Contextual Attention Network: Transformer Meets U-Net

Reza Azad[1], Moein Heidari[2], Yuli Wu[1], and Dorit Merhof[1,3]

[1] Institute of Imaging and Computer Vision, RWTH Aachen University, Germany
[2] School of Electrical Engineering, Iran University of Science and Technology, Iran,
moein_heidari@elec.iust.ac.ir
[3] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
{azad, yuli.wu,dorit.merhof}@lfb.rwth-aachen.de

Abstract. Currently, convolutional neural networks (CNN) (e.g., U-Net) have become the de facto standard and attained immense success in medical image segmentation. However, as a downside, CNN based methods are a double-edged sword as they fail to build long-range dependencies and global context connections due to the limited receptive field that stems from the intrinsic characteristics of the convolution operation. Hence, recent articles have exploited Transformer variants for medical image segmentation tasks which open up great opportunities due to their innate capability of capturing long-range correlations through the attention mechanism. Although being feasibly designed, most of the cohort studies incur prohibitive performance in capturing local information, thereby resulting in less lucidness of boundary areas. In this paper, we propose a contextual attention network to tackle the aforementioned limitations. The proposed method uses the strength of the Transformer module to model the long-range contextual dependency. Simultaneously, it utilizes the CNN encoder to capture local semantic information. In addition, an object-level representation is included to model the regional interaction map. The extracted hierarchical features are then fed to the contextual attention module to adaptively recalibrate the representation space using the local information. Then, they emphasize the informative regions while taking into account the long-range contextual dependency derived by the Transformer module. We validate our method on several large-scale public medical image segmentation datasets and achieve state-of-the-art performance. We have provided the implementation code in github.

Keywords: Transformer · semantic segmentation · attention · medical image.

1 Introduction

Convolutional neural networks (CNNs), and more specifically, fully convolutional networks, have shown prominence in the majority of medical image segmentation applications. As a variant of these architectures, U-Net [25] has rendered notable performance and has been extensively utilized across a wide range of medical domains [28,20,7,15,4]. In spite of their superb performance, CNN-based approaches suffer from a limitation in modeling the long-range semantic dependencies due to a confined receptive field size (even with dilated/atrous sampling [11]) and due to the nature of the convolution layer. Therefore, such a deficiency in capturing multi-scale information yields a performance degradation in the segmentation of complex structures, with variation in shapes and scales. Different methods have been proposed to solve the problem of the restricted receptive field of regular CNNs in recent years [11,22]. Wang et al. [29] extended the self-attention concept into the spatial domain to model non-local properties of images by devising a non-local module that can be easily integrated into existing network designs. As a result of the prominent performance of the attention mechanism, a line of research has studied bridging the gap between the attention mechanism and CNNs in medical image segmentation [26,8,3]. To overcome the aforementioned limitation of CNNs, recently proposed Transformer-based architectures that leverage the self-attention mechanism to construct the contextual representations have been utilized. Transformers, unlike regular CNNs, are not only capable of modeling global contexts but also of leveraging to the versatile local information. Inspired by this, numerous studies have attempted to adapt Transformers for various image recognition tasks [14,9]. Moreover, Transformer-based models have recently gained growing attention ahead of their CNN counterpart in medical image segmentation [27,10]. As an example of an alternative perspective by treating semantic segmentation using Transformers, [31] proposed to model semantic segmentation as a sequence-to-sequence prediction task. Although the network is well designed to model the global contextual representation, it pays less attention to the local information and is, consequently, less precise in the boundary area. Additionally, case studies have been established to investigate the amalgamation of Transformers and U-Net in medical image segmentation [10,17]. Valanarasu et al. [27] introduce a Local-Global training methodology for Transformers to learn both global and local features, respectively. However, this approach fails to model object-level interaction and renders a poor performance in the case of overlapping objects of interests.

What all these methods have in common is their limitation in designing a specific mechanism to adaptively combine the global and local contextual representations. Particularly, a mechanism to jointly model the local semantic CNN representation along with the global contextual features derived from the Transformer module is critical for a task-specific purpose. To address this limitation, we propose the contextual attention network. Our design offers a two-stream pipeline, where in the first stream we utilize a CNN module to extract local semantic information and the object-level interaction map, while the second path incorporates the Transformer module to capture long-range contextual representations. Our Transformer module produces an image-level contextual representation *(ICR)* to construct the spatial dependency map in the image level and it produces regional importance coefficients *(RIC)* to model the importance of each region. In contrast to [27] which simply concatenates the local and global features, our method utilizes a contextual attention module to adaptively scale the feature maps and emphasizes on the important regions. Through extensive experiments, our empirical findings validate that our method is able to pay more attention to the overlapped boundary area while providing a strong semantic segmentation map. Our main contributions are summarized as follows:

- A contextual attention mechanism to adaptively aggregate pixel, object and image level features
- Coupling Transformer module with the CNN encoder to model object-level interaction
- State-of-the-art results on the public datasets along with a publicly-available implementation source code

2 Proposed Method

The proposed network architecture is depicted in Figure 1. Our proposed architecture offers an end-to-end training strategy to adaptively incorporate the global contextual representation into local representative features derived from the CNN module. Our design proposes a contextual attention mechanism for feature recalibration and boundary-aware semantic segmentation. We will discuss each part in the following subsections.

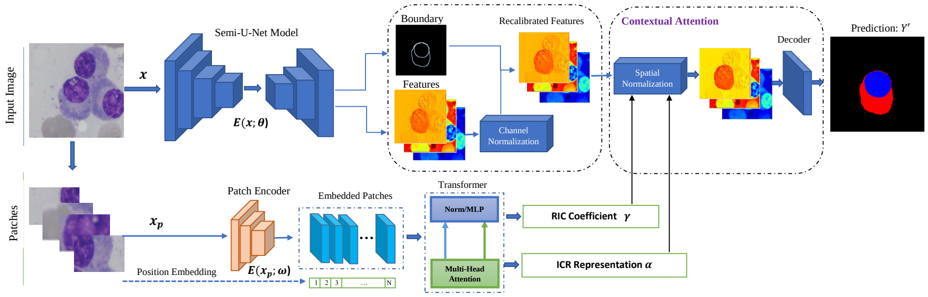

Fig. 1. Illustration of the proposed approach for medical image segmentation with the incorporation of the global contextual representation into local representative features.

2.1 CNN Representation

As illustrated in Figure 1, our proposed method consists of two encoding streams, where in the first path we deploy the semi U-Net structure [25] to extract CNN representation. Given an image $\mathbf{x} \in \mathcal{R}^{H\times W\times C}$ with spatial dimension H and W, and C channels, our CNN encoder E_θ applies a series of convolutional blocks to model the pixel-level contextual representations. The locality nature of the convolutional operation usually limits the strength in modelling the object-level interaction. To include such representations, we model the object-level interaction by learning the boundary heatmap:

$$B = \sigma(Conv_b(E(x;\theta))), B \in \mathcal{R}^{H\times W\times 1} \tag{1}$$

where σ shows the sigmoid activation and the $Conv_b(.)$ shows a 1×1 kernel convolutional operation. This additional head enables the regional interaction and provides a surrogate signal for modelling the regional contextual dependency.

2.2 Long-range Contextual Representation

To learn the long-range contextual dependency we include the Transformer module in the bottom stream of our proposed pipeline. To prepare the input for the Transformer module, we divide the input image $\mathbf{x} \in \mathcal{R}^{H\times W\times C}$ into flattened uniform non-overlapping patches $\mathbf{x}_p \in \mathcal{R}^{N\times\left(p^2\cdot C\right)}$, where $p\times p$ denotes the dimension of each patch and $N = [\frac{HW}{p^2}]$ is the length of image sequence. Afterwards, using a patch encoder $E(x_p;\omega)$, we project the patches into a K dimensional embedding space. In order to maintain the spatial information of each patch, we learn a 1-D positional embedding $I_{\text{pos}} \in \mathcal{R}^{N\times K}$ which is subsequently added to the patch-embedding to preserve positional information $t_0 = \left[x_p^1 I; x_p^2 I; \cdots; x_p^N I\right] + I_{pos}$, where $I \in \mathcal{R}^{\left(p^2\cdot C\right)\times K}$ designates the projected patch embedding. We then exploit a stack of Transformer blocks encompassing a multi-head self-attention (MSA) and a multilayer perceptron (MLP) to learn the long-range contextual representation. An MSA layer is composed of M parallel self-attention heads to scale the embedded patches: $t_i' = \operatorname{MSA}\left(\operatorname{Norm}\left(\mathbf{t}_{i-1}\right)\right) + \mathbf{t}_{i-1}, \quad i = 1\ldots L$. Next, the MLP modules learn the long-range contextual dependency by: $t_i = \operatorname{MLP}\left(\operatorname{Norm}\left(t_i'\right)\right) + t_i', \quad i = 1\ldots L$, where Norm() denotes layer normalization [5], and $t_i \in \mathcal{R}^{\frac{HW}{p^2}\times d}$ shows the encoded semantic representation in d dimensional space. In addition to the encoded features, we model the image-level contextual representation *(ICR)* by reshaping *(Re)* the feature and applying a 1×1 convolutional operation *($Conv_I$)* :

$$ICR = \sigma(Conv_I(Re(t_L))), ICR \in \mathcal{R}^{H\times W\times 1} \tag{2}$$

We use ICR to construct the spatial dependency map in the image level to later normalize the feature set generated by the CNN module. We further define the region importance coefficients *(RIC)* to model the distribution of foreground pixels in each region. The objective of the RIC coefficient is to provide a supervisory signal to guide the contextual attention module in determining the important regions (Eq. 3). $Conv_R$ shows a 1×1 convolutional operation.

$$RIC = \sigma(Conv_R(t_L)),\ RIC \in \mathcal{R}^{\frac{HW}{p^2}\times 1} \tag{3}$$

2.3 Contextual Attention Module

To adaptively aggregate the extracted features, we propose the contextual attention module. The importance of each feature set should be in line with the task at hand, thus, our proposed module utilizes the following two-level normalization steps: First, it recalibrates the CNN representation for a pixel-level object understanding, then it performs a spatial normalization to selectively emphasize the long-range contextual dependencies inside the feature set. Following the squeeze and excitation [19], we define the channel-wise normalization weights (w_{ch}) as:

$$w_{ch} = \sigma\left(\mathbf{W}_2 \delta\left(\mathbf{W}_1 GAP(f)\right)\right) \tag{4}$$

where GAP shows the global average pooling operation applied to the CNN features (f), **W1** and **W2** are the learning parameters, and δ and σ are the Sigmoid and ReLU activation functions. We form the normalized features by: $f' = w_{ch} \cdot f$. To emphasize the boundary area, we add the boundary representation to the normalized feature, $\tilde{f} = f' + B$. The objective of the boundary feature is to emphasize the boundary regions and guide the model to precisely separate the overlapping objects (e.g. object-level interaction). Next, using the feature set derived from the Transformer module, we perform the spatial normalization. To this end, first we multiply the RIC coefficient with the corresponding regions in $\tilde{f}$ to scale the representation based on the regional importance, $f_{sn} = RIC \cdot \tilde{f}$. To further incorporate the long-range dependency, we concatenate the ICR representation with the f_{sn} and then apply the convolutional kernel followed by the batch normalization (BN) and activation function to perform a non-linear aggregation:

$$\tilde{f_{sn}} = \delta(BN(Conv(ICR, f_{sn})) \tag{5}$$

The $Conv$ is the 1×1 convolutional operation. The resulting feature set contains both local semantic and global contextual representations which are selectively combined to perform the semantic segmentation task. Subsequently, we apply the decoder block to the extracted features to predict the segmentation mask, $Y' = D(\tilde{f_{sn}}; \gamma)$. The joint objective loss function that we optimize during the training is as follows:

$$\mathcal{L}_{\text{joint}} = \lambda_1 \mathcal{L}_{\text{segmentation}} + \lambda_2 \mathcal{L}_{\text{boundary}} + \lambda_3 \mathcal{L}_{\text{RIC}} \tag{6}$$

where $\mathcal{L}_{\text{segmentation}}$ calculates the Cross-entropy loss between the predicted mask and the ground truth, $\mathcal{L}_{\text{boundary}}$ shows the binary cross-entropy loss for the boundary prediction, and $\mathcal{L}_{\text{RIC}}$ calculates the MSE loss between the distribution of foreground pixels in each image patch and the corresponding predicted one. We use coefficients $\lambda_i, i \in \{1, 2, 3\}$ to weight each loss.

3 Experiments

3.1 Dataset

Skin Lesion Segmentation: Automatic skin lesion segmentation is one of the most demanding tasks in medical image analysis for accurate diagnosis and treatment. In this respect, we focus on three challenge benchmarks: ISIC 2017

[13], ISIC 2018 [12] and PH2 [23]. Following the literature work [1,2] for each series, we divide the dataset into train, validation, and test sets accordingly. We use the same setting for a fair evaluation and downsize the original images from the resolution of 576×767 pixels to 256×256 pixels in the pre-processing step.
Multiple Myeloma Segmentation: The proposed method is also evaluated on multiple myeloma cell segmentation grand challenges [16], which are provided by the SegPC 2021 (Segmentation of Multiple Myeloma Plasma Cells in Microscopic Images). Images in this dataset were captured from bone marrow aspirate slides of patients diagnosed with Multiple Myeloma (MM), a type of white blood cell cancer. Using the pipeline from [6], we follow the same strategy as [3] and split the original training dataset (290 images) into a training and validation set, and evaluate our method on the original validation set as our new test set.

3.2 Experimental Set-up

Network Details and Training Process: As depicted in Figure 1, our model uses both U-Net and Transformer modules to semantically label the input image. For the U-Net model, we use the Resent encoder [18] pre-trained on ImageNet and a four-blocks decoder module to generate the segmentation mask. Simultaneously, the Transformer structure follows the common implementation of the Vision Transformer with M (experimentally 4) heads. The implementation is performed in PyTorch and the results are carried out on a single GPU system with Nvidia RTX 3090. The model is trained end-to-end employing the Adam optimizer, batch size 4 and a learning rate 10^{-4} for 100 epochs.
Evaluation Protocol: Our evaluation takes into account the evaluation metrics used in the respective challenges, which comprises several well-known segmentation metrics, including sensitivity $= \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, specificity $= \frac{\mathrm{TN}}{\mathrm{TN+FP}}$, accuracy $= \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$, $mIOU = \frac{TP}{TP+FP+FN}$, and $DSC = \frac{2TP}{2TP+FP+FN}$ scores, where TP indicates true samples which are correctly classified as true, TN stands for the correct classification of the negative samples, FP and FN show the wrongly classified samples respectively.

3.3 Results

Quantitative results: We provide qualitative comparisons on the benchmarks introduced in Section 3.1. Starting from the skin lesion segmentation scenario, Table 1 reports the comparison results on the three datasets, namely, ISIC 2017, ISIC 2018 and PH2. We exploited different evaluation metrics to accomplish a general and fair comparison. The baseline approach is simply the U-Net method without any of the proposed modules. Overall, our method attains a superior global performance across all datasets and evaluation metrics and most of the outperforming margins are statistically significant. We also observed that our method outperforms the Transformer [30,10,27] counterparts in almost all skin lesion segmentation benchmarks, which further proves the effectiveness of our design compared to other Transformer-based models. Note that, in contrast to [30],

Table 1. Performance comparison of the proposed method vs. state-of-the-art methods on skin lesion segmentation benchmarks.

Articles	**ISIC 2017**				**ISIC 2018**				**PH2**			
	DSC	**SE**	**SP**	**ACC**	**DSC**	**SE**	**SP**	**ACC**	**DSC**	**SE**	**SP**	**ACC**
U-Net [25]	0.8159	0.8172	0.9680	0.9164	0.8545	0.8800	0.9697	0.9404	0.8936	0.9125	0.9588	0.9233
Att U-Net [24]	0.8082	0.7998	0.9776	0.9145	0.8566	0.8674	0.9863	0.9376	0.9003	0.9205	0.9640	0.9276
DAGAN [21]	0.8425	0.8363	0.9716	0.9304	0.8807	0.9072	0.9588	0.9324	0.9201	0.8320	0.9640	0.9425
TransUNet [10]	0.8123	0.8263	0.9577	0.9207	0.8499	0.8578	0.9653	0.9452	0.8840	0.9063	0.9427	0.9200
MCGU-Net [1]	0.8927	0.8502	**0.9855**	0.9570	0.895	0.848	**0.986**	0.955	0.9263	0.8322	0.9714	0.9537
MedT [27]	0.8037	0.8064	0.9546	0.9090	0.8389	0.8252	0.9637	0.9358	0.9122	0.8472	0.9657	0.9416
FAT-Net [30]	0.8500	0.8392	0.9725	0.9326	0.8903	**0.9100**	0.9699	0.9578	**0.9440**	**0.9441**	0.9741	**0.9703**
Proposed	**0.9164**	**0.9128**	0.9789	**0.9660**	**0.9059**	0.9038	0.9746	**0.9603**	0.9414	0.9395	**0.9756**	0.9647

we attained eminent performance without exploiting any augmentation strategy. Additionally, Table 2 lists the quantitative results of different alternative methods and the proposed network on the SegPC dataset. The results demonstrate that the proposed network attains better results than the other approaches by achieving significant performance gains over the baseline. It is worth mentioning that the SegPC dataset contains samples with high overlaps and the effectiveness of our method is statistically significant in comparison to the SOTA approaches.

Table 2. Performance evaluation on the SegPC challenge (best result is highlighted).

Methods	**mIOU**
Frequency recalibration U-Net [3]	0.9392
XLAB Insights [6]	0.9360
DSC-IITISM [6]	0.9356
Multi-scale attention deeplabv3+ [6]	0.9065
U-Net [25]	0.7665
Baseline	0.9172
Proposed	**0.9395**

Qualitative results: Visual segmentation results on both tasks are illustrated in Figure 2. Clearly, our proposed model produces smooth segmentation results for both tasks and performs well in the boundary area for separating the object of interest from the background. This fact reveals the importance of both Transformer modules for long-range contextual dependency (encouraging object learning) and the combination of boundary modules with the convolutional features maps in precise boundary recovery. Specifically, for the SegPC dataset, we observed that the proposed method segments the myeloma instances from a highly overlapped background with high precision. During our experimental visualization, we also observed that, compared to the U-Net model, the proposed structure produces robust segmentation results even with a noisy annotation, which is a common scenario in the medical domain. Comparable to recent work [27,10], our results also suggest that coupling the Transformer module with the CNN segmentation model can provide an additional input signal for a reliable and robust segmentation architecture.

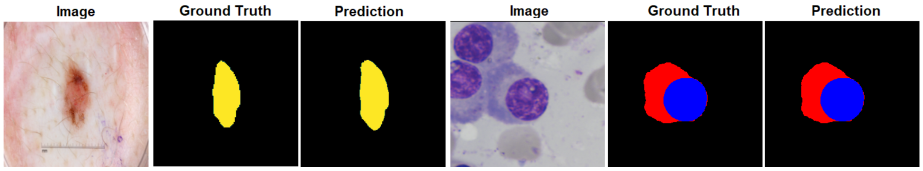

Fig. 2. Prediction results of the proposed method on both skin lesion segmentation and multiple Myeloma instance segmentation.

3.4 Ablation Study

The suggested network comprises the Transformer, the boundary and the attention add-on modules that are included for learning both local and global feature sets. To experimentally evaluate the effect and contribution of each module in the generalization performance, we selectively remove any of the modules, as shown in Table 3. The qualitative finding suggests that removing any of the modules from the architecture results in a performance loss. More precisely, it can be observed that removing the Transformer module largely decreases the importance of the global attention maps and results in a clear performance drop. However, by including the Transformer module the network takes into account the strength of the global attention mechanism in combination with the local effectiveness of the convolution feature maps in learning generic and rich feature sets for precise localization ability. It is also worthwhile to mention that the combination of the proposed modules decreases the number of wrong predictions and the number of isolated FP (cf. Table 1) due to the strength of long-range contextual dependency maps.

Table 3. Effect of eliminating each module on the overall performance of the proposed method. We report the results on ISIC 2018 dataset.

Module			
Boundary	Transformer	Contextual Attention	DSC
(×)	(√)	(√)	0.905
(√)	(×)	(√)	0.896
(√)	(√)	(×)	0.901
(√)	(√)	(√)	**0.906**

4 Conclusion

In this work, we invoke the inherent frailty of regular convolutional neural networks in capturing long-range contextual dependencies. Specifically, the presented methodology is a novel contextual attention network which exploits the

Transformer module along with the CNN encoder in order to concomitantly combine local and global representations for a further performance boost. The results presented in this paper validate that our proposal achieves substantial improvement over many architectures in semantic segmentation tasks.

References

1. Asadi-Aghbolaghi, M., Azad, R., Fathy, M., Escalera, S.: Multi-level context gating of embedded collective knowledge for medical image segmentation. arXiv preprint arXiv:2003.05056 (2020)
2. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densely connected convolutions. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 406–415 (2019). https://doi.org/10.1109/ICCVW.2019.00052
3. Azad, R., Bozorgpour, A., Asadi-Aghbolaghi, M., Merhof, D., Escalera, S.: Deep frequency re-calibration u-net for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3274–3283 (2021)
4. Azad, R., Khosravi, N., Merhof, D.: Smu-net: Style matching u-net for brain tumor segmentation with missing modalities (2021)
5. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
6. Bozorgpour, A., Azad, R., Showkatian, E., Sulaiman, A.: Multi-scale regional attention deeplab3+: Multiple myeloma plasma cells segmentation in microscopic images. arXiv preprint arXiv:2105.06238 (2021)
7. Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y., Chen, G.: Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. Quantitative imaging in medicine and surgery **10**(6), 1275 (2020)
8. Cai, Y., Wang, Y.: Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. arXiv preprint arXiv:2012.10952 (2020)
9. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366 (2021)
10. 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)
11. 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. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
12. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., 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)
13. 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)
14. 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)
15. Feyjie, A.R., Azad, R., Pedersoli, M., Kauffman, C., Ayed, I.B., Dolz, J.: Semi-supervised few-shot learning for medical image segmentation. arXiv preprint arXiv:2003.08462 (2020)
16. Gupta, A., Mallick, P., Sharma, O., Gupta, R., Duggal, R.: Pcseg: Color model driven probabilistic multiphase level set based tool for plasma cell segmentation in multiple myeloma. PloS one **13**(12), e0207908 (2018)
17. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. 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)
20. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059. IEEE (2020)
21. Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., Wang, S.: Skin lesion segmentation via generative adversarial networks with dual discriminators. Medical Image Analysis **64**, 101716 (2020)
22. Li, M., Lian, F., Wang, C., Guo, S.: Accurate pancreas segmentation using multi-level pyramidal pooling residual u-net with adversarial mechanism. BMC Medical Imaging **21**(1), 1–8 (2021)
23. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: Ph 2-a dermoscopic image database for research and benchmarking. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 5437–5440. IEEE (2013)
24. 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)
25. 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)
26. Sinha, A., Dolz, J.: Multi-scale self-guided attention for medical image segmentation. IEEE journal of biomedical and health informatics **25**(1), 121–130 (2020)
27. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2021)
28. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 363–373. Springer (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. pp. 7794–7803 (2018)
30. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Medical Image Analysis **76**, 102327 (2022)
31. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)

5 Appendix

In this part, we intend to provide some additional details regarding our approach, which allow a deeper understanding into our experiments.

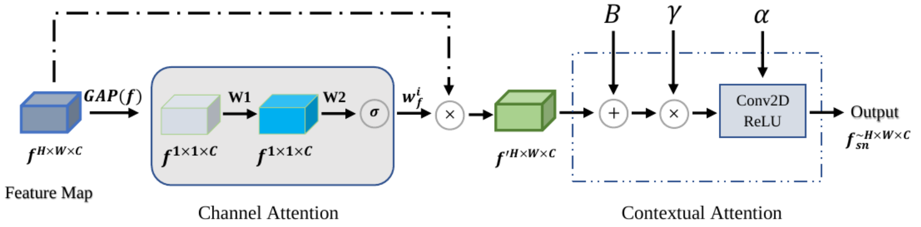

Fig. 3. Perceptual visualization of the proposed Contextual Attention module. The proposed structure applies a channel-wise normalization along with the boundary (B) representation to recalibrate the feature space and then uses RIC(γ) and ICR(α) features to incorporate the spatial contextual information inside the feature set.

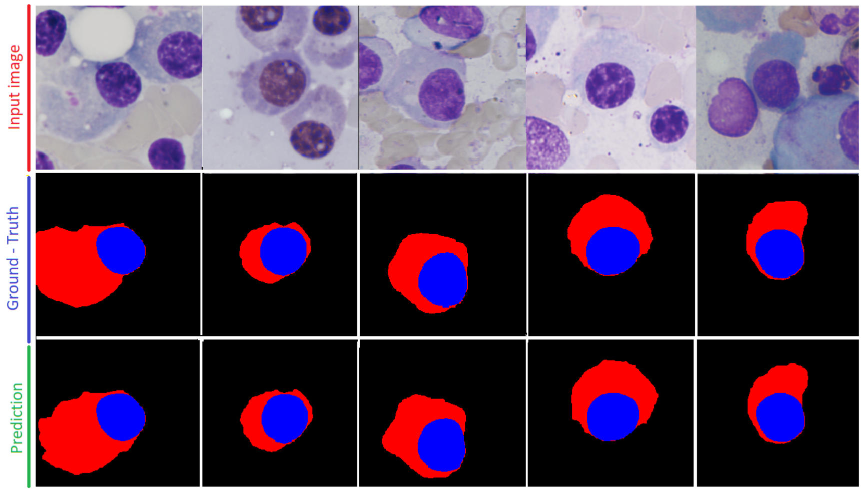

Fig. 4. More results of the proposed method for multiple Mylomia segmentation on the SegPC2021 dataset. The first row shows the input image, the second row indicates the ground truth for each image and the third row shows the prediction of the network.

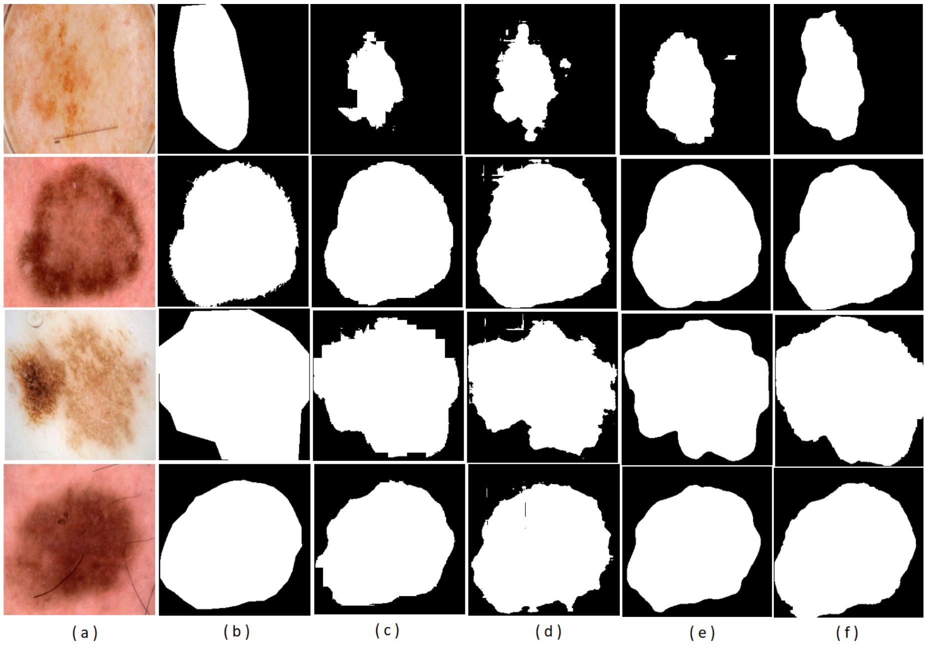

Fig. 5. Visual comparisons of different methods for skin lesion segmentation task. (a) Input images. (b) Ground truth. (c) U-Net [25]. (d) Gated Axial-Attention paper [27]. (e) Proposed method without a contextual attention module and (f) Proposed method.