Dbahnet: Dual-Branch Attention-Based Hybrid Network for High-Resolution 3d Micro-Ct Bone Scan Segmentation

Amine Lagzouli, Peter Pivonka, David Ml Cooper, Vittorio Sansalone, Alice Othmani

▶ **To cite this version:**

Amine Lagzouli, Peter Pivonka, David Ml Cooper, Vittorio Sansalone, Alice Othmani. Dbahnet: Dual-Branch Attention-Based Hybrid Network for High-Resolution 3d Micro-Ct Bone Scan Segmentation. 2024 IEEE International Symposium on Biomedical Imaging (ISBI), May 2024, Athens, Greece. pp.1-5, ⟨10.1109/ISBI56570.2024.10635241⟩. ⟨hal-04761185⟩

HAL Id: hal-04761185

https://hal.science/hal-04761185v1

Submitted on 31 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635241

DBAHNET: DUAL-BRANCH ATTENTION-BASED HYBRID NETWORK FOR HIGH-RESOLUTION 3D MICRO-CT BONE SCAN SEGMENTATION

Amine Lagzouli[⋆†] *Peter Pivonka*[⋆] *David ML Cooper*[‡]
Vittorio Sansalone [†] *Alice Othmani*[§]

[⋆] School of Mechanical, Medical, and Process Engineering, Queensland University of Technology, Brisbane, Australia
[†] Univ Paris Est Creteil, Univ Gustave Eiffel, CNRS, UMR 8208, MSME, F-94010 Créteil, France
[‡] Department of Anatomy, Physiology, and Pharmacology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[§] Université Paris-Est Creteil (UPEC), LISSI, Vitry sur Seine 94400, France

ABSTRACT

The precise segmentation of cortical and trabecular bone compartments in high-resolution micro-computed tomography (μCT) scans is crucial for evaluating bone structure and understanding how different medical treatments and mechanical loadings affect bone morphology, offering valuable insights into osteoporosis. In this work, we propose a novel hybrid neural network architecture named Dual-Branch Attention-based Hybrid Network (DBAHNet) for 3D μCT segmentation. DBAHNet combines both transformers and convolution neural networks in a dual-branch fashion and fuses their respective information at each hierarchical level, to better capture long-range dependencies and local features, and for a better understanding of the contextual representation. We train and evaluate DBAHNet on three datasets of high-resolution (<5μm) μCT scans of mouse tibiae. The results show that the proposed DBAHNet achieves state-of-the-art performance by surpassing several popular architectures. Our model also achieves a precise segmentation of the cortical and trabecular bone compartments along different regions of the bone, demonstrating a comprehensive understanding of the bone. Models and code are available at GitHub.

Index Terms— High-Resolution, Micro-Computed Tomography (μCT), 3D Segmentation, Hybrid Network, Attention Mechanism

1. INTRODUCTION

Osteoporosis represents a significant healthcare challenge, arising from an imbalance in the natural processes of bone remodeling, where bone resorption surpasses bone formation. Imaging techniques measure a variety of bone characteristics, and tracking them over time helps researchers understand disease progression and the efficacy of drug treatments.

Manual segmentation of high-resolution μCT bone scans is labor-intensive, often requiring several hours for a single scan due to image size and precision demands, especially in the metaphysis or when delineating regions with increased bone porosity or treatment-induced variations.

Recently, deep learning has revolutionized medical segmentation extending across various imaging modalities. This success is largely attributed to the use of architectures based on Convolutional Neural Networks (CNNs), such as U-Net [1], that excel at capturing local features and dependencies between pixels.

The attention mechanism is also widely used for medical image segmentation, particularly due to its ability to filter the feature maps. For instance, the Squeeze-and-Excitation block was introduced in [2], which emphasized relevant channels in feature maps by investigating their inter-dependencies channel-wise, while ignoring irrelevant ones. Similarly, SA-UNet [3] used a spatial attention that filters the spatial context of relevant features in the three-dimensional space. Alternatively, [4] presented the attention gates that filter the feature maps generated in the encoder and transmitted through skip connections to the decoder.

Furthermore, transformers are increasingly popular for use in computer vision tasks, due to their ability to understand long-range dependencies within a 3D scan. The Vision Transformer (ViT) was introduced in [5], marking the first fully transformer-based architecture to achieve state-of-the-art performance. To improve efficiency, Swin Transformers [6] employ shifted windows for enhanced global attention by applying self-attention to non-overlapping windows and enabling cross-window connections in subsequent layers.

There is growing interest in hybrid architectures that combine transformers and CNNs, such as UNETR [7] and

Authorized licensed use limited to: University of Queensland. Downloaded on October 30,2024 at 23:10:03 UTC from IEEE Xplore. Restrictions apply.

SwinUNETR [8]. The hybrid architectures excel at capturing both global and local context within a 3D scan, providing a more comprehensive data representation. Similarly, TransUNet [9] integrated transformers into the bottleneck of a U-Net architecture, focusing on capturing the long-range dependencies of high-level features.

Despite the effective application of hybrid networks in various studies for medical imaging segmentation, challenges persist when dealing with high-resolution µCT scans characterized by detailed anatomical structures. In such scenarios, the delineation of classes is complicated due to the complex topological features, further increased by the substantial computational resources required to process the large scan sizes.

In this paper, we propose a novel architecture called Dual-Branch Attention-based Hybrid Network (DBAHNet) for high-resolution 3D µCT scan segmentation, which consists of an hybrid combination of transformers and CNNs, fusing both their respective feature maps in each hierarchical layer, and integrating channel and spatial attention within the convolution blocks. Our contributions can be summarized as follows: (1) We employ a dual-branch setup in the encoder and decoder, integrating convolutional and transformer architectures for local and global context, which is further enhanced with channel-wise attention in the encoder and spatial-wise attention in the decoder. (2) We combine both branches feature maps using a feature fusion module (TCFFM) that merges and encodes their information, before setting them to the next hierarchical level. (3) State-of-the-art results on high-resolution (<5µm) 3D µCT bone scans of mouse tibia segmentation dataset, along with publicly available implementation source code via GitHub.

2. PROPOSED APPROACH

2.1. Overall architecture

Our proposed architecture Dual-Branch Attention-based Hybrid Network (DBAHNet, see Fig . 1) features a dual-branch hybrid design that incorporates both CNNs and transformers in both the encoder and decoder pathways.

Initially, a Patch Embedding block project the 3D scan into an embedding space C = 96, using successive convolutions and resulting in a lower-dimensional patch embedding $(C, \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4})$, where H, W, and D are the height, width and depth of the input 3D scan. This embedding serve as the input in parallel to the transformer and convolutions branches consisting of three hierarchical levels. Each level comprises of two sequential Swin transformer blocks in the transformer branch, and a Channel Attention-based Convolution Module (CACM, see Fig. 2.a) in the convolution branch. The output of each level in both branches are fused and further encoded with the Transformer-Convolution Feature Fusion Module (TCFFM, see Fig. 2.c), where down-sampling is performed for subsequent use in the next layer.

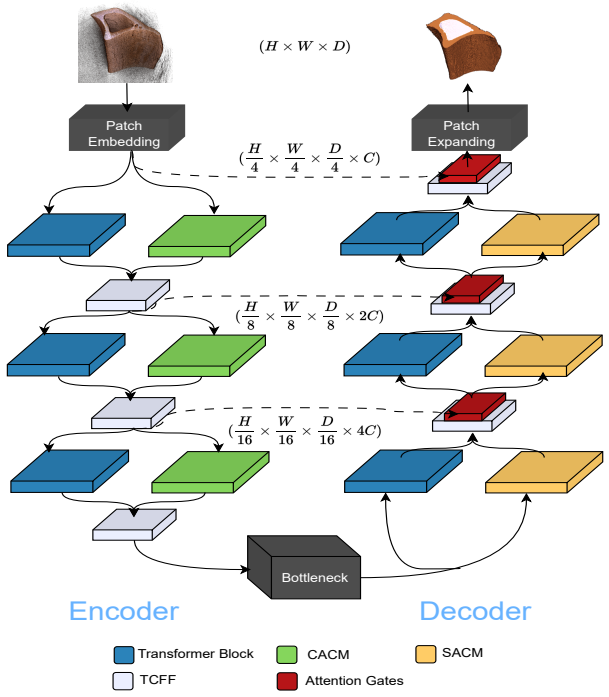

Fig. 1. Global Architecture of the Dual-Branch Attention-based Hybrid Network (DBAHNet)

The resulting feature maps from the encoder of size $(8C, \frac{H}{32} \times \frac{W}{32} \times \frac{D}{32})$ are sufficiently down-scaled and are forwarded to the bottleneck. The bottleneck performs global attention and aggregate information from the entire encoded feature maps for the decoder.

Similarly, the decoder is symmetrically mirroring the encoder, with the Spatial Attention-based Convolution Module (SACM, see Fig. 2.b) instead of the CACM, which enhances the relevant spatial features. The feature maps from both branches are fused and encoded using the TCFFM module, which performs up-sampling in the decoder to restore the original volume size. Along the decoder, feature maps from all the layers are filtered using attention gates and the residual skip connections from the encoder. Finally, the feature maps are further decoded with the Patch Expanding block to reconstruct the segmentation masks. In the following sections, each component of the DBAHNet will be described in details.

2.2. Transformer block

We utilize 3D-adapted SWIN Transformers for feature map processing at multiple scales, which aims to capture global long-range dependencies within the volume. Each transformer block consists of two transformers. The first transformer is the Local Volume Multi-Head Self-Attention

Authorized licensed use limited to: University of Queensland. Downloaded on October 30,2024 at 23:10:03 UTC from IEEE Xplore. Restrictions apply.

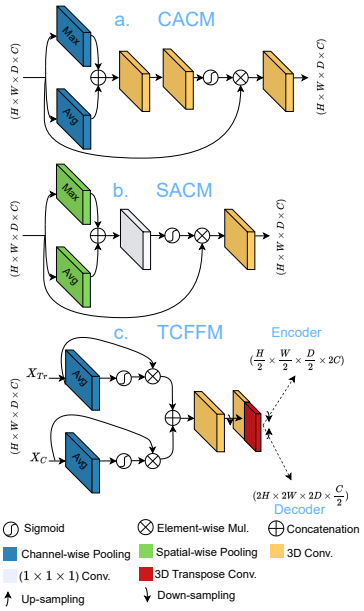

Fig. 2. a) **CACM**: Channel Attention-based Convolution Module, b) **SACM**: Spatial Attention-based Convolution Module, and c) **TCFFM**: Transformer-Convolution Feature Fusion Module.

block (LV-MHSA) that employs regular volume partitioning. The second transformer is a Shifted version of LV-MHSA, denoted as SLV-MHSA, which uses shifted partitioning for enhanced layer-to-layer connectivity. The Transformer block for a layer l can be expressed as

$$\begin{aligned} \hat{x}^{l} &= \text{LV-MHSA}(\text{LN}(x^{l-1})) + x^{l-1} \\ x^{l} &= \text{MLP}(\text{LN}(\hat{x}^{l})) + \hat{x}^{l} \\ \hat{x}^{l+1} &= \text{SLV-MHSA}(\text{LN}(x^{l})) + x^{l} \\ x^{l+1} &= \text{MLP}(\text{LN}(\hat{x}^{l+1})) + \hat{x}^{l+1} \end{aligned} \tag{1}$$

Where LN stands for Layer Normalization, and MLP stands for a GeLU-activated Multi Layer Perceptron.

The self-attention is computed as follows

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^{T}}{\sqrt{K}})V \tag{2}$$

Where Q, K, and V represent queries, keys, and values respectively, and K represents the dimension of the key and query.

2.3. Channel Attention-based Convolution Module

In the encoder, we employ a Channel Attention-based Convolution Module (CACM, see Fig. 2.a) that enhances cross-channel interaction. We first apply global average pooling, followed by two GeLU-activated 3D convolutions to create an attention map. This map modulates the initial feature map through element-wise multiplication. A final 3D convolution further encodes the output for use in subsequent layers.

2.4. Spatial Attention-based Convolution Module

In the decoder, we employ the Spatial Attention-based Convolution Module (SACM, see Fig. 2.b) for focused reconstruction of the segmentation mask, enhancing the salient features and aiding in the preservation of detailed structures. We first apply max-pooling and average-pooling, concatenate the results, and use a $1 \times 1 \times 1$ convolution to create an attention map. The input feature map is modulated by the attention map, and further processed by a final 3D convolution.

2.5. Transformer Convolution Feature Fusion Module

In the Transformer Convolution Feature Fusion Module (TCFFM, see Fig. 2.c), the feature maps obtained from both the transformer and convolution branches, denoted as x_{Tr} and x_{C}, are fused at each hierarchical level. Initially, a channel-wise average pooling is applied to x_{Tr} and x_{C} to extract a representative value for each channel of the feature maps followed by a sigmoid function, generating an attention mask that filters the channels. Subsequently, the results are concatenated and encoded through a 3D convolution layer. The resulting feature maps are then either down-sampled in the encoder, or up-sampled in the decoder.

2.6. Bottleneck

Having reduced the dimensionality of the resulting feature maps with the encoder, we employ a series of four Global 3D transformer blocks in the bottleneck. Global 3D transformer blocks perform global attention over all the downsampled feature maps. They excel at aggregating information from the entire feature map, which allows an understanding of the global context and offers a comprehensive representation to the decoder.

3. EXPERIMENTS AND RESULTS

Table 1. Performance comparison of the proposed method on the tibia μCT test dataset (C : Cortical, T: Trabecular).

Methods	Dice score			HD95 (mm)
	Avg	C	T	
3D-UNet	0.901	0.905	0.896	0.412
Att-UNet	0.951	0.963	0.938	0.193
UNETR	0.966	0.983	0.949	0.113
Swin-UNet	0.973	0.990	0.957	0.050
DBAHNet	**0.984**	**0.991**	**0.977**	**0.019**

Authorized licensed use limited to: University of Queensland. Downloaded on October 30,2024 at 23:10:03 UTC from IEEE Xplore. Restrictions apply.

Table 2. Performance evaluation of the proposed method at different bone regions: Middle (50%), Proximal-75% (75%), and Proximal-85% (85%).

Methods	Avg DSC			DSC Cortical			DSC Trabecular		
	50%	75%	85%	50%	75%	85%	50%	75%	85%
3D-UNet	0.909	0.901	0.882	0.923	0.889	0.885	0.896	0.914	0.878
Attention 3D-UNet	0.958	0.966	0.914	0.977	0.969	0.932	0.939	0.965	0.895
UNETR	0.971	0.973	0.953	0.989	0.985	0.980	0.953	0.962	0.926
Swin-UNet	0.976	0.975	0.972	0.993	0.990	0.990	0.958	0.960	0.953
DBAHNet (Ours)	**0.986**	**0.983**	**0.983**	**0.993**	**0.990**	**0.992**	**0.979**	**0.977**	**0.974**

3.1. Dataset

The 3D μCT tibia dataset is based on three separate studies [10, 11, 12]. All samples feature tibiae obtained from C57BL/6 virgin female mice. For the scope of this research, we focused solely on the 74 control tibiae that were not subjected to any treatments in the referenced medical experiments. High-resolution μCT scans of these tibiae were performed using the SkyScan 1172 (SkyScan, Kontich, Belgium) with a resolution of 4.8 μm - 5 μm. These scans were manually segmented following standard guidelines [13].

3.2. Implementation Details

All training and experiments were conducted with 4 NVIDIA V100 32GB GPUs. We employed the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.99, a batch size of 2, and a cosine annealing learning rate scheduler starting at 10^{-4}. The input scans are randomly cropped into (320, 320, 32) subvolumes and subjected to data augmentation, including flipping on all axes, random contrast adjustments, noise removal, and normalization. We evaluated the performances of our model using Sørensen-Dice score coefficient (DSC) and the 95th percentile of the Hausdorff distance (HD95). We used a Dice Cross Entropy loss function for the training.

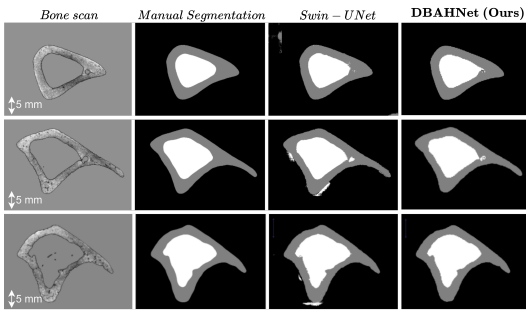

Fig. 3. Segmentation results at different bone regions. Middle-50% at the top, Proximal-75% at the middle, and Proximal-85% at the bottom.

3.3. Quantitative and qualitatives results

The model achieved state-of-the-art performances with an average DSC of 98.40%, a DSC of 99.12% for the cortical bone and a DSC of 97.68% for the trabecular bone prior to any post-processing. To further analyse the model's capability, we investigate the performance of the model in different bone regions within all the test dataset. Our model achieves an average DSC of 99.30 % on the Middle-50% region, an average DSC of 98.33% on the Proximal-75% region, and an average DSC of 98.28% on the Proximal-85% region.

We trained several popular architectures on our dataset and showed that our method surpasses CNNs and Transformer-based architectures, including 3D-UNet, Attention 3D-UNet, UNETR, and Swin-UNet. We surpass the state-of-the-art models on both cortical and trabecular compartment, as presented in Table 1. We also obtained an increase of the performance of the model over multiple regions of the bone (see Table. 2), proving the global understanding of the high-resolution 3D scan. We showcased the robustness of our model with a qualitative visualization of the μCT tibia segmentation results in Fig. 3, at different bone regions (Middle (50%), Proximal-75% and Proximal-85%).

4. CONCLUSION AND FUTURE WORK

We proposed a novel architecture, DBAHNet, which merges convolution and transformer outputs throughout all stages, harnessing their respective strengths in capturing short-range and long-range dependencies, for a robust contextual representation. The integrated channel and spatial attention mechanisms refine the model's performance by emphasizing relevant features. Our model demonstrated its proficiency by setting state-of-the-art results on 3D high resolution μCT tibia scans.

In future work, we intend to assess the architecture's robustness in segmenting tibiae subjected to varying medical treatments. Our goal is to develop a versatile model applicable to diverse studies for bone structure and morphology analysis. Additionally, we aim to evaluate DBAHNet's adaptability across different medical imaging modalities, to build a robust architecture for general purposes.

Authorized licensed use limited to: University of Queensland. Downloaded on October 30,2024 at 23:10:03 UTC from IEEE Xplore. Restrictions apply.

Acknowledgments

Mr Lagzouli acknowledges the support of a PhD scholarship from Queensland University of Technology. Profs Pivonka and Cooper gratefully acknowledge funding support from the Canadian New Frontiers in Research Fund Exploration (NFRFE). Prof. Pivonka also acknowledges funding support from the Australian Research Council (IC190100020, DP230101404).

Compliance with Ethical Standards

This research study was conducted retrospectively using animal data collected by the University of Bristol, Bristol, UK. All procedures complied with the UK Animals (Scientific Procedures) Act 1986 and were reviewed and passed by the ethics committee of The Royal Veterinary College (London, UK).

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "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*. Springer, 2015, pp. 234–241.

[2] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[3] Changlu Guo, Márton Szemenyei, Yugen Yi, Wenle Wang, Buer Chen, and Changqi Fan, "Sa-unet: Spatial attention u-net for retinal vessel segmentation," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 1236–1242.

[4] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[7] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[8] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[10] Toshihiro Sugiyama, Leanne K Saxon, Gul Zaman, Alaa Moustafa, Andrew Sunters, Joanna S Price, and Lance E Lanyon, "Mechanical loading enhances the anabolic effects of intermittent parathyroid hormone (1–34) on trabecular and cortical bone in mice," *Bone*, vol. 43, no. 2, pp. 238–248, 2008.

[11] Toshihiro Sugiyama, Lee B Meakin, Gabriel L Galea, Brendan F Jackson, Lance E Lanyon, Frank H Ebetino, R Graham G Russell, and Joanna S Price, "Risedronate does not reduce mechanical loading-related increases in cortical and trabecular bone mass in mice," *Bone*, vol. 49, no. 1, pp. 133–139, 2011.

[12] Toshihiro Sugiyama, Lee B Meakin, William J Browne, Gabriel L Galea, Joanna S Price, and Lance E Lanyon, "Bones' adaptive response to mechanical loading is essentially linear between the low strains associated with disuse and the high strains associated with the lamellar/woven bone transition," *Journal of bone and mineral research*, vol. 27, no. 8, pp. 1784–1793, 2012.

[13] Mary L Bouxsein, Stephen K Boyd, Blaine A Christiansen, Robert E Guldberg, Karl J Jepsen, and Ralph Müller, "Guidelines for assessment of bone microstructure in rodents using micro–computed tomography," *Journal of bone and mineral research*, vol. 25, no. 7, pp. 1468–1486, 2010.

Authorized licensed use limited to: University of Queensland. Downloaded on October 30,2024 at 23:10:03 UTC from IEEE Xplore. Restrictions apply.