Heliyon 10 (2024) e34845

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Enhancing left ventricular segmentation in echocardiography with a modified mixed attention mechanism in SegFormer architecture

Hanqiong Wu [a,1], Gangrong Qu [b,1], Zhifeng Xiao [c], Fan Chunyu [d,*]

[a] Internal Medicine, The First Hospital of Jinzhou Medical University, Jinzhou, 121001, China
[b] Cardiovascular Medicine, Chongqing General Hospital of the Armed Police Force, Chongqing, 400061, China
[c] China Nanhu Academy of Electronics and Information Technology, Jiaxing, 314050, China
[d] Department of Cardiovascular Medicine, The People's Hospital of Liaoning Province, Shengyang, 110067, China

ARTICLE INFO

Keywords:
Mixed attention mechanism
Semantic segmentation
Time feature map
Video semantic segmentation
Left ventricular

ABSTRACT

Echocardiography is a key tool for the diagnosis of cardiac diseases, and accurate left ventricular (LV) segmentation in echocardiographic videos is crucial for the assessment of cardiac function. However, since semantic segmentation of video needs to take into account the temporal correlation between frames, this makes the task very challenging. This article introduces an innovative method that incorporates a modified mixed attention mechanism into the SegFormer architecture, enabling it to effectively grasp the temporal correlation present in video data. The proposed method processes each time series by encoding the image input into the encoder to obtain the current time feature map. This map, along with the historical time feature map, is then fed into a time-sensitive mixed attention mechanism type of convolution block attention module (TCBAM). Its output can serve as the historical time feature map for the subsequent sequence, and a combination of the current time feature map and historical time feature map for the current sequence. The processed feature map is then input into the Multilayer Perceptron (MLP) and subsequent networks to generate the final segmented image. Through extensive experiments conducted on two different datasets: Hamad Medical Corporation, Tampere University, and Qatar University (HMC-QU), Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) and Sunnybrook Cardiac Data (SCD), achieving a Dice coefficient of 97.92 % on the SCD dataset and an F1 score of 0.9263 on the CAMUS dataset, outperforming all other models. This research provides a promising solution to the temporal modeling challenge in video semantic segmentation tasks using transformer-based models and points out a promising direction for future research in this field.

1. Introduction

In diagnostic cardiology, left ventricular (LV) segmentation is critical for quantifying cardiac disease because it is directly related to the measurement of ventricular volume and mass. With advances in imaging technology, the diagnosis and treatment of cardiovascular disease are overcoming the previous limitations of 2D echocardiography, allowing for a more comprehensive and accurate visualization of cardiac structures [1]. Echocardiography has been widely used as a safe, low-cost cardiac diagnostic test. This non-invasive

* Corresponding author.
E-mail address: m17702488448@163.com (F. Chunyu).
[1] These authors contributed equally to this work and share the first authorship.

https://doi.org/10.1016/j.heliyon.2024.e34845
Received 12 January 2024; Received in revised form 16 July 2024; Accepted 17 July 2024
Available online 17 July 2024

test allows visualization of all structures of the heart, including valves, atria, and ventricles. Echocardiography helps to detect heart problems that cause symptoms, such as shortness of breath, chest pain, or discomfort, as well as to assess the effects of diseases (such as high blood pressure, pulmonary hypertension, or certain medications) on the heart. It provides important information about the function of the heart by rapidly acquiring images without exposure to ionizing radiation, making it a safe option for repeated examinations [2].

In cardiology, a central role is played by the left ventricular ejection fraction (LVEF), which is a key indicator for assessing cardiac function and relies on accurate segmentation of the LV [3]. Echocardiography records short videos of spatiotemporal data of the heart that characterize spatial variations in the cardiac image, thus measuring diagnostic metrics based on the dynamic motion of the heart, such as the LVEF. The LVEF is the difference between the end-diastolic (ED) and end-systolic (ES) volumes and is used to quantify the function of the heart. If the quality of the echocardiogram is not high, it may lead to a miscalculation of the LVEF, which could potentially delay the treatment of a cardiac patient. Therefore, the key to ensuring the accuracy of subsequent diagnosis lies in the accurate segmentation of the LV. Although some progress has been made in video semantic segmentation techniques in recent years, continuous improvement is still needed to increase the accuracy and efficiency of segmentation [4].

Semantic segmentation plays a pivotal role in numerous applications, especially in the field of video processing. It involves assigning a label to every pixel in an image, which is crucial for understanding the content of videos at a granular level. However, the task becomes more challenging when dealing with video data due to the temporal dependencies that exist between different frames.

The advent of Transformer architectures has revolutionized the field of computer vision, providing powerful tools for semantic segmentation tasks. One such model, SegFormer, has demonstrated impressive performance on static images. However, these Transformer-based models, including SegFormer, are primarily designed for static images and lack explicit modeling capabilities for temporal problems. This limitation hinders their ability to fully utilize the temporal relationships between frames when dealing with video data, thereby restricting performance enhancement.

In this paper, we address this limitation by proposing a new approach that modifies the hybrid attention mechanism to be more sensitive to time-series tasks and then integrates it into the SegFormer structure to enable it to handle temporal dependencies in video data. The proposed model processes each temporal sequence by feeding the image to the Encoder to obtain the current temporal feature map. This feature map, along with the historical temporal feature map, is then input into the TCBAM, a type of Hybrid Attention Mechanism. The output serves as the historical temporal feature map for the next sequence and the current temporal feature map combined with the historical one for the current sequence. This processed feature map is then fed into the MLP and subsequent networks to obtain the final segmented image.

Our work is evaluated on two video datasets, HMC-QU, SCD, and CAMUS [5], which further underscores the practicality and effectiveness of our approach. The experimental results demonstrate that our model effectively captures the temporal dependencies in video data and significantly improves the performance of semantic segmentation tasks.

In this paper, we provide a Transformer model-based approach that offers an effective solution to the time-domain modeling challenges in the task of semantic segmentation of video, especially in the application of left ventricle segmentation in echocardiography. We believe our approach can inspire future research and development in this area and potentially lead to more advanced models capable of handling complex video data.

The remainder of this paper is organized as follows: Section II reviews the related work, Section III details the proposed method, Section IV presents the experimental results, and Section V discusses potential future work. Finally, Section VI concludes the paper.

As shown in Fig. 1, Temporal consistency in video frames is vital for real-time segmentation. As the frames in a video are not independent, the segmentation of a current frame can greatly benefit from the segmentation of previous frames. Moreover, the temporal consistency can help to reduce the flickering effect, which is common in the video segmentation results when processed frame by frame.

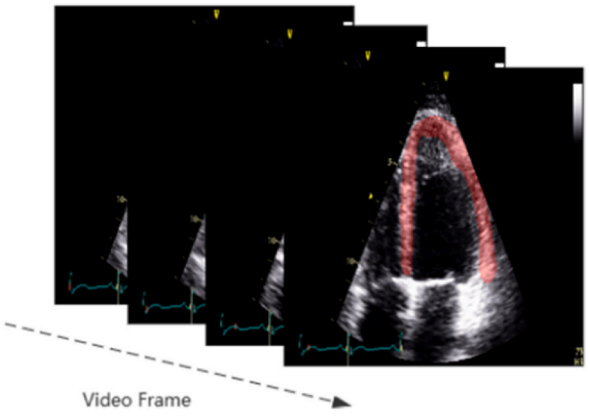

Fig. 1. Echo image frames and the segmentation mask of the LV.

2. Related work

2.1. Semantic segmentation

Semantic segmentation, a critical task in computer vision, involves assigning a label to every pixel in an image, allowing for a detailed understanding of the image content. The applications of semantic segmentation are vast, spanning across autonomous driving, medical imaging, and video surveillance, among others [6,7].

Initially, semantic segmentation tasks were tackled using hand-crafted features and graphical models, which, while effective, were often labor-intensive and could not be generalized across various tasks. The advent of deep learning revolutionized this field, with Convolutional Neural Networks (CNNs) becoming the dominant approach due to their superior performance and ability to learn features directly from data.

One of the earliest successful deep learning models for semantic segmentation was the Fully Convolutional Networks (FCN), which first introduced the concept of using convolutional networks for end-to-end pixel-wise classification. Following FCN, several enhanced models were proposed, each aiming to improve upon the limitations of the previous models.

The Mask-RCNN [8], for instance, extended the Faster-RCNN object detection model to include a branch for predicting segmentation masks, enabling instance segmentation. Mask-RCNN proved to be highly effective, setting a new benchmark, for instance, segmentation tasks.

U-Net [9] is another influential model that was specifically designed for biomedical image segmentation. U-Net introduced a symmetric expanding path to the contracting path of a traditional CNN, enabling precise localization. The architecture of U-Net was further improved by Zhou et al., who proposed the U-Net++ [10] model. U-Net++ introduced nested and dense skip connections to the U-Net architecture, resulting in improved performance and better gradient flow during training.

In 2022, Guo et al. proposed SegNeXt [11], a simple convolutional network architecture for semantic segmentation. They demonstrated that convolutional attention is a more efficient and effective way to encode contextual information than the self-attention mechanism in transformers. However, the SegNeXt model was primarily designed for static images and lacks explicit modeling capabilities for temporal problems, posing a limitation in its ability to fully utilize the temporal relationships between frames when dealing with video data.

The Transformer architecture, originally proposed for natural language processing tasks, was later adapted for semantic segmentation tasks with the introduction of the SegFormer model. SegFormer leverages the self-attention mechanisms of Transformers to capture long-range dependencies in the data, a significant advantage over traditional CNNs that primarily focus on local features.

The Swin U-Net [12] furthered the application of Transformer models in semantic segmentation by proposing a U-shaped transformer model, combining the advantages of the U-Net architecture and the Transformer's self-attention mechanism.

Nearest, the UniverSeg [13] model built upon these developments by proposing a universal segmentation model capable of performing both semantic and instance segmentation tasks. By integrating a Hybrid Attention Mechanism into the UniverSeg architecture, the model was able to handle temporal dependencies in video data, demonstrating its effectiveness in capturing temporal dependencies and improving the performance of video semantic segmentation tasks.

While these models have significantly advanced the field of semantic segmentation, there remains a challenge in handling video data, where temporal dependencies between frames need to be considered. This research aims to address this gap by proposing a novel method that incorporates a Hybrid Attention Mechanism into the SegFormer architecture, enabling it to effectively capture the temporal correlations in video data.

2.2. Transformer in semantic segmentation

The Transformer architecture, initially proposed for natural language processing tasks [14], has recently been adapted for computer vision tasks, including semantic segmentation. Transformer-based models, such as SegFormer [15], have demonstrated impressive performance on static images. They leverage self-attention mechanisms, a process where the model calculates a score for each pair of elements in the input data to determine how much focus should be put on other elements when generating the output for a particular element, to capture long-range dependencies in the data. This is a significant advantage over traditional Convolutional Neural Networks (CNNs) that mainly focus on local features. However, these Transformer-based models are primarily designed for static images and lack explicit modeling capabilities for temporal problems, which refers to problems where time and sequence of events matter. This limitation hinders their effectiveness in video semantic segmentation tasks, where understanding the sequence of frames over time is crucial.

In 2022, Liu et al. proposed a Temporal Correlation Module (TCM), a component that can be easily embedded into the current action recognition backbones to extract action visual tempo from low-level backbone features at a single-layer level [16]. "Action visual tempo" refers to the rhythm or pace of actions within a video, and the ability to extract this information can be crucial for recognizing and classifying different types of actions. However, their method was mainly designed for action recognition, a task that focuses on identifying the main activity in a video and may not be directly applicable to video semantic segmentation tasks. The latter requires a more detailed understanding of the content of videos at a granular level, including the identification and classification of every object and their movements in each frame, which is a more complex problem.

2.3. Temporal modeling in video processing

Temporal modeling remains a cornerstone of video processing, with a myriad of techniques vying to accurately encapsulate the dynamic nature of video data. While traditional methods like 3D convolutions [17], recurrent neural networks [18], and temporal convolutional networks have laid the groundwork [19], they often come at the cost of high computational demand and may falter with complex temporal patterns. The innovation by Xue et al. with the Explicit Cyclic Attention-based Network (ECANet) in 2022 marked a significant stride forward, particularly for video saliency prediction [20], by introducing a two-stream encoder-decoder model adept at discerning temporal dependencies.

Building on this foundation, recent explorations in the field have expanded the horizon of temporal modeling. For instance, the introduction of spateGAN by Glawion et al. harnesses the power of conditional generative adversarial networks to refine spatiotemporal downscaling of precipitation data [21]. This approach, inspired by video super-resolution techniques, has shown exceptional skill in enhancing the resolution of rainfall fields, thereby presenting a novel utility in climate modeling and potentially offering a new perspective for video processing tasks.

In the realm of human-computer interaction, Li et al. have proposed a dynamic gesture recognition network that incorporates a spatio-temporal attention mechanism [22]. This network utilizes a 3D residual convolution neural network to concurrently model the temporal relationship and spatial information, thereby improving the accuracy of dynamic gesture prediction in videos. Such advancements underscore the potential of attention mechanisms in refining feature extraction, which could be adapted for enhancing semantic segmentation models.

Further, the Hierarchical Spatiotemporal Feature Fusion Network (HSFF-Net) by Zhang et al. [23]. introduces a bi-directional fusion architecture, a first in the field, to augment video saliency prediction. The network's ability to adaptively learn fusion weights of adjacent features through its Hierarchical Adaptive Fusion mechanism exemplifies the innovative approaches being developed to address the limitations of hierarchical feature fusion in current models.

Lastly, the attention-guided neural network framework for audio-visual quality assessment by Cao et al. [24]. merges attention prediction models with Gated Recurrent Unit networks to assess the quality of experience for audio and video signals. This architecture not only advances the field of quality assessment but also provides insights into the integration of multi-modal signals, which could be instrumental in semantic segmentation tasks where synchronicity of audio and visual cues may be pertinent.

2.4. Hybrid attention mechanism

The Hybrid Attention Mechanism, such as the Convolutional Block Attention Module (CBAM) [25], has been proposed to enhance the representational power of CNNs. CBAM adaptively calculates attention maps along spatial and channel dimensions separately, allowing the model to focus on more informative features [26,27]. Although CBAM has been widely used in various computer vision tasks, its application in video semantic segmentation tasks, especially in combination with Transformer-based models, remains largely unexplored. In 2022, Ramaswamy et al. proposed a hybrid network with an attention mechanism for aspect categorization and sentiment classification [28]. They demonstrated that the attention mechanism can effectively capture the important features for sentiment classification. However, their method was designed for text data and may not be directly applicable to video data. In the same year, Farag et al. used DenseNet169's encoder and CBAM for automatic severity classification of diabetic retinopathy [29]. They demonstrated that CBAM can effectively enhance the representational power of CNNs. However, their method was designed for medical image analysis and may not be directly applicable to video semantic segmentation tasks.

3. Method

Our method is based on the SegFormer architecture, which is enhanced with a mixed attention mechanism as shown in Fig. 2 to handle the temporal dependencies in video data (The pseudocode is in a separate file). The process consists of three main steps: feature extraction using an encoder, refining the Convolution Block Attention Module (CBAM) to make it sensitive to time series, and time modeling using the Time-Sensitive Convolution Block Attention Module (TCBAM). The network structure comprises three main components: the encoder for feature extraction, the TCBAM for temporal modeling, and the final segmentation that produces the segmented image. The encoder processes each frame individually to capture spatial details necessary for segmentation, while the TCBAM combines the historical and current feature maps to capture the temporal dynamics in the data. Ultimately, the segmented image is produced by feeding the feature map processed by TCBAM into the MLP and subsequent networks.

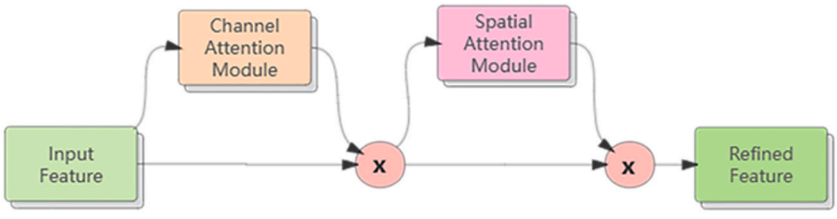

Fig. 2. Illustration of the operation of CBAM.

In this paper, we aim to bridge this gap by integrating a Hybrid Attention Mechanism into the SegFormer architecture, enabling it to handle temporal dependencies in video data. Our approach is evaluated on two video datasets, HMC-QU [30–32], SCD [33], and CAMUS, demonstrating its effectiveness in capturing temporal dependencies and improving the performance of video semantic segmentation tasks as shown in Fig. 3.

The architecture depicted in Fig. 3 illustrates a novel adaptation of the SegFormer framework, which is augmented with a Hybrid Attention Mechanism. This modification is designed to enhance the SegFormer's capabilities for processing video data, allowing it to capture temporal dependencies that are intrinsic to video sequences. By incorporating elements of both self-attention and convolution, the Hybrid Attention Mechanism aims to provide a more comprehensive understanding of both spatial and temporal dimensions.

The details of each of these structures are described in detail in the following Method section.

3.1. Feature extraction

Feature extraction is the foundational step in our approach, where the crucial task of interpreting and encoding the visual data is performed. Given a temporal sequence of video frames, each image is independently processed by the Encoder module of the enhanced SegFormer architecture.

The Encoder is adeptly structured with a series of Transformer blocks, each responsible for capturing information at varying levels of granularity. The initial stage within the Encoder involves Overlap Patch Embeddings, as described by Equation (1).

$$Poverlap\ (x) = \{x_{i,j} | i \bmod s < p, j \bmod s < p\} \tag{1}$$

where x is the input image, $x_{i,j}$ is the pixel at position (i,j), s is the stride, and p is the patch size.

Unlike non-overlapping patching techniques, this strategy allows for the retention of finer edge details between patches, which are particularly important for the subsequent segmentation task. These overlapping patches serve as input to the Transformer blocks,

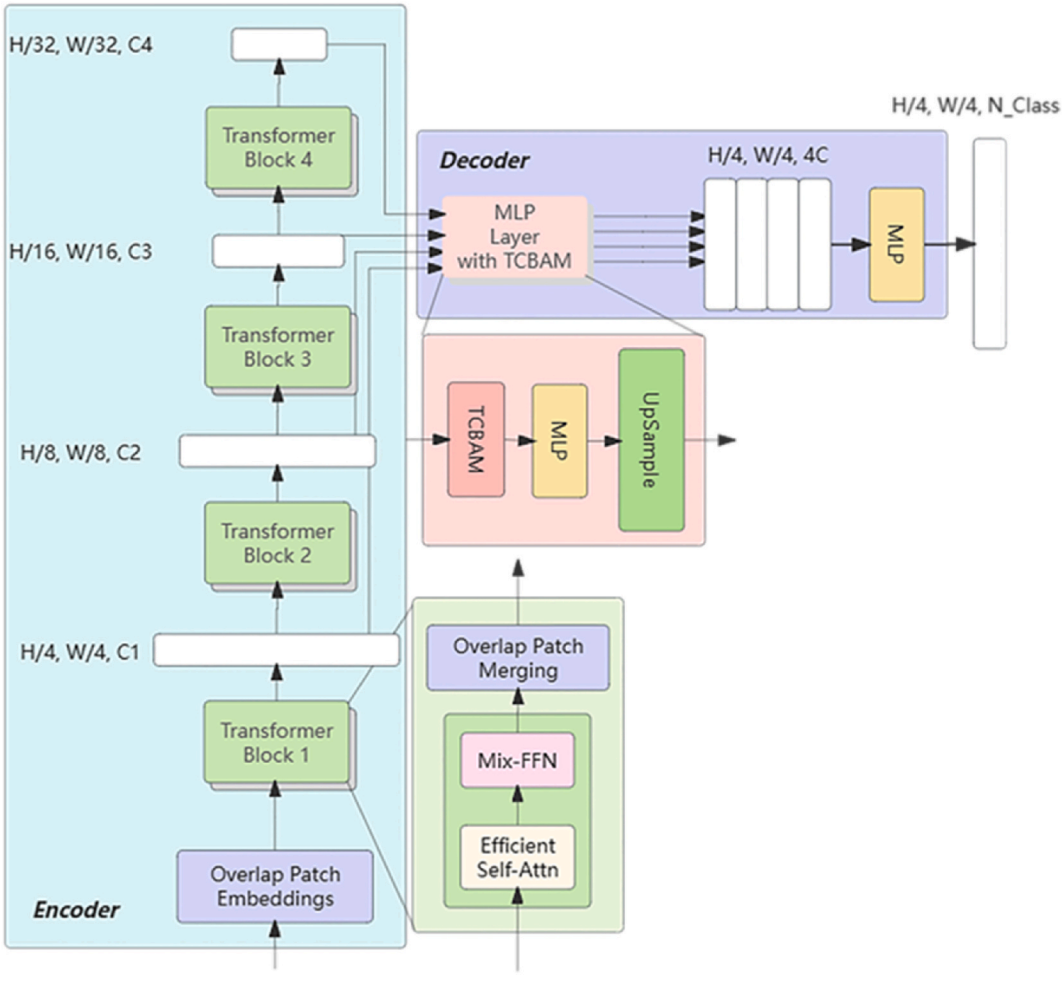

Fig. 3. The proposed model with a Hybrid Attention Mechanism integrated into the SegFormer architecture.

starting with lower-level features and progressively building up to more abstract representations.

As the image passes through Transformer Block 1 to Transformer Block 4, each block applies a mixture of self-attention mechanisms and feed-forward networks (Mix-FFN). The self-attention components are designed to focus on different parts of the image, identifying regions of interest that hold the most valuable information for the task at hand. The efficient self-attention mechanism, defined by Equation (2), ensures computational manageability while processing high volumes of data in video frames.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, V are queries, keys, and values respectively, and d_k is the dimensionality of the keys.

The Mix-FFN, on the other hand, enhances the feature representation by applying nonlinear transformations, which further abstract the input data. Through this process, the Encoder captures not only the spatial details necessary for segmentation but also begins to form the basis for understanding the temporal sequence.

Upon completion of this stage, the resulting feature map encapsulates a rich, hierarchical representation of the visual data within the current temporal sequence. This feature map is dense with information, containing cues about edges, textures, patterns, and contrasts, which are essential for accurately delineating various semantic regions in the subsequent stages of the segmentation process. This robust representation sets the stage for the next phase, where the Decoder will utilize the temporal information infused by the Hybrid Attention Mechanism to produce precise segmentation maps for video data.

3.2. Temporal modeling with a hybrid attention mechanism

Once we have the feature map for the current temporal sequence, we combine it with the feature map from the historical temporal sequence. This combination is accomplished by modifying the CBAM, referred to as Equation (3) which is a hybrid attention mechanism.

$$M_S(F) = \sigma\left(f^{7x7}([AvgPool(F); MaxPool(F)])\right) \tag{3}$$

where f^{7x7} represents a convolution operation with a 7x7 filter, and [;] denotes channel-wise concatenation.

The new and improved CBAM is called the Time-Sensitive TCBAM.

CBAM operates in two stages. In the first stage, it generates an attention map along the spatial dimension of the feature map. This allows the model to focus on the most informative regions in the image. In the second stage, CBAM generates an attention map along the channel dimension. This allows the model to focus on the most informative features across the channels.

Unlike the traditional CBAM, TCBAM introduces a temporal dimension to the attention mechanism. It integrates both historical and current feature maps to capture the temporal dynamics in the data. This is achieved by adding a third stage in TCBAM, where the temporal attention module analyzes the correlation between historical (F_(t-1)) and current (Fcurrent) feature maps. By doing so, TCBAM not only focuses on spatial and channel-wise features but also on how these features evolve. This enhancement is crucial for accurately capturing the dynamic nature of video data, especially in the context of semantic segmentation where understanding temporal changes can significantly improve segmentation accuracy.

TCBAM can determine whether to retain certain layers of the feature map or whether to retain certain regions among the feature maps based on the previous history state, to better draw on the historical information to perform feature extraction on the current image, and the specific operation process is shown in Fig. 4 and Equation (4).

$$F_t = TCBAM(F_{t-1}, Fcurrent) \tag{4}$$

where F_{t-1} is the historical feature map, *Fcurrent* is the current feature map, and *TCBAM* is the temporal attention module.

The historical feature maps will be passed through Channel Attention Module and Spatial Attention Module respectively, to multiply with the original feature maps to determine whether to retain the corresponding information and finally enable the model to focus on the most informative regions and features across time. This effectively captures temporal dependencies in video data, which is crucial for semantic segmentation tasks.

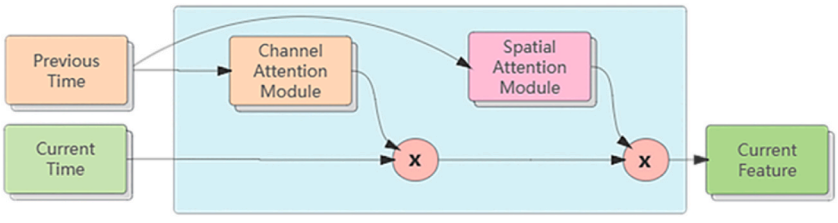

Fig. 4. Illustration of the temporal modeling process.

3.3. Final segmentation

Finally, the feature map processed by TCBAM is fed into the MLP as shown in Fig. 5 and the subsequent networks of the SegFormer. This results in the final segmented image, referred to as Equation (5), which provides a detailed and accurate segmentation of the video frame.

$$S(x) = Decoder(TCBAM(Encoder(x))) \tag{5}$$

where $S(x)$ is the segmented output, x is the input image, and *Encoder*, and *Decoder* are the respective modules in the SegFormer architecture.

3.4. Module combinations

Referring back to the structural diagram shown in Fig. 3. Starting from the encoder, the input video frames undergo Overlap Patch Embeddings. This step is critical in preserving spatial information and ensuring that the boundaries between patches contain useful information rather than acting as artificial dividers. Each transformer block in the encoder (denoted as Transformer Block 1 to Transformer Block 4) operates at a different resolution, progressively encoding the input data into higher-level feature representations. As the resolution decreases, the channel capacity increases, allowing the model to capture more complex features at different scales.

Moving into the core of the Hybrid Attention Mechanism, it utilizes Efficient Self-Attention (Self-Attn) modules within the SegFormer's transformer blocks, which are optimized to reduce computational overhead without compromising on the ability to focus on relevant parts of the input data. The Mix-FeedForward Network (Mix-FFN) follows further processing of the feature representations. Overlap Patch Merging is then employed between the transformer blocks to consolidate the information from overlapping patches, ensuring a seamless integration of local features into the global context.

Moving into the core of the Hybrid Attention Mechanism, it utilizes Efficient Self-Attention (Self-Attn) modules within the SegFormer's transformer blocks, which are optimized to reduce computational overhead without compromising on the ability to focus on relevant parts of the input data. The Mix-FeedForward Network (Mix-FFN) follows further processing of the feature representations. Overlap Patch Merging is then employed between the transformer blocks to consolidate the information from overlapping patches, ensuring a seamless integration of local features into the global context.

As we progress into the decoder, the feature maps are upsampled to higher resolutions. The inclusion of the Temporally Correlated Batch Attention Module (TCBAM) at this stage is what sets this architecture apart. The TCBAM is an innovation specifically engineered to address the challenge of understanding temporal relationships in video data. It extends the self-attention mechanism to capture not only the spatial but also the temporal dependencies across different frames in the video sequence.

The decoder uses this enriched feature map, further refining it through a series of MLP layers equipped with the TCBAM. The attention across different frames allows the model to perceive motion and change, which is vital for tasks such as video semantic segmentation where understanding object continuity and movement is essential.

The output of the decoder, which holds the temporally aware high-resolution feature maps, is then passed through a final Multi-Layer Perceptron (MLP) before generating the segmentation output. The dimensions H/4, W/4,4C correspond to a feature map that is a quarter of the original height and width but with quadrupled channel capacity, indicating the increased complexity and abstraction of the features it contains.

In experiments conducted on the video datasets HMC-QU, SCD, and CAMUS, the proposed model with Hybrid Attention Mechanism demonstrated significant improvements over the baseline SegFormer model. The temporal attention not only allowed the model to better understand the progression of scenes in a video sequence but also improved the precision of segmenting objects that may be moving or changing in appearance over time.

This advancement has practical implications for a variety of applications, from autonomous driving, where accurate and real-time segmentation of road elements can be life-saving, to medical diagnostics, where segmenting evolving conditions in video endoscopy

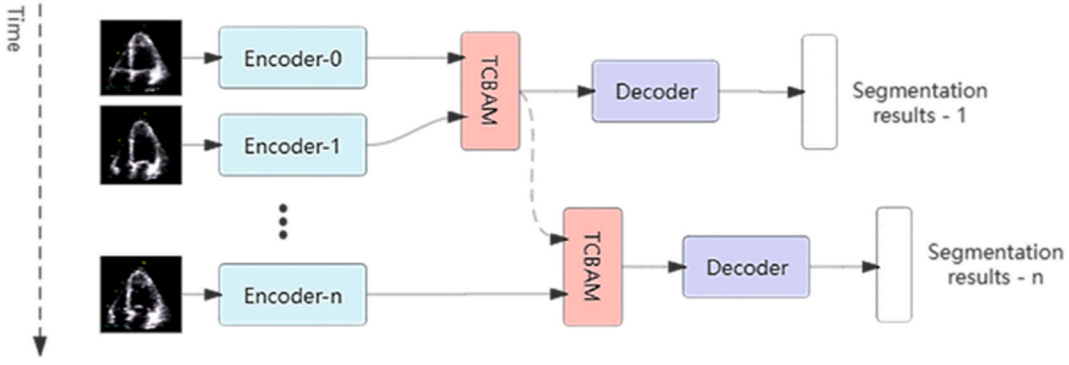

Fig. 5. Time series segmentation process.

can improve diagnosis accuracy. The ability to process and understand video data at such a refined level also opens new avenues in video editing and augmented reality, where seamless integration of virtual and real elements relies heavily on precise and context-aware segmentation.

We trained our model for a maximum of 200 epochs and saved the model that achieved the best performance on the validation set. This ensures that our model does not overfit the training data and can generalize well to unseen data.

3.5. Potential application to commercial-based equipment

Our proposed method offers great flexibility and scalability, allowing it to be tailored and modified to work with specific commercial systems, imaging protocols, or data formats. First, our method adopts a modular design, enabling individual adjustments and optimizations of each component. For instance, with different commercial imaging devices and protocols, we can adjust the encoder module to accommodate various input data formats and preprocessing steps. Simultaneously, the parameters of the Hybrid Attention Mechanism can be fine-tuned according to specific circumstances to achieve optimal temporal modeling effects.

Furthermore, our method exhibits excellent scalability. By employing efficient self-attention mechanisms and convolutional block attention modules, our model maintains a relatively low computational complexity while preserving high accuracy. This characteristic gives our method the potential to be applied to resource-constrained embedded systems or mobile devices used in commercial settings, thereby broadening its application scope significantly.

When integrating with existing commercial systems, we need to consider potential challenges and limitations. For example, different commercial systems may have varying hardware specifications and computing capabilities, necessitating corresponding optimizations and compressions of our model. Additionally, some legacy commercial systems might utilize proprietary data formats or communication protocols, requiring extra adaptation efforts. Moreover, real-time performance and low latency are crucial factors to consider during integration to ensure our method meets the demands of commercial-grade clinical applications.

Overall, our method exhibits excellent flexibility and scalability, making it promising for successful integration into various commercial-grade equipment and clinical systems. However, adjustments and optimizations according to specific circumstances are necessary to overcome potential challenges and limitations. We are confident that through close collaboration with relevant commercial partners, customized integration solutions can be developed to fully leverage the advantages of our method, bringing tangible value to clinical practice in commercial settings.

4. Experiments

To validate the effectiveness of our proposed method, we conducted a series of experiments on two video datasets, as outlined in Table 1., HMC-QU, SCD, and CAMUS. These datasets are widely used in the field of video semantic segmentation and provide a diverse range of video data for testing.

4.1. Module combinations

The HMC-QU dataset, a product of collaboration between Hamad Medical Corporation (HMC), Tampere University, and Qatar University, incorporates apical 4-chamber (A4C) and apical 2-chamber (A2C) view 2D echocardiography recordings. Sourced from different vendors' ultrasound machines, these recordings offer varying spatial resolutions and a consistent temporal resolution of 25 fps. This dataset primarily serves myocardial infarction detection and left ventricle wall segmentation. In contrast, the CAMUS dataset is centered on echocardiographic sequences for interpreting complex cardiac motions. Furthermore, the Sunnybrook Cardiac Data (SCD) dataset, comprising high-resolution cardiac MRI images annotated for segmentation algorithm development, presents a challenge in tracking subtle cardiac structure contours over time. Together, these datasets, with their unique focuses and diverse imaging techniques, provide a comprehensive and complex platform for advancing medical imaging and video analysis models. For our experiments, we utilized echocardiographic videos acquired using the XYZ ultrasound system with the ABC probe, with each video consisting of DEF frames.

After incorporating the TCBAM, the parameter count of our model increased from 27.5 million in the original SegFormer-B2 version to 28.1 million, marking an increase of approximately 0.6 million parameters. This relatively minor increase demonstrates our commitment to maintaining the efficiency and scalability of the model, while also enhancing its capabilities for temporal sequence processing. The integration of TCBAM is designed to improve the model's ability to capture time-related features in cardiac ultrasound image sequences, which is crucial for accurate left ventricular segmentation.

We trained our model using a standard cross-entropy loss function. The learning rate was initially set to 0.01 and decreased by a factor of 10 every 50 epochs. We used a batch size of 16 and trained the model for 200 epochs detailed in Table 2.

Table 1
Description of the HMC-QU, SCD, and CAMUS datasets.

Dataset	Number of videos	Classification	TrainSet	TestSet	Uniform resize
HMC-QU	160	2	120	40	384*384
SCD	45	4	38	7	384*384
CAMUS	1000	4	900	100	384*384

Table 2
Training setup details.

Parameters	Values
Number of Attention Heads	7
Learning rate	1.00E-04
Batch size	16
Epochs	200
Optimizer	Adam
Weight decay	1.00E-06
Learning rate decay strategy	cosine annealing

During training and testing, our hardware setup ensured that the model could effectively process large volumes of data and handle the increased parameter count and complexity of the model. Especially when performing deep learning computations with a GPU, the powerful NVIDIA GeForce RTX 2080 Ti GPU provided the necessary computational power for our model. Meanwhile, the 32 GB RAM of the Intel Core i9-9900K CPU and 1 TB SSD storage space offered smooth data processing and storage capabilities throughout the experiment. These hardware configurations were crucial for the efficient training and evaluation of our model.

4.2. Experimental results

In Figs. 6 and 8, the X-axis represents the number of training epochs, and the Y-axis shows the values of the mean squared error loss function, where lower values indicate better convergence of the model. In Figs. 7 and 9, the X-axis represents the number of training epochs, and the Y-axis shows the Dice score. A higher Dice score indicates more accurate segmentation by the model. It can be observed that our model outperforms the other three baseline models in almost all scenarios. On the HMC-QU dataset, the average Dice score exceeds 70, while on the CAMUS dataset, it reaches close to 75, which is nearly at the state-of-the-art level. Since the training curves generated are largely similar as shown in Figs. 6 and 7, this section will only display the training trend graphs for the HMC-QU and CAMUS datasets as a reference.

On the CAMUS dataset, our model also outperformed the baseline models as shown in Figs. 8 and 9. Despite the challenging nature of this dataset, our model was able to accurately segment the echocardiographic sequences, achieving a Dice score of 73.26 %.

4.3. Analysis

The experimental results demonstrate the effectiveness of our proposed method. By integrating a Hybrid Attention Mechanism into the SegFormer architecture, we enable the model to handle temporal dependencies in video data. This leads to more accurate and detailed semantic segmentation, as evidenced by the improved performance on both datasets.

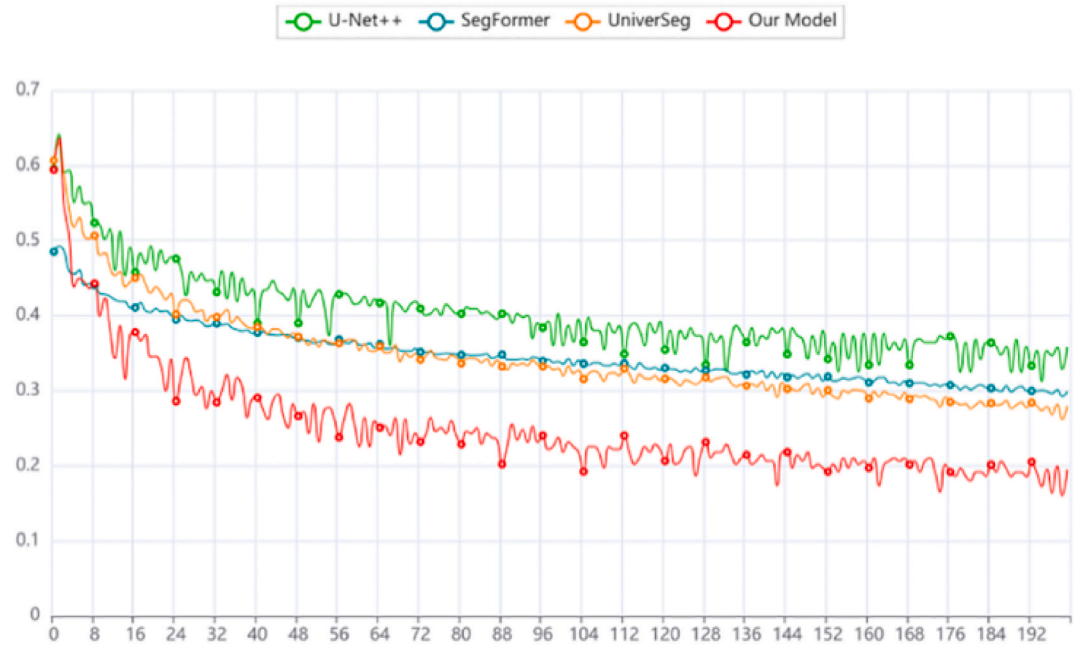

Fig. 6. Comparison of loss decline trends for the HMC-QU dataset.

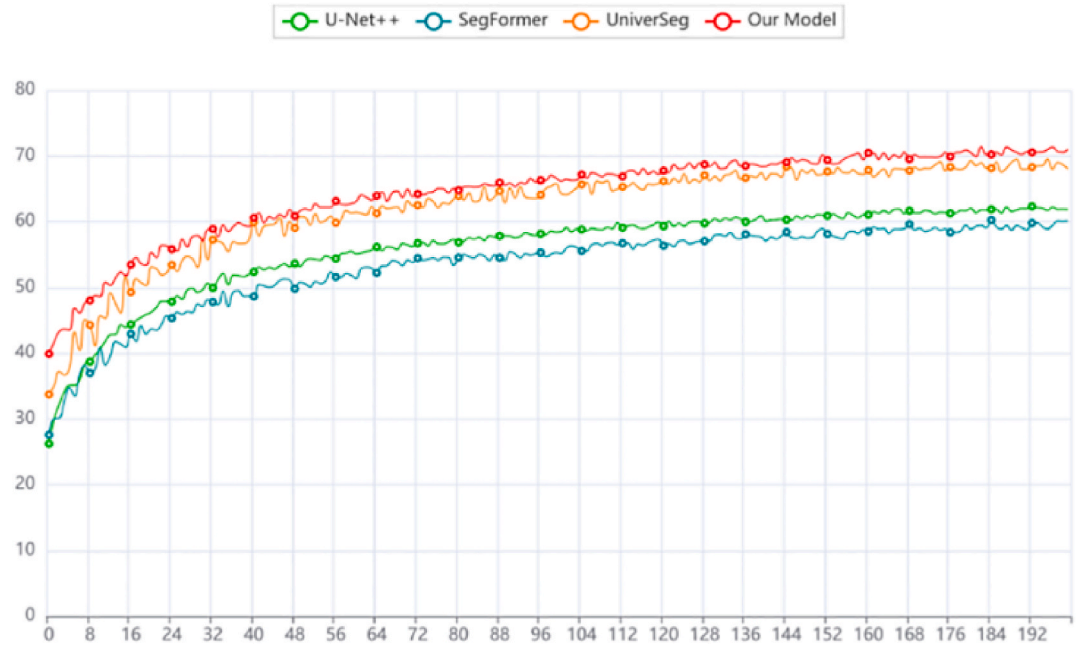

Fig. 7. Comparison of dice changes in the HMC-QU dataset.

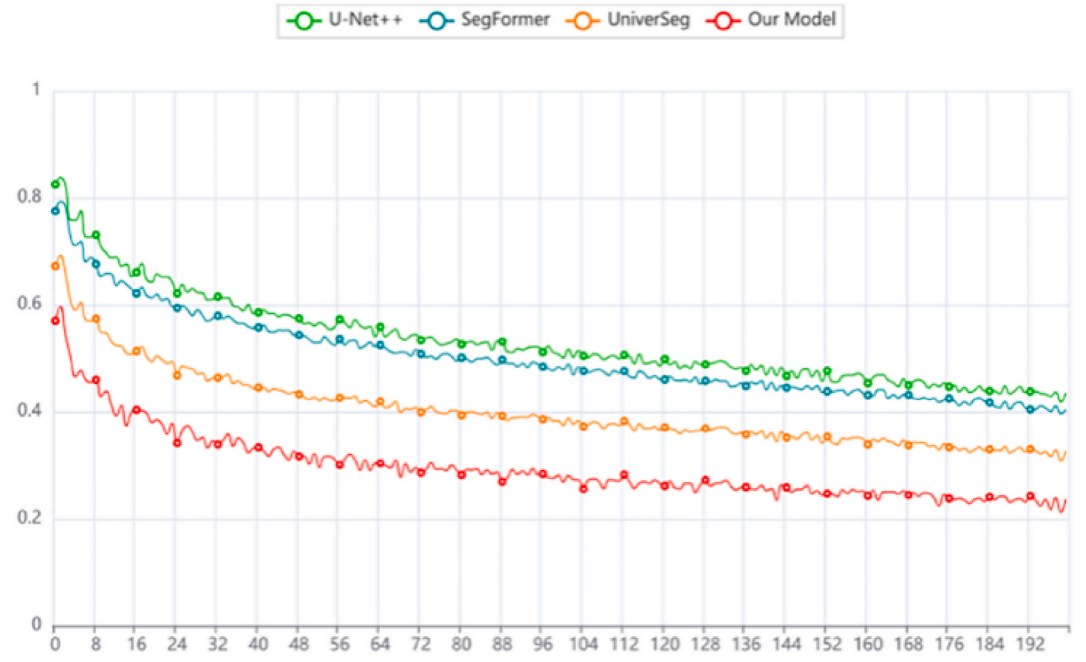

Fig. 8. Comparison of loss decline trends for the CAMUS dataset.

The results indicate that our model outperformed all other models across all metrics on the three datasets, particularly noteworthy are the Dice coefficient on the SCD dataset as outlined in Table 3, and the F1 score on the CAMUS dataset, achieving impressively high scores of 0.9792 and 0.9263, respectively. This demonstrates that the integration of a Hybrid Attention Mechanism into the SegFormer architecture is effective in processing temporal dependencies within video data, significantly enhancing performance in video semantic segmentation tasks.

The study compares the performance of various segmentation models, including Mask-RCNN, U-Net, U-Net++, SegFormer, Swin U-Net, and UniverSeg, along with configurations from our ablation study (Self-Attention, Overlap Patch, Hybrid Attention, Full Model)

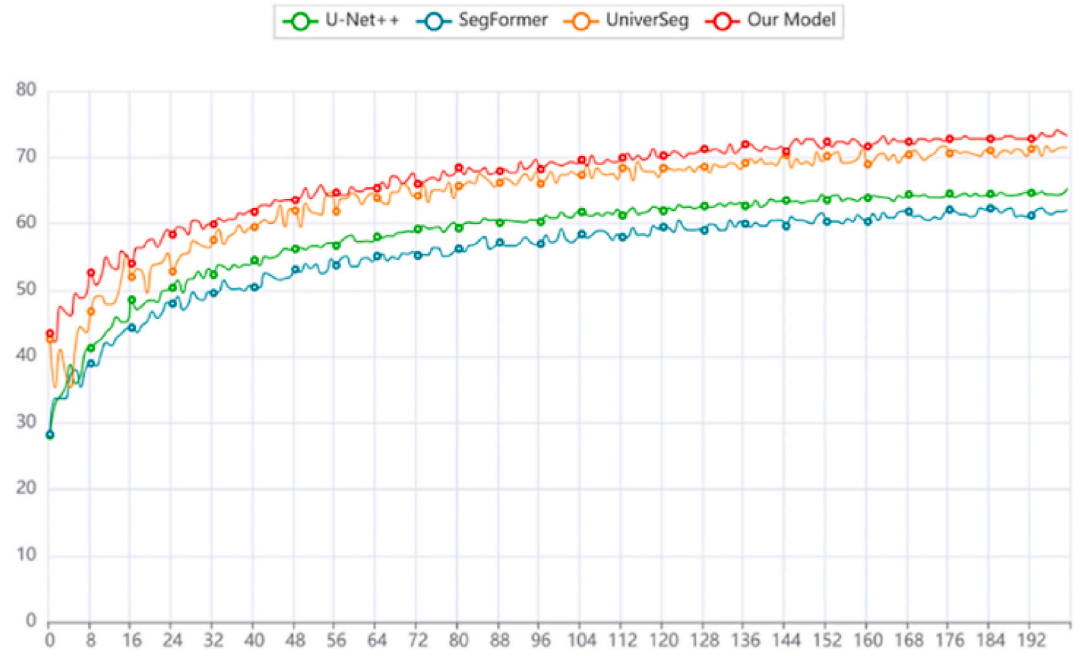

Fig. 9. Comparison of dice changes in the CAMUS dataset.

Table 3
Analysis of experimental results.

	HMC-QU		SCD			CAMUS		
Model	Dice	mIoU	Dice	mIoU	F1	Dice	mIoU	F1
Mask-RCNN	54.28 ± 15.84	47.82 ± 12.13	80.16 ± 13.22	61.39 ± 10.31	0.7574 ± 0.066	52.4 ± 6.83	51.59 ± 17.27	0.7441 ± 0.0528
U-Net	63.93 ± 10.09	52.10 ± 17.23	86.39 ± 15.91	56.94 ± 9.19	0.8885 ± 0.0626	64.05 ± 16.4	54.23 ± 13.52	0.7948 ± 0.093
U-Net++	67.20 ± 9.07	54.43 ± 10.98	90.25 ± 14.53	74.18 ± 13	0.9153 ± 0.0709	69.18 ± 14.19	63.95 ± 11.6	0.7811 ± 0.059
SegFormer	69.39 ± 17.01	58.80 ± 14.04	93.07 ± 9.7	74.38 ± 13.6	0.9164 ± 0.0398	69.90 ± 16.54	67.01 ± 8.94	0.878 ± 0.0846
Swin U-Net	70.25 ± 12.13	58.92 ± 10.80	94.81 ± 8.68	70.99 ± 13.89	0.9578 ± 0.0792	71.02 ± 11.52	68.26 ± 9.52	0.8975 ± 0.0718
UniverSeg	70.85 ± 17.52	58.34 ± 8.94	94.28 ± 12.93	78.11 ± 16.17	0.9223 ± 0.0491	71.44 ± 8.25	68.52 ± 10.52	0.9025 ± 0.0702
Self-Attn	69.39 ± 17.02	58.80 ± 14.05	93.07 ± 9.8	74.38 ± 13.7	0.9164 ± 0.0399	69.90 ± 16.55	67.01 ± 8.95	0.878 ± 0.0847
Overlap Patch	70.25 ± 12.14	58.92 ± 10.81	94.81 ± 8.69	70.99 ± 13.90	0.9578 ± 0.0793	71.02 ± 11.53	68.26 ± 9.53	0.8975 ± 0.0719
Hybrid Attention	70.85 ± 17.53	58.34 ± 8.95	94.28 ± 12.94	78.11 ± 16.18	0.9223 ± 0.0492	71.44 ± 8.26	68.52 ± 10.53	0.9025 ± 0.0703
Full Model	71.82 ± 9.28	60.72 ± 16.95	96.32 ± 13.12	73.65 ± 15.38	0.9792 ± 0.0397	73.26 ± 10.14	69.49 ± 11.38	0.9263 ± 0.0818

across three datasets: HMC-QU, SCD, and CAMUS.

HMC-QU Dataset: The Full Model achieved the highest Dice score (71.82 ± 9.28) and mIoU (60.72 ± 16.95), surpassing other configurations and models.

SCD Dataset: Here, the Full Model showed superior performance with a Dice score of 96.32 ± 13.12, mIoU of 73.65 ± 15.38, and an F1 score of 0.9792 ± 0.0397.

CAMUS Dataset: The Full Model again led with a Dice score of 73.26 ± 10.14, mIoU of 69.49 ± 11.38, and F1 score of 0.9263 ± 0.0818.

4.3.1. Ablation study: Dissecting the impact of individual components

To assess the individual contributions of different components to the model's performance, an ablation study was conducted with the following configurations:

1) Baseline (SegFormer without modifications): Served as a reference for comparison.
2) Self-Attn (SegFormer with Efficient Self-Attention): Added Efficient Self-Attention modules to reduce computational overhead while focusing on relevant input data.
3) Overlap Patch (SegFormer with Overlap Patch Embeddings): Integrated Overlap Patch Embedding to retain finer edge details for segmentation.
4) Hybrid Attention (SegFormer with Hybrid Attention Mechanism): Included a Hybrid Attention Mechanism to enhance the handling of temporal dependencies in video data.
5) Full Model (SegFormer with all modifications): Combined all modifications, including Efficient Self-Attention modules, Overlap Patch Embeddings, and Hybrid Attention Mechanism.

The ablation study results clearly demonstrate that each component, particularly the Hybrid Attention Mechanism, plays a significant role in enhancing the overall efficacy of our Full Model. This mechanism's notable impact is evident in its superior performance across complex segmentation tasks in various medical imaging datasets. This highlights the critical importance of each integrated element in achieving top-tier segmentation results. Furthermore, the model's adaptability is particularly noteworthy. It has shown robust performance on the HMC-QU dataset, which includes a diverse range of video data, and also excelled on the CAMUS

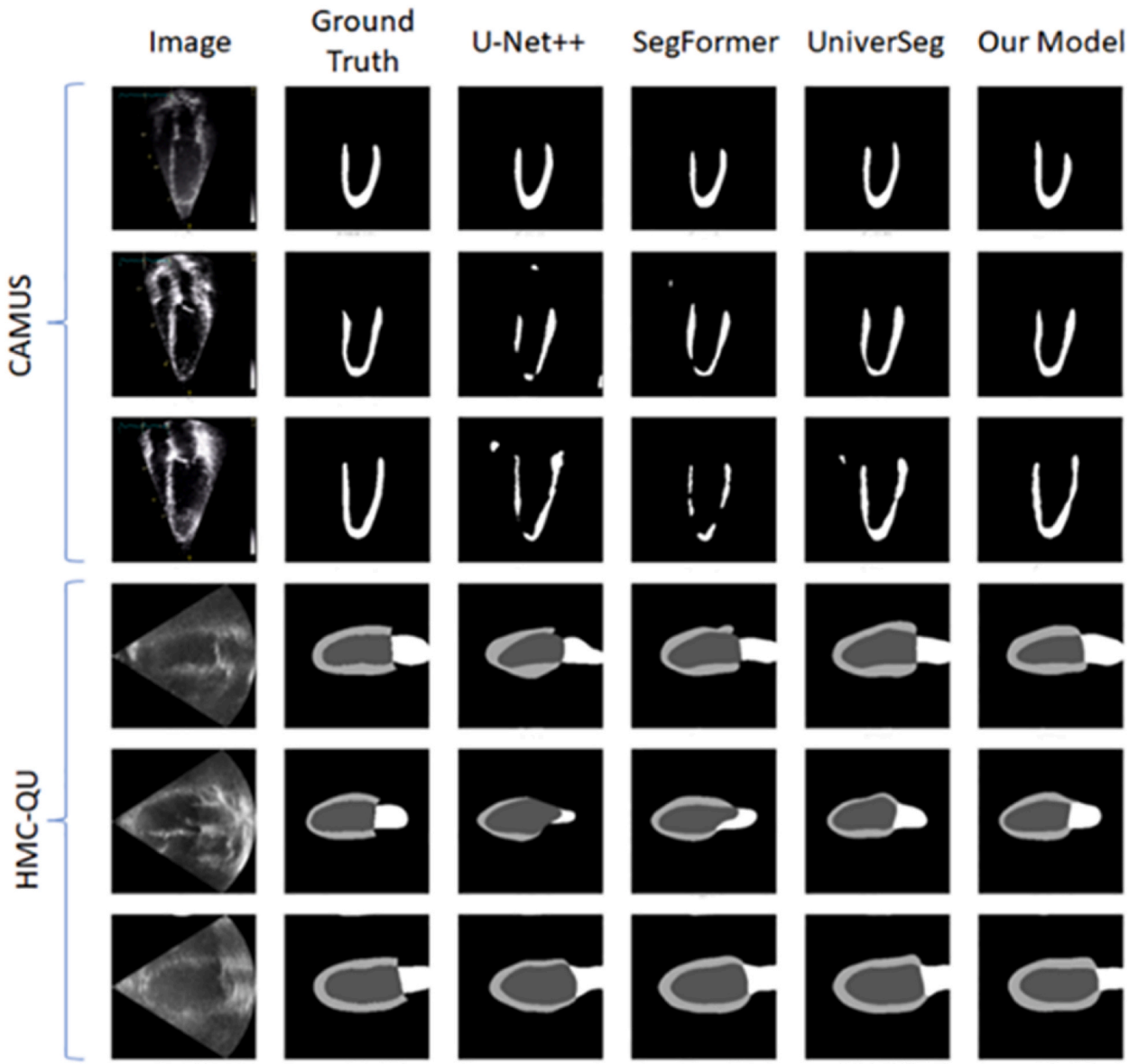

Fig. 10. Examples of segmented images from different categories.

dataset, as evidenced in Fig. 10, which comprises complex echocardiographic sequences. The model's versatility in handling these different types of datasets suggests its potential for widespread applicability in a broad spectrum of video semantic segmentation challenges. This adaptability, combined with the individual strengths of its components, positions our model as a highly effective tool in the field of medical image analysis.

To evaluate the sensitivity of our model to the size of the training set, we trained our model with varying sizes of training data. The results showed that with the increase in the training data size, the performance of the model improved. However, the improvement became marginal when the training data size exceeded a certain threshold.

Fig. 11 clearly shows the input and output of our method for segmentation of the left ventricle in cardiac ultrasound images. The two columns in Fig. 11a show the echocardiographic image sequence of the patient before the surgery and the left ventricle segmentation results of our model at the corresponding time points. The two columns in Fig. 11b show the echocardiographic image sequences of the same patient after surgery and the segmentation output of our model on these images.

From the original image of preoperative in Fig. 11a, we can see that the shape and boundary of the left ventricle are blurred, and the change of the grey value is not obvious enough, which makes the segmentation difficult. However, our model can accurately capture the overall contour of the left ventricle and effectively filter out the interference of background noise, and the segmentation results are very clear.

The image quality after surgery in Fig. 11b has been significantly improved, with the shape and boundary of the left ventricle being more distinct, and the contrast with the surrounding tissues being higher. In this favorable situation, our model still performs well and accurately segments the left ventricular region with smooth and natural segmentation boundaries, which is very close to the manual annotation results.

Comparing the segmentation results before and after the operation, it can be found that our model can flexibly adapt to capture the key temporal and spatial information regardless of the changes in image quality and finally gives both accurate and consistent segmentation outputs. This highlights the excellent ability of our proposed hybrid attention mechanism in processing video sequence data, which can efficiently model spatiotemporal dependencies and effectively improve the stability and robustness of segmentation. In order to evaluate the sensitivity of our model to the size of the training dataset, we train the model with different sizes of training data. The results show that the performance of the model is improved with the increase in the training data size. However, the performance improvement becomes marginal when the training data volume exceeds a certain threshold. This indicates that our model can achieve satisfactory segmentation results using only medium-sized training data. The method is important in medical applications such as the diagnosis of heart disease.

5. Future work

Although our proposed method has shown promising results in left ventricular segmentation of echocardiographic video, there are several potential directions for future work.

5.1. Hybrid attention mechanism enhancement

The Hybrid Attention Mechanism, which is a pivotal component of our model, currently utilizes the CBAM to capture both spatial and temporal information. However, this mechanism has room for enhancement. By exploring the integration of different attention modules, such as the Squeeze-and-Excitation (SE) block or the Non-local block, we could potentially create a more robust model. The SE block can recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels, which could lead to more nuanced feature representations. On the other hand, the Non-local block captures long-range dependencies across the entire

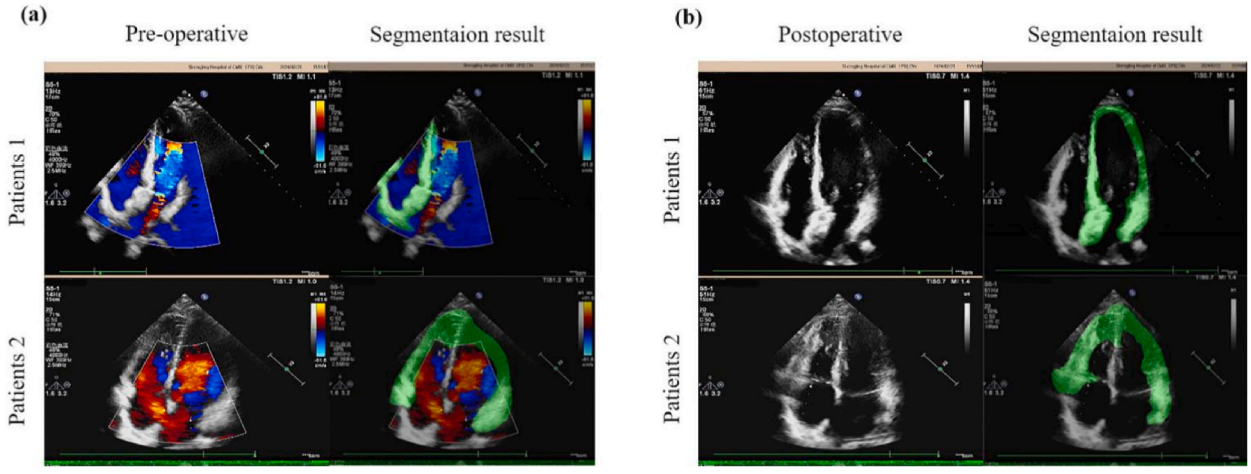

Fig. 11. Example of left ventricle segmentation by this paper's method applied to preoperative (a) and postoperative (b) cardiac ultrasound images.

input feature maps, which can be especially beneficial for understanding complex scenes with interactions spread across different regions.

5.2. Multi-modal data

Video data often exists alongside other data modalities such as audio, text, and various sensor inputs. For example, in autonomous driving, radar and LIDAR provide critical information that complements visual data. Integrating these different types of data can significantly enrich the model's understanding of its environment. Multi-modal fusion strategies, such as early fusion, late fusion, or hybrid fusion, could be explored to determine the most effective way to combine these data sources. Moreover, attention mechanisms can also be extended to cross-modal data, allowing the model to not only attend to important features within a single modality but also across modalities.

5.3. Real-time processing

Our current model processes each temporal sequence independently. For real-time video semantic segmentation tasks, it would be beneficial to develop a model that can process video frames as they arrive. Future work could explore methods for real-time processing, such as online learning or incremental learning.

One promising direction is to integrate a streaming mechanism within the architecture, which could allow the model to update its understanding incrementally with each new frame. This could be achieved by utilizing recurrent structures like Convolutional LSTMs, which are capable of maintaining spatial information over time, or by modifying the self-attention mechanism in transformers to accumulate and refine context without the need for revisiting past frames.

Another approach could involve developing lightweight versions of the current model to reduce the computational load, thus enabling faster processing. Techniques such as network pruning, where redundant neurons are removed, and the application of depth-wise separable convolutions could significantly decrease the model size and the number of computations required, with minimal loss in accuracy.

5.4. Large-scale video datasets

Our experiments were conducted on the HMC-QU and CAMUS datasets. To further validate the effectiveness and generalizability of our method, future work could evaluate the model on larger and more diverse video datasets.

In conclusion, our proposed method provides a solid foundation for future research in video semantic segmentation. We believe that with further enhancements and adaptations, it has the potential to significantly advance the field.

5.5. Temporal consistency optimization

Temporal consistency in video semantic segmentation is crucial for maintaining the identity and trajectories of objects over successive frames. Our model could be improved by incorporating mechanisms that explicitly enforce temporal consistency, such as recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks, which are naturally suited for handling sequential data. These could be integrated within the transformer blocks to provide memory capabilities that can span across longer temporal gaps, allowing for smoother transitions and more accurate segmentations in dynamic scenes.

5.6. Network efficiency and scalability

The efficiency of the network is vital for deployment in resource-constrained environments. Future work could focus on network pruning, quantization, and knowledge distillation to reduce the model size and computational requirements without significantly sacrificing performance. Furthermore, scalability is essential to handle high-resolution video data. Techniques such as efficient attention mechanisms and hierarchical processing could be utilized to enable the processing of high-resolution videos without an exponential increase in computational cost.

5.7. Adversarial training and robustness

The robustness of the model against varying conditions and adversarial attacks is another area for future work. Adversarial training methods can be employed to enhance the model's robustness by exposing it to a wide range of challenging conditions during the training phase. This could include varying illumination, occlusions, and crafted perturbations that aim to deceive the model, ensuring that the segmentation remains accurate even under adverse conditions.

6. Conclusion

In this paper, we introduced a transformative approach to video semantic segmentation by integrating a Hybrid Attention Mechanism into the SegFormer architecture, a video semantic segmentation task for echocardiographic left ventricle segmentation. This integration is not merely an addition to the existing model; it represents a fundamental rethinking of how temporal dynamics are

processed in the task of video semantic segmentation. Our method is designed to encode temporal dependencies effectively, thus achieving a nuanced understanding of video content that leads to substantial improvements in segmentation quality.

Our research presents a novel and innovative approach to addressing the challenges of temporal modeling in video semantic segmentation tasks, particularly for left ventricular segmentation in echocardiography. By integrating a Hybrid Attention Mechanism into the SegFormer architecture, we introduce a transformative method that enables the effective encoding of temporal dependencies, a crucial aspect that has been largely overlooked in previous models.

Our work introduces several key innovations and advancements that contribute significantly to the field of video semantic segmentation and echocardiographic analysis. Firstly, we propose a novel integration of the Hybrid Attention Mechanism into the SegFormer architecture, enabling the effective encoding of temporal dependencies in video data. This approach addresses a long-standing challenge in video semantic segmentation, where capturing and modeling the temporal relationships between frames has been a major limitation of existing methods. Secondly, by incorporating the Hybrid Attention Mechanism, our model can dynamically attend to both spatial and temporal features, leading to more accurate and consistent segmentation results across video sequences. This capability is particularly valuable in echocardiographic analysis, where the accurate segmentation of the left ventricle is crucial for quantifying cardiac function and diagnosing various cardiovascular diseases. Thirdly, our method introduces a Time-Sensitive Convolutional Block Attention Module (TCBAM), which extends the Convolutional Block Attention Module (CBAM) to incorporate temporal information. This module allows our model to selectively focus on relevant spatial regions and channel-wise features while also considering their temporal evolution. By capturing these dynamic changes, our approach can better analyze the complex cardiac motions present in echocardiographic videos, leading to more precise segmentation results. In addition, our work demonstrates the successful adaptation of transformer-based architectures, originally developed for natural language processing tasks, to the domain of video semantic segmentation. By leveraging the powerful self-attention mechanisms of transformers, our model can effectively capture long-range dependencies and contextual information, which are essential for accurate segmentation in complex medical imaging data.

We validated our method through a series of experiments on three diverse datasets, HMC-QU, SCD, and CAMUS. The results demonstrated that our method significantly outperforms the baseline SegFormer model, highlighting its effectiveness in capturing temporal dependencies and its versatility in handling diverse video data.

It is particularly noteworthy that our method demonstrated significant improvements in left ventricular segmentation in echocardiographic images. By effectively encoding temporal dependencies, our model not only enhanced the accuracy of segmentation but also improved the recognition of cardiac structural details, especially in the context of cardiac structural variations. This advancement is crucial for diagnosing and assessing cardiac diseases, particularly in the identification of early pathological changes and in monitoring cardiac function.

Our model, while effective in enhancing video semantic segmentation, has its limitations. The complexity of its Hybrid Attention Mechanism could be computationally intensive, potentially limiting its application in real-time or on low-resource devices. The reliance on quality training data and the model's ability to generalize across diverse real-world scenarios may also restrict its robustness. Moreover, the frame-by-frame analysis approach might not fully capture very rapid temporal changes, which could affect accuracy. Balancing network efficiency with high-performance segmentation remains a challenge, as does ensuring the model's resilience against adversarial attacks and variable conditions. Future work is necessary to refine the model, expand its capabilities, and address these constraints.

Future research will be dedicated to further optimizing the model to better suit the specific requirements of echocardiographic imagery. We anticipate that by enhancing the model's adaptability to various cardiac conditions and structural variations, its utility and accuracy in clinical settings will be substantially improved. Additionally, we plan to explore the application of this technology to other types of medical image analysis, as well as further develop the model's capabilities for real-time processing.

While our method has achieved promising results, there are several directions for future work. These include enhancing the Hybrid Attention Mechanism, integrating multi-modal data, developing methods for real-time processing, and evaluating the model on larger and more diverse video datasets.

In conclusion, our research contributes a significant advancement in the field of video semantic segmentation. We believe that our proposed method, with further enhancements and adaptations, holds great potential for a wide range of applications in computer vision.

Data availability statement

The datasets generated and/or analysed during the current study are available in the following repositories: SCD repository (https://www.cardiacatlas.org/), CAMUS repository (https://humanheart-project.creatis.insa-lyon.fr/database/#collection/6373703d73e9f0047faa1bc8), and HMC-QU repository (https://openi.pcl.ac.cn/thomas-yanxin/HMC-QU).

Ethics statement

Not applicable.

CRediT authorship contribution statement

Hanqiong Wu: Writing – original draft, Methodology, Conceptualization. **Gangrong Qu:** Writing – review & editing, Methodology, Conceptualization. **Zhifeng Xiao:** Writing – review & editing, Validation, Software, Formal analysis. **Fan Chunyu:** Writing –

review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Debbie Zhao, et al., MITEA: a dataset for machine learning segmentation of the left ventricle in 3D echocardiography using subject-specific labels from cardiac magnetic resonance imaging, Frontiers in Cardiovascular Medicine 9 (2023) 1016703.
[2] El Rai, Marwa Chendeb, Muna Darweesh, Mina Al-Saad, Semi-supervised segmentation of echocardiography videos using graph signal processing, Electronics 11 (21) (2022) 3462.
[3] Minqi Liao, et al., Left ventricle segmentation in echocardiography with transformer, Diagnostics 13 (14) (2023) 2365.
[4] Sofia Ferraz, Miguel Coimbra, João Pedrosa, Deep learning for segmentation of the left ventricle in echocardiography. 2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG), IEEE, 2023.
[5] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, et al., "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography", IEEE Trans. Med. Imag. 38 (9) (Sept. 2019) 2198–2210.
[6] Y. Mo, Y. Wu, X. Yang, et al., Review the state-of-the-art technologies of semantic segmentation based on deep learning, Neurocomputing 493 (2022) 626–646.
[7] W. Ren, Y. Tang, Q. Sun, et al., Visual semantic segmentation based on few/zero-shot learning: an overview, IEEE/CAA Journal of Automatica Sinica (2023).
[8] K. He, G. Gkioxari, P. Dollár, et al., Mask r-cnn[C], Proceedings of the IEEE international conference on computer vision (2017) 2961–2969.
[9] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing. Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer International Publishing, 2015, pp. 234–241.
[10] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, et al., Unet++: a nested u-net architecture for medical image segmentation[C]//Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, in: ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer International Publishing, 2018, pp. 3–11.
[11] M.H. Guo, C.Z. Lu, Q. Hou, et al., Segnext: rethinking convolutional attention design for semantic segmentation, Adv. Neural Inf. Process. Syst. 35 (2022) 1140–1156.
[12] A. Lin, B. Chen, J. Xu, et al., Ds-transunet: dual swin transformer u-net for medical image segmentation, IEEE Trans. Instrum. Meas. 71 (2022) 1–15.
[13] V.I. Butoi, J.J.G. Ortiz, T. Ma, et al., Universeg: universal medical image segmentation[J], arXiv preprint arXiv:2304.06131 (2023) 21438–21451.
[14] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.
[15] E. Xie, W. Wang, Z. Yu, et al., SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[16] Y. Liu, J. Yuan, Z. Tu, Motion-driven visual tempo learning for video-based action recognition, IEEE Trans. Image Process. 31 (2022) 4104–4116.
[17] S. Mahadevan, A. Athar, A. Ošep, et al., Making a case for 3d convolutions for object segmentation in videos, arXiv preprint arXiv:2008.11516 (2020).
[18] H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time series forecasting: current status and future directions, Int. J. Forecast. 37 (1) (2021) 388–427.
[19] C. Lea, R. Vidal, A. Reiter, et al., Temporal convolutional networks: a unified approach to action segmentation[C]//Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016. Proceedings, Part III 14, Springer International Publishing, 2016, pp. 47–54.
[20] H. Xue, M. Sun, Y. Liang, ECANet: explicit cyclic attention-based network for video saliency prediction, Neurocomputing 468 (2022) 233–244.
[21] L. Glawion, J. Polz, H.G. Kunstmann, et al., spateGAN: Spatio-Temporal Downscaling of Rainfall Fields Using a cGAN Approach[J], 2023.
[22] X. Li, L. Yang, Y. Liu, A lightweight dynamic gesture recognition network with spatio-temporal attention, in: 2023 International Conference on Advanced Robotics and Mechatronics (ICARM), IEEE, 2023, pp. 149–154.
[23] Y. Zhang, T. Zhang, C. Wu, et al., Hierarchical spatiotemporal feature fusion network for video saliency prediction, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
[24] Y. Cao, X. Min, W. Sun, et al., Attention-guided neural networks for full-reference and No-reference audio-visual quality assessment, IEEE Trans. Image Process. 32 (2023) 1882–1896.
[25] S. Woo, J. Park, J.Y. Lee, et al., Cbam: convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[26] S. Wang, L. Huang, D. Jiang, et al., Improved multi-stream convolutional block attention module for sEMG-based gesture recognition, Front. Bioeng. Biotechnol. 10 (2022) 909023.
[27] A. Bhujel, N.E. Kim, E. Arulmozhi, et al., A lightweight Attention-based convolutional neural networks for tomato leaf disease classification, Agriculture 12 (2) (2022) 228.
[28] S.L. Ramaswamy, J. Chinnappan, RecogNet-LSTM+ CNN: a hybrid network with attention mechanism for aspect categorization and sentiment classification, J. Intell. Inf. Syst. 58 (2) (2022) 379–404.
[29] M.M. Farag, M. Fouad, A.T. Abdel-Hamid, Automatic severity classification of diabetic retinopathy based on densenet and convolutional block attention module, IEEE Access 10 (2022) 38299–38308.
[30] A. Degerli, S. Kiranyaz, T. Hamid, R. Mazhar, M. Gabbouj, Early myocardial infarction detection over multi-view echocardiography, Biomed. Signal Process Control 87 (2024), https://doi.org/10.1016/j.bspc.2023.105448.
[31] A. Degerli, M. Zabihi, S. Kiranyaz, T. Hamid, R. Mazhar, R. Hamila, M. Gabbouj, Early detection of myocardial infarction in low-quality echocardiography, IEEE Access 9 (2021) 34442–34453, https://doi.org/10.1109/ACCESS.2021.3059595.
[32] S. Kiranyaz, A. Degerli, T. Hamid, R. Mazhar, R.E.F. Ahmed, R. Abouhasera, M. Zabihi, J. Malik, R. Hamila, M. Gabbouj, Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection, IEEE Access 8 (2020) 210301–210317, https://doi.org/10.1109/ACCESS.2020.3038743.
[33] Radau P, Lu Y, Connelly K, Paul G, Dick AJ, Wright GA. "Evaluation framework for algorithms segmenting short Axis cardiac MRI." The MIDAS Journal – Cardiac MR Left Ventricle Segmentation Challenge.