

TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation

Yundong Zhang¹, Huiye Liu^{1,2(✉)}, and Qiang Hu¹

¹ Rayicer, Guangzhou, China
huiyeliu@rayicer.com

² Georgia Institute of Technology, Atlanta, GA, USA

Abstract. U-Net based convolutional neural networks with deep feature representation and skip-connections have significantly boosted the performance of medical image segmentation. In this paper, we study the more challenging problem of improving efficiency in modeling global contexts without losing localization ability for low-level details. TransFuse, a novel two-branch architecture is proposed, which combines Transformers and CNNs in a parallel style. With TransFuse, 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 proposed to fuse the multi-level features from each branch. TransFuse achieves the newest state-of-the-arts on polyp segmentation task, with 20% fewer parameters and the fastest inference speed at about 98.7 FPS.

Keywords: Medical Image Segmentation · Transformers · Convolutional neural networks · Fusion

1 Introduction

Convolutional neural networks (CNNs) have achieved unprecedented performance in numerous medical image segmentation tasks, such as pelvis segmentation [3], cardiac diagnosis [11] and polyp segmentation [5,10]. Among all the variants of CNNs, the U-Net like encoder-decoder based networks have demonstrated superb performance, where coarse-grained deep features are progressively upsampled and fused with fine-grained shallow features using skip-connection, such that both the global context and high-resolution details are incorporated. Nowadays, the U-Net family keeps expanding and prospered thanks to the advances of those newly integrated components, such as residual connection [10], attention gate [12], and etc.

However, it is still challenging for U-Net to improve its efficiency on modeling global contexts. Traditionally, both stacked convolution layers and consecutive down-sampling are used in the encoders to generate sufficiently large receptive fields of deep layers, so that this problem is circumvented rather than resolved to some extent. Unfortunately, these strategies bring several drawbacks: (1) the training of very deep nets is affected by the diminishing feature reuse

problem [16,23] that low-level features are washed out by consecutive multiplications; (2) the local information crucial to dense prediction tasks, e.g., pixel-wise segmentation, is discarded, since the spatial resolution is reduced gradually; (3) training parameter-heavy deep nets with small medical image datasets tends to be unstable and easily overfitting. Besides, some studies [21] propose using the non-local self-attention mechanism to model global dependencies. Since the computational complexity of these modules typically grows quadratically with respect to spatial size, thus they may only be appropriately applied on low-resolution maps.

Recently, Transformer, originally used to model sequence-to-sequence predictions in NLP tasks [19], has attracted tremendous interests in the computer vision community. The first purely self-attention based vision transformers (ViT) for image recognition is proposed in [4], which obtained competitive results on ImageNet with the prerequisite of being pretrained on a large external dataset. SETR [25] replaces the encoders with transformers in the conventional encoder-decoder based networks to successfully achieve state-of-the-arts (SOTAs) results on the natural image segmentation task. Inspired by those works, we apply transformers into medical image segmentation. Interestingly, we find that SETR-like pure transformer-based segmentation network produces unsatisfactory performance, due to lack of spatial inductive-bias in modelling local information.

Some researchers are working on combining the CNNs with Transformers to create a hybrid structure. For example, our concurrent work TransUnet [2] utilizes CNNs to extract low-level features, which are then passed through transformers to model global interaction. With skip-connections incorporated, TransUnet set new records in the CT multi-organ segmentation task. However, those existing 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 the area of medical image segmentation, we propose a different architecture called TransFuse to combine them in this paper. TransFuse 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) it utilizes the strengths of both CNNs and Transformers since low-level spatial features and high-level semantic context can be effectively captured by these two parallel branches, respectively; (2) it does not need to build very deep nets and alleviates gradient vanishing and feature diminishing reuse problems; (3) it is highly efficient in both the model sizes and inference speed, thanks to the parallel structure. Extensive experiments demonstrate the superior performance of our proposed models.

2 Method

As shown in Fig. 1, TransFuse consists of two parallel branches that process information differently: (1) CNN branch, where it gradually increases the receptive field and encodes the feature from local to global; (2) Transformer branch,

where it starts with global self-attention and recover the local details at the end. Besides, features of the same resolution extracted from each branch are fused together using our proposed BiFusion Module. Finally, the multi-level fused feature maps are combined using gated skip-connection [12] and generate the segmentation. There are two main benefits of the proposed two-branch approach: firstly, we explicitly avoid building very deep nets to preserve low-level context, which are beneficial to segmentation; secondly, the different characteristics of CNNs and Transformers on extracting features make the fusion scheme effective, compared to others two-stream CNN-based fusion. We introduce each of the components in the following sections.

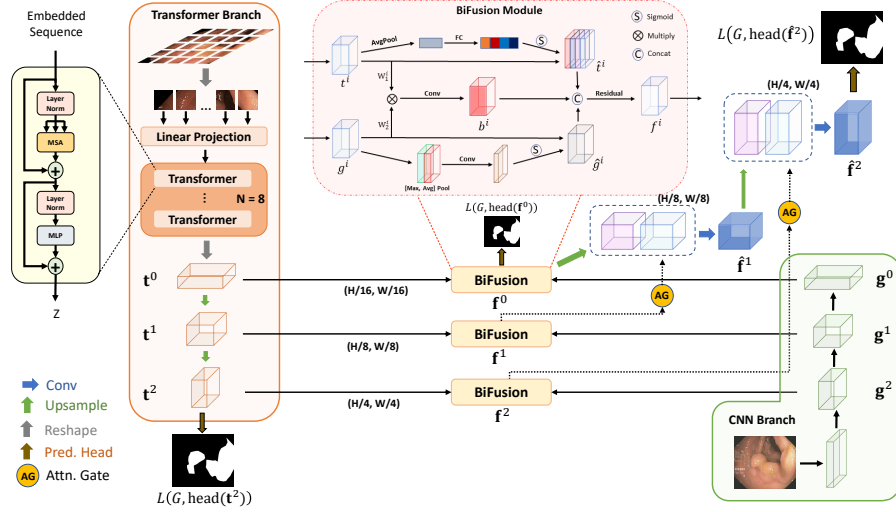


Fig. 1. Overview of TransFuse, which consists of parallel CNN branch and transformer branch. The two branches are fused with our customized BiFusion module.

2.1 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 shape is added to \mathbf{e} . The resulting embeddings $\mathbf{z} \in \mathbb{R}^{N \times D_0}$, is treated as the input to Transformer encoder, which contains L layers of multiheaded self-attention (MSA) and Multi-layer 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:

$$\text{SA}(\mathbf{z}_i) = \text{softmax} \left(\frac{\mathbf{q}_i \mathbf{k}^T}{\sqrt{D_h}} \right) \mathbf{v}, \quad (1)$$

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 $\mathbf{z}_i, \mathbf{q}_i \in \mathbb{R}^{1 \times D}$ are the i^{th} row of \mathbf{z} and \mathbf{q} , respectively. MSA is a simple extension of SA that concatenates multiple SAs and projects the latent dimension back to \mathbb{R}^{D_0} . For brevity, please refer to [4] for further details of MSA and MLP. Finally, layer normalization is applied to the output of last transformer layer, obtaining the encoded sequence $\mathbf{z}^L \in \mathbb{R}^{N \times D_0}$.

For the decoder part, we use simple progressive upsampling (PUP) method, which is adopted from SETR [25]. 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 difference scales \mathbf{t}^0 , \mathbf{t}^1 and \mathbf{t}^2 are saved for late fusion with the CNN branch.

2.2 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 downsample the feature maps by a factor of two. We take the outputs from 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.

2.3 BiFusion Module

To effectively combine the encoded features from CNNs and Transformers, we propose a new BiFusion module 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:

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

where $W_1^i \in \mathbb{R}^{D_i \times L_i}$, $W_1^i \in \mathbb{R}^{C_i \times L_i}$, $|\odot|$ is element-wise dot product and Conv is a 3x3 convolution layer. The channel attention is implemented as SE-Block proposed in [7] to promote the global information from the Transformer branch. The spatial attention is adopted from CBAM [22] block to enhance local details and suppress irrelevant regions. The dot product then models cross relationship between features of the two branch. 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. For generating final segmentation, \mathbf{f}^i s are combined using the attention gates (AG) adopted from Attention-Unet [12], where we have $\hat{\mathbf{f}}^{i+1} = \text{Conv}([\text{Up}(\mathbf{f}^i), \text{AG}(\mathbf{f}^{i+1})])$ for $i = 0, 1$, as shown in Fig. 1.

2.4 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 pixel received larger weights [13]. 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 [5], 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 $\mathcal{L} = \alpha L(G, \text{head}(\hat{\mathbf{f}}^2)) + \gamma L(G, \text{head}(\mathbf{t}^2)) + \beta L(G, \text{head}(\mathbf{f}^0))$, where α, γ, β are tunnable hyperparameters and G is groundtruth.

3 Experiments

3.1 Datasets and Settings

We use Pytorch for implementing all the experiments, which is run on an NVIDIA RTX 2080Ti GPU. Gradient accumulation is used in case memory is insufficient for single step of full batch update. Code and models will be released soon.

Polyp Segmentation We test our proposed models on polyp segmentation task. Five public polyp datasets are chosen: Kvasir [9], CVC-ClinicDB [1], CVC-ColonDB [17], EndoScene [20] and ETIS [15]. We use the same split and training settings as in [5,8], i.e. only the training sets of Kvasir and CVC-ClinicDB are used for training, while testing is done on all the datasets. The training sets contain 1450 images in total and the test sets consist of 100 images from Kvasir, 62 images from ClinicDB, 380 images from ColonDB, 60 images from EndoScene and 196 images from ETIS. Images are resized into 352×352 and no data augmentation except multi-scale training is used. Adam optimizer with learning rate of $1e-4$ is chosen. Models are trained for 30 epochs with batchsize equals to 16. The values of α, β and γ are set to 0.5, 0.3, 0.2 empirically.

Table 1. Quantitative results on polyp segmentation datasets compared to previous SOTAs. ‘†’ represents scores taken from [5] and ‘-’ means results not applicable.

Methods	Kvasir		ClinicDB		ColonDB		EndoScene		ETIS	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net [14]	0.818	0.746	0.823	0.750	0.512	0.444	0.710	0.627	0.398	0.335
U-Net++ [26]	0.821	0.743	0.794	0.729	0.483	0.410	0.707	0.624	0.401	0.344
ResUNet-mod [24]	0.791	-	0.779	-	-	-	-	-	-	-
ResUNet++ [10]	0.813	0.793	0.796	0.796	-	-	-	-	-	-
SFA† [6]	0.723	0.611	0.700	0.607	0.469	0.347	0.467	0.329	0.297	0.217
PraNet [5]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567
HarDNet-MSEG [8]	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

3.2 Results

In this section, we compare our proposed model with state-of-the-art segmentation methods as well as present ablation study.

Results on Polyp Segmentation Table. 1 shows the performance of our proposed models, in which two variants are provided: the smaller version (TransFuse-S) use ResNet-34 (R34) and DeiT-Small (DeiT-S) [18] with 8 transformer layers as backbones for CNN branch and Transformer branch, respectively; the larger version (TransFuse-L) use Res2Net-50 (R50) and 10-layer DeiT-Base (DeiT-B) as backbone. We do not use released weights from ViT [4] as they are pretrained on ImageNet-21K and thus direct comparison with previous works will be unfair. DeITs have the same architecture with ViT but use an improved training strategy, which achieves similar performance by using ImageNet-1K only. Following [5,10], we use mean Dice and mean IoU for quantitative evaluation. Our TransFuse shows superior performance on the challenging Kvasir dataset. When using a larger version, our TransFuse-L sets a new record on the two in-domain datasets. For unseen datasets (ColonDB, EndoScene and ETIS), TransFuse consistently outperform all SOTAs by a large margin. TransFuse-S achieves on average about 5.2% improvement in terms of the mean Dice score with respect to the latest HarDNet-MSEG [8] model, while the larger version slightly suffers from overfitting on ColonDB. This demonstrates the strong learning and generalization ability of our proposed model.

Table. 2 compares the number of parameters and inference speed between our proposed models and previous arts. TransFuse-S achieves SOTAs performance with 20% less parameters. Also, by leveraging the parallel design, the two branches can be executed concurrently to achieve further acceleration, achieving 98.7 FPS on RTX2080Ti. if no parallelism is applied, TransFuse-S is still able to run at 84.5 FPS, similar to HarDNet-MSEG.

Table 2. Analysis on the efficiency of TransFuse. FPS is evaluated on RTX2080Ti with Xeon(R) Gold 5218 CPU.

Methods	Backbones	#Param.	FPS	Kvasir	ColonDB
PraNet [5]	Res2Net-50	32.5M	63.4	0.898	0.709
HarDNet-MSEG [8]	HardNet-68	33.3M	85.3	0.912	0.731
SETR-PUP [25]	DeiT-B	89.4M	63.0	0.911	0.756
TransFuse-S	R34+DeiT-S	26.3M	84.5(98.7)	0.918	0.773

Ablation Study We further present ablation study in Table. 3 to evaluate the design choices of our proposed model, by varying the encoders and decoders part of our model. Specifically, we first test sequential-based structure and evaluate SETR [25] with different backbone (Pure vs. Hybrid), from which we can see that incorporating CNNs help learning local information better. Secondly, we investigate the influence of decoding scheme, where we compare PUP with gated skip-connection [12] (Hybrid-AttnUnet). For the parallel-based structure, we implement a dual-stream CNN (DualCNN) method, where the transformer branch is replaced with another CNN while keeping the others same. We also try to use an ensemble-based methods (Hybrid-Ensemble), where two branches generate segmentation predictions independently, followed by direct averaging. Finally, we compare the BiFusion module to a concatenate and residual module (TransFuse-Concat). We find that: (1) when replacing the transformer branch with another CNN network (DualCNN), the model performance degrades significantly, due to lack of ability on modelling global context; (2) the SETR-based [25] sequential structures have limited learning ability for dense segmentation; we hypothesis that this is because the transformer encoders will destroy the low-level features and make the networks unnecessarily deep; (3) comparing to concat and residual fusion methods, our proposed BiFusion module combines the two branch more effectively.

Table 3. Ablation study on polyp segmentation.

Methods	Backbones	Encoder	Decoder	Kvasir	ColonDB
Pure-PUP	DeiT-S	Sequential	PUP	0.886	0.711
Hybrid-PUP	R34+DeiT-S	Sequential	PUP	0.898	0.758
Hybrid-AttnUnet	R34+DeiT-S	Sequential	Gated Skip.	0.908	0.749
DualCNN	R34+VGG16	Parallel	BiFusion	0.896	0.651
Hybrid-Ensemble	R34+DeiT-S	Parallel	Ensemble	0.891	0.727
TransFuse-Concat	R34+DeiT-S	Parallel	Concat+Res.	0.912	0.764
TransFuse-S	R34+DeiT-S	Parallel	BiFusion	0.918	0.773

4 Conclusion

In this paper, we present a novel strategy to combine Transformers and CNNs with simple 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 SOTAs performance on polyp segmentation meanwhile 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 in our model by using more compact attention mechanisms.

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilarinho, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* **43**, 99–111 (2015)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Deniz, C.M., Xiang, S., Hallyburton, R.S., Welbeck, A., Babb, J.S., Honig, S., Cho, K., Chang, G.: Segmentation of the proximal femur from mr images using deep convolutional neural networks. *Scientific reports* **8**(1), 1–14 (2018)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranut: Parallel reverse attention network for polyp segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 263–273. Springer (2020)
6. Fang, Y., Chen, C., Yuan, Y., Tong, K.y.: Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 302–310. Springer (2019)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141 (2018)
8. 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)
9. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: *International Conference on Multimedia Modeling*. pp. 451–462. Springer (2020)
10. 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)*. pp. 225–2255. IEEE (2019)

11. Khened, M., Kollerathu, V.A., Krishnamurthi, G.: Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical image analysis* **51**, 21–45 (2019)
12. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018)
13. 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*. pp. 7479–7489 (2019)
14. 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*. pp. 234–241. Springer (2015)
15. 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* **9**(2), 283–293 (2014)
16. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. *arXiv preprint arXiv:1505.00387* (2015)
17. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging* **35**(2), 630–644 (2015)
18. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877* (2020)
19. 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)
20. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering* **2017** (2017)
21. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7794–7803 (2018)
22. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 3–19 (2018)
23. Zagoruyko, S., Komodakis, N.: Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016)
24. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* **15**(5), 749–753 (2018)
25. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840* (2020)
26. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11. Springer (2018)