Multimedia Tools and Applications (2024) 83:6093–6111
https://doi.org/10.1007/s11042-023-15524-5

TC-SegNet: robust deep learning network for fully automatic two-chamber segmentation of two-dimensional echocardiography

Shyam Lal[1]

Received: 9 March 2022 / Revised: 3 October 2022 / Accepted: 19 April 2023 /
Published online: 3 June 2023

Abstract

Heart chamber quantification is an essential clinical task to analyze heart abnormalities by evaluating the heart volume estimated through the endocardial border of the chambers. A precise heart chamber segmentation algorithm using echocardiography is essential for improving the diagnosis of cardiac disease. This paper proposes a robust two chamber segmentation network (TC-SegNet) for echocardiography which follows a U-Net architecture and effectively incorporates the proposed modified skip connection, Atrous Spatial Pyramid Pooling (ASPP) modules and squeeze and excitation modules. The TC-SegNet is evaluated on the open-source fully annotated dataset of cardiac acquisitions for multi-structure ultrasound segmentation (CAMUS). The proposed TC-SegNet obtained an average value of F1-score of 0.91, an average Dice score of 0.9284 and an IoU score of 0.8322 which are higher than the reference models used here for comparison. Further, Pixel error (PE) of 1.5109 which are significantly less than the comparison models. The segmentation results and metrics show that the proposed model outperforms the state-of-the-art segmentation methods.

Keywords Atrous Spatial Pyramid Pooling (ASPP) · Cardiac segmentation · Deep learning · Echocardiography · Left atrium · Left ventricle · Myocardium · Residual path connections · Squeeze and Excitation

1 Introduction

Two-dimensional echocardiography is a widely used imaging modality for a non-invasive assessment of cardiac structures due to its short acquisition times [38] and good temporal resolution [14]. However, interpretation of echocardiographic images are operator dependent and hence suffers from variability in image acquisition, subjective interpretation [23]

✉ Shyam Lal
shyam.mtec@gmail.com

[1] Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Mangaluru 575025, Karnataka, India

and inter and intra-observer variability. Thus, the assessment of 2D echocardiography has yet remained unsatisfactory. The primary interest in medical imaging is segmenting the image into partitions of anatomically significant regions, based on which various clinical and geometric parameters can be interpreted. This will aid in the evaluation of the diagnosis and prognosis of the patient. These parameters play a paramount role in accuracy and efficiency in computer-aided diagnosis. Segmentation of cardiac structures is the first step towards further quantitative analysis like calculating the clinical indices such as ejection fraction and LV volumes.

Semi-automatic methods were widely used for echocardiographic segmentation before the advent of deep learning techniques. These include atlas-based methods [20], active appearance model [31], motion-based method [28], deformable models (BEAS, level-set) [1, 34] and graph-based methods [2]. Although these approaches generally achieve ideal results, semantic segmentation remains challenging due to the intricacies in feature representation. Moreover, these semi-automatic methods are time-consuming and are subjective, making them prone to intra- and inter-observer variability [15]. A lot of effort is put in significant feature engineering and scripting of hard written rules, and they are not robust when the data quality is poor [16].

With the advent of deep learning techniques, fully automatic methods could be developed for the process of semantic segmentation. Deep neural networks have been highly effective in extracting complex interpretative features from the underlying data for computer vision tasks. These features are leveraged to learn complex, meaningful representations in an end-to-end manner, making deep learning algorithms easily applicable to various tasks. Olaf Ronneberger et al. [25] introduced U-Net for bio-medical image segmentation. The major contribution of this model was the idea of skip connection and this model led to the development of many good performing models in segmentation. The skip connections helped to retain the information while generating the segmentation maps in the decoder part. Modifications in skip connections are introduced in various works such as [5, 26] to avoid the semantic gap between low and high resolution features. A. K. Sharma et al. introduced a cascaded ensembled network [27] for the dermatologist-level classification of skin cancer. This integrates handcrafted features (color moments and Gray Level Co-occurrence Matrix (GLCM) features) with the Convolutional Neural Network. S.S.Verma et al. proposed Cov-Xmlc [32] for the detection of COVID-19 from X-Ray images. This model effectively incorporated the classifier, SVM with the model VGG16.

The efficiency of deep learning segmentation models made them to surpass the previous state-of-the-art methods developed for cardiac segmentation. Smistad et al. [29] developed a method to segment left ventricle in 2D ultrasound images using U-Net CNN architecture. The network was, however, trained with the output of a deformable model segmentation method [28] due to lack of training data and obtained a dice score of 0.87. Oktay et al. [21] proposed a segmentation model to segment the 3D LV$_{endo}$ structure using an approach called anatomically constrained neural network (ACNN). The Challenge on Endocardial Three-dimensional Ultrasound Segmentation (CETUS) dataset was used to assess the performance of their method. The training phase, however, used only 15 patients and obtained a dice score of 0.912 (End Dyastolic (ED)) and 0.873 (End Systolic (ES)) on the testing set of 30 patients [21]. Kim T. et al. [13] presented segmentation of the left ventricle from echocardiographic images by using both U-Net and segAN. The segAN consists of a segmenter and a discriminator. The segmenter used here is the U-Net architecture. The problem with this is the discriminator output,final output, depends on the ground-truth image and the predicted image of the segmenter. Y. Yang et al. [36] introduced shape constraints into the U-Net architecture. They incorporated three levels of constraints: global level, regional level and pixel level with the U-Net architecture. Fei Liu et al. [17] proposed a deep pyramid local attention neural network (PLANet) which included a pyramid local attention module which enhances features by capturing supporting information. Recent research in semantic segmentation is focused on improving the semantic representation of the image by making full use of spatial and contextual features. Specialized modules have been introduced in convolutional neural networks such as atrous convolution, squeeze and excitation modules and pyramid pooling to improve the performance on semantic segmentation tasks. While recognizing the success of U-Net and its variants, this work meticulously investigate opportunities for development and propose a modified architecture to incorporate contemporary ideas for the task of echocardiogram segmentation. It has been observed that U-Net architecture may not be entirely robust when applied to tasks that require better extraction and richer representation of features [10, 41, 42]. This work shows that incorporating modern research ideas in computer vision tasks improves model performance and makes it more generalizable to different tasks and forms of data.

The major contributions the work are the following:

1. The proposed modified skip connection to increase the recognition ability of proposed TC-SegNet model.
2. The proposed TC-SegNet architecture mainly focuses on improving the spatial representation of features, capturing channel-wise relationships and bridging the semantic gap between the encoder and decoder while keeping in mind issues faced during training such as vanishing/exploding gradient and degradation problem.
3. The experimental results of proposed TC-SegNet model has been compared with recent benchmark models on publicly available and fully annotated CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) dataset.

This paper's organization is as follows: Section 2 discusses methodology of the proposed model. The experimental configuration is presented in Sections 3. The evaluation metrics are are presented in Section 4. The experimental results are explained in Section 5. Finally, the conclusion is given in Section 6.

2 Methodology

Having briefly discussed the evolution of the state-of-the-art image segmentation architectures and their pitfalls in heart chamber segmentation, we further dive deeper into the U-Net architecture and its variants to address these issues.

2.1 U-Net backbone

U-Net [25] has shown incredibly promising performance in the medical domain with limited training data. This success is attributed to the use of skip connections which combines low-level features from the encoder to the decoder. A naive intuition behind this is that passing extracted features from a lower level to the latter stages allows the transfer of information between the encoder and decoder stages. The passing of low level finer details can help to bridge the semantic gap for precise reconstruction of segmentation masks. In addition to the novel skip connections, employing data augmentation allows the model to be invariant to transformations in the training corpus. The U-Net network follows a symmetric architecture; the encoder breaks the input down to capture meaningful spatial patterns, while the decoder aims to reconstruct desired segmentation outputs. Repeated 3×3 convolution operations are performed on each encoder step, followed by a pooling layer. In each decoder step, input feature maps are upsampled and concatenated with feature maps from the corresponding encoder level via skip connections [18]. This helps to preserve the information lost due to pooling operations. The augmented features are then propagated to the successive layers.

2.2 Proposed architecture - TC-SegNet

The workflow of the proposed model is given in Fig. 1. The preprocessed images and ground-truth images are given to the network for training. For the input image, the model is predicting the result with initial random weights and calculating the loss with respect to the ground-truth images. The weights are given for updation in order to minimize the loss for the better prediction. Training the model will return optimized weights, and with these weughts, the model will be able to segment a new test image.

The proposed architecture is build on the U-Net architecture by reviewing contemporary research ideas and considering specific drawbacks of U-Net. The proposed architecture mainly focuses on improving the spatial representation of features, capturing channel-wise relationships, bridging the semantic gap between the encoder and decoder while keeping

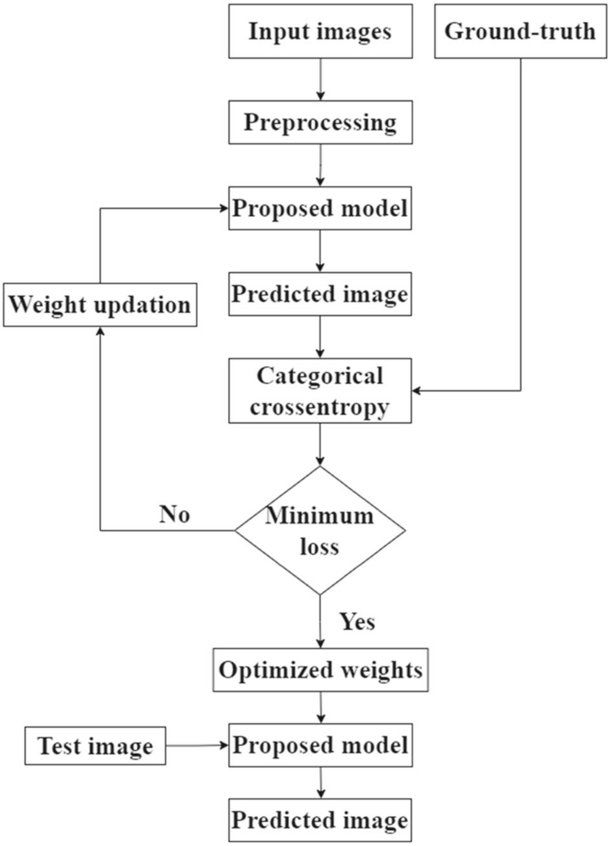

Fig. 1 Workflow of the proposed model

in mind issues faced during training such as vanishing/exploding gradient and degradation problem and reducing the computational complexity.

The network follows a U-Net backbone comprising of three modules: encoder, bridge and decoder. Figure 2 depicts the proposed architecture. The proposed model used a modified residual block inspired from [39] to each step of the encoder and decoder. The encoder consists of three residual blocks separated by the squeeze and excitation module and max

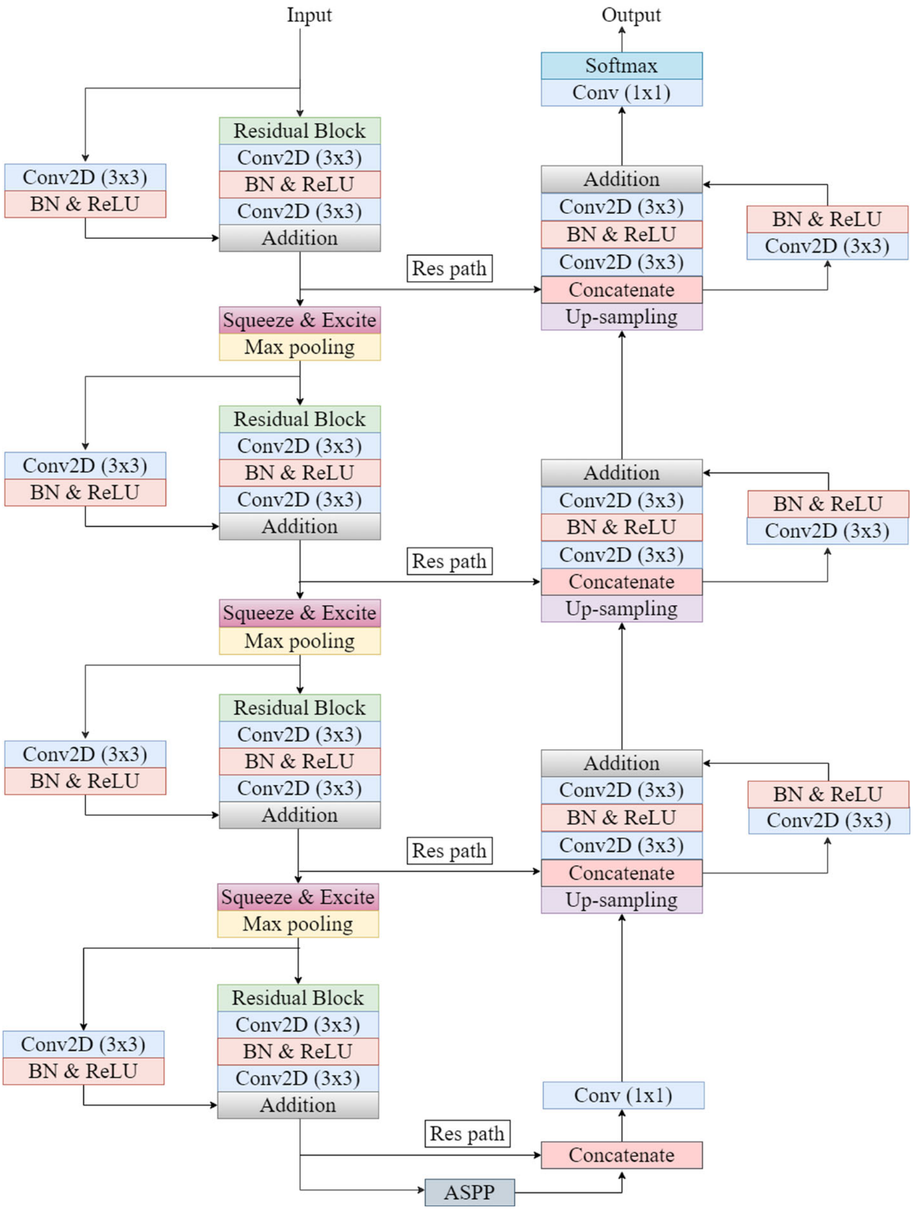

Fig. 2 Proposed TC-SegNet model

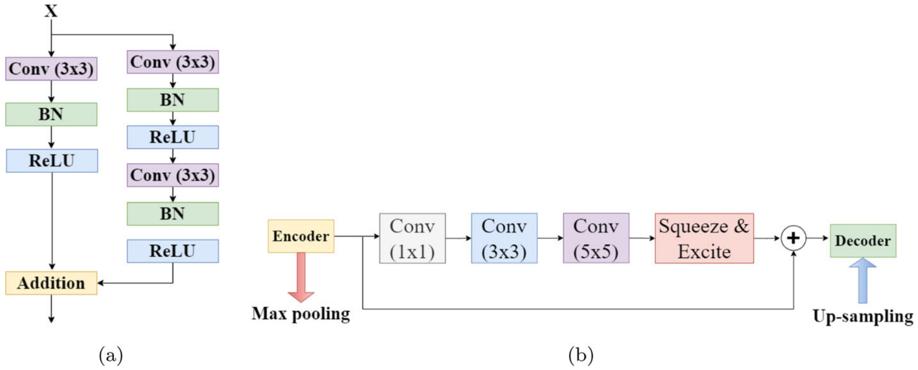

Fig. 3 (a) Residual block (b) Residual skip path (Res Path)

pooling. The residual block is shown in Fig. 3b. It performs two 3 × 3 convolution operations. It also consists of an identity mapping that connects the input and output of the residual unit. Before each residual unit in the decoding path, the input feature is up-sampled. A proposed modified skip connection from the corresponding encoder level concatenates the low-level features as in the U-Net architecture. A 1 × 1 convolution with softmax activation function is included to produce the desired segmentation masks.

The ASPP block acts as an information bottleneck between the encoder and decoder. This helps in the extraction of multiple scale features by expanding the convolutional layers' receptive field. Modified skip connections, inspired from [10] are incorporated between the corresponding levels of the encoder and decoder in the proposed model. At each successive step, the path size is decreasing by reducing the number of convolutional blocks along the skip connections from 4 to 3 and finally to 2. The modified skip path from the encoder to the decoder is depicted in Fig. 3b. The architecture is based on ideas from [12] which further modified and applied for the task of echocardiogram segmentation. In the following sections, each sub-module of the proposed network is analyzed in a greater depth.

2.2.1 Residual blocks

Studies have suggested that the depth of the network is crucial in performance, and state-of-the-art results are obtained by exploiting such "deep" architectures. Experiments had shown accuracy saturation and, moreover, rapid degradation with increased network depth due to the vanishing/exploding gradient problem [11]. He et al. [8] addressed this by introducing the residual learning framework.

The idea behind residual learning is that instead of approximating an underlying representation $H(x)$ using stacked layers, explicitly approximate a residual function $F(x)$: $H(x) - x$. This simple reformulation is based on the phenomenon of degradation. The degradation problem arose when the network faced a problem with approximating identity mappings by the stacked non-linear layers. Hence, if the hypothesis function is closer to the identity function, residual learning should ease the learning process. Zhang et al. [39] further combined the residual learning framework with a U-Net backbone for the task of Road Extraction. Let x_t be the input at the t^{th} step and x_{t+1} be the output at the same step, it is possible to represent the residual connection with the activation, g mathematically with equations (1) and (2).

$$y_t = h(x_t) + F(x_t, W_t) \tag{1}$$

$$x_{t+1} = g(y_t) \tag{2}$$

where F is the residual function, and the identity mapping function $h(x_t)$ passes x_t through a convolution, a batch normalization layer and ReLU activation layer.

The shortcut connection or skip connection of the residual block is actually taking the activations from one layer and adding to another layer which helps to sustain the learning parameters of the network in the proceeding layers. The residual unit not only eases training, the skip connections between the encoder and decoder stage will facilitate information transfer without degradation. This plays a vital role in complexity reduction enabling us to design a shallower architecture with lesser parameters.

2.2.2 Atrous spatial pyramid pooling

The existence of objects at multiple scales in an image has posed a significant challenge in image segmentation. It is required deep learning architectures to be scale-invariant. Traditionally this was overcome by aggregating feature maps of FCN at different scales of the same input. This approach improves performance; however, it comes at a computational expense.

RCNN [6] and PSPNet [40] have shown success using spatial pyramid pooling (SPP) [7]. Liang et al. (2018) further proposed a modified SPP using atrous convolutional layers: Atrous Spatial Pyramid Pooling [3], with different sampling rates to increase the receptive field of convolution filters. This improved model invariance to objects at different scales and made use of contextual information.

The SPP divides spatial information into bins by pooling features captured by filters of different sizes; this allows extraction and representation of features at varying levels of granularity and arbitrary scales. Atrous convolution [4] allows convolution kernels to make use of broader context by increasing the kernel size without increasing model parameters. Generally, conventional CNN architectures use small convolution filters to reduce computational and model complexity. However, small filters present a difficulty as features undergo several pooling and convolution operations, thereby reducing the feature resolution and making models sensitive to local image transforms. The ASPP helps to extract features at multiple scales by using different field-of-views and thus helps to segment different objects at various scales accurately.

The ASPP block in the proposed model consists of three parallel convolution layers, with dilation rates of 2, 4 and 8. The concatenation of outputs of each convolution layer is taken as the final output of the ASPP block.

2.2.3 Squeeze and excitation

The previous modules improved the spatial representation of features. The U-Net backbone is further modified to emphasize the channel relationship using the “Squeeze-and-Excitation” module proposed by Hu et al. [9].

A photographer chooses the best frame from all the available frames taken when capturing a single photograph, which depends on various factors like contrast, blur, distortion, etc. Abstractly, the photographer chooses the frame that perfectly captures the information conveyed in the photograph.

Analogously in CNN architectures, frames can be thought of as the channels in a feature map computed by a convolutional layer. The squeeze and excitation module applies a weighting function on the channels based on their importance; the notion is to render more "attention" to feature maps that capture relevant and essential features.

The Squeeze module decomposes spatial information of each feature map by a Global Average Pooling operation and squeezes it into a channel descriptor. The output of the squeeze module $X_s \in R^{1\times 1\times C}$ can be represented by the equation (3).

$$X_s^{c_i} = \frac{1}{H \times W}\left\{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{s_{in}}(h, w)\right\} \tag{3}$$

The Excitation module is a fully connected neural network with two single hidden layer. The excitation block is viewed as a self-attention mechanism on channels whose relationships cannot be captured by convolutional filters as they are not defined in a local receptive field. The processes included in the Excitation module can be represented by the Equations (4) to (6).

$$X_{d1} = \varphi\left\{F_d\left(X_s, W\right)\right\} \tag{4}$$

$$X_{d2} = \sigma\left\{F_d\left(X_{e1}, W\right)\right\} \tag{5}$$

$$X_{se} = X_{d2} \times X_{s_{in}} \tag{6}$$

In Equation (4) and (5), $F_d(.)$ represents the fully connected layer, W is the weight vector, φ is the ReLU activation function and σ is the Sigmoid activation function. In Equation (6), X_{se} represents the feature obtained from the Squeeze and Excitation module. In a nutshell, the Squeeze and Excitation block decomposes a feature map by Global Average Pooling and the aggregated information (capturing channel-wise dependencies) is fed into a fully connected network which returns a vector of weights. This amplifies the effect of relevant features on the network.

2.2.4 Proposed modified skip connections

The skip connections in U-Net ease the propagation of low-level features for the construction of segmentation masks, thus preserving and considering spatial features that might have diminished due to pooling operations. However, a semantic gap exists between the features at the lower levels of the encoder and corresponding feature maps at the decoder level that extracts much more complex features that undergo more processing. Nabil et al. [10] conjectured that a simple concatenation of such feature maps that are incomparable might not suffice for accurate prediction.

To reduce this imbalance in complexity of features at the encoder and decoder, three convolutional layers with kernels 1 × 1, 3 × 3 and 5 × 5 and the squeeze and excitation block are introduced along the skip connections. The output of the squeeze and excitation block is added with the original encoder feature. This is taken for the concatenation with the features at the decoder. The modified skip connection path is shown in Fig. 3b. The intention behind this is that adding more non-linear transformations on the more "callow" feature maps at the encoder should compensate for the high-level features at the decoder stage.

The three convolutions with different kernel sizes, which improves the feature extraction. A 5x5 kernel can extract more information than a 3x3 convolution. This helps to extract features at different levels. The squeeze and excitation block will give attention to the more relevant features extracted with these convolutions. Addition of this output with the original encoder output will help to retain the information extracted from the encoder stages and ease the training process.The basic block of modified skip connection can be represented as in Equations (7) and (8).

$$X_1 = F\left(X_{enc}, W_{1\times1}\right)\ ;\ X_2 = F\left(X_1, W_{3\times3}\right)\ ;\ X_3 = F\left(X_2, W_{5\times5}\right) \tag{7}$$

$$X_{skip} = F_{SE}\left(X_3\right) + X_{enc} \tag{8}$$

In Equation (7), $F(.)$ represents the convolution operation, X_{enc} is the encoder feature, $W_{1\times1}$, $W_{3\times3}$ and $W_{5\times5}$ are the filter kernels of point-wise, 3 × 3 and 5 × 5 convolutions respectively.In Equation (8), X_{skip} is the feature from the modified skip connection and $F_{SE}(.)$ represents the squeeze and excitation operation.

3 Experimental configuration

3.1 Dataset

The dataset used for the purpose of 2D echocardiographic assessment is the largest publicly available and fully annotated CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) dataset [15].

The CAMUS dataset consists of 500 patients' echocardiography images of both ED and ES with both two-chamber and four-chamber views. Half of the patients have an EF lower than 45% and are considered to be at a pathological risk [15]. 19% of the images are of poor quality. All images were resized to 256 x 256. Out of the 500 patients, 450 patients for whom the ground-truths were accessible are used for training, validation and testing purposes. Each patient has 2 images (ED and ES) for each of two-chamber and four-chamber views. Thus, a total of 1800 images (ES and ED combined) were used, out of which 1080 images were used for training, 360 were used for validation, and 360 images were used for testing. Figure 4 shows the image quality distribution (Good, Medium, Poor) in the train, validation and test datasets. The whole dataset is available for download at https://camus.creatis.insa-lyon.fr/challenge/.

On analyzing the data, it is found that the echocardiogram dataset is very noisy and is corrupted with speckle noise, so the images are preprocessed by wiener filtering, applying

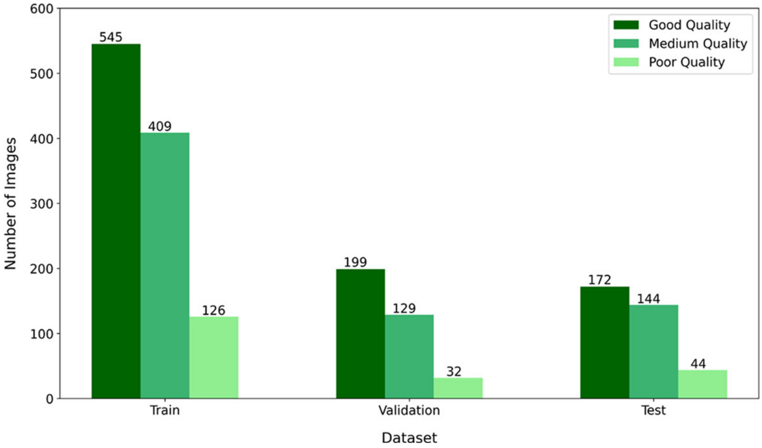

Fig. 4 Image quality distribution in train, validation and test datasets

morphological operations and anisotropic diffusion [19]. A 2D CS adaptive Wiener Filter [30] is employed for removing noise in the satellite images. It is based on the statistics computed from the local neighbourhood of each pixel and adjusts itself to the local image variance.

Next, two morphological techniques, Opening by reconstruction followed by Closing by reconstruction are used for further removal of speckle noise from the echocardiograms. For Opening by reconstruction [24], the Wiener filtered image is first eroded and then its morphological reconstruction is obtained by using the original filtered image as the mask. Opening by reconstruction restores precisely the size and shape of the objects that remain after eroding the image. Finally, Closing by reconstruction is done by complementing the image and computing its Opening by reconstruction. The final result is complemented again to obtain the clean image.

Anisotropic diffusion (Perona-Malik diffusion) [37] is a partial differential equation based method for smoothing and restoration of the clean image. This filter removes image noise without affecting significant parts of the image. For our dataset, we have used the Perona-Malik equation to perform the filtering. The Perona-Malik model reduces the diffusivity at locations which have a higher likelihood to be an edge by performing a non-linear diffusion.

The results of each step of the preprocessing procedure is shown in Fig. 5. To improve the generalization capability the model, data augmentation techniques like random flips and rotations, elastic transforms and contrast variation are also performed before training.

3.2 Training details

The proposed model is implemented on the NVIDIA Quadro RTX4000 GPU with 8GB onboard memory. The model is trained with a batch size of 4, for 100 epochs and use the Adam optimizer with an initial learning rate of 10^{-3}. The activation function used is ReLU between model layers while the final layer has a softmax activation. Batch normalization layers are also used for regularisation.

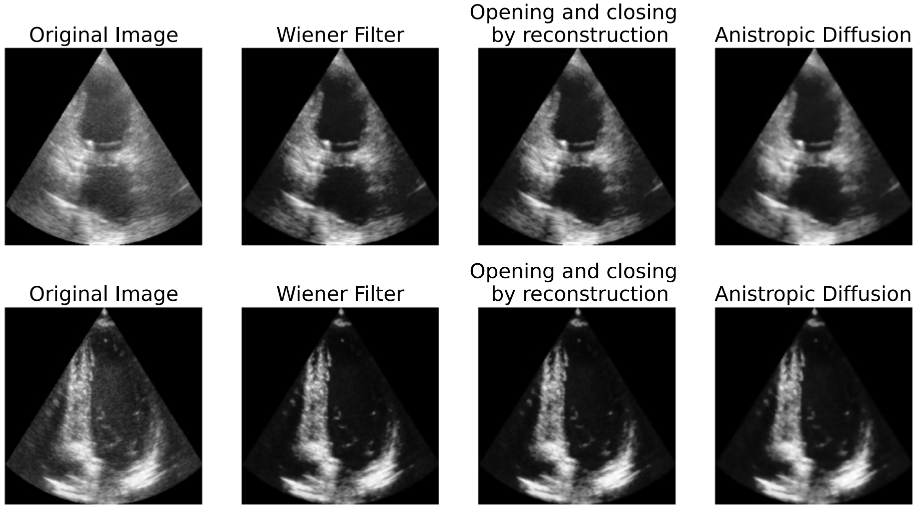

Fig. 5 Preprocessing steps

3.3 Training loss function

The loss function used is categorical cross entropy. The ground truth consists of 4 classes - Background, Left ventricle, Myocardium and Left atrium. For multi-class problem, the loss function is defined mathematically in Equation(8).

$$Loss = -\sum_{n=1}^{4} y_{n,\ true} \log(y_{n,\ pred}) \tag{9}$$

where $y_{n,\ true}$ is the true pixel class and $y_{n,\ pred}$ is the predicted pixel class.

4 Evaluation metrics

One of the most simplistic metrics to evaluate a segmentation model is Pixel Error (PE), which gives the percentage of misclassified pixels. This metric, however, will provide misleading results when the class representation is small within an image, as it will be more biased towards the negative class. This is why segmentation algorithms are generally evaluated using more robust metrics which measure the relative spatial overlap between the predictions and the ground truth. In this paper, the models are evaluated using Dice coefficient and Jaccard Index - a measure of the spatial overlap of images, F1-score and Pixel Error. In the following section the above mentioned metrics are described in further detail.

4.1 Pixel error

Pixel error is the ratio of misclassified pixels to the total pixels in the image for a given class and it is given in Equation(9).

$$\text{Pixel Error (PE)} = \frac{FP + FN}{TP + TN + FP + FN} \tag{10}$$

where,
TP (True Positive) represents a pixel correctly identified as belonging to the given class
TN (True Negative) represents a pixel correctly predicted as not belonging to the particular class
FP (False Positive) represents a pixel incorrectly identified as belonging to the given class
FN (False Negative) represents a pixel incorrectly predicted as not belonging to the given class.

The pixel error for each class and the global pixel error (excluding background class) are calculated.

4.2 Dice similarity coefficient

The Dice similarity coefficient (or Dice score) is a measure of the similarity between two sets. In general, for two sets X and Y, the Dice similarity coefficient is defined as in Equation (10).

$$\text{DSC}(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{11}$$

For image segmentation problems, the dice score can be calculated as Equation (11).

$$\mathrm{DSC}(f, x, y) = \frac{2 \times \sum_{i,j} f(x)_{ij} \times y_{ij} + \epsilon}{\sum_{i,j} f(x)_{ij} + \sum_{i,j} y_{ij} + \epsilon} \tag{12}$$

where the variables in the formula are,

- x : the input image
- $f(x)$: the predicted output
- y : the ground truth
- ϵ is a small number that is added to avoid division by zero

For the multi-class problem, the equation becomes:

$$DSC(f, x, y) = \frac{1}{4} \sum_{c=1}^{4} (DSC_c(f, x, y)) \tag{13}$$

In Equation (12), DSC_c is the Dice similarity coefficient of the c^{th} class.

4.3 IoU coefficient

The Intersection over Union, also known as the Jaccard index, is an evaluation metric to identify the overlap between the ground truth of segmentation and the model's predicted output. The IoU coefficient is the ratio of the pixels common to both the predicted and ground truth to the total pixels present in both the masks and it is given in Equation (13).

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{14}$$

The intersection (target ∩ prediction) consists of pixels found in both the ground truth and the mask predicted by the model and the union (target ∪ prediction) is comprised of all pixels found in either the predicted or ground truth. For multi-class segmentation, the global IoU score is computed by calculating the IoU scores for each class individually and then averaging the IoU scores over all the classes.

4.4 F1-score

F1 score is the harmonic mean of precision and recall. A good F1 score indicates that there are less number of false positives and false negatives. The F1-score can be calculated form the Equation (15).

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{15}$$

5 Ablation study

The proposed TC-SegNet is based on the U-Net architecture, on which different experimentations have been done for the development of the TC-SegNet. The model effectively utilizes the advantages of modified skip connection, residual blocks, SE block and ASPP. The effectiveness of each block is evaluated by detaching them from the proposed model and the obtained results are given in Table 1.

The proposed modified skip connections are replaced with the normal skip connection in the model and the results are given as variation-1. Variation-2, variation-3 and variation-4

Table 1 Overall quality comparison of five variants of proposed TC-SegNet

Model variations	Skip path	Residual block	ASPP	SE block	F1	Mean IoU	Dice	PE [‡](%)
Variation-1	×	✓	✓	✓	0.9033	0.8261	0.9232	1.5552
Variation-2	✓	×	✓	✓	0.9067	0.8271	0.9195	1.5463
Variation-3	✓	✓	×	✓	0.9067	0.8283	0.9246	1.5350
Variation-4	✓	✓	✓	×	0.9067	0.8298	0.9256	1.5333
Proposed TC-SegNet	✓	✓	✓	✓	**0.91**	**0.8322**	**0.9284**	**1.5109**

[‡] PE - Pixel Error

Bold entries signify the best results

are representing the models without residual blocks, without ASPP and without SE block respectively. Each variation leads to a reduction in the results and the proposed model, where contributions of all these blocks are involved, is giving the best result than the variations.

The variation-1 results in a reduction of 0.67% in F1-score, 0.61% in mean IoU and 0.52% in dice coefficient from the proposed model. Among all the variations, least value for all metrics, except dice coefficient (second least value) and more prediction error are given by the variation-1. The variation-2 gives the least dice coefficient value which is 0.9195 which is 0.89% less than that of the proposed model. The second least result is given by the variation-2 and it is giving the second highest score for prediction error. From the ablation study, it is possible to infer that the modified skip connection contributed more to the performance of the proposed model.

6 Results and discussion

6.1 Baseline considerations

The evaluations and inferences were performed by setting up the vanilla U-Net architecture as the baseline model. Every convolution block of the U-Net-32 encoder uses 32, 64, 128 and 256 number of filters respectively. The choice for filter setting is crucial for model size and performance. Using a higher number of filters allows us to extract more features and allows us to represent complex functions which would definitely aid model performance, but at the same time, increases the model complexity. This is the trade-off this work tried to optimize..

The proposed model is developed based on U-Net16 architecture and effectively added other efficient modules and modified skip connection which gives better performance than the existing models with less model complexity. Experimental results of proposed TC-SegNet model is compared with recent state-of-the-art models ASPPU-Net [33] and HMEDN [41] along with common variants of the U-Net architecture: Recurrent U-Net [35], Attention U-Net [22] and Residual U-Net [39].

6.2 Discussion

A partial ablation study is performed on the model to demonstrate the effect of suggested modifications on model accuracy. Vanilla U-Net architecture integrated with the specialized modules - Residual Block (Res-U-Net), Attention Block (Attention-U-Net) and ASPP Block (ASPP U-Net) and are evaluated. Table 2 depicts the results of the evaluated models.

Table 2 Results comparison of evaluated models

Architecture	Section	*F1*	*Dice*	*IoU*	*PE*[‡](%)
U-Net (2015) [25]	Endocardium Cavity	0.93	0.9438	0.8733	1.2623
	Myocardium Cavity	0.86	0.8792	0.7510	2.8065
	Left Atrium Cavity	0.90	0.9133	0.8141	1.0175
	Average	0.8967	0.9121	0.8128	1.6955
Attention U-Net (2018) [22]	Endocardium Cavity	0.93	0.9482	0.8692	1.2796
	Myocardium Cavity	0.86	0.8900	0.7513	2.9015
	Left Atrium Cavity	0.90	0.9222	0.8149	0.9998
	Average	0.8966	0.9201	0.8118	1.7270
Res-U-Net (2018) [39]	Endocardium Cavity	0.94	0.9452	0.8792	1.1958
	Myocardium Cavity	0.87	0.8865	0.7688	2.5769
	Left Atrium Cavity	0.90	0.9143	0.8222	0.9697
	Average	0.9033	0.9153	0.8234	1.5808
ASPPU-Net (2020)[33]	Endocardium Cavity	0.93	0.9461	0.8700	1.2822
	Myocardium Cavity	0.86	0.8862	0.7507	2.8624
	Left Atrium Cavity	0.89	0.9128	0.8054	1.0575
	Average	0.8933	0.9150	0.8087	1.7340
HMEDN (2020) [41]	Endocardium Cavity	0.93	0.9416	0.8691	1.3008
	Myocardium Cavity	0.86	0.8782	0.7517	2.7952
	Left Atrium Cavity	0.89	0.9041	0.8024	1.0871
	Average	0.8933	0.9079	0.8077	1.7278
segAN (2021) [39]	Endocardium Cavity	0.92	0.9339	0.8572	1.4319
	Myocardium Cavity	0.85	0.8740	0.7418	2.9756
	Left Atrium Cavity	0.75	0.7894	0.5989	2.4259
	Average	0.84	0.8658	0.7326	2.2778
Proposed TC-SegNet	Endocardium Cavity	**0.94**	**0.9559**	**0.8882**	**1.0987**
	Myocardium Cavity	**0.88**	**0.9003**	**0.7767**	**2.5135**
	Left Atrium Cavity	**0.91**	**0.9289**	**0.8317**	**0.9205**
	Average	**0.91**	**0.9284**	**0.8322**	**1.5109**

[‡]PE - Pixel Error

Bold entries signify the best results

The ResU-Net shows best result whereas segAN shows poor performance from the reference models. The proposed TC-SegNet is showing better performance compared to the reference models. The average values of F1-score, Dice coefficient, mean IoU and prediction error given by the proposed model are 0.91, 0.9284, 0.8322 and 1.5109 respectively. These values obtained from ResU-Net and segAN are (0.9033, 0.9153, 0.8234, 1.5808) and (0.84, 0.8658, 0.7326, 2.2778) respectively. The best Dice coefficient value from all

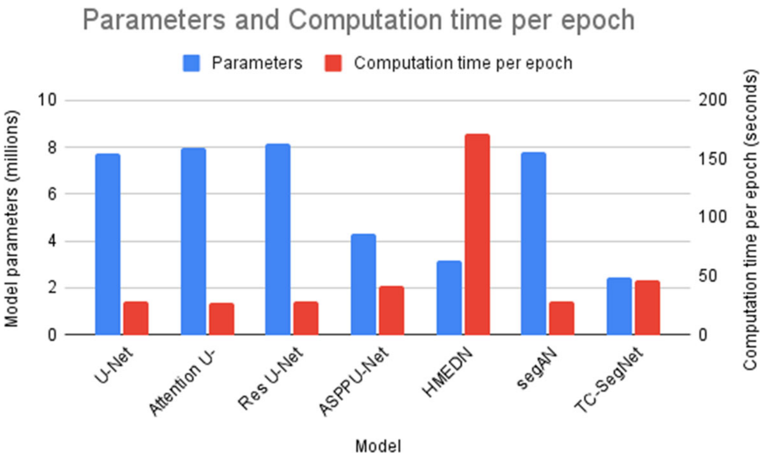

Fig. 6 Comparison of model parameters and computation time per epoch

reference models is given by the Attention U-Net and which is 0.9201 but other metric values are less than that of ResU-Net. The class-wise metric values are also higher for the proposed model. The class endocardium cavity is giving best values among all the three classes and least values are given by the class myocardium cavity for all models except segAN.

The complexity of a model can be defined using the number of parameters. The number of parameters and training time per epoch for all models are included in the Fig. 6. The proposed model required 2.41 million parameters and which is less than all other models used here. The ResU-Net requires more parameters which is 8.13 million. From the reference models, the least number of parameters is used by the model HMEDN and which is 3.14 million. The training time per epoch is also shown in Fig. 6. The training time per epoch is less for the model Attention U-Net which is 27 seconds and the proposed model is taking 46 seconds. This time is higher for the model HMEDN which is 171 seconds. The proposed model is giving best performance with less complexity.

Despite using a comparably smaller model, we account for this improvement in performance due to the various modifications to the model which resulted in extraction of relevant features and improved feature representation. This can be simplistically summarized as the quality of features over the quantity of features.

Previous studies [15] have evaluated the results on a test set consisting of only medium and good quality images. This results in data mismatch between the training and test distribution and the model performance on poor quality images is not completely assessed. The test set contains a significant fraction of poor quality images as seen in Fig. 4 to show that the proposed model is generalizable to poor quality data.

Now for further emphasizing the robustness and the ability of the proposed architecture to capture subtle nuances on various types of ultrasound images. Figure 7 shows, the predictions of evaluated models on sample two chamber and four chamber echocardiograms of varying quality. On clean and good quality images, the predictions made by other models are comparable to the proposed model to a certain degree. On a pixel level, however, the proposed model is observed to be more accurate. When evaluating on more challenging images with relatively poor quality, inconspicuous outlines and overlap of unwanted artefacts, the effectiveness of the proposed architecture is evident. Results produced by the

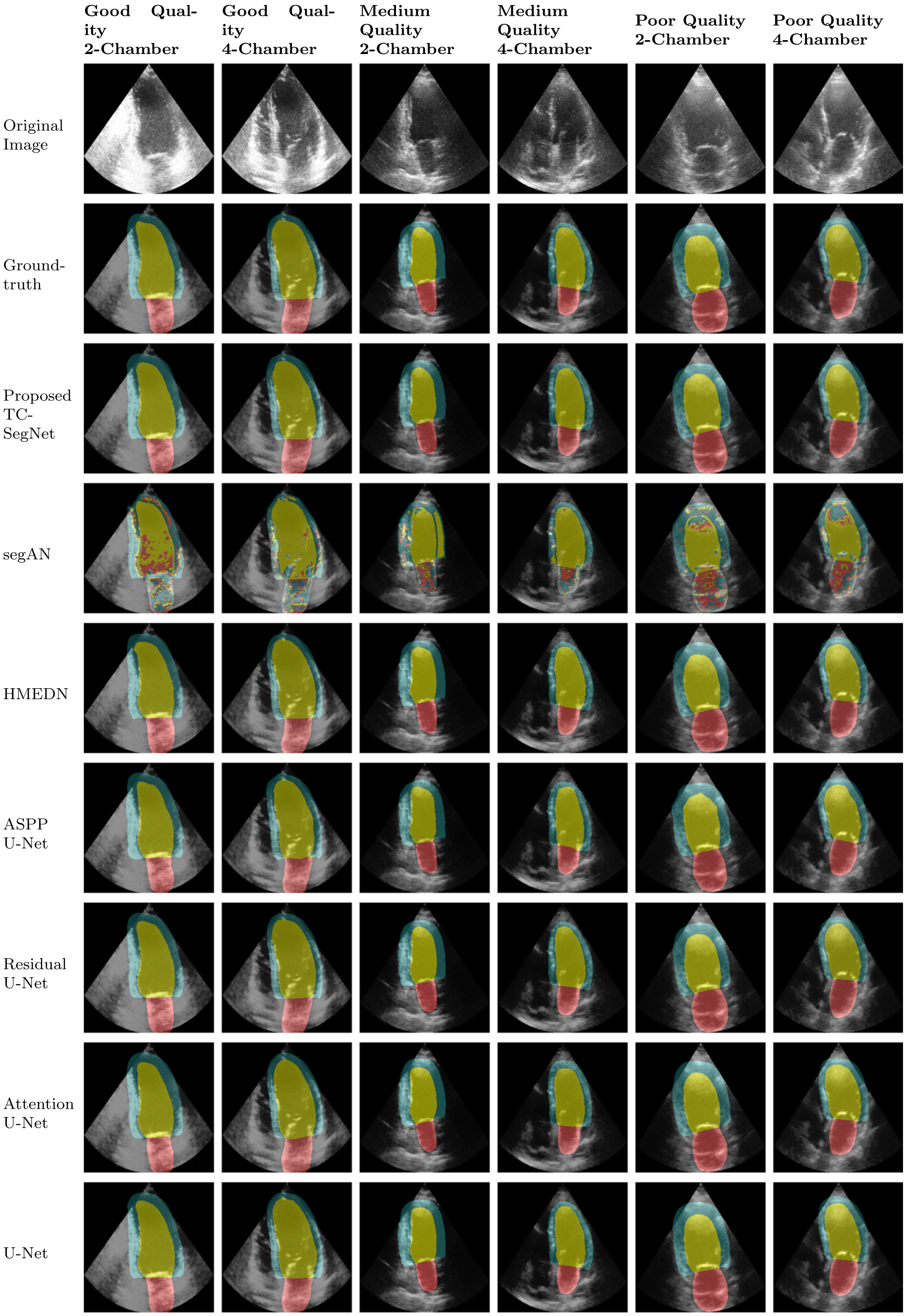

Fig. 7 Comparison of predictions made on 2-chamber and 4-chamber views of varying quality

proposed model proves to be robust to faint boundaries, variations in orientation and size, and perturbations in data.

7 Conclusion

This work proposed a robust encoder-decoder deep learning architecture: TC-SegNet for segmenting 2D echocardiographic images. The suggested architecture successfully overcomes certain drawbacks of the UNet model by improving feature representation using specialized modules and reducing the disparity between the encoder and decoder features. The proposed model significantly outperforms baseline models, especially when tested on low-quality images with perturbations and indefinite boundaries with fewer number of parameters. The proposed model is giving an F1-score of 0.91, dice coefficient of 0.9284, mean IoU of 0.8322 and prediction error of 1.5109%. The TC-SegNet is showing an improvement of 0.61% in F1-score, 1.31% in Dice coefficient, 0.88% in mean IoU and a decrease of 0.07% in prediction error compared to the best performing model, ResU-Net among the reference models. The TC-SegNet is less complex as the number of parameters is 2.41 millions and the ResU-Net has 8.136 millions. Thus the proposed model appears to be highly efficient in terms of the trade-off between the model performance and the number of parameters. Considering the wide range of image quality involved in the CAMUS dataset, the proposed model is observed to be robust to variability, especially to image quality. Further experiments involving hyperparameter optimization and evaluation on medical data obtained from various domains and modalities can improve the generalizability and model performance. Segmentation of class myocardium is better with the proposed model than the reference models, but still there is a scope of improvement. In the future, extraction of class specific features and extension of the dataset will be focused.

Acknowledgements Author would like to thank the reviewers and the Associate Editor, Editor-in-Chief for their valuable suggestions and feedback toward improving the quality of the manuscript. Author also would like to thank Sravya N, Varun Kumar M and Niran K Rao for helping during simulation and preparation of the the manuscript.

Declarations

Ethics approval This article does not contain any studies involving human participants and/or animals conducted by any of the authors.

Informed consent Not applicable

Conflict of Interests The author declares no competing interests.

References

1. Barbosa D, Friboulet D, Dhooge J, Bernard O (2014) Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching. In: Proceedings of the MICCAI challenge on endocardial three-dimensional ultrasound segmentation-CETUS, pp 17–24
2. Bernier M, Jodoin P-M, Lalande A (2014) Automatized evaluation of the left ventricular ejection fraction from echocardiographic images using graph cut. In: Proc. MICCAI Challenge Echocardiogr. Three Dimensional Ultrasound Segmentation (CETUS), pp 25–32
3. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587

4. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
5. Chen X, Zhang R, Yan P (2019) Feature fusion encoder decoder network for automatic liver lesion segmentation. In: Proc. IEEE International symposium on biomedical imaging (ISBI), pp 430–433
6. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on computer vision and pattern recognition, pp 580–587
7. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on computer vision and pattern recognition (CVPR), pp 770–778
9. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 7132–7141
10. Ibtehaz N, Rahman MS (2020) Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Netw 121:74–87
11. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on international conference on machine learning - volume 37 ICML'15. JMLR.org, pp 448–456
12. Jha D, Smedsrud P, Riegler M, Johansen D, de Lange T, Halvorsen P, Johansen H, Simulamet (2019) Resunet++: an advanced architecture for medical image segmentation
13. Kim T, Hedayat M, Vaitkus VV, Belohlavek M, Krishnamurthy V, Borazjani I (2021) Automatic segmentation of the left ventricle in echocardiographic images using convolutional neural networks, 1763–1781
14. Lang RM, Badano LP, Tsang W, Adams DH, Agricola E, Buck T, Faletra FF, Franke A, Hung J, Pérez de Isla L, Kamp O, Kasprzak JD, Lancellotti P, Marwick TH, McCulloch ML, Monaghan MJ, Nihoyannopoulos P, Pandian NG, Pellikka PA, Pepi M, Roberson DA, Shernan SK, Shirali GS, Sugeng L, Ten Cate FJ, Vannan MA, Zamorano JL, Zoghbi WA (2012) Eae/ase recommendations for image acquisition and display using three-dimensional echocardiography. J Am Soc Echocardiogr 25(1):3–46
15. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, Espeland T, Berg E, Jodoin P-M, Grenier T, Lartizien C, Drhooge J, Løvstakken L, Bernard O (2019) Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Trans Med Imaging PP:1–1
16. Lei T, Wang R, Wan Y, Zhang B, Meng H, Nandi AK (2020) Medical image segmentation using deep learning: a survey
17. Liu F, Wang K, Liu D, Yang X, Tian J (2021) Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography, 67: 101873
18. Mulliqi N (2020) The importance of skip connections in encoder-decoder architectures for colorectal polyp detection
19. Nakphu N, Dewi DEO, Rizqie MQ, Supriyanto E, Mohd Faudzi AA, Kho DCC, Kadiman S, Rittipravat P (2014) Apical four-chamber echocardiography segmentation using marker-controlled watershed segmentation. In: 2014 IEEE Conference on biomedical engineering and sciences (IECBES), pp 644–647
20. Oktay O, Shi W, Keraudren K, Caballero J, Rueckert D (2014) Learning shape representations for multi-atlas endocardium segmentation in 3d echo images. The MIDAS Journal - Challenge on Endocardial Three-dimensional Ultrasound Segmentation, 10
21. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, Cook SA, de Marvao A, Dawes T, O'Regan DP, Kainz B, Glocker B, Rueckert D (2018) Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE Trans Med Imaging 37(2):384–395
22. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention u-net: learning where to look for the pancreas
23. Pinto A, Pinto F, Faggian A, Rubini G, Caranci F, Macarini L, Genovese E, Brunese L (2013) Sources of error in emergency ultrasonography. Critical Ultrasound J 5 Suppl 1:S1
24. Robinson K, Whelan PF (2004) Efficient morphological reconstruction: a downhill filter. Pattern Recogn Lett 25(15):1759–1767
25. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation, 9351:234–241
26. Seo H, Huang C, Bassenne M, Xiao R, Xing L (2019) Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. IEEE Trans Med Imaging 39(5):1316–1325

27. Sharma AK, Tiwari S, Aggarwal G, Goenka N, Kumar A, Chakrabarti, Prasun, Chakrabarti T, Gono R, Leonowicz Z, Jasinski M (2022) Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network and handcrafted features based deep neural network, 10, 17920–17932
28. Smistad E, Lindseth F (2014) Real-time tracking of the left ventricle in 3d ultrasound using Kalman filter and mean value coordinates
29. Smistad E, stvik A, Haugen B, Olav, Lvstakken L (2017) 2d left ventricle segmentation using deep learning. In: 2017 IEEE International Ultrasonics Symposium (IUS), pp 1–4
30. Suresh S, Lal S (2017) Two-dimensional cs adaptive fir wiener filtering algorithm for the denoising of satellite images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing PP:1–13
31. van, Stralen M, Haak A, Leung K, Burken G, Bosch J, G (2014) Segmentation of multi-center 3d left ventricular echocardiograms by active appearance models. MIDAS, 73–80
32. Verma SS, Prasad A, Kumar A (2022) Covxmlc: high performance covid-19 detection on x-ray images using multi-model classification, biomedical signal processing and control, 71(Part B)
33. Wan T, Zhao L, Feng H, Li D, Tong C, Qin Z (2020) Robust nuclei segmentation in histopathology using asppu-net and boundary refinement. Neurocomputing, 408
34. Wang C, Smedby O (2014) Model-based left ventricle segmentation in 3d ultrasound using phase image, 10: 81–88
35. Wang W, Yu K, Hugonot J, Fua P, Salzmann M (2019) Recurrent u-net for resource-constrained segmentation. In: 2019 IEEE/CVF International conference on computer vision (ICCV), pp 2142–2151
36. Yang Y, Sermesant M (2021) Shape constraints in deep learning for robust 2d echocardiography analysis (hal-03371358)
37. Yang J, Guo Z, Zhang D, Wu B, Du S (2022) An anisotropic diffusion system with nonlinear time-delay structure tensor for image enhancement and segmentation. Comput Math Appl 107:29–44
38. Yodwut C, Weinert L, Klas B, Lang RM, Mor-Avi V (2012) Effects of frame rate on three-dimensional speckle-tracking-based measurements of myocardial deformation. J Am Soc Echocardiogr 25(9):978–985
39. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. IEEE Geosci Remote Sens Lett 15(5):749–753
40. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), pp 6230–6239
41. Zhou S, Nie D, Adeli E, Yin J, Lian J, Shen D (2020) High-resolution encoder decoder networks for low-contrast medical image segmentation. IEEE Trans Image Process 29:461–475
42. Zhou Z, Siddiquee MR, Tajbakhsh N, Liang J (2020) Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39:1856–1867

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.