PeerJ Computer Science

MCV-UNet: a modified convolution & transformer hybrid encoder-decoder network with multi-scale information fusion for ultrasound image semantic segmentation

Zihong Xu[1] and Ziyang Wang[2]

[1] Department of Mechanical Engineering, Columbia University, New York, United States of America
[2] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ABSTRACT

In recent years, the growing importance of accurate semantic segmentation in ultrasound images has led to numerous advances in deep learning-based techniques. In this article, we introduce a novel hybrid network that synergistically combines convolutional neural networks (CNN) and Vision Transformers (ViT) for ultrasound image semantic segmentation. Our primary contribution is the incorporation of multi-scale CNN in both the encoder and decoder stages, enhancing feature learning capabilities across multiple scales. Further, the bottleneck of the network leverages the ViT to capture long-range high-dimension spatial dependencies, a critical factor often overlooked in conventional CNN-based approaches. We conducted extensive experiments using a public benchmark ultrasound nerve segmentation dataset. Our proposed method was benchmarked against 17 existing baseline methods, and the results underscored its superiority, as it outperformed all competing methods including a 4.6% improvement of Dice compared against TransUNet, 13.0% improvement of Dice against Attention UNet, 10.5% improvement of precision compared against UNet. This research offers significant potential for real-world applications in medical imaging, demonstrating the power of blending CNN and ViT in a unified framework.

Subjects Artificial Intelligence, Computer Vision, Neural Networks
Keywords Ultrasound imaging, Vision transformer, Image semantic segmentation, Convolutional neural network

Submitted 14 February 2024
Accepted 30 May 2024
Published 24 June 2024

Corresponding author
Ziyang Wang,
ziyang.wang17@gmail.com

Academic editor
Syed Hassan Shah

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj-cs.2146**

INTRODUCTION

Persistent postsurgical pain, defined as pain lasting more than 3–6 months after surgery (*Merskey, 1986*), presents a significant challenge in patient care. Management strategies for postsurgical pain typically encompass symptom control and disease modification (*Kehlet, Jensen & Woolf, 2006*). In clinical practice, the focus often shifts toward symptom control, which primarily relies on the use of narcotics and inhibitors (*Baby & Jereesh, 2017*). The frequent use of narcotics is associated with a range of unwanted side effects, including respiratory depression, nausea, vomiting, and other opioid-related adverse events (*Bajwa & Haldar, 2015*). Moreover, increased narcotic usage has been

How to cite this article Xu Z, Wang Z. 2024. MCV-UNet: a modified convolution & transformer hybrid encoder-decoder network with multi-scale information fusion for ultrasound image semantic segmentation. *PeerJ Comput. Sci.* **10**:e2146 http://doi.org/10.7717/peerj-cs.2146

linked to extended hospital stays and heightened risk of depression (*Armaghani et al., 2016*). Some studies have suggested that indwelling catheters represent an alternative, safe, and effective method for postsurgical pain management (*Wijayasinghe et al., 2016*; *Sola et al., 2012*; *Pacik, Nelson & Werner, 2008*). However, the accurate placement of catheters is crucial, as incorrect placement can lead to unanticipated pain, opioid use, and potential complications, such as readmission or delayed hospital discharge (*Hauritz et al., 2019*). To address these challenges, various methods have been explored to enhance the precision of catheter placement and nerve location identification.

Nerve stimulation (NS) techniques have emerged to enhance the safety and precision of medical procedures, particularly in situations where traditional anatomic landmark techniques may lead to unintended punctures or cannulations (*Pham-Dang et al., 2003*; *Kick et al., 1999*; *Copeland & Laxton, 2001*). NS involves the stimulation of sensory nerves, inducing non-noxious sensations that effectively compete with and attenuate pain signals, thereby reducing pain perception. Additionally, NS has the potential to trigger the release of endorphins, natural pain-relieving chemicals, and modulate nerve activity implicated in pain signaling. Ultrasound technology has gained recognition as an alternative method to elevate the safety and quality of catheter placements in medical practice (*Chan et al., 2003*). Ultrasound techniques offer multifaceted advantages by enabling the visualization of nerve structures before injection, guiding the needle precisely to target nerves, and providing real-time visualization of the local anesthetic's dispersion pattern. Numerous studies have demonstrated the superior performance of ultrasound-guided techniques over traditional anatomic landmarks and NS methods (*Brass et al., 2015*; *Schnabel et al., 2013*). In response to the ongoing pursuit of enhanced nerve identification and catheter placement precision, more advanced techniques have been proposed to further optimize these procedures.

Deep learning-based networks for ultrasound segmentation have emerged as the predominant choice, delivering remarkable segmentation performance at the pixel level. Convolutional neural networks (CNNs) have demonstrated the efficient capacity for extracting intricate features from grid-like data (*Long, Shelhamer & Darrell, 2015*; *Ronneberger, Fischer & Brox, 2015*; *Huang et al., 2020*; *Oktay et al., 2018*; *Li et al., 2022*). The UNet architecture revolutionized CNN-based segmentation with its symmetric encoder–decoder design, enabling impressive results even with limited datasets (*Ronneberger, Fischer & Brox, 2015*; *Wang, Zhang & Voiculescu, 2021*). However, CNNs have inherent limitations due to the localized nature of their convolutional operations, which can lead to under- or over-segmentation in complex ultrasound images. Addressing this challenge, the Attention UNet was introduced, showcasing its efficacy in handling variable small-sized organs by incorporating attention gates (AGs) within the UNet (*Oktay et al., 2018*). Further advancements by researchers like *Chen, Yao & Zhang (2020)* involved the fusion of the ResNet architecture (*He et al., 2016*) with attention mechanisms, thereby enhancing feature extraction and generating high-quality segmentation results for complex features. Another innovative approach, the AAUNet, adaptively selects receptive fields of varying scales from channel and spatial dimensions, leading to substantial improvements in breast lesion segmentation in ultrasound images (*Chen et al., 2022*).

Ultrasound images inherently incorporate non local features, resulting in ambiguous boundaries between target regions and backgrounds. Traditional UNet-based models face challenges in capturing long-range semantic dependencies within ultrasound images (*Fang et al., 2023*). Atrous or dilated convolution methods were introduced in these scenarios (*Chen et al., 2017b*; *Chen et al., 2018*). These methods expand the receptive field of convolutions without increasing the parameter count, allowing them to effectively aggregate multiscale contextual information. Atrous CNN have proven instrumental in achieving more accurate segmentation, particularly in scenarios involving intricate spatial structures and scales (*Yu, Koltun & Funkhouser, 2017*). For instance, *Zhou, He & Jia (2020)* applied atrous convolution to preserve resolution information in feature maps when segmenting brain tumor ultrasound images (*Zhou, He & Jia, 2020*), showcasing the versatility of these techniques in addressing segmentation challenges.

Recent advancements have seen the successful integration of CNN and transformer blocks to preserve global semantic information, with transformers demonstrating exceptional prowess in capturing intricate patterns and relationships in both natural language processing (*Devlin et al., 2018*; *Vaswani et al., 2023*) and computer vision domains (*Parmar et al., 2018*; *Liu et al., 2021*). A novel adaptation of the Transformer for computer vision, known as the ViT, eliminates the need for convolutions to extract features from images (*Dosovitskiy et al., 2020*). The ViT segments images into discrete non-overlapping patches. Spatial positioning information is then introduced to these patches through position encodings, and they are subsequently passed through standard transformer layers. This allows the ViT to effectively model both local and global semantic dependencies. Further augmenting this progress, the Segformer incorporates a Bilinear Fusion mechanism to efficiently merge multi-level feature maps, enhancing both receptive field and resolution for optimized segmentation results (*Xie et al., 2021*). The TransUNet offers a compelling solution with remarkable segmentation performance, effectively marrying high-resolution spatial details from CNN features with the contextual breadth of transformers to address inherent locality limitations and mitigate feature resolution loss, typically associated with pure transformers (*Chen et al., 2021*). The Swin-UNet, by combining a symmetric encoder–decoder structure with skip connections and integrating local-to-global self-attention, marks a significant advancement in image segmentation, optimizing transformer computations and enhancing segmentation efficiency (*Cao et al., 2022*). Lin and collaborators have integrated Swin Transformers and Multi-scale Vision Transformers (*Chen, Fan & Panda, 2021*) into the UNet, fostering excellent long-range dependencies between features of different scales (*Lin et al., 2022*). Additionally, CSwin-PNet was proposed to further enhance long-range dependency modeling, particularly tailored for ultrasound breast segmentation (*Yang & Yang, 2023*).

Given the considerations highlighted, we recognized the significance of global modeling within CNN. Drawing inspiration from TransUNet and atrous convolutions, we introduce our novel approach, referred to as the Modified CNN & ViT hybrid Encoder-Decoder segmentation network with multi-scale information fusion approach (MCV-UNet). To our knowledge, this marks the first endeavor to integrate CNN and ViT explicitly for

ultrasound nerve segmentation. Our contributions in this study can be delineated as follows:

1. Inspired by the burgeoning success of the Vision Transformer in the domain of computer vision, we have further integrated the ViT-layer within the Encoder-Decoder segmentation paradigm, enhancing its feature extraction prowess.
2. Recognizing the importance of capturing intricate details that span from local nuances to broader patterns, we introduce various atrous CNN layers. These layers augment the network's receptive field, bolstering its ability to discern and process multi-scale spatial hierarchies.
3. To validate the efficacy of our approach, we compared MCV-UNet against an array of established baseline methods. The empirical evaluations underscored our network's superior capabilities on a public dataset, yielding competitive results against 15 baseline methods.

The remainder of this article is structured as follows: 'Related work' reviews the relevant literature, highlighting key developments in CNN and ViT utilized to medical image segmentation. 'Approach' details the proposed approach, MCV-UNet, including the network framework, analytical techniques, and related equations. 'Results' discusses the results obtained with MCV-UNet, covering data sources, implementation details, and evaluation criteria. It also provides an in-depth discussion of these results from different perspectives. 'Conclusion' is the conclusion including remarks, summarizing the superior performance of MCV-UNet and suggesting ideas for future research in this field. To aid in the clarity and readability of this article, a table of abbreviations is provided in Table 1.

RELATED WORK

Medical image segmentation with CNN

CNN has initially emerged as the predominant methods for image processing tasks (*Milletari, Navab & Ahmadi, 2016*; *Chen et al., 2018*; *Lv et al., 2020*; *Ali, Qureshi & Shah, 2023*). In the domain of medical image processing, where the desired output extends beyond a single class label, the need for precise segmentation of organs or tumors is paramount. Pioneering efforts by *Cireşan, Meier & Schmidhuber (2012)* leveraged deep neural network networks trained on GPUs, leading to substantial improvements in recognition rates on medical image datasets. The introduction of the fully convolutional network (FCN) marked a crucial development by striking a balance between capturing global and local information through the integration of multi-resolution layers (*Long, Shelhamer & Darrell, 2015*). To further enhance training efficiency with limited data, Ronneberger introduced the symmetric UNet architecture, extending the contracting network by incorporating successive layers with skip connections (*Ronneberger, Fischer & Brox, 2015*). UNet quickly gained popularity for its remarkable ability to learn invariance from medical images. LinkNet innovatively directly linked the encoder to the corresponding decoder, ensuring precise predictions without compromising network processing speed (*Chaurasia & Culurciello, 2017*). Subsequent advancements in UNet-based networks, like the Attention UNet with its AGs mechanisms, enabled networks to focus on targets

Table 1 Abbreviation instructions.

Abbreviation	Full form
CNN	Convolutional neural networks
ViT	Vision transformers
NS	Nerve stimulation
AGs	Attention gates
FCN	Fully convolutional network
Dice	Dice coefficient
Acc	Accuracy
Pre	Precision
Sen	Sensitivity
Spec	Specificity
Cost	Computational cost
LN	Layer normalization
MSA	Multi-head self-attention
MLP	Multilayer perceptron
TP	True positive
FP	False positive
TN	True negative
FN	False negative

of complex shape and size, expanding its applications in ultrasound image segmentation (*Oktay et al., 2018*). Res-UNet was specifically designed for ultrasound nerve segmentation, enhancing accuracy through the incorporation of dense atrous convolutions and residual multiple posing modules compared to the traditional UNet (*Wang, Shen & Zhou, 2019*). Furthermore, researchers have explored combining recurrent neural networks with residual neural networks to achieve improved organ segmentation performance (*Alom et al., 2018*). Addressing the growing demand for precise medical image segmentation, UNet3+ maximized feature map utilization through full-scale connections (*Huang et al., 2020*). Transfer Learning techniques were also incorporated with the UNet architecture, as demonstrated by *Cheng & Lam (2021)* who applied their network successfully to lung ultrasound segmentation, leveraging mechanisms for detecting edges, shapes, and textures from ultrasound images. Additionally, the Dense-PSP-UNet introduced an innovative Pyramid Scene Parsing (PSP) module, surpassing skip connection settings in performance and employing Contrast Limited Adaptive Histogram Equalization (CLAHE) (*Reza, 2004*) to reduce image noise levels during training (*Ansari et al., 2023*).

Medical image segmentation with transformers

The transformer architecture, initially pivotal in sequential processing, marked a paradigm shift with its self-attention mechanism, enabling unprecedented performance in various classification tasks (*Vaswani et al., 2017*). In computer vision, the ViT replaced traditional convolutional layers with a novel approach of segmenting images into non-overlapping patches, treated as linear embeddings. This method facilitated contextual relationships between patches through self-attention, enhancing the network's comprehension of the

entire image (*Dosovitskiy et al., 2020*; *Wang, Zhao & Ni, 2022*; *Chen et al., 2021*; *Liu, Hu & Chen, 2023*). ViT has set new benchmarks in object detection (*Fang et al., 2021*), rivaling state-of-the-art CNN architectures, particularly when pre-trained on extensive datasets (*Dosovitskiy et al., 2020*). The introduction of axial (*Ho et al., 2019*) and hierarchical attention (*Yang et al., 2016*) further refined ViT, enabling more precise segmentation. A significant advantage of ViT is its capacity to handle varied image sizes, crucial for intricate ultrasound images. In medical image segmentation, where precision is critical for diagnosis and treatment, traditional methods face challenges like varying contrasts and subtle pathological indicators. TransUNet combined CNN's local detail capture with transformers' holistic view, enhancing the understanding of medical images (*Chen et al., 2021*). Swin-Unet, leveraging transformer blocks, adeptly handles high-resolution medical scans (*Liu et al., 2021*; *Cao et al., 2022*). The hybrid CNN-Transformer network further innovated by integrating large-kernel convolution, effectively capturing multi-scale information (*Liu, Hu & Chen, 2023*). The ViT-Patch introduced a secondary task on the patch tokens, in addition to the primary task on the class token, demonstrating superior performance compared to the standard ViT in breast ultrasound segmentation. token (*Feng et al., 2023*). HA-UNet's introduction of local–global transformer blocks represents a significant step in reducing computational complexity without sacrificing segmentation efficiency (*Zhang et al., 2024*). The inclusion of a cross attention block in HA-UNet not only improved feature integration but also demonstrated significant advancements in ultrasound breast lesion segmentation.

APPROACH

Architecture overview

In the domain of deep learning applied to image segmentation, the objective is to map an input image x to its segmented inference y. This mapping is denoted as $y_{\text{pred}} = f(x;\theta)$, where f is the deep learning network, θ represents the network's parameters, and y_{pred} is the predicted segmentation of each pixel, where pred $\in [0,1]$. The corresponding ground truth for the input image x is represented as y_{gt}. During the training phase, we use a dataset consisting of batches of paired data represented as $(x, y_{\text{gt}}) \in D_{\text{train}}$. Our primary aim during training is to optimize the parameters θ to minimize the difference between y_{pred} and y_{gt}. For evaluation on unseen data, we use $(x, y_{\text{gt}}) \in D_{\text{test}}$ and assess the network's performance by comparing y_{pred} to y_{gt}. The MCV-UNet, a novel approach in medical image segmentation for ultrasound images, is depicted in Fig. 1. The architecture, which integrates CNN and ViT, consists of an encoder, bottleneck, decoder, and skip connections. Built upon the foundational UNet structure (*Ronneberger, Fischer & Brox, 2015*), MCV-UNet innovates with key components atrous convolutional and ViT layers. The process begins with two 3×3 atrous convolutional layers in the encoder, designed to extract multi-scale features while expanding the network's receptive field without significantly increasing computations (*Chen et al., 2017a*). This is followed by standard convolution-based encoders and max-pooling layers, effectively balancing spatial dimension reduction and computational efficiency. In the symmetric design, the feature maps from the encoder aid each upsampling

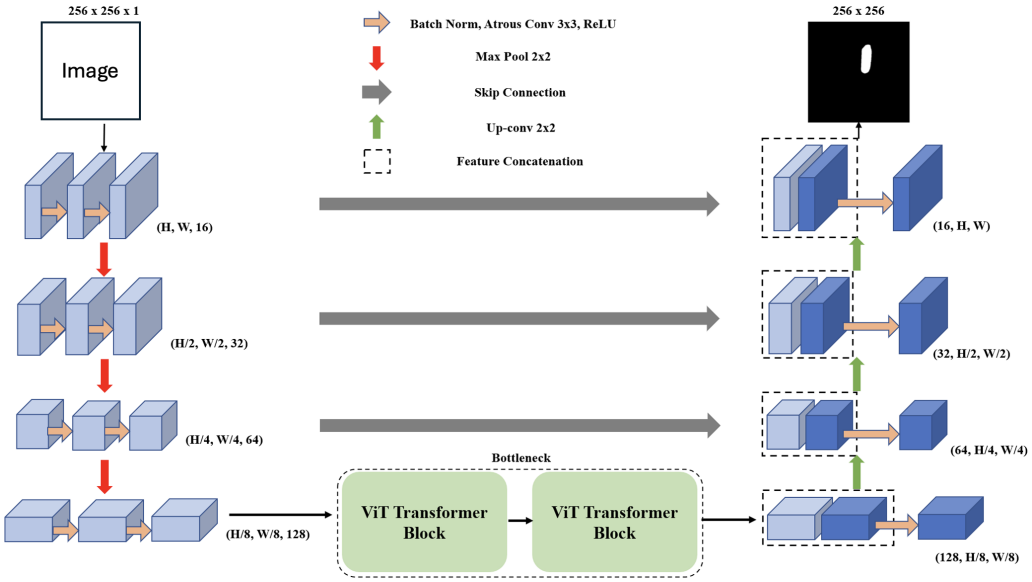

Figure 1 **The proposed encoder-decoder segmentation MCV-UNet based on atrous CNN and ViT blocks.**

Full-size DOI: 10.7717/peerjcs.2146/fig-1

step in the decoder. A 2×2 deconvolution layer halves the feature channels, and these are merged with corresponding encoder feature maps *via* skip connections, preserving crucial information.

We introduce a bottleneck with two transformer blocks, leveraging ViT for its segmentation accuracy (*Dosovitskiy et al., 2020*). This unique combination of atrous convolution and ViT allows MCV-UNet to capture both local details and global context effectively, a crucial requirement in medical image segmentation. The final expanding layer in the decoder then maps feature vectors to the desired class numbers, ensuring the output matches the input resolution. MCV-UNet's design is a strategic evolution from conventional UNet, inspired by TransUNet's hybrid CNN-Transformer approach (*Chen et al., 2021*). The specific functionalities of ViTs and atrous convolutions, crucial to MCV-UNet's performance, are further detailed in the subsequent sections.

Vision transformer layer

Two successive Vision Transformer blocks serves as a key element in the bottleneck between the encoder and decoder is illustrated in Fig. 2. Within each Vision Transformer block, we applied a layer normalization (LN), multi-head self-attention (MSA), a two-layer multilayer perceptron (MLP) with GELU (*Ba, Kiros & Hinton, 2016*; *Hendrycks & Gimpel, 2016*). A residual connection was applied each module (*He et al., 2016*). The computation within these continuous Transformer blocks is illustrated as follows:

$$\hat{z}^l = \mathrm{MSA}\left(\mathrm{LN}(z^{l-1})\right) + z^{l-1} \tag{1}$$

$$z^l = \mathrm{MLP}\left(\mathrm{LN}(\hat{z}^l)\right) + \hat{z}^l \tag{2}$$

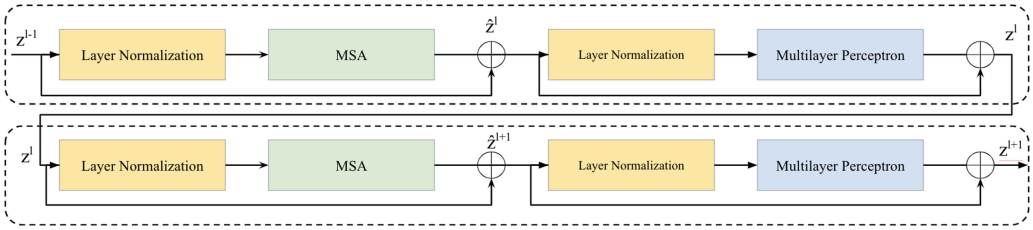

Figure 2 **Two successive Vision Transformer blocks.**
Full-size DOI: 10.7717/peerjcs.2146/fig-2

$$\hat{z}^{l+1} = \mathrm{MSA}\left(\mathrm{LN}(z^{l})\right) + z^{l} \tag{3}$$

$$z^{l+1} = \mathrm{MLP}\left(\mathrm{LN}(\hat{z}^{l+1})\right) + \hat{z}^{l+1} \tag{4}$$

here, $\hat{z}^{l}$ and z^{l} represent the outputs of the MSA module and the MLP module of the lth block, respectively.

Drawing inspiration from previous works (*Pan et al., 2022*; *Keles, Wijewardena & Hegde, 2023*; *Wang & Ma, 2023*; *Qin et al., 2022*; *Zhang et al., 2023*), our self-attention computation strategy adheres to the principles of scaled dot-product attention (*Vaswani et al., 2023*). This approach leverages the efficiency of dot-product attention, optimizing the use of matrix multiplication (*Bahdanau, Cho & Bengio, 2014*). The self-attention computation can be illustrated as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \tag{5}$$

where Q, K, and V are matrices representing queries, keys, and values, respectively, with dimensions $Q, K, V \in \mathbb{R}^{M^2 \times d}$, where M^2 denotes the number of patches in a window and d represents the query or key dimension. The bias term B is derived from the bias matrix $\hat{B}$, with $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$.

Atrous convolution layer

In the classical encoder–decoder architecture, the repeated operations of max-pooling and striding at consecutive layers often lead to a substantial reduction in the spatial resolution of the resulting feature maps. While skip connections and deconvolutional layers can help recover some of this lost spatial information, MCV-UNet takes a further step to mitigate this issue by incorporating atrous convolution within the encoder–decoder architecture.

Atrous convolution, initially introduced for the computation of the undecorated wavelet transform in the "algorithme à trous" scheme (*Holschneider et al., 1990*), has demonstrated high performance in various applications, including semantic segmentation (*Chen et al., 2017a*). In the context of two-dimensional data, atrous convolution is defined as follows:

$$y[i] = \sum_{k}^{K} x[i + rk]w[k] \tag{6}$$

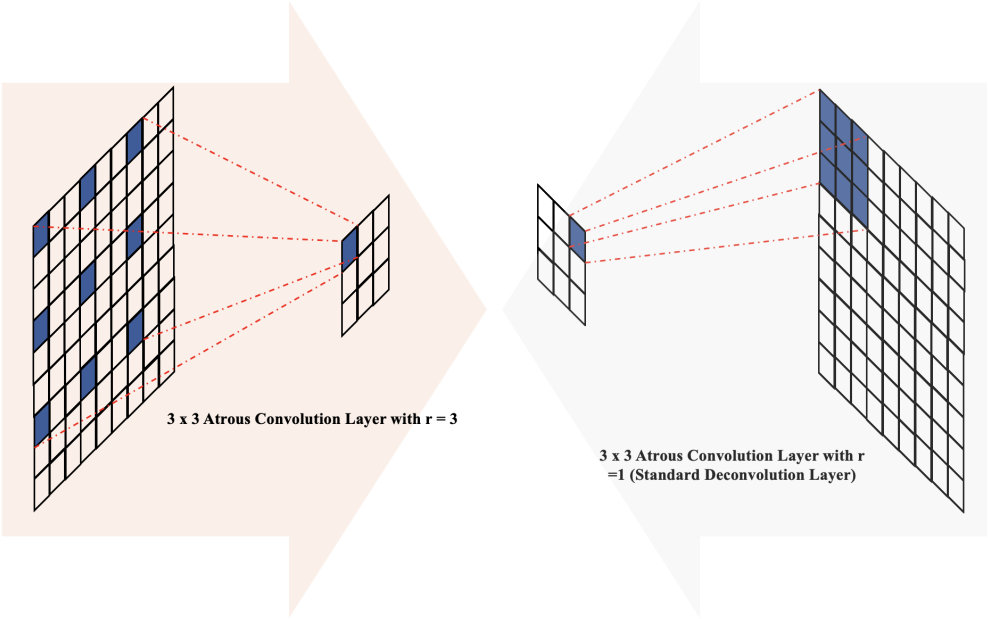

Figure 3 **The illustration of atrous convolution-based block.**
Full-size DOI: 10.7717/peerjcs.2146/fig-3

here, the rate parameter r corresponds to the stride of the sampled input signal, $x[i]$ represents the two-dimensional input signal, and the atrous rate r convolves the input signal with the upsampled filter $w[k]$ by introducing $r-1$ zeros between consecutive filter values along each spatial dimension (*Chen et al., 2017b*; *Gu et al., 2019*). The parameter K denotes the length of the filter $w[k]$, and the standard convolution corresponds to the special case of an atrous rate $r=1$. Adjusting the atrous rate provides the network with flexibility in terms of the field of view and enables the generation of larger outputs without significantly increasing computational demands (*Chen et al., 2017a*). Previous works have incorporated atrous convolution in various ways, including within encoder–decoder blocks (*Chen et al., 2018*; *Chen et al., 2017b*), skip connections (*Wang & Voiculescu, 2021*), and the module bridging the encoding and decoding stages to extract dense features (*Lv et al., 2020*; *Gu et al., 2019*; *Pan et al., 2019*; *Ma, Gu & Wang, 2024*).

In the architecture of MCV-UNet, inspired by the encoder–decoder structure with atrous convolution (*Chen et al., 2018*), we placed batch normalization layers before atrous convolution operations with atrous rates $r=3$ in the encoder blocks and $r=1$ (standard convolution) in the decoder blocks, as illustrated in Fig. 3. The introduction of holes (with $r=3$) in the down-sampling process facilitates the computation of responses at all image positions while introducing zeros between filter values. This increases the size of the filter compared to the standard convolution layer, but computations only consider the values of non-zero filter elements, ensuring a constant number of filter parameters and computational operations. Overall, this approach offers the advantage of controlling

Table 2 The hyper-parameter setting for MCV-UNet and all baseline methods.

Epoch	Optimizer	Learning rate	Batch size	Dataset
50	Adam	10^{-4}	8	$5640 \times 256 \times 256$

feature resolution, enhancing the receptive field of the network without sacrificing image resolution.

RESULTS

Dataset

In our experiments, we employed the Nerve Segmentation database, a publicly available resource from the Kaggle Competition platform (*Anna Montoya et al., 2016*). This database is integral to medical imaging, particularly for the analysis of the brachial plexus nerve, a critical area often studied in ultrasound imaging. This dataset comprises 5,640 256×256 ultrasound images, distinctly split into 1,128 testing and 4,512 validation samples, with no overlap between training, validation, and testing sets. Each image encompasses a 2D ultrasound scan of the nerve alongside a meticulously manually annotated 2D segmentation mask, serving as ground truth. The images present a unique challenge due to the random distribution of the nerve within them, demanding precise segmentation skills. To facilitate uniform analysis, all images underwent normalization, scaling pixel values to the range [0, 1], thereby simplifying the task of differentiating the nerve from surrounding tissues. This approach ensures accurate segmentation by leveraging expert annotations and standardized image processing techniques.

Implementation details

The implementation of our approach was developed using Python 3 and TensorFlow (*Abadi et al., 2015*). Our experiments were conducted on a robust computing setup, featuring an Intel Xeon CPU with 2 vCPUs and 13 GB of RAM, and significantly accelerated with an NVIDIA A100 GPU, equipped with 40 GB of VRAM, known for its high computational efficiency in deep learning tasks. We adapted several networks from established sources, specifically segmentation models and Keras-UNet-Collections, applying necessary modifications to optimize them for our specific dataset. These adaptations were crucial in handling our dataset's unique characteristics.

During the training phase, consistency in parameters across all networks was maintained to ensure fair comparative analysis. We employed the Adam Optimizer (*Kingma & Ba, 2014*), renowned for its efficiency in computing gradients, setting the learning rate to to 10^{-4}, batch size to 8, and the number of training epochs to 50. Batch normalization layers (*Ioffe & Szegedy, 2015*) were strategically incorporated to enhance training speed and stability. The details of the hyper-parameter setting in the experiment is illustrated in Table 2.

The network's performance was evaluated using the Dice coefficient-based loss, a standard metric in image segmentation tasks, which quantifies the similarity between predicted and ground truth segmentation. We saved the network from the epoch showing

the best performance for testing phase segmentation. On our specified hardware, training a single network typically required 3 to 5 h, depending on the network's complexity and architecture.

Metrics

To comprehensively evaluate the performance of MCV-UNet, a diverse set of evaluation metrics are utilized. These metrics encompass a range of criteria, including the Dice coefficient (Dice), Accuracy (Acc), Precision (Pre), Sensitivity (Sen), Specificity (Spec), and the parameters of network as computational cost metrics (Cost). Each of these metrics offers a unique perspective on the effectiveness of MCV-UNet in segmenting medical images. The details of our evaluation metrics can be outlined as follows:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{7}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

where, TP represents the number of true positives, TN denotes the number of true negatives, FP signifies the number of false positives, and FN stands for the number of false negatives.

By employing these diverse metrics, a comprehensive assessment of MCV-UNet's segmentation performance could be achieved. Each metric contributes valuable insights into different aspects of the network's performance, enabling us to evaluate the effectiveness and accuracy of MCV-UNet in the context of medical image segmentation.

Comparison with state-of-the-arts

To evaluate the performance of MCV-UNet in the context of medical image segmentation, we conducted an extensive comparison with 17 baseline networks, including a diverse range of architectural designs, each with its own strengths and characteristics. To demonstrate the effectiveness of MCV-UNet, we use ViT block bridge the encoder and decoder, with comparisons made against classical CNN networks and their variants. Furthermore, we evaluated the impact of the encoder and decoder design, incorporating atrous convolution layers, by contrasting the result with existing hybrid CNN-ViT networks. The networks compared include: UNet (*Ronneberger, Fischer & Brox, 2015*), UNet-ResNet34

(*Ronneberger, Fischer & Brox, 2015*), UNet-MobileNet (*Ronneberger, Fischer & Brox, 2015*), UNet-InceptionV3 (*Ronneberger, Fischer & Brox, 2015*), Linknet (*Chaurasia & Culurciello, 2017*), Linknet-MobileNet (*Chaurasia & Culurciello, 2017*), FPN-ResNet34 (*Lin et al., 2017*), FPN-MobileNet (*Lin et al., 2017*), TransUNet (*Chen et al., 2021*), FPN-InceptionV3 (*Lin et al., 2017*), VNet (*Milletari, Navab & Ahmadi, 2016*), AttentionUNet (*Oktay et al., 2018*), UNet3+ (*Huang et al., 2020*), U2-Net (*Qin et al., 2020*), RARUNet (*Wang, Zhang & Voiculescu, 2021*), QAPNet (*Wang & Voiculescu, 2021*), and R2UNet (*Alom et al., 2018*). This comprehensive comparison allows us to demonstrate the unique strengths and capabilities of MCV-UNet in the context of medical image segmentation.

Qualitative results

Figure 4 displays qualitative comparison results with three randomly chosen ultrasound medical images alongside their corresponding ground truths. The original raw images are removed due to Kaggle Policy. For each example, predictions generated by 17 baseline methods are compared with the results from MCV-UNet. The results of these visual analyses yield valuable insights into the performance of each approach. Classical CNN-based methods, including UNet (*Ronneberger, Fischer & Brox, 2015*), Attention UNet (*Oktay et al., 2018*), and V-Net (*Milletari, Navab & Ahmadi, 2016*), tend to exhibit issues of over-segmentation or under-segmentation. For instance, in the first example, UNet-MobileNet over-segments the nerve while V-Net under-segments it. This observation underscores the superior capability of MCV-UNet in effectively encoding global contexts and distinguishing the semantics. In addition, in the context of existing hybrid CNN-ViT networks, the predictions generated by TransUNet (*Chen et al., 2021*) demonstrate coarser characteristics than those by MCV-UNet, particularly with regard to boundary and shape. In the third example, MCV-UNet displays excellent alignment with nerve boundary of the ground truth, whereas TransUNet predicts more false positives. These visual comparisons proves the superior performance of our network, characterized by its capacity to preserve detailed shape information, resulting in fewer false positives and false negatives compared to the baseline methods. This superiority is attributed to the successful combination of CNN and ViT architectures in preserving high-level global information and low-level details, while minimizing spatial information loss with atrous convolution layers.

Quantitative results

Table 3 presents a comprehensive quantitative evaluation of our MCV-UNet in comparison to the 17 baseline methods in Tables 3, 4, 5. The quantitative results proved the exceptional performance of MCV-UNet on most evaluation metrics. For the main criterion metric of dice coefficient, MCV-UNet achieves a remarkable result of 62.51%, surpassing the second-ranked network by 0.47%. In terms of accuracy and precision, MCV-UNet outperforms all competitors, with a 0.18% increase in Acc and a 1.16% increase in Pre compared to the second-best network. MCV-UNet exhibits competitive performance in sensitivity and specificity metrics, aligning with the top-performing networks. Regarding computational cost, MCV-UNet falls within the median range of all trained networks, showing a slight advantage over our derived architecture, TransUNet. These quantitative results indicate

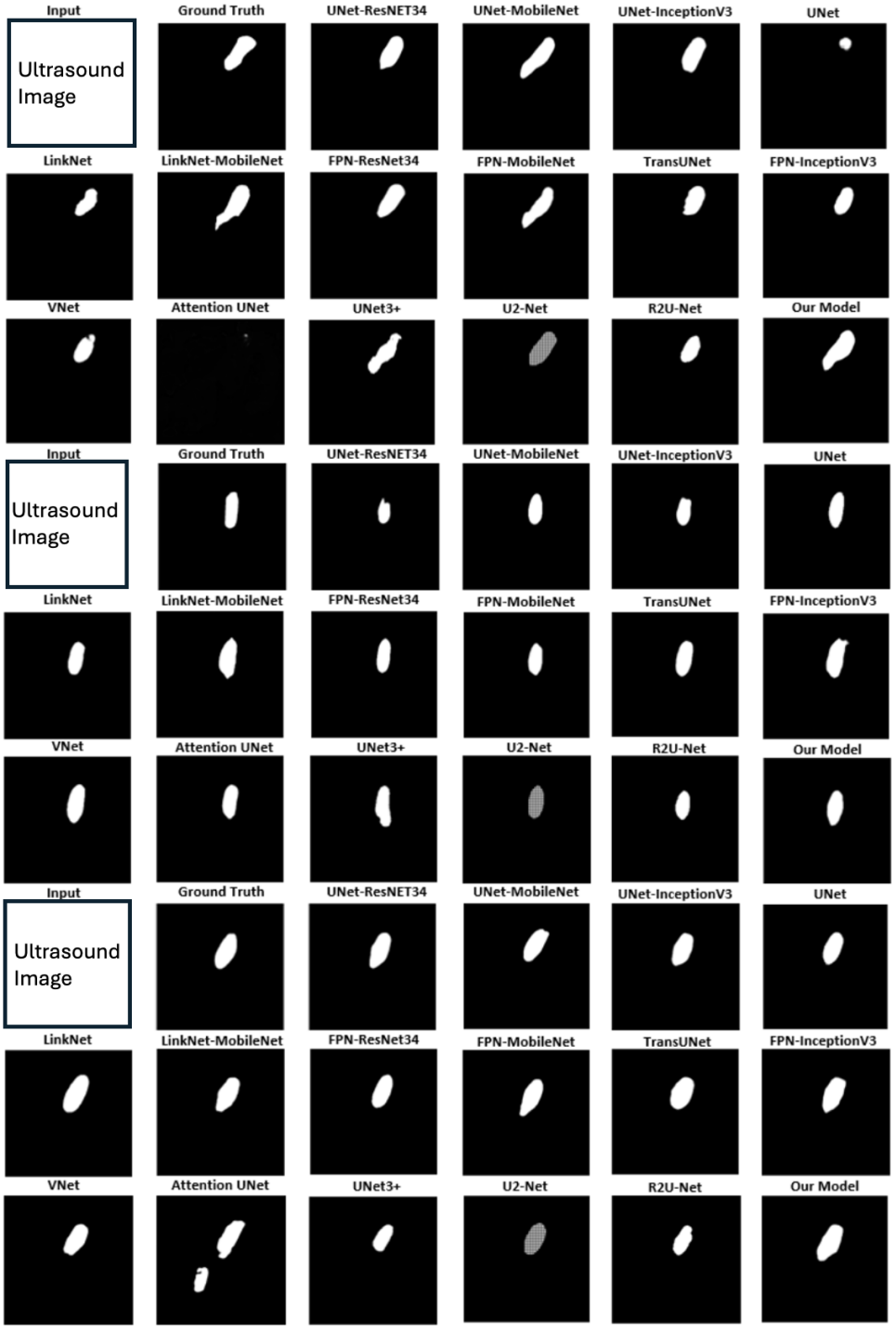

Figure 4 The segmentation results of different networks on the brachial plexus nerve testing dataset.
Full-size DOI: 10.7717/peerjcs.2146/fig-4

Table 3 The performance of MCVUNet with other baseline methods on ultrasound nerve segmentation test set.

Network	Dice	Acc	Pre	Sen	Spec	Cost
UNet	0.6217	0.9910	0.6179	0.6256	0.9953	1,967,041
UNet-ResNet34	0.6165	0.9910	0.6218	0.6130	0.9955	24,456,160
UNet-MobileNet	0.6184	0.9907	0.6038	0.6349	0.9950	8,336,343
UNet-InceptionV3	0.6173	0.9899	0.5599	0.6997	0.9934	29,933,111
LinkNet	0.6218	0.9911	0.6286	0.6168	0.9956	20,325,137
LinkNet-MobileNet	0.6067	0.9896	0.5474	0.6840	0.9932	4,546,071
FPN-ResNet34	0.6172	0.9906	0.5941	0.6435	0.9947	23,930,960
FPN-MobileNet	0.5971	0.9907	0.6099	0.5859	0.9955	6,103,111
TransUNet	0.5976	0.9888	0.5193	**0.7037**	0.9922	4,675,335
FPN-InceptionV3	0.6222	0.9904	0.5803	0.6745	0.9942	25,029,287
VNet	0.5955	0.9896	0.5533	0.6491	0.9937	3,690,129
AttentionUNet	0.5533	0.9887	0.5184	0.5934	0.9934	638,322
UNet3+	0.5890	0.9899	0.5695	0.6099	0.9945	497,848
U2-Net	0.4366	0.9892	0.5738	0.3523	**0.9969**	3,775,677
R2UNet	0.6115	0.9910	0.6230	0.6004	0.9956	1,448,215
RARUNet	0.6219	0.9910	0.6181	0.6257	0.9953	11,793,638
QAPNet	0.6218	0.9910	0.6180	0.6257	0.9954	9,472,052
Ours	**0.6251**	**0.9928**	**0.6359**	0.6912	0.9960	4,675,329

Notes.
The best performance results are highlighted in bold. The second-best performance of MCV-UNet is highlighed with an underline.

MCV-UNet's competitive edge across multiple evaluation metrics relative to the 17 baseline networks. The parameters of MCV-UNet is 43%, 76&, 49% lower than UNet with mobilenet as network backbone, LinkNet, and QAPNet. Notably, it is also slightly lower than the current advanced ViT-based TransUNet due to the modified atrous CNN is utilized. They also emphasize capabilities of MCV-UNet of combining the strengths of classical CNN and hybrid CNN-ViT networks.

Sensitivity analysis & ablation study

In addition to our primary experiments, we carried out a sensitivity analysis focused on the hyper-parameter setting related to the dilated rate in our multi-scale CNN. This analysis is essential to determine the optimal configuration for effectively segmenting nerves in ultrasound images. The detailed results of this analysis are presented in Table 4. These findings validate the effectiveness of the dilated rate settings in our proposed MCV-UNet.

To study the individual and collective impact of the multi-scale modules proposed in our network, we conducted an ablation study, the results of which are detailed in Table 5. This study methodically explores the effects of omitting or modifying various components of our network. These findings not only validate the efficacy of each proposed contribution but also highlight their synergistic effect in enhancing the accuracy and robustness of the ultrasound nerve segmentation process.

Table 4 The sensitivity analysis of atrous CNN setting.

Dilation rate	1	2	3	4	5	6
Dice	0.5976	0.6105	0.6251	0.6248	0.6296	0.6175
Acc	0.9888	0.9894	**0.9928**	0.9912	0.9907	0.9910
Pre	0.5193	0.5423	**0.6359**	0.6323	0.5955	0.6186
Spec	0.9922	0.9929	**0.9960**	0.9957	0.9946	0.9954

Notes.
The best performance results are highlighted in bold. The second-best performance of MCV-UNet is highlighed with an underline.

Table 5 The ablation study of MCV-UNet on ultrasound nerve segmentation test set.

Multi-Scale CNN	Self-Attention	Dice	Acc	Pre	Sen	Spec
		0.6217	0.9910	0.6179	0.6256	0.9953
✓		0.6220	0.9903	0.5780	0.6706	0.9941
	✓	0.5976	0.9888	0.5193	0.7037	0.9922
✓	✓	**0.6251**	**0.9928**	**0.6359**	**0.6912**	**0.9960**

Notes.
The best performance results are highlighted in bold.

CONCLUSION

In this study, we studied the combination of modified CNN and ViT to address the intricate challenge of ultrasound nerve segmentation. Recognizing the limitations inherent in conventional CNN-based networks, especially their restricted capacity to exploit long-range semantic dependencies in ultrasound images, we proposed the MCV-UNet. This novel design represents a modified encoder–decoder framework that seamlessly combines the robust capabilities of both CNN and ViT, while it integrates atrous convolution layers to effectively recover lost spatial information. This kind of multi-scale feature information extraction is valuable in ultrasound imaging, because the nerve structure is complex, and the location, size of nerve can be different in each of ultrasound image. Considering other modalities images, such as CT, MRI, and PET, the MCV-UNet is also valuable to be explored especially when the region of interest (ROI) is complex and should be recognized based on both of the local- and global-based features.

The qualitative and quantitative evaluations indicates that the proposed network outperformed 17 classical baseline methods, exhibiting fewer FP and FN—a testament to its robustness and precision. The integration of various dilated CNN layers further amplified its feature extraction capabilities, bridging the gap between local and global contextual understanding in the images.

Future work might consider refining the network architecture, introducing novel attention mechanisms, or expanding the approach to other challenging medical imaging domains. The computational burden should also be further decreased, because the current parameters of network is still high due to the utilization of ViT. Different types of CNN, network pruning, and knowledge distillation can also be studied to enable the efficiency of the network.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Zihong Xu conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Ziyang Wang conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in the Supplemental File.

The raw data is publicly available at Kaggle: https://www.kaggle.com/competitions/ultrasound-nerve-segmentation.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2146#supplemental-information.

REFERENCES

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. 2015. TensorFlow: large-scale machine learning on heterogeneous systems. *Available at https://www.tensorflow.org*.

Ali H, Qureshi R, Shah Z. 2023. Artificial intelligence–based methods for integrating local and global features for brain cancer imaging: scoping review. *JMIR Medical Informatics* **11(1)**:e47445 DOI 10.2196/47445.

Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. 2018. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. ArXiv arXiv:1802.06955 DOI 10.48550/arXiv.1802.06955.

Ansari MY, Yang Y, Meher PK, Dakua SP. 2023. Dense-PSP-UNet: a neural network for fast inference liver ultrasound segmentation. *Computers in Biology and Medicine* **153**:106478 DOI 10.1016/j.compbiomed.2022.106478.

Anna Montoya H, kaggle446, shirzad, Cukierski W, yffud. 2016. Ultrasound Nerve Segmentation. Kaggle. *Available at https://kaggle.com/competitions/ultrasound-nerve-segmentation*.

Armaghani SJ, Lee DS, Bible JE, Shau DN, Kay H, Zhang C, McGirt MJ, Devin CJ. 2016. Increased preoperative narcotic use and its association with postoperative complications and length of hospital stay in patients undergoing spine surgery. *Journal of Spinal Disorders and Techniques* **29(2)**:e93–e98.

Ba JL, Kiros JR, Hinton GE. 2016. Layer normalization. ArXiv arXiv:1607.06450.

Baby M, Jereesh A. 2017. Automatic nerve segmentation of ultrasound images. In: *2017 International conference of electronics, communication and aerospace technology (ICECA), volume 1*. Piscataway: IEEE, 107–112.

Bahdanau D, Cho K, Bengio Y. 2014. Neural machine translation by jointly learning to align and translate. ArXiv arXiv:1409.0473.

Bajwa SJS, Haldar R. 2015. Pain management following spinal surgeries: an appraisal of the available options. *Journal of Craniovertebral Junction & Spine* **6(3)**:105 DOI 10.4103/0974-8237.161589.

Brass P, Hellmich M, Kolodziej L, Schick G, Smith AF. 2015. Ultrasound guidance versus anatomical landmarks for internal jugular vein catheterization. *Cochrane Database of Systematic Reviews* **1(1)**:CD006962 DOI 10.1002/14651858.CD006962.pub2.

Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. 2022. Swin-Unet: unet-like pure transformer for medical image segmentation. In: *European conference on computer vision*. Cham: Springer Nature Switzerland.

Chan VW, Perlas A, Rawson R, Odukoya O. 2003. Ultrasound-guided supraclavicular brachial plexus block. *Anesthesia & Analgesia* **97(5)**:1514–1517.

Chaurasia A, Culurciello E. 2017. Linknet: exploiting encoder representations for efficient semantic segmentation. In: *2017 IEEE visual communications and image processing (VCIP)*. Piscataway: IEEE, 1–4.

Chen C-FR, Fan Q, Panda R. 2021. Crossvit: cross-attention multi-scale vision transformer for image classification. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Piscataway: IEEE, 357–366.

Chen G, Li L, Dai Y, Zhang J, Yap MH. 2022. AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. *IEEE Transactions on Medical Imaging* **42(5)**:1289–1300 DOI 10.1109/TMI.2022.3226268.

Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. 2021. Transunet: transformers make strong encoders for medical image segmentation. ArXiv arXiv:2102.04306.

Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. 2017a. 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 Epub 2017 Apr 27 DOI 10.1109/TPAMI.2017.2699184.

Chen L-C, Papandreou G, Schroff F, Adam H. 2017b. Rethinking atrous convolution for semantic image segmentation. ArXiv arXiv:1706.05587.

Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

Chen X, Yao L, Zhang Y. 2020. Residual attention u-net for automated multi-class segmentation of COVID-19 chest CT images. ArXiv arXiv:2004.05645.

Cheng D, Lam EY. 2021. Transfer learning U-Net deep learning for lung ultrasound segmentation. ArXiv arXiv:2110.02196.

Cireşan D, Meier U, Schmidhuber J. 2012. Multi-column deep neural networks for image classification. In: *2012 IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE.

Copeland SJ, Laxton MA. 2001. A new stimulating catheter for continuous peripheral nerve blocks. *Regional Anesthesia and Pain Medicine* **26(6)**:589 DOI 10.1053/rapm.2001.26215.

Devlin J, Chang M.-W., Lee K, Toutanova K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv arXiv:1810.04805.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. 2020. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv arXiv:2010.11929.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M. 2020. An image is worth 16x16 words: transformers for image recognition at scale. ArXiv arXiv:2010.11929.

Fang Y, Liao B, Wang X, Fang J, Qi J, Wu R, Niu J, Liu W. 2021. You only look at one sequence: rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems* **34**:26183–26197.

Fang C, Wang Q, Cheng L, Gao Z, Pan C, Cao Z, Zheng Z, Zhang D. 2023. Reliable mutual distillation for medical image segmentation under imperfect annotations. *IEEE Transactions on Medical Imaging* **42(6)**:1720–1734 DOI 10.1109/TMI.2023.3237183.

Feng H, Yang B, Wang J, Liu M., Yin L, Zheng W, Yin Z, Liu C. 2023. Identifying malignant breast ultrasound images using ViT-patch. *Applied Sciences* **13(6)**:3489 DOI 10.3390/app13063489.

Gu Z, Cheng J, Fu H, Zhou K., Hao H, Zhao Y, Zhang T, Gao S, Liu J. 2019. Ce-net: context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging* **38(10)**:2281–2292 DOI 10.1109/TMI.2019.2903562.

Hauritz RW, Hannig KE, Balocco AL, Peeters G, Hadzic A, Børglum J, Bendtsen TF. 2019. Peripheral nerve catheters: a critical review of the efficacy. *Best Practice & Research Clinical Anaesthesiology* **33(3)**:325–339 DOI 10.1016/j.bpa.2019.07.015.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE, 770–778.

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

Ho J, Kalchbrenner N, Weissenborn D, Salimans T. 2019. Axial attention in multidimensional transformers. ArXiv arXiv:1912.12180.

Holschneider M, Kronland-Martinet R, Morlet J, Tchamitchian P. 1990. A real-time algorithm for signal analysis with the help of the wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P, eds. *Wavelets. Inverse problems and theoretical imaging*. Berlin, Heidelberg: Springer, 286–297 DOI 10.1007/978-3-642-75988-8_28.

Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J. 2020. Unet 3+: a full-scale connected unet for medical image segmentation. In: *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Piscataway: IEEE, 1055–1059.

Ioffe S, Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*.

Kehlet H, Jensen TS, Woolf CJ. 2006. Persistent postsurgical pain: risk factors and prevention. *The Lancet* **367(9522)**:1618–1625 DOI 10.1016/S0140-6736(06)68700-X.

Keles FD, Wijewardena PM, Hegde C. 2023. On the computational complexity of self-attention. In: *International conference on algorithmic learning theory*.

Kick O, Blanche E, Pham-Dang C, Pinaud M, Estebe JP. 1999. A new stimulating stylet for immediate control of catheter tip position in continuous peripheral nerve blocks. *Anesthesia & Analgesia* **89(2)**:533–534 DOI 10.1213/00000539-199908000-00062.

Kingma DP, Ba J. 2014. Adam: a method for stochastic optimization. ArXiv arXiv:1412.6980.

Li W, Tang YM, Wang Z, Yu KM, To S. 2022. Atrous residual interconnected encoder to attention decoder framework for vertebrae segmentation via 3D volumetric CT images. *Engineering Applications of Artificial Intelligence* **114**:105102 DOI 10.1016/j.engappai.2022.105102.

Lin A, Chen B, Xu J, Zhang Z, Lu G, Zhang D. 2022. Ds-transunet: dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement* **71**:1–15.

Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. 2017. Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE.

Liu X, Hu Y, Chen J. 2023. Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron. *Biomedical Signal Processing and Control* **86**:105331 DOI 10.1016/j.bspc.2023.105331.

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*. Piscataway: IEEE, 10012–10022.

Long J, Shelhamer E, Darrell T. 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE,.

Lv Y, Ma H, Li J, Liu S. 2020. Attention guided U-Net with atrous convolution for accurate retinal vessels segmentation. *IEEE Access* **8**:32826–32839 DOI 10.1109/ACCESS.2020.2974027.

Ma C, Gu Y, Wang Z. 2024. TriConvUNeXt: a pure CNN-Based lightweight symmetrical network for biomedical image segmentation. *Journal of Imaging Informatics in Medicine* Epub ahead of print Apr 23 2024 DOI 10.1007/s10278-024-01116-8.

Merskey HE. 1986. Classification of chronic pain: descriptions of chronic pain syndromes and definitions of pain terms. *Pain* **3**:S1–S226.

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

Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B. 2018. Attention u-net: learning where to look for the pancreas. ArXiv arXiv:1804.03999.

Pacik PT, Nelson CE, Werner C. 2008. Pain control in augmentation mammaplasty using indwelling catheters in 687 consecutive patients: data analysis. *Aesthetic Surgery Journal* **28(6)**:631–641 DOI 10.1016/j.asj.2008.09.001.

Pan X, Ge C, Lu R, Song S, Chen G, Huang Z, Huang G. 2022. On the integration of self-attention and convolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Piscataway: IEEE, 815–825.

Pan X, Li L, Yang D, He Y, Liu Z, Yang H. 2019. An accurate nuclei segmentation algorithm in pathological image based on deep semantic network. *IEEE Access* **7**:110674–110686 DOI 10.1109/ACCESS.2019.2934486.

Parmar N, Vaswani A, Uszkoreit J, Łukasz Kaiser , Shazeer N, Ku A, Tran D. 2018. Image transformer. *ArXiv*. DOI 10.48550/arXiv.1802.05751.

Pham-Dang C, Kick O, Collet T, Gouin F, Pinaud M. 2003. Continuous peripheral nerve blocks with stimulating catheters. *Regional Anesthesia and Pain Medicine* **28(2)**:83–88 DOI 10.1097/00115550-200303000-00002.

Qin Z, Sun W, Deng H, Li D, Wei Y, Lv B, Yan J, Kong L, Zhong Y. 2022. cosFormer: rethinking softmax in attention. ArXiv arXiv:2202.08791.

Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. 2020. U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognition* **106**:107404 DOI 10.1016/j.patcog.2020.107404.

Reza AM. 2004. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology* **38**:35–44 DOI 10.1023/B:VLSI.0000028532.53893.82.

Ronneberger O, Fischer P, Brox T. 2015. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical image computing and computer-assisted intervention—MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351*. Cham: Springer, 234–241 DOI 10.1007/978-3-319-24574-4_28.

Schnabel A, Meyer-Frießem C, Zahn P, Pogatzki-Zahn E. 2013. Ultrasound compared with nerve stimulation guidance for peripheral nerve catheter placement: a meta-analysis of randomized controlled trials. *British Journal of Anaesthesia* **111(4)**:564–572 DOI 10.1093/bja/aet196.

Sola C, Raux O, Savath L, Macq C, Capdevila X, Dadure C. 2012. Ultrasound guidance characteristics and efficiency of suprazygomatic maxillary nerve blocks in infants: a descriptive prospective study. *Pediatric Anesthesia* **22(9)**:841–846 DOI 10.1111/j.1460-9592.2012.03861.x.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. In: *31st conference on neural information processing systems (NIPS 2017), Long Beach, CA, USA*.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2023. Attention is all you need. ArXiv arXiv:1706.03762.

Wang Z, Ma C. 2023. Dual-contrastive dual-consistency dual-transformer: a semi-supervised approach to medical image segmentation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Piscataway: IEEE, 870–879.

Wang R, Shen H, Zhou M. 2019. Ultrasound nerve segmentation of brachial plexus based on optimized ResU-Net. In: *2019 IEEE international conference on imaging systems and techniques (IST)*. Piscataway: IEEE, 1–6.

Wang Z, Voiculescu I. 2021. Quadruple augmented pyramid network for multi-class COVID-19 segmentation via CT. In: *2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*. Piscataway: IEEE, 2956–2959.

Wang Z, Zhang Z, Voiculescu I. 2021. RAR-U-Net: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels. In: *2021 IEEE international conference on image processing (ICIP)*. Piscataway: IEEE, 21–25.

Wang Z, Zhao W, Ni Z. 2022. Adversarial vision transformer for medical image semantic segmentation with limited annotations. In: *BMVC*. 1002.

Wijayasinghe N, Duriaud HM, Kehlet H, Anderson KG. 2016. Ultrasound guided intercostobrachial nerve blockade in patients with persistent pain after breast cancer surgery: a pilot study. *Pain Physician* **19(2)**:e309 DOI 10.36076/ppj/2016.19.E309.

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. In: *35th conference on neural information processing systems (NeurIPS 2021)*.

Yang H, Yang D. 2023. CSwin-PNet: a CNN-Swin transformer combined pyramid network for breast lesion segmentation in ultrasound images. *Expert Systems with Applications* **213**:119024 DOI 10.1016/j.eswa.2022.119024.

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. 2016. Hierarchical attention networks for document classification. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

Yu F, Koltun V, Funkhouser T. 2017. Dilated residual networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway: IEEE.

Zhang H, Lian J, Yi Z, Wu R, Lu X, Ma P, Ma Y. 2024. HAU-Net: hybrid CNN-transformer for breast ultrasound image segmentation. *Biomedical Signal Processing and Control* **87**:105427 DOI 10.1016/j.bspc.2023.105427.

Zhang L, Lu J, Zhang J, Zhu X, Feng J, Xiang T. 2023. Softmax-free linear transformers. ArXiv arXiv:2207.03341v3.

Zhou Z, He Z, Jia Y. 2020. AFPNet: a 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images. *Neurocomputing* **402**:235–244 DOI 10.1016/j.neucom.2020.03.097.