Review



# Ultrasound Image Analysis with Vision Transformers – Review

Majid Vafaeezadeh , Hamid Behnam * , Parisa Gifani

Posted Date: 4 January 2024

doi: 10.20944/preprints202401.0309.v1

Keywords: Transformer; Ultrasound (US); Deep Learning; Convolutional Neural Network(CNN); Vision transformer(ViT); Swin Transformer

*Review*

# Ultrasound Image Analysis with Vision Transformers—Review

**Majid Vafaeezadeh [1], Hamid Behnam [1,*] and Parisa Gifani [2]**

[1] Biomedical Engineering Department, School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran; majvaf@gmail.com

[2] Medical Sciences and Technologies Department, Science and Research Branch, Islamic Azad University, Tehran, Iran; p.gifani@gmail.com

* Correspondence: behnam@iust.ac.ir; Tel.: (9821+)77451500

**Abstract:** Ultrasound (US) has become a widely used imaging modality in clinical practice, characterized by its rapidly evolving technology, advantages, and unique challenges such as low imaging quality and high variability. There is a critical need to develop advanced automatic US image analysis methods to enhance diagnostic accuracy and objectivity. Vision transformer, a recent innovation in machine learning, has demonstrated significant potential in various research fields, including general image analysis and computer vision, due to its capacity to process large datasets and learn complex patterns. Its suitability for automatic US image analysis tasks, such as classification, detection, and segmentation, has been recognized. This review provides an introduction to vision transformer and discusses its applications in specific US image analysis tasks, while also addressing the open challenges and potential future trends in its application in medical US image analysis. Vision transformer has shown promise in enhancing the accuracy and efficiency of ultrasound image analysis and is expected to play an increasingly important role in the diagnosis and treatment of medical conditions using ultrasound imaging as technology progresses.

**Keywords:** transformer; ultrasound (US); deep learning; convolutional neural network(CNN); vision transformer(ViT); Swin transformer

## 1. Introduction

Ultrasound (US) is a versatile imaging technique that has become a fundamental resource in medical diagnosis and screening. Its wide acceptance and use by both physicians and radiologists underscore its importance and reliability. US is widely utilized due to its safety, affordability, non-invasive nature, real-time visualization capabilities, and the comfort it provides to those performing the procedure. It stands out among other imaging techniques like X-ray, MRI, and CT scans because of its several significant benefits, including absence of ionizing radiation, portability, ease of access, and cost-efficiency [1]. US is applied in various medical fields, such as breast US, echocardiography, transrectal US, intravascular US (IVUS), prenatal diagnostic US, and abdominal US. It is particularly prevalent in obstetrics [2]. However, despite its numerous benefits, US also presents certain challenges. These include lower image quality due to noise and artifacts, a high dependence on the operator or diagnostician's experience, and significant variations in performance across different institutions and manufacturers' US systems [3].

Artificial intelligence (AI) methods, particularly deep learning models, have revolutionized ultrasound imaging by automating image processing and enabling automated disease diagnosis and detection of abnormalities. Convolutional neural networks (CNNs) have played a vital role in this transformation, demonstrating remarkable improvements in various medical imaging modalities [4–9]. However, the limitations of CNNs in capturing long-range dependencies and contextual information led to the development of vision transformers (ViTs) [10] in image processing. The self-attention mechanism, a key part of the Transformer, possesses the capability to establish relationships between sequence elements, thus facilitating the learning of long-range interactions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



Significant strides have been made in the vision community to integrate attention mechanisms into architectures inspired by CNNs. Recent research has shown that these transformer modules can potentially substitute standard convolutions in deep neural networks by working on a sequence of image patches culminating in the creation of ViTs.

In recent years, the integration of Vision Transformers into medical US analysis has encompassed a diverse range of methodological tasks. These tasks include traditional diagnostic functions like segmentation, classification, biometric measurements, detection, quality assessment, and registration, as well as innovative applications like image-guided interventions and therapy.

Notably, segmentation, detection, and classification stand out as the fundamental tasks, with widespread utilization across various anatomical structures in medical US analysis. Figure 1 illustrates the distribution of methodological tasks and anatomical application categories in the context of medical ultrasound image analysis. The figure provides insights into the prevalence of these tasks across different anatomical structures, including but not limited to heart [11], prostate [12], liver [13], breast [14], brain [15], lymph node [16], lung [17], pancreatic [18], carotid [19], thyroid [11], intravascular [20], fetus [21], urinary bladder [22], gallbladder, other structures [23].

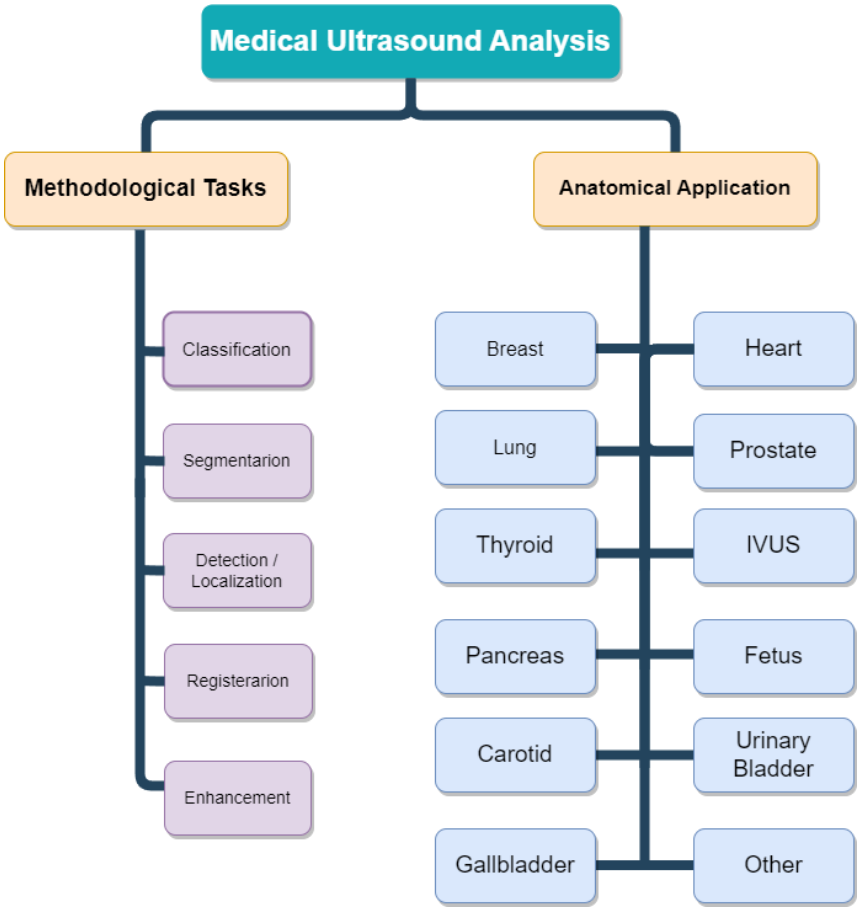


**Figure 1.** Overview of the applications and organs covered in this review.

The review in [24]offers a broad perspective on the applications of vision transformers in medical imaging, encompassing various modalities and imaging techniques. Nonetheless, it lacks a specific focus on ultrasound imaging applications, which are crucial for understanding the unique challenges and opportunities within this specialized field.

This review seeks to fill the knowledge gap in the field of ultrasound imaging by providing a comprehensive examination of Vision Transformer-based AI methods that are specifically designed for this application. Given the unique attributes and diagnostic needs of ultrasound imaging, such an overview can offer invaluable insights for those working in this specialized area. The purpose of this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



review is to offer a thorough analysis of Transformer models that have been specially developed for ultrasound imaging and its associated analysis applications

Figure 2 depicts the trend of published papers related to the intersection of ultrasound and vision transformers since 2017, based on the search query "Ultrasound AND vision transformers" on Google Scholar. The plot clearly demonstrates a substantial increase in the number of publications in this field over the specified period. The rising trend suggests a growing interest and research activity in the application of vision transformers in ultrasound imaging.

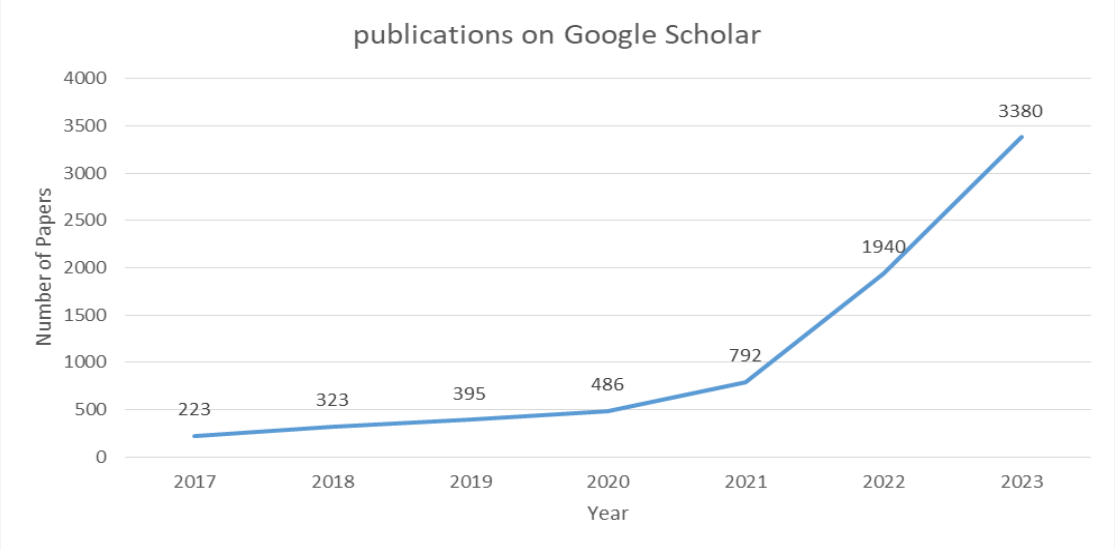


**Figure 2.** numbers of publications of "Ultrasound and vision transformers" on Google Scholar.

This survey paper offers an exhaustive examination of the utilization of transformers in the field of ultrasound imaging. As the first of its kind, it serves to connect the vision and ultrasound imaging communities in this rapidly advancing domain. This analysis involves an extensive review of 231 relevant papers to summarize recent advancements and identify the most pertinent ones for this topic.

Our review paper is divided based on the organs and provides an in-depth analysis of the different tasks such as classification, segmentation, object detection, and image enhancement. In Figure 3, the pie plot visually represents the distribution of the number of papers for each organ considered in the review, providing a comprehensive overview of the focus areas within the medical ultrasound analysis literature.

In conclusion, we offer a comprehensive critique of the current state of the field, pinpointing major challenges, emphasizing unresolved issues, and suggesting potential future paths. The structure of the paper is as follows. Section 2 provides foundational information on the field, with an emphasis on the key principles that underpin transformers. Our review is then segmented based on the organs, covered in Sections 3-1 to 3-14. Lastly, we engage in a thorough discussion of the overall state of the field, identifying significant challenges, spotlighting open problems, and charting promising directions for the future.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



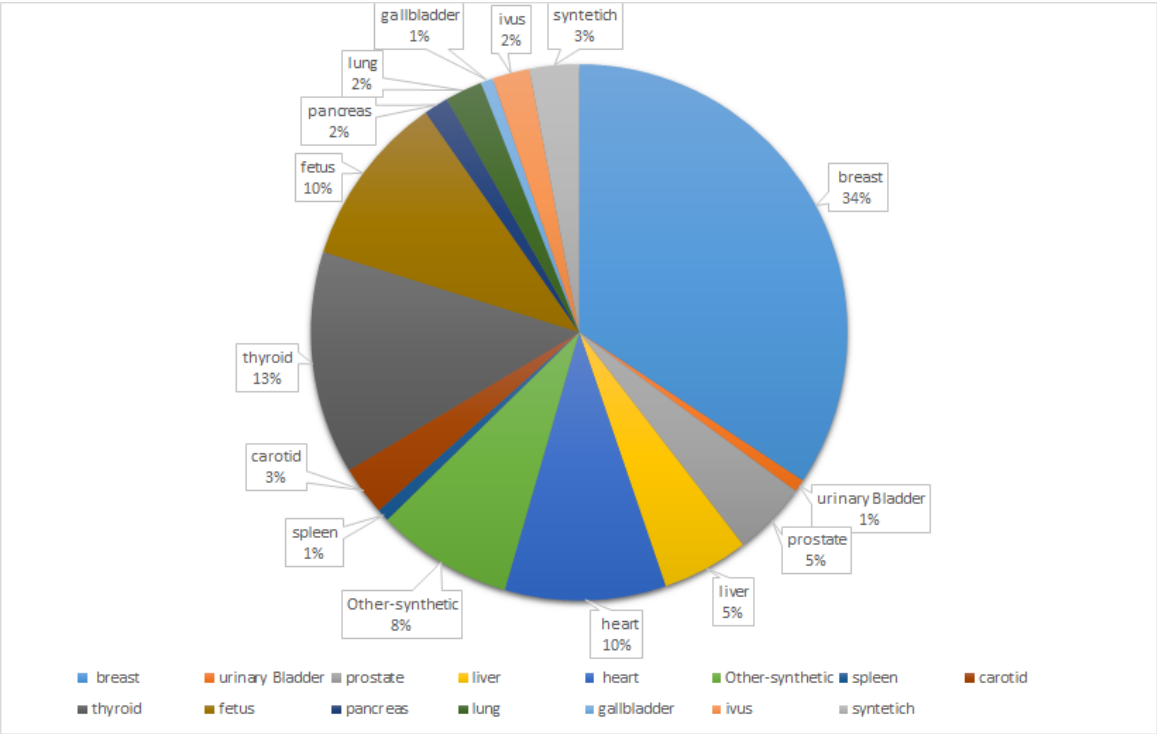


**Figure 3.** Distribution of organs considered in the review paper, with percentages of each organ's representation in the literature.

## 2. background

### *2.1. Fundamentals of Transformer*

Transformers have revolutionized the field of Natural Language Processing (NLP) by providing a fundamental framework for processing and understanding language. Originally introduced by Vaswani et al. in 2017 [25], transformers have since become a cornerstone in NLP, particularly with the development of models such as BERT, GPT-3, and T5.

Fundamentally, transformers represent a kind of neural network architecture that doesn't rely on convolutions. They excel at identifying long-range dependencies and relationships in sequential data, which makes them especially effective for tasks related to language. The groundbreaking aspect of transformers is their attention mechanism, which empowers them to assign weights to the significance of various words in a sentence. This capability allows them to process and comprehend context more efficiently than earlier NLP models.

In addition to their significant impact on NLP, transformers have also shown promise in the field of computer vision. Vision transformers (ViTs) have emerged as a novel approach for image recognition tasks, challenging the traditional convolutional neural network (CNN) architectures.

By applying the self-attention mechanism to image patches, vision transformers can effectively capture global dependencies in images, enabling them to understand context and relationships between different parts of an image. This has led to impressive results in tasks such as image classification, object detection, and image segmentation.

The introduction of vision transformers has also opened up opportunities for cross-modal learning, where transformers can be applied to tasks that involve both text and images, such as image captioning and visual question-answering. This demonstrates the versatility of transformers in handling multimodal data and their potential to drive innovation at the intersection of NLP and computer vision.

Overall, the application of transformers in computer vision showcases their adaptability and potential to revolutionize not only NLP but also other domains of artificial intelligence, paving the way for new advancements in multimodal learning and understanding of complex data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



2.1.1. Self-attention

In transformers, self-attention is a critical element of the attention mechanism that allows the model to focus on various segments of the input sequence and identify dependencies among them [25]. This process involves converting the input sequence into three vectors: queries, keys, and values. The queries are employed to extract pertinent data from the keys, while the values are utilized to generate the output. The attention weights are determined based on the correlation between the queries and keys. The final output is produced by summing the weighted values.

This mechanism is especially potent in detecting long-term dependencies within the input sequence, thereby making it a valuable instrument for natural language processing and similar sequence-based tasks.

To elaborate, before the input sentence is fed into the self-attention block, it is first converted into an embedding vector. This process is known as "word embedding" or "sentence embedding," and it forms the basis for many Natural Language Processing (NLP) tasks. After the embedding vector, the positional information of each word is also included because the position can alter the meaning of the word or sentence. This is done to allow the model to track the position of each vector or word. Once the vectors are prepared, the next step is to calculate the similarity between any two vectors. The dot product is commonly used for this purpose due to its computational efficiency and space optimization. The dot product provides scalar results, which are suitable for our needs. After obtaining the similarity scores, the next step involves normalizing and applying the softmax function to obtain the attention weights. These weights are then multiplied with the original input vector to adjust the values according to the weights received from the softmax function.

2.1.2. Multi-head self-attention

Multi-head self-attention is a strategy used in transformers to boost the model's capacity to grasp a variety of relationships and dependencies within the input sequence [25]. This methodology involves executing self-attention multiple times concurrently, each time with different sets of learned queries, keys, and values. Each set originates from a linear projection of the initial input, offering multiple unique viewpoints on the input sequence (Figure 4).

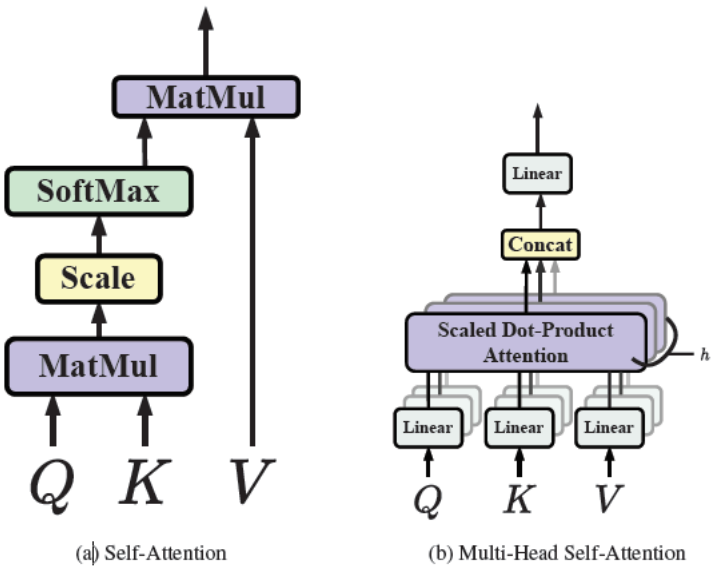


**Figure 4.** Self-attention and multi-head self-attention [25].

Utilizing multiple attention heads allows the model to pay attention to different portions of the input sequence and collect various types of information simultaneously. Once the self-attention process is independently carried out for each head, the outcomes are amalgamated and subjected to a linear transformation to yield the final output. This methodology empowers the model to effectively

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



identify intricate patterns and relationships within the input data, thereby enhancing its overall representational capability.

Multi-head self-attention is a key innovation in transformers, contributing to their effectiveness in handling diverse and intricate sequences of data, such as those encountered in natural language processing and other sequence-based tasks.

### *2.2. transformer architecture*

The architecture of transformers consists of both encoder and decoder blocks(Figure 5), which are fundamental components in sequence-to-sequence models, particularly in tasks such as machine translation [25].

**Encoder**: The encoder is responsible for processing the input sequence. It typically comprises multiple layers, each containing self-attention mechanisms and feedforward neural networks. In each layer, the input sequence is transformed through self-attention, allowing the model to capture dependencies and relationships within the sequence. The outputs from the self-attention are then passed through position-wise feedforward networks to further process the information. The encoder's role is to create a rich representation of the input sequence, capturing its semantic and contextual information effectively.

**Decoder**: The decoder, on the other hand, is tasked with generating the output sequence based on the processed input. Similar to the encoder, it consists of multiple layers, each containing self-attention mechanisms and feedforward neural networks. However, the decoder also includes an additional cross-attention mechanism that allows it to focus on the input sequence (encoded representation) while generating the output. This enables the decoder to leverage the information from the input sequence to produce a meaningful output sequence.

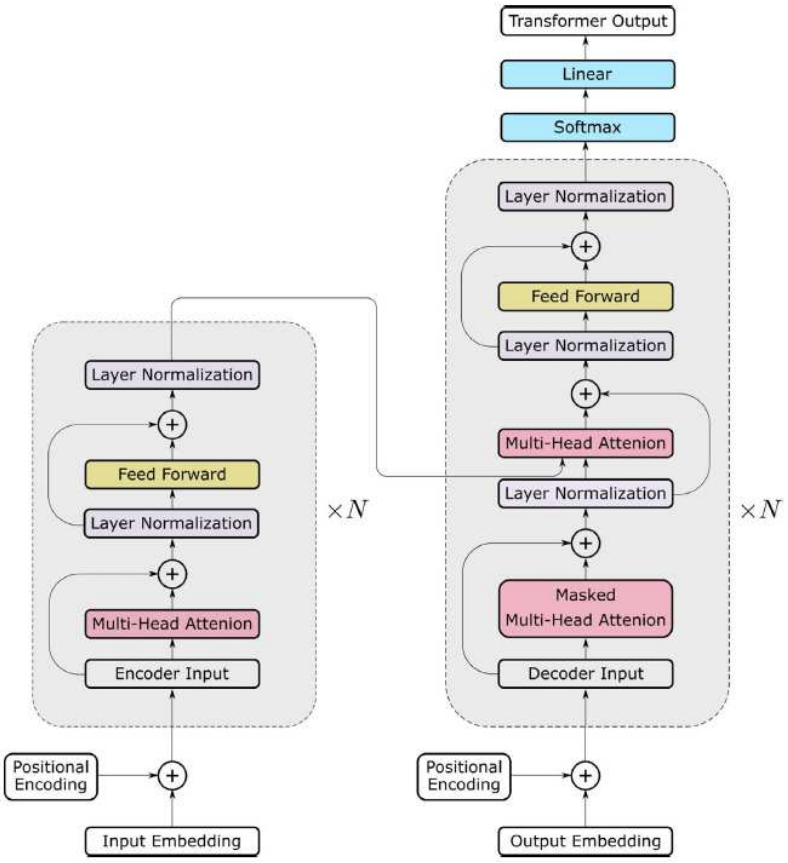


**Figure 5.** Transformer architecture [25].

The encoder-decoder architecture in transformers enables the model to effectively handle sequence-to-sequence tasks, such as machine translation and text summarization. It allows for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



capturing complex dependencies within the input sequence and leveraging that information to generate accurate and coherent output sequences.

### *2.3. vision transformers*

The achievements of transformers in natural language processing have influenced the computer vision research community, leading to numerous endeavors to modify transformers for vision-related tasks. Transformer-based models specifically designed for vision applications have been rapidly developed, with notable examples including the detection transformer (DETR) [26], Vision Transformer (ViT), data-efficient image transformer (DeiT) [27], and Swin-Transformer [28]. These models represent significant advancements in leveraging transformers for computer vision and have gained recognition for their contributions to tasks such as object detection, image classification, and efficient image comprehension.

**DETR**: DETR, standing for DEtection TRansformer, has brought a major breakthrough in the realm of computer vision, particularly in the area of object detection tasks. Created by Carion et al. [26], DETR represents a departure from conventional methods that depended heavily on manual design processes, and demonstrates the potential of transformers in revolutionizing object detection within the field of computer vision. This approach replaces the complex, hand-crafted object detection pipeline with a simpler one based on Transformers. This method simplifies the intricate, manually crafted object detection pipeline by substituting it with a Transformer.

DETR uses a transformer encoder to comprehend the relationships between image features derived from a Convolutional Neural Network (CNN) backbone. The transformer decoder generates object queries, and a feed-forward network is responsible for assigning labels and determining bounding boxes around the objects. This involves a set-based global loss mechanism that ensures unique predictions through bipartite matching, along with a Transformer encoder-decoder architecture. With a fixed small set of learned object queries, DETR considers the relationships between objects and the global image context to directly produce the final set of predictions in parallel.

**ViT**: Following the introduction of DETR, Dosovitskiy et al. [10] introduced the Vision Transformer (ViT), a model that employs the fundamental architecture of the traditional transformer for image classification tasks. As depicted in Figure 6, ViT operates similarly to a BERT-like encoder-only transformer, utilizing a series of vector representations to classify images.

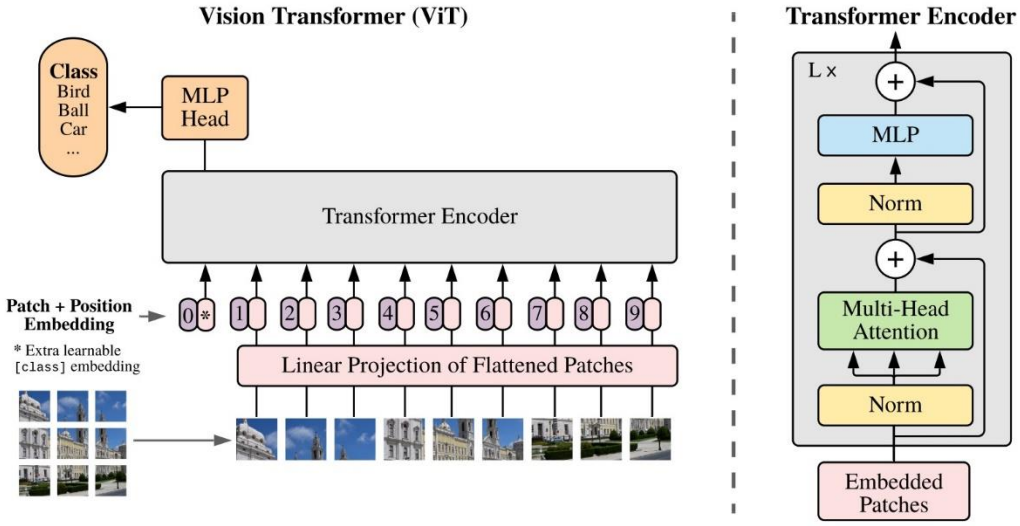


**Figure 6.** Vision transformer overview [10].

The process begins with the input image being converted into a sequence of patches. Each patch is paired by a positional encoding technique, which encodes the spatial positions of the patches to provide spatial information. These patches, along with a class token, are then fed into the transformer.

 doi:10.20944/preprints202401.0309.v1



This process computes the Multi-Head Self-Attention (MHSA) and generates the learned embeddings of the patches. The class token's state from the ViT's output serves as the image's representation. Lastly, a multi-layer perceptron (MLP) is used to classify the learned image representation.

Moreover, ViT can also accept feature maps from CNNs as input for relational mapping, in addition to raw images. This flexibility allows for more nuanced and complex image analyses.

**DeiT:** To address the issue of ViT requiring vast amounts of training data, Touvron et al. [30] introduced Data-efficient Image Transformer (DeiT) to achieve high performance on small-scale data.

In the context of knowledge distillation, a teacher-student framework was implemented, incorporating a distillation token, a term used in transformer terminology. This token followed the input sequence and enabled the student model to learn from the output of the teacher model. They hypothesized that using a Convolutional Neural Network (CNN) as the teacher model could assist in training the transformer as the student network, allowing the student network to inherit inductive bias.

**Swin Transformer** Introduced by Ze Liu et al. in 2021 [28], the Swin Transformer is a transformer architecture known for its ability to generate a hierarchical feature representation. This architecture exhibits linear computational complexity relative to the size of the input image. It's particularly useful in various computer vision tasks due to its ability to serve as a versatile backbone. These tasks include instance segmentation, semantic segmentation, image classification, and object detection.

The Swin Transformer is based on the standard transformer architecture, but it uses shifted windows to process images at different scales. The Swin Transformer is designed to be more efficient than other transformer architectures, such as the ViT, with smaller datasets.

**PVT**: The Pyramid Vision Transformer (PVT) [29] is a transformer variant adept at handling dense prediction tasks. It employs a pyramid structure, enabling detailed inputs (4 x 4 pixels per patch) and reducing the sequence length of the Transformer as it deepens, thus lowering computational cost. PVT comprises several key components: Dense Connections for learning complex patterns, Feedforward Networks for data processing, Layer Normalization for stabilizing learning, Residual Connections (Skip Connections) for mitigating the vanishing gradients problem, and Scaled Dot-Product Attention for calculating input data relevance.

**CvT:** The Convolutional Vision Transformer (CvT) [30] is an innovative architecture that enhances the Vision Transformer (ViT) by integrating convolutions. This enhancement is realized through two primary alterations: a hierarchical structure of Transformers with a new convolutional token embedding, and a convolutional Transformer block that employs a convolutional projection. The convolutional token embedding layer provides the ability to modify the token feature dimension and the quantity of tokens at each level, allowing the tokens to depict progressively intricate visual patterns across wider spatial areas, similar to feature layers in Convolutional Neural Networks (CNNs). The convolutional Transformer block replaces the linear projection in the Transformer module with a convolutional projection, capturing local spatial context and reducing semantic ambiguity in the attention mechanism.

The CvT architecture has been found to exhibit superior performance compared to other Vision Transformers and ResNets on ImageNet-1k. Interestingly, the CvT model demonstrates that positional encoding, a crucial element in existing Vision Transformers, can be safely discarded. This simplification allows the model to handle higher resolution vision tasks more effectively.

**HVT:** The Hybrid Vision Transformer (HVT) [31] is a unique architecture that merges the advantages of convolutional neural networks (CNNs) and transformers for image processing. It capitalizes on transformers' ability to concentrate on global relationships in images and CNNs' capacity to model local correlations, resulting in superior performance across various computer vision tasks. HVTs typically blend both the convolution operation and self-attention mechanism, enabling the exploitation of both local and global image representations. They have demonstrated impressive results in vision applications, providing a viable alternative to traditional CNNs, and have been successfully deployed in tasks such as image segmentation, object detection, and surveillance anomaly detection. However, the specific implementation of the HVT can vary significantly

 doi:10.20944/preprints202401.0309.v1



depending on the task and requirements, with some HVTs incorporating additional components or modifications to further boost performance. In essence, the Hybrid Vision Transformer is a potent tool for image processing tasks, amalgamating the strengths of both CNNs and transformers to achieve high performance

## 3. Organs

In this section, we explain the use of transformer-based methods in ultrasound images, focusing on commonly used organs.

### *3.1. Breast*

Breast cancer is the most common cancer. Providing breast cancer screening and introducing new treatment methods since 1990's reduced mortality of this cancer [32]. The gold standard imaging method for breast cancer detection is mammography. But mammography uses ionizing radiations. Also, mammography is not proper for detecting cancer in dense breasts. Sonography is another imaging system that routinely is used for breast screening. It is harmless, cheap, uses portable systems and provide better results in dense breast cases. Table 1 provides a detailed comparison of transformer-based models for breast ultrasound image analysis.

Two dimensional ultrasound images are mostly used, but recently automatic breast ultrasound systems (ABUS) are provided that produce 3-D scans of the breast containing many 2-D slices. Due to speckle noise and artifacts and also various shapes of breast nodules, analyzing these images is a challenge and need a lot of experience.

There are many attempts to use artificial intelligence for analyzing ultrasound breast images. In recent years vision transformers were considered for this problem.

In [33] 3D U-Net is used with attention mechanism and transformer layers for segmenting ABUS 3-D images. The transformers inserted to consider long distance relations.

In [14] Semi-supervised vision transformer is used for breast cancer classification in 2-D breast ultrasound images. They tackle the problem of image scarcity of breast cancer data bases by using semi supervised vision transformer. They adopted an adaptive token sampler to select informative tokens to reduce computation cost.

Vision transformer with various augmentation strategies is used for breast nodule classification in [34]. They considered three classes of benign, malignant and normal. They showed that transfer learning in the vision transformer has comparable or better performance than transfer learning the CNN. Also, they found that augmentation is not very helpful in transferring pre-trained vision transformers.

Transfer learning of vision transformer by imageNet and histology datasets were used for classifying breast ultrasound images in [35]. The trained Vision Transformer, Vision Transformer with transfer learning, and CNN with transfer learning are compared with their model, and it got the best results. The patch sizes of 16*16 and 32*32 are compared and the smaller patch size provide better results.

Transformer and information bottlenecks based on the UNet model (IB-TransUNet) is used in [36] for ultrasound breast image segmentation. The bottlenecks remove the redundant features and prevent the overfitting.

[37] provides a relatively large breast ultrasound image dataset including 2405 images. They used the fact that most benign tumors growths horizontally and most malignant tumors expand vertically to the deeper tissues, and applied horizontal and vertical transformer to distinguish them without using predefined region of interest.

[38] considered a cross image modelling and a cross image dependency loss to take to account the common features of tumors in different images for segmentation purpose. They combined a CNN based encoder and a transformer-based encoder for getting near and far dependencies. They suggest that this idea can be used for combining different information like elastography and attenuation.

 doi:10.20944/preprints202401.0309.v1



A relatively large dataset consisting 21332 images is used in [39] with a vision transformer for localizing and BI-RADS classification of the nodules. Its outputs are compared with the diagnoses of 6 experienced radiologists.

A single Transformer layer and multiple information bottleneck (IB) blocks are used in [40] for segmenting ultrasound breast images instead of using many transformers that increase complexity and becomes vulnerable to overfitting. They got better results in comparison with TransUNet that uses 12 transformer layers.

A deep supervised transformer full-resolution residual network was presented in [41]. Its feature fusion is better and also suppress irrelevant features while the deep supervision mechanism reduces the gradient vanishing problem. Augmentation is used in training dataset. It took 33msec for segmenting every image, using GPU, which is an acceptable time.

Supervised learning and unsupervised learning are used together in [42] for segmentation and classification of the breast ultrasound images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



**Table 1.** Detailed description of transformer-based breast US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| [33] | 2021 | segmentation | 3D Deep Attentive U-Net with Transformer | Self collected data set | Dice: 0.7636<br>Jc: 0.6214<br>HD:15.47<br>Prec: 0.7895<br>Se: 0.7542<br>Sp: 0.9885 | 3D deep convolution NN |
| [14] | 2022 | classification | Semi-supervised vision transformer | DBUI<br>BreakHis | Acc: 0.981<br>Prec:0.981<br>Rec: 0.986<br>F1-core: 0.984 | a semi-supervised learning ViT |
| [34] | 2022 | classification | Vision transformer/Transfer learning | BUSI<br>UDIAT<br>(goo.gl/SJmoti) | Acc: 0.867<br>AUC: 0.95 | Using various augmentation methods |
| [35] | 2022 | classification | Vision transformer/Transfer learning | BUSI<br>Mendeley breast ultrasound | Acc: 0.919<br>AUC: 0.937<br>F1-core: 0.919<br>MCC score:0.924<br>Kappa score:0.919 | Transfer learning from cancer cell classification |
| [36] | 2023 | segmentation | Transformer and information bottlenecks | Synapse dataset<br>BUSI | DSC: 0.8195<br>HD: 20.35 | multi-resolution fusion to skip connections |
| [37] | 2023 | classification | Horizontal and vertical transformers | UDIAT<br>BUSI<br>GDPH&SYSUCC | AUC: 0.92<br>Acc: 0.893<br>Spe: 0.836<br>Prec: 0.906<br>Rec: 0.926<br>F1-score:0.916 | Deriving horizontal and vertical spatial information |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| [38] | 2023 | Segmentation | Cross-Image Dependency Modeling | BUSI<br>UDIAT | Dice: 0.8577<br>Jc: 0.7899<br>Acc: 0.9733<br>Sp: 0.9894<br>Se: 0.8584 | cross-image dependency module , cross-image contextual modeling; and a cross-image dependency loss |
| [39] | 2023 | Localization/BI-RADS classifications | Vision transformer | Self collected data set | Acc: 0.9489<br>Sp: 0.9509<br>Se: 0.941 | BI-RADS classification |
| [40] | 2023 | segmentation | Transformer and information bottlenecks | BUSI | F1: 0.8078<br>IoU: 0.6775 | Using one transformer layer |
| [41] | 2023 | segmentation | a full-resolution residual stream/ TransU-Net/ transformer | open BUS dataset from the Sun Yat-sen University Cancer Center/ UDIAT | Dice: 0.9104 | deep supervised transformer U-shaped full-resolution<br>residual network |
| [42] | 2024 | segmentation /Classification | Using both supervised and unsupervised learning | BUSI<br>UDIAT | Acc: 0.99907<br>Sp: 0.9766<br>Se: 0.9977 | Tackle the problem of mask unavailability |

Acc: Accuracy, DSC: Dice similarity coefficient, HD: Hausdorff Distance, Jc: Jaccard index, Se: Sensitivity, Sp: specificity, Prec: Precision, Rec: Recall, MCC: Matthews Correlation Coefficient, AUC: Area Under Curve, IoU: Intersection over Union.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



### 3.2. urinary bladder

Bacterial infection can cause cystitis in the urinary bladder. Ultrasound imaging is one of the diagnostic methods of cystitis. The urinary Bladder Wall Thickness (BWT) estimated in ultrasound (US) images to diagnose cystitis [22]. Deep learning is used for segmenting the urinary wall bladder, then feature extraction for classification is used to detect cystitis. CNN is compared with vision transformer. 250 subjects were used where half of them had cystitis. Image augmentation is applied to increase the images to 1000 images.

A U-Net is used for urinary bladder wall, then eight features are derived from the segmented area, and five notable features are chosen between them. A CNN model is applied for classifying the normal and cystitis cases. The result is compared with the results of pre trained CNN and vision transformer.

The best CNN model got 95% for precision, recall, F1 score and accuracy. While the vision transformer got 94%for precision, 89% for recall, 93% for F1 score and 92.1% for accuracy.

### 3.3. Pancreatic

Diagnosis of pancreatic cancer is the most difficult among all cancers. Endoscopic ultrasound (EUS) is the best diagnostic method for this cancer. But EUS images are difficult to analyze. In [18] a ViT based dual self-supervised network (DSN) is presented for classifying EUS images to pancreatic and non-pancreatic cancer. At first a ROI is selected by a multi-operator transformation, and then DSN transfer the unlabeled images to features. In this research also a huge EUS-based pancreas image dataset (LEPset) is gathered that includes 3500 pathologically confirmed EUS images with labels and 8,000 EUS images without label that is publically available [43]. The method is also applied to BUID [44] data set.

### 3.4. Prostate

Prostate cancer is a common cancer between men. Early diagnosis will help the treatment and mortality reduction. A usual method for its initial diagnosis is transrectal sonography. These sonography images are difficult to be labeled, because of its low resolution, noises and artifacts. Therefore, in [45] an unsupervised network is designed to extract features from these images.

ROI of the ultrasound image and biopsy core image are used together for improving cancer detection [12].

A “ROI-scale” network using self-supervised learning extracts features from small ROIs and a “core-scale” transformer model derives a series of features from several ROIs in the needle trace region of prostate biopsy tissue to predict the tissue type. Attention maps are used for localizing the cancer at the ROI region.

A dataset of micro-ultrasound gathered from 578 patients who had prostate biopsy is used for evaluating the method. The model performs better compared to ROI-scale-only models. It got 80.3% AUROC, a statistically significant improvement over ROI-scale classification. The method is compared to other studies on prostate cancer detection with various imaging modalities. Table 2 illustrates an extensive comparison of transformer-based models applied to the analysis of prostate ultrasound images.

**Table 2.** Detailed description of transformer-based prostate US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| [45] | 2023 | Classification | Online-Net and Target-Net. | Self-collected data | Acc:0.8046; Malignant: Precision:0.8267;Recall:0.8662; F1-score:0.7907; Benign: Precision:0.7500;Recall:0.6364; F1-score:0.6885; | a self-supervised dual-head attentional bootstrap learning network (SDABL), including Online-Net and Target-Net. |

 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| [12] | 2023 | Classification | ROI scale and Core scale feature extraction | Self-collected data | Precision: 0.787; Se: 0.880; Sp: 0.512 AUROC: 0.803; | a micro-ultrasound data set with biopsyresults |

Acc: Accuracy, Se: Sensitivity, Sp: specificity, Prec: Precision, Rec: Recall, AUROC: area under the receiver operating characteristic.

### *3.5. Thyroid*

The prevalence of thyroid nodules in adults is between 19% to 67%. There are many attempts to segment and classification of these nodules. Table 3 presents a comprehensive comparison of transformer-based models used in the analysis of Thyroid ultrasound images.

In [46], a boundary attention transformer net (BTNet), by incorporating CNN and transformer short and long range features are fused. A boundary attention block is designed for improving edge information learning. The features are fused at different scales.

In [47] ultrasound images and infrared thermal images are used simultaneously. The features are derived separately by two CNN and transformer encoders to capture local and global features respectively, and these features are fused using a vision transformer. Ultrasound images provide anatomical information and infrared thermal images provide thermodynamic information of the nodules.

A hybrid model of CNN and ViT for diagnosing thyroid nodules is presented in [48]. GAN model is used for data augmentation to overcome the problem of data shortage. They showed that the hybrid model, that combines ResNet50 and ViT_B16, has better performance compared to CNN or ViT when using independently.

To protect parathyroid glands in thyroid surgery using ultrasound images, a network including transformer for considering long range dependency is introduced in [49]. It consists of two encoding and one decoding network for parathyroid glands segmentation. The two branches extract local and global features.

Contrast-enhanced ultrasound (CEUS) can be used for monitoring microvascular perfusion. [50] provides a segmentation and classification method for thyroid nodules using CEUS images. It used a spatiotemporal transformer based CEUS analysis.

In [51] shallow and deep features are fused for classification of thyroid nodules. ROI of the nodule is fed to a CNN network for extracting shape and texture features. Whole image is fed to a Swin Transformer to derive deep features. Then these two group features are combined and fed to a fully connected layer to classify the nodule.

**Table 3.** Detailed description of transformer-based thyroid US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| [46] | 2023 | Segmentation | CNN, Vision Transformer, | Self-collected data the DDTI data set the Breast Ultrasound Images Data Set(BUID) | IoU:0.810, Dice:0.892; | boundary attention transformer net |
| [47] | 2023 | Classification | CNN, Vision Transformer, | Self-collected data | Accuracy:0.9738; Precision:0.9699; Specificity:0.9739; Sensitivity:0.9736; F1-score:0.9717; F2-score:0.9738; | Using ultrasound images and infrared thermal images simultaneously, Using CNN and Transformer for feature extraction and Vision Transformer for feature fusion. |
| [48] | 2023 | Segmentation | CNN, Vision Transformer, | Self-collected data | Dice: 84.76; Jaccard:74.39; Miou:86.5; | Using residual bottlenecks, Transformer bottlenecks, and two |

 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | Recall:83.9; Precision:86.5; | branch down-sampling blocks, and the long-range feature extractor composed of the Vision Transformer. |
| [49] | 2023 | Classification | Hybrid CNN and ViT | Public CIM@LAB | F1:96.67, Recall:95.01, Precision:98.51, Accuracy:97.63, | A hybrid ViT model with a backbone CNN. |
| [50] | 2023 | Segmentation Classification | Swin Transformer | Self-collected data | Dice:82.41; Accuracy:86.59; | The dynamic swin-transformer encoder and multi-level feature collaborative learning are combined into U-net |
| [51] | 2023 | Classification | Hybrid CNN and Swin Transformer, | public dataset DDTI provided by the National University of Colombia, | Accuracy:0.954; Specificity:0.958; Sensitivity:0.975; AUC:0.974; | shallow and deep features are fused for classification. |

*3.6. Heart*

Understanding the heart's complexity and dynamism presents considerable challenges due to its intricate and constantly changing characteristics. These characteristics include detailed structures such as chambers, valves, and vessels that undergo transformations throughout the cardiac cycle.

Despite issues such as speckle noise, shadows, and changes in patient anatomy, several deep learning models have been designed for applications like single image classification [63]. However, these models often fail to consider the dynamic nature of the heart and struggle with signal loss in ultrasound images. Transformers play a crucial role in ultrasound heart imaging, particularly in the analysis of complex temporal dependencies in patient data, which can enhance the prediction of various tasks. Their successful application in heart imaging is evident in various ways, including the detection of End-Systolic (ES) and End-Diastolic (ED) frames in ultrasound videos, heart chamber segmentation, predicting left ventricular ejection fraction (LVEF), Aortic stenosis (AS) detection and severity classification, and assessing the size and function of the right ventricular (RV) in cardiovascular patients. In the field of cardiac imaging, transformers are used to process Patient histories are organized as time series, encompassing a variety of clinical events, enabling the models to decipher intricate temporal patterns over time.

In cardiac imaging, transformers are employed to handle patient histories, which are structured as time series encompassing diverse clinical events. This enables the models to comprehend progressively intricate temporal relationships. The slow changes of the heart's structure and background in echo imaging, along with the resemblance of following frames, emphasizes the necessity to understand the local temporal context and minor spatial modifications in the heart's chambers, valves, and walls for a thorough diagnosis. Table 4 summarized an exhaustive comparison of transformer-based models utilized in the analysis of heart ultrasound images.

Zeng and et al. [52] developed the Multi-Attention Efficient Feature Fusion Network (MAEF-Net), a system that automatically detect ES and ED frames and segments the left ventricle in all cardiac cycles to calculate the LVEF. The system employs a multi-attention mechanism for effective heartbeat feature capture and noise suppression, and integrates a deep supervision mechanism and spatial pyramid feature fusion for improved feature extraction. The method was tested and proven effective on the publicly accessible EchoNet-Dynamic dataset and a private clinical dataset, showing promising results. The mean absolute error (MAE) for detecting ED and ES frames, as well as for predicting LVEF on the public EchoNet-Dynamic dataset, was particularly noteworthy.

Ahmadi et al. [53] examined aortic stenosis severity by focusing on the temporal localization of the opening and closing of the valve, and the shape and mobility of the aortic valve. They applied Temporal Deformable Attention (TDA) in frame-level embedding to enhance the transformers' understanding of locality and a temporal coherent loss to increase sensitivity to minor aortic valve

 doi:10.20944/preprints202401.0309.v1



movements. Finally, they adopted attention weights to identify echo frames with significant clinical importance, prioritizing these frames in the weighted aggregation for the final classification.

Vafeezadeh and his colleagues [54] introduced the CarpNet network to account for the time information of all echocardiography video frames. This was accomplished by combining the transformer network and the Inception_Resnet_V2 convolutional network as a feature extractor. As a result, the performance of mitral valve classification based on the Carpentier criteria was enhanced, surpassing the performance of single-image acquisition.

A novel model is proposed by Qurri [55] that merges the advantages of Convolutional Neural Networks (CNNs) and Transformers in the Unet framework for segmenting the heart in ultrasound images from the CAMUS Dataset. A Transformer is placed at the Unet bottleneck to connect the encoder and the decoder and to capture long-range contextual information. The model presents a new attention module named Three-Level Attention (TLA) at the decoder side, that consists of an Attention Gate (AG), channel attention, and spatial normalization technique. The TLA module enriched the feature map derived from the skip connections. For the encoder, Squeeze-and-Excitation (SE) is applied to the skip connections leaving the encoder, as another type of attention.

Luo et al [56] present a new method for segmenting the heart in images that utilizes multi-scale features and a position-aware attention mechanism. Their approach, based on an inverted pyramid structure, is aimed at extracting contextual information from low-resolution ultrasound images. The network is trained with images at different scales and combines prediction results to improve its contextual awareness. An attention module enhanced with positional encoding information is presented to help the network learn important positional clues, thereby increasing segmentation accuracy. This method is able to capture contextual information at various resolutions, which is especially helpful in comprehending the complexities of the heart's structure and its varying changes throughout the cardiac cycle. The method is verified through rigorous experiments on the EchoNet-Dynamic dataset.

A Co-Attention Spatial Transformer Network (STN) that exploits interframe correlations to improve left ventricle motion tracking between ED and ES frames and strain analysis in noisy 3D echocardiography is introduced by Ahn et al. [57]. This method enhances feature extraction through the utilization of feature cross-correlations, drawing inspiration from speckle tracking techniques. The team introduces an innovative temporal constraint aimed at normalizing the motion field, thereby facilitating the generation of smooth and realistic paths of cardiac displacement over time, all without the need for preconceived notions about cardiac motion. This objective is accomplished by integrating a temporal consistency regularization component into the loss function. Both a synthetic echocardiography dataset and an in vivo porcine 3D+time echocardiography dataset were utilized for thorough performance evaluations.

Tang et al. [58] proposed a novel approach that merges a deformable model with a medical transformer neural network for image segmentation, addressing the challenge of data scarcity in medical imaging. The axial-attention and dual-scale training strategy are applied to mine long-range feature information. The image augmentation strategy effectively applies these techniques to enhance the performance of deep neural networks in medical image processing.

Liao et al. [59] suggested two different Transformer models for LV segmentation in echocardiography. One model utilizes Segformer, while the other combines Swin Transformer and K-Net. The performance of the models on challenging samples that were not easily segmented was also examined. The results confirmed the superiority of the proposed Transformer models over CNN models, even for samples that were not easily segmented by the CNN model. To achieve precise segmentation results, post-processing such as filtering out unnecessary parts is applied.

Zhao et al. [60] have developed an Interactive Fusion Transformer Network (IFT-Net) for the quantitative analysis of pediatric echocardiography. This network constructs a dual-attention pyramid transformer (DPT) branch that models long-range dependencies from spatial and channels, thereby enhancing the learning of global context information. The IFT-Net also incorporates a bidirectional interactive fusion (BIF) unit that merges local and global features interactively. This approach maximizes their preservation and refines the segmentation process. The BIF consists of two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



independent modules: the group feature learning (GFL) and the channel squeeze-excitation (CSE) unit. The anatomical structures are segmented through the decoder network, and the clinical anatomical parameters are measured through key point positioning.

Fazry and his team [61] introduced a new deep learning method for estimating the ejection fraction from echocardiogram videos, eliminating the need for left ventricle segmentation. This approach, known as UltraSwin, leverages hierarchical vision Transformers and Swin Transformers to extract spatio-temporal features. UltraSwin comprises two primary modules: the Transformers Encoder (TE), which serves as a feature extractor, and the EF Regressor, which functions as a regressor head. The method was evaluated on the EchoNet-Dynamic dataset.

Hagberg et al. [62] created a deep learning model that employs Natural Language Processing (NLP) to evaluate the size and functionality of the right ventricle (RV) from echocardiographic images. They established a pipeline for automatic annotation of video loops, which formed the basis for constructing two image classification models. These models were trained on labels generated through a combination of manual annotation and NLP models. The models were then employed to assess RV function and size. The RV size and function models were 12-layer BERT models, which were pre-trained on a large Swedish dataset.

Ultrasound videos, which can have varying lengths and cardiac cycles of different durations, often require more sophisticated processing methods than traditional frame-by-frame approaches. This is because such methods can overlook the temporal information encoded within the videos. In order to incorporate spatio-temporal support within deep convolutional networks, heuristic frame sampling methods are typically applied to create a stack of chosen frames from videos. In a research conducted by Reynaud et al. [11], a transformer architecture known as the Residual Auto-Encoder Network was utilized along with a BERT model to automatically identify the Early Systole (ES) and Early Diastole (ED) frames in ultrasound videos. This was done to compute the Left Ventricular Ejection Fraction (LVEF).

**Table 4.** Detailed description of transformer-based heart US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| MAEF-Net [52] | 2023 | Segmentation and Detection | dual attention (DA) mechanism + atrous spatial pyramid pooling (EASPP) | EchoNet-Dynamic (10,030 videos) Private clinical dataset (2129 images) | DSC: 0.9310 MAE: 0.9281 | Captured heartbeat features, minimized noise, integrated a deep supervision mechanism, and employed spatial pyramid feature fusion |
| [53] | 2023 | Aortic stenosis (AS) detection and severity classification | Temporal Deformable Attention (TDA)+MLP+Transformer | Private AS Dataset: 2247 patients and 9117 videos public dataset: TMED-2 577 patients | Acc(AS detection on private and dataset): 0.952 and 0.915 Acc(classification on private and dataset): 0.781 and 0.838% | Implemented a temporal loss method to boost sensitivity towards subtle movements of the Autonomic Vascular (AV) system. Applied temporal attention mechanisms to merge spatial data with temporal contextual information Automatically identified key echo frames for classifier |
| CarpNet [54] | 2023 | classification | transformer network +Inception_Resnet_V2 | Private Dataset: 1773 case | Acc: 0. 71 | The initial public unveiling of the application of the Carpentier functional classification in echocardiographic videos of the mitral valve |
| Improved UNet [55] | 2023 | Segmentation | CNNs(Squeeze-and-Excitation (SE)) and Transformer | CAMUS Dataset | DSC(for ED): 0.9252 HD(for ED): 11.04mm | The proposed network architecture includes the introduction of the Three-Level Attention (TLA) |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | DSC(for ES): 0.9264<br>HD(for ES): 12.35mm | module, utilizing attention mechanisms<br>The TLA module boosts the feature embedding.<br>A Transformer is integrated at the bottleneck. |
| Position Attention [56] | 2023 | Segmentation | Position Attention Block + Atrous Spatial Pyramid Pooling (ASPP) | EchoNet-Dynamic dataset | DSC: 0.9145<br>Precision: 0.9079;<br>Recall: 0.9278;<br>F1-score: 0.9177<br>Jc: 0.8847 | employing bicubic interpolation to produce high-resolution images,<br>It integrates a position-aware attention to capture positional knowledge. |
| Co-attention spatial transformer [57] | 2023 | Tracking | Co-Attention Spatial Transformer Network (STN) | synthetic dataset + an in vivo 3D echocardiography dataset | MTE: 0.99 | implementation of a Spatial-Temporal Co-Attention Module within 3D Echocardiography |
| [58] | 2023 | Segmentation | gated axial attention | 480 transverse images | DSC: 0.919 | The network leveraged axial attention and dual-scale training to obtain detailed insights from long-range features, enabling the model to focus on important areas. ensuring its applicability across a wide range of medical imaging scenarios. |
| Segformer + Swin Transformer and K-Net [59] | 2023 | Segmentation | Mixed Vision Transformer + Lightweight Segformer | EchoNet-Dynamic dataset | DSC(for Swin and Segformer): 0.9292 and 0.9279 | The technique employs basic post-processing by discarding segments with the largest pixel square, leading to more accurate segmentation outcomes.<br>Two exclusive Transformer automated deep-learning strategies are introduced for Left Ventricle (LV) segmentation in echocardiography. These strategies aim to enhance missegmented outcomes via post-processing. |
| IFT-Net [60] | 2022 | Segmentation | interactive fusion transformer network (IFT-Net) | 4485 A4C and 1623 PSAX echocardiography of pediatric dataset + CAMUS | Acc:0.954<br>DSC($LV_{Endo}$ and $LV_{Epi}$): 0.9049 and 0.8046 | The novel interaction established between the convolution branch and the transformer branch enables bidirectional fusion of local features and global context information.<br>A parallel network of Dual-Path Transformers (DPT) and Convolutional Neural Networks (CNN) was introduced, enabling effective fusion of local and global features through full-process dual-branch feature interactive learning.<br>this system has been applied to perform an automatic quantitative analysis of pediatric echocardiography. |
| UltraSwin [61] | 2022 | estimate the ejection fraction | hierarchical vision Transformers | EchoNet-Dynamic dataset | MAE: 5.59 | calculating ejection fraction without requiring left ventricle segmentation |

 doi:10.20944/preprints202401.0309.v1



| | | | | | |
|---|---|---|---|---|---|
| Semi-supervised learning with NLP [62] | 2022 | right ventricular (RV) function and size classification | text classification with 12-layer BERT model | 12,684 examinations with Swedish text dataset | Se and Sp (Text classifier for RV size): 0.98 and 0.98<br>Se and Sp (Text classifier for RV function): 0.99 and 0.98<br>Acc (A4C and view classification): 0.92 and 0.73 Se and Sp (The image classifier for RV size and function): 0.8 and 0.85<br>Se and Sp (The image classifier for RV function): 0.93 and 0.72 | Developed a pipeline for automatic image assessment using NLP models.<br>Utilized model-annotated data from written echocardiography reports for training.<br>Achieved significant improvement in sensitivity and specificity for identifying impaired RV function and enlarged RV.<br>Demonstrated the potential of integrating auto-annotation within NLP applications.<br>Showcased the capability for fast and cost-effective expansion of the training dataset. |
| Ultrasound Video Transformers [11] | 2021 | ES/ED detection and LVEF estimation | BERT model and Residual Auto-Encoder Network | Echonet-Dynamic dataset | Average Frame Distances of 3.36 Frames for ES and 7.17 Frames for ED, MAE(LVEF): 5.95<br>R2(LVEF): 0.52 | Developed an end-to-end learnable approach that allows for ejection fraction estimation without the need for segmentation.<br>Introduced a modified transformer architecture capable of processing image sequences of varying lengths. |

Acc: Accuracy, DSC: Dice similarity coefficient, MAE: mean absolute error, ED: End-diastolic, ES: end-systolic, LVEF: left ventricular ejection fraction, HD: Hausdorff Distance, Jc: Jaccard index, MTE: median tracking error, $LV_{Endo}$: left ventricular endocardium, $LV_{Epi}$:left ventricular epicardium, Se: Sensitivity, Sp: specificity.

### *3.7. Fetal*

Researchers have developed innovative methods for analyzing fetal obstetric ultrasound imagery, leveraging the power of transformer and Convolutional Neural Network (CNN) architectures. Rahman and his team [64] have enhanced the precision of identifying fetal planes from ultrasound images by training the Swin Transformer. They have also improved image quality through the use of Histogram Equalization and Fuzzy Logic-based contrast enhancement. Table 5 provides a thorough evaluation of transformer-based models employed in the analysis of fetal obstetric ultrasound images.

A transformer-based image classification approach using a newly designed residual cross-variance attention (R-XCA) block named COMFormer was introduced for categorizing maternal-fetal and brain anatomical structures within 2D fetal US images [65]. The structures are divided into two primary categories: maternal-fetal (which includes the brain, abdomen, thorax, femur, and the mother's cervix among others), and brain anatomical structures (such as trans-ventricular, trans-cerebellum, trans-thalamic, and non-brain structures). A significant feature of the R-XCA block is the use of residual connections, which help mitigate the vanishing gradients problem and enhance the learning process of COMFormer. The performance of this architecture was assessed using a widely accessible dataset known as "BCNatal" for two separate classification tasks.

In another study, Arora et al. [66] explored the application of the vision transformer as a machine learning method to analyze the texture of placental ultrasound images during the first, second, and third trimesters of pregnancy. This was achieved through a prospective observational study that involved the collection of 2D placental US images at different stages of pregnancy.

Chen and his team [67] have introduced the Children Intussusception Diagnosis Network (CIDNet), a comprehensive artificial intelligence algorithm designed for swift diagnosis of intussusception in children using ultrasound images. The system utilizes a transformer-based approach and a Multi-Instance Deformable Transformer Classification (MI-DTC) module, which includes a pre-processing component. This module is engineered to precisely identify and locate abnormal regions related to intussusception in ultrasound images. The team also incorporated several Convolutional Neural Network (CNN)-based algorithms as the backbone networks.

Qiao and colleagues [68] proposed a dual-path chain multi-scale gated axial-transformer network (DPC-MSGATNet) that models both global dependencies and local visual cues for fetal US four-chamber (FC) views for segment heart chambers, supporting clinicians in studying cardiac anatomy and aiding in the identification of fetal congenital heart defects (CHD). This model enables precise segmentation of the four chambers, assisting clinicians in analyzing cardiac morphology and aiding in the diagnosis of fetal congenital heart defects (CHD). The DPC-MSGATNet consists of a local and a global branch that operate concurrently on an entire FC view and image patches to learn multi-scale representations. To enhance the interactions between these branches, an interactive dual-path chain gated axial-transformer (IDPCGAT) module has been designed.

In a landmark study (For the first time) [69], Płotka et al. introduced an innovative system for predicting fetal birth weight (FBW) known as BabyNet. This system leverages multimodal data and a visual data processing component, effectively integrating Transformers and CNNs. The hybrid model enhances the 3D ResNet-18 architecture by incorporating a Residual Transformer Module (RTM). This module refines features through a global self-attention mechanism, residual connections, and facilitates both local and global feature representation. The architecture of BabyNet was further developed in their subsequent research [71]. The convolutional component identifies local image patterns and interactions, while the transformer component models long-term dependencies and relationships. A module is implemented in the deeper layers of BabyNet to conditionally shift feature maps based on non-imaging data, such as gestational age. Following up on their initial work, Płotka et al. unveiled BabyNet++ [70], a unique network specifically engineered for FBW prediction using multimodal data. This network uses a custom RTM and incorporates Dynamic Affine Feature Transform Maps (DAFT) to efficiently incorporate clinical data within the model structure. This approach evaluates 2D+ t spatio-temporal features in fetal US videos using tabular clinical data.

Yang et al. [21] proposed a one-stage network for automatic measurement of fetal head circumference (HC) using ultrasound images, without any post-processing. This system detects the fetal head position and ellipse parameters utilizing an anchor-free method. Their network combines a simple transformer with a CNN to extract global and local features, and uses a soft stage-wise regression (SSR) strategy and an IOU loss term to improve the accuracy of rotating elliptic object detection. The network is the first of its kind to directly measure fetal HC, marking a significant advancement in the field.

Finally, Zhao et al. [72] designed a landmark retrieval-based method for guiding US-probe movement, which constructs a set of landmarks around a virtual 3D fetal model and compares the current ultrasound image to the landmarks' global descriptors using a deep neural network (DNN) model. Their method uses a Transformer-VLAD network to learn the global descriptors, and avoids human annotation by using a KD-tree search of 3D probe positions to generate training data in a self-supervised way. This approach is intuitive and suitable for human operators, and it avoids costly human annotation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



**Table 5.** Detailed description of transformer-based fetal US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| fetal plane detection [64] | 2023 | Classification | Swin Transformer+ Evidential Dempster–Shafer Based CNN | BCNatal: 12400 images | Acc: 0.889 | Utilized an Evidentiary classifier, specifically the Dempster Shafer Layer, in conjunction with a custom-designed CNN for fetal plane detection. Implemented an end-to-end learnable approach for sample classification exploring the effects of the Swin Transformer, infrequently used in ultrasound fetal planes analysis |
| COMFormer [65] | 2023 | Classification | residual cross-variance attention (R-XCA) | BCNatal: 12, 400 images | Acc(maternal-fetal): 0.9564 Acc(brain anatomy): 0.9633 | The COMFormer model employs a R-XCA block, leveraging residual connections to decrease gradients and boost the learning process. |
| placental ultrasound image texture evolution [66] | 2023 | Classification | vision transformer (ViT) | 1008 cases | Acc(T1&T2 images): 0.6949 Acc(T2&T3 images): 0.7083 Acc(T1&T3 images): 0.8413 | Evaluated three deep learning models and found that the transfer learning model achieved the highest accuracy. |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| CIDNet [67] | 2023 | Classification | MI-DTC (multi-stance deformable transformer classification) | 9999 images | balance Acc(BACC): 0.8464<br>AUC: 0.9716 | Utilized four CNN based model as backbone networks for pre-processing.<br>Implemented an effective cropping procedure in the pre-processing module.<br>multi-weighted new loss function led to improvement<br>application of Gaussian blurring curriculum was confirmed to fix the texture bias. |
| DPC-MSGATNet [68] | 2023 | Segmentation | Interactive dual-path chain gated axial-transformer (IDPCGAT) | 556 FC views | F1 score: 0.9687<br>IoU: 0.9399 | DPC-MSGATNet was developed with a global and a local branch network allows for the simultaneous handling of the full image and its smaller segments |
| BabyNet [69] | 2022 | regression | Residual Transformer Module in the 3D ResNet | 225 2D fetal ultrasound videos | MAPE: 7.5 + 0.66 | present a new methodology for predicting birth weight, which is derived directly from fetal ultrasound video scans.<br>leverages a novel Residual Transformer Module |
| BabyNet++ [70] | 2023 | regression | Residual Transformer with Dynamic Affine Feature Transform Maps (DAFT) | 582 2D fetal ultrasound videos | MAPE: 5.1 + 0.6 | Demonstrated that BabyNet++ outperforms expert clinicians<br><br>Proved that BabyNet++ is less sensitive to clinical data |
| [71] | 2023 | regression | BabyNet | 900 routine fetal ultrasound examinations | MAPE: 3.75+ 2.00%. | There is no significant difference observed between fetal weight predictions made by human |

 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | experts and those generated by a deep network |
| RDHCformer [21] | 2022 | Segmentation | Integrating Transformer and CNN | HC18 dataset | MAE ± std (mm): 1.97±1.89 | rotating ellipse detection method was employed for skull edge detection, based on the anchorfree method. To address the challenge of angle regression, a Soft Stagewise Regression (SSR) strategy was introduced Kullback-Leibler Divergence (KLD) loss was incorporated into the total loss function to enhance the regression accuracy |
| Transformer-VLAD [72] | 2021 | image retrieval | Transformer-VLAD(vector of locally aggregated descriptors) | ScanTrainer Simulator(535,775 US images) | recall@top1: 0.834 | The task of directing the movement of the US probe was addressed as a landmark retrieval issue, utilizing a learned descriptor search method. A Transformer-VLAD network was specifically developed to facilitate automatic landmark retrieval. |

Acc: Accuracy, MAE: mean absolute error, AUC: Area Under Curve, IoU: Intersection over Union, MAPE: Mean Absolute Percentage Error, std: standard deviation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



*3.8. Carotic*

Atherosclerosis, a common cause of ischemic heart disease and stroke, is typically monitored by physicians through the analysis of various anatomical and biomechanical properties of carotid plaques over several cardiac cycles. Ultrasound (US) imaging plays a crucial role in this process, providing a non-invasive method for visualizing, evaluating, and screening carotid atherosclerotic plaque. It enables radiologists to accurately segment these plaques and extract key features such as size, shape, and echo strength, thereby significantly improving early diagnosis and treatment strategies for carotid atherosclerosis. Despite the computational challenges associated with using transformers for analyzing carotid US videos, the advent of several transformer-based networks for ultrasound medical video analysis signals a promising advancement in this field.

LIN et al. [73] developed a model called the U-shaped CSWin Transformer (U-CSWT) for the purpose of automatically segmenting the lumen-intima boundary (LIB) and media-adventitia boundary (MAB) in 3-D ultrasound images of the carotid artery (CA) from 3-D ultrasound images. The U-CSWT, which is composed of hierarchical CSWT modules in both its encoder and decoder, is designed to extract comprehensive global context information from the 3-D image. The U-CSWT's U-shaped structure and the inclusion of the CSWin transformer in the encoder and decoder allow to model long-range dependence while reducing the model's computational complexity. This process involved descriptor learning via contrastive learning, using self-constructed anchor-positive-negative ultrasound image pairs.

Li et al. [74] proposed a new video analysis transformer-based network, known as BP-Net, which is guided by target boundary and perfusion features and is designed to assess the integrity of the fibrous cap using B-mode US and contrast-enhanced US (CEUS) videos. Building on their previously proposed plaque auto-tracking network, they introduced a plaque edge attention module and reverse mechanism to focus the dual video analysis on the fiber cap of plaques. To extract the most valuable features from the fibrous cap, they proposed a feature fusion module. Finally, they integrated a multi-head convolution attention into a transformer-based network to evaluate the integrity of fibrous caps accurately. This approach captures both semantic features and global context information.

Lastly, Hu et al. developed the RMFG_Net [19], a network designed for automatic segmentation of atherosclerotic carotid plaques in ultrasound videos. This network uses a transformer-based algorithm for stable plaque positioning, extracts spatial and temporal features across video frames for high-quality segmentation, integrates a spatial-temporal feature filter to suppress noise and enhance target area detail, applies multi-layer gated computing for feature fusion and adequate feature map aggregation, and is trained end-to-end, eliminating the need for additional operations. Furthermore, it can process at a speed of 68 frames per second. Table 6 illustrate a detailed assessment of transformer-based models used in analyzing carotic ultrasound images.

**Table 6.** Detailed description of transformer-based carotic US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| U-CSWT [73] | 2023 | Segmentation | U-shaped CSWin transformer | 213 3-D ultrasound Images | DSC (MAB in the common carotid artery): 0.946<br>DSC (LIB in the common carotid artery): 0.908 | This method employs a novel approach to descriptor learning, which is accomplished through contrastive learning. This technique makes use of self-constructed anchor-positive-negative pairs of ultrasound images. |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| BP-Net [74] | 2023 | classification | boundary and perfusion network (BP-Net) + multi-modal fusion block | 245 US and CEUS videos | Acc: 0.9235<br>AUC: 0.935 | a multi-modal fusion block has been incorporated to delve deeper into the internal/external characteristics of the plaque and highlight more influential features across US and contrast-enhanced ultrasound (CEUS) videos. It capitalizes on the sturdiness of CNN and the refined global modeling of Transformers, leading to more precise classification results. |
| RMFG_Net [19] | 2023 | Segmentation | Transformer-based Cross-scale Spatial Location (TCSL) | DT dataset: 157 | DSC: 0.8598<br>IoU: 0.7922<br>HD (mm): 11.66 | A proposed Spatial–Temporal Feature Filter (STFF) learns more target information from low-level features<br>a multilayer gated fusion model is introduced for efficient information propagation, reducing noise during fusion. |

Acc: Accuracy, DSC: Dice similarity coefficient, HD: Hausdorff Distance, Se: Sensitivity, Sp: specificity, AUC: Area Under Curve, IoU: Intersection over Union.

### *3.10. Lung*

Various studies have explored the use of transformers in the lung organ, particularly in relation to COVID-19 data. Nehary et al. [75] discuss the application of deep learning and hand-crafted features for classifying lung ultrasound images to detect COVID-19. Their proposed method involves a fusion of Histogram of Oriented Gradients (HOG) features with abstract features from deep learning models like VGG16 and Vision Transformer (ViT) to enhance detection accuracy. The effectiveness of this fusion technique is demonstrated using a public COVID-19 dataset, showing improved classification accuracy when HOG features are fused with abstract features from VGG16 and ViT .

Perera et al. introduced POCFormer [17], a lightweight transformer architecture designed for COVID-19 detection using point-of-care ultrasound. The architecture, consisting of a vision transformer and a linear transformer, is compact with around 2 million parameters, making it suitable for deployment on low-power devices like smartphones. It can run in real-time and has potential for use in rural and underserved areas. POCFormer outperforms other architectures in binary and multiclass classification experiments, demonstrating high accuracy in distinguishing between COVID-19 and healthy patients, as well as COVID-19 and bacterial pneumonia.

Xing et al. [76] proposed a semi-supervised, frame-to-video-based lung ultrasound (LUS) scoring model for diagnosing respiratory diseases. The model consists of two components: a frame-level (FL) scoring model and a video-level (VL) scoring model. The FL model uses a dual attention vision transformer (DaViT) to extract local and global features from LUS frames, which are manually scored by clinicians. The VL model employs a frame-to-video approach, using a 40-channel input with a patch embedding layer and transferring DaViT parameters from the FL model to each channel. It uses a long-short term memory (LSTM) module for correlation analysis of the 40-channel output and a final MLP head for video scoring. The model achieves high accuracy in both FL and VL scoring, with 95.08% and 92.59% accuracy, respectively. Table 7 provides a comprehensive evaluation of transformer-based models that have been applied in the analysis of lung ultrasound images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



**Table 7.** Detailed description of transformer-based lung US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| Nehary [75] | 2023 | Classification | Vision transformer (ViT) | lung ultrasound images (LUS) dataset: 202 | Acc: 0.8666 | the advantages of ViT models include their ability to extract abstract features, leverage transfer learning, utilize transformer encoding for spatial context understanding, and perform accurate final classification |
| POCFormer [17] | 2023 | classification | vision transformer and a linear transformer | 212 US videos | Acc: 0.939 | lightweight transformer architecture |
| DaViT [76] | 2023 | Segmentation | a dual attention vision transformer (DaViT) | LUS dataset: 202 | Acc(FL scoring): 0.9508 Acc(VL scoring): 0.9259 | using a long-short term memory (LSTM) module for correlation analysis |

Acc: Accuracy

*3.11. Liver*

Transformer models have indeed been used for tasks related to liver ultrasounds, specifically for the classification of liver lesions. One notable example is the TransLiver model, a hybrid Transformer model designed for multi-phase liver lesion classification.

Zhang et al. [77] disscused the use of deep learning techniques, specifically a Vision-Transformer (ViT)-based classification method, for the automatic recognition of standard liver sections in ultrasound images. The research aims to address subjective errors in traditional manual scanning and standardize the medical examination of the liver in adults. The authors collect 12 common liver ultrasound standard sections and train the ViT model on these, achieving an accuracy of 92.9% in the available ultrasound dataset. The ViT model outperforms other deep learning frameworks and shows promising results for the recognition of standard liver sections. The research contributes to the study of adult organs, as previous research has mainly focused on fetal organs. The dataset used in this research was partitioned into a 4:1 ratio for model training, validation, prediction, and testing. The document also mentions the use of visual attention mechanisms and targeted histogram equalization to enhance the recognition and contour information in the ultrasound images. Table 8 gives a comprehensive overview of how transformer-based models have been utilized in the analysis of liver ultrasound images.

Dadoun et al. [13] discussed a study on the use of deep learning networks, specifically Faster R-CNN and DETR, for detecting, localizing, and characterizing focal liver lesions (FLLs) on abdominal ultrasound images. The networks were trained on a dataset of 1026 patients and tested on 48 additional patients. DETR outperformed Faster R-CNN and was comparable to or exceeded the performance of three caregivers in detecting FLLs, localizing lesions, and characterizing FLLs as benign or malignant. The study suggests that these networks, particularly DETR, could assist non-expert caregivers in screening patients at high risk of malignancy, potentially improving early detection of hepatocellular carcinoma. However, the study had limitations, including a limited number of images in the test set and the retrospective nature of the study. Further research is needed to validate these findings and explore the integration of clinical information in the screening process.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



Zhang et al. [78] introduced the use of an Ultra-Attention structured perception strategy for the automatic recognition of standard liver sections in ultrasound imaging. This deep learning approach, inspired by natural language processing attention mechanisms, amplifies small features in ultrasound images that may be overlooked. The Ultra-Attention model, guided by a convolutional neural network, addresses the challenge of accurately identifying standard sections by considering the coupling of anatomic structures within the images. It uses a modularized approach where each local piece of information contributes to the final decision, rather than focusing solely on local areas like traditional convolutional neural networks. The Ultra-Attention structure consists of multiple encoder layers, each performing attention operations on the ultrasound images. It uses a modularized approach where each local piece of information contributes to the final decision. The model incorporates dropout mechanisms and Part-Transfer Learning to enhance robustness and convergence. With a classification accuracy of 93.2%, the Ultra-Attention model outperforms traditional convolutional neural network methods, offering a promising solution for improving the accuracy and efficiency of ultrasound diagnosis.

**Table 8.** Detailed description of transformer-based liver US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| [77] | 2022 | Classification | Vision transformer (ViT) | 13970 images | Acc: 0.929 | standardize the medical examination of the liver in adults |
| DETR [13] | 2022 | Detection | vision transformer and a linear transformer | 1026 patients | Sp: 0.90 Se: 0.97 | detecting, localizing, and characterizing focal liver lesions |
| Ultra-Attention [78] | 2023 | Classification | Transformer | 14900 images | Acc: 0.932 | accurately identifying standard sections by considering the coupling of anatomic structures within the images |

Acc: Accuracy, DSC: Dice similarity coefficient, HD: Hausdorff Distance, Se: Sensitivity, Sp: specificity, AUC: Area Under Curve, IoU: Intersection over Union

### *3.12. IVUS*

Transformer models have indeed found applications in the analysis of intravascular ultrasound (IVUS) images.

Huang et al. [79] proposed a framework, POST-IVUS, for automated segmentation of lumen and external elastic membrane (EEM) boundaries in intravascular ultrasound (IVUS) images. This framework addresses the challenges of IVUS segmentation, such as inter-observer variability and the presence of artifacts, by combining Fully Convolutional Networks (FCNs) with temporal context-based feature encoders, a selective transformer module, and a temporal constraining and fusion module. The POST-IVUS framework has shown superior performance compared to state-of-the-art methods, with a Jaccard measure of 0.92 for lumen and 0.94 for EEM segmentation. It has been integrated into a software called QCU-CMS for user-friendly automated IVUS image segmentation, demonstrating its potential for practical applications.

The proposed framework for IVUS segmentation includes two temporal context-based feature encoders, the rotational alignment encoder and the visual persistence encoder, which focus on relevant vessel movement and encode residual visual features, respectively. The Selective Transformer module in the STR U-Net enhances the inference ability of the segmentation model,

 doi:10.20944/preprints202401.0309.v1



particularly in regions with little visual information, by mimicking the perceptual organization property of human vision and capturing long-range dependencies and global context. The SWIN transformer, a key component of the framework, is used as the backbone of the inference branch in the STR U-Net. It introduces connections between areas by dividing images into different patches and calculating hierarchical representations, thereby improving the accuracy of boundary prediction in challenging areas.

The document discusses the Multilevel Structure-Preserved Generative Adversarial Network (MSP-GAN) [20], a proposed method for domain adaptation in intravascular ultrasound (IVUS) analysis. The MSP-GAN addresses the poor generalizability of IVUS analysis methods due to the diversity of IVUS datasets by integrating a vision transformer, a superpixel-wise multiscale contrastive (SMC) constraint, and an uncertainty-aware teacher-student consistency (TSC) constraint. These components work together to effectively preserve structures at global, local, and fine levels, improving the generalizability of IVUS analysis methods. The vision transformer, incorporated into the generator of the MSP-GAN, maintains global pathology information during the image translation process by capturing long-range dependencies and understanding the global context of the images. This enhances the structural similarity between the synthetic and source images, improving the accuracy of downstream IVUS analysis methods such as vessel and lumen segmentation and stenosis-related parameter quantification. The document also discusses the Transformer-incorporated generator, a key component of the MSP-GAN, which preserves global pathology information during the image translation process by combining the strengths of convolutional networks and transformers. It captures both local interactions and long-range dependencies in IVUS images. The generator comprises a convolution-based encoder for efficient visual feature learning and a vision transformer for modeling complex relations of feature components and extracting global information. The outputs of the encoder and transformer are fused to generate context-rich features, which are then decoded into the synthetic image. By incorporating the vision transformer, the generator can interpret the global context of IVUS images, maintain the global pathology information presented in the source images, and improve the structural similarity between the synthesized and source images. Table 9 provides an exhaustive summary of the application of transformer-based models in the analysis of IVUS ultrasound images.

**Table 9.** Detailed description of transformer-based IVUS image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| POST-IVUS [79] | 2023 | segmentation | selective transformer | IVUS-2011 | Jac: 0.92 | Segmentation by combining Fully Convolutional Networks (FCNs) with temporal context-based feature encoders |
| MSP-GAN [20] | 2023 | classification | vision transformer and a linear transformer | 212 US videos | Acc: 0.939 | domain adaptation in IVUS |

Acc: Accuracy, DSC: Dice similarity coefficient, HD: Hausdorff Distance, Se: Sensitivity, Sp: specificity, AUC: Area Under Curve, IoU: Intersection over Union

 doi:10.20944/preprints202401.0309.v1



*3.13. Gallbladder*

Basu et al. proposed RadFormer [23], a novel deep neural network architecture for accurate and interpretable detection of Gallbladder Cancer (GBC) from Ultrasound (USG) images. RadFormer combines global and local attention mechanisms using a transformer-based approach. It outperforms human radiologists in detection accuracy and provides interpretable explanations for its decisions. These explanations are based on visual bag-of-words style feature embeddings that can be mapped to radiological features used in medical literature. The model demonstrates high sensitivity and specificity in detecting GBC from USG images and allows for the discovery of new visual features relevant to GBC diagnosis.

RadFormer uses a global branch to extract deep features from the entire ultrasound image and a local branch to generate a region of interest (ROI) and extract deep features using a bag-of-features (BOF) technique. These features are fused using a transformer-based architecture, enhancing GBC detection performance. RadFormer's performance is evaluated against several baseline models, demonstrating superior accuracy. By mapping the neural features to radiological lexicons, RadFormer provides precise and interpretable explanations for GBC detection. The architecture addresses challenges of ultrasound images, such as sensor noise, artifacts, and visual similarities between non-malignant regions and cancerous gallbladder. Overall, RadFormer presents a significant advancement in the field of medical imaging and cancer detection.

*3.14. Other-Synthetic*

This section delves deeper into the wider application of transformer technology beyond the specific analysis of ultrasound images for certain organs, as discussed in previous sections. The field of imaging and tracking has witnessed substantial advancements through the use of transformer networks. For example, Qu et al. [87] developed the Complex Transformer Network (CTN), which integrates complex self-attention (CSA) and complex convolution modules for zero-degree single-angle polarization waveform imaging (PWI) beamforming. This technique maps delayed in-phase and quadrature (IQ) data directly to an image, with the CSA module assigning dynamic weights to reconstruction features based on their coherence. Table 10 gives a thorough review of the utilization of transformer-based models in the examination of Other-Synthetic ultrasound images.

In the realm of microbubble (MB) localization, Liu et al. [15] have introduced a Swin transformer-based neural network for end-to-end mapping of MBs. They further refined this method with a Super-Resolution Modified Transformer (SR-MT), improving MB localization and scaling the input dimension. They proposed a transformer-based neural network to replace the MB localization step in generating Ultra-Structure-Super-Resolution (US-SR) images.

Another transformer-based approach, the Depthwise Separable Convolutional Swin Transformer, was introduced by Liu et al. [16]. This transformer is designed for cervical lymph-node-level classification in ultrasound images. The network includes a deepwise separable convolution branch in the self-attention mechanism to capture discriminative local features. To tackle data imbalance issues, a new loss function was proposed to enhance the performance of the classification network.

Yan et al. [80] utilized a transformer-based network for motion prediction in their needle tip tracking system. This approach helped them estimate the target's current position from its past position data, addressing the issue of the target's temporary disappearance. The transformer network processes the entire data sequence at each instance, capturing both long- and short-term dependencies to fully understand the internal relationships within the input data sequence.

Zhao et al. trained an automatic segmentation Medical Transformer (MedT) network for ultrasound images of the distal humeral cartilage [81]. This research represents the first application of multiple deep learning algorithms for dynamic, volumetric ultrasound images in distal humeral cartilage segmentation, which are critical for minimally invasive surgeries.

Zhou et al. introduced the Lightweight Attention Encoder–Decoder Network (LAEDNet) [82], an innovative and efficient asymmetrical encoder–decoder network, for the segmentation of Head Circumference Ultrasound Images Dataset (HCUS).

 doi:10.20944/preprints202401.0309.v1



Katakis and his team [83] evaluated the potential of vision transformers for the automated segmentation of a muscle's cross-sectional area (CSA) and its mean grey level value, aiming to estimate the echogenicity of muscle architecture.

Lo [84] employed a pre-trained Vision Transformer (ViT) model to extract image features for the purpose of diagnosing septic arthritis from gray-scale and Power Doppler ultrasound images. Leveraging the deep learning capabilities of the ViT, the system autonomously and efficiently gathers significant image features for classification purposes.

Zhang et al. [85] have introduced a novel Pyramid Convolutional Transformer (PCT) architecture for the segmentation of parotid gland tumors. This architecture employs a shrinking pyramid framework to capture dense pixel features effectively and leverages multi-scale image dependencies. A Fusion Attention Transformer CNN (FTC) block is also incorporated to manage the complex and variable contour characteristics of parotid gland tumors. This block merges the Transformer with CNN, forming a dual branch structure to extract both global and local image features.

Finally, Manzari et al. [86] suggested an innovative hybrid model that integrates the strengths of CNNs and Transformers, mitigating the high quadratic complexity of the self-attention mechanism. They use an efficient convolution operation to attend to information across various representation spaces. Additionally, they aim to enhance the model's resistance to adversarial attacks by learning smoother decision boundaries. The hybrid model, known as Medical Vision Transformer (MedViT), combines local representations and global features using robust components. A novel patch moment changer augmentation has also been developed to add diversity and affinity to the training data.

**Table 10.** Detailed description of transformer-based Other-Synthetic US image analysis.

| Methods /References | Publication Year | Task | Architecture | Dataset | Evaluation Metrics | Highlights |
|---|---|---|---|---|---|---|
| CTN [23] | 2023 | plane-wave imaging (PWI) | CTN: complex transformer network | 1700 samples | contrast ratio: 11.59 dB<br>contrast-to-noise ratio: 1.16<br>generalized contrast-to-noise ratio: 0.68 | A CTN was developed using complex convolution to manage envelope information and extract complex reconstruction features from complex IQ data. This resulted in higher spatial resolution and contrast at significantly reduced computational costs.<br>The Complex Self-Attention (CSA) module, was developed based on the principles of the self-attention mechanism.<br>This module assists in eliminating irrelevant complex reconstruction features, thus enhancing image quality. |
| SR-MT [15] | 2023 | localization | Swin transformer | 11 000 realistic synthetic datasets | Lateral localization precision (LP)(MB= 1.6 MBs/mm2):15.0<br>DSC: 0.8<br>IoU: 0.66 | The research confirmed the effectiveness of the proposed method in precisely locating Microbubbles (MB) in synthetic data and in vivo visualization of brain structures. |
| Depthwise Swin Transformer [16] | 2022 | classification | Swin Transformer | 2268 ultrasound images(1146 cases) | Acc: 0.8065<br>Se: 0.8068<br>Sp: 0.7873<br>F1 value: 0.7942 | introducing a comprehensive approach for categorizing cervical lymph node levels in ultrasound images.<br>Employing model that combines depthwise separable convolutions withs transformer architecture, along with a novel loss function. |
| tip tracking [80] | 2023 | Tracking | visual tracking network | 3,000 US images | tracking success rate: 78% | Implemented a motion prediction system, based on the Transformer network<br><br>Constructed a visual tracking module leveraging dual mask sets to pinpoint the needle tip and minimize background noise.<br><br>Constructed a robust data fusion system that combines the results from the motion prediction and visual tracking systems |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2024 doi:10.20944/preprints202401.0309.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| [81] | 2023 | Segmentation | Medical Transformer (MedT) | 5,321 ultrasound images | DSC: 0.894 | Developed image-guided therapy (IGT) for visualization of distal humeral cartilage |
| LAEDNet [82] | 2022 | Segmentation | Lightweight Attention Encoder–Decoder Network+ Lightweight Residual Squeeze-and-Excitation (LRSE) | Brachial Plexus (BP) Dataset<br>Breast Ultrasound Images Dataset (BUSI)<br>Head Circumference Ultrasound (HCUS) Dataset | DSC (BP): 0.73<br>DSC(BUSI): 0.738<br>DSC(HCUS): 0.913 | The LAEDNet's unique asymmetrical structure plays a crucial role in minimizing network parameters, thereby accelerating the inference process. A compact decoding block named LRSE has been developed, which employs an attention mechanism for smooth integration with the LAEDNet backbone. |
| TMUNet [83] | 2023 | Segmentation | Vision transformer+The contextual attention network (TMUNet) | 2005 transverse ultrasound | DSC: 0.96 | Providing additional knowledge to ensure the execution of the previously mentioned tasks. |
| [84] | 2023 | Feature extraction+Classification | vision transformer (ViT) | 278 images | Acc: 0.92<br>AUC: 0.92 | Vision Transformer is employed as a feature extractor, while Support Vector Machine (SVM) acts as the classifier |
| PCT [85] | 2023 | Segmentation | Pyramid Convolutional Transformer (PCT) | PGTSeg(parotid gland tumor segmentation ) dataset: 365 images | IoU: 0.8434<br>DSC: 0.9151 | The Transformer branch incorporates an enhanced version of the multi-head attention mechanism, referred to as the multi-head fusion attention (MHFA) module. |
| MedViT [86] | 2023 | Classification | Medical Vision Transformer (MedViT) | BreastMNIST: 780 breast ultrasound | AUC: 0.938<br>Acc: 0.897 | To improve both generalization performance and adversarial resilience, the authors aim to increase a model's reliance on global structure features rather than texture information. They do this by calculating the mean and variance of training examples along the channel dimensions in the feature space and mixing them together. This method enables the exploration of new regions in the feature space that are mainly associated with global structure features. |

Acc: Accuracy, DSC: Dice similarity coefficient, AUC: Area Under Curve, IoU: Intersection over Union

 doi:10.20944/preprints202401.0309.v1



## 4. Discussion

Transformers, a unique type of convolutional-free neural network architecture, are designed to excel in capturing long-range dependencies within sequential data, making them suitable for language-related and computer vision tasks. They utilize an attention mechanism, specifically self-attention, which allows the model to focus on different parts of the input sequence and identify relationships between them. This makes self-attention a powerful tool for tasks involving sequences, such as video processing. Another technique used in transformers is multi-head self-attention, which enhances the model's ability to discern varied relationships and dependencies within the input sequence. This technique provides multiple unique viewpoints on the input sequence, enabling the model to focus on different segments of the input sequence and capture a variety of information simultaneously.

Transformers are surpassing conventional methods in medical image analysis. Despite their promising results, there are still challenges due to the high computational demand of transformers. Therefore, improvements to the transformer architecture are needed to make it more lightweight and efficient.

Despite their potential, transformers face challenges in ultrasound imaging due to data sparsity and the complexity of medical data. However, they offer new possibilities for accurate and efficient analysis, emphasizing the need for future research to address these challenges.

Research using transformers is rapidly growing in the field of ultrasound medical image analysis. Most current transformer-based methods can be easily applied to ultrasound imaging problems without significant changes.

Current transformer-based models for ultrasound diagnosis only use 2-D cross-sectional images for making predictions. However, 3-D ultrasound data or spatiotemporal data could potentially improve diagnostic accuracy. Despite the promising outcomes demonstrated by transformer methods for ultrasound, the advancement of AI-powered ultrasound lags behind that of AI-powered CT and MRI. This is primarily due to the significant intra- and inter-reader variability encountered during the acquisition and interpretation of ultrasound images.

## 5. Conclusions

Given the unique characteristics and diagnostic needs of ultrasound imaging, a comprehensive review of AI methods based on vision transformers, specifically designed for ultrasound imaging, can offer crucial insights for researchers and practitioners in this particular field. Hence, this review seeks to fill this void by providing an in-depth examination of Transformer models that have been specifically developed for ultrasound imaging and its related image analysis applications.

In conclusion, this review paper has provided a comprehensive overview of the application of vision transformers in medical ultrasound. It includes a detailed analysis of over 231 relevant studies, highlighting the most recent developments in this field.

We began by elucidating the fundamental structures of transformers, followed by an introduction to the most significant architectures of vision transformers. Subsequently, we categorized the paper based on different organs, providing a structured approach to understanding the diverse applications of these technologies. We reviewed the most pivotal papers that have utilized vision transformers in medical ultrasound, highlighting the transformative impact of these methodologies in the field. As the landscape of medical ultrasound continues to evolve, the role of vision transformers is anticipated to become increasingly prominent, paving the way for more sophisticated and precise diagnostic tools. This review underscores the potential of vision transformers to revolutionize medical ultrasound, marking a significant stride towards the future of healthcare.

**Author Contributions:** M.Vafaeezadeh: Preparation of the original draft for heart, fetal, carotid, and other synthetic sections. P.gifani: Review and editing of the lung, liver, IVUS, and gallbladder sections, as well as the introduction and transformer description sections. H. Behnam: Supervision and review and editing of the breast, pancreatic, urinary bladder, prostate, and thyroid sections.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Koutras, A.; Perros, P.; Prokopakis, I.; Ntounis, T.; Fasoulakis, Z.; Pittokopitou, S.; Samara, A.A.; Valsamaki, A.; Douligeris, A.; Mortaki, A. Advantages and Limitations of Ultrasound as a Screening Test for Ovarian Cancer. *Diagnostics* **2023**, *13*, 2078.
2. Leung, K.-Y. Applications of Advanced Ultrasound Technology in Obstetrics. *Diagnostics* **2021**, *11*, 1217.
3. Brunetti, N.; Calabrese, M.; Martinoli, C.; Tagliafico, A.S. Artificial intelligence in breast ultrasound: from diagnosis to prognosis—a rapid review. *Diagnostics* **2022**, *13*, 58.
4. Gifani, P.; Vafaeezadeh, M.; Ghorbani, M.; Mehri-Kakavand, G.; Pursamimi, M.; Shalbaf, A.; Davanloo, A.A. Automatic diagnosis of stage of COVID-19 patients using an ensemble of transfer learning with convolutional neural networks based on computed tomography images. *Journal of Medical Signals & Sensors* **2023**, *13*, 101-109.
5. Ait Nasser, A.; Akhloufi, M.A. A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics* **2023**, *13*, 159.
6. Shalbaf, A.; Gifani, P.; Mehri-Kakavand, G.; Pursamimi, M.; Ghorbani, M.; Davanloo, A.A.; Vafaeezadeh, M. Automatic diagnosis of severity of COVID-19 patients using an ensemble of transfer learning models with convolutional neural networks in CT images. *Polish Journal of Medical Physics and Engineering* **2022**, *28*, 117-126.
7. Qian, J.; Li, H.; Wang, J.; He, L. Recent Advances in Explainable Artificial Intelligence for Magnetic Resonance Imaging. *Diagnostics* **2023**, *13*, 1571.
8. Vafaeezadeh, M.; Behnam, H.; Hosseinsabet, A.; Gifani, P. A deep learning approach for the automatic recognition of prosthetic mitral valve in echocardiographic images. *Computers in Biology and Medicine* **2021**, *133*, 104388.
9. Gifani, P.; Shalbaf, A.; Vafaeezadeh, M. Automated detection of COVID-19 using ensemble of transfer learning with deep convolutional neural network based on CT scans. *International journal of computer assisted radiology and surgery* **2021**, *16*, 115-123.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
11. Reynaud, H.; Vlontzos, A.; Hou, B.; Beqiri, A.; Leeson, P.; Kainz, B. Ultrasound video transformers for cardiac ejection fraction estimation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24, 2021; pp. 495-505.
12. Gilany, M.; Wilson, P.; Perera-Ortega, A.; Jamzad, A.; To, M.N.N.; Fooladgar, F.; Wodlinger, B.; Abolmaesumi, P.; Mousavi, P. TRUSformer: improving prostate cancer detection from micro-ultrasound using attention and self-supervision. *International Journal of Computer Assisted Radiology and Surgery* **2023**, 1-8.
13. Dadoun, H.; Rousseau, A.-L.; de Kerviler, E.; Correas, J.-M.; Tissier, A.-M.; Joujou, F.; Bodard, S.; Khezzane, K.; de Margerie-Mellon, C.; Delingette, H. Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal US images. *Radiology: Artificial Intelligence* **2022**, *4*, e210110.
14. Wang, W.; Jiang, R.; Cui, N.; Li, Q.; Yuan, F.; Xiao, Z. Semi-supervised vision transformer with adaptive token sampling for breast cancer classification. *Frontiers in Pharmacology* **2022**, *13*, 929755.
15. Liu, X.; Almekkawy, M. Ultrasound Localization Microscopy Using Deep Neural Network. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **2023**.
16. Liu, Y.; Zhao, J.; Luo, Q.; Shen, C.; Wang, R.; Ding, X. Automated classification of cervical lymph-node-level from ultrasound using Depthwise Separable Convolutional Swin Transformer. *Computers in Biology and Medicine* **2022**, *148*, 105821.
17. Perera, S.; Adhikari, S.; Yilmaz, A. Pocformer: A lightweight transformer architecture for detection of covid-19 using point of care ultrasound. In Proceedings of the 2021 IEEE international conference on image processing (ICIP), 2021; pp. 195-199.

18. Li, J.; Zhang, P.; Wang, T.; Zhu, L.; Liu, R.; Yang, X.; Wang, K.; Shen, D.; Sheng, B. DSMT-Net: Dual Self-supervised Multi-operator Transformation for Multi-source Endoscopic Ultrasound Diagnosis. *IEEE Transactions on Medical Imaging* **2023**.
19. Hu, X.; Cao, Y.; Hu, W.; Zhang, W.; Li, J.; Wang, C.; Mukhopadhyay, S.C.; Li, Y.; Liu, Z.; Li, S. Refined feature-based Multi-frame and Multi-scale Fusing Gate network for accurate segmentation of plaques in ultrasound videos. *Computers in Biology and Medicine* **2023**, 107091.
20. Xia, M.; Yang, H.; Qu, Y.; Guo, Y.; Zhou, G.; Zhang, F.; Wang, Y. Multilevel structure-preserved GAN for domain adaptation in intravascular ultrasound analysis. *Medical Image Analysis* **2022**, *82*, 102614.
21. Yang, C.; Liao, S.; Yang, Z.; Guo, J.; Zhang, Z.; Yang, Y.; Guo, Y.; Yin, S.; Liu, C.; Kang, Y. RDHCformer: Fusing ResDCN and Transformers for Fetal Head Circumference Automatic Measurement in 2D Ultrasound Images. *Frontiers in Medicine* **2022**, *9*, 848904.
22. Sankari, V.R.; Raykar, D.A.; Snekhalatha, U.; Karthik, V.; Shetty, V. Automated detection of cystitis in ultrasound images using deep learning techniques. *IEEE Access* **2023**.
23. Basu, S.; Gupta, M.; Rana, P.; Gupta, P.; Arora, C. RadFormer: Transformers with global–local attention for interpretable and accurate Gallbladder Cancer detection. *Medical Image Analysis* **2023**, *83*, 102676.
24. Shamshad, F.; Khan, S.; Zamir, S.W.; Khan, M.H.; Hayat, M.; Khan, F.S.; Fu, H. Transformers in medical imaging: A survey. *Medical Image Analysis* **2023**, 102802.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European conference on computer vision, 2020; pp. 213-229.
27. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International conference on machine learning, 2021; pp. 10347-10357.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 10012-10022.
29. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 568-578.
30. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 22-31.
31. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 12179-12188.
32. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
33. Liu, Y.; Yang, Y.; Jiang, W.; Wang, T.; Lei, B. 3d deep attentive u-net with transformer for breast tumor segmentation from automated breast volume scanner. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021; pp. 4011-4014.
34. Gheflati, B.; Rivaz, H. Vision transformers for classification of breast ultrasound images. In Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022; pp. 480-483.
35. Ayana, G.; Choe, S.-W. BUVITNET: Breast ultrasound detection via vision transformers. *Diagnostics* **2022**, *12*, 2654.
36. Li, G.; Jin, D.; Yu, Q.; Qi, M. IB-TransUNet: Combining Information Bottleneck and Transformer for Medical Image Segmentation. *Journal of King Saud University-Computer and Information Sciences* **2023**, *35*, 249-258.
37. Mo, Y.; Han, C.; Liu, Y.; Liu, M.; Shi, Z.; Lin, J.; Zhao, B.; Huang, C.; Qiu, B.; Cui, Y. Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging* **2023**.
38. Wu, H.; Huang, X.; Guo, X.; Wen, Z.; Qin, J. Cross-image Dependency Modelling for Breast Ultrasound Segmentation. *IEEE Transactions on Medical Imaging* **2023**.
39. Ji, H.; Zhu, Q.; Ma, T.; Cheng, Y.; Zhou, S.; Ren, W.; Huang, H.; He, W.; Ran, H.; Ruan, L. Development and validation of a transformer-based CAD model for improving the consistency of BI-RADS category 3–5 nodule classification among radiologists: a multiple center study. *Quantitative Imaging in Medicine and Surgery* **2023**, *13*, 3671.
40. Li, G.; Jin, D.; Yu, Q.; Zheng, Y.; Qi, M. MultiIB-TransUNet: Transformer with multiple information bottleneck blocks for CT and ultrasound image segmentation. *Medical Physics* **2023**.
41. Zhou, J.; Hou, Z.; Lu, H.; Wang, W.; Zhao, W.; Wang, Z.; Zheng, D.; Wang, S.; Tang, W.; Qu, X. A deep supervised transformer U-shaped full-resolution residual network for the segmentation of breast ultrasound image. *Medical Physics* **2023**, *50*, 7513-7524.

42. Song, M.; Kim, Y. Optimizing proportional balance between supervised and unsupervised features for ultrasound breast lesion classification. *Biomedical Signal Processing and Control* **2024**, *87*, 105443.
43. https://zenodo.org/records/8041285.
44. https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset.
45. Lu, X.; Liu, X.; Xiao, Z.; Zhang, S.; Huang, J.; Yang, C.; Liu, S. Self-supervised dual-head attentional bootstrap learning network for prostate cancer screening in transrectal ultrasound images. *Computers in Biology and Medicine* **2023**, *165*, 107337.
46. Li, C.; Du, R.; Luo, Q.; Wang, R.; Ding, X. A novel model of thyroid nodule segmentation for ultrasound images. *Ultrasound in Medicine & Biology* **2023**, *49*, 489-496.
47. Zhang, N.; Liu, J.; Jin, Y.; Duan, W.; Wu, Z.; Cai, Z.; Wu, M. An adaptive multi-modal hybrid model for classifying thyroid nodules by combining ultrasound and infrared thermal images. *BMC bioinformatics* **2023**, *24*, 315.
48. JERBI, F.; ABOUDI, N.; KHLIFA, N. Automatic classification of ultrasound thyroids images using vision transformers and generative adversarial networks. *Scientific African* **2023**, *20*, e01679.
49. Liu, Q.; Ding, F.; Li, J.; Ji, S.; Liu, K.; Geng, C.; Lyu, L. DCA-Net: Dual-branch contextual-aware network for auxiliary localization and segmentation of parathyroid glands. *Biomedical Signal Processing and Control* **2023**, *84*, 104856.
50. Chen, F.; Han, H.; Wan, P.; Liao, H.; Liu, C.; Zhang, D. Joint Segmentation and Differential Diagnosis of Thyroid Nodule in Contrast-Enhanced Ultrasound Images. *IEEE Transactions on Biomedical Engineering* **2023**.
51. Zhao, X.; Li, H.; Xu, J.; Wu, J. Ultrasonic Thyroid Nodule Benign-Malignant Classification with Multi-level Features Fusions. In Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC), 2023; pp. 907-912.
52. Zeng, Y.; Tsui, P.-H.; Wu, W.; Zhou, Z.; Wu, S. MAEF-Net: multi-attention efficient feature fusion network for deep learning segmentation. In Proceedings of the 2021 IEEE International Ultrasonics Symposium (IUS), 2021; pp. 1-4.
53. Ahmadi, N.; Tsang, M.; Gu, A.; Tsang, T.; Abolmaesumi, P. Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series. *IEEE Transactions on Medical Imaging* **2023**.
54. Vafaeezadeh, M.; Behnam, H.; Hosseinsabet, A.; Gifani, P. CarpNet: Transformer for mitral valve disease classification in echocardiographic videos. *International Journal of Imaging Systems and Technology* **2023**.
55. Al Qurri, A.; Almekkawy, M. Improved UNet with Attention for Medical Image Segmentation. *Sensors* **2023**, *23*, 8589.
56. Luo, J.; Wang, Q.; Zou, R.; Wang, Y.; Liu, F.; Zheng, H.; Du, S.; Yuan, C. A Heart Image Segmentation Method Based on Position Attention Mechanism and Inverted Pyramid. *Sensors* **2023**, *23*, 9366.
57. Ahn, S.S.; Ta, K.; Thorn, S.L.; Onofrey, J.A.; Melvinsdottir, I.H.; Lee, S.; Langdon, J.; Sinusas, A.J.; Duncan, J.S. Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3D echocardiography. *Medical Image Analysis* **2023**, *84*, 102711.
58. Tang, Z.; Duan, J.; Sun, Y.; Zeng, Y.; Zhang, Y.; Yao, X. A combined deformable model and medical transformer algorithm for medical image segmentation. *Medical & Biological Engineering & Computing* **2023**, *61*, 129-137.
59. Liao, M.; Lian, Y.; Yao, Y.; Chen, L.; Gao, F.; Xu, L.; Huang, X.; Feng, X.; Guo, S. Left Ventricle Segmentation in Echocardiography with Transformer. *Diagnostics* **2023**, *13*, 2365.
60. Zhao, C.; Chen, W.; Qin, J.; Yang, P.; Xiang, Z.; Frangi, A.F.; Chen, M.; Fan, S.; Yu, W.; Chen, X. IFT-net: Interactive fusion transformer network for quantitative analysis of pediatric echocardiography. *Medical Image Analysis* **2022**, *82*, 102648.
61. Fazry, L.; Haryono, A.; Nissa, N.K.; Hirzi, N.M.; Rachmadi, M.F.; Jatmiko, W. Hierarchical Vision Transformers for Cardiac Ejection Fraction Estimation. In Proceedings of the 2022 7th International Workshop on Big Data and Information Security (IWBIS), 2022; pp. 39-44.
62. Hagberg, E.; Hagerman, D.; Johansson, R.; Hosseini, N.; Liu, J.; Björnsson, E.; Alvén, J.; Hjelmgren, O. Semi-supervised learning with natural language processing for right ventricle classification in echocardiography—a scalable approach. *Computers in Biology and Medicine* **2022**, *143*, 105282.
63. Vafaeezadeh, M.; Behnam, H.; Hosseinsabet, A.; Gifani, P. Automatic morphological classification of mitral valve diseases in echocardiographic images based on explainable deep learning methods. *International Journal of Computer Assisted Radiology and Surgery* **2022**, *17*, 413-425.
64. Rahman, R.; Alam, M.G.R.; Reza, M.T.; Huq, A.; Jeon, G.; Uddin, M.Z.; Hassan, M.M. Demystifying evidential Dempster Shafer-based CNN architecture for fetal plane detection from 2D ultrasound images leveraging fuzzy-contrast enhancement and explainable AI. *Ultrasonics* **2023**, *132*, 107017.
65. Sarker, M.M.K.; Singh, V.K.; Alsharid, M.; Hernandez-Cruz, N.; Papageorghiou, A.T.; Noble, J.A. COMFormer: Classification of Maternal-Fetal and Brain Anatomy using a Residual Cross-Covariance

Attention Guided Transformer in Ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **2023**.

66. Arora, U.; Sengupta, D.; Kumar, M.; Tirupathi, K.; Sai, M.K.; Hareesh, A.; Chaithanya, E.S.S.; Nikhila, V.; Bhavana, N.; Vigneshwar, P. Perceiving placental ultrasound image texture evolution during pregnancy with normal and adverse outcome through machine learning prism. *Placenta* **2023**.
67. Chen, X.; You, G.; Chen, Q.; Zhang, X.; Wang, N.; He, X.; Zhu, L.; Li, Z.; Liu, C.; Yao, S. Development and evaluation of an artificial intelligence system for children intussusception diagnosis using ultrasound images. *Iscience* **2023**, *26*.
68. Qiao, S.; Pang, S.; Luo, G.; Sun, Y.; Yin, W.; Pan, S.; Lv, Z. DPC-MSGATNet: dual-path chain multi-scale gated axial-transformer network for four-chamber view segmentation in fetal echocardiography. *Complex & Intelligent Systems* **2023**, 1-17.
69. Płotka, S.; Grzeszczyk, M.K.; Brawura-Biskupski-Samaha, R.; Gutaj, P.; Lipa, M.; Trzciński, T.; Sitek, A. BabyNet: residual transformer module for birth weight prediction on fetal ultrasound video. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022; pp. 350-359.
70. Płotka, S.; Grzeszczyk, M.K.; Brawura-Biskupski-Samaha, R.; Gutaj, P.; Lipa, M.; Trzciński, T.; Išgum, I.; Sánchez, C.I.; Sitek, A. BabyNet++: Fetal birth weight prediction using biometry multimodal data acquired less than 24 hours before delivery. *Computers in Biology and Medicine* **2023**, *167*, 107602.
71. Płotka, S.S.; Grzeszczyk, M.K.; Szenejko, P.I.; Żebrowska, K.; Szymecka-Samaha, N.A.; Łęgowik, T.; Lipa, M.A.; Kosińska-Kaczyńska, K.; Brawura-Biskupski-Samaha, R.; Išgum, I. Deep learning for estimation of fetal weight throughout the pregnancy from fetal abdominal ultrasound. *American journal of obstetrics & gynecology MFM* **2023**, *5*, 101182.
72. Zhao, C.; Droste, R.; Drukker, L.; Papageorghiou, A.T.; Noble, J.A. Visual-assisted probe movement guidance for obstetric ultrasound scanning using landmark retrieval. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, 2021; pp. 670-679.
73. Lin, Y.; Huang, J.; Xu, W.; Cui, C.; Xu, W.; Li, Z. Method for carotid artery 3-D ultrasound image segmentation based on cswin transformer. *Ultrasound in Medicine & Biology* **2023**, *49*, 645-656.
74. Li, L.; Hu, Z.; Huang, Y.; Zhu, W.; Zhao, C.; Wang, Y.; Chen, M.; Yu, J. BP-Net: Boundary and perfusion feature guided dual-modality ultrasound video analysis network for fibrous cap integrity assessment. *Computerized Medical Imaging and Graphics* **2023**, *107*, 102246.
75. Nehary, E.; Rajan, S.; Rossa, C. Lung Ultrasound Image Classification Using Deep Learning and Histogram of Oriented Gradients Features for COVID-19 Detection. In Proceedings of the 2023 IEEE Sensors Applications Symposium (SAS), 2023; pp. 1-6.
76. Xing, W.; Liu, Y.; He, C.; Liu, X.; Li, Y.; Li, W.; Chen, J.; Ta, D. Frame-to-video-based Semi-supervised Lung Ultrasound Scoring Model. In Proceedings of the 2023 IEEE International Ultrasonics Symposium (IUS), 2023; pp. 1-4.
77. Zhang, J.; Chen, Y.; Liu, P. Automatic Recognition of Standard Liver Sections Based on Vision-Transformer. In Proceedings of the 2022 IEEE 16th International Conference on Anti-counterfeiting, Security, and Identification (ASID), 2022; pp. 1-4.
78. Zhang, J.; Chen, Y.; Zeng, P.; Liu, Y.; Diao, Y.; Liu, P. Ultra-Attention: Automatic Recognition of Liver Ultrasound Standard Sections Based on Visual Attention Perception Structures. *Ultrasound in Medicine & Biology* **2023**, *49*, 1007-1017.
79. Huang, X.; Bajaj, R.; Li, Y.; Ye, X.; Lin, J.; Pugliese, F.; Ramasamy, A.; Gu, Y.; Wang, Y.; Torii, R. POST-IVUS: A perceptual organisation-aware selective transformer framework for intravascular ultrasound segmentation. *Medical Image Analysis* **2023**, *89*, 102922.
80. Yan, W.; Ding, Q.; Chen, J.; Yan, K.; Tang, R.S.-Y.; Cheng, S.S. Learning-based needle tip tracking in 2D ultrasound by fusing visual tracking and motion prediction. *Medical Image Analysis* **2023**, *88*, 102847.
81. Zhao, W.; Su, X.; Guo, Y.; Li, H.; Basnet, S.; Chen, J.; Yang, Z.; Zhong, R.; Liu, J.; Chui, E.C.-s. Deep learning based ultrasonic visualization of distal humeral cartilage for image-guided therapy: a pilot validation study. *Quantitative Imaging in Medicine and Surgery* **2023**, *13*, 5306.
82. Zhou, Q.; Wang, Q.; Bao, Y.; Kong, L.; Jin, X.; Ou, W. Laednet: A lightweight attention encoder–decoder network for ultrasound medical image segmentation. *Computers and Electrical Engineering* **2022**, *99*, 107777.
83. Katakis, S.; Barotsis, N.; Kakotaritis, A.; Tsiganos, P.; Economou, G.; Panagiotopoulos, E.; Panayiotakis, G. Muscle Cross-Sectional Area Segmentation in Transverse Ultrasound Images Using Vision Transformers. *Diagnostics* **2023**, *13*, 217.
84. Lo, C.-M.; Lai, K.-L. Deep learning-based assessment of knee septic arthritis using transformer features in sonographic modalities. *Computer Methods and Programs in Biomedicine* **2023**, *237*, 107575.
85. Zhang, G.; Zheng, C.; He, J.; Yi, S. PCT: Pyramid convolutional transformer for parotid gland tumor segmentation in ultrasound images. *Biomedical Signal Processing and Control* **2023**, *81*, 104498.

86. Manzari, O.N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S.B.; Ayatollahi, A. MedViT: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine* **2023**, *157*, 106791.
87. Qu, X.; Ren, C.; Wang, Z.; Fan, S.; Zheng, D.; Wang, S.; Lin, H.; Jiang, J.; Xing, W. Complex transformer network for single-angle plane-wave imaging. *Ultrasound in Medicine & Biology* **2023**, *49*, 2234-2246.