

# Automatic regional image quality scoring for echocardiography

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

<sup>a</sup>*Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway*

<sup>b</sup>*Health Research, SINTEF, Trondheim, Norway*

<sup>c</sup>*St. Olavs hospital hospital, Trondheim, Norway*

---

## Abstract

## Objective

To develop and compare methods for the automatic estimation of regional cardiac ultrasound image quality 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

---

## 1 Introduction

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

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

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

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

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

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

50 available in practice and does not give regional metrics.

51

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

67

68 All of the automatic methods mentioned above have the limitation that  
69 they only provide a global image quality evaluation and/or do not assess the  
70 image quality separate from the view correctness. The novelty of this work is  
71 automatic quality estimation on the regional level. To the best of our knowl-  
72 edge, this is the first study to propose a method for automatic regional image  
73 quality assessment in echocardiography. Our method focuses specifically on  
74 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 assess 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.

#### 100 *HUNT4*

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

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

110 • **Regional image quality dataset** For this work, we created an addi-  
111 tional dataset of image quality labels. The local image quality labels  
112 are manual annotations that asses the image quality of the cardiac re-  
113 gions of interest on a subset of the HUNT4 dataset in ALAX, A2C,  
114 and A4C views.

#### 115 *Regional image quality annotation on HUNT4*

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

---

<sup>1</sup><https://github.com/smistad/annotationweb>

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

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

132 For the first round of annotations, each of the three clinicians annotated  
 133 the same 10 frames from 5 recordings of 2 study participants. We used this  
 134 dataset to calculate the inter-observer variability. For the second round of  
 135 annotations, the three clinicians collectively annotated 458 frames from 158  
 136 recordings of 65 study participants. The annotations from the second round  
 137 form the **regional image quality dataset**. This dataset was split randomly  
 138 at the study participant level into train, validation, and test sets, allocating  
 139 70%, 15%, and 15% of the data to each set respectively. Table 2 shows the  
 140 consistency of each split across the ALAX, A2C, and A4C views.



141 *Regional image quality estimation*

142 *Classical ultrasound image quality metrics*

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

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

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

152 where  $\mu_{\text{segment}}$  is the average intensity in each region and  $\mu_{\text{LV}}$  is the  
153 average intensity inside the LV lumen.

- 154 • **Contrast to Noise Ratio (*CNR*)** [5] is defined as

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

155 where  $\sigma_{\text{segment}}$  is the standard deviation in each region and  $\sigma_{\text{LV}}$  is the  
156 standard deviation inside the LV lumen.

- 157 • **Generalized CNR (*gCNR*)** [6] is defined as the maximum perfor-  
158 mance that can be expected from a hypothetical pixel classifier based  
159 on intensity using a set of optimal thresholds. It is calculated as

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

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

#### 166 *Local, deep-learning predicted image coherence as quality metric*

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

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

171 where  $S_i$  is the delayed signal for the  $i$ -th transducer element. This is equiv-  
 172 alent to taking the coherent sum of the signal and dividing it by the incoher-  
 173 ent sum of each signal. Thus, the coherence factor measures of how well the  
 174 complex signals of all transducer elements align. The remainder of the signal-  
 175 processing chain is the native processing of the GE HealthCare Vivid E95  
 176 system<sup>2</sup> but without the log compression. The result is a **coherence image**  
 177 with the same dimension as the B-Mode image. The final preprocessing step  
 178 applies gamma normalization with  $\gamma = 0.5$  on the coherence images,

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

---

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

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

182

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

#### 200 *End-to-end deep-learning quality prediction*

201 The end-to-end learning approach trains a convolutional neural network  
 202 on the regional image quality dataset to predict the quality of each region  
 203 directly. The network predicts the image quality labels of all regions si-

204 multaneously, as illustrated in Fig. B.1 (a). As architecture, we start from  
 205 MobileNetV2 [20] and replace the final dense layer with a dense layer with  
 206 eight outputs, one for each region. The weights of MobileNetV2 are initialized  
 207 using a model pretrained on ImageNet [21], with all weights set as trainable  
 208 during training. We treat the problem as a regression task where the model  
 209 predicts a score for each segment. Table 1 shows the correspondence between  
 210 quality scores and annotation labels. The loss function is the sum of the mean  
 211 squared errors of each output and the model with the best validation loss is  
 212 selected. Table 4 lists the remaining configuration details used during train-  
 213 ing. Appendix B describes the ablation study conducted to justify this setup.

214

## 215 **Experimental setup**

### 216 *Evaluation of coherence prediction from B-mode*

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

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

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

### 222 *Evaluation of quality metrics*

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

---

<sup>3</sup>The maximum pixel value for coherence images is 1.

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

### 229 *Comparison to inter-observer variability*

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

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

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

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

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

### 238 *Relation to variability in clinical measurements*

239 This experiment evaluates whether there is a relation between the pre-  
 240 dicted quality and the agreement between different methods for clinical mea-  
 241 surements. The hypothesis is that with lower image quality the variability,  
 242 and thus the uncertainty, of the measurements between methods and between  
 243 experts increases. More specifically, this analysis compares peak global lon-  
 244 gitudinal strain (GLS) and ejection fraction (EF) measurements obtained

245 either fully automatically with AI tools or manually by using GE Health-  
246 Care EchoPAC software on HUNT4 [23, 13]. For AI estimation of GLS and  
247 EF, the deep-learning methods proposed by Østvik et al. [24] and Smistad  
248 et al. [25] were used respectively. The study participants in HUNT4 used for  
249 model development were excluded from the analysis. For GLS, the predicted  
250 quality is the average quality of all segments over the full recording. For EF,  
251 the predicted quality is the average quality of all segments in the end-diastole  
252 (ED) and end-systole (ES) frames of all cycles in the recording.

### 253 *Evaluation on CAMUS*

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

## 258 **Results**

### 259 *Results of coherence prediction from B-mode*

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

### 268 *Results of quality metrics*

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

### 276 *Results of comparison to inter-observer variability*

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

### 285 *Results of relation to variability in clinical measurements*

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

## 291 *Results of evaluation on CAMUS*

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

## 300 **Discussion**

### 301 *Challenges and considerations*

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

313



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

323  
 324 One reason for the weak correlation between the annotations and the  
 325 pixel-based methods is the rough selection of ROI and background region.  
 326 Fig. 4 shows how the average metrics of the classic pixel-based and coherence  
 327 metrics increase for each quality label until the *good* label, and then drop  
 328 again for *excellent*. This is because on the one hand in these high-quality  
 329 images, the blood speckles can be visible inside the LV lumen, which is used  
 330 as background region, and on the other hand the myocardium tissue, which  
 331 is used as ROI, is less blurred resulting in a smaller spread of pixels with  
 332 high intensity. This can be seen for the anterolateral wall and apex in the  
 333 rightmost column of Fig. 5. One possibility to only select regions belong-  
 334 ing to the tissue is to perform automatic pixel selection methods like Otsu  
 335 thresholding [28] or percentile filters, but in our experiments this reduced the  
 336 performance even further.

337  
 338 It can be argued that classic metrics like the (g)CNR measure something

339 conceptually different than the qualitative assessment of the clinicians. This  
340 work proposes a method to automate the extraction of these classic metrics  
341 and studies how well these align with subjective ratings of clinicians. The  
342 weak correlation does not necessarily mean that these metrics are inferior.  
343 Instead, the correct approach to measure image quality depends on the ap-  
344 plication, and we show that the classical metrics do not correlate well with  
345 the qualitative labels of the cardiologists in echocardiography.

#### 346 *Design choices*

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

361

362 The ablation study of the end-to-end learning model showed that increas-  
363 ing the complexity of the model did not improve the performance. This is a

364 result of the relatively small dataset size and the specific task of the model.  
365 For a more general model of image quality prediction with more varied in-  
366 put, e.g. one model for all cardiac views, a larger dataset and more complex  
367 model may be required.

### 368 *Clinical use*

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

376

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

### 384 **Computational complexity**

385 Table 8 compares the processing time of the three quality estimation  
386 methods described in this study. The runtimes are dominated by the infer-  
387 ence times of the relevant neural networks, shown in Table 9. Both the gCNR

388 method and local coherence method run the nnU-Net segmentation model  
389 at inference. The coherence method additionally runs the coherence predic-  
390 tion model, assuming the coherence image is not available. The end-to-end  
391 method only runs the dedicated end-to-end model.

## 392 Real-time demo application and examples

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

406 A video with more examples of image quality predictions on recordings  
407 from study participants in the HUNT4 study with high ( $BMI > 30$ ) and nor-  
408 mal ( $20 < BMI < 25$ ) body mass index (BMI) can be accessed at [https://](https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_additional_examples/27730251?file=50487105)  
409 [figshare.com/articles/media/Regional\\_quality\\_estimation\\_for\\_echocardiography\\_](https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_additional_examples/27730251?file=50487105)  
410 [using\\_deep\\_learning\\_-\\_additional\\_examples/27730251?file=50487105](https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_additional_examples/27730251?file=50487105).  
411 The study participants with high BMI have lower image quality on average.

412 However, many other factors influence image quality as well.

## 413 Conclusion

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

426

## 427 Acknowledgments

428 This work was supported in part by the Research Council of Norway under Project  
429 237887. During the preparation of this work the authors used Grammarly text editor  
430 and ChatGPT for assistance in sentence editing. After using these tools, the authors  
431 reviewed and edited the content as needed and take full responsibility for the content of  
432 the publication.

## 433 Conflict of interest

434 The authors declare no conflict of interest.

## 435 *Data availability*

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

## 437 **Appendix A. Extraction of cardiac regions of interest**

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

448

449 The segmentation of the MYO is divided into eight regions using the following algo-  
450 rithm:

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

- 463 6. Draw circles<sup>4</sup> with a radius of 2 millimeters around the annulus points, points A  
464 and B in Fig. A.1. We use these additional two regions to asses the local image  
465 quality of the annulus points in the image. The result are the eight regions, as in  
466 Fig. A.1.
- 467 7. Remove any parts of regions that fall outside of the sector. The apical top regions  
468 in Fig. 1 (a), i.e., the yellow and white masks, are examples of this. If more than  
469 50% of all pixels inside the region fall outside the sector, we exclude the region from  
470 analysis.

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

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

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

---

<sup>4</sup>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.

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

490

491 The ablation study consists of two parts. Both parts used the training configuration  
492 listed in Table 4 and data from the regional image quality dataset. In the first part, we  
493 examined the effect of changing the convolutional backbone and the effect of framing the  
494 problem as a classification or regression task. Table B.1 compares the predictions with the  
495 annotations on the test set for the different configurations using the default end-to-end  
496 model. In the second part of the ablation study, the backbone was fixed to MobileNetV2  
497 [20] and the problem was framed as a regression task. Next, the different variants shown  
498 in Fig. B.1 were compared. Table B.2 summarizes the results on the test set. *The default*  
499 *end-to-end model with no attention, the MobileNetV2 [20] architecture, and the problem*  
500 *framed as a regression task gave the best results and were thus used for this paper.*



## 501 References

- 502 [1] Padeloup D, Olaisen SH, Østvik A, Sabo S, Pettersen HN, Holte E, et al. Real-  
503 time echocardiography guidance for optimized apical standard views. *Ultrasound in*  
504 *Medicine & Biology*. 2023;49(1):333-46.
- 505 [2] Droste R, Drukker L, Papageorghiou AT, Noble JA. Automatic probe movement  
506 guidance for freehand obstetric ultrasound. In: *Medical Image Computing and Com-*  
507 *puter Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima,*  
508 *Peru, October 4–8, 2020, Proceedings, Part III* 23. Springer; 2020. p. 583-92.
- 509 [3] Østvik A, Smistad E, Aase SA, Haugen BO, Lovstakken L. Real-time standard view  
510 classification in transthoracic echocardiography using convolutional neural networks.  
511 *Ultrasound in medicine & biology*. 2019;45(2):374-84.
- 512 [4] Smith S, Lopez H, Bodine Jr W. Frequency independent ultrasound contrast-detail  
513 analysis. *Ultrasound in medicine & biology*. 1985;11(3):467-77.
- 514 [5] Patterson M, Foster F. Improvement and Quantitative Assessment of B-mode Images  
515 Produced by Annular Array/Cone Hybrids. In: *Acoustical Imaging*. Springer; 1984.  
516 p. 477-7.
- 517 [6] Rodriguez-Molares A, Rindal OMH, D’hooge J, Måsøy SE, Austeng A, Bell MAL,  
518 et al. The generalized contrast-to-noise ratio: A formal definition for lesion de-  
519 tectability. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*.  
520 2019;67(4):745-59.
- 521 [7] Rindal OMH, Bjåstad TG, Espeland T, Berg EAR, Måsøy SE. A very large cardiac  
522 channel data database (VLCD) used to evaluate Global Image Coherence (GIC) as  
523 an in-vivo image quality metric. *IEEE Transactions on Ultrasonics, Ferroelectrics,*  
524 *and Frequency Control*. 2023.
- 525 [8] Abdi AH, Luong C, Tsang T, Jue J, Gin K, Yeung D, et al. Quality assessment of  
526 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 Sys-  
tems 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, Pasedeloup D, Østvik A, et al. Au-  
tomatic measurements of left ventricular volumes and ejection fraction by artifi-  
cial 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

- 553 comparison. IEEE transactions on ultrasonics, ferroelectrics, and frequency control.  
554 2020;68(5):1487-95.
- 555 [16] Gonzalez RC. Digital image processing. Pearson education india; 2009.
- 556 [17] Rigby KW. Method and apparatus for coherence filtering of ultrasound images.  
557 Google Patents; 1999. US Patent 5,910,115.
- 558 [18] Gundersen EL, Smistad E, Jahren TS, Måsøy SE. Hardware-Independent Deep Sig-  
559 nal Processing: A Feasibility Study in Echocardiography. IEEE Transactions on  
560 Ultrasonics, Ferroelectrics, and Frequency Control. 2024.
- 561 [19] Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep  
562 learning for segmentation using an open large-scale dataset in 2D echocardiography.  
563 IEEE transactions on medical imaging. 2019;38(9):2198-210.
- 564 [20] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted  
565 residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer  
566 vision and pattern recognition; 2018. p. 4510-20.
- 567 [21] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hier-  
568 archical image database. In: 2009 IEEE conference on computer vision and pattern  
569 recognition. Ieee; 2009. p. 248-55.
- 570 [22] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from  
571 error visibility to structural similarity. IEEE transactions on image processing.  
572 2004;13(4):600-12.
- 573 [23] Nyberg J, Jakobsen EO, Østvik A, Holte E, Stølen S, Lovstakken L, et al. Echocar-  
574 diographic reference ranges of global longitudinal strain for all cardiac chambers using  
575 guideline-directed dedicated views. JACC: Cardiovascular Imaging. 2023;16(12):1516-  
576 31.

- 577 [24] Østvik A, Salte IM, Smistad E, Nguyen TM, Melichova D, Brunvand H, et al. My-  
578 ocardial function imaging in echocardiography using deep learning. *ieee transactions*  
579 *on medical imaging*. 2021;40(5):1340-51.
- 580 [25] Smistad E, Østvik A, Salte IM, Melichova D, Nguyen TM, Haugaa K, et al. Real-  
581 Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learn-  
582 ing. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*.  
583 2020;67(12):2595-604.
- 584 [26] Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep  
585 Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiog-  
586 raphy. *IEEE Transactions on Medical Imaging*. 2019 Sep;38(9):2198-210. Conference  
587 Name: *IEEE Transactions on Medical Imaging*.
- 588 [27] Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*.  
589 1945;1(6):80-3. Available from: <http://www.jstor.org/stable/3001968>.
- 590 [28] Otsu N. A threshold selection method from gray-level histograms. *IEEE transactions*  
591 *on systems, man, and cybernetics*. 1979;9(1):62-6.
- 592 [29] Smistad E, Bozorgi M, Lindseth F. FAST: framework for heterogeneous medical image  
593 computing and visualization. *International Journal of computer assisted radiology*  
594 *and surgery*. 2015;10:1811-22.
- 595 [30] Vyver GVD. Regional quality estimation for echocardiography using deep learning -  
596 real-time demo. 2024 7. Available from: [https://figshare.com/articles/media/](https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_real-time_demo/26413984)  
597 [Regional\\_quality\\_estimation\\_for\\_echocardiography\\_using\\_deep\\_learning\\_](https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_real-time_demo/26413984)  
598 [-\\_real-time\\_demo/26413984](https://figshare.com/articles/media/Regional_quality_estimation_for_echocardiography_using_deep_learning_-_real-time_demo/26413984).
- 599 [31] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-  
600 configuring method for deep learning-based biomedical image segmentation. *Nature*  
601 *methods*. 2021;18(2):203-11.
- 602 [32] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnUNet. GitHub; 2023.  
603 <https://github.com/MIC-DKFZ/nnUNet>.

- 604 [33] Vyver GVD, Thomas S, Ben-Yosef G, Olaisen SH, Dalen H, Løvstakken L, et al.  
605       Toward Robust Cardiac Segmentation Using Graph Convolutional Networks. IEEE  
606       Access. 2024;12:33876-88.
- 607 [34] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural net-  
608       works. In: International conference on machine learning. PMLR; 2019. p. 6105-14.
- 609 [35] Eppel S. Classifying a specific image region using convolutional nets with an ROI  
610       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 9.1. The 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

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

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

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

657 **Figure 8:** Box plots of the predicted quality scores for each quality category in the CA-  
658 MUS dataset, based on the end-to-end model’s predictions.

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

## 663 **Video Captions**

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

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