U-NET V2: RETHINKING THE SKIP CONNECTIONS OF U-NET FOR MEDICAL IMAGE SEGMENTATION

Yaopeng Peng[1] *Milan Sonka*[2] *Danny Z. Chen*[1]

[1]University of Notre Dame [2]University of Iowa

ABSTRACT

In this paper, we introduce U-Net v2, a new robust and efficient U-Net variant for medical image segmentation. It aims to augment the infusion of semantic information into low-level features while simultaneously refining high-level features with finer details. For an input image, we begin by extracting multi-level features with a deep neural network encoder. Next, we enhance the feature map of each level by infusing semantic information from higher-level features and integrating finer details from lower-level features through Hadamard product. Our novel skip connections empower features of all the levels with enriched semantic characteristics and intricate details. The improved features are subsequently transmitted to the decoder for further processing and segmentation. Our method can be seamlessly integrated into any Encoder-Decoder network. We evaluate our method on several public medical image segmentation datasets for skin lesion segmentation and polyp segmentation, and the experimental results demonstrate the segmentation accuracy of our new method over state-of-the-art methods, while preserving memory and computational efficiency. Code is available at: https://github.com/yaoppeng/U-Net_v2.

***Index Terms*—** Medical image segmentation, U-Net, Skip connections, Semantics and detail infusion

1. INTRODUCTION

With the advance of modern deep neural networks, significant progress has been made in semantic image segmentation. A typical paradigm for semantic image segmentation involves an Encoder-Decoder network with skip connections [1]. In this framework, the Encoder extracts hierarchical and abstract features from an input image, while the decoder takes the feature maps generated by the encoder and reconstructs a pixel-wise segmentation mask or map, assigning a class label to each pixel in the input image. A series of studies [2, 3, 4, 5, 6, 7, 8, 9] have been conducted to incorporate global information into the feature maps and enhance multi-scale features, resulting in substantial improvements in segmentation performance.

In the field of medical image analysis, accurate image segmentation plays a pivotal role in computer-aided diagnosis and analysis. U-Net [10] was originally introduced for medical image segmentation, utilizing skip connections to connect the encoder and decoder stages at each level. The skip connections empower the decoder to access features from earlier encoder stages, hence preserving both high-level semantic information and fine-grained spatial details. This approach facilitates precise delineation of object boundaries and extraction of small structures in medical images. Additionally, a dense connection mechanism to reduce dissimilarities between features in the encoders and decoders by concatenating features from all levels and all stages [11]. A mechanism was designed to enhance features by concatenating features of different scales from both higher and lower levels [12].

However, these connections in U-Net based models may not be sufficiently effective in integrating low-level and high-level features, and the overly complex design may make the model overfit and introduce noise. For example, in ResNet [13], a deep neural network was formed as an ensemble of multiple shallow networks, and an explicitly added residual connection illustrated that the network can struggle to learn the identity map function, even when trained on a million-scale image dataset [**?**], which demonstrate that the difficulty for a network to work as expected.

Regarding the features extracted by the encoders, the low-level features usually preserve more details but lack sufficient semantic information and may contain undesired noise. In contrast, the high-level features contain more semantic information but lack precise details (e.g., object boundaries) due to the significant resolution reduction. Simply fusing features through concatenation will heavily rely on the network's learning capacity, which is often proportional to the training dataset size. This is a challenging issue, especially in the context of medical imaging, which is commonly constrained by limited data. Such information fusion, accomplished by concatenating low-level and high-level features across multiple levels through dense connections, may limit the contribution of information from different levels and potentially introduce noise.

2. RELATED WORK

The fusion of features from different levels has been extensively studied in previous works. In In [11, 12], a dense connection was proposed to concatenate features from different levels of the encoder and decoder. Even though these proposed methods do not significantly increase the number of parameters, GPU memory consumption will rise a lot because all intermediate feature maps and the corresponding gradients must be stored for forward passes and backward gradient computations. This leads to an increase in both GPU memory usage and floating point operations (FLOPs). In [14], reverse attention was utilized to explicitly establish connections among multi-scale features. In [15], ReLU activation was applied to higher-level features and the activated features were multiplied with lower-level features. Additionally, in [16], the authors proposed to extract features from CNN and Transformer models separately, combining the features from both the CNN and Transformer branches at multiple levels to enhance the feature maps. However, these approaches are complex, and their performance remains not very satisfactory, thus desiring further improvement. In EGE-UNet [17], the authors proposed a Group multi-axis Hadamard Product Attention module (GHPA) to compute attention in the xy, yz, and

THIS RESEARCH WAS SUPPORTED IN PART BY NIH NIBIB GRANT R01-EB004640.

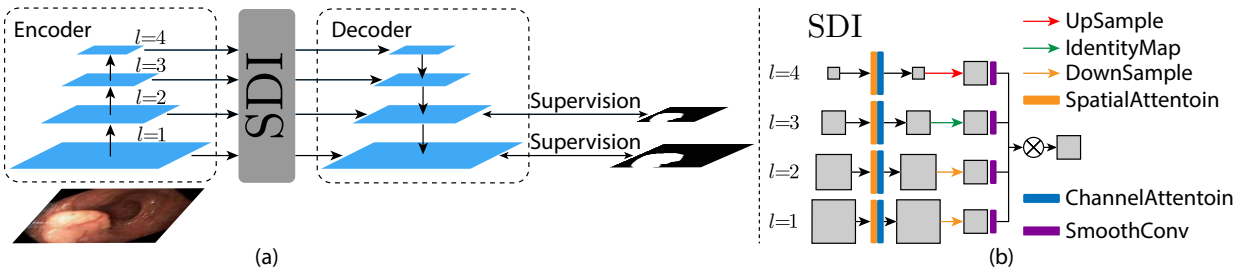

Fig. 1. (a) The overall architecture of our U-Net v2 model, which consists of an Encoder, the SDI (semantics and detail infusion) module, and a Decoder. (b) The architecture of the SDI module. For simplicity, we only show the refinement of the third level features ($l = 3$). SmoothConv denotes a 3×3 convolution for feature smoothing. $\bigotimes$ denotes the Hadamard product.

xz planes, and a Group Aggregation Bridge module (GAB) that groups the features of two consecutive levels and uses concatenation to combine them. Our method differs from EGE-UNet in that we explicitly use elementwise product to incorporate the semantics and local details of all levels, enabling the features of each level to have more global semantic information and local details. Experiments demonstrate that our method is simpler and performs better.

In this paper, we present U-Net v2, a new U-Net based segmentation framework with straightforward and efficient skip connections. Our model first extracts multi-level feature maps using a CNN or Transformer encoder. Next, for a feature map at the i-th level, we explicitly infuse higher-level features (which contain more semantic information) and lower-level features (which capture finer details) through a simple Hadamard product operation, thereby enhancing both the semantics and details of i-th level features. Subsequently, the refined features are transmitted to the decoder for resolution reconstruction and segmentation. Our method can be seamlessly integrated into any Encoder-Decoder network.

We evaluate our new method on two medical image segmentation tasks, Skin Lesion Segmentation and Polyp Segmentation, using publicly available datasets. The experimental results demonstrate that our U-Net v2 consistently outperforms state-of-the-art methods in these segmentation tasks while preserving FLOPs and GPU memory efficiency.

Dataset	Method	DSC (%)	IoU (%)
ISIC 2017	U-Net [10]	86.99	76.98
	TransFuse [16]	88.40	79.21
	MALUNet [18]	88.13	78.78
	SANet [15]	88.25	79.12
	EGE-UNet [17]	88.77	79.81
	U-Net v2 (ours)	**90.21**	**82.17**
ISIC 2018	U-Net [10]	87.55	77.86
	UNet++ [11]	87.83	78.31
	TransFuse [16]	89.27	80.63
	MALUNet [18]	89.04	80.25
	SANet [15]	89.09	80.34
	EGE-UNet [17]	89.46	80.94
	U-Net v2 (ours)	**91.52**	**84.15**

Table 1. Experimental comparison with state-of-the-art methods on the two ISIC datasets.

3. METHOD

3.1. Overall Architecture

The overall architecture of our U-Net v2 is shown in Fig. 1(a). It comprises three main modules: the encoder, the SDI (Semantic and Detail Infusion) module, and the decoder.

Given an input image I, with $I \in R^{H \times W \times C}$, the encoder produces features in M levels. We denote the i-th level features as f_i^0, $1 \leq i \leq M$. These collected features, $\{f_1^0, f_2^0, \ldots, f_M^0\}$, are then transmitted to the SDI module for further refinement.

3.2. Semantics and Detail Infusion (SDI) Module

With the hierarchical feature maps generated by the encoder, we first apply the spatial and channel attention mechanisms [8] to the features f_i^0 of each level i. This process enables the features to integrate both local spatial information and global channel information, as formulated below:

$$f_i^1 = \phi_i^c(\varphi_i^s(f_i^0)), \tag{1}$$

where f_i^1 represents the processed feature map at the i-th level, and φ_i^s and ϕ_i^c denote the parameters of spatial and channel attentions at the i-th level, respectively. Furthermore, we apply a 1×1 convolution to reduce the channels of f_i^1 to c, where c is a hyper-parameter. This resulted feature map is denoted as f_i^2, with $f_i^2 \in R^{H_i \times W_i \times c}$, where H_i, W_i, and c represent the width, height, and channels of f_i^2, respectively.

Next, we need to send the refined feature maps to the decoder. At each decoder level i, we use f_i^2 as the target reference. Then, we adjust the sizes of the feature maps at every j-th level to match the same resolution as f_i^2, formulated as:

$$f_{ij}^3 = \begin{cases} \textcircled{D}(f_j^2, (H_i, W_i)) & \text{if } j < i\,, \\ \textcircled{I}(f_j^2) & \text{if } j = i, \\ \textcircled{U}(f_j^2, (H_i, W_i)) & \text{if } j > i, \end{cases} \tag{2}$$

where $\textcircled{D}$, $\textcircled{I}$, and $\textcircled{U}$ represent adaptive average pooling, identity mapping, and bilinearly interpolating f_j^2 to the resolution of $H_i \times W_i$, respectively, with $1 \leq i, j \leq M$.

Afterwards, a 3×3 convolution is applied in order to smooth each resized feature map f_{ij}^3, formulated as:

$$f_{ij}^4 = \theta_{ij}(f_{ij}^3), \tag{3}$$

where θ_{ij} represents the parameters of the smooth convolution, and f_{ij}^4 is the j-th smoothed feature map at the i-th level.

Datasets	Method	DSC (%)	IoU (%)	MAE
Kvasir-SEG	U-Net [10]	81.8	74.6	0.055
	UNet++ [11]	82.1	74.3	0.048
	PraNet [14]	89.8	84.0	0.030
	SANet [15]	90.4	84.7	0.028
	TransFuse [16]	91.8	86.8	0.023
	Polyp-PVT [19]	91.7	86.4	0.023
	nnUNet [20]	91.9	86.8	0.022
	U-Net v2 (ours)	**92.8**	**88.0**	**0.019**
ClinicDB	U-Net [10]	82.3	75.5	0.019
	UNet++ [11]	79.4	72.9	0.022
	PraNet [14]	89.9	84.9	0.009
	SANet [15]	91.6	85.9	0.012
	TransFuse [16]	93.4	88.6	0.007
	Polyp-PVT [19]	93.7	88.9	**0.006**
	nnUNet [20]	93.6	88.7	0.007
	U-Net v2 (ours)	**94.4**	**89.6**	**0.006**
ColonDB	U-Net [10]	51.2	44.4	0.061
	UNet++ [11]	48.3	41.0	0.064
	PraNet [14]	71.2	64.0	0.043
	SANet [15]	75.3	67.0	0.043
	TransFuse [16]	74.4	67.6	0.049
	Polyp-PVT [19]	80.8	72.7	0.031
	nnUNet [20]	80.2	72.5	0.032
	U-Net v2 (ours)	**81.2**	**73.1**	**0.030**
ETIS	U-Net [10]	39.8	33.5	0.036
	UNet++ [11]	40.1	34.4	0.035
	PraNet [14]	62.8	56.7	0.031
	SANet [15]	75.0	65.4	0.015
	TransFuse [16]	73.7	67.1	0.021
	Polyp-PVT [19]	78.7	**70.6**	**0.013**
	nnUNet [20]	78.6	70.7	0.013
	U-Net v2 (ours)	**79.0**	70.5	**0.013**
Endoscene	U-Net [10]	71.0	62.7	0.022
	UNet++ [11]	70.7	62.4	0.018
	PraNet [14]	87.1	79.7	0.010
	SANet [15]	88.8	81.5	0.008
	TransFuse [16]	**90.4**	**83.8**	**0.007**
	Polyp-PVT [19]	90.0	83.3	0.007
	nnUNet [20]	88.5	81.9	0.008
	U-Net v2 (ours)	89.7	83.1	**0.007**

Table 2. Experimental comparison with state-of-the art methods on the Polyp datasets.

After resizing all the i-th level feature maps into the same resolution, we apply the element-wise Hadamard product to all the resized feature maps to enhance the i-th level features with both more semantic information and finer details, as:

$$f_i^5 = H([f_{i1}^4, f_{i2}^4, \ldots, f_{iM}^4]), \tag{4}$$

where $H(\cdot)$ denotes the Hadamard product (see Fig. 1(b)). Afterwards, f_i^5 is dispatched to the i-th level decoder for further resolution reconstruction and segmentation.

4. EXPERIMENTS

4.1. Datasets

We evaluate our new U-Net v2 using the following datasets.

Dataset	Method	DSC (%)	AVD
Retouch	U-Net [10]	84.2	0.021
	nnUNet [20]	84.3	0.023
	DconnNet [21]	87.7	0.020
	U-Net v2 (ours)	**89.1**	**0.018**

Table 3. Experimental comparison with state-of-the-art methods on the Retinal Fluid datasets.

Dataset	Method	DSC (%)	IoU (%)
ISIC 2017	UNet++ (PVT) [11]	89.60±0.17	81.16±0.07
	U-Net v2 w/o SDI	89.85±0.14	81.57±0.06
	U-Net v2 w/o SC	90.20±0.13	82.16±0.05
	U-Net v2 (ours)	**90.21±0.13**	**82.17±0.05**
ColonDB	UNet++ (PVT) [11]	78.0±4.3	69.6±3.9
	U-Net v2 w/o SDI	79.2±4.1	71.5±3.7
	U-Net v2 w/o SC	**81.3±3.7**	72.8±4.0
	U-Net v2 (ours)	81.2±3.9	**73.1±4.4**

Table 4. Ablation study on the ISIC 2017 and ColonDB datasets. SC denotes spatial and channel attentions. We use PVT as the encoder for all methods.

ISIC Datasets: Two datasets of skin lesion segmentation are used: ISIC 2017 [22, 23], which comprises 2050 dermoscopy images, and ISIC 2018 [22], which contains 2694 dermoscopy images. For fair comparison, we follow the train/test split strategy as outlined in [17]; and use the same evaluation metrics (DSC and mIoU) to measure the model performance..

Polyp Segmentation Datasets: Five datasets are used: Kvasir-SEG [24], ClinicDB [25], ColonDB [26], Endoscene [27], and ETIS [28]. For fair comparison, we use the train/test split strategy in [14]; and use the same evaluation metrics (DSC, mIoU and MAE) to measure the model performance . Specifically, 900 images from ClinicDB and 548 images from Kvasir-SEG are used as the training set, while the remaining images serve as the test set.

Retinal Fluid Dataset: [29] including 70 OCT volumes from three scanners, i.e., Cirrus, Spectralis, and Topcon. We follow the train/test split strategy outlined in [21] and use the same evaluation metrics (DSC and AVD) to measure the model's performance..

4.2. Experimental Setup

We conduct experiments on an NVIDIA P100 GPU with PyTorch. Our network is optimized using the Adam optimizer, with an initial learning rate = 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We employ a polynomial learning rate decay with a power of 0.9. The maximum number of training epochs is set to 300. For the hyperparameter c, we tried the values 16, 24, 32, 64, and 128, and found that 32 offers the best trade-off between performance and computation. Each experiment is run 5 times, and the averaged results are reported. We use the Pyramid Vision Transformer (PVT) [30] as the encoder for feature extraction. Following the practice of EGE-UNet [17], we train all the models on the ISIC dataset for 80 epochs. Similarly, following the approach of Polyp-PVT [19], we train all the models on the PolyP dataset for 100 epochs. Following the method of DconnNet [21], we train all the models on the Retouch dataset for 50 epochs.

4.3. Results and Analysis

Comparison results with state-of-the-art methods on the ISIC datasets are presented in Table 1. As shown, our proposed U-

Method	DSC (ISIC 2017)	Input size	# Params (M)	GPU memory usage (MB)	FLOPs (G)	FPS
U-Net (PVT)	89.85	(1, 3, 256, 256)	28.15	478.82	8.433	39.678
UNet++ (PVT)	89.60	(1, 3, 256, 256)	29.87	607.31	19.121	34.431
U-Net v2 (ours)	90.21	(1, 3, 256, 256)	25.02	411.42	5.399	36.631

Table 5. Comparison of computational complexity, GPU memory usage, and inference time, using an NVIDIA P100 GPU. We use PVT as the encoder for all methods.

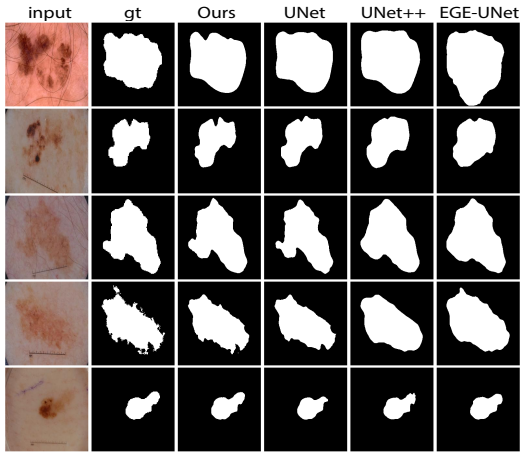

Fig. 2. Example segmentations from ISIC 2017 dataset. We use PVT as the encoder for U-Net and UNet++.

Net v2 improves the DSC scores by 1.44% and 2.06%, and the IoU scores by 2.36% and 3.21% on the ISIC 2017 and ISIC 2018 datasets, respectively, compared to EGE-UNet [17]. This improvement is due to the fact that our model incorporates the semantics and local details of all levels, allowing the features at each level to gain more global semantic cues and local details. This is in contrast to EGE-UNet, which only incorporates features from two consecutive levels. Additionally, our proposed elementwise Hadamard product is more effective compared to EGE-UNet, which uses concatenation and convolution to do the feature fusion. These improvements demonstrate the effectiveness of our proposed method for infusing semantic information and finer details into each feature map.

Comparison results with state-of-the-art methods on the polyp segmentation datasets are presented in Table 2. As shown, our proposed U-Net v2 outperforms Poly-PVT [19] on the Kavasir-SEG, ClinicDB, ColonDB, and ETIS datasets, with DSC score improvements of 1.1%, 0.7%, 0.4%, and 0.3%, respectively. This underscores the consistent effectiveness of our proposed method in infusing semantic information and finer details into feature maps at each level.

Comparison experimental results on the Retouch Retinal Fluid Datasets are shown in Table 3. As shown, our proposed U-Net v2 improved the DSC scores by 1.4% and AVD score by 0.002 on the Retouch Dataset, respectively, compared to DconnNet [21]. This underscores the consistent effectiveness of our proposed method in infusing semantic information and local details into the feature maps at each level.

4.4. Ablation Study

We conduct ablation study using the ISIC 2017 and ColonDB datasets to examine the effectiveness of our U-Net v2, as reported in Table 4. Specifically, we use the PVT [30] model as the encoder for UNet++ [11]. Note that U-Net v2 is reverted to a vanilla U-Net with a PVT backbone when our SDI module is removed. SC denotes spatial and channel attentions within the SDI module. One can see from Table 4 that UNet++ exhibits a slight performance reduction compared to U-Net v2 without SDI (i.e., U-Net with the PVT encoder). This decrease may be attributed to the simple concatenation of multi-level features generated by dense connections, which could confuse the model and introduce noise. Table 4 demonstrates that the SDI module contributes the most to the overall performance, highlighting that our proposed skip connections (i.e., SDI) consistently yield performance improvements.

4.5. Qualitative Results

Some qualitative examples on the ISIC 2017 dataset are given in Fig. 2, which demonstrate that our U-Net v2 is capable of incorporating semantic information and finer details into the feature maps at each level. Consequently, our segmentation model can capture finer details of object boundaries.

4.6. Computation, GPU Memory, and Inference Time

To examine the computational complexity, GPU memory usage, and inference time of our U-Net v2, we report the parameters, GPU memory usage, FLOPs, and FPS (frames per second) for our method, U-Net [10], and UNet++ [11] in Table 5. The experiments use float32 as the data type, which results in 4B of memory usage per variable. The GPU memory usage records the size of the parameters and intermediate variables that are stored during the forward/backward pass. $(1, 3, 256, 256)$ represents the size of the input image. All the tests are conducted on an NVIDIA P100 GPU.

In Table 5, one can observe that UNet++ introduces more parameters, and its GPU memory usage is larger due to the storage of intermediate variables (e.g., feature maps) during the dense forward process. Typically, such intermediate variables consume much more GPU memory than the parameters. Furthermore, the FLOPs and FPS of U-Net v2 are also superior to those of UNet++. The FPS reduction by our U-Net v2 compared to U-Net (PVT) is limited.

5. CONCLUSIONS

A new U-Net variant, U-Net v2, was introduced, which features a novel and straightforward design of skip connections for improved medical image segmentation. This design explicitly integrates semantic information from higher-level features and finer details from lower-level features into feature maps at each level produced by the encoder using a Hadamard product. Experiments conducted on Skin Lesion and Polyp Segmentation datasets validated the effectiveness of our U-Net v2. Complexity analysis suggested that U-Net v2 is also efficient in FLOPs and GPU memory usage.

6. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[2] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 2881–2890.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *arXiv preprint arXiv:1412.7062*, 2014.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.

[7] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[8] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[9] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018, pp. 8759–8768.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[11] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *DLMIA*. Springer, 2018, pp. 3–11.

[12] Jiawei Zhang, Yuzhen Jin, Jilan Xu, Xiaowei Xu, and Yanchun Zhang, "MDU-Net: Multi-scale densely connected U-Net for biomedical image segmentation," *arXiv preprint arXiv:1812.00352*, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[14] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao, "PraNet: Parallel reverse attention network for polyp segmentation," in *MICCAI*. Springer, 2020, pp. 263–273.

[15] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui, "Shallow attention network for polyp segmentation," in *MICCAI*. Springer, 2021, pp. 699–708.

[16] Yundong Zhang, Huiye Liu, and Qiang Hu, "TransFuse: Fusing Transformers and CNNs for medical image segmentation," in *MICCAI, Proceedings, Part I 24*. Springer, 2021, pp. 14–24.

[17] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu, "EGE-UNet: An efficient group enhanced UNet for skin lesion segmentation," *arXiv preprint arXiv:2307.08473*, 2023.

[18] Jiacheng Ruan, Suncheng Xiang, Mingye Xie, Ting Liu, and Yuzhuo Fu, "MALUNet: A multi-attention and light-weight UNet for skin lesion segmentation," in *BIBM*. IEEE, 2022, pp. 1150–1156.

[19] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao, "Polyp-PVT: Polyp segmentation with Pyramid Vision Transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[20] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[21] Ziyun Yang and Sina Farsiu, "Directional connectivity-based segmentation of medical images," in *Proceedings of the IEEE/CVF CVPR*, 2023, pp. 11525–11535.

[22] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019.

[23] Matt Berseth, "ISIC 2017-skin lesion analysis towards melanoma detection," *arXiv preprint arXiv:1703.00523*, 2017.

[24] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen, "Kvasir-SEG: A segmented polyp dataset," in *MMM, Part II 26*, 2020, pp. 451–462.

[25] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *CMIG*, vol. 43, pp. 99–111, 2015.

[26] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *TMI*, vol. 35, no. 2, pp. 630–644, 2015.

[27] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, Aaron Courville, et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, 2017.

[28] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Journal of CARS*, vol. 9, pp. 283–293, 2014.

[29] Hrvoje Bogunović, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, et al., "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *TMI*, vol. 38, no. 8, pp. 1858–1874, 2019.

[30] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions," in *CVPR*, 2021, pp. 568–578.