ACC-UNet: A Completely Convolutional UNet Model for the 2020s

Nabil Ibtehaz[1] and Daisuke Kihara[1,2(✉)]

[1] Department of Computer Science, Purdue University, West Lafayette, IN, USA
[2] Department of Biological Sciences, Purdue University, West Lafayette, IN, USA
dkihara@purdue.edu

Abstract. This decade is marked by the introduction of Vision Transformer, a radical paradigm shift in broad computer vision. A similar trend is followed in medical imaging, UNet, one of the most influential architectures, has been redesigned with transformers. Recently, the efficacy of convolutional models in vision is being reinvestigated by seminal works such as ConvNext, which elevates a ResNet to Swin Transformer level. Deriving inspiration from this, we aim to improve a purely convolutional UNet model so that it can be on par with the transformer-based models, e.g., Swin-Unet or UCTransNet. We examined several advantages of the transformer-based UNet models, primarily long-range dependencies and cross-level skip connections. We attempted to emulate them through convolution operations and thus propose, ACC-UNet, a completely convolutional UNet model that brings the best of both worlds, the inherent inductive biases of convnets with the design decisions of transformers. ACC-UNet was evaluated on 5 different medical image segmentation benchmarks and consistently outperformed convnets, transformers, and their hybrids. Notably, ACC-UNet outperforms state-of-the-art models Swin-Unet and UCTransNet by $2.64 \pm 2.54\%$ and $0.45 \pm 1.61\%$ in terms of dice score, respectively, while using a fraction of their parameters (59.26% and 24.24%). Our codes are available at https://github.com/kiharalab/ACC-UNet.

Keywords: UNet · image segmentation · fully convolutional network

1 Introduction

Semantic segmentation, an essential component of computer-aided medical image analysis, identifies and highlights regions of interest in various diagnosis tasks. However, this often becomes complicated due to various factors involving image modality and acquisition along with pathological and biological variations [18]. The application of deep learning in this domain has thus certainly benefited in this regard. Most notably, ever since its introduction, the UNet model [19] has
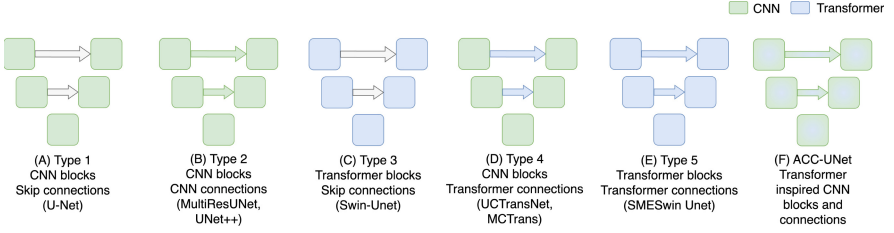
Fig. 1. Developments and innovations in the UNet architecture.

demonstrated astounding efficacy in medical image segmentation. As a result, UNet and its derivatives have become the de-facto standard [25].

The original UNet model comprises a symmetric encoder-decoder architecture (Fig. 1a) and employs skip-connections, which provide the decoder spatial information probably lost during the pooling operations in the encoder. Although this information propagation through simple concatenation improves the performance, there exists a likely semantic gap between the encoder-decoder feature maps. This led to the development of a second class of UNets (Fig. 1b). U-Net++ [26] leveraged dense connections and MultiResUNet [11] added additional convolutional blocks along the skip connection as a potential remedy.

Till this point in the history of UNet, all the innovations were performed using CNNs. However, the decade of 2020 brought radical changes in the computer vision landscape. The long-standing dominance of CNNs in vision was disrupted by vision transformers [7]. Swin Transformers [15] further adapted transformers for general vision applications. Thus, UNet models started adopting transformers [5]. Swin-Unet [9] replaced the convolutional blocks with Swin Transformer blocks and thus initiated a new class of models (Fig. 1c). Nevertheless, CNNs still having various merits in image segmentation, led to the development of fusing those two [2]. This hybrid class of UNet models (Fig. 1d) employs convolutional blocks in the encoder-decoder and uses transformer layers along the skip connections. UCTransNet [22] and MCTrans [24] are two representative models of this class. Finally, there have also been attempts to develop all-transformer UNet architectures (Fig. 1e), for instance, SMESwin Unet [27] uses transformer both in encoder-decoder blocks and the skip-connection.

Very recently, studies have begun rediscovering the potential of CNNs in light of the advancements brought by transformers. The pioneering work in this regard is 'A ConvNet for the 20202020ss' [16], which explores the various ideas introduced by transformers and their applicability in convolutional networks. By gradually incorporating ideas from training protocol and micro-macro design choices, this work enabled ResNet models to outperform Swin Transformer models.

In this paper, we ask the same question but in the context of UNet models. We investigate if a UNet model solely based on convolution can compete with the transformer-based UNets. In doing so, we derive motivations from the transformer architecture and develop a purely convolutional UNet model. We propose

a patch-based context aggregation contrary to window-based self-attention. In addition, we innovate the skip connections by fusing the feature maps from multiple levels of encoders. Extensive experiments on 5 benchmark datasets suggest that our proposed modifications have the potential to improve UNet models.

2   Method

Firstly, we analyze the transformer-based UNet models from a high-level. Deriving motivation and insight from this, we design two convolutional blocks to simulate the operations performed in transformers. Finally, we integrate them in a vanilla UNet backbone and develop our proposed ACC-UNet architecture.
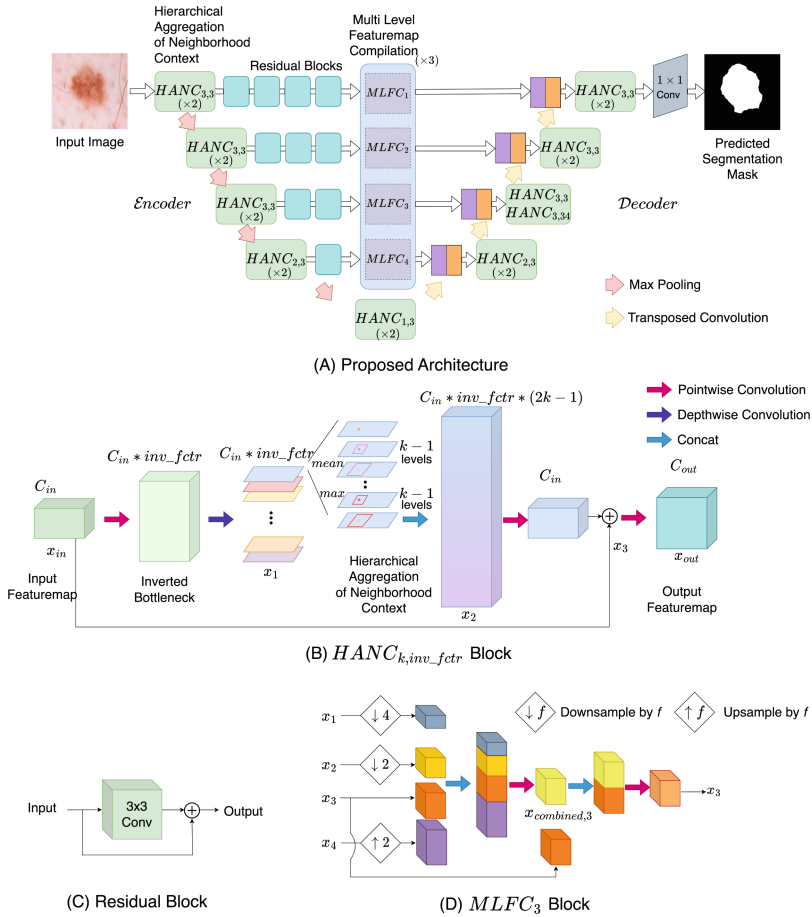

Fig. 2. (A) Architecture of the proposed ACC-UNet. (B) A generalized view of $HANC_{k,inv\_fctr}$ block. (C) A generic residual block used in skip connection. (D) An example view of the 3rd level $MLFC$ block

2.1   A High-Level View of Transformers in UNet

Transformers apparently improve UNet models in two different aspects.

Leveraging the Long-Range Dependency of Self-attention. Transformers can compute features from a much larger view of context through the use of (windowed) self-attention. In addition, they improve expressivity by adopting inverted bottlenecks, i.e., increasing the neurons in the MLP layer. Furthermore, they contain shortcut connections, which facilitate the learning [7].

Adaptive Multi-level Feature Combination Through Channel Attention. Transformer-based UNets fuse the feature maps from multiple encoder levels adaptively using channel attention. This generates enriched features due to the combination of various regions of interest from different levels compared to simple skip-connection which is limited by the information at the current level [22].

Based on these observations, we modify the convolutional blocks and skip-connections in a vanilla UNet model to induce the capabilities of long-range dependency and multi-level feature combinations.

2.2   Hierarchical Aggregation of Neighborhood Context (HANC)

We first explore the possibility of inducing long-range dependency along with improving expressivity in convolutional blocks. We only use pointwise and depthwise convolutions to reduce the computational complexity [8].

In order to increase the expressive capability, we propose to include inverted bottlenecks in convolutional blocks [16], which can be achieve by increasing the number of channels from c_{in} to $c_{inv} = c_{in} * inv\_fctr$ using pointwise convolution. Since these additional channels will increase the model complexity, we use 3×3 depthwise convolution to compensate. An input feature map $x_{in} \in \mathbb{R}^{c_{in},n,m}$ is thus transformed to $x_1 \in \mathbb{R}^{c_{inv},n,m}$ as (Fig. 2b)

$$x_1 = DConv_{3\times3}(PConv_{c_{in}\rightarrow c_{inv}}(x_{in})) \tag{1}$$

Next, we wish to emulate self-attention in our convolution block, which at its core is comparing a pixel with the other pixels in its neighborhood [15]. This comparison can be simplified by comparing a pixel value with the mean and maximum of its neighborhood. Therefore, we can provide an approximate notion of neighborhood comparison by appending the *mean* and *max* of the neighboring pixel features. Consecutive pointwise convolution can thus consider these and capture a contrasting view. Since hierarchical analysis is beneficial for images [23], instead of computing this aggregation in a single large window, we compute this in multiple levels hierarchically, for example, $2 \times 2, 2^2 \times 2^2, \cdots, 2^{k-1} \times 2^{k-1}$ patches. For $k = 1$, it would be the ordinary convolution operation, but as we increase the value of k, more contextual information will be provided, bypassing the need for larger convolutional kernels. Thus, our proposed hierarchical neighborhood context aggregation enriches feature map $x_1 \in \mathbb{R}^{c_{inv},n,m}$ with

contextual information as $x_2 \in \mathbb{R}^{c_{inv}*(2k-1),n,m}$ (Fig. 2b), where $||$ corresponds to concatenation along the channel dimension

$$x_2 = (x_1||mean_{2\times2}(x_1)||mean_{2^2\times2^2}(x_1)||\cdots||mean_{2^{k-1}\times2^{k-1}}(x_1) \atop ||max_{2\times2}(x_1)||max_{2^2\times2^2}(x_1)||\cdots||max_{2^{k-1}\times2^{k-1}}(x_1)) \quad (2)$$

Next, similar to the transformer, we include a shortcut connection in the convolution block for better gradient propagation. Hence, we perform another pointwise convolution to reduce the number of channels to c_{in} and add with the input feature map. Thus, $x_2 \in \mathbb{R}^{c_{inv}*(2k-1),n,m}$ becomes $x_3 \in \mathbb{R}^{c_{in},n,m}$ (Fig. 2b)

$$x_3 = PConv_{c_{inv}*(2k-1)\to c_{in}}(x_2) + x_{in} \quad (3)$$

Finally, we change the number of filters to c_{out}, as the output, using pointwise convolution (Fig. 2b)

$$x_{out} = PConv_{c_{in}\to c_{out}}(x_3) \quad (4)$$

Thus, we propose a novel Hierarchical Aggregation of Neighborhood Context (HANC) block using convolution but bringing the benefits of transformers. The operation of this block is illustrated in Fig. 2b.

2.3   Multi Level Feature Compilation (MLFC)

Next, we investigate the feasibility of multi-level feature combination, which is the other advantage of using transformer-based UNets.

Transformer-based skip connections have demonstrated effective feature fusion of all the encoder levels and appropriate filtering from the compiled feature maps by the individual decoders [22,24,27]. This is performed through concatenating the projected tokens from different levels [22]. Following this approach, we resize the convolutional feature maps obtained from the different encoder levels to make them equisized and concatenate them. This provides us with an overview of the feature maps across the different semantic levels. We apply pointwise convolution operation to summarize this representation and merge with the corresponding encoder feature map. This fusion of the overall and individual information is passed through another convolution, which we hypothesize enriches the current level feature with information from other level features.

For the features, x_1, x_2, x_3, x_4 from 4 different levels, the feature maps can be enriched with multilevel information as (Fig. 2d)

$$x_{comb,i} = PConv_{c_{tot}\to c_i}(resize_i(x1)||resize_i(x2)||resize_i(x3)||resize_i(x4)) \quad (5)$$

$$x_i = PConv_{2c_i\to c_i}(x_{comb,i}||x_i), \qquad i = 1,2,3,4 \quad (6)$$

Here, $resize_i(x_j)$ is an operation that resizes x_j to the size of x_i and $c_{tot} = c_1 + c_2 + c_3 + c_4$. This operation is done individually for all the different levels.

We thus propose another novel block named Multi Level Feature Compilation (MLFC), which aggregates information from multiple encoder levels and enriches the individual encoder feature maps. This block is illustrated in Fig. 2d.

2.4 ACC-UNet

Therefore, we propose fully convolutional ACC-UNet (Fig. 2a). We started with a vanilla UNet model and reduced the number of filters by half. Then, we replaced the convolutional blocks from the encoder and decoder with our proposed HANC blocks. We considered $inv\_fctr = 3$, other than the last decoder block at level 3 ($inv\_fctr = 34$) to mimic the expansion at stage 3 of Swin Transformer. $k = 3$, which considers up to 4×4 patches, was selected for all but the bottleneck level ($k = 1$) and the one next to it ($k = 2$). Next, we modified the skip connections by using residual blocks (Fig. 2c) to reduce semantic gap [11] and stacked 3 MLFC blocks. All the convolutional layers were batch-normalized [12], activated by Leaky-RELU [17] and recalibrated by squeeze and excitation [10].

To summarize, in a UNet model, we replaced the classical convolutional blocks with our proposed HANC blocks that perform an approximate version of self-attention and modified the skip connection with MLFC blocks which consider the feature maps from different encoder levels. The proposed model has 16.77 M parameters, roughly a 2M increase than the vanilla UNet model.

3 Experiments

3.1 Datasets

In order to evaluate ACC-UNet, we conducted experiments on 5 public datasets across different tasks and modalities. We used ISIC-2018 [6,21] (dermoscopy, 2594 images), BUSI [3](breast ultrasound, used 437 benign and 210 malignant images similar to [13]), CVC-ClinicDB [4] (colonoscopy, 612 images), COVID [1] (pneumonia lesion segmentation, 100 images), and GlaS [20] (gland segmentation, 85 training, and 80 test images). All the images and masks were resized to 224×224. For the GlaS dataset, we considered the original test split as the test data, for the other datasets we randomly selected 20% of images as test data. The remaining 60% and 20% images were used for training and validation and the experiments were repeated 3 times with different random shuffling.

3.2 Implementation Details

We implemented ACC-UNet model in PyTorch and used a workstation equipped with AMD EPYC 7443P 24-Core CPU and NVIDIA RTX A6000 (48G) GPU for our experiments. We designed our training protocol identical to previous works [22], except for using a batch size of 12 throughout our experiments [27]. The models were trained for 1000 epochs [27] and we employed an early stopping patience of 100 epochs. We minimized the combined cross-entropy and dice loss [22] using the Adam [14] optimizer with an initial learning rate of 10^{-3}, which was adjusted through cosine annealing learning rate scheduler [13][1]. We performed online data augmentations in the form of random flipping and rotating [22].

---

[1] Swin-UNet-based models were trained with SGD [9] for poor performance of Adam.

3.3   Comparisons with State-of-the-Art Methods

We evaluated ACC-UNet against UNet, MultiResUNet, Swin-Unet, UCTransnet, SMESwin-Unet, i.e., one representative model from the 5 classes of UNet, respectively (Fig. 1). Table 1 presents the dice score obtained on the test sets.

Table 1. Comparison with the state-of-the-art models. The first and second best scores are styled as bold and italic, respectively. The subscripts denote the standard deviation.

Model	params	FLOPs	ISIC-18	ClinicDB	BUSI	COVID	GlaS
UNet	14M	37G	$87.97_{0.11}$	$90.66_{0.92}$	$72.27_{0.86}$	$71.21_{1.4}$	$87.99_{1.32}$
MultiResUNet	7.3 M	1.1G	$88.55_{0.24}$	$88.20_{1.67}$	$72.43_{0.91}$	$71.33_{3.59}$	$88.34_{1.05}$
Swin-Unet	27.2 M	6.2G	$89.24_{0.14}$	$90.69_{0.50}$	$76.06_{0.43}$	$68.56_{1.07}$	$86.45_{0.28}$
UCTransnet	66.4 M	38.8G	$89.08_{0.44}$	$92.57_{0.39}$	$76.56_{0.2}$	$73.09_{3.63}$	$87.17_{0.85}$
SMESwin-Unet	169.8 M	6.4G	$88.57_{0.13}$	$89.62_{0.08}$	$73.94_{2.06}$	$58.4_{0.03}$	$83.72_{0.18}$
ACC-UNet	16.8 M	38G	$89.37_{0.34}$	$92.67_{0.57}$	$77.19_{0.87}$	$73.99_{0.53}$	$88.61_{0.61}$

The results show an interesting pattern. Apparently, for the comparatively larger datasets (ISIC-18) transformer-based Swin-Unet was the 2nd best method, as transformers require more data for proper training [2]. On the other end of the spectrum, lightweight convolutional model (MultiResUNet) achieved the 2nd best score for small datasets (GlaS). For the remaining datasets, hybrid model (UCTransnet) seemed to perform as the 2nd best method. SMESwin-Unet fell behind in all the cases, despite having such a large number of parameters, which in turn probably makes it difficult to be trained on small-scale datasets.

However, our model combining the design principles of transformers with the inductive bias of CNNs seemed to perform best in all the different categories with much lower parameters. Compared to much larger state-of-the-art models, for the 5 datasets, we achieved $0.13\%, 0.10\%, 0.63\%, 0.90\%, 0.27\%$ improvements in dice score, respectively. Thus, our model is not only accurate, but it is also efficient in using the moderately small parameters it possesses. In terms of FLOPs, our model is comparable with convolutional UNets, the transformer-based UNets have smaller FLOPs due to the massive downsampling at patch partitioning.

3.4   Comparative Qualitative Results on the Five Datasets

In addition to, achieving higher dice scores, apparently, ACC-UNet generated better qualitative results. Figure 3 presents a qualitative comparison of ACC-UNet with the other models. Each row of the figure comprises one example from each of the datasets and the segmentation predicted by ACC-UNet and the ground truth mask are presented in the rightmost two columns. For the 1st example from the ISIC-18 dataset, our model did not oversegment but rather followed the lesion boundary. In the 2nd example from CVC-ClinicDB, our model managed to distinguish the finger from the polyp almost perfectly. Next in the
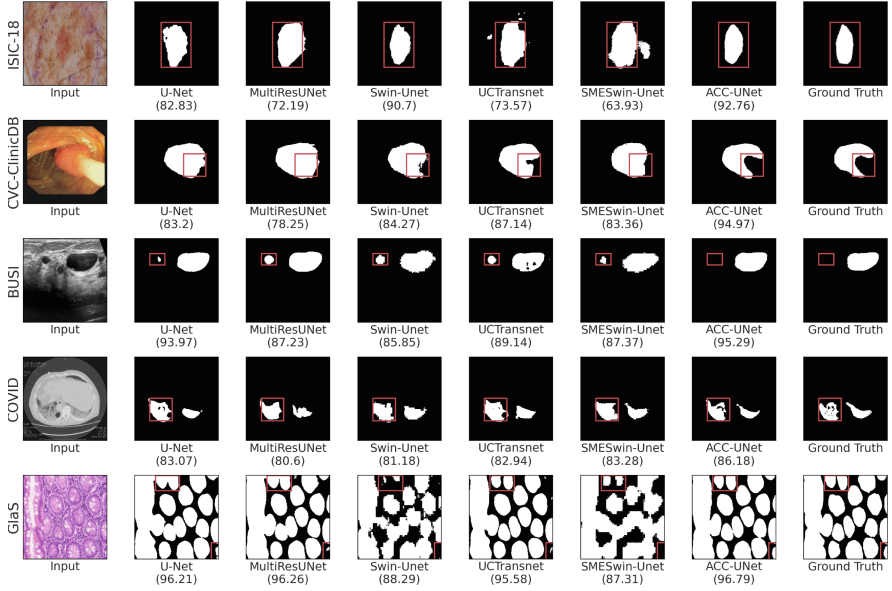
Fig. 3. Comparative qualitative results, with dice score provided inside the parenthesis.

3rd example from BUSI, our prediction filtered out the apparent nodule region on the left, which was predicted as a false positive tumor by all the other models. Similarly, in the 4th sample from the COVID dataset, we were capable to model the gaps in the consolidation of the left lung visually better, which in turn resulted in 2.9% higher dice score than the 2nd best method. Again, in the final example from the GlaS dataset, we not only successfully predicted the gland at the bottom right corner but also identified the glands at the top left individually, which were mostly missed or merged by the other models, respectively.

3.5   Ablation Study

We performed an ablation study on the CVC-ClinicDB dataset to analyze the contributions of the different design choices in our roadmap (Fig. 4). We started with a UNet model with the number of filters halved as our base model, which results in a dice score of 87.77% with 7.8M parameters. Using depthwise convolutional along with increasing the bottleneck by 4 raised the dice score to 88.26% while slightly reducing the parameters to 7.5M. Next, HANC block was added with $k = 3$ throughout, which increased the number of parameters by 340% for an increase of 1.1% dice score. Shortcut connections increased the performance by 2.16%. We also slowly reduced both k and inv_fctr which reduced the number of parameters without any fall in performance. Finally, we added the MLFC blocks (4 stacks) and gradually optimized k and inv_fctr along with dropping one MLFC stage, which led to the development of ACC-UNet. Some other inter-
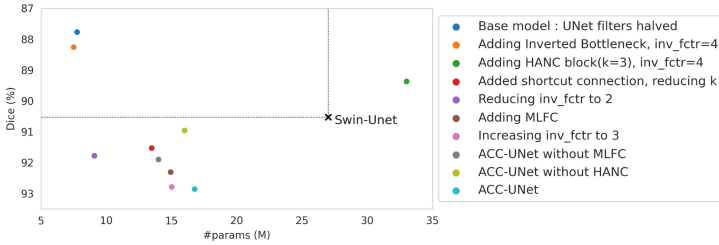
Fig. 4. Ablation study on the CVC-ClinicDB dataset.

esting ablations were ACC-UNet without MLFC (dice 91.9%) or HANC (dice 90.96%, with 25% more filters to keep the number of parameters comparable).

4   Conclusions

Acknowledging the benefits of various design paradigms in transformers, we investigate the suitability of similar ideas in convolutional UNets. The resultant ACC-UNet possesses the inductive bias of CNNs infused with long-range and multi-level feature accumulation of transformers. Our experiments reveals this amalgamation indeed has the potential to improve UNet models. One limitation of our model is the slowdown from concat operations (please see supplementary materials), which can be solved by replacing them. In addition, there are more innovations brought by transformers [16], e.g., layer normalization, GELU activation, AdamW optimizer, these will be explored further in our future work.

References

1. Covid-19 ct segmentation dataset. https://medicalsegmentation.com/covid19/. Accessed 20 Aug 2022
2. Ailiang, L., Xu, J., Jinxing, L., Guangming, L.: ConTrans: improving transformer with convolutional attention for medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 297–307. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_29
3. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
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. Comput. Med. Imaging Graph. **43**, 99–111 (2015)

5. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation (2021)
6. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC) (2019)
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
9. Hu, C., Wang, Y., Joy, C., Dongsheng, J., Xiaopeng, Z., Qi, T., Manning, W.: Swin-Unet: unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. IEEE (2018)
11. Ibtehaz, N., Rahman, M.: MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. **121**, 74–87 (2020)
12. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, vol. 37. pp. 448–456. JMLR.org (2015)
13. Jose, V.J.M., Patel, V.M.: UNeXt: MLP-based rapid medical image segmentation network. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 23–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_3
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2015)
15. Liu, Z., et al.: Swin transformer: hierarchical vis ion transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002. IEEE (2021)
16. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976. IEEE (2022)
17. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the 30th International Conference on Machine Learning (2013)
18. Olabarriaga, S., Smeulders, A.: Interaction in the segmentation of medical images: a survey. Med. Image Anal. **5**(2), 127–142 (2001)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Sirinukunwattana, K., et al.: Gland segmentation in colon histology images: the glas challenge contest. Med. Image Anal. **35**, 489–502 (2017)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 180161 (2018)
22. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: UCTransNet: rethinking the skip connections in U-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 3, pp. 2441–2449 (2022)

23. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks (2022)
24. Ji, Y., et al.: Multi-compound transformer for accurate biomedical image segmentation. In: de Bruijn, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 326–336. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_31
25. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE **109**(5), 820–838 (2021)
26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
27. Ziheng, W., et al.: SMESwin Unet: merging CNN and transformer for medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 517–526. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_50