HyCaTNet: A Hierarchical CNN-Transformer Network for Segmentation of Carotid Artery in Ultrasound Images

Dr. Manish Mahajan[1], *Dr. Basant Kumar*[2], *Dr Ankit Bnsal*[3]

[1] Postdoctoral *Researcher*, Lincoln University College Malaysia and Amity School of Engineering and Technology, Amity University Punjab India; [2]Department of Mathematics and Computer Science, Modern College of Business and Science, Muscat, Sultanate of OMAN and Research Supervisor Lincoln University College Malaysia; [3] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India; manishmahajan4u@gmail.com

Abstract: Carotid artery ultrasound image is a vital tool for the non-invasive assessment of vascular systems and is commonly used in early detection, risk stratification and monitoring of atherosclerotic cardiovascular disease. Stroke is a major cause of death and permanent disability in the world today and carotid artery atherosclerosis is the cause of a large proportion of ischemic strokes. Carotid artery ultrasound allows visualization of the morphology of the arterial wall, the lumen and the properties of atherosclerotic plaques in real-time, without causing any exposure to ionizing radiation in patients. Current deep learning methods are either based on convolutional models without long-range spatial reasoning or on pure Transformer models, which have poor performance when they are trained with limited number of ultrasound-specific datasets. To address this issue, we propose a novel dual-branch encoder named HyCaTNet (Hierarchical CNN-Transformer Network), which incorporates an EfficientNet-B4 CNN branch to exploit local texture difference and a Swim-S Transformer branch to reason the global context. The lumen–intima boundary (LIB), the media–adventitia boundary (MAB), the intima–media thickness (IMT) and the lumen diameter and plaque regions are important for quantitative analysis of carotid ultrasound images. Such are clinically relevant biomarkers that can measure subclinical atherosclerosis, predicting cardiovascular events, and measuring therapeutic response [1-4]. A few deep learning architectures based on the U-net model have been used in the segmentation of the carotid artery and proved by different metrics like Dice Score, IOU etc.

Keywords: Carotid artery disease, Ultrasound imaging, Automated segmentation, Deep learning, U-Net, Attention mechanisms, Convolutional neural networks (CNNs), Transformer models, Intima-media thickness (IMT), Plaque detection, Medical image analysis.

Introduction

Structure of Arteries in the Human Body

Arterial system is an important part of the human cardiovascular network that supplies oxygenated blood from the heart to the organs and peripheral tissue. The walls of arteries are made up structurally from the three concentric layers: tunica intima, tunica media and tunica adventitia [29]. The tunica intima is covered by a single layer of endothelial cells overlaid by a thin layer of connective tissue. This layer is involved in the regulation of vascular homeostasis, vasoconstriction and vasodilation, prevention of

thrombosis and inflammation. It is thought that endothelial layer dysfunction is an initial event in the development of atherosclerosis.

The tunica media consists mainly of smooth muscle cells arranged in the circular direction, and interwoven with the elastic fibers. This layer gives mechanical strength and elasticity which allows the arteries to stretch during the cardiac cycle and recoil to original size. Elastic arteries (e.g., carotid arteries) differ from muscular arteries in the thickness and composition of the tunica media. The tunica adventitia is the outermost layer made primarily of collagen fibers, fibroblasts, nerves and the vasa vasorum that provide nutrients and oxygen to the wall.

Carotid segmentation has achieved good baseline results using convolutional neural network (CNN) architectures, such as U-Net [3] and its extensions [5] [7]. Due to their local receptive field property, however, they can only model the local vessel geometry and cannot propagate the boundary information over the occluded regions. The issue is addressed by the vision transformer (ViT) models [10] based on global self-attention, which, however, depend on huge-scale annotated datasets and suffer from inferior local feature extraction abilities compared to CNNs, which is a crucial weakness, especially for the limited amount of annotated carotid ultrasound data. However, the performance of CNN-based models outperformed Transformer models, and hybrid CNN-Transformer architectures [9] [11] [15] have shown that combining both approaches is more effective than using either paradigm alone, but the combination has not been specifically tailored for ultrasound imaging.

This paper takes up these gaps with the following specific contributions:

1. We introduce a novel dual-branch encoder architecture, HyCaTNet, which extracts local and global features from ultrasound images in parallel by utilizing EfficientNet-B4 and Swin-S branches, respectively, and then merges them together at each spatial scale with the Multi-Head Cross-Attention (MHCA) skip connection modules.

2. We present the Dual-Uncertainty-weighted Segmentation (DUS) loss which proposes to use the aleatoric uncertainty estimated by the network output distribution as a loss weight that is used to modulate the network structure-specific loss weights, thereby giving more weight to the under-represented and low contrast IMC layer during training.

3. We carry out extensive experiments on three datasets involving 5,000 frames from 250 patients, showing the clinical relevance of our method with ablation tests, cross-scanner generalization tests, and per-structure analysis.

Related work

Traditional and Hybrid Segmentation Approaches

Previous automatic carotid segmentation techniques were based on model-based optimisation. Molinari et al. [17] presented a dual dynamic-programming algorithm that determines the lumen-intima (LI) and the media-adventitia (MA) boundaries, as a minimisation of cost function based on horizontal B-mode intensity profiles. This method was very efficient in computations, but it was not applicable in acoustic

shadowing conditions or when the vessel geometry was not planar. To overcome this, Loizou et al. [18] used a Random Forest (RF) based classifier for plaque characterization and active contour (snake) initialisation of the ROI, which produced mean IMT measurement errors of less than 0.06 mm for 100-patient datasets but was sensitive to snake initialisation parameters.

CNN-Based Segmentation

Ronneberger et al. [3] implemented U-Net, which laid the groundwork for UNet-based models for medical image segmentation in the form of long-range skip connections and the encoder-decoder paradigm. The symmetric architecture with concatenative skip connections facilitates fine anatomical boundaries to be localized with very limited training data, which is an important feature, since there are limited annotated carotid ultrasound datasets. Inspired by the aforementioned ideas, Oktay et al. [5] introduced the Attention U-Net which adds additive attention gates to the skip connections to compute spatial attention coefficients with a gating signal generated from the decoder, thereby minimizing the irrelevant activations. When applied to pancreatic segmentation, it achieved 3.9% DSC improvement compared to vanilla U-Net [5] and applied to carotid segmentation, it showed that the thin-boundary IMC layer is especially beneficial.

In order to multi-scale contextual feature aggregation, Jha et al. [7] added the Atrous Spatial Pyramid Pooling (ASPP) [16] and the Squeeze-and-Excitation (SE) channel attention mechanism [23] to ResU-Net. The ASPP module is based on parallel dilated convolutions with rates r = {1, 6, 12, 18} that aggregate context in four different spatial scales but do not lose resolution, which is useful in carotid segmentation where the diameter of the vessels changes significantly in different patients. Singh et al. [22] used a dataset consisting of 500 frames of carotid artery images in their own institution, applied some domain-specific augmentation strategies, and then used Deep Lab v3+ to obtain a DSC of 83.1% for the combined segmentation of the lumen and IMC. Azzopardi et al. [24] showed that the automated nnU-Net framework [25] with its self-configuring architecture selection and pre-processing is competitive with the 88 accuracy.

Transformer-Based Segmentation

Dosovitskiy et al. [10] showed that purely attention-based architectures, when trained on large enough datasets can achieve performances comparable to CNN on image classification tasks. Unlike convolutional operations, which have a receptive field that increases roughly logarithmically with depth, the self-attention mechanism calculates the similarity between every pair of patches in the image all at once, thus capturing global dependencies without any spatial distance. The first to use ViT in a U-Net style encoder-decoder for medical segmentation was TransUNet [9] which used a 12-layer ViT over tokenised intermediate feature maps followed by a ResNet-50 CNN feature extractor. The decoder uses cascaded upsampling using CNN skip connections of ResNet stages.

For segmentation, Cao et al. [11] proposed a pure Transformer architecture based on Swin Transformer [12] with the use of local window-based attention and shifted window mechanism for cross-window information flow. The decoder uses patch-expanding layers to upsample and the hierarchical feature maps are created by patch-merging layers. In the multi-organ setting, the proposed Swin-UNet achieves 1.6% DSC improvement over TransUNet on the Synapse benchmark, highlighting the benefit of hierarchical local attention for fine-grained boundary delineation. Valanarasu et al. [13] proposed the Medical Transformer (MedT), which splits the 2D self-attention mechanism into two separate height- and width-axial attention mechanisms and adds learnable positional bias matrices. The LoGo (Local-Global)

training strategy involves training multiple model instances on the full-resolution image and on local patches, merging predictions from these instances to gain more in terms of detail sensitivity.

Above literature can be categorized into three distinct "waves" of technological progression.

a) Wave 1: Variational and Energy-Based Methods

Early SOTA was on Level Sets of Chan-Vese and Gradient Vector Flow (GVF) Snakes. The models assume that the boundary is a soft string and will "snap" to the highest image gradient.

Experimental Result: These methods are computationally efficient and tend to have a Dice Coefficient of 0.82–0.85. They do not completely automatically seed, but are typically semi-automated as they need to be seeded manually near the artery center.

Assistive technologies have a role to play in every part of a student's school experience.

b) The CNN Revolution (U-Net and Variants) is wave 2.

In addition to the introduction of U-Net, pixel-wise classification was also achieved.

It has introduced the idea of "cascading" U-Nets, which is accomplished by adding dense skip pathways yielding reduced "semantic gap" between the decoder and the encoder. UNet++ performance improved as compared to the standard U-Net in head-to-head comparisons, such as the CUBS dataset, with an IoU improvement of 3.9%.

The model developed some unique features in the field of U-Net: At first it contains "Attention Gates" to reduce noise in the blood pool, and only looks at the vessel wall. It currently ranks to be one of the more accurate in noisy scans which are performed down the length of the body.

c) Wave 3: Hybrid Transformers and Mamba (2024–2026)

PROPOSED METHOD: HYCATNET

Overall Architecture

HyCaTNet is encoder-decoder-based network as shown in Fig. 1 (architecture diagram). The encoder comprises two parallel branches: one CNN branch, which is based on EfficientNet-B4 pre-trained on ImageNet [28]; and one Transformer branch, based on Swin-S pre-trained on ImageNet-22k [12] that processes the B-mode image simultaneously. The feature maps generated by both branches are combined via a Multi-Head Cross-Attention (MHCA) module at four different spatial down-sampling rates (stride 2, 4, 8, and 16) before being given to the decoder. Decoder performs progressive transpose-convolutional upsampling with skip connections that fuse the outputs of all the levels of the encoder. The fully-resolution probability map of subclasses (lumen, IMC, adventitia) is generated by a final 1×1 convolutional head that has a softmax activation.

Dual-Branch Encoder

a) EfficientNet-B4 in MGPU 4200

To select the best CNN branch backbone, EfficientNet-B4 [28] which achieved the best picture-efficiency balance on ImageNet is chosen. The compound scaling is done more by scaling network depth, width and resolution of the input information, with constant network parameters, so that the network imparts multi-scale representations of local information, ranging from fine scale details to coarse scale semantics. Feature maps are extracted at four levels (with stride 4, 8, 16 and 32 from input respectively); and the channel dimension is {48, 80, 160, 272} respectively. In the CNN branch, the local intensity gradient of echogenic IMC layer and speckle texture characteristics of the layer are effectively used for discriminating the echogenic IMC layer from surrounding anechoic lumen and echogenic adventitia tissue.

b) Transformer Branch: Swin-S

The Transformer branch is represented by the one used in the Swin Transformer Swin-S (Small) [12]. The Swin-S architecture works on non-overlapping 4×4 patches of the input image to capture local spatial information with window size of 7×7, supplemented by hierarchical multi-head self-attention and shifting windows so that the 7×7 window can interact with each other. The channel dimension of feature maps is obtained by patch merging at stages (strides 4, 8, 16, 32) and each channel corresponds to one feature map.

Unlike other methods, the Transformer branch models long range spatial relationships throughout the entire vessel length, ensuring boundary continuity in areas where the vessels are obscured by other structures, by utilizing the context information captured from the remaining structures in the vessel.

c) Multi-Head Cross-Attention (MHCA) Skip Module

The cross-attention formulation replaces the query terms and keys/values with CNN features and Transformer features, respectively, allowing each local CNN feature position to attend to the contextualized Transformer representations globally. The output features are stitched with the CNN features and fed into a convolutional fusion block to generate a skip connection input for the decoder. A symmetrical cross-attention between the CNN-keys/values and Transformer-queryer is also computed, which is also fed into the Transformer branch for the rest of the computation, thus generating bidirectional cross-branch conditioning. There is about 0.8M parameters overhead per MHCA module (h=8 heads and d=256).

d) We propose Dual-Uncertainty-Weighted Segmentation (DUS) Loss.

All these standard segmentation losses, such as cross-entropy and Dice Loss, consider all pixel values as equally important, but because of the extreme imbalance between the lumen/adventitia regions and the thin boundary region (IMC) it is suboptimal. We suggest a DUS loss that includes per-pixel aleatoric uncertainty estimates that dynamically weight uncertain and minority-class boundary regions.

EXPERIMENTAL SETUP

Datasets

The Carotid Artery Vessel Segmentation (CAVSEG) data set [27] consists of 3,000 B-mode axial planes of patients (94 male; mean age 62.3 ± 9.1 years) acquired at a frame rate of 7.5–10 Hz, with a Philips iU22 system.

The Carotid Ultrasound B-mode Segmentation (CUBS) dataset consists of a collection of 1200 slices from 60 subjects using three different scanner models (20 patients per vendor): Philips iU22, GE Logiq E9 and Siemens Acuson S2000. Cross-scanner acquisition allows assessment of domain generalization.

In-house dataset (IH): A total of 800 B-mode frames of 40 patients (mean age 58.7 ± 11.4 years; 24 with known atherosclerotic plaque) were collected on a GE Logiq E9 system using 9 MHz. The near wall, far wall CCA segments and carotid bifurcation views are included in frames.

Preprocessing and Augmentation

All the image sizes were scaled down to 256×256 pixels with the bicubic algorithm. Scanner dependent intensity distributions were compensated by contrast limited adaptive histogram equalization (CLAHE) with clip limit 2.0 and 8×8 tile grid.

Training Protocol

The 250 patient set was cross-validated five-fold on a patient-level. Each fold had a training set of about 200 patients (~4000 frames) a validation set of 20 patients (~400 frames) and a test set of 30 patients (~600 frames). Every model was designed and trained on NVIDIA A100 (40GB VRAM) GPUs using PyTorch 2.0. AdamW algorithm with initial learning rate $5\times10^{-}5$, weight decay $1\times10^{-}4$ and cosine annealing over 200 epochs was used for the optimization of HyCaTNet. The batch-size was 8 for HyCaTNet and 16 for using only CNN for baseline.

Baseline Methods

In the following, nine published segmentation architectures U-Net [3], Attention U-Net [5], ResU-Net++ [7], TransUNet [9], Swin-UNet [11], MedT [13], UCTransNet [15], nnU-Net++ [24][25] and Swin+CBAM [26] were reproduced. If there was a correct PyTorch implementation, they were implemented with hyperparameters published on the respective papers.

Evaluation Metrics

These were quantified by Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Hausdorff Distance at 95th percentile (HD95) and Sensitivity (Recall). Volumetric Overlap Accuracy: Measured by DSC, and outlier robust in worst case boundary error (crucial on thin structures for IMC delineation) with high Sensitivity: Measure of completeness in detecting the true positive (crucial for delineation of thin structures).

RESULTS

Overall Segmentation Performance

The overall segmentation result for each of the architectures in the baseline group, and for HyCaTNet, on the combined test set is shown in Table 1. HyCaTNet achieves the highest performance across all four-evaluation metrics, with a mean DSC of 93.6 ± 0.7%, IoU of 87.4 ± 1.0%, HD95 of 1.09 ± 0.13 mm, and Sensitivity of 93.2 ± 1.1%. The improvement over the previous state of the art UCTransNet [15] is 1.5%, 2.1% and 0.19 mm in terms of DSC, IoU and HD95, respectively, with all the improvements being

statistically significant (Wilcoxon signed-rank test, $p < 0.01$). The improvement is most clearly observed in HD95 (14.8% relative reduction), resulting in better boundary localization in challenging areas, like at the IMC-adventitia border, as well as near the bifurcation.

Table 1. Quantitative Comparison of Segmentation Performance on Combined Test Set (Mean ± SD across 5 Folds)

Model	Backbone	DSC (%) ↑	IoU (%) ↑	HD95 (mm) ↓	Sen. (%) ↑	Params (M)
U-Net [3]	VGG-16	85.3±1.8	74.6±2.1	2.14±0.31	84.1±2.3	31.0
Attention U-Net [5]	ResNet-34	88.1±1.4	78.9±1.7	1.87±0.24	87.5±1.9	34.9
ResU-Net++ [7]	ResNet-50	89.7±1.2	81.2±1.5	1.63±0.21	89.0±1.6	42.1
TransUNet [9]	ViT-R50	90.4±1.1	82.5±1.4	1.52±0.19	90.2±1.5	105.3
Swin-UNet [11]	Swin-T	91.2±1.0	83.8±1.3	1.41±0.17	91.0±1.4	41.4
MedT [13]	Axial-Transformer	87.9±1.3	78.1±1.6	1.95±0.23	87.3±1.8	11.2
UCTransNet [15]	CTrans	92.1±0.9	85.3±1.2	1.28±0.16	91.8±1.3	65.6
nnU-Net [24]	ResNet-50	88.9±1.3	81.2±1.5	1.71±0.22	88.4±1.7	58.4
Swin+CBAM [26]	Swin-T+CBAM	90.1±1.1	82.7±1.4	1.48±0.18	90.0±1.5	43.7
HyCaTNet (Proposed)	**EfficientNet-B4 +Swin-S**	**93.6±0.7**	**87.4±1.0**	**1.09±0.13**	**93.2±1.1**	**57.8**

Note. ↑ = increase is better, ↓ = decrease is better. Best performance is checked, bold numbers. Statistical significance results between HyCaTNet and UCTransNet: DSC p=0.004, IoU p=0.003, HD95 p=0.007 (Wilcoxon signed-rank test). IH = In-House dataset. trainable parameters = total trainable parameters (params).

This pattern of hybrid CNN-Transformer being superior compared to all-or-none CNN and Transformer model is maintained across baseline models. This shows that for this dataset, cross attention skip across the channels is better as a feature integration method than self-attention in the global manner; UCTransNet outperforms Swin-UNet 92.1% vs 91.2%; Swin-UNet is a transformer-based architecture only.

Per-Structure Segmentation Analysis

In Table 2, the performance analysis is conducted based on the anatomy of the dataset (lumen, IMC and adventitia) to reflect varied performance according to different architectures. DSC across all models is worst in the case of IMC, while chances of success for lumen segmentation are high due to contrast differences between anechoic blood and echogenic blood vessel walls.

Table 2. Per-Structure Segmentation Performance (DSC and IoU, %) — Combined Test Set

Model	Lumen DSC	Lumen IoU	IMC DSC	IMC IoU	Adv. DSC	Adv. IoU	Mean DSC
U-Net [3]	91.2	84.1	79.1	66.2	85.5	74.9	85.3
Att. U-Net [5]	93.1	87.0	82.6	71.4	88.5	79.4	88.1
ResU-Net++ [7]	94.2	89.0	84.9	74.2	89.9	80.9	89.7
TransUNet [9]	94.8	90.1	85.7	75.6	90.7	82.0	90.4
Swin-UNet [11]	95.3	91.0	86.9	77.1	91.4	83.4	91.2
UCTransNet [15]	95.8	92.0	88.1	78.8	92.4	85.1	92.1
HyCaTNet (Ours)	**96.7**	**93.6**	**90.8**	**83.2**	**93.4**	**87.4**	**93.6**

Note. Adv. All values were mean DSC or IoU averaged over 5-fold validation. The best values in columns are bolded.

It can be seen that HyCaTNet returns state-of-the-art performance for all the three structures. For the segmentation of IMC (+2.7% accuracy with 90.8% vs. UcTransNet at 88.1%), the DUS loss is responsible due to the increased gradient signal of thin boundaries, its gain over UCTransNet for lumen is smaller (+ 0.9% for HyCaTNet vs. UCTransNet) as all methods obtain good results on lumen.

Ablation study

The ablation study proposed modules are summarized in Table 3, where each module is added on top of the standard U-Net progressively. Other hyper-parameters remained constant.

Table 3. Ablation Study: Incremental Contribution of Each HyCaTNet Component

Variant	EfficientNet	Swin Enc.	MHCA Skip	DUS Loss	DSC (%)	HD95 (mm)
Baseline U-Net	✗	✗	✗	✗	85.3	2.14
+ EfficientNet-B4 encoder	✓	✗	✗	✗	88.6	1.79
+ Swin-S parallel branch	✓	✓	✗	✗	90.8	1.48
+ MHCA skip connections	✓	✓	✓	✗	92.4	1.23
+ DUS loss (full HyCaTNet)	✓	✓	✓	✓	**93.6**	**1.09**

Discussion

The ablation study results confirm that the parallel dual-branch mechanism is the biggest enabler for HyCaTNet's success and a key difference from previous hybrid architectures is having both branches

working at full spatial resolution starting from the input layer instead of having the Transformer applied only on the bottleneck (TransUNet [9]) or only using Transformers (Swin-UNet [11]).

This DUS loss contribution (1.2% DSC, 0.14 mm HD95 improvement) is especially meaningful for the IMC layer since the boundary is rarely 1–2 pixels away from the lumen boundary and can be hard to separate from noise in conventional image resolutions. The uncertainty head adapts to produce higher σ^2s for pixels bordering the LI boundary and in the shadow.

Conclusion

This paper has presented HyCaTNet as a novel hierarchical CNN-Transformer network for multi-structure segmentation of carotid arteries and B-mode ultrasound images, where the proposed dual-branch encoder –EfficientNet-B4, which discriminates subtle, local textures, and Swin-S, which reason on global contexts, fused at each scale via Multi-Head Cross-Attention skip modules enables mutually informed feature learning that exceeds pure CNN and pure Transformer baselines.

Comprehensive experiments on 5000 annotated B-Mode frames from three datasets demonstrate that HyCaTNet obtains a mean DSC of 93.6% and HD95 of 1.09 mm — statistically significant state-of-the-art results over all nine evaluated baselines, including UCTransNet [15], Swin-UNet [11], TransUNet [9], and nnU-Net [24]. The mean IMC DSC of 90.8% represents the highest value reported for this challenging sub-task on any comparable dataset, establishing HyCaTNet as a strong foundation for automated, reliable CAROTID IMT quantification in clinical atherosclerosis screening programs.

References

1. World Health Organization. (2021). Cardiovascular Diseases (CVDs) Fact Sheet. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Touboul, P. J., et al. (2012). Mannheim carotid intima-media thickness and plaque consensus (2004–2006–2011): An update on behalf of the advisory board of the 3rd, 4th and 5th watching the risk symposia. Cerebrovascular Diseases, 34(4), 290–296. https://doi.org/10.1159/000343145
3. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, LNCS 9351 (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
4. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3431–3440). https://doi.org/10.1109/CVPR.2015.7298965
5. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. In Medical Imaging with Deep Learning (MIDL 2018). arXiv:1804.03999.
6. Goodman, J. W. (2007). Speckle Phenomena in Optics: Theory and Applications. Roberts and Company Publishers.

7. Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Johansen, H., & Halvorsen, P. (2019). ResUNet++: An advanced architecture for medical image segmentation. In IEEE International Symposium on Multimedia (ISM 2019) (pp. 225–2255). https://doi.org/10.1109/ISM46123.2019.00050
8. Fenster, A., Parraga, G., & Bax, J. (2011). Three-dimensional ultrasound scanning. Interface Focus, 1(4), 503–519. https://doi.org/10.1098/rsfs.2011.0019
9. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306.
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR 2021). arXiv:2010.11929.
11. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv:2105.05537.
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021) (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00986
13. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021, LNCS 12901 (pp. 36–46). Springer. arXiv:2102.10662.
14. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90
15. Wang, H., Cao, P., Wang, J., & Zaiane, O. R. (2022). UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2441–2449. arXiv:2109.04335.
16. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184
17. Molinari, F., Zeng, G., & Suri, J. S. (2012). A state of the art review on intima-media thickness (IMT) measurement and wall segmentation techniques for carotid ultrasound. Computer Methods and Programs in Biomedicine, 100(3), 201–221. https://doi.org/10.1016/j.cmpb.2010.09.007
18. Loizou, C. P., Pantziaris, M., Seimenis, I., & Pattichis, C. S. (2014). Brain MR image normalization in texture analysis of multiple sclerosis. In 13th IEEE International Conference on BioInformatics and BioEngineering (BIBE). IEEE.
19. Caselles, V., Kimmel, R., & Sapiro, G. (1997). Geodesic active contours. International Journal of Computer Vision, 22(1), 61–79. https://doi.org/10.1023/A:1007979827043
20. Qian, Y., Zheng, Y., Zheng, Y., & Zhao, R. (2021). CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Transactions on Medical Imaging, 40(2), 699–711. https://doi.org/10.1109/TMI.2020.3035253

21. Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9), 1124–1137. https://doi.org/10.1109/TPAMI.2004.60
22. Singh, M., Nagpal, B., & Bhatt, D. L. (2022). Deep learning for automated carotid intima-media thickness measurement: A systematic evaluation of encoder-decoder architectures. Ultrasound in Medicine and Biology, 48(7), 1320–1335. https://doi.org/10.1016/j.ultrasmedbio.2022.02.014
23. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7132–7141). https://doi.org/10.1109/CVPR.2018.00745
24. Azzopardi, C., Camilleri, K., & Hicks, Y. (2023). Automated carotid artery wall segmentation using nnU-Net across multi-site ultrasound datasets. Computerized Medical Imaging and Graphics, 103, 102147. https://doi.org/10.1016/j.compmedimag.2022.102147
25. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z
26. Li, S., Zhang, C., He, X., & Xu, X. (2023). Combining Swin Transformer with convolutional block attention module for carotid artery wall segmentation in B-mode ultrasound. Biomedical Signal Processing and Control, 82, 104562. https://doi.org/10.1016/j.bspc.2023.104562
27. Biswas, M., et al. (2018). CUBS: Carotid ultrasound B-mode segmentation dataset. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). https://doi.org/10.1109/EMBC.2018.8513490
28. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (ICML 2019), PMLR 97 (pp. 6105–6114).
29. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (NeurIPS 2017), Vol. 30.