GCEENet: A Global Context Enhancement and Exploitation for Medical Image Segmentation

Nguyen Tuan Hung[1], Phan Ngoc Lan[1], Nguyen Thi Oanh[1], Nguyen Thi Thuy[2], and Dinh Viet Sang[1]

[1] School of Information and Communication Technology Hanoi University of Science and Technology, Hanoi, Vietnam
[2] Faculty of Information Technology, Vietnam National University of Agriculture, Hanoi, Vietnam

Abstract. Despite advancements in deep learning and computer vision, medical image segmentation is still a challenging problem. A major challenge for many segmentation models is the inherent complexity and interconnectivity of pixels in medical images. These characteristics require modeling not only local features but also a global understanding of image semantics. In this paper, we propose a deep convolutional neural network called GCEENet to effectively address the above challenges. GCEENet features a combination of global context encoders and local distribution modules, working in conjunction to preserve the global image context. Our experiments on several medical image segmentation datasets show that GCEENet outperforms current state-of-the-art models in all measured metrics.

Keywords: Deep learning · Semantic segmentation · Medical image analysis.

1 Introduction

Medical image segmentation is an essential task in medical image analysis, which is widely used in various problems such as blood vessel detection in retinal images, cell segmentation in electron microscopic recordings, brain segmentation in magnetic resonance imaging (MRI), and lung segmentation in computed tomography (CT). Algorithms that can automate these tasks can greatly aid doctors and medical professionals during diagnostics and treatment.

The proposal of convolutional neural networks (CNNs) has made deep feature learning a feasible approach for medical image segmentation. Deep learning-based segmentation methods formulate the segmentation problem as a pixel-wise classification problem whose input is the raw image matrices. Spatial components and their interactions play a central role in how these networks derive their insights. Many deep network architectures for image segmentation in general and medical image segmentation, in particular, are based on Fully Convolutional Network (FCN) [22]. U-Net [25], in particular, is a popular FCN variant that has become the de-facto standard for image segmentation, especially biomedical

image segmentation. However, U-Net and many of its variants have limitations in dealing with long-range dependencies, owing to the inherently local nature of convolution blocks. This limitation degrades performance on complex segmentation tasks that require a strong global understanding of the input image.

The addition of global context modules to the models has yielded promising results in recent years, but their utilization is still limited. In this paper, we seek to more effectively address the problem with a neural network design focused on preserving and aggregating global information. Our main contributions are:

- A novel architecture called GCEENet, capable of enhancing and exploiting global context information for improving segmentation;
- Extensive experiments to compare our proposed GCEENet with other state-of-the-art models on several public benchmark datasets. We also validate different components of GCEENet through a series of ablation studies.

The rest of the paper is organized as follows. Section 2 reviews the literature regarding convolutional neural networks and semantic segmentation in medical imaging. In Section 3, we describe the proposed network architecture in detail. Section 4 outlines our experiment settings. The results are presented and discussed in Section 5. Finally, we conclude the paper and discuss future works in Section 6.

2 Related work

2.1 Convolutional neural networks for semantic segmentation

Most early works on CNNs focused on image classification problems, using benchmarks such as the CIFAR and ImageNet datasets. Fully Convolutional Network (FCN) [22] was the first successful attempt at adopting CNNs for the semantic segmentation problem. FCN replaces the final fully connected layers in traditional CNNs with another series of convolutions that produce the segmentation map. U-Net [25] was one of the first models to address this demand by introducing an encoder-decoder architecture. This architecture featured skip connections between encoder and decoder blocks, allowing low-level information to flow to deeper layers. Many later works further refined and improved U-Net such as UNet++ [35], Attention-UNet [24], AG-ResUNet++ [16], and NeoUNet [23]. While these works focused on different aspects of U-Net, such as backbones and skip mechanism, a key bottleneck of U-Net and many other U-Net based approaches is preserving and aggregating contextual information in the extracted features.

2.2 Contextual information modeling

Small receptive fields are often the bottleneck in FCNs, preventing models from perceiving larger fields with semantic context. Initial efforts to enlarge the receptive fields while preserving spatial resolution include Global average pooling [21] and dilated convolution [31]. The DeepLab papers [6], [8], [7] proposed

atrous spatial pyramid pooling (ASPP) as an alternative. In PSPnet [34], pooling features from multiple window sizes are upsampled to the same size and then concatenated together to fuse convolutional features from multiple scales.

Self-attention is an efficient technique for various tasks, including visual recognition, machine translation, and generative modeling. In computer vision, NLNet [29] is among the first attempts at adopting self-attention to model pixel-level relations. Huang et al. [15] address a shortcoming in NLNet, which spends excessive amounts of computation on generating attention maps for every query position with CCNet. CCNet creates the surrounding context for each pixel by stacking two criss-cross blocks. Finally, CGNL-Net [32] adds channel information to attention maps and reduces computation costs via approximation.

Many of the works mentioned above focus exclusively on global context modeling. While this is indeed a major issue for segmentation problems, global information also tends to over-smooth smaller objects and boundaries. In this paper, we examine a more comprehensive approach that takes full advantage of both high-level global context and low-level local information.

3 Proposed architecture

3.1 Overview

We propose a novel neural network architecture named GCEENet (Global context enhancement and exploitation). GCEENet is built upon the ideas of U-Net, comprising an encoder and a decoder module. Convolutional and pooling blocks in the encoder extract features from the input image, creating increasingly more abstract representations while reducing the feature map's size. Output from the final encoder block is passed to the Global Context Encoder (GCE), then enhanced via the Local Distribution module (LD). The decoder, or Feature Aggregation Module (FAM), combines high-level, low-level, and global context information and upsamples feature maps to eventually form the segmentation map. Fig. 1 describes this overall architecture in detail.

3.2 Global Context Encoder module

Long-range contextual information (i.e., lower-level to high-level blocks) is critical for improving segmentation. Features containing rich global context information can help subsequent layers to learn more robust task-specific representations. The Global Context Encoder for GCEENet needs to be able to distill such information and pass it to the decoder. In this section, we investigate a number of existing building blocks that are potential candidates for the GCE module. We will perform ablation studies in Section 4 and provide conclusions on the best selection.

The first candidate is the Pyramid Pooling Module (PPM), proposed by Zhao et al. [34]. Intuitively, Pyramid Pooling is a hierarchical global prior that fuses features under four different pyramid scales with bin sizes of 1×1, 2×2, 3×3,

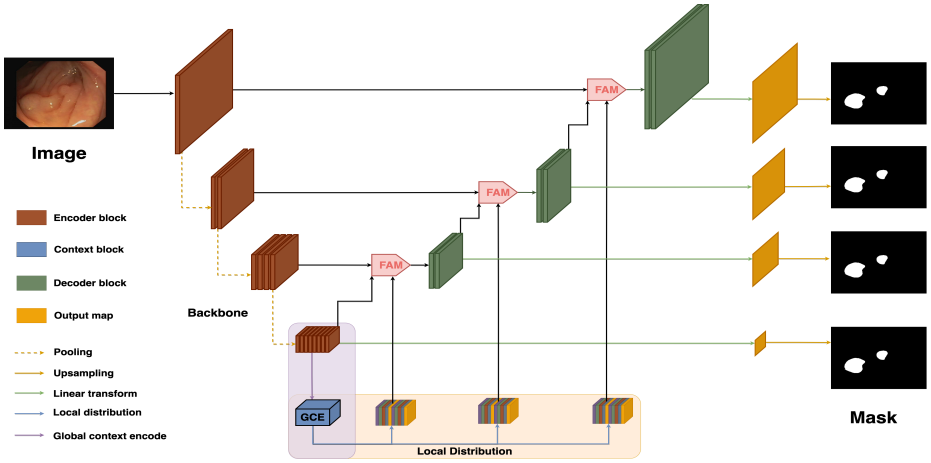


Fig. 1: Model overview

and 6×6, respectively. After each pyramid level, a 1×1 convolution layer is used to reduce the dimension of context representation. And then, the low-dimension feature maps are upsampled to get the same size as the original feature map by bilinear interpolation. PSPNet provides an effective global contextual prior for pixel-level scene parsing.

Our next candidate is Atrous Spatial Pyramid Pooling (ASPP), proposed by Chen et al. [7] in the seminal DeepLab paper. ASPP is a convolutional feature layer with filters at multiple sampling rates. This enables arbitrarily large field-of-views for convolutional filters. An Atrous convolution with rate r pads $r-1$ zeros between each consecutive value, thus inflating a k-kernel filter to $k+(k-1)(r-1)$ filters without increasing the number of parameters. As a result, the model can aggregate more context information while trading off accurate localization.

Our third candidate is the Compact Generalized Non-local network (CGNL) [32]. This module is based on Non-local Network [29], which aims to learn the relation of each position to all other positions in the input image. Such relations are achieved using the non-local mean operation:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f\left(\mathbf{x}_i, \mathbf{x}_j\right) g\left(\mathbf{x}_j\right)$$

This can be seen as direct modeling of long-range dependencies. CGNL further models relations among different channels by flattening the output of the linear transform. In addition, the representation is divided into groups and approximated via the Taylor series, which helps decrease model size. Firstly, the output is flattened after transforming feature maps. Then, the non-local mean is computed based on these vectors. Each vector has a shape of HWC, so the pairwise matrix is now (HWC)x(HWC) - much higher complexity than the original

non-local operation. To mitigate this problem, this section proposes a compact representation that leads to an affordable approximation for GNL. The use of Taylor approximation aims to reduce the model's complexity.

The final candidate module for GCE is Criss-Cross Attention, proposed by the authors of the Criss-Cross Network [15]. Criss-Cross attention stems from graph neural network designs, seeking to improve convolutional with an estimated full-image context. Context information is generated both horizontally and vertically using the affinity operation.

3.3 Local Distribution

A common problem for all global context encoders is that while large objects with long-range dependencies are accounted for, smaller objects and boundaries are over-smoothed. We remedy this issue by appending a Local Distribution (LD) module after GCE.

The Local Distribution module [20] consists of a downsample filter followed by immediate upsampling to the original size before feeding to a sigmoid function. The output mask M is sensitive to both spatial and channel. This mask is multiplied element-wise with the input feature map and then concatenated with the input.

$$\mathbf{M} = \sigma\left(\text{ upsample }\left(\mathbf{W}_d \mathbf{F}_{GA}\right)\right)$$
$$\mathbf{F}_{GALD} = \mathbf{M} \odot \mathbf{F}_{GA} + \mathbf{F}_{GA}$$

3.4 Aggregator module

While low-level features contain more detailed information such as textures, boundaries, and spatial structures, they also contain more background noise. In contrast, higher-level features can provide cleaner, more abstract semantic information. Therefore, a more comprehensive view of the image can be produced by a combination of these types of feature maps. Thus, we use the aggregator module proposed in [9]. In this work, Chen et al. proposed a module to fully integrate the three feature-level features, thereby creating a more comprehensive feature than the global perspective. Specifically, as shown in the Fig. 2, the information integration module receives three-part input, including the high-level features from the output of the previous layer, the low-level features from the corresponding bottom layer, and the context feature generated by the **GCE** module after enhancing with LD module.

3.5 Loss function

GCEENet's loss function is a sum of the weighted IoU loss and binary cross entropy loss, similar to the loss employed by PraNet [11] and F3Net [30]. The weighted IoU loss assigns higher weight values to "harder" pixels located at object boundaries. Specifically, each prediction mask S goes through a large

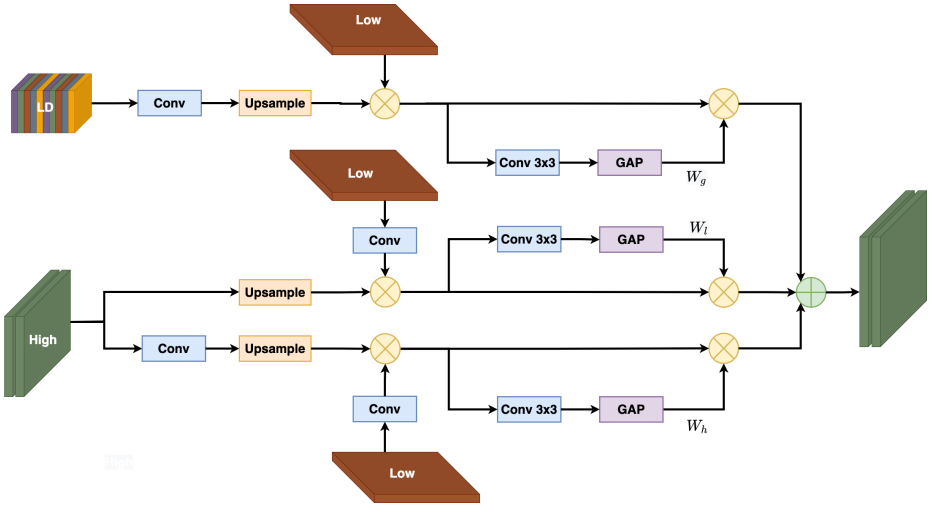


Fig. 2: Feature Aggregator Module (FAM)

average pooling layer with a kernel size of 31, forming a mask denoted as S'. The weight matrix is calculated as:

$$\mathcal{W} = 1 + 5 \times |S' - S|$$

The combined loss is formally defined as:

$$\mathcal{L}(S, G) = \mathcal{L}_{wIoU}(S, G) + \mathcal{L}_{wBCE}(S, G)$$

where S is the model's prediction mask, and G is the ground truth mask.

We also use deep supervision during training to improve robustness. Output from each decoder block (S_2 through S_5) is scaled to the original mask size and put through the loss function L. The final loss is the sum of all sub-losses:

$$\mathcal{L}_{total} = \sum_{i=2}^{5} \mathcal{L}\left(upsample(S_i), G\right)$$

4 Experiments and discussion

4.1 Benchmark datasets

We perform experiments with a wide range of benchmark datasets for medical image segmentation. All the datasets are described in Table **??**. We use five polyp segmentation datasets: the ETIS-Larib dataset [26], the CVC ClinicDB dataset [4], the CVC ColonDB dataset [27], the EndoScene-CVC300 dataset [28] and the Kvasir-SEG dataset [19]. Following the setup in PraNet [11], we use the training sets for Kvasir-SEG and CVC-ClinicDB as training data, while the

remaining images are used as test data. ISIC 2018 dataset [10] is a lesion segmentation dataset introduced by the International Skin Imaging Collaboration (ISIC). This is a dermoscopy dataset that is useful in the diagnosis of skin cancer and contains skin lesions and their corresponding annotations. It includes 2594 images: 1815 for training, 259 for validation, and 520 for testing. Finally, Kvasir Instrument dataset [17] contains 590 images of endoscopic tools and ground truth masks annotating each tool's position. Image resolution varies from 720x576 to 1280x1024.

4.2 Experiment settings

GCEENet is trained using the Adam optimizer. We set the initial learning rate to 10^{-4}, and use a hybrid warm-up and cosine annealing schedule. The learning rate increases linearly to 8.10^{-4} in the first eight epochs, then decreases along a cosine function in the remaining epochs. We train GCEENet with a batch size of 16 for 200 epochs during each experiment, and the results are the average of 5 training sessions. Data augmentations are applied to improve robustness. We perform augmentations online during training, with a probability of 0.7. The following augmentations are applied: Horizontal/vertical flip, random rotation, motion blur, change of brightness, contrast, saturation, random cropping. Images are resized to 3 different scales for training: 264×264, 352×352 and 440×440.

GCEENet is implemented in Python 3.8, using the PyTorch framework. Training is performed on a single machine running Ubuntu Linux 20.04, with an AMD Ryzen 3970X 3.7GHz CPU, 126GB of RAM, and an NVIDIA GeForce RTX3090 GPU.

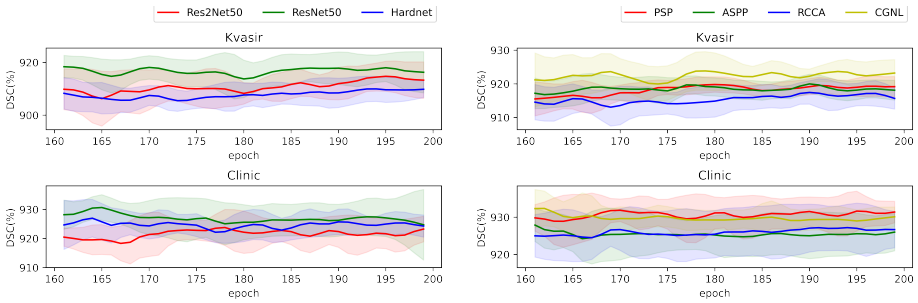
We evaluate models using the Dice coefficient (DSC), the mean Intersection over Union (mIoU), Precision, and Recall.

5 Results and discussion

5.1 Ablation study

We perform a series of ablation studies to determine the effectiveness of each individual module in GCEENet. For these experiments, we use the Kvasir and CVC-Clinic datasets for training and testing.

Effectiveness of encoder backbones We compare the effectiveness of three encoder backbones for GCEENet: ResNet50, Res2Net50, and HardNet68. The three backbones are tested in conjunction with the CGNL global context encoder Table 1 shows evaluation results for each of the three backbone. We also demonstrate each model's performance on the training set through 200 training epochs in Figure 3a. Overall, we find that ResNet50 achieves the most stable results on both datasets. As a result, we shall only incorporate ResNet50 as the backbone for later experiments.



(a) Backbone comparison (b) Global context encoder comparison

Fig. 3: Ablation study over 200 training epochs

Table 1: Comparison of GCEENet with the ResNet50, Res2Net50 and HarDNet68 backbones

Encoder backbone	No.params	Model size (MB)	Kvasir DSC	Clinic DSC
Res2Net50 [13]	68,550,468	261.50	90.85% ± 0.26%	91.98% ± 0.71%
ResNet50 [12]	68,711,788	262.11	91.42% ± 0.81%	92.24% $pm 1.43%
HarDNet68 [5]	50,261,440	191.73	91.00% ± 0.27%	92.42% ± 0.42%

Effectiveness of global context encoder In this experiment, we evaluate the effectiveness of four candidates for GCEENet's global context encoder: CGNL, CC, PSP, and ASPP. The four candidates with two approaches are: widening the perceptive field with convolutions (PSP and ASPP) and modeling correlations between regions and pixels using attention (CGNL and CC).

As seen in Figure 3b and Table 2, the CGNL module performs best on the Kvasir dataset but drops off slightly in the CVC-Clinic dataset. The other wide-receptive-field method, PSP, is much more stable on both datasets, having the best DSC for CVC-Clinic and the second best for Kvasir. Following these results, we shall perform the next ablation experiment with the two best-performing GCEs: PSP and CGNL.

Table 2: Comparison of GCEENet with different global context encoder modules

GCE module	No.params	Model size (MB)	Kvasir DSC	Clinic DSC
PSP [34]	55,728,452	212.59	91.92% ± 0.29%	93.16% ± 0.38%
ASPP [7]	64,183,364	244.84	91.84% ± 0.27%	92.56% ± 0.52%
RCCA [15]	63,999,684	244.14	91.67% ± 0.35%	92.67% ± 0.61%
CGNL [32]	46,653,764	177.97	92.26% ± 0.41%	92.94% ± 0.30%

At the end of our ablation study, we find that the optimal combination of components for GCEENet is the ResNet50 backbone, the Compact Generalized

Non-local global context encoder module, and the feature aggregation module (FAM). This is the version of GCEENet that will be used to compare with other existing models.

5.2 Comparison to baseline models

Polyp segmentation Table 3 shows performance metrics for our proposed GCEENet compared to 6 existing polyp segmentation models (U-Net, U-Net++, SFA, PraNet, HarDNet-MSEG and TransFuse) on 5 polyp datasets. Note that models are only trained on subsets of the Kvasir and CVC-ClinicDB datasets. GCEENet outperforms all existing models on the Kvasir, CVC-Colon, and ETIS-Larib datasets. Compared to the best baseline model - TransFuse-L - GCEENet achieves $\approx 3\%$ improvement on CVC-ColonDB (whose image types are not seen during training). This shows that GCEENet can generalize well to unseen data distributions.

Table 3: Quantitative results on the Kvasir, CVC-ColonDB and EndoScene

Method	**Kvasir**		**CVC-Clinic**		**CVC-Colon**		**ETIS**		**EndoScene**	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
UNet [25]	0.818	0.746	0.823	0.755	0.512	0.444	0.398	0.335	0.71	0.627
U-Net++ [35]	0.821	0.743	0.794	0.729	0.483	0.410	0.401	0.344	0.707	0.624
SFA [14]	0.723	0.661	0.700	0.607	0.469	0.374	0.297	0.217	0.467	0.329
PraNet [11]	0.898	0.840	0.899	0.849	0.709	0.640	0.628	0.567	0.871	0.797
HarDNet-MSEG [5]	0.912	0.857	0.932	0.882	0.731	0.660	0.677	0.613	0.887	0.821
TransFuse-S [33]	0.918	0.868	0.918	0.868	0.773	0.696	0.733	0.659	0.902	0.833
TransFuse-L [33]	0.918	0.868	**0.934**	0.886	0.744	0.676	0.737	0.661	**0.904**	**0.838**
GCEENet (Ours)	**0.923**	**0.871**	0.928	**0.882**	**0.776**	**0.703**	**0.727**	**0.652**	0.879	0.809

Kvasir Instrument Table 4 shows quantitative results obtained by GCEENet in Kvasir Instrument dataset. The model was trained five times in the same setting as in previous experiments. The result is averaged over five last epochs in each training time. GCEENet outperforms other methods on all the ratings with a dice score of 96% (increase 5.5% compared with DoubleUnet).

Table 4: Quantitative results on the Kvasir-Instrument dataset

Method	**mIoU**	**mDice**	**F2-score**	**Precision**	**Recall**	**Overall Acc**
U-Net [25]	0.858	0.916	0.932	0.899	0.949	0.986
DoubleUNet [18]	0.843	0.904	0.915	0.897	0.928	0.984
GCEENet (Ours)	**0.929**	**0.960**	**0.960**	**0.963**	**0.964**	**0.993**

Skin lesion segmentation Table 5 shows quantitative results obtained by GCEENet and five previous models for the ISIC 2018 skin lesion segmentation dataset. The model was trained five times in the same setting as in previous experiments. The result is averaged over five last epochs in each training time. GCEENet achieves state-of-the-art performance on the mDice, sensitivity, and accuracy metrics. Notably, our model outperforms the second-best MCGU-Net by 6% in the sensitivity metric.

Table 5: Quantitative results on the ISIC 2018 dataset

Method	mDice	Sensitivity	Specificity	Accuracy	Precision
U-Net [25]	0.647	0.708	0.964	0.89	0.779
Attention U-net [24]	0.665	0.717	0.967	0.897	0.787
R2U-net [1]	0.679	0.792	0.928	0.88	0.741
Attention R2U-Net	0.691	0.726	0.971	0.904	0.822
BCDU-Net (d=1) [3]	0.847	0.783	0.98	0.936	0.922
BCDU-Net (d=3) [3]	0.851	0.785	0.982	0.937	0.928
MCGU-Net (d=1) [2]	0.889	0.845	0.984	0.952	0.938
MCGU-Net (d=3) [2]	0.895	0.848	0.986	0.955	0.947
GCEENet (Ours)	**0.898**	**0.908**	0.970	**0.962**	0.917

6 Conclusion

This paper has proposed GCEENet, a convolutional neural network designed for the semantic segmentation problem in medical images. We show that the combination of global context encoder modules and feature aggregation alongside a strong backbone network can help enhance and exploit global context information. Our experiments show that GCEENet outperforms existing state-of-the-art models on several benchmark datasets.

7 Acknowledgment

This work was funded by Vingroup Innovation Foundation (VINIF) under project code VINIF.2020.DA17.

References

1. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation (2018)
2. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation (2020)

3. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional ConvLSTM U-net with densley connected convolutions pp. 406–415 (2019)
4. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, 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)
5. Chao, P., Kao, C.Y., Ruan, Y.S., Huang, C.H., Lin, Y.L.: Hardnet: A low memory traffic network. In: ICCV. pp. 3552–3561 (2019)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI **40**(4), 834–848 (2017)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI. vol. 34, pp. 10599–10606 (2020)
10. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., 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). pp. 168–172. IEEE (2018)
11. 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: MICCAI. pp. 263–273. Springer (2020)
12. Gao, S., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.H.: Res2net: A new multi-scale backbone architecture. PAMI (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Huang, S.C., Hoang, Q.V., Le, T.H.: Sfa-net: A selective features absorption network for object detection in rainy weather conditions. IEEE Transactions on Neural Networks and Learning Systems (2022)
15. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: ICCV. pp. 603–612 (2019)
16. Hung, N.B., Duc, N.T., Van Chien, T., Sang, D.V.: Ag-resunet++: An improved encoder-decoder based method for polyp segmentation in colonoscopy images. In: 2021 RIVF International Conference on Computing and Communication Technologies (RIVF). pp. 1–6. IEEE (2021)
17. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., Johansen, D., Halvorsen, P.: Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: MultiMedia Modeling. pp. 218–229. Springer International Publishing, Cham (2021)
18. Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D.: Doubleu-net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS). pp. 558–564. IEEE (2020)

19. 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: ICMM. pp. 451–462. Springer (2020)
20. Li, X., Zhang, L., You, A., Yang, M., Yang, K., Tong, Y.: Global aggregation then local distribution in fully convolutional networks. arXiv preprint arXiv:1909.07229 (2019)
21. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
23. Ngoc Lan, P., An, N.S., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V.: Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In: International Symposium on Visual Computing. pp. 15–28. Springer (2021)
24. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas (2018)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
26. 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)
27. 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)
28. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017** (2017)
29. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
30. Wei, J., Wang, S., Huang, Q.: F^3net: Fusion, feedback and focus for salient object detection. In: AAAI. vol. 34, pp. 12321–12328 (2020)
31. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
32. Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., Xu, F.: Compact generalized non-local network. arXiv preprint arXiv:1810.13125 (2018)
33. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: MICCAI. pp. 14–24. Springer (2021)
34. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
35. Zhou, Z., Rahman Siddiquee, M.M., 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)