UNeXt: MLP-based Rapid Medical Image Segmentation Network

Jeya Maria Jose Valanarasu and Vishal M. Patel

Johns Hopkins University

Abstract. UNet and its latest extensions like TransUNet have been the leading medical image segmentation methods in recent years. However, these networks cannot be effectively adopted for rapid image segmentation in point-of-care applications as they are parameter-heavy, computationally complex and slow to use. To this end, we propose UNeXt which is a Convolutional multilayer perceptron (MLP) based network for image segmentation. We design UNeXt in an effective way with an early convolutional stage and a MLP stage in the latent stage. We propose a tokenized MLP block where we efficiently tokenize and project the convolutional features and use MLPs to model the representation. To further boost the performance, we propose shifting the channels of the inputs while feeding in to MLPs so as to focus on learning local dependencies. Using tokenized MLPs in latent space reduces the number of parameters and computational complexity while being able to result in a better representation to help segmentation. The network also consists of skip connections between various levels of encoder and decoder. We test UNeXt on multiple medical image segmentation datasets and show that we reduce the number of parameters by **72x**, decrease the computational complexity by **68x**, and improve the inference speed by **10x** while also obtaining better segmentation performance over the state-of-the-art medical image segmentation architectures. Code is available at https://github.com/jeya-maria-jose/UNeXt-pytorch

Keywords: Medical Image Segmentation. MLP. Point-of-Care.

1 Introduction

Medical imaging solutions have played a pivotal role for diagnosis and treatment in the healthcare sector. One major task in medical imaging applications is segmentation as it is essential for computer-aided diagnosis and image-guided surgery systems. Over the past decade, many works in the literature have focused on developing efficient and robust segmentation methods. UNet [17] is a landmark work which showed how efficient an encoder-decoder convolutional network with skip connections can be for medical image segmentation. UNet has became the backbone of almost all the leading methods for medical image segmentation in recent years. Following UNet, a number of key extensions like UNet++ [29], UNet3+ [13], 3D UNet [7], V-Net [16], Y-Net [15] and KiUNet

[21,22] have been proposed. Recently, many transformer-based networks have been proposed for medical image segmentation as they learn a global under-standing of images which can be helpful in segmentation. TransUNet [6] modi-fies the ViT architecture [10] into an UNet for 2D medical image segmentation. Other transformer-based networks like MedT [20], TransBTS [25], and UNETR [11] have also been proposed for medical image segmentation. Note that almost all the above works have focused on improving the performance of the network but do not focus much on the computational complexity, inference time or the number of parameters, which are essential in many real-world applications. As most of these are used for analysis in laboratory settings, they are tested using machines with high compute power (like GPUs). This helps accelerate the speed of inference and also help accommodate a large number of parameters.

In recent times, there has been a translation of medical imaging solutions from laboratory to bed-side settings. This is termed as point-of-care imaging as the testing and analysis is done by the side of the patient. Point-of-care imaging [23] helps clinicians with expanded service options and improved patient care. It helps in reducing the time and procedures involved in patients having to go visit radiology centers. Technology improvements around point-of-care imaging are leading to greater patient satisfaction. The use of point-of-care devices has been increasing in recent years. For example, point-of-care ultrasound (POCUS) de-vices [1] have shown to be useful to quickly check pleural irregularities in lungs, cardiac hemodynamic flow and automatic bladder volume calculation. Phone-camera based images are also being used to detect and diagnose skin conditions [2]. Magnetic resonance imaging (MRI) machines have also been developed for bed-side operation and fast analysis [3]. These recent diagnostic developments have helped in clear and fast acquisition of medical images at point-of-care as seen in Fig. 1. Tasks like segmentation, classification and registration are also being integrated along with these appliances to help patients and clinicians ac-celerate the diagnosis process. The major deep-learning based solutions for these tasks (like UNet and TransUNet) come with an inherent computation overhead
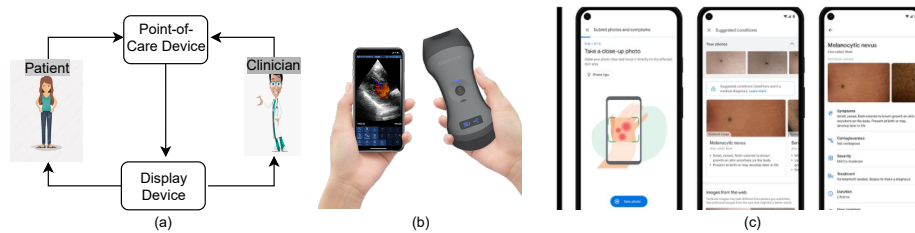

Fig. 1. Motivation for UNeXt: As medical imaging solutions become more applicable at point-of-care, it is important to focus on making the deep networks light-weight and fast while also being efficient. (a) Point-of-Care medical intervention workflow. (b) Recent medical imaging developments: POCUS device [1] and (c) Phone-based skin lesion detection and identification application [2].

and a large number of parameters making them difficult to use in point-of-care applications. In this work, we focus on solving this problem and design an efficient network that has less computational overhead, low number of parameters, a faster inference time while also maintaining a good performance. Designing such a network is essential to suit the shifting trends of medical imaging from laboratory to bed-side. To this end, we propose UNeXt which is designed using convolutional networks and (multilayer perceptron) MLPs.

Recently, MLP-based networks [27,19,14,18] have also been found to be competent in computer vision tasks. Especially MLP-Mixer [18], an all-MLP based network which gives comparable performance with respect to transformers with less computations. Inspired by these works, we propose UNeXt which is a convolutional and MLP-based network. We still follow a 5-layer deep encoder-decoder architecture of UNet with skip connections but change the design of each block. We have two stages in UNeXt- a convolutional stage followed by an MLP stage. We use convolutional blocks with less number of filters in the initial and final blocks of the network. In the bottleneck, we use a novel Tokenized MLP (Tok-MLP) block which is effective at maintaining less computation while also being able to model a good representation. Tokenized MLP projects the convolutional features into an abstract token and then uses MLPs to learn meaningful information for segmentation. We also introduce shifting operation in the MLPs to extract local information corresponding to different axial shifts. As the tokenized features are of the less dimensions and MLPs are less complicated than convolution or self-attention and transformers; we are able to reduce the number of parameters and computational complexity significantly while also maintaining a good performance. We evaluate UNeXt on ISIC skin lesion dataset [8] and Breast UltraSound Images (BUSI) dataset [4] and show that it obtains better performance than recent generic segmentation architectures. More importantly, we reduce the number of parameters by **72x**, decrease the computational complexity by **68x** and increase the inference speed by **10x** when compared to TransUNet making it suitable for point-of-care medical imaging applications.

In summary, this paper makes the following contributions: 1) We propose UNeXt, the first convolutional MLP-based network for image segmentation. 2) We propose a novel tokenized MLP block with axial shifts to efficiently learn a good representation at the latent space. 3) We successfully improve the performance on medical image segmentation tasks while having less parameters, high inference speed, and low computational complexity.

2   UNeXt

Network Design: UNeXt is an encoder-decoder architecture with two stages: 1) Convolutional stage, and a 2) Tokenized MLP stage. The input image is passed through the encoder where the first 3 blocks are convolutional and the next 2 are Tokenized MLP blocks. The decoder has 2 Tokenized MLP blocks followed by 3 convolutional blocks. Each encoder block reduces the feature resolution by 2 and each decoder block increases the feature resolution by 2. Skip connections are

also included between the encoder and decoder. The number of channels across each block is a hyperparameter denoted as $C1$ to $C5$. For the experiments using UNeXt architecture, we follow $C1 = 32$, $C2 = 64$, $C3 = 128$, $C4 = 160$, and $C5 = 256$ unless stated otherwise. Note that these numbers are actually less than the number of filters of UNet and its variants contributing to the first change to reduce parameters and computation.
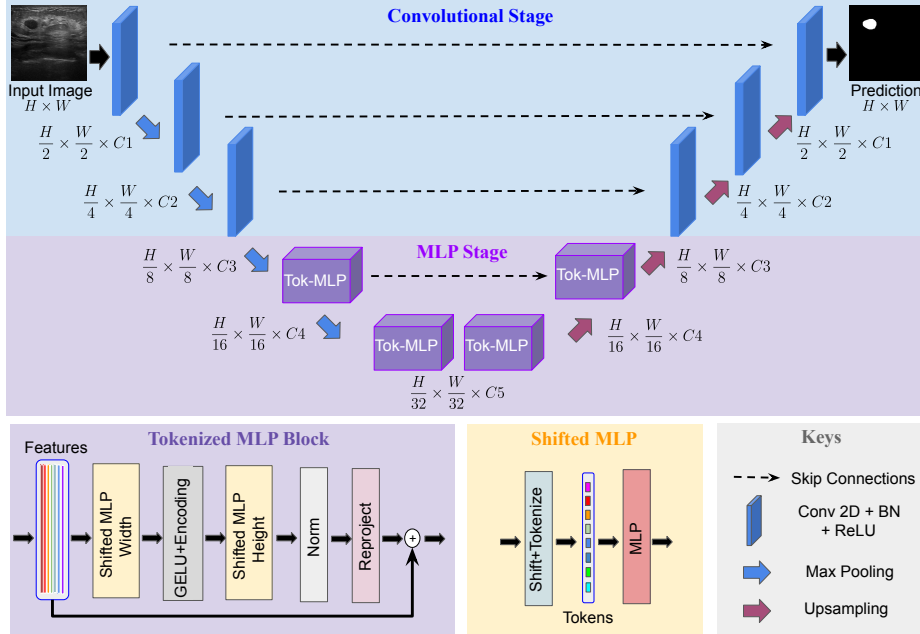

Fig. 2. Overview of the proposed UNeXt architecture.

Convolutional Stage: Each convolutional block is equipped with a convolution layer, a batch normalization layer and ReLU activation. We use a kernel size of 3×3, stride of 1 and padding of 1. The conv blocks in the encoder use a max-pooling layer with pool window 2×2 while the ones in the decoder consist of a bilinear interpolation layer to upsample the feature maps. We use bilinear interpolation instead of transpose convolution as transpose convolution is basically learnable upsampling and contributes to more learnable parameters.
Shifted MLP: In shifted MLP, we first shift the axis of the channels of conv features before tokenizing. This helps the MLP to focus on only certain locations of the conv features thus inducing locality to the block. The intuition here is similar to Swin transformer [5] where window-based attention is introduced to add more locality to an otherwise completely global model. As the Tokenized MLP block has 2 MLPs, we shift the features across width in one and across height in another like in axial-attention [24]. We split the features to h different

partitions and shift them by j locations according to the specified axis. This helps us create random windows introducing locality along an axis.
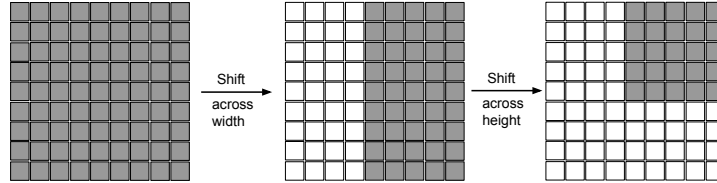

Fig. 3. Shifting operation. The features are shifted sequentially across height and width before tokenizing to induce window locality in the network.

Tokenized MLP Stage: In the tokenized MLP block, we first shift the features and project them into tokens. To tokenize, we first use a kernel size of 3 and change the number of channels to E, where E is the embedding dimension (number of tokens) which is a hyperparameter. We then pass these tokens to a shifted MLP (across width) where hidden dimensions of the MLP is a hyperparameter H. Next, the features are passed through a depth wise convolutional layer (DW-Conv). We use DWConv in this block for two reasons: 1) It helps to encode a positional information of the MLP features. It is shown in [26] that Conv layer in an MLP block is enough to encode the positional information and it actually performs better than the standard positional encoding techniques. Positional encoding techniques like the ones in ViT need to be interpolated when the test and training resolutions are not the same often leading to reduced performance. 2) DWConv uses less number of parameters and hence increases efficiency. We then use a GELU [12] activation layer. We use GELU instead of RELU as it is a more smoother alternative and is found to perform better. In addition, most recent architectures like ViT [10] and BERT [9] have successfully used GELU to obtain improved results. We then pass the features through another shifted MLP (across height) that converts the dimensions from H to O. We use a residual connection here and add the original tokens as residuals. We then apply a layer normalization (LN) and pass the output features to the next block. LN is preferred over BN as it makes more sense to normalize along the tokens instead of normalizing across the batch in the Tokenized MLP block.

The computation in the Tokenized MLP block can be summarized as:

$$X_{shift} = Shift_W(X); T_W = Tokenize(X_{shift}), \tag{1}$$

$$Y = f(DWConv((MLP(T_W)))), \tag{2}$$

$$Y_{shift} = Shift_H(Y); T_H = Tokenize(Y_{shift}), \tag{3}$$

$$Y = f(LN(T + MLP(GELU(T_H)))), \tag{4}$$

where T denotes the tokens, H denotes height, W denotes width, $DWConv$ denotes depth-wise convolution and LN denotes layer normalization. Note that

all of these computations are performed across the embedding dimension H which is significantly less than the dimensionality of the feature maps $\frac{H}{N} \times \frac{H}{N}$ where N is a factor of 2 depending on the block. In our experiments, we set H to 768 unless stated otherwise. This way of designing the Tokenized MLP block helps in encoding meaningful feature information and not contribute much in terms of computation or parameters.

3   Experiments and Results

Datasets: To make our experiments as close to point-of-care imaging as possible, we pick International Skin Imaging Collaboration (ISIC 2018) [8] and Breast UltraSound Images (BUSI) [4] datasets to benchmark our results. The ISIC dataset contains camera-acquired dermatologic images and corresponding segmentation maps of skin lesion regions. The ISIC 2018 dataset consists of 2594 images. We resize all the images to a resolution of 512×512. BUSI consists of ultrasound images of normal, benign and malignant cases of breast cancer along with the corresponding segmentation maps. We use only benign and mailgnant images which results in a total of 647 images resized to a resolution of 256×256.

Implementation Details: We develop UNeXt using Pytorch framework. We use a combination of binary cross entropy (BCE) and dice loss to train UNeXt. The loss $\mathcal{L}$ between the prediction $\hat{y}$ and the target y is formulated as:

$$\mathcal{L} = 0.5BCE(\hat{y}, y) + Dice(\hat{y}, y) \tag{5}$$

We use an Adam optimizer with a learning rate of 0.0001 and momentum of 0.9. We also use a cosine annealing learning rate scheduler with a minimum learning rate upto 0.00001. The batch size is set equal to 8. We train UNeXt for a total of 400 epochs. We perform a 80-20 random split thrice across the dataset and report the mean and variance.

Performance Comparison: We compare the performance of UNeXt with recent and widely used medical image segmentation frameworks. In particular, we compare with convolutional baselines like UNet [17], UNet++ [29] and ResUNet [28]. We also compare with very recent transformer baselines like TransUNet [6] and MedT [20]. Note that we have focused on comparing against the baselines in terms of segmentation performance (F1 score and IoU) as well as number of parameters, computational complexity (in GFLOPs) and inference time (in ms).

We tabulate the results in Table 1. It can be observed that UNeXt obtains better segmentation performance than all the baselines with close second being TransUNet. The improvements are statistically significant with $p < 10^{-5}$. However, the most compelling point to note here is that UNeXt has very less number of computation compared to TransUNet as UNeXt does not have any attention blocks. The computation is calculated in terms of the number of floating point operators (FLOPs). We note that UNeXt has the least GFLOPs of 0.57 compared to TransUNet's 38.52 and UNet's 55.84. It is also the most lightweight network compared to all baselines. In particular, we note that UNeXt has only 1.58 M parameters compared to 105.32 M parameters of TransUNet.

We also present the average inference time while operating on a CPU. Note that we have specifically bench-marked the inference time in CPU instead of GPU as point-of-care devices mostly operate on low-compute power and often do not have the computing advantage of GPU. We perform feed forward for 10 images of resolution 256×256 and report the average inference time. The CPU used for bench-marking was an Intel Xeon Gold 6140 CPU operating at 2.30 GHz. It can be noted that we experimented with Swin-UNet [5] but found have problems with convergence on small datasets resulting in poor performance. However, Swin-UNet is heavy with 41.35 M parameters and also computationally complex with 11.46 GFLOPs.

Table 1. Performance Comparison with convolutional and transformer baselines.

Networks	Params (in M)	Inference Speed (in ms)	GFLOPs	ISIC [8]		BUSI [4]	
				F1	IoU	F1	IoU
UNet [17]	31.13	223	55.84	84.03 ± 0.87	74.55 ± 0.96	76.35 ± 0.89	63.85 ± 1.12
UNet++ [29]	9.16	173	34.65	84.96 ± 0.71	75.12 ± 0.65	77.54 ± 0.74	64.33 ± 0.75
ResUNet [28]	62.74	333	94.56	85.60 ± 0.68	75.62 ± 1.11	78.25 ± 0.74	64.89 ± 0.83
MedT [20]	1.60	751	21.24	87.35 ± 0.18	79.54 ± 0.26	76.93 ± 0.11	63.89 ± 0.55
TransUNet [6]	105.32	246	38.52	88.91 ± 0.63	80.51 ± 0.72	79.30 ± 0.37	66.92 ± 0.75
UNeXt	**1.47**	**25**	**0.57**	**89.70 ± 0.96**	**81.70 ± 1.53**	**79.37 ± 0.57**	**66.95 ± 1.22**

In Figure 4, we plot the comparison charts of F1 score vs. GLOPs, F1 score vs. Inference time and F1 Score vs. Number of Parameters. The F1 score used here corresponds to the ISIC dataset. It can be clearly seen from the charts that UNeXt and TransUNet are the best performing methods in terms of the segmentation performance. However, UNeXt clearly outperforms all the other networks in terms of computational complexity, inference time and number of parameters which are all important characteristics to consider for point-of-care imaging applications. In Figure 5, we present sample qualitative results of UNeXt along with other baselines. It can be observed that UNeXt produces competitive segmentation predictions compared to the other methods.
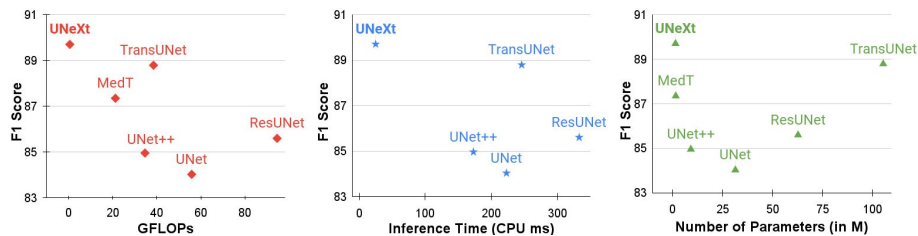

Fig. 4. Comparison Charts. Y-axis corresponds to F1 score (higher the better). X-axis corresponds to GFLOPs, inference time and number of parameters (lower the better). It can be seen that UNeXt is the most efficient network compared to the others.
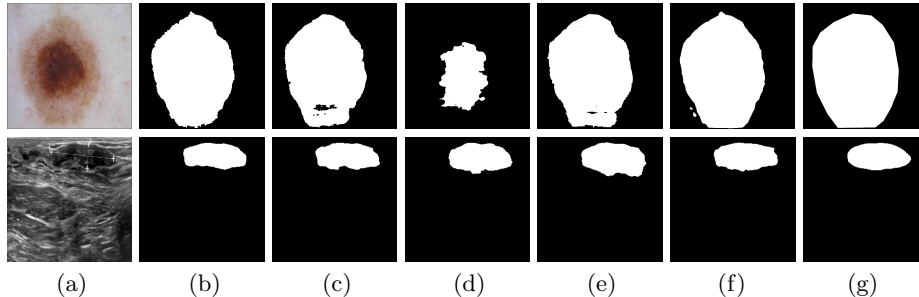
Fig. 5. Qualitative comparisons. Row 1 - ISIC dataset, Row 2 - BUSI dataset. (a) Input. Predictions of (b) UNet (c) UNet++ (d) MedT (e) TransUNet (f) UNeXt and (g) Ground Truth.

4   Discussion

Ablation Study: We conduct an ablation study (shown in Table 2) to understand the individual contribution of each module in UNeXt. We first start with the original UNet and then just reduce the number of filters to reduce the number of parameters and complexity. We see a reduction of performance with not much reduction in parameters. Next, we reduce the depth and use only a 3-level deep architecture which is basically the Conv stage of UNeXt. This reduces the number of parameters and complexity significantly but also reduces the performance by 4%. Now, we introduce the tokenized MLP block which improves the performance significantly while increasing the complexity and parameters by a minimal value. Next, we add the positional embedding method using DWConv as in [26] and see some more improvement. Next, we add the shifting operation in the MLPs and show that shifting the features before tokenizing improves the performance without any addition to parameters or complexity. As the shift operation does not contribute to any addition or multiplication it does not add on to any FLOPs. We note that shifting the features across both axes results in the best performance which is the exact configuration of UNeXt with minimal parameters and complexity. Note that all of the above experiments were conducted using a single fold of the ISIC dataset.

Table 2. Ablation Study.

Network	Params	Inf. Time (in ms)	GFLOPs	F1	IoU
Original UNet	31.19	223	55.84	84.03	74.55
Reduced UNet	7.65	38	9.36	83.65	72.54
Conv Stage	0.88	9	0.36	80.12	67.75
Conv Stage + Tok-MLP w/o PE	1.46	22	0.57	88.78	79.32
Conv Stage + Tok-MLP + PE	1.47	23	0.57	89.25	80.76
Conv Stage + Shifted Tok-MLP (W) + PE	1.47	24	0.57	89.38	82.01
Conv Stage + Shifted Tok-MLP (H) + PE	1.47	24	0.57	89.25	81.94
Conv Stage + Shifted Tok-MLP (H+W) + PE	1.47	25	0.57	90.41	82.78

Analysis on number of channels: The number of channels is a main hyperparamter of UNeXt which affects the number of parameters, complexity and the performance of the network. In Table 3, we conduct experiments on single fold of ISIC to show two more different configurations of UNeXt. It can be observed that increasing the channels (UNeXt-L) further improves the performance while adding on to computational overhead. Although decreasing it (UNeXt-S) reduces the performance (the reduction is not drastic) but we get a very lightweight model.

Table 3. Analysis on the number of channels.

Network	C1	C2	C3	C4	C5	Params	Inf. Speed (in ms)	GFLOPs	F1	IoU
UNeXt-S	8	16	32	64	128	0.32	22	0.10	89.62	81.40
UNeXt-L	32	64	128	256	512	3.99	82	1.42	90.65	83.10
UNeXt	16	32	128	160	256	1.47	25	0.57	90.41	82.78

Difference from MLP-Mixer: MLP-Mixer uses an all-MLP architecture for image recognition. UNeXt is a convolutional and MLP-based network for image segmentation. MLP-Mixer focuses on channel mixing and token mixing to learn a good representation. In contrast, we extract convolutional features and then tokenize the channels and use a novel tokenized MLPs using shifted MLPs to model the representation. It is worthy to note that we experimented with MLP-Mixer as encoder and a normal convolutional decoder. The performance was not optimal for segmentation and it was still heavy with around 11 M parameters.

5   Conclusion

In this work, we have proposed a new deep network architecture UNeXt for medical image segmentation focussed for point-of-care applications. UNeXt is a convolutional and MLP-based architecture where there is an initial conv stage followed by MLPs in the latent space. Specifically, we propose a tokenized MLP block with shifted MLPs to efficiently model the representation with minimal complexity and parameters. We validated UNeXt on multiple datasets where we achieve faster inference, reduced complexity and less number of parameters while also achieving the state-of-the-art performance.

References

1. https://www.butterflynetwork.com/iq
2. https://blog.google/technology/health/ai-dermatology-preview-io-2021/
3. https://hyperfine.io/
4. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)

5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)

6. 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)

7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)

8. 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)

9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

10. 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)

11. 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)

12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

13. 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)

14. Lian, D., Yu, Z., Sun, X., Gao, S.: As-mlp: An axial shifted mlp architecture for vision. arXiv preprint arXiv:2107.08391 (2021)

15. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 893–901. Springer (2018)

16. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)

17. 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)

18. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems **34** (2021)

19. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. arXiv preprint arXiv:2105.03404 (2021)
20. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 36–46. Springer International Publishing, Cham (2021)
21. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. pp. 363–373. Springer (2020)
22. Valanarasu, J.M.J., Sindagi, V.A., Hacihaliloglu, I., Patel, V.M.: Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. IEEE Transactions on Medical Imaging (2021)
23. Vashist, S.K.: Point-of-care diagnostics: Recent advances and trends. Biosensors **7**(4), 62 (2017)
24. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2020)
25. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119. Springer (2021)
26. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203 (2021)
27. Yu, T., Li, X., Cai, Y., Sun, M., Li, P.: S2-mlp: Spatial-shift mlp architecture for vision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 297–306 (2022)
28. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018)
29. 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)