CENet: Context Enhancement Network for Medical Image Segmentation

Afshin Bozorgpour[1], Sina Ghorbani Kolahi[2], Reza Azad[1], Ilker Hacihaliloglu[3], and Dorit Merhof[1,4]

[1] Faculty of Informatics and Data Science, University of Regensburg, Germany
[2] Department of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[3] Department of Radiology & Department of Medicine, University of British Columbia, Vancouver, BC, Canada
[4] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
{dorit.merhof@ur.de}

Abstract. Medical image segmentation, particularly in multi-domain scenarios, requires precise preservation of anatomical structures across diverse representations. While deep learning has advanced this field, existing models often struggle with accurate boundary representation, variability in organ morphology, and information loss during downsampling, limiting their accuracy and robustness. To address these challenges, we propose the Context Enhancement Network (CENet), a novel segmentation framework featuring two key innovations. First, the Dual Selective Enhancement Block (DSEB) integrated into skip connections enhances boundary details and improves the detection of smaller organs in a context-aware manner. Second, the Context Feature Attention Module (CFAM) in the decoder employs a multi-scale design to maintain spatial integrity, reduce feature redundancy, and mitigate overly enhanced representations. Extensive evaluations on both radiology and dermoscopic datasets demonstrate that CENet outperforms state-of-the-art (SOTA) methods in multi-organ segmentation and boundary detail preservation, offering a robust and accurate solution for complex medical image analysis tasks. The source code is publicly available at https://github.com/xmindflow/cenet.

Keywords: Medical Image Segmentation · Feature Enhancement · Multi-scale Representation

1 Introduction

Medical image segmentation, driven by advancements in deep learning and computer vision, is a critical tool for extracting semantically meaningful information from raw medical datasets. It enables the precise pixel-wise delineation of anatomical structures, organs, and lesions, which are often characterized by diverse shapes, appearances, and pathological conditions [1]. This capability is

essential for clinical applications and computer-aided diagnosis systems. Among the most prominent approaches for segmentation are Fully Convolutional Neural Networks (FCNs), particularly the U-Net architecture [20] and its variants. These models leverage an encoder-decoder structure to capture multiscale representations: the encoder extracts contextual information, while the decoder upsamples compressed features to produce precise, localized predictions. Skip connections further enhance this process by preserving fine-grained spatial details [18], such as boundaries, and enabling the decoder to reconstruct predictions more accurately using high-quality, contextualized features from the encoder.

Despite their strengths, Convolutional Neural Networks (CNNs) inherently struggle to model global contextual relationships due to the limited receptive field of convolutional kernels. This limitation often leads to suboptimal performance in multiscale segmentation tasks involving complex structures [22]. To address this, various strategies have been proposed, including deformable convolutions [2], dilated convolutions [8], spatial pyramid pooling [6], and the integration of attention mechanisms into high-level semantic feature maps [23]. More recently, the Vision Transformer (ViT) [7] has emerged as a promising alternative, utilizing self-attention mechanisms to effectively model long-range dependencies and achieve SOTA performance. However, while ViTs excel at capturing global context, they often underperform in modeling local representations and context. Additionally, their quadratic computational complexity makes them inefficient for large-scale applications [12].

Despite advances in CNNs and Transformer networks, their hierarchical structures, reliance on downsampling, and self-attention mechanisms often compromise boundary details and fine-grained semantic representation, limiting their ability to capture multiscale features essential for complex organ and lesion morphologies in medical images. Although hybrid CNN-Transformer architectures [5,9,13] and localized self-attention mechanisms [4,10] have been explored, their focus on global representations and fixed receptive fields restricts accurate segmentation of deformable structures across scales. Many approaches, such as [27], focus on body features over edge information, which is crucial for accurate segmentation and detail reconstruction. While studies like [17,14] separately integrate fine-grained features (e.g., boundaries), the absence of proper control mechanisms often results in noise, decoder-stage degradation, and inefficient network learning with strong inductive bias.

To address these challenges, we propose the Context Enhancement Network (CENet), a novel framework for medical image segmentation. ❶ The Dual Selective Enhancement Block (DSEB) refines fine-grained features by leveraging coarse guidance from the previous decoder, amplifying salient regions while filtering irrelevant information to maintain contextual balance. ❷ The decoder includes the Context Feature Attention Module (CFAM), which uses depth-wise dilated convolutions and a context-aware gating mechanism for multiscale feature representation while addressing over-enhancement through adaptive rectification. ❸ Evaluations on radiology (*Synapse*, *ACDC*) and dermoscopic datasets

(PH^2, $HAM10000$) show that CENet outperforms SOTA methods in precise and context-aware segmentation.

2 Method

In this study, we propose a novel Context Enhancement Network (CENet) designed to enhance feature representation and improve segmentation accuracy by leveraging hierarchical feature extraction and refined contextual information to effectively capture global contextual dependencies and local spatial details. The network is systematically structured into three key components: a Pyramid Vision Transformer V2 (PvT-V2) [27] backbone for multi-scale feature extraction, DSEB to contextually enrich skip connections, and CFAM for multi-scale representation and semantic feature refinement. The overall architecture of the proposed method is illustrated in Figure 1.
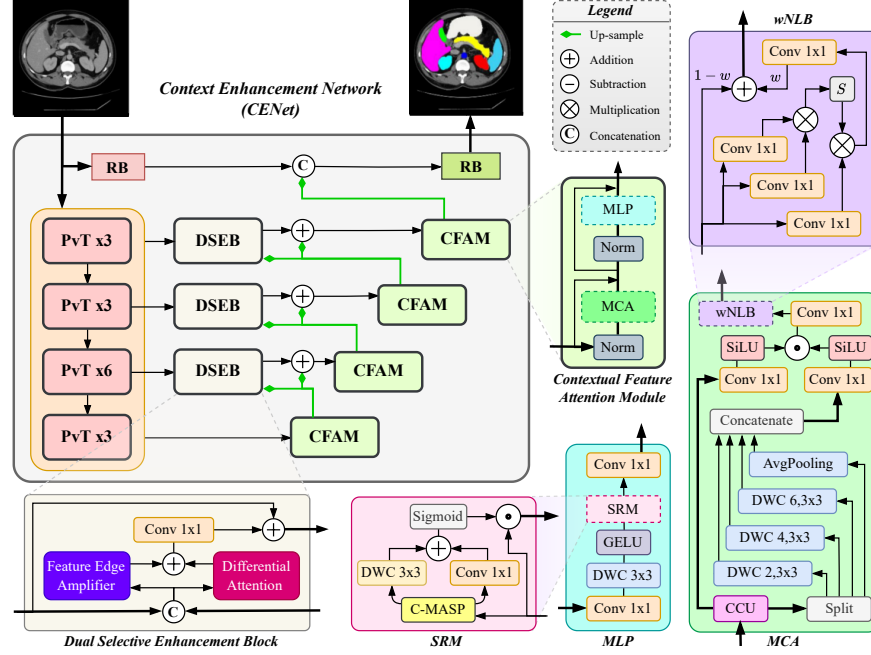

Fig. 1: The Contextual Enhancement Network (CENet) uses a pretrained encoder to generate multi-resolution features, processed by the DSEB as skip connections. The decoder refines these features via the CFAM, which includes a CCU, MCA, wNLB, and MLP.

2.1 Dual Selective Enhancement Block (DSEB)

The DSEB improves CENet's skip connections for precise, boundary-sensitive medical image segmentation. After concatenating encoder features with upsampled CFAM outputs (using EUC Block [19]) as decoder signals, the input is processed through two stages: the Feature Edge Amplifier (FEA) and Differential Attention (DiffAtt) [29]. The FEA enhances skip connection feature maps by refining edges and spatial details crucial for boundary detection and the localization of small structures. It begins by performing multi-scale downsampling $\mathcal{D}(F, s)$ and upsampling $U(F, s)$ on the input feature map F, defined as:

$$F_{u1} = U(\mathcal{D}(F, s_1), s_0), \quad F_{u2} = U(\mathcal{D}(F, s_2), s_0), \tag{1}$$

where $s_1 = 0.75$, $s_2 = 0.5$, and $s_0 = 1.0$ are the scales. The difference between these upsampled features isolates refined edge details, computed as $F_{\text{edge}} = |F_{u1} - F_{u2}|$. These features are added back to the original feature map F, weighted by $\lambda \in \mathbb{R}^d$, resulting in the enhanced feature map:

$$\tilde{F} = F + \lambda F_{\text{edge}}. \tag{2}$$

The DiffAtt, inspired by NLP, reduces attention noise and selectively enhances meaningful context from the concatenated input features. Query and key vectors are split into two groups, and separate attention maps are computed. By subtracting these maps, issues such as imbalanced token importance in visual contexts (e.g., boundary regions or small structures) are addressed, while redundant attention is removed. The refined attention weights are combined with the value vector, reducing irrelevant context and focusing on critical structures, including fine details like edges. By combining FEA and DiffAtt, DSEB improves feature detail, reinforces localized boundaries while highlighting salient context, surpassing irrelevant regions, and improves segmentation of intricate structures, achieving better overall accuracy.

2.2 Contextual Feature Attention Module (CFAM)

The CFAM, a Transformer-based decoder, improves hierarchical feature processing in CENet while avoiding over-enhanced representations. It refines features through four interconnected components for coherent information flow. First, the **Channel Calibration Unit (CCU)** processes the input $F \in \mathbb{R}^{H \times W \times C}$ to recalibrate channel-wise features, enhancing their capacity to capture diverse characteristics for precise segmentation. The CCU employs Global Multi-Aspect Pooling (G-MASP), combining average pooling ($\mathcal{P}_{\text{avg}}(F)$), max pooling ($\mathcal{P}_{\text{max}}(F)$), and standard deviation pooling ($\mathcal{P}_{\text{std}}(F)$) to generate a global descriptor $\mathbf{g} \in \mathbb{R}^{3C}$. This descriptor drives a channel-wise attention mechanism that reduces the channel capacity to C and adaptively reweights feature maps, prioritizing significant channels for improved diversity and representation. The CCU's formulation is given as:

$$\begin{aligned} \mathbf{g} &= [\mathcal{P}_{\text{avg}}(F); \mathcal{P}_{\text{max}}(F); \mathcal{P}_{\text{std}}(F)], \\ \mathbf{s} &= \sigma\left(f_{1 \times 1}\left(\text{GELU}\left(f_{1 \times 1}(\mathbf{g})\right)\right)\right), \quad F' = F \odot \mathbf{s}, \end{aligned} \tag{3}$$

where $\mathbf{s} \in \mathbb{R}^C$ is weights, $F' \in \mathbb{R}^{H \times W \times C}$ is the reweighted output, $\odot$ is point-wise multiplication, $f_{1 \times 1}$ is point-wise convolution, $[;]$ is concatenation, and σ is sigmoid.

Following the CCU, CFAM utilizes the **Multi-scale Contextual Aggregator (MCA)** to refine F' by capturing spatial context across multiple scales. The input $F' \in \mathbb{R}^{C \times H \times W}$ is split into four parts: $F'_i \in \mathbb{R}^{C_i \times H \times W}$, where F'_1, F'_2, F'_3 share equal channel dimensions ($C_1 = C_2 = C_3$), and F'_4 contains the less than 10 percent of all the channels, ensuring $C_1 + C_2 + C_3 + C_4 = C$. Most channels are allocated to the first three branches for the convolution operations, while the fourth branch handles global patterns via average pooling. The splits F'_1, F'_2, F'_3 are processed with parallel dilated depth-wise convolutions (f_{dk}) using dilation rates (e.g., 3, 5, and 8), respectively, while F'_4 undergoes average pooling. The outputs are concatenated and refined through a $f_{1 \times 1}$ convolution with SiLU activation. A context-aware gating mechanism then adjusts feature importance, suppressing redundancy and emphasizing salient features, resulting in robust multi-scale representations for improved segmentation. The MCA is expressed as:

$$
\begin{aligned}
F_{\text{cat}} &= \left[f_{d_3}(F'_1), f_{d_5}(F'_2), f_{d_8}(F'_3), \mathcal{C}_{avg}(F'_4) \right], \\
F_{\text{MCA}} &= \left(\text{SiLU}(f_{1 \times 1}(F_{\text{cat}})) \odot \text{SiLU}(f_{1 \times 1}(F')) \right) + F,
\end{aligned}
\tag{4}
$$

where $\mathcal{C}_{avg}$ represrnts channel-wise average pooling, f_{d_k} denotes the depth-wise convolution operator with a kernel size of $k \times k$ and dilation rate $d_k \in \{6, 8, 12\}$.

To adjust feature enhancement based on the accumulation of the noise from overly enhanced representations that come from previous layers and corresponding DSEB blocks after feature aggregation, the **weighted Non-local Block (wNLB)** prevents noise accumulation by modelling long-range spatial dependencies and adaptively denoising features while preserving critical details. It uses a self-attention mechanism with a learnable weighting parameter as a specific instance of non-local (NL) operations [28]. The final stage of the CFAM uses an enhanced MLP block, equipped with **Spatial Calibration Module (SRM)**. SRM applies Channel-wise Multi-Axis Spatial Pooling (C-MASP) using parallel pooling ($\mathcal{C}_{avg}, \mathcal{C}_{max}, \mathcal{C}_{std}$) to create a spatial descriptor $G \in \mathbb{R}^{3 \times H \times W}$. This descriptor is processed by parallel point-wise convolution that captures pixel-wise relations, and depth-wise convolution (f_k^{dw}) captures neighborhood interactions. The combined output (S) recalibrates the feature map via element-wise multiplication, then passes through a linear layer in the enhanced MLP for rich feature representations. Formally:

$$
\begin{aligned}
S &= \sigma(f_{1 \times 1}(G) + f_k^{dw}(G)), \quad G = \left[\mathcal{C}_{avg}(F_{\text{MCA}}), \mathcal{C}_{max}(F_{\text{MCA}}), \mathcal{C}_{std}(F_{\text{MCA}}) \right], \\
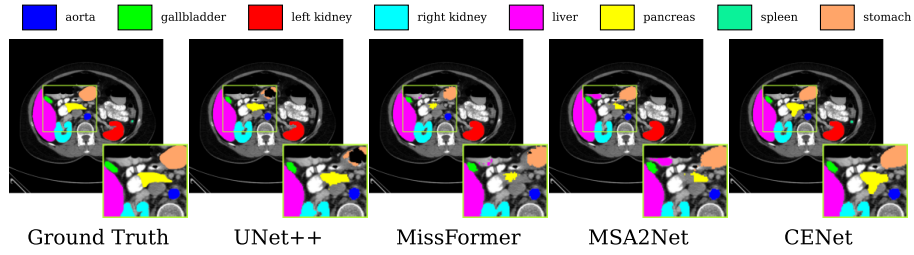F_{\text{recal}} &= F_{\text{MCA}} \odot S.
\end{aligned}
\tag{5}
$$

Fig. 2: Visual comparison of the proposed method versus others on the Synapse dataset.

3 Experiments and Results

3.1 Datasets and Implementation Details

The performance of CENet was evaluated on four datasets. The model, developed in PyTorch and trained on an NVIDIA A100 GPU (80GB), used the ImageNet-pretrained PVTv2-b2 encoder [27] at a 224 × 224 input resolution. For the *Synapse* dataset (30 CT scans), 18 scans were used for training and 12 for validation [15], following TransUNet's protocol [5]. Training involved 250 epochs, a batch size of 16, and an SGD optimizer with a 0.05 learning rate. On the *ACDC* [3] dataset (100 cardiac MRI scans), the split was 70 training, 10 validation, and 20 testing cases, with 150 epochs, a batch size of 12, and an Adam optimizer (learning rate 0.0001). For skin lesion segmentation, the *PH2* dataset (80 samples) and *HAM10000* (10,015 images) were trained for 100 epochs with a batch size of 16 and an Adam optimizer (learning rate 0.0001), using preprocessing/augmentation from [1]. The model also integrates BDoU Loss [24].

3.2 Results

Radiology: The performance of CENet on radiological datasets was evaluated, with Table 1 presenting the quantitative results on the Synapse dataset using DSC and HD metrics. Our approach significantly outperforms existing CNN-based methods. CENet also shows enhanced learning capabilities compared to Hybrid models, with improvements of 0.29% over MSA^{2}Net, respectively. These results underscore CENet's ability in segmenting various organs. Figure 2 provides a visual representation of CENet's performance in segmenting various organs, demonstrating CENet's accuracy in multi-scale segmentation of the kidneys, pancreas, and stomach. Furthermore, Table 2 emphasizes the effectiveness of our method against SOTA approaches on the ACDC dataset for cardiac segmentation in MRI images. CENet achieves the highest average DICE score of 92.18%. Moreover, CENet excels in all three organ segmentation tasks, regardless of morphological variations.

Table 1: Evaluation results on the Synapse dataset (blue indicates the best and red the second best results).

| Methods	Params	FLOPs	Spl.	RKid.	LKid.	Gal.	Liv.	Sto.	Aor.	Pan.	Average DSC↑	HD95↓
R50 U-Net [5]	30.42 M	-	85.87	78.19	80.60	63.66	93.74	74.16	87.74	56.90	74.68	36.87
TransUNet [5]	96.07 M	88.91 G	85.08	77.02	81.87	63.16	94.08	75.62	87.23	55.86	77.49	31.69
Swin-UNet [4]	27.17 M	6.16 G	90.66	79.61	83.28	66.53	94.29	76.60	85.47	56.58	79.13	21.55
HiFormer-B [9]	25.51 M	8.05 G	90.99	79.77	85.23	65.23	94.61	81.08	86.21	59.52	80.39	14.70
VM-UNet [21]	44.27 M	6.52 G	89.51	82.76	86.16	69.41	94.17	81.40	86.40	58.80	81.08	19.21
PVT-EMCAD-B2 [19]	26.76 M	5.60 G	92.17	84.10	88.08	68.87	95.26	83.92	88.14	68.51	83.63	15.68
MSA^{2}Net [13]	112.77 M	15.56 G	92.69	84.24	88.30	74.35	95.59	84.03	89.47	69.30	84.75	13.29
2D D-LKA Net [2]	101.64 M	19.92 G	91.22	84.92	88.38	73.79	94.88	84.94	88.34	67.71	84.27	20.04
CENet (Ours)	33.39 M	12.76 G	94.58	86.33	91.12	68.10	96.19	83.89	89.56	70.54	85.04	8.84
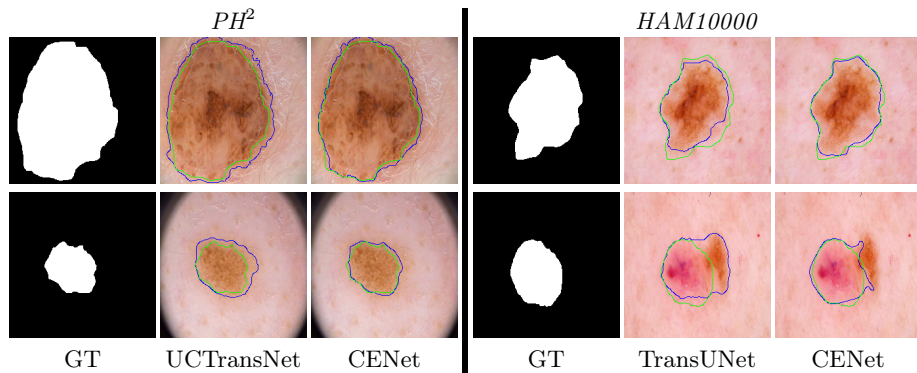
Fig. 3: Qualitative comparison of CENet and previous methods across skin benchmarks.

Dermoscopy: Table 2 evaluates our proposed network on two skin lesion segmentation datasets using DSC and ACC metrics. CENet outperforms CNN-based, Transformer-based, and hybrid methods, demonstrating superior performance and generalization. It surpasses Swin-Unet [16], U-Net [20], and UC-TransNet [25], achieving DSC score improvements of 0.58% and 1.45% on PH^2 and HAM10000 datasets, respectively. Moreover, Figure 3 showcases CENet's superiority in capturing intricate structures and producing precise boundaries through effective boundary integration.

Ablation Study. To evaluate CENet's components, ablation studies were conducted to assess their efficiency, performance, and placement within the DSEB and CFAM modules (Table 3). Results show that optimal performance occurs when the wNLB is positioned at the end of the MCA and CCU, while the DiffAttn and FEA operate in parallel within the DSEB, outperforming alternative configurations. Feature analyses (Figure 4) on the *ACDC* and *PH^2* datasets further demonstrate that in the first row, attention maps before applying the

Table 2: Evaluation results on the skin benchmarks (PH^2 and $HAM10000$) and $ACDC$ dataset.

| Methods	PH^2		$HAM10000$		Methods	aDSC	RV	MYO	LV
	DSC	ACC	DSC	ACC					
U-Net [20]	89.36	92.33	91.67	95.67	R50+UNet [5]	87.55	87.10	80.63	94.92
TransUNet [5]	88.40	92.00	93.53	96.49	R50+AttnUNet [5]	86.75	87.58	79.20	93.47
Swin-Unet [4]	94.49	96.78	92.63	96.16	TransUNet[5]	89.71	88.86	84.53	95.73
DeepLabv3+ [6]	92.02	95.03	92.51	96.07	ViT+CUP [5]	81.45	81.46	70.71	92.18
Att-UNet [18]	90.03	92.76	92.68	96.10	Swin-UNet [4]	90.00	88.55	85.62	95.83
UCTransNet [25]	90.93	94.08	93.46	96.84	R50+ViT+CUP [5]	87.57	86.07	81.88	94.75
MissFormer [11]	85.50	90.50	92.11	96.21	MT-UNet [26]	90.43	86.64	89.04	95.62
					MISSFormer [11]	90.86	89.55	88.04	94.99
CENet	95.04	97.19	94.71	97.04	**CENet**	92.18	90.90	89.63	95.99
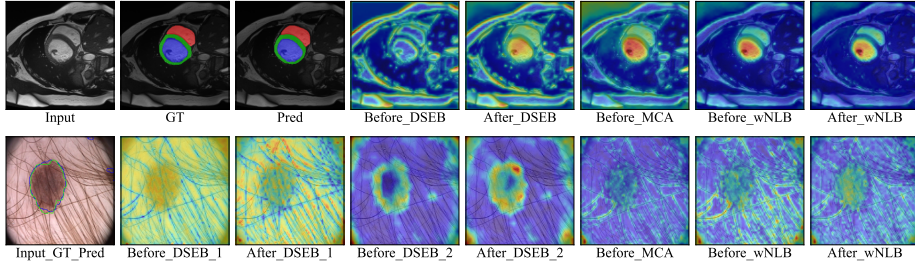
Fig. 4: Feature visualization in CENet: first row shows an $ACDC$ sample, second row a PH^2 example

DSEB are diffuse, failing to focus on key regions. After applying the DSEB, attention becomes more focused, and the wNLB refines MCA multiscale recalibrated feature maps by suppressing irrelevant details and emphasizing salient regions. In the second row, the DSEB enhances feature focus at different CENet levels, while the wNLB acts as a denoising mechanism, suppressing over-attended details (e.g., hairs) and directing attention to critical regions.

Table 3: Effect of different components of CENet on PH^2 dataset. #FLOPs are reported in (G), and P indicates the Parameters in Millions (M). All results are averaged over three runs. The best results are shown in bold.

Components				FLOPs (G)	Params (M)	Performance (DICE)
DSEB		CFAM				PH^2
FEA	DiffAtt	wNLB	CCU			
No	No	No	No	7.53	29.86	94.08
Yes	No	Yes	No	9.22	31.41	94.27
Yes	Yes	No	No	11.16	31.83	94.42
Yes	Yes	Yes	No	12.84	33.37	94.79
Yes	Yes	Yes	Yes	12.84	33.39	**95.04**

4   Conclusion

The proposed CENet improves context-aware medical image segmentation with a DSEB for boundary-sensitive and context amplification and a CFAM for multiscale representation and reduces feature redundancy. Evaluations demonstrate CENet's superior accuracy and boundary preservation over SOTA methods, providing a robust solution for complex segmentation tasks.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. Azad, R., Niggemeier, L., Huttemann, M., Kazerouni, A., Aghdam, E.K., Velichko, Y., Bagci, U., Merhof, D.: Beyond self-attention: Deformable large kernel attention for medical image segmentation (2023)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision Workshops(ECCVW) (2022)
5. 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)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
7. 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)
8. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. Computational visual media **9**(4), 733–752 (2023)

9. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6202–6212 (2023)

10. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging pp. 1–1 (2022). https://doi.org/10.1109/TMI.2022.3230943

11. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging (2022)

12. Jamil, S., Jalil Piran, M., Kwon, O.J.: A comprehensive survey of transformers for computer vision. Drones **7**(5), 287 (2023)

13. Kolahi, S.G., Chaharsooghi, S.K., Khatibi, T., Bozorgpour, A., Azad, R., Heidari, M., Hacihaliloglu, I., Merhof, D.: Msa^{2}net: Multi-scale adaptive attention-guided network for medical image segmentation. arXiv preprint arXiv:2407.21640 (2024)

14. Kuang, H., Wang, Y., Liang, Y., Liu, J., Wang, J.: Bea-net: Body and edge aware network with multi-scale short-term concatenation for medical image segmentation. IEEE Journal of Biomedical and Health Informatics **27**(10), 4828–4839 (2023). https://doi.org/10.1109/JBHI.2023.3304662

15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)

16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

17. Ma, Q., Mao, K., Wang, G., Xu, L., Zhao, Y.: Lcaunet: A skin lesion segmentation network with enhanced edge and body fusion. arXiv preprint arXiv:2305.00837 (2023)

18. 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)

19. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024)

20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

21. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)

22. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis **88**, 102802 (2023)

23. Sohn, K., Hao, Y., Lezama, J., Polania, L., Chang, H., Zhang, H., Essa, I., Jiang, L.: Visual prompt tuning for generative transfer learning (2022)

24. Sun, F., Luo, Z., Li, S.: Boundary difference over union loss for medical image segmentation (2023)

25. 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, pp. 2441–2449 (2022)
26. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 2390–2394. IEEE (2022)
27. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
28. 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)
29. Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential transformer. arXiv preprint arXiv:2410.05258 (2024)