Automatic regional image quality scoring for echocardiography

Gilles Van De Vyver,[a,*], Svein-Erik Måsøy[a], Håvard Dalen[a,c], Bjørnar Leangen Grenne[a,c], Espen Holte[a,c], Sindre Hellum Olaisen[a], John Nyberg[a], Andreas Østvik[a,b], Lasse Løvstakken[a], and Erik Smistad[a,b]

[a]*Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway*
[b]*Health Research, SINTEF, Trondheim, Norway*
[c]*St. Olavs hospital hospital, Trondheim, Norway*

Abstract

Objective

To develop and compare methods automatically estimate the regional ultrasound image quality for echocardiography separate from view correctness.

Methods

Three methods for estimating image quality were developed:

1. **Classic Pixel-based Metric**: The generalized contrast-to-noise ratio (gCNR), computed on myocardial segments (region of interest) and left ventricle lumen (background), extracted by a U-Net segmentation model.

*Corresponding Author: Gilles Van De Vyver, NTNU, Norwegian University of Science and Technology, 7491 Trondheim, Norway. E-mail: gilles.van.de.vyver@ntnu.no

2. **Local Image Coherence**: The average local coherence as predicted by a U-Net model that predicts image coherence from B-Mode ultrasound images at the pixel level.
3. **Deep Convolutional Network**: An end-to-end deep learning model that predicts the quality of each region in the image directly.

These methods were evaluated against manual regional quality annotations provided by three experienced cardiologists.

Results

The results indicate poor performance of the gCNR metric, with Spearman correlation to the annotations of $\rho = 0.24$. The end-to-end learning model obtains the best result, $\rho = 0.69$, comparable to the inter-observer correlation, $\rho = 0.63$. Finally, the coherence-based method, with $\rho = 0.58$, outperformed the classical metrics and is more generic than the end-to-end approach.

Conclusion

The deep convolutional network provides the most accurate regional quality prediction, while the coherence-based method offers a more generalizable solution. The generalized contrast-to-noise ratio (gCNR) showed limited effectiveness in this study. The image quality prediction tool is available as an open-source Python library at `https://github.com/GillesVanDeVyver/arqee`.

Keywords: Cardiac segmentation, Ultrasound, image quality, Coherence, Signal-To-Noise Ratio

Introduction

Image quality is one of the main challenges in ultrasound imaging and can differ significantly between patients and imaging equipment. In echocardiography, many factors influence image quality such as the ultrasound scanner, the patient, and the probe. Several quantitative measurements using the images are performed. However, this requires image quality good enough for the given measurement. Different measurements have different image quality requirements, for instance, left ventricular (LV) volume, ejection fraction (EF), and strain measurements require good image quality in the entire myocardium. On the other hand, mitral annular plane systolic excursion (MAPSE) only requires good image quality in the annulus. We also believe it is important to estimate regional image quality for each frame. For instance, end-diastole and end-systole frames are used for ejection fraction measurements, thus good image quality is required for both those frames. For measurements such as strain, which uses frame-by-frame tracking, the image quality of each segment of every frame is important as low quality in some frames may ruin the tracking in that segment. Good image quality should generally provide measurement values with low uncertainty. Estimating image quality can be useful in the following ways:

- To guide operators to achieve as good image quality as possible while scanning.
- To automatically select the best images, recordings, and the best cardiac cycles to use for a given measurement.
- As quality assurance, e.g. to warn the user when an image is not good

enough for a measurement, and to automatically approve/disapprove individual myocardial segments based on image quality.

- In data mining projects, to exclude cases with insufficient quality for reliable measurements.

Image quality is subjective and can vary depending on the application or the specific measurement being performed. What qualifies as ”good” or ”acceptable” depends on the context. In this study, we use the definitions outlined in Table 1.

We distinguish between two types of quality of ultrasound images: view quality/correctness and image quality. In this work, we will focus on image quality specifically. For view correctness, previous work has demonstrated that 3D ultrasound can serve as training data to automatically identify the transducer rotation and tilt in relation to the desired standard view and can guide the user to the correct position [1, 2, 3]. For image quality, the classic ultrasound signal-processing metrics are the contrast ratio (CR)[4], contrast-to-noise ratio (CNR)[5], and generalized CNR (gCNR)[6]. These three metrics need a region of interest (ROI) and a background region to compare against. More recently, global image coherence (GIC)[7] has been proposed as a general quality metric that does not require the selection of these two regions. The image coherence measures how well the signals of the transducer elements align after delay compensation, with more alignment corresponding to clearer and sharper images. From the above mentioned methods, only the GIC can be used directly and automatically for measuring image quality separately as it does not require selecting an ROI and noise region. However, this approach requires channel data, which is not readily available in practice and does not give regional metrics.

Several automatic methods for measuring ultrasound image quality have been published applicable to cardiac imaging. Abdi et al.[8] used a recurrent neural network to predict the global quality of cardiac cine loops. The criteria for quality assessment take both image quality and view correctness into account. In subsequent studies[9, 10], they used an architecture that performs both view classification and global quality prediction simultaneously. The image quality metric is a global criterion based on the manual judgment of the clarity of the blood-tissue interface. Labs et al.[11] used a multi-stream neural network architecture where each stream takes in a sequence of frames and predicts a specific quality criterion. The criteria are global and take both view correctness and image quality into account. Karamalis et al.[12] detect attenuated shadow regions with random walks resulting in a pixel-level confidence map. Unlike the other methods above, this method is not based on deep learning. It provides a local, pixel-level metric, but it only measures the visibility of regions and not the quality of their content.

All of the automatic methods mentioned above have the limitation that they only provide a global image quality evaluation and/or do not assess the image quality separate from the view correctness. The novelty of this work is automatic quality estimation on the regional level. To the best of our knowledge, this is the first study to propose a method for automatic regional image quality assessment in echocardiography. Our method focuses specifically on quantifying image quality within different myocardial segments, which can provide more granular and precise insights into the suitability of specific regions for various measurements.

Methods

In this work, we developed and compared three fully automatic methods to asses regional image quality in cardiac ultrasound separate from the view correctness:

- Classical ultrasound image quality metrics, such as CR and CNR, applied in cardiac regions automatically extracted using deep-learning segmentation.
- Deep-learning predicted ultrasound coherence, which is a measure of how coherent a signal is received by the transducer elements, together with deep-learning segmentation.
- End-to-end prediction of regional image quality.

The rest of this section first presents the datasets used to develop these methods, and then presents each of the three methods.

Datasets

VLCD

The **Very Large Cardiac Channel Data Database (VLCD)** consists of channel data from 33280 frames from 538 recordings of 106 study participants[7]. It contains parasternal short axis (PSAX), parasternal long axis (PLAX), apical long axis (ALAX), apical two-chamber (A2C), and apical four-chamber (A4C) views. We split the VLCD dataset on the study participant level into train, validation, and test sets, 70%, 15%, and 15% respectively.

HUNT4

The Nord-Trøndelag Health Study dataset (HUNT4Echo) is a clinical dataset including among others PSAX, PLAX, ALAX, A2C, and A4C views acquired using a GE Vivid E95 scanner and GE M4S probe. Each recording contains 3 cardiac cycles. We use two subsets of the HUNT4Echo dataset.

- **Segmentation annotation dataset** A fraction of 311 study participant exams, the segmentation annotation set[13], contains single frame segmentation annotations in both ED and ES as pixel-wise labels of the left ventricle (LV), left atrium (LA), and myocardium (MYO) in ALAX, A2C, and A4C views.

- **Regional image quality dataset** For this work, we created an additional dataset of image quality labels. The local image quality labels are manual annotations that asses the image quality of the cardiac regions of interest on a subset of the HUNT4 dataset in ALAX, A2C, and A4C views.

Regional image quality annotation on HUNT4

An annotation tool was developed specifically for this project using the open-source Annotation Web software[1][14]. The tool was made to enable

[1] `https://github.com/smistad/annotationweb`

clinicians to annotate regional image quality as efficiently and accurately as possible. The tool is freely available and can be adapted to other image quality projects. Table 1 defines the quality levels used in this work. Three cardiologists, each of whom performed more than 10,000 echocardiographic examinations and is European Association of Cardiovascular Imaging (EACVI) certified in transthoracic echocardiography, performed the quality annotations. The cardiologists used the following protocol:

1. Annotate the end-diastole (ED) and end-systole (ES) frame of each recording, and optionally other frames if the image quality changes significantly during the recording.
2. If the majority of the cardiac regions of interest is out-of-sector, label it as out-of-sector. Otherwise, label the part of the region that is inside the sector according to the definitions in Table 1. We ignore the out-of-sector regions in the remainder of this work.

For the first round of annotations, each of the three clinicians annotated the same 10 frames from 5 recordings of 2 study participants. We used this dataset to calculate the inter-observer variability. For the second round of annotations, the three clinicians collectively annotated 458 frames from 158 recordings of 65 study participants. The annotations from the second round form the **regional image quality dataset**. This dataset was split randomly at the study participant level into train, validation, and test sets, allocating 70%, 15%, and 15% of the data to each set respectively. Table 2 shows the consistency of each split across the ALAX, A2C, and A4C views.

Regional image quality estimation

Classical ultrasound image quality metrics

For the classical image quality metrics, deep-learning segmentation is used to extract the annulus regions and each of the myocardial segments as regions of interest and the LV as the background region. Appendix A gives more details about the procedure for dividing the segmentation into regions. The four classical ultrasound image quality metrics below were tested. We apply histogram matching[15, 16] to a Gaussian distribution ($\mu = 127, \sigma = 32$) for the B-Mode grayscale images before calculating pixel-based quality metrics.

- **Pixel intensity** is the average pixel intensity value in each region.
- **Contrast Ratio (*CR*)**[4] is defined as

$$CR = \frac{\mu_{\text{segment}}}{\mu_{\text{LV}}}$$

  where μ_{segment} is the average intensity in each region and μ_{LV} is the average intensity inside the LV lumen.
- **Contrast to Noise Ratio (*CNR*)[5]** is defined as

$$CNR = \frac{\mu_{\text{segment}} - \mu_{\text{LV}}}{\sqrt{\sigma^2_{\text{segment}} + \sigma^2_{\text{LV}}}}$$

  where σ_{segment} is the standard deviation in each region and σ_{LV} is the standard deviation inside the LV lumen.
- **Generalized CNR (*gCNR*)**[6] is defined as the maximum performance that can be expected from a hypothetical pixel classifier based on intensity using a set of optimal thresholds. It is calculated as

$$gCNR = 1 - \frac{1}{2} \sum_{i=0}^{MAX_i} \min\{p_{\text{segment}}(i), p_{\text{LV}}(i)\}$$

where $p_{\text{segment}}(x)$ is the probability density function of the pixel intensities inside the region the $gCNR$ is calculated for, $p_{\text{LV}}(x)$ the probability density function of the pixel intensities inside the LV lumen, and MAX_i the maximum possible pixel intensity. Fig. 1 shows an example of the probability density functions for one of the regions as ROI and the LV lumen as background.

Local, deep-learning predicted image coherence as quality metric

We use the VLCD dataset to calculate the coherence factor[17] for each pixel in the ultrasound image. This factor is the ratio between the amplitude of the sum of the received signals to the sum of the amplitudes of those signals,

$$CF = \frac{\sum_{n=1}^{N} S_i}{\sum_{n=1}^{N} |S_i|}$$

where S_i is the delayed signal for the i-th transducer element. This is equivalent to taking the coherent sum of the signal and dividing it by the incoherent sum of each signal. Thus, the coherence factor measures of how well the complex signals of all transducer elements align. The remainder of the signal-processing chain is the native processing of the GE HealthCare Vivid E95 system[2] but without the log compression. The result is a **coherence image** with the same dimension as the B-Mode image. The final preprocessing step applies gamma normalization with $\gamma = 0.5$ on the coherence images,

$$t_{i,normalized} = t_i^{\gamma}$$

[2] Gundersen et al.[18] describe this signal-processing chain in more detail.

where t_i are the pixels of the target coherence image. The corresponding B-mode images are generated from the channel data using the same, native signal-processing pipeline.

The HUNT4 dataset, like most ultrasound datasets, does not include channel data. Therefore, VLCD was used to train an image-to-image network that takes as input the grayscale B-mode image and predicts the coherence image, which is then used to calculate local image coherence. We use a lightweight U-Net architecture inspired by the U-Net 1 architecture in[19], with characteristics listed in Table 3. As coherence is related to image quality, we only apply augmentations that do not influence the quality of the image. Furthermore, the coherence should be invariant to different gain settings, so we additionally augment with brightness adjustments on the B-mode image while keeping the target coherence image unchanged. During training and validation of the coherence prediction model, we sample a random frame from each recording in the train and validation set respectively during each epoch. During testing, we use all frames in the test set. The local image coherence quality metric of a region is the average pixel value of all pixels corresponding to the region in the coherence image. This is the same as the pixel intensity metric above but applied to the coherence image instead of the B-Mode image.

End-to-end deep-learning quality prediction

The end-to-end learning approach trains a convolutional neural network on the regional image quality dataset to predict the quality of each region directly. The network predicts the image quality labels of all regions simultaneously, as illustrated in Fig. B.1 (a). As architecture, we start from MobileNetV2[20] and replace the final dense layer with a dense layer with eight outputs, one for each region. The weights of MobileNetV2 are initialized using a model pretrained on ImageNet [21], with all weights set as trainable during training. We treat the problem as a regression task where the model predicts a score for each segment. Table 1 shows the correspondence between quality scores and annotation labels. The loss function is the sum of the mean squared errors of each output and the model with the best validation loss is selected. Table 4 lists the remaining configuration details used during training. Appendix B describes the ablation study conducted to justify this setup.

Experimental setup

Evaluation of coherence prediction from B-mode

We use the structural similarity index (SSIM)[22], peak signal-to-noise ratio (PSNR)[3], and relative pixel error (RPE) to evaluate the coherence image prediction network. We define the RPE as

$$RPE = \frac{|t_i - p_i|}{max(t_i, \epsilon)}$$

where t_i are the pixel values of the target coherence image, p_i the pixel values of the predicted coherence image, and $\epsilon = 1e-4$.

Evaluation of quality metrics

The correlation and accuracy of each quality metric were measured by comparing them to the expert annotations on the test set of the regional

[3]The maximum pixel value for coherence images is 1.

image quality dataset. For the classic image quality methods and regional coherence method, linear regression models were used to map the quality metric values to image quality labels. The train and validation set were used together to fit the linear regression model and evaluate it on the test set.

Comparison to inter-observer variability

We compare the end-to-end, local image coherence, and gCNR-based model to the inter-observer variability on the data obtained in the first round of annotation. The inter-observer variability is calculated from the aggregate of the three unique pairwise score errors between each of the three annotators:

$$e_{\text{inter-observer}} = e_{12} \cup e_{23} \cup e_{13}$$

where e_{ij} is the score difference between operator i and j. The error metrics of the automatic methods are calculated from the aggregate of the pairwise score errors between the output of the method and each of the three annotators:

$$e_M = e_{1M} \cup e_{2M} \cup e_{3M}$$

where e_{iM} is the score difference between operator i and method M. .

Relation to variability in clinical measurements

This experiment evaluates whether there is a relation between the predicted quality and the agreement between different methods for clinical measurements. The hypothesis is that with lower image quality the variability, and thus the uncertainty, of the measurements between methods and between experts increases. More specifically, this analysis compares peak global longitudinal strain (GLS) and ejection fraction (EF) measurements obtained either fully automatically with AI tools or manually by using GE HealthCare EchoPAC software on HUNT4[23, 13]. For AI estimation of GLS and EF, the deep-learning methods proposed by Østvik et al.[24] and Smistad et al.[25] were used respectively. The study participants in HUNT4 used for model development were excluded from the analysis. For GLS, the predicted quality is the average quality of all segments over the full recording. For EF, the predicted quality is the average quality of all segments in the end-diastole (ED) and end-systole (ES) frames of all cycles in the recording.

Evaluation on CAMUS

To test the generalizability of the end-to-end model, we apply the model to each recording in the public CAMUS dataset [26] and compare the predictions to the reference quality labels. The predicted quality is the average quality of all segments over the full recording.

Results

Results of coherence prediction from B-mode

Table 5 summarizes the average metric values on the test set. Fig. 2 shows an example of the best, median, and worst-case predictions according to the relative pixel error. Fig. 3 illustrates how the predicted coherence images are almost independent of the brightness/gain and contrast/dynamic range of the input B-Mode images. The main finding is that the difference between the estimated and ground truth coherence images is small and thus the predicted coherence images can be used to obtain coherence-based quality metrics for B-mode for which the channel data is not available.

Results of quality metrics

Table 6 summarizes the results of the evaluation of the quality metrics. Fig. 4 shows box plots of the quality metrics per image quality label for the end-to-end, coherence, gCNR, and intensity models. Fig. 5 shows examples of B-mode images with varying quality together with labels from the annotators and automatic quality metrics. The main finding is that the end-to-end model performs the best, followed by the local image coherence metric. The classical ultrasound image quality metrics perform poorly.

Results of comparison to inter-observer variability

Fig. 6 shows the bar plot comparing the automatic methods to the inter-observer variability and Table 7 lists the corresponding average metric values. Using the Wilcoxon signed-rank test[27] and a significance level of $p = 0.05$, we find that the difference in mean absolute error (MAE) between each of the methods is statistically significant. The difference between the inter-observer MAE and the MAE of each of the methods is also statistically significant, i.e. the inter-observer MAE is higher than the MAE of the end-to-end model and lower than the MAE of the other two models.

Results of relation to variability in clinical measurements

Fig. 7 shows box plots visualizing the agreement between the measurements obtained automatically and with EchoPAC for each predicted quality category. The standard deviations in these plots represent how well the AI estimates agree with the manual references. The main finding is that the limits of agreement are narrower for higher qualities.

Results of evaluation on CAMUS

Fig. 8 shows the box plots of the average qualities over all frames as predicted by the end-to-end model for each quality category in CAMUS. Although the image quality categories in the CAMUS have different definitions compared to the ones in this study, better quality on one scale should on average relate to better quality on the other scale as well. Fig. 8 confirms this is indeed the case. The difference in average predicted quality between the different quality categories in CAMUS is statistically significant using the independent two-sample t-test and a significance level of $p = 0.05$.

Discussion

Challenges and considerations

Assessing image quality based on human perception is inherently a subjective task, even when supported by clear definitions of image quality categories. This creates challenges for training and evaluation as there is no ground truth as the reference labels are a subjective estimation themselves. Therefore, it is not realistic to expect the automatic models to agree with reference labels as well as on tasks with well-defined ground truth labels. This, together with a rather fine scale of image quality categories, explains the low accuracies in Tables 6 and 7. Fig. 6 shows how the end-to-end model has on average less error than the annotators between each other. This indicates the model has learned to produce quality labels that strike a middle ground between the subjective assessments of the annotators.

The end-to-end learning model overestimates low-quality regions and underestimates high-quality regions, as can be seen in Fig. 4a. This means that the model can only explain a limited amount of variability in the image quality labels and is a result of minimizing the mean squared error (MSE) while dealing with subjective, and thus noisy reference values with fixed boundaries. We can eliminate this effect by fitting a linear model on the validation set that maps predicted image quality to image quality labels and applying it when doing inference on the test set. This increases MSE but gives more uniform performance over the image quality labels.

One reason for the weak correlation between the annotations and the pixel-based methods is the rough selection of ROI and background region. Fig. 4 shows how the average metrics of the classic pixel-based and coherence metrics increase for each quality label until the *good* label, and then drop again for *excellent*. This is because on the one hand in these high-quality images, the blood speckles can be visible inside the LV lumen, which is used as background region, and on the other hand the myocardium tissue, which is used as ROI, is less blurred resulting in a smaller spread of pixels with high intensity. This can be seen for the anterolateral wall and apex in the rightmost column of Fig. 5. One possibility to only select regions belonging to the tissue is to perform automatic pixel selection methods like Otsu thresholding[28] or percentile filters, but in our experiments this reduced the performance even further.

It can be argued that classic metrics like the (g)CNR measure something conceptually different than the qualitative assessment of the clinicians. This work proposes a method to automate the extraction of these classic metrics and studies how well these align with subjective ratings of clinicians. The weak correlation does not necessarily mean that these metrics are inferior. Instead, the correct approach to measure image quality depends on the application, and we show that the classical metrics do not correlate well with the qualitative labels of the cardiologists in echocardiography.

Design choices

The different methods in this study have a trade-off between accuracy and versatility. The default end-to-end network gives the best results but requires specific image quality labels for the task. Next, the coherence-based method is more generic and can potentially be applied more generally without the need for view-specific image quality annotations. Rindal et al.[7] showed that the GIC is not significantly different between apical views, but is higher for apical views than parasternal views. Thus, while a single image-to-image model can learn to predict coherence for different views, the mapping from coherence to image quality should be done for each group of views separately. Another advantage is that coherence can be used to give a global image quality metric without the need for a segmentation model. Finally, the pixel-based methods can be applied automatically in the most general way given a segmentation model to select ROI and background regions but also give the lowest accuracy.

The ablation study of the end-to-end learning model showed that increasing the complexity of the model did not improve the performance. This is a result of the relatively small dataset size and the specific task of the model. For a more general model of image quality prediction with more varied input, e.g. one model for all cardiac views, a larger dataset and more complex model may be required.

Clinical use

Image quality estimation can be the first step towards a method for giving reliability estimates to clinical measurements and quality control of fully automatic methods. Fig. 7 shows that the variability in clinical measurements goes down with higher predicted quality. However, image quality is only one source of variability, so a reliability model would also need to include view correctness and other factors that determine whether a given input is difficult to assess.

More direct use cases of the quality prediction model include the automatic selection of the best frame to perform a clinical measurement when multiple options are available, data cleansing in data mining, and automatic disapproval of segments for regional strain analysis. All the methods explored in this work are computationally efficient and can be run in real-time while scanning, and can thus be used as a guidance tool to enable clinicians to acquire images with better image quality.

Computational complexity

Table 8 compares the processing time of the three quality estimation methods described in this study. The runtimes are dominated by the inference times of the relevant neural networks, shown in Table 9. Both the gCNR method and local coherence method run the nnU-Net segmentation model at inference. The coherence method additionally runs the coherence prediction model, assuming the coherence image is not available. The end-to-end method only runs the dedicated end-to-end model.

Real-time demo application and examples

To showcase the functionality of the end-to-end, real-time quality network, a real-time application was developed using the FAST framework[29]. The demo is a split-screen application that shows the B-Mode input to the left and the segmentation regionally color-coded by the quality as predicted by the end-to-end network to the right. Fig. 9 provides a screenshot of the application in use. We provide a demo video[30] illustrating the application in action while a clinician operates a GE Vivid E95 scanner. The video can be accessed at `https://doi.org/10.6084/m9.figshare.26413984.v2`. The video demonstrates the output of the end-to-end neural network and how it reacts to different scenarios such as lung obstruction. The video also shows that the proposed end-to-end method can run in real-time while scanning and thus may be used to guide operators to achieve good image quality while scanning.

A video with more examples of image quality predictions on recordings from study participants in the HUNT4 study with high ($BMI > 30$) and normal ($20 < BMI < 25$) body mass index (BMI) can be accessed at `https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_additional_examples/27730251?file=50487105`. The study participants with high BMI have lower image quality on average.

However, many other factors influence image quality as well.

Conclusion

In this work, we developed and compared different deep-learning methods for regional image quality estimation in cardiac ultrasound. We show that classic pixel-based methods, such as (g)CNR, together with automatic image segmentation, give low agreement with the quality assessment of cardiologists. We developed a U-Net model to predict the coherence factor for each pixel in the ultrasound image and showed that the resulting coherence image can be used to assess the image quality in a pixel-based way with better performance than the classic measures. The best results, below inter-observer variability, are obtained by using an end-to-end deep-learning model. Finally, we show higher predicted quality is associated with lower limits of agreement between fully automatic and manual methods for the clinical measurements EF and GLS.

Acknowledgments

This work was supported in part by the Research Council of Norway under Project 237887. During the preparation of this work the authors used Grammarly text editor and ChatGPT for assistance in sentence editing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflict of interest

The authors declare no conflict of interest.

Data availability

The data used in this study is confidential and is therefore not publicly available.

Appendix A. Extraction of cardiac regions of interest

We use the nnU-Net[31, 32] architecture to segment the cardiac images. The nnU-Net is used out of the box using the default configuration but without the final ensemble step. Instead, we train and validate on a single predefined 80% train, 10% validation, and 10% test split from the HUNT4 segmentation annotation dataset. Table A.1 summarizes the characteristics of the nnU-Net architecture. This model is described in more detail and compared to other segmentation models in our previous work[33]. We use two nnU-Nets, one for apical two- (A2C) and four-chamber (A4C) views, and one for apical long axis (ALAX) views. The nnU-Net for A2C and A4C views segments the left ventricle (LV), left atrium (LA) and myocardium (MYO). The nnU-Net for ALAX views additionally segments the aorta (AO).

The segmentation of the MYO is divided into eight regions using the following algorithm:

1. Extract the annulus points. For A2C/A4C views, these are the points where the MYO meets the LA. For ALAX views, these are the points where the MYO meets the LA and AO. Points A and B are the annulus points in Fig. A.1.
2. Extract the apex of the LV, defined as the furthest points from the base points within the lv lumen. This is point C in Fig. A.1.
3. Divide both the left and right part of the endocardium border, defined as the border between the LV and MYO regions, into three parts with equal length. This gives points D, E, F and G in Fig. A.1.
4. Find the closest points on the outer MYO border for points C, D, E, F and G. These are points H, I, J, K and L in Fig. A.1.
5. Fill in the regions by connecting the points via the contour, resulting in the MYO divided into six regions.

6. Draw circles[4] with a radius of 2 millimeters around the annulus points, points A and B in Fig. A.1. We use these additional two regions to asses the local image quality of the annulus points in the image. The result are the eight regions, as in Fig. A.1.
7. Remove any parts of regions that fall outside of the sector. The apical top regions in Fig. 1 (a), i.e., the yellow and white masks, are examples of this. If more than 50% of all pixels inside the region fall outside the sector, we exclude the region from analysis.

The goal is to automatically quantify the image quality in each of these eight regions. The LV lumen is used as background region.

Appendix B. Ablation study of the end-to-end learning model

In the ablation study, we evaluated the impact of modifying the architecture of the end-to-end learning model. Three different network architectures were tested: Cardiac View Classification (CVC) network[3], MobileNetV2[20], and EfficientNet[34]. We tested approaching the problem both as a classification and regression task, with only the final dense layer and loss function being changed accordingly. Additionally, three basic network attention variations were tested using the automatic segmentation output. The default model did not use any attention and predicts each label directly, as in Fig. B.1 (a). For the other two versions, the region masks extracted from the segmentation were dilated with a square dilation filter of size 50x50 pixels and used as an additional input to the networks which then predict the label of one region at a time. The dilation filter reveals the direct vicinity around each region so the boundary between tissue and background becomes visible. The first variant used this dilated mask as hard attention by blacking out the other parts of the image, as shown in Fig. B.1 (b). For the second variant, the

[4]Due to the unequal pixel spacing in depth and width, the annulus regions become ovals in the 256x256 segmentation maps. When plotting the images with equal spacing in width and depth, these regions become circles again.

masks were used as soft attention as input to a side branch of the network, as proposed by Eppel[35]. For this version, we created an attention map and added it element-wise to the output of the first layer, corresponding to version 'c' in Eppel[35]. Fig. B.1 (c) shows this configuration.

The ablation study consists of two parts. Both parts used the training configuration listed in Table 4 and data from the regional image quality dataset. In the first part, we examined the effect of changing the convolutional backbone and the effect of framing the problem as a classification or regression task. Table B.1 compares the predictions with the annotations on the test set for the different configurations using the default end-to-end model. In the second part of the ablation study, the backbone was fixed to MobileNetV2[20] and the problem was framed as a regression task. Next, the different variants shown in Fig. B.1 were compared. Table B.2 summarizes the results on the test set. *The default end-to-end model with no attention, the MobileNetV2[20] architecture, and the problem framed as a regression task gave the best results and were thus used for this paper.*

References

[1] Pasdeloup D, Olaisen SH, Østvik A, Sabo S, Pettersen HN, Holte E, et al. Real-time echocardiography guidance for optimized apical standard views. Ultrasound in Medicine & Biology. 2023;49(1):333-46.

[2] Droste R, Drukker L, Papageorghiou AT, Noble JA. Automatic probe movement guidance for freehand obstetric ultrasound. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. Springer; 2020. p. 583-92.

[3] Østvik A, Smistad E, Aase SA, Haugen BO, Lovstakken L. Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. Ultrasound in medicine & biology. 2019;45(2):374-84.

[4] Smith S, Lopez H, Bodine Jr W. Frequency independent ultrasound contrast-detail analysis. Ultrasound in medicine & biology. 1985;11(3):467-77.

[5] Patterson M, Foster F. Improvement and Quantitative Assessment of B-mode Images Produced by Annular Array/Cone Hybrids. In: Acoustical Imaging. Springer; 1984. p. 477-7.

[6] Rodriguez-Molares A, Rindal OMH, D'hooge J, Måsøy SE, Austeng A, Bell MAL, et al. The generalized contrast-to-noise ratio: A formal definition for lesion detectability. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control. 2019;67(4):745-59.

[7] Rindal OMH, Bjåstad TG, Espeland T, Berg EAR, Måsøy SE. A very large cardiac channel data database (VLCD) used to evaluate Global Image Coherence (GIC) as an in-vivo image quality metric. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control. 2023.

[8] Abdi AH, Luong C, Tsang T, Jue J, Gin K, Yeung D, et al. Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard

view planes. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. Springer; 2017. p. 302-10.

[9] Luong C, Liao Z, Abdi A, Girgis H, Rohling R, Gin K, et al. Automated estimation of echocardiogram image quality in hospitalized patients. The International Journal of Cardiovascular Imaging. 2021;37:229-39.

[10] Van Woudenberg N, Liao Z, Abdi AH, Girgis H, Luong C, Vaseli H, et al. Quantitative echocardiography: real-time quality estimation and view classification implemented on a mobile android device. In: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings. Springer; 2018. p. 74-81.

[11] Labs RB, Vrettos A, Loo J, Zolgharni M. Automated assessment of transthoracic echocardiogram image quality using deep neural networks. Intelligent Medicine. 2023;3(03):191-9.

[12] Karamalis A, Wein W, Klein T, Navab N. Ultrasound confidence maps using random walks. Medical image analysis. 2012;16(6):1101-12.

[13] Olaisen S, Smistad E, Espeland T, Hu J, Pasdeloup D, Østvik A, et al. Automatic measurements of left ventricular volumes and ejection fraction by artificial intelligence: clinical validation in real time and large databases. European Heart Journal - Cardiovascular Imaging. 2023 10:jead280. Available from: `https://doi.org/10.1093/ehjci/jead280`.

[14] Smistad E, Østvik A, Løvstakken L. Annotation Web-An open-source web-based annotation tool for ultrasound images. In: 2021 IEEE International Ultrasonics Symposium (IUS). IEEE; 2021. p. 1-4.

[15] Bottenus N, Byram BC, Hyun D. Histogram matching for visual ultrasound image

comparison. IEEE transactions on ultrasonics, ferroelectrics, and frequency control. 2020;68(5):1487-95.

[16] Gonzalez RC. Digital image processing. Pearson education india; 2009.

[17] Rigby KW. Method and apparatus for coherence filtering of ultrasound images. Google Patents; 1999. US Patent 5,910,115.

[18] Gundersen EL, Smistad E, Jahren TS, Måsøy SE. Hardware-Independent Deep Signal Processing: A Feasibility Study in Echocardiography. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control. 2024.

[19] Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE transactions on medical imaging. 2019;38(9):2198-210.

[20] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4510-20.

[21] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248-55.

[22] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004;13(4):600-12.

[23] Nyberg J, Jakobsen EO, Østvik A, Holte E, Stølen S, Lovstakken L, et al. Echocardiographic reference ranges of global longitudinal strain for all cardiac chambers using guideline-directed dedicated views. JACC: Cardiovascular Imaging. 2023;16(12):1516-31.

[24] Østvik A, Salte IM, Smistad E, Nguyen TM, Melichova D, Brunvand H, et al. Myocardial function imaging in echocardiography using deep learning. ieee transactions on medical imaging. 2021;40(5):1340-51.

[25] Smistad E, Østvik A, Salte IM, Melichova D, Nguyen TM, Haugaa K, et al. Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control. 2020;67(12):2595-604.

[26] Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. IEEE Transactions on Medical Imaging. 2019 Sep;38(9):2198-210. Conference Name: IEEE Transactions on Medical Imaging.

[27] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80-3. Available from: http://www.jstor.org/stable/3001968.

[28] Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979;9(1):62-6.

[29] Smistad E, Bozorgi M, Lindseth F. FAST: framework for heterogeneous medical image computing and visualization. International Journal of computer assisted radiology and surgery. 2015;10:1811-22.

[30] Vyver GVD. Regional quality estimation for echocardiography using deep learning - real-time demo. 2024 7. Available from: https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_real-time_demo/26413984.

[31] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203-11.

[32] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnUNet. GitHub; 2023. https://github.com/MIC-DKFZ/nnUNet.

[33] Vyver GVD, Thomas S, Ben-Yosef G, Olaisen SH, Dalen H, Løvstakken L, et al. Toward Robust Cardiac Segmentation Using Graph Convolutional Networks. IEEE Access. 2024;12:33876-88.

[34] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019. p. 6105-14.

[35] Eppel S. Classifying a specific image region using convolutional nets with an ROI mask as input. arXiv preprint arXiv:181200291. 2018.

Figure Captions

Figure 1: Calculating gCNR for a region of the myocardium. The left side of the figure shows the segmentation with the MYO and annulus points divided into regions. The right side shows the probability density functions of the segment and background pixels used to calculate the gCNR. In this example, we use the mid region on the left side as ROI and the full LV lumen as background. These correspond to the green and red masks respectively in the left part of the figure.

Figure 2: Best, median, and worst case of image-to-image coherence prediction task with relative pixel errors of 4.0, 5.4, and 9.1the coherence image as predicted by the image-to-image network. The third column shows the ground truth coherence image as calculated from the channel data. Finally, the rightmost column shows the color-coded difference of the target minus the predicted image. The images come from the VLCD dataset, which is only used to train the coherence network on the pixel level. Thus, the correctness of the view and its alignment are not crucial in this context.

Figure 3: Effect of brightness on coherence prediction. The first row shows a B-Mode image from the regional image quality dataset, brightened and darkened with gamma correction ($\gamma = 0.9$ and 1.1). The second row shows the predicted coherence images generated by giving the corresponding input from the first row to the network. The predicted coherence is unaffected by the adjustments in brightness, apart from the saturation effect in the brightened image reducing the information in the input, as can be seen in the basal part of the inferolateral wall

Figure 4: Box plots of quality metrics versus regional quality labels on the test set of the regional image quality dataset. The predictions of the end-to-end model have the strongest correlation to the quality labels. The dotted line represents the linear regression model that maps the quality metrics to quality labels. The inference output of the end-to-end model can be used directly without additional linear model.

Figure 5: Example cases of annotations and automatically predicted regional quality from the test set. The visualization uses the regional quality metrics to color-code

the output of divided segmentation output. The end-to-end model predicts the regional qualities directly from the B-mode without using the segmentation output. The local image coherence metric uses the segmentation output to select ROI. The gCNR metric uses the segmentation output to select ROI and background region. The background region, i.e., the LV lumen, is not shown in the image

Figure 6: Bar plot comparing inter-observer variability to automatic methods. A method with lower variability will have the most occurrences with low score errors. Here we can observe that the variability of the end-to-end model is on par with the inter-observer variability, while the two other methods are not

Figure 7: Box plots of the difference between clinical measurement values obtained automatically by AI [24, 25] and reference measurements obtained manually using GE HealthCare EchoPAC on the HUNT4 data [23, 13], per image quality category, as predicted by the end-to-end model. The decrease in standard deviation with better image quality indicates a better agreement between the methods for higher image quality. Additionally, there is a noticeable change in bias between different quality categories. We believe this effect is partly caused by physiological differences correlated with image quality and is out of scope for this work.

Figure 8: Box plots of the predicted quality scores for each quality category in the CAMUS dataset, based on the end-to-end model's predictions.

Figure 9: Screenshot of the real-time demo application. The left side shows the input B-Mode image. The right side shows the output of the segmentation color-coded by the output of the end-to-end quality network. The color codes are the same as in Fig. 5.

Video Captions

Video 1: Real-time demo of automatic, regional image quality estimation. The demo uses the end-to-end image quality model.

Video 2: Additional examples of image quality predictions on recordings from study participants in the HUNT4 study with normal and medium BMI. The study participants with high BMI have lower image quality on average.