Noname manuscript No.
(will be inserted by the editor)

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

Niranjan Ramesh Rao [1,2] · Varun Kumar M [1,2] · Shyam Lal [1,*]

the date of receipt and acceptance should be inserted later

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 a subject of intense research. This paper proposes a robust two chamber segmentation network (TC-SegNet) for echocardiography. Our proposed segmentation network follows a U-Net architecture with atrous spatial pyramid pooling (ASPP) modules, squeeze and excitation modules and modified skip connections to segment heart chambers. TC-SegNet is evaluated and tested on the open-source fully annotated CAMUS dataset for echocardiographic assessment. The model's results are evaluated using the Precision, Recall, Dice coefficient and Intersection of Union (IoU) score.Our proposed TC-SegNet obtained an average Dice score of 0.9265 and an IoU score of 0.809 . Further, average value of Hausdorff distance(HD) of 5.0359 and Pixel error (PE) of 5.0359. Segmentation results and metrics show that our 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 [1] and good temporal resolution [2]. However, interpretation of echocardiographic images are operator dependent and hence suffers from variability in image acquisition, subjective interpretation [3] 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

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

* Corresponding author
Email addresses: niranjanrao1509@gmail.com; varun.mkmr@gmail.com; shyam.mtec@gmail.com

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 [4], active appearance model [5], motion-based method [6], deformable models (BEAS, level-set) [7], [8] and graph-based methods [9]. 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 [10]. 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 [11].

With the advent of deep learning techniques, fully automatic methods could be developed, which has been observed to surpass the previous state-of-the-art methods developed for cardiac 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. Smistad et al. [12] 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 [6] due to lack of training data and obtained a dice score of 0.87. Oktay et al. [13] 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 (ED) and 0.873 (ES) on the testing set of 30 patients [13]. 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. In this work, while recognizing the success of U-Net and its variants, we 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 [14] [15] [16]. We show 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. We have introduced a 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 [17] 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 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 preserve information lost due to pooling operations. The augmented features are then propagated to the successive layers.

2.2 Proposed architecture - TC-SegNet

We build on the U-Net architecture by reviewing contemporary research ideas and considering specific drawbacks of U-Net. Our proposed 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.

The network follows a U-Net backbone comprising of three modules: encoder, bridge and decoder. Figure 1 depicts our proposed architecture. We introduce a modified residual block as proposed by Zhang et al. [19] to each step of the encoder and decoder. The encoder consists of three residual blocks separated by the squeeze and excitation module. The residual block is shown in Figure 2a. It performs two 3×3 convolution operations with the application of batch normalization and ReLU activation before each convolution. It also consists of an identity mapping that connects the input and output of the residual unit. In each encoder step, a strided convolution layer (stride 2) is used to downsample the feature map. Before each residual unit in the decoding path, there is an attention module [20] followed by up-sampling of feature maps. A modified skip connection from the corresponding encoder level concatenates the low-level features as in the U-Net architecture. The final decoder block is connected to the Atrous Spatial Pyramid Pooling (ASPP) block, followed by a 1×1 convolution to produce desired segmentation masks.

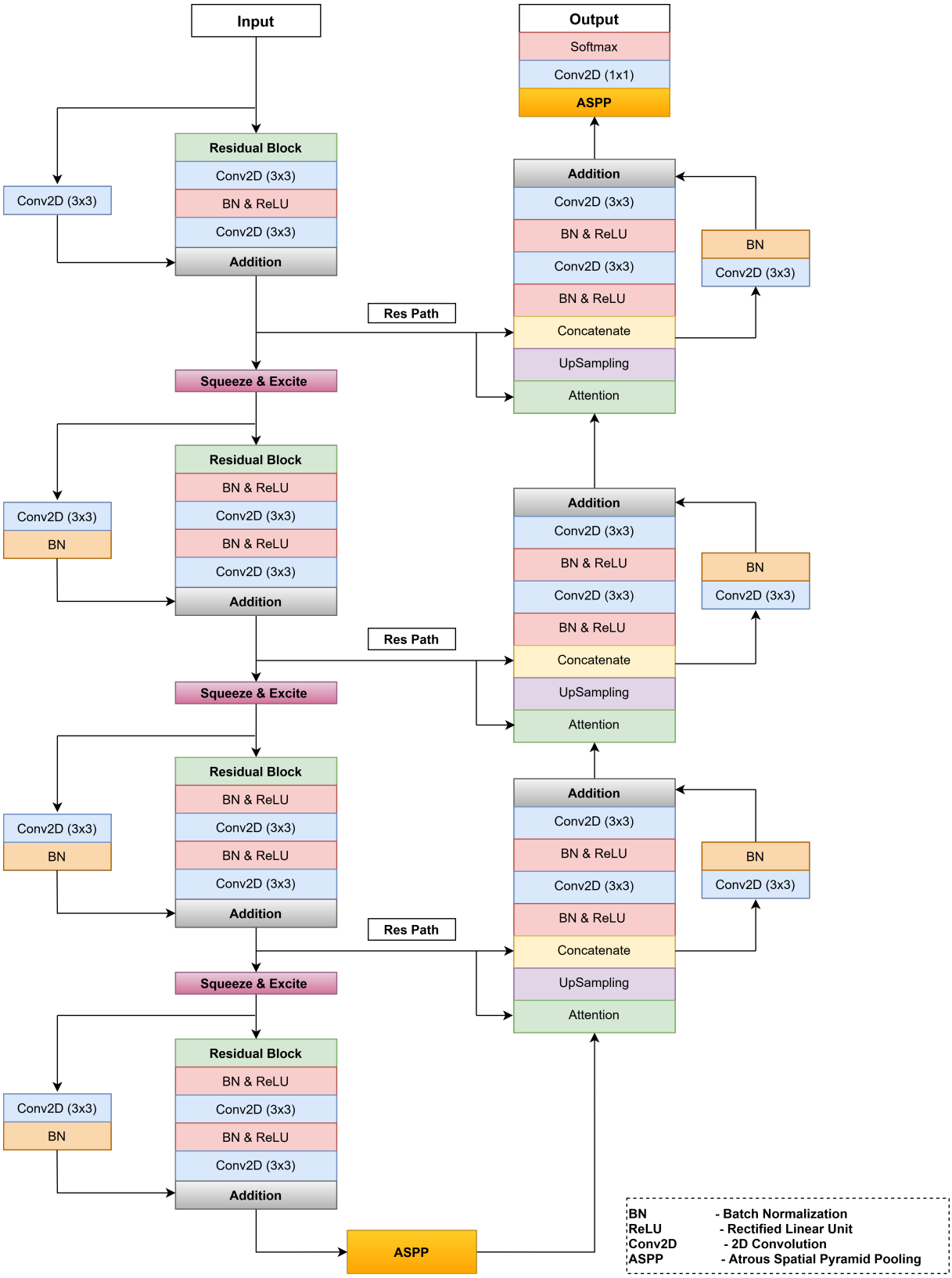

Fig. 1: Proposed TC-SegNet Model

The ASPP block acts as an information bottleneck between the encoder and decoder. This helps in the efficient extraction of features across multiple scales and expands the convolutional layers' receptive field.

We incorporate modified skip connections [14] between the corresponding levels of the encoder and decoder. Features from the encoder stage are convolved repeatedly before concatenation with the decoder. At each successive step, we reduce the path size 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 Figure 2b.

The architecture is based on ideas from [21] which we further modify and apply for the task of echocardiogram segmentation.

In the following sections, we analyze each sub-module of our network in greater depth.

2.2.1 Residual Blocks

Studies have suggested that the depth of the network is crucial for 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 [22]. He et al. (2016) 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, we 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. (2018) further combined the residual learning framework with a U-Net backbone for the task of Road Extraction. If we consider x_t to be the input at the t^{th} step and x_{t+1} to be the output at the same step, we can mathematically represent the residual connection with activation g as:

$$y_t = h(x_t) + F(x_t, W_t)$$
$$x_{t+1} = g(y_t)$$

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

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. We require 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 [23] and PSPNet [24] have shown success using spatial pyramid pooling (SPP) [25]. Liang et al. (2018) further proposed a modified SPP using atrous convolutional layers: Atrous Spatial Pyramid Pooling [26], 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.

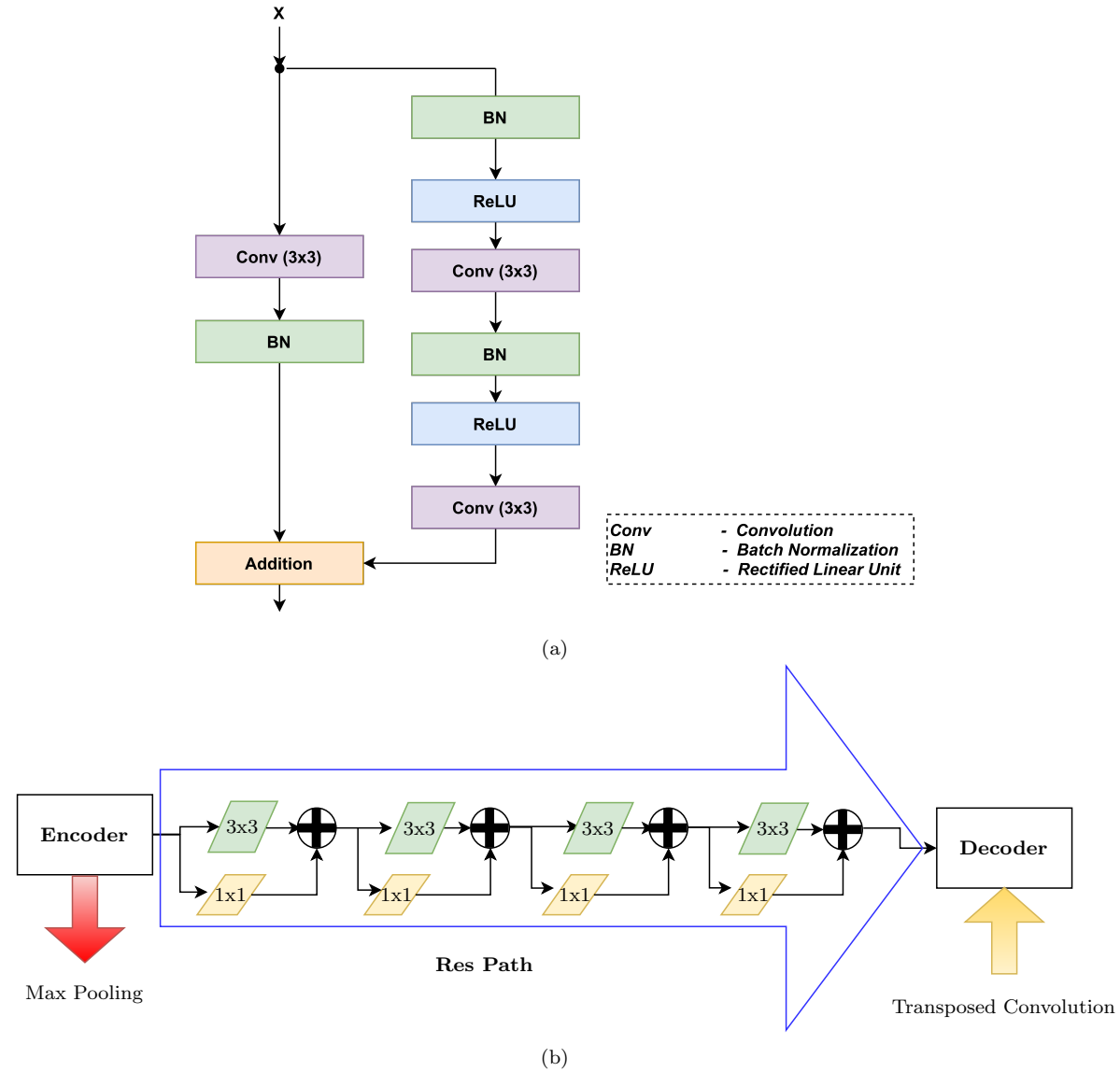

Fig. 2: (a) Residual Block (b) Residual Skip Path (Res Path)

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 [27] 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 block in our model consists of three parallel convolution blocks, with dilation rates of 6, 12, and 18, and batch normalization and ReLU activation. The output of each convolution block is concatenated and passed to a final convolutional layer.

2.2.3 Squeeze and Excitation

The previous modules improved the spatial representation of features. We further modify the U-Net backbone to emphasize the channel relationship using the "Squeeze-and-Excitation" module proposed by Hu et al. [28].

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 Excitation module is a fully connected neural network with a 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.

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 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. [14] 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, convolutional layers are introduced along the skip connections. The modified skip connection path is shown in Fig. 2b. 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. Furthermore, these convolutional layers are supplemented with residual connections to ease the training process.

3 Experimental Configuration

3.1 Dataset

We have used the largest publicly available and fully annotated CAMUS (Cardiac Acquisitions for Multi-structure Ultrasound Segmentation) dataset for the purpose of 2D echocardiographic assessment [10]. The CAMUS dataset consists of 500 patients' echocardiography images of both ED and ES views 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 [10]. 19% of the images are of poor quality. All images were resized to 256 x 256. Out of the 500 patients, we have used 450 patients for whom ground truths were accessible for training, validation and testing purposes. Each patient has 2 images (End Dyastolic and End Systolic phase) 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 3 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/.

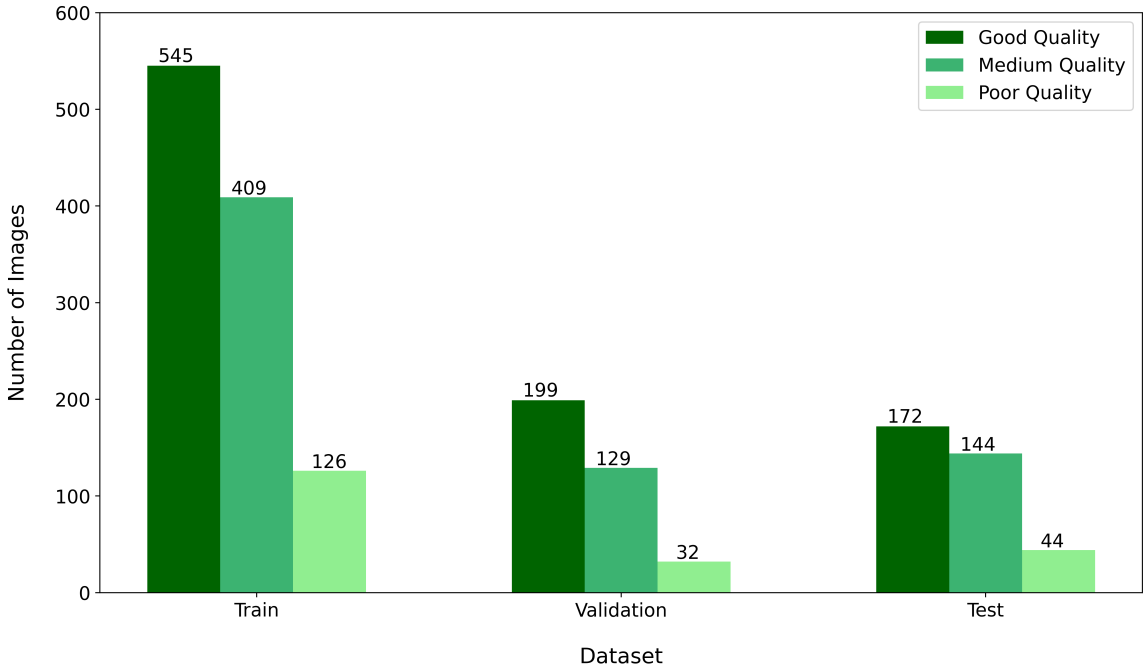

Fig. 3: Image Quality distribution in train, validation and test datasets

On analyzing the data, we find that the echocardiogram dataset is very noisy and is corrupted with speckle noise. We preprocess the images by wiener filtering, applying morphological operations and anisotropic diffusion [29]. We employed a 2D pixel-wise adaptive Wiener Filter [30] for removing

noise in the echocardiogram images. It is based on the statistics computed from the local neighbourhood of each pixel and adjusts itself to the local image variance.

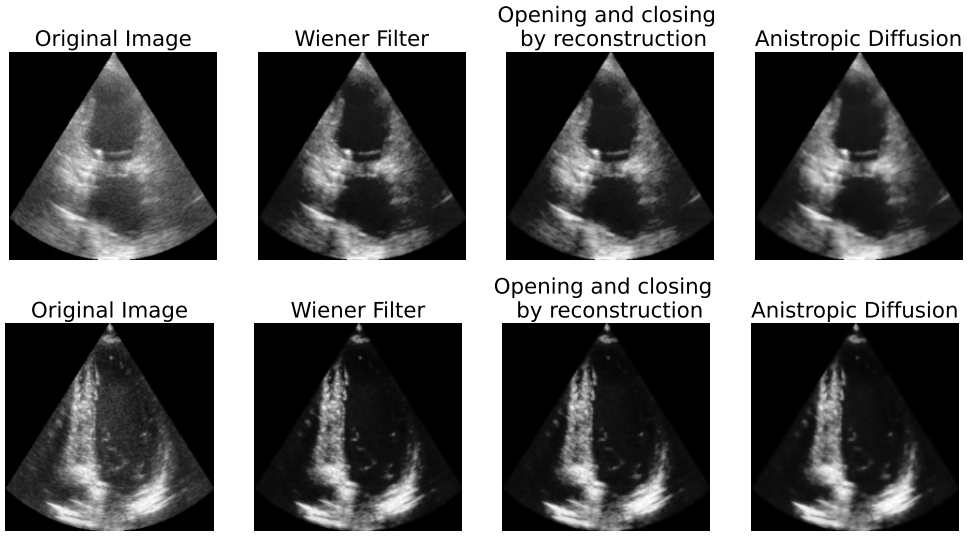

Fig. 4: Preprocessing steps

Next, we used two morphological techniques, Opening by reconstruction followed by Closing by reconstruction, for further removal of speckle noise from the echocardiograms. For Opening by reconstruction [31], 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) 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 [32] 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.

All the images are resized to 256 x 256 and we perform preprocessing for noise removal by using Wiener Filter followed by Morphological Operations and anisotropic diffusion to smoothen the final image. The results of each step of the preprocessing procedure is shown in Figure 4. To improve the generalization capability our model, we also perform data augmentation techniques like random flips and rotations, elastic transforms and contrast variation before training.

3.2 Training Details

Our model was trained on a cloud platform (Google Colab) using Tesla T4 GPU. We train the model with a batch size of 16, 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. We have also used batch normalization layers for regularisation.

3.3 Training Loss Function

We have used the categorical cross entropy as our loss function. The ground truth consists of 4 classes - Background, Left ventricle, Myocardium and Left atrium . For our multi-class problem, the loss function is defined as follows,

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

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. Additionally, geometric metrics are of specific interest to account for discrepancies between segmentation contours. In this paper, we evaluate our models using Dice coefficient and Jaccard Index - a measure of the spatial overlap of images, Hausdorff Distance - a measure of the geometric accuracy of segmentation boundaries 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.

$$\text{Pixel Error (PE)} = \frac{FP + FN}{TP + TN + FP + FN} \tag{2}$$

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.

We have calculated the pixel error for each class and the global pixel error (excluding background class).

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:

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

For image segmentation problems, the dice score can be calculated as follows:

$$\mathrm{DSC}(f, x, y) = \frac{2 \times \sum_{i,j} f(x)_{ij} \times y_{ij} + \epsilon}{\sum_{i,j} f(x)_{ij} + \sum_{i,j} y_{ij} + \epsilon} \tag{3}$$

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 our multi-class problem, the equation becomes:

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

where 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.

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{5}$$

The intersection (target $\cap$ prediction) consists of pixels found in both the ground truth and the mask predicted by the model and the union (target $\cup$ 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 Hausdorff Distance

2D Hausdorff Distance (HD) [33] is a geometric metric, used to measure the degree of resemblance between two objects that are superimposed on each other. Given two sets A and B, the Hausdorff distance is calculated as follows,

$$\mathrm{H(A,\ B)} = \max(h(A,B), h(B,A)) \tag{6}$$

where

$$h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{7}$$

and $\|\cdot\|$ is some underlying norm like L_2 norm on the points of set A and B. We evaluated the Hausdorff distance between shapes defined by the contours of predicted masks and ground truth, for each of LV_{endo}, LV_{epi} and LA regions.

5 Results and Discussion

5.1 Baseline Considerations

We perform evaluations and inferences 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 a trade-off we want to optimize.

We find that the larger model, U-Net-32, performs better when compared to its counterpart U-Net-16. The U-Net-32 baseline is chosen to demonstrate the improvement in accuracy of the proposed model, while reducing the model size.

Additionally, experimental results of proposed TC-SegNet model is compared with recent state-of-the-art models ASPPU-Net [35] and HMEDN [15] along with common variants of the U-Net architecture: Recurrent U-Net [34], Attention U-Net [20] and Residual U-Net [19].

5.2 Discussion

A partial ablation study is performed on the model to demonstrate the effect of suggested modifications on model accuracy. We evaluate 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). Table 1 depicts the results of evaluated models. The per class and average segmentation scores of the proposed TC-SegNet model is clearly the best among the evaluated models. An equally significant feat along with the model performance is the model complexity. We find that the model trained with filter numbers 16, 32, 64 and 128, outperforms the baseline U-Net-32, while simultaneously reducing the number of trainable parameters. Figure 5 depicts a clear comparison of number of parameters and computation time of each evaluated model. The average computation time per epoch for our proposed model (Dice score: 0.93) is 78 seconds which is greater than second best model, Res U-Net (Dice Score: 0.90), by 9 seconds and less than the latest model, HMEDN (Dice score: 0.88), by 50 seconds.

Table 1: Results comparison of evaluated models

Architecture	**Section**	***Dice***	***IoU***	***HD*†(mm)**	***PE*‡(%)**
U-Net [17]	Endocardium Cavity	0.8921	0.8428	5.6637	6.3520
	Myocardium Cavity	0.8631	0.8194	6.3010	10.0913
	Left Atrium Cavity	0.8845	0.8251	5.4407	8.7118
	Average	0.8799	0.8291	5.8011	8.3850
Recurrent-U-Net [34]	Endocardium Cavity	0.8879	0.8140	6.1061	6.2852
	Myocardium Cavity	0.8623	0.7982	6.2995	9.1001
	Left Atrium Cavity	0.8660	0.7977	6.6244	8.5733
	Average	0.8721	0.8033	6.3433	7.9862
Attention U-Net [20]	Endocardium Cavity	0.8898	0.8434	5.4507	7.9177
	Myocardium Cavity	0.8683	0.7662	6.6899	10.8901
	Left Atrium Cavity	0.8912	0.8141	5.7133	8.0674
	Average	0.8831	0.8079	5.9513	8.9584
Res-U-Net [19]	Endocardium Cavity	0.9124	0.8440	5.3044	5.2419
	Myocardium Cavity	0.8893	0.8161	6.3145	8.8021
	Left Atrium Cavity	0.9037	0.8302	5.6352	6.5090
	Average	0.9018	0.8301	5.7514	6.8511
ASPPU-Net (2020) [35]	Endocardium Cavity	0.9007	0.8232	7.0438	7.9171
	Myocardium Cavity	0.8138	0.6914	9.7880	10.8903
	Left Atrium Cavity	0.8534	0.7621	8.2386	8.0619
	Average	0.8559	0.7589	8.3568	8.9564
HMEDN (2020) [15]	Endocardium Cavity	0.9185	0.8517	5.7184	5.9100
	Myocardium Cavity	0.8522	0.7449	6.6023	9.8959
	Left Atrium Cavity	0.8728	0.7896	7.3265	8.0891
	Average	0.8811	0.7954	6.5491	7.9650
Proposed TC-SegNet	Endocardium Cavity	**0.9512**	**0.8535**	**4.3509**	**3.7011**
	Myocardium Cavity	**0.9320**	**0.8652**	**5.6224**	**7.5929**
	Left Atrium Cavity	**0.8963**	**0.8340**	**5.1344**	**6.2349**
	Average	**0.9265**	**0.8509**	**5.0359**	**5.8420**

†HD - Hausdorff Distance. ‡PE - Pixel Error.

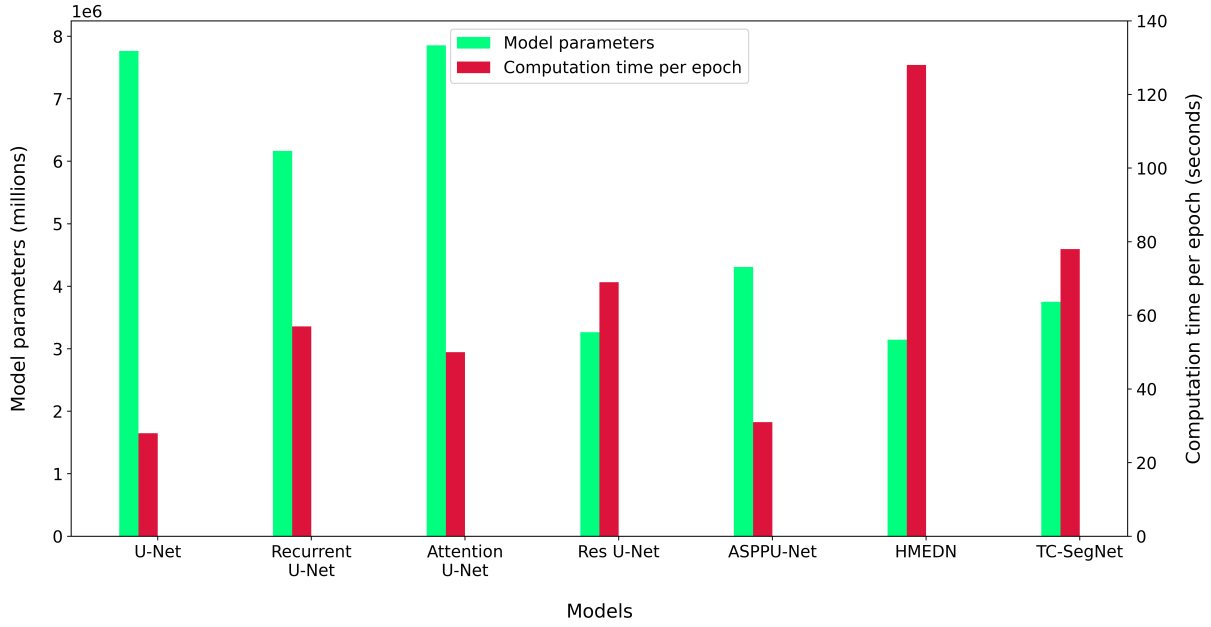

Fig. 5: Comparison of model parameters and computation time per epoch

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 [10] 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. Our test set contains a significant fraction of poor quality images as seen in Figure 3 to show our model is generalizable to poor quality data.

Now we further emphasize the robustness and the ability of the proposed architecture to capture subtle nuances on various types of ultrasound images. In Figure 6, we compare the predictions of evaluated models on sample two chamber and four chamber echocardiograms of varying quality. We see that on clean and good quality images, the predictions made by other models are to a certain degree comparable to the proposed model. On a pixel level, however, our 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 our architecture is evident. Results produced by the proposed model proves to be robust to faint boundaries, variations in orientation and size, and perturbations in data.

6 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

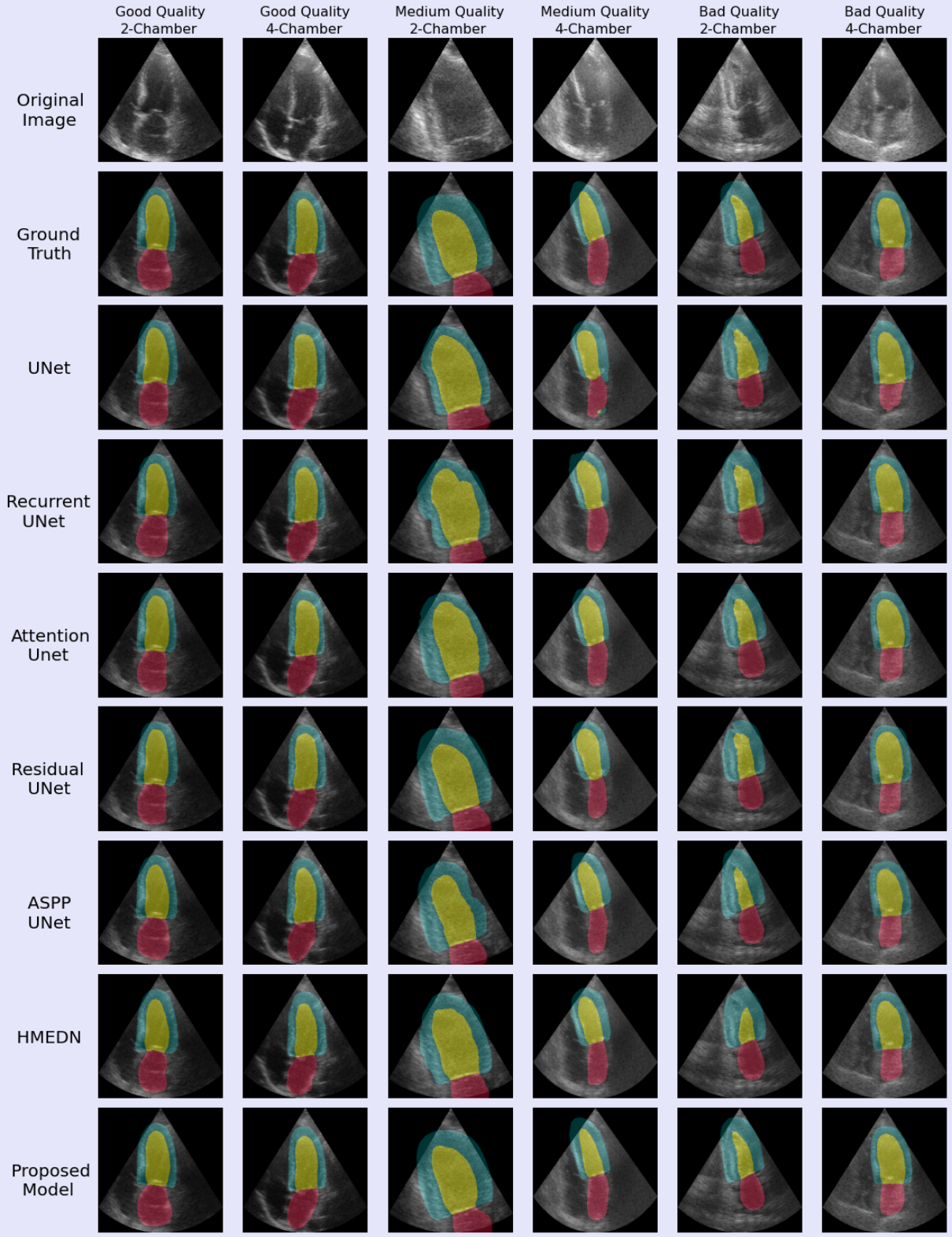

Fig. 6: Comparison of predictions made on 2-Chamber and 4-Chamber views of varying quality

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. Along with a higher Dice score, the segmentation outputs of TC-SegNet more alike to the ground truth segmentation masks. Our 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, our model is observed to be robust to variability, especially to image quality. We believe further experiments involving hyperparameter optimization and evaluation on medical data obtained from various domains and modalities can improve the generalizability and model performance.

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 interest The authors declare no competing interests.

References

[1] C. Yodwut et al. "Effects of frame rate on three-dimensional speckle-tracking-based measurements of myocardial deformation". In: *J Am Soc Echocardiogr* 25.9 (Sept. 2012), pp. 978–985. DOI: https://10.1016/j.echo.2012.06.001.

[2] Roberto M. Lang et al. "EAE/ASE Recommendations for Image Acquisition and Display Using Three-Dimensional Echocardiography". In: *Journal of the American Society of Echocardiography* 25.1 (2012), pp. 3–46. ISSN: 0894-7317. DOI: https://doi.org/10.1016/j.echo.2011.11.010.

[3] António Pinto et al. "Sources of error in emergency ultrasonography". In: *Critical ultrasound journal* 5 Suppl 1 (July 2013), S1. DOI: https://doi.org/10.1186/2036-7902-5-S1-S1.

[4] O. Oktay et al. "Learning Shape Representations for Multi-Atlas Endocardium Segmentation in 3D Echo Images". In: *The MIDAS Journal - Challenge on Endocardial Three-dimensional Ultrasound Segmentation* (Oct. 2014). DOI: https://doi.org/10.13140/2.1.3767.5522.

[5] Marijn van Stralen et al. "Segmentation of Multi-Center 3D Left Ventricular Echocardiograms by Active Appearance Models". In: *MIDAS* (Oct. 2014), pp. 73–80.

[6] Erik Smistad and Frank Lindseth. "Real-time Tracking of the Left Ventricle in 3D Ultrasound Using Kalman Filter and Mean Value Coordinates". In: Sept. 2014. DOI: 10.13140/2.1.1330.6888.

[7] Daniel Barbosa et al. "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* (Jan. 2014), pp. 17–24.

[8] C. Wang and O. Smedby. "Model-based left ventricle segmentation in 3D ultrasound using phase image". In: Oct. 2014, pp. 81–88.

[9] Michael Bernier, Pierre-Marc Jodoin, and Alain Lalande. "Automatized Evaluation of the Left Ventricular Ejection Fraction from Echocardiographic Images Using Graph Cut". In: *Proc. MIC-*

CAI Challenge Echocardiogr. Three Dimensional Ultrasound Segmentation (CETUS) (Sept. 2014), pp. 25–32.

[10] Sarah Leclerc et al. "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography". In: *IEEE Transactions on Medical Imaging* PP (Feb. 2019), pp. 1–1. DOI: https://doi.org/10.1109/TMI.2019.2900516.

[11] Tao Lei et al. "Medical Image Segmentation Using Deep Learning: A Survey". In: (2020). arXiv: 2009.13120 [eess.IV].

[12] Erik Smistad et al. "2D left ventricle segmentation using deep learning". In: *2017 IEEE International Ultrasonics Symposium (IUS)*. 2017, pp. 1–4. DOI: 10.1109/ULTSYM.2017.8092573.

[13] Ozan Oktay et al. "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation". In: *IEEE Transactions on Medical Imaging* 37.2 (2018), pp. 384–395. DOI: 10.1109/TMI.2017.2743464.

[14] Nabil Ibtehaz and M. Sohel Rahman. "MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation". In: *Neural Networks* 121 (2020), pp. 74–87. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.08.025.

[15] Sihang Zhou et al. "High-Resolution Encoder–Decoder Networks for Low-Contrast Medical Image Segmentation". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 461–475. DOI: 10.1109/TIP.2019.2919937.

[16] Zongwei Zhou et al. "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation". In: *IEEE Transactions on Medical Imaging* 39 (2020), pp. 1856–1867.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: vol. 9351. Oct. 2015, pp. 234–241. ISBN: 978-3-319-24573-7. DOI: https://doi.org/10.1007/978-3-319-24574-4_28.

[18] Nita Mulliqi. "The Importance Of Skip Connections In Encoder-Decoder Architectures For Colorectal Polyp Detection". In: Sept. 2020. DOI: https://doi.org/10.1109/ICIP40778.2020.9191310.

[19] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. "Road Extraction by Deep Residual U-Net". In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753. DOI: https://doi.org/10.1109/LGRS.2018.2802944.

[20] Ozan Oktay et al. *Attention U-Net: Learning Where to Look for the Pancreas*. 2018. arXiv: 1804.03999 [cs.CV].

[21] Debesh Jha et al. "ResUNet++: An Advanced Architecture for Medical Image Segmentation". In: Dec. 2019. DOI: https://doi.org/10.1109/ISM46123.2019.00049.

[22] Sergey Ioffe and Christian Szegedy. "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. Lille, France: JMLR.org, 2015, pp. 448–456. DOI: https://doi.org/10.5555/3045118.3045167.

[23] Ross Girshick et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587. DOI: https://doi.org/10.1109/CVPR.2014.81.

[24] Hengshuang Zhao et al. "Pyramid Scene Parsing Network". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6230–6239. DOI: https://doi.org/10.1109/CVPR.2017.660.

[25] Kaiming He et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2015), pp. 1904–1916. DOI: https://doi.org/10.1109/TPAMI.2015.2389824.

[26] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[27] Liang-Chieh Chen et al. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2018), pp. 834–848. DOI: https://doi.org/10.1109/TPAMI.2017.2699184.

[28] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132–7141. DOI: https://doi.org/10.1109/CVPR.2018.00745.

[29] Nonthaporn Nakphu et al. "Apical four-chamber echocardiography segmentation using Marker-controlled Watershed segmentation". In: *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES)*. 2014, pp. 644–647. DOI: https://doi.org/10.1109/IECBES.2014.7047583.

[30] Jae S. Lim. *Two-Dimensional Signal and Image Processing*. USA: Prentice-Hall, Inc., 1990. ISBN: 0139353224.

[31] Kevin Robinson and Paul F. Whelan. "Efficient morphological reconstruction: a downhill filter". In: *Pattern Recognition Letters* 25.15 (2004), pp. 1759–1767. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2004.07.002.

[32] Joachim Weickert. "Image smoothing and restoration by PDEs". In: *Anisotropic diffusion in image processing*. BG Teubner Stuttgart, 1998, pp. 1–53.

[33] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. "Comparing images using the Hausdorff distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9 (1993), pp. 850–863. DOI: 10.1109/34.232073.

[34] Wei Wang et al. "Recurrent U-Net for Resource-Constrained Segmentation". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 2142–2151. DOI: https://doi.org/10.1109/ICCV.2019.00223.

[35] Tao Wan et al. "Robust nuclei segmentation in histopathology using ASPPU-Net and boundary refinement". In: *Neurocomputing* 408 (Mar. 2020). DOI: 10.1016/j.neucom.2019.08.103.