Echocardiography View Classification Using Quality Transfer Star Generative Adversarial Networks

Zhibin Liao[1⋆], Mohammad H. Jafari[1⋆], Hany Girgis[2⋆], Kenneth Gin[2], Robert Rohling[1], Purang Abolmaesumi[1⋆⋆], and Teresa Tsang[2⋆⋆]

[1]The University of British Columbia, Canada
2Vancouver General Hospital, Canada

Abstract. 2D echocardiography (echo) is the most widely used imaging technique to identify cardiac disease. In addition to anatomical variability in patients, the quality of acquired echo image can vary significantly depending on the ultrasound (US) machine and the experience level of the operator, where a poor image quality can affect the diagnosis. This variability can also result in reduced performance of machine learning models trained on these data. With the recent advances in generative adversarial networks (GAN), we demonstrate that it is possible to transfer the image quality of echo images to a user-defined quality level with the use of a multi-domain transfer approach referred as StarGAN. The proposed quality transfer StarGAN (QT-StarGAN) requires no pairs of low- and high-quality echo images and incorporates the temporal information of echo images during the training phase. We evaluate the proposed approach using 16,612 echo cine series obtained from 3,157 patients. Using a standard echo view classification task, we demonstrate that the accuracy of classification is significantly improved using QT-StarGAN.

Keywords: Echocardiography · Ultrasound · Image Quality · View Classification · Generative Network · Domain Transfer · GAN.

1 Introduction

2D echocardiography (echo) is the most widely used imaging technique to identify cardiac disease. Echo examination is low-cost and non-invasive, and given its portability specifically using the latest generation mobile ultrasound technology, it can be used as first line imaging for emergency diagnoses and point-of-care. The echo image quality is determined by the presence and contrast of target heart anatomies (*i.e.*, valve and chamber wall), proper centering , imaging depth, and imaging gain settings. Hence, in addition to inter-patient anatomical variations, the experience level of a sonographer and the imaging system itself are the main factors contributing to echo image quality.

To improve quality, conventional methods have investigated speckle reduction and edge enhancement [2,6,16,20], which are difficult to generalize across patients

⋆ Joint first-authors
⋆⋆ Joint corresponding authors

and imaging systems. With the increasing popularity of deep learning in medical imaging, neural networks-based image de-nosing and super-resolution methods have been shown to improve image quality of CT and PET [3, 13]. However, these methods typically require paired source and target domain images (*i.e.*, low and high quality counterparts showing the exact patient anatomy) to train the transfer model, which is virtually impossible to obtain during echo acquisition given above-mentioned factors contributing to variations in image quality. On the other hand, deep neural networks have been also used to assess ultrasound image quality [1, 21, 22] with the supervision of expert score labels.

Image quality transfer task can be viewed as an image-to-image translation problem conditioned on the quality attribute. For echo images, the difficulty mainly stems from the limited availability of source and target image pairs. The recently proposed Cycle-Consistent Generative Adversarial Networks (GAN) [24] enforces the image transfer consistency in a GAN framework to avoid the need for image pairs. CycleGAN has fixed source and target domains; thus, multiple CycleGAN models are needed for multi-domain transfer. StarGAN [5] combines the CycleGAN and conditional GAN [12] to encapsulate multi-domain transfer in one model.

In this paper, we propose an echo image quality transfer network, trained with echo images and corresponding quality labels, which enables the translation of echo images towards a user-defined quality level. To achieve this, we first start with StarGAN [5]. We demonstrate that using the original StarGAN structure to transfer image quality leads to significant reconstruction artifacts. To alleviate this, we present several critical modifications to StarGAN's generator and discriminator networks and introduce Quality Transfer StarGAN (QT-StarGAN), which can produce images that are visually very close to clinically obtained data. For quantitative evaluation, we use echo view classification task as a testbed, and demonstrate that our proposed QT-StarGAN results in significantly improved classification accuracy.

2 Dataset

To collect the gold standard labels, a cardiologist with 30 years clinical experience was given the task to assign the quality assessment and also the view classification labels[1] for a total of 16,612 echo cine series from 3,157 unique patients, each echo series containing 40-70 image frames[2]. For each cine series, all image frames inherit the assigned labels. The quality labels were defined in four coarse quality categories: *Poor* ($0\% \sim 25\%$), *Fair* ($25\% \sim 50\%$), *Good* ($50\% \sim 75\%$), and *Excellent* ($75\% \sim 100\%$), with the percentages reference to the lowest to highest image qualities within each category. Given that our goal is to freely transfer the echo image quality to any level between 0% to 100%,

[1] 14 standard cardiac views: A{2-5}C (4 views), PLAX, RVIF, S{4, 5}C (2 views), IVC, PSAX-{A, M, PM, APEX} (4 views), SUPRA, and due to the extreme similarity, PSAX-M/PSAX-PM views are combined as one class.

[2] This dataset was randomly collected from **** for this project, under the approvals from ****.

we modify the aleatoric uncertainty modelling method proposed by [14, 17] to directly supervise the training of a numerical quality estimation module.

3 Methodology

Let us note the collected dataset as $\mathcal{D} = \{\mathbf{x}_i, q_i, v_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i$ is a 2D echo image, $q_i \in \mathcal{C} = \{c : \textit{Poor}, \ldots, \textit{Excellent}\}$ and $v_i \in \{\text{A2C}, \ldots, \text{SUPRA}\}$ are the corresponding image quality and view classification labels, and $\mathbf{q}_i$ and $\mathbf{v}_i$ are vectors representing the one-hot version of respective label. For simplifying the notations, images $\mathbf{x}_{i-1}$ and $\mathbf{x}_{i+1}$ in the dataset represent the previous and the next frames of $\mathbf{x}_i$ from one echo cine series, where the special cases of boundary frame samples are not discussed.

The QT-StarGAN is constructed by a generator G and a discriminator D, and the network architecture can be view in Fig. 1. An image $\tilde{\mathbf{x}}_i = G(\mathbf{x}_i, t_i)$ is generated by forwarding $\mathbf{x}_i$ and a user-defined target quality level $t_i \in (0, 1]$. The quality estimation branch D_q needs to handle the fine-grained numerical quality level t_i for $\tilde{\mathbf{x}}_i$ and the coarse categorical quality labels q_i for $\mathbf{x}_i$. By following Kendall and Gal [14], our D_q emits two numerical outputs $D_{q^\mu}, D_{q^\sigma} \in (0, 1]$ to model the quality level as a Gaussian distribution $\mathcal{N}(D_{q^\mu}, D_{q^\sigma})$, instead of a typical regression module with only a numerical output. To accept the one-hot categorical target label $\mathbf{q}_i$, we compute the probability distribution $\mathbf{p}_i$ of each coarse quality class using the Gaussian Cumulative Density Function (CDF):

$$\begin{aligned} \text{cdf}(z) &= \frac{1}{2}\big(1 + \text{erf}(\frac{z - D_{q^\mu}(\mathbf{x})}{D_{q^\sigma}(\mathbf{x})\sqrt{2}})\big), \\ p_c^* &= \text{cdf}(u_c) - \text{cdf}(l_c), \text{ and,} \\ p_c &= \frac{p_c^*}{\sum_{c \in C} p_c^*}, \end{aligned} \tag{1}$$

where erf(.) denotes the error function, l_c and u_c are the respective lower and upper bound for c, and the subscript i is dropped for clarity. For training with the the fine-grained numerical value t_i, only the $D_{q^\mu}(\tilde{\mathbf{x}}_i)$ is trained with the mean absolute loss function. Hence, the loss function for D_q is defined as:

$$\ell_q(\mathbf{x}_i) = H(\mathbf{q}_i, \mathbf{p}_i) + |t_i - D_{q^\mu}(\tilde{\mathbf{x}}_i)|, \tag{2}$$

where $H(.,.)$ represents the cross-entropy loss, and $\mathbf{p}_i$ denotes an one-hot vector of the probability distribution $\mathbf{p}_i = [p_{c,i}]_{c \in \mathcal{C}}^{\text{T}}$.

The second job of D is to make sure that appearance of $\tilde{\mathbf{x}}_i$ is close to a real 2D echo image. This is commonly achieved by following the adversarial training [7]: $\ell_{adv} = \log D_{adv}(\mathbf{x}_i) + \log\big(1 - D_{adv}(1 - \tilde{\mathbf{x}}_i)\big)$. The original StarGAN model implements the Wasserstein GAN objective [4] with gradient penalty [8] to improve the training stability, which is unchanged in our network:

$$\ell_{adv}(\mathbf{x}_i) = D_{adv}(\mathbf{x}_i) - D_{adv}(\tilde{\mathbf{x}}_i) - \lambda_{gp}(||\nabla D_{adv}(\hat{\mathbf{x}}_i)||_2 - 1)^2, \tag{3}$$

where D_{adv} represents the second output branch of D, $\hat{\mathbf{x}}_i$ denotes a sample uniformly from a straight line between $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$, and $\lambda_{gp} = 10$ is a weighting factor determined by [5].

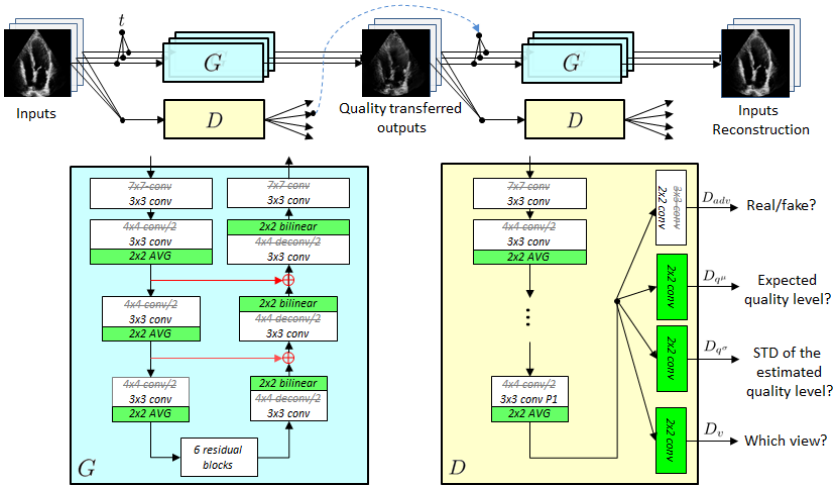


Fig. 1. The network structure of the proposed QT-StarGAN (c3). In general, we strip the down/up-sampling duties from the conv/deconv layers and replaced with explicit average-pooling (AVG) and bilinear interpolation layers. "/#" denotes the stride other than 1. The green blocks represent the added layers; the red connections represent the introduced parameter-free skip connections; and the blue dotted connection represents the use of D estimated quality level of the original images for transferring the generated images back to the original quality level.

Here, we add a regularization loss to D to ensure $\tilde{\mathbf{x}}_i$ can preserve the correct heart anatomical patterns that are necessary to fulfill the view classification task, via a third output branch D_v:

$$\ell_v(\mathbf{x}_i) = H\big(\mathbf{v}_i, D_v(\mathbf{x}_i)\big) + H\big(\mathbf{v}_i, D_v(\tilde{\mathbf{x}}_i)\big). \tag{4}$$

Finally, the generator should be able to re-generate $\mathbf{x}_i$, given $\tilde{\mathbf{x}}_i$ and q_i, following the cycle consistent reconstruction [15, 24]. Since q_i is categorical, we replace q_i with $D_{q_\mu}(\mathbf{x}_i)$ (treated as a constant in the computation graph):

$$\ell_{rec}(\mathbf{x}_i) = |\mathbf{x}_i - G\big(\tilde{\mathbf{x}}_i, D_{q_\mu}(\mathbf{x}_i)\big)|. \tag{5}$$

Therefore, the overall training objective is a combination of the above loss components:

$$\ell(\mathbf{x}_i) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\Big(\lambda_q \ell_q(\mathbf{x}_i) + \ell_{adv}(\mathbf{x}_i) + \lambda_v \ell_v(\mathbf{x}_i) + \lambda_{rec}\ell_{rec}(\mathbf{x}_i)\Big), \tag{6}$$

where we find $\lambda_q = 10$, $\lambda_v = 1$, and $\lambda_{rec} = 10$ yield the optimal performance.

3.1 Network Modifications

In the original StarGAN network, G is in a form of auto-encoder with a series of down-sampling convolution (conv) layers (contraction path), a series of residual

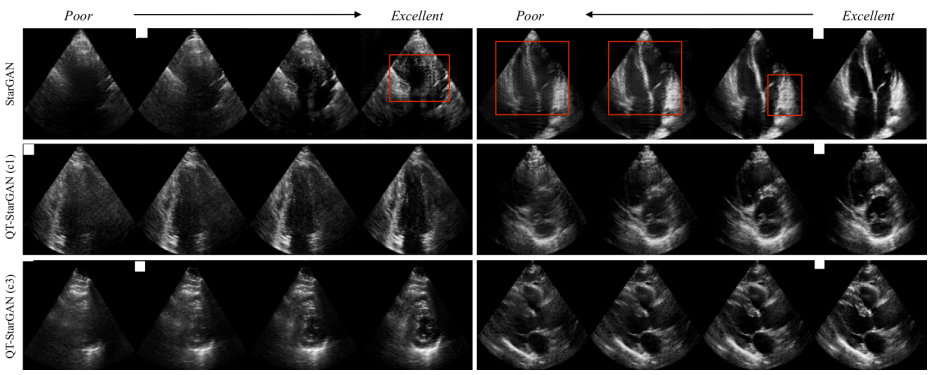


Fig. 2. The quality level altered samples generated by StarGAN, QT-StarGAN (c1), QT-StarGAN (c3), featuring the low quality to high quality transfer on the left panel and high to low on the right panel. The sub-images with a white square on the top-left corner are the original images before the translation.

layers (feature tuning path), and a series of up-sampling deconvolution (deconv) layers (expansion path), and D is a multi-layer conv network.

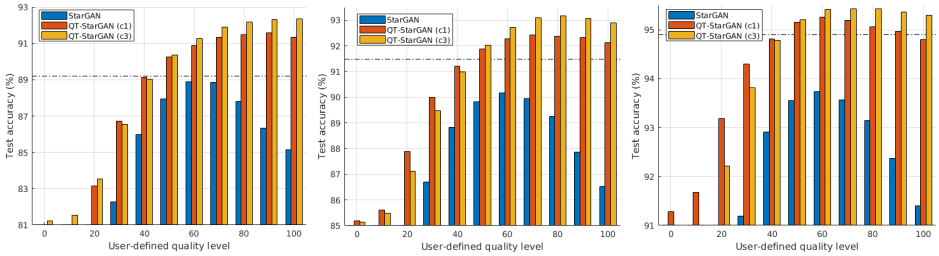
Bilinear Up-sampling: Using the original StarGAN network, we observe that the generated images can contain visually noticeable artifacts (see the first row of Fig. 2). These artifacts have grid appearance patterns, which is a result of using deconv layers with stride 2 to up-sample the spatial size of the feature tensor. Hence, we replaced the deconv layers with bilinear up-sampling layer followed by conv layer with stride 1 during the up-sampling phase.

Skip Connections: Motivated by U-Net [18] and ResNet [9] designs, we added two parameter-free skip-addition links between the contraction path and the expansion path, allowing a chance for the high-frequency information to be utilized in image generation. These modifications have been illustrated in Fig. 1.

Temporal Discriminator: While keeping G on single frame generation, we extend a temporal window of three frames for D to allow temporal information to aid to the optimization of the objectives. In Fig. 2, we show examples of the single-frame and three-frame variations of QT-StarGAN on the second and third rows as QT-StarGAN (c1)/(c3), respectively. This modification is simply an alteration to the input of D to a stack of three frames, noted as $D(\{\mathbf{z}_{i-1}, \mathbf{z}_i, \mathbf{z}_{i+1}\})$ for $\mathbf{z} \in \{\mathbf{x}, \tilde{\mathbf{x}}, \hat{\mathbf{x}}\}$, where the respective task labels remain the same for all three frames.

4 Experiments

For quantitative evaluation, we use a state-of-the-art view classification model proposed by Zhang *et al.* [23] (*i.e.*, a VGG-16 [19] network), a point-of-care mobile ultrasound view classification and quality assessment model [21] (*i.e.*, a combination of a DenseNet [11] with three dense-blocks followed by an LSTM [10] of a 10-frame temporal window, where each dense block has exactly one densely-connected layer), and a standalone DenseNet structure extracted from the above



(a) DenseNet (b) DenseNet+LSTM (c) VGG-16

Fig. 3. The view classification performance on (a) DenseNet, (b) DenseNet+LSTM, and (c) VGG-16, with quality level *altered* test set, and the dotted lines indicate the classification performance of the *original* test set.

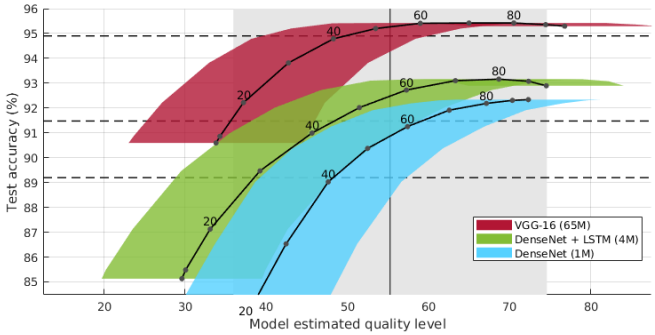


Fig. 4. The quality distribution of the *altered* test set (by QT-StarGAN (c3)) ("# M" indicates the number of model parameters), where the values marked on the center black lines of the colored shapes indicate the user-defined quality levels used to transfer the image quality, and the x-axis coordinate of the black lines indicates the mean *estimated* quality levels (by the view classification models). The width of the colored areas at any point on a center black line indicates the STD of the *altered* test set. The gray background area indicates the mean and STD of the *original* test set.

DenseNet+LSTM model. For training these networks, we uniformly adjusted the input size of these three models to 128×128 to match the image size of QT-StarGAN. The same image quality estimation module D_{q^μ} and D_{q^σ} are also added to estimate image quality level. For statistically analyzing the view classification performance, we trained five instances of QT-StarGANs and each of the view classification models was also repetitively trained five times. For testing, one instance of QT-StarGAN (only the generator is used) is paired with one instance of the view classification model to alter the images with regard to the user-defined quality level.

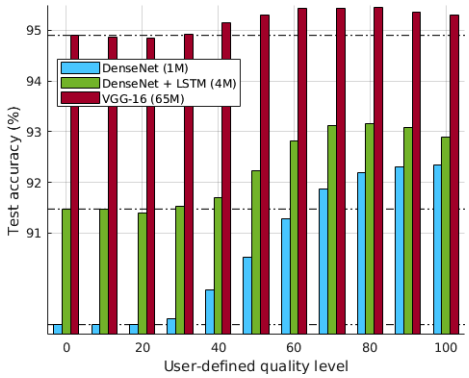


Fig. 5. The performance of the QT-StarGAN (c3) by only transfer the test images that with estimated image quality lower than the user-defined quality level.

5 Results and Discussion

It can be observed in Fig. 3: 1) compared to the off-the-shelf StarGAN, the introduced modifications in QT-StarGAN (c1)/(c3) can lead to significant improvement in view classification accuracy, and it can outperform testing the original test set (the dashed line) when the user-defined quality level is set to above 50%; and, 2) the benefit of the temporal discriminator can be observed with the improved view classification accuracy of c3 *vs.* c1. From Fig. 4, we observe a phenomenon that the estimated quality from the view classification model does not always match the user defined quality level. We suspect that this due to the correlation between the defined echo image quality and ultrasound imaging parameters such as the imaging depth and proper centering. From Fig. 2, we see that QT-StarGAN is able to alter the image contrast by enhancing the anatomical edges and changing the speckle patterns over the quality translation, while the imaging depth and centering are preserved. As a result, QT-StarGAN cannot fully adjust the image quality level to a user-defined value. Intuitively for echo imaging analysis tasks such as view classification, transferring the quality to a lower level does not have any value. In Fig. 5, we only transfer the images with the estimated quality lower than the user-defined level, which improves classification accuracy with statistical significance at above 60% user-defined quality level for DenseNet and DenseNet+LSTM, and at 80% quality for VGG-16.

In summary, we present QT-StarGAN to transfer echo image quality by a user-defined quality level. We demonstrate the effectiveness of this network by testing the echo view classification accuracy, where the classification accuracy can be significantly improved using quality transfer. We anticipate that QT-StarGAN can be used in many other classification tasks to reduce the impact of image quality variability.

References

1. Abdi, A.H., et al.: Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes. In: MICCAI. pp. 302–

310. Springer (2017)
2. Achim, A., et al.: Novel bayesian multiscale method for speckle removal in medical ultrasound images. IEEE TMI **20**(8), 772–783 (2001)
3. Alexander, D.C., et al.: Image quality transfer and applications in diffusion MRI. NeuroImage **152**, 283–298 (2017)
4. Arjovsky, M., et al.: Wasserstein generative adversarial networks. In: ICML. pp. 214–223 (2017)
5. Choi, Y., et al.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE CVPR. pp. 8789–8797 (2018)
6. Coupé, P., et al.: Nonlocal means-based speckle filtering for ultrasound images. IEEE TIP **18**(10), 2221–2229 (2009)
7. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
8. Gulrajani, I., et al.: Improved training of wasserstein gans. In: NIPS. pp. 5767–5777 (2017)
9. He, K., et al.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
11. Huang, G., et al.: Densely connected convolutional networks. In: IEEE CVPR. vol. 1-2, p. 3 (2017)
12. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: IEEE CVPR. pp. 1125–1134 (2017)
13. Kang, E., et al.: A deep CNN using directional wavelets for low-dose X-ray CT reconstruction. Medical Physics **44**(10), 360–375 (2017)
14. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS. pp. 5574–5584 (2017)
15. Kim, T., et al.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML. pp. 1857–1865. JMLR. org (2017)
16. Loizou, C.P., et al.: Quality evaluation of ultrasound imaging in the carotid artery based on normalization and speckle reduction filtering. MBEC **44**(5), 414 (2006)
17. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Neural Networks. vol. 1, pp. 55–60. IEEE (1994)
18. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
19. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR. pp. 1–14 (2015)
20. Tsantis, S., et al.: Multiresolution edge detection using enhanced fuzzy c-means clustering for ultrasound image speckle reduction. Medical Physics **41**(7), 72903–1–11 (2014)
21. Van Woudenberg, N., et al.: Quantitative echocardiography: Real-time quality estimation and view classification implemented on a mobile android device. In: POCUS, pp. 74–81. Springer (2018)
22. Wu, L., et al.: FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Trans. Cyber. **47**(5), 1336–1349 (2017)
23. Zhang, J., et al.: Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation **138**(16), 1623–1635 (2018)
24. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE ICCV. pp. 2223–2232 (2017)