms. Specifically, we obtain the fused feature representation $\mathbf{f}^i, i = 0, 1, 2$ by the following operations:

$$
\begin{aligned}
\hat{\mathbf{t}}^i &= \text{ChannelAttn}(\mathbf{t}^i) & \hat{\mathbf{g}}^i &= \text{SpatialAttn}(\mathbf{g}^i) \\
\hat{\mathbf{b}}^i &= \text{Conv}(\mathbf{t}^i \mathbf{W}_1^i \odot \mathbf{g}^i \mathbf{W}_2^i) & \mathbf{f}^i &= \text{Residual}([\hat{\mathbf{b}}^i, \hat{\mathbf{t}}^i, \hat{\mathbf{g}}^i])
\end{aligned}
\tag{2}
$$

where $W_1^i \in \mathbb{R}^{D_i \times L_i}$, $W_2^i \in \mathbb{R}^{C_i \times L_i}$, $|\odot|$ is the Hadamard product and Conv is a 3x3 convolution layer. The channel attention is implemented as SE-Block proposed in [10] to promote global information from the Transformer branch. The spatial attention is adopted from CBAM [30] block as spatial filters to enhance local details and suppress irrelevant regions, as low-level CNN features could be noisy. The Hadamard product then models the fine-grained interaction between features from the two branches. Finally, the interaction features $\hat{\mathbf{b}}^i$ and attended features $\hat{\mathbf{t}}^i, \hat{\mathbf{g}}^i$ are concatenated and passed through a Residual block. The resulting feature $\mathbf{f}^i$ effectively captures both the global and local context for the current spatial resolution. To generate final segmentation, $\mathbf{f}^i$s are combined using the attention-gated (AG) skip-connection [20], where we have $\hat{\mathbf{f}}^{i+1} = \text{Conv}([\text{Up}(\hat{\mathbf{f}}^i), \text{AG}(\mathbf{f}^{i+1}, \text{Up}(\hat{\mathbf{f}}^i))])$ and $\hat{\mathbf{f}}^0 = \mathbf{f}^0$, as in Fig. 1.

Loss Function. The full network is trained end-to-end with the weighted IoU loss and binary cross entropy loss $L = L_{IoU}^w + L_{bce}^w$, where boundary pix-

els receive larger weights [17]. Segmentation prediction is generated by a simple head, which directly resizes the input feature maps to the original resolution and applies convolution layers to generate M maps, where M is the number of classes. Following [8], We use deep supervision to improve the gradient flow by additionally supervising the transformer branch and the first fusion branch. The final training loss is given by $\mathcal{L} = \alpha L\left(G, \text{head}(\hat{\mathbf{f}}^2)\right) + \gamma L\left(G, \text{head}(\mathbf{t}^2)\right) + \beta L\left(G, \text{head}(\mathbf{f}^0)\right)$, where α, γ, β are tunnable hyperparameters and G is groundtruth.

3   Experiments and Results

Data Acquisition. To better evaluate the effectiveness of TransFuse, four segmentation tasks with different imaging modalities, disease types, target objects, target sizes, etc. are considered: 1) *Polyp Segmentation*, where five public polyp datasets are used: Kvasir [14], CVC-ClinicDB [2], CVC-ColonDB [24], EndoScene [27] and ETIS [21]. The same split and training setting as described in [8,11] are adopted, i.e. 1450 training images are solely selected from Kvasir and CVC-ClinicDB while 798 testing images are from all five datasets. Before processing, the resolution of each image is resized into $352{\times}352$ as [8,11]. 2) *Skin Lesion Segmentation*, where the publicly available 2017 International Skin Imaging Collaboration skin lesion segmentation dataset (ISIC2017) [5] is used[3]. ISIC2017 provides 2000 images for training, 150 images for validation and 600 images for testing. Following the setting in [1], we resize all images to $192{\times}256$. 3) *Hip Segmentation*, where a total of 641 cases are collected from a hospital with average size of $2942{\times}2449$ and pixel spacing as 0.143mm[4]. Each image is annotated by a clinical expert and double-blind reviewed by two specialists. We resized all images into 352×352, and randomly split images with a ratio of 7:1:2 for training, validation and testing. 4)*Prostate Segmentation*, where volumetric Prostate Multi-modality MRIs from the Medical Segmentation Decathlon [22] are used. The dataset contains multi-modal MRIs from 32 patients, with a median volume shape of $20 \times 320 \times 319$. Following the setting in [12], we reshape all MRI slices to 320×320, and independently normalize each volume using z-score normalization.

Implementation Details. TransFuse was built in PyTorch framework [16] and trained using a single NVIDIA-A100 GPU. The values of α, β and γ were set to 0.5, 0.3, 0.2 empirically. Adam optimizer with learning rate of 1e-4 was adopted

---

[3] Another similar dataset ISIC2018 was not used because of the missing test set annotation, which makes fair comparison between existing works can be hardly achieved.

[4] All data are from different patients and with ethics approval, which consists of 267 patients of Avascular Necrosis, 182 patients of Osteoarthritis, 71 patients of Femur Neck Fracture, 33 patients of Pelvis Fracture, 26 patients of Developmental Dysplasia of the Hip and 62 patients of other dieases.

and all models were trained for 30 epochs as well as batch size of 16, unless otherwise specified.

In polyp segmentation experiments, no data augmentation was used except for multi-scale training, as in [8,11]. For skin lesion and hip segmentation, data augmentation including random rotation, horizontal flip, color jittering, etc. were applied during training. A smaller learning rate of 7e-5 was found useful for skin lesion segmentation. Finally, we follow the nnU-Net framework [12] to train and evaluate our model on Prostate Segmentation, using the same data augmentation and post-processing scheme. As selected pretrained datasets and branch backbones may affect the performance differently, three variants of TransFuse are provided to 1) better demonstrate the effectiveness as well as flexibility of our approach; 2) conduct fair comparisons with other methods. *TransFuse-S* is implemented with ResNet-34 (R34) and 8-layer DeiT-Small (DeiT-S) [25] as backbones of the CNN branch and Transformer branch respectively. Similarly, *TransFuse-L* is built based on Res2Net-50 and 10-layer DeiT-Base (DeiT-B), while *TransFuse-L** uses ResNetV2-50 and ViT-B [7]. Note that ViTs and DeiTs have the same backbone architecture and they mainly differ in the pre-trained strategy and dataset: the former is trained on ImageNet21k while the latter is trained on ImageNet1k with heavier data augmentation.

Evaluation Results TransFuse is evaluated on both 2D and 3D datasets to demonstrate the effectiveness. As different medical image segmentation tasks serve different diagnosis or operative purposes, we follow the commonly used evaluation metrics for each of the segmentation tasks to quantitatively analyze the results. Selected visualization results of *TransFuse-S* are shown in Fig. 2.

    Results of Polyp Segmentation. We first evaluate the performance of our proposed method on polyp segmentation against a variety of SOTA methods, in terms of mean Dice (mDice) and mean Intersection-Over-Union (mIoU). As in Tab. 1, our *TransFuse-S/L* outperform CNN-based SOTA methods by a large margin. Specifically, TransFuse-S achieves 5.2% average mDice improvement on the *unseen* datasets (ColonDB, EndoSene and ETIS). Comparing to other transformer-based methods, *TransFuse-L** also shows superior learning ability on Kvasir and ClinicDB, observing an increase of 1.3% in mIoU compared to TransUnet. Besides, the efficiency in terms of the number of parameters as well as inference speed is evaluated on an RTX2080Ti with Xeon(R) Gold 5218 CPU. Comparing to prior CNN-based arts, *TransFuse-S* achieves the best performance while using only 26.3M parameters, about 20% reduction with respect to HarDNet-MSEG (33.3M) and PraNet (32.5M). Moreover, *TransFuse-S* is able to run at 98.7 FPS, much faster than HarDNet-MSEG (85.3 FPS) and PraNet (63.4 FPS), thanks to our proposed parallel-in-branch design. Similarly, *TransFuse-L** not only achieves the best results compared to other Transformer-based methods, but also runs at 45.3 FPS, about 12% faster than TransUnet.

    Results of Skin Lesion Segmentation. The ISBI 2017 challenge ranked methods according to Jaccard Index [5] on the ISIC 2017 test set. Here, we use Jaccard Index, Dice score and pixel-wise accuracy as evaluation metrics. The

Table 1: Quantitative results on polyp segmentation datasets compared to previous SOTAs. The results of [4] is obtained by running the released code and we implement SETR-PUP. '-' means results not available.

Methods	Kvasir mDice	Kvasir mIoU	ClinicDB mDice	ClinicDB mIoU	ColonDB mDice	ColonDB mIoU	EndoScene mDice	EndoScene mIoU	ETIS mDice	ETIS mIoU
U-Net [18]	0.818	0.746	0.823	0.750	0.512	0.444	0.710	0.627	0.398	0.335
U-Net++ [33]	0.821	0.743	0.794	0.729	0.483	0.410	0.707	0.624	0.401	0.344
ResUNet++ [13]	0.813	0.793	0.796	0.796	-	-	-	-	-	-
PraNet [8]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567
HarDNet-MSEG [11]	0.912	0.857	0.932	0.882	0.731	0.660	0.887	0.821	0.677	0.613
TransFuse-S	**0.918**	**0.868**	0.918	0.868	**0.773**	**0.696**	0.902	0.833	0.733	0.659
TransFuse-L	**0.918**	**0.868**	0.934	0.886	0.744	0.676	**0.904**	**0.838**	0.737	0.661
SETR-PUP [32]	0.911	0.854	0.934	0.885	0.773	0.690	0.889	0.814	0.726	0.646
TransUnet [4]	0.913	0.857	0.935	0.887	0.781	0.699	0.893	0.824	0.731	0.660
TransFuse-L*	**0.920**	**0.870**	**0.942**	**0.897**	**0.781**	**0.706**	**0.894**	**0.826**	**0.737**	**0.663**

Table 2: Quantitative results on ISIC 2017 test set. Results with backbones use weights pretrained on ImageNet.

Methods	Backbones	Epochs	Jaccard	Dice	Accuracy
CDNN [31]	-	-	0.765	0.849	0.934
DDN [15]	ResNet-18	600	0.765	0.866	0.939
FrCN [1]	VGG16	200	0.771	0.871	0.940
DCL-PSI [3]	ResNet-101	150	0.777	0.857	0.941
SLSDeep [19]	ResNet-50	100	0.782	**0.878**	0.936
Unet++ [33]	ResNet-34	30	0.775	0.858	0.938
TransFuse-S	R34+DeiT-S	30	**0.795**	0.872	**0.944**

Table 3: Results on in-house hip dataset. All models use pretrained backbones from ImageNet and are of similar size ($\sim$ 26M). HD and ASD are measured in mm.

Methods	Pelvis HD	Pelvis ASD	L-Femur HD	L-Femur ASD	R-Femur HD	R-Femur ASD
Unet++ [33]	14.4	1.21	9.33	0.932	5.04	0.813
HRNetV2 [28]	14.2	1.13	6.36	0.769	5.98	0.762
TransFuse-S	**9.81**	**1.09**	**4.44**	**0.767**	**4.19**	**0.676**

comparison results against leading methods are presented in Tab. 2. *TransFuse-S* is about 1.7% better than the previous SOTA SLSDeep [19] in Jaccard score, without any pre- or post-processing and converges in less than 1/3 epochs. Besides, our results outperform Unet++ [33] that employs pretrained R34 as backbone and has comparable number of parameters with TransFuse-S (26.1M vs 26.3M). Again, the results prove the superiority of our proposed architecture.

Results of Hip Segmentation. Tab. 3 shows our results on hip segmentation task, which involves three human body parts: Pelvis, Left Femur (L-Femur) and Right Femur (R-Femur). Since the contour is more important in dianosis and THA preoperative planning, we use Hausdorff Distance (HD) and Average Surface Distance (ASD) to evaluate the prediction quality. Compared to the two advanced segmentation methods [33,28], *TransFuse-S* performs the best on both metrics and reduces HD significantly (30% compared to HRNetV2 as well as 34% compared to Unet++ on average), indicating that our proposed method is able to capture finer structure and generates more precise contour.

Results of Prostate Segmentation. We compare TransFuse-S with nnU-Net [12], which ranked 1st in the prostate segmentation challenge [22]. We follow the same preprocessing, training as well as evaluation schemes of the publicly

Table 4: Quantitative results on prostate MRI segmentation. PZ, TZ stand for the two labeled classes (peripheral and transition zone) and performance (PZ, TZ and mean) is measure by dice score.

Methods	PZ	TZ	Mean	Params	Throughput
nnUnet-2d [12]	0.6285	0.8380	0.7333	29.97M	0.209s/vol
nnUnet-3d_full[12]	0.6663	0.8410	0.7537	44.80M	0.381s/vol
TransFuse-S	**0.6738**	**0.8539**	**0.7639**	**26.30M**	**0.192s/vol**

Table 5: Ablation study on parallel-in-branch design. Res: Residual.

Index	Backbones	Composition	Fusion	Kvasir	ColonDB
E.1	R34	Sequential	-	0.890	0.645
E.2	DeiT-S	Sequential	-	0.889	0.727
E.3	R34+DeiT-S	Sequential	-	0.908	0.749
E.4	R34+VGG16	Parallel	BiFusion	0.896	0.651
E.5	R34+DeiT-S	Parallel	Concat+Res	0.912	0.764
E.6	R34+DeiT-S	Parallel	BiFusion	0.918	0.773

Table 6: Ablation study on BiFusion module. Res: Residual; TFM: Transformer; Attn: Attention.

Fusion	Jaccard	Dice	Accuracy
Concat+Res	0.778	0.857	0.939
+CNN Spatial Attn	0.782	0.861	0.941
+TFM Channel Attn	0.787	0.865	0.942
+Dot Product	0.795	0.872	0.944

available nnU-Net framework[5] and report the 5-fold cross validation results in Tab. 4. We can find that TransFuse-S surpasses nnUNet-2d by a large margin (+4.2%) in terms of the mean dice score. Compared to nnUNet-3d, TransFuse-S not only achieves better performance, but also reduces the number of parameters by ∼41% and increases the throughput by ∼50% (on GTX1080).

Ablation Study. An ablation study is conducted to evaluate the effectiveness of the parallel-in-branch design as well as BiFusion module by varying design choices of different backbones, compositions and fusion schemes. A *seen* (Kvasir) and an *unseen* (ColonDB) datasets from polyp are used, and results are recorded in mean Dice. In Tab. 5, by comparing E.3 against E.1 and E.2, we can see that combining CNN and Transformer leads to better performance. Further, by comparing E.3 against E.5, E.6, we observe that the parallel models perform better than the sequential counterpart. Moreover, we evaluate the performance of a double branch CNN model (E.4) using the same parallel structure and fusion settings with our proposed E.6. We observe that E.6 outperforms E.4 by 2.2% in Kvasir and 18.7% in ColonDB, suggesting that the CNN branch and transformer branch are complementary to each other, leading to better fusion results. Lastly, performance comparison is conducted between another fusion module comprising concatenation followed by a residual block and our proposed BiFusion module (E.5 and E.6). Given the same backbone and composition setting, E.6 with BiFusion achieves better results. Additional experiments conducted on ISIC2017 are presented in Tab. 6 to verify the design choice of BiFusion module, from which we find that each component shows its unique benefit.

---

[5] https://github.com/MIC-DKFZ/nnUNet

Polyp Segmentation

Skin Lesion Segmentation

Hip Segmentation
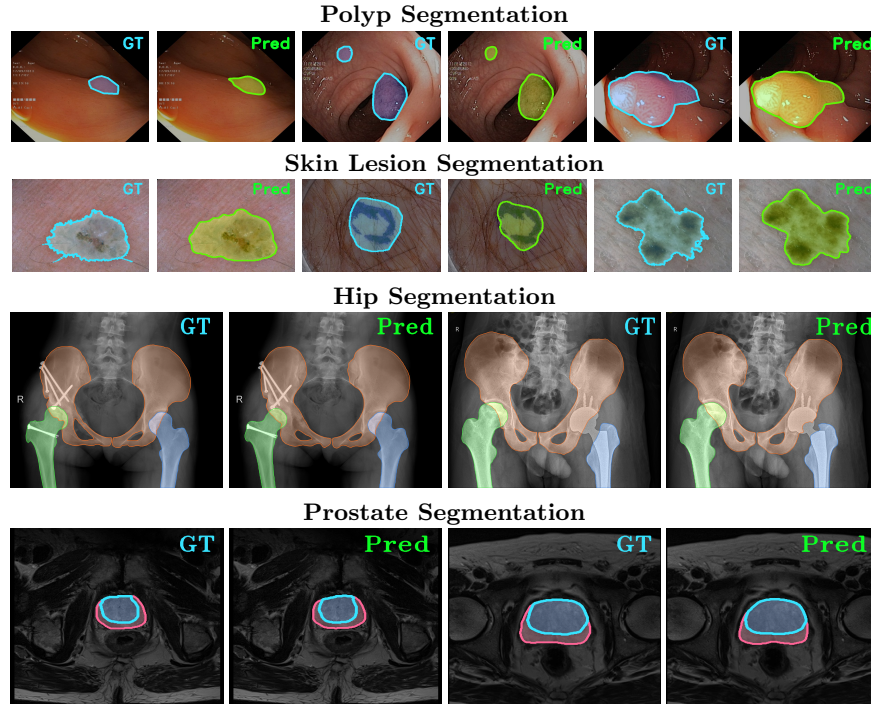

Prostate Segmentation

Fig. 2: Results visualization on all three tasks (best viewed in color). Each row follows the repeating sequence of ground truth (GT) and predictions (Pred).

4   Conclusion

In this paper, we present a novel strategy to combine Transformers and CNNs with late fusion for medical image segmentation. The resulting architecture, TransFuse, leverages the inductive bias of CNNs on modeling spatial correlation and the powerful capability of Transformers on modelling global relationship. TransFuse achieves SOTA performance on a variety of segmentation tasks whilst being highly efficient on both the parameters and inference speed. We hope that this work can bring a new perspective on using transformer-based architecture. In the future, we plan to improve the efficiency of the vanilla transformer layer as well as test TransFuse on other medical-related tasks such as landmark detection and disease classification.

References

1. Al-Masni, M.A., Al-Antari, M.A., et al.: Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Computer methods and programs in biomedicine (2018)
2. Bernal, J., Sánchez, F.J., et al.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics (2015)
3. Bi, L., Kim, J., et al.: Step-wise integration of deep class-specific learning for dermoscopic image segmentation. Pattern recognition (2019)
4. Chen, J., Lu, Y., et al.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Codella, N.C., Gutman, D., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (2018)
6. Deng, J., Dong, W., et al.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition (2009)
7. Dosovitskiy, A., Beyer, L., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2020)
9. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging (2019)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
11. Huang, C.H., Wu, H.Y., Lin, Y.L.: Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
12. Isensee, F., Jäger, P.F., et al.: Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 (2019)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM) (2019)
14. Jha, D., Smedsrud, P.H., et al.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling (2020)
15. Li, H., He, X., et al.: Dense deconvolutional network for skin lesion segmentation. IEEE journal of biomedical and health informatics (2018)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
17. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (2015)

19. Sarker, M.M.K., Rashwan, H.A., et al.: Slsdeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2018)
20. Schlemper, J., Oktay, O., et al.: Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis (2019)
21. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery (2014)
22. Simpson, A.L., Antonelli, M., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
23. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint arXiv:1505.00387 (2015)
24. Tajbakhsh, N., et al.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging (2015)
25. Touvron, H., Cord, M., et al.: Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 (2020)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
27. Vázquez, D., Bernal, J., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering (2017)
28. Wang, J., Sun, K., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020)
29. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
30. Woo, S., Park, J., et al.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV) (2018)
31. Yuan, Y., Lo, Y.C.: Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. IEEE journal of biomedical and health informatics (2017)
32. Zheng, S., Lu, J., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv preprint arXiv:2012.15840 (2020)
33. Zhou, Z., et al.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging (2019)