Enhancing Echocardiogram Video Quality via Latent Space Editing

David Choi[1], Milos Vukadinovic[2,3], Bryan He[4], David Ouyang[1,3]

[1] Cedars-Sinai Medical Center, Los Angeles, CA, USA
[2] UCLA Bioengineering, Los Angeles, CA, USA
[3] Kaiser Permanente, Pleasanton, CA, USA
[4] Stanford Computer Science, Palo Alto, CA, USA
david.ouyang@kp.org

Abstract. Echocardiography (echo), or cardiac ultrasound, is the most widely used imaging modality for cardiac form and function due to its relatively low cost, rapid acquisition time, and non-invasive nature. However, ultrasound acquisitions are often limited by artifacts, noise, and low-quality acquisitions that hinder diagnostic interpretation. Existing techniques for enhancing echos consist of traditional filter-based algorithms, deep-learning approaches developed on radiofrequency (RF) data, or approaches that have strong priors such as manual segmentation labels which limits both clinical applicability and scalability. To address these limitations, we introduce a data-driven approach for enhancing echo videos using a generative model trained on historical images. We learn a latent space representation for echo images using a generative model and use self-supervision from a synthetic dataset of high quality (HQ) and simulated low quality (LQ) image pairs to estimate a direction vector in the latent space from the LQ to HQ domain. In both held-out internal and external test sets, our approach resulted in echo videos with higher gCNR (0.60-0.62 vs. 0.48-0.53) and quality score (P_{hq}) (0.99-0.99 vs. 0.92-0.96) compared to original LQ videos. Furthermore, we leverage previously developed models for echo to show preservation of key clinical characteristics such as LVEF (MAE 4.74-6.82) and left ventricle segmentation (Dice 0.92-0.93), suggesting potential for future clinical use to improve the quality of echo videos.

Keywords: Video enhancement, Generative Adversarial Networks, Echocardiography.

1 Introduction

Transthoracic echocardiography (TTE) is a non-invasive imaging modality used for monitoring cardiac function and diagnosing abnormalities. In many cases, TTE is the first and most common option for screening heart diseases such as endocarditis, valvular heart disease, etc. Unfortunately, TTE acquisitions are prone to a form of noise referred to as acoustic clutter caused by reverberations in echogenic structures such as subcutaneous fat, bones, and lungs blocking the propagation of ultrasound waves [1].

This can result in low contrast between cardiac chambers and tissue in B-mode echo which pose challenges for clinical diagnosis [2], measurement reliability [3], and post-formation image processing [4]. Altogether, these challenges motivate the development of techniques for enhancing LQ echo videos.

Recent advances in generative modeling have allowed models to learn complex and high-dimensional distributions of data. This has led to data-driven methods achieving greater capabilities for domain translation and inverse problems in medical imaging [5-7] by leveraging generative models e.g. Generative Adversarial Networks (GANs) [8] as data priors. Promisingly, developments in GAN architectures and techniques e.g. StyleGAN2 [9-10] have demonstrated the ability to edit real videos [11-13] by learning a well-structured latent space of data, inverting data into latent embeddings, and interpolating in semantically meaningful directions. These techniques have been applied to medical imaging for video generation, physiological guidance, and super-resolution of cardiac MRIs [14]; however, their efficacy in echo remains unexplored.

To solve the problem of enhancing LQ echo videos, we adopt the StyleGAN2 architecture to learn a latent space representation of echo images and leverage it to encode the underlying relationship between LQ and HQ images. Then, we interpret the problem of enhancing LQ echo videos as a domain translation problem and estimate a direction vector from the LQ to HQ domain in the latent space by using a synthetic dataset of paired HQ and simulated LQ image pairs. Lastly, we use this estimated direction vector to edit LQ videos into HQ videos. We demonstrate that our proposed method is capable of enhancing LQ echo videos in both held-out internal and external experiments.

Related Works. In comparison to CT and MRI, the application of deep learning to ultrasound enhancement is understudied mainly due to the infeasibility of collecting LQ and HQ image pairs, difficulty of developing a statistical model for noise, and intrinsic noisiness of acquisitions. Nonetheless, some works have demonstrated successful use of generative models to tackle the problem. In [15], Stevens *et al.* trained score-based diffusion models [16] to model tissue and noise separately on RF data and implemented joint posterior sampling to separate cardiac tissue from noise. Despite successful outcomes, RF data is often discarded in clinical practice due to its lack of clinical relevance and large memory footprint. Therefore, the application of this approach to existing acquisitions is limited. On the other hand, Escobar *et al.* [17] implemented a variation of a CycleGAN [18] to translate LQ apical-view echo images to HQ images, enforcing anatomical consistency by incorporating manual segmentation masks into the training process. However, the requirement of manual labels for training limits the scalability of this approach to different standard views and other larger datasets.
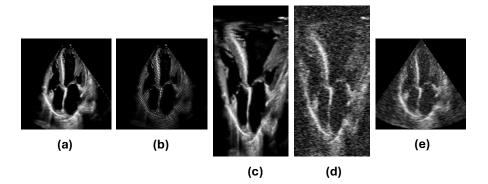
Contributions. In this article, we present a scalable, data-driven method for enhancing B-mode echo videos using a generative model and self-supervision from a synthetic dataset of paired HQ and simulated LQ images. Furthermore, we propose automatable metrics for measuring the quality and consistency of echo videos to facilitate further

research into enhancement methods for echo, leveraging domain-specific tasks and image quality metrics. We make code and weights available at https://github.com/echonet/image_quality.

2 Methods

Datasets. For our internal dataset, we collect 25,212 apical 4-chamber view (A4C) echo videos across 21,327 subjects from Cedars-Sinai Medical Center. To train the StyleGAN2 model, we extract 2,620,978 frames from videos in the entire dataset. For all other purposes, we split 25,117 videos into the training set, 45 to the validation set, and 40 to the test set. For our external test set, we collect 40 A4C echo videos in the publicly available MIMIC-IV-ECHO [19-22] dataset. For validation and test sets, we collect videos from studies with LQ indications in sonographer reports and truncate videos to 32 frames. We downsize the resolution of all videos to 256×256.

Models. We adopt the StyleGAN2 architecture and training configurations specified in [10]. After training, we only use the synthesis component $\mathbb{G}_s : \mathbb{W}^+ \to \mathbb{X}$ of the generator where $\mathbb{W}^+ \subseteq \mathbb{R}^{14 \times 512}$ is the learned latent space and $\mathbb{X} \subseteq \mathbb{R}^{H \times W \times C}$ is the image space. For inversion, we adopt the e4e encoder architecture and training configurations specified in [11] to learn a projection $\mathcal{E} : \mathbb{X} \to \mathbb{W}^+$ from images to latent embeddings which has been shown to promote temporally consistent editing [12]. To ensure near-identical reconstruction of video frames from latent embeddings, we use a technique known as pivotal tuning inversion (PTI) [13] to finetune a generator $\mathbb{G}_s'$ for each video during inference to be able to reconstruct frames $\{x_i\}_{i=1}^N$ from latent embeddings $\{\mathcal{E}(x_i)\}_{i=1}^N$.

Fig. 1. Steps to obtain a LQ simulation. (a) Original HQ image (b) Step 1: Radial-polar sampling with M=224, N=100 (c) Step 2: Rectification (d) Step 3: Gaussian noise corruption with $\sigma = 0.2$ (e) Step 4: Bilinear interpolation
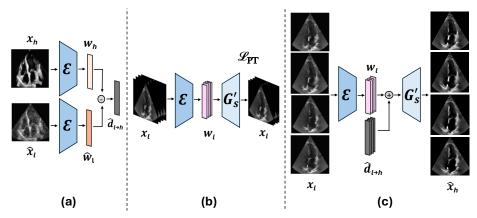
Synthetic Dataset Curation. To estimate a direction vector $\hat{d}_{l \to h} \in \mathbb{W}^+$ from the LQ to HQ domain, we curate a minimal dataset of HQ and simulated LQ image pairs ($P =$

40) for supervision. Using a minimal dataset reduces the amount of manual labeling and has shown sufficiency for mitigating bias [23].

Given a HQ image $x_h \in \mathbb{X}$, we simulate anisotropic noise in LQ acquisitions based on the framework defined in [24] with steps illustrated in Figure 1. First, we use radial-polar sampling to sample M points along N radial lines in the ultrasound sector, where M and N determine the axial and lateral resolution. Then, the sampled points are rearranged into an $M \times N$ rectangular grid $I_H \in \mathbb{R}^{M \times N}$ for image processing. Next, I_H is corrupted with Gaussian noise i.e.

$$\hat{I}_L = I_H + n \tag{1}$$

where $n \sim \mathcal{N}(0, \sigma) \in \mathbb{R}^{M \times N}$. Lastly, coordinates in $\hat{I}_L$ are remapped to the ultrasound sector using bilinear interpolation resulting in a LQ estimate $\hat{x}_l$ of x_h.

Fig. 2. Overview of our proposed method. (a) Estimation of edit direction from LQ to HQ domain (b) Fine-tuning of the StyleGAN2 synthesis network (c) Latent space editing using the estimated direction from LQ to HQ domain

Edit Direction. Given a curated dataset $X = \{(x_{h,i}, \hat{x}_{l,i})\}_{i=1}^{P}$ of HQ and simulated LQ image pairs, we project it to latent embedding pairs $W = \{(w_{h,i}, \hat{w}_{l,i})\}_{i=1}^{P} = \mathcal{E}(X)$. To estimate $\hat{d}_{l \to h}$, we compute the mean of the direction vectors corresponding to each embedding pair $(w_{h,i}, \hat{w}_{l,i})$ i.e.

$$\hat{d}_{l \to h} = \frac{1}{P} \sum_{i=1}^{P} (w_{h,i} - \hat{w}_{l,i}) \tag{2}$$

Now, our goal is to edit frames of a LQ video $x_l = \{x_{l,i}\}_{i=1}^{N}$ to a HQ version $\hat{x}_h = \{\hat{x}_{h,i}\}_{i=1}^{N}$. To accomplish this, we project x_l to latent embeddings $w_l = \{w_{l,i}\}_{i=1}^{N} = \mathcal{E}(x_l)$ and interpolate in the direction $\hat{d}_{l \to h}$ i.e.

$$\hat{w}_h = \{\hat{w}_{h,i}\}_{i=1}^{N} = w_l + \alpha \cdot \hat{d}_{l \to h} \tag{3}$$

where α is a scalar that controls the edit level. Then, we project $\hat{w}_h$ to corresponding images $\hat{x}_h = \{\hat{x}_{h,i}\}_{i=1}^N = \mathbb{G}'_s(\hat{w}_h)$ which gives us a HQ version of the original LQ video x_l. An illustration of the overall process is shown in Figure 2.
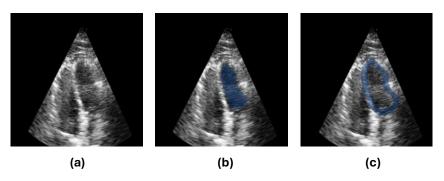
3 Experiments

Fig. 3. Overview of considered regions for gCNR metric. (a) Original image (b) LV chamber region (c) LV myocardium region

Metrics. The core objective of video enhancement for echo is (1) to increase the contrast between cardiac chambers and tissue and (2) to preserve important clinical attributes relating to cardiac anatomy and motion.

To measure contrast in video frames, we compute the mean of generalized contrast-to-noise ratios (gCNR) [25] which is an unsupervised image quality metric widely used in medical imaging. Simply stated, gCNR interprets the area of the overlapping region between the pixel densities of two regions of interest (ROIs) as a measure of contrast. For evaluating A4C echo, we choose the left ventricular (LV) chamber and myocardium as the ROIs. We leverage EchoNet-Dynamic (EchoNet) [26] to automate the segmentations for the ROIs. By default, EchoNet provides a segmentation of the LV chamber. For the LV myocardium, we dilate the segmentation of the LV chamber and remove the original segmentation portion. An example of the considered regions is shown in Figure 3. In addition, we compute the mean of video quality scores (P_{hq}) given by a pretrained R(2+1)D [27] network developed for binary classification of HQ videos.

To measure the consistency of cardiac anatomy and motion, we use clinically important measurements commonly used to assess cardiac function, namely LV segmentation and LV ejection fraction (LVEF). For anatomical consistency, we compute the mean of Dice scores between LV segmentations in the original LQ and enhanced HQ video frames traced by EchoNet. For motion consistency, we compute the mean absolute error (MAE) between LVEF measurements of the original LQ and enhanced HQ videos predicted by EchoNet.
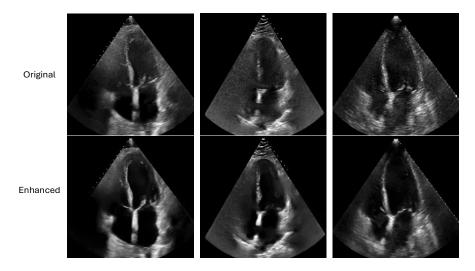
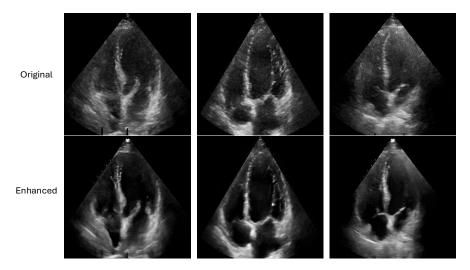**Table 1.** Quantitative results for internal test set. Mean estimate and 95% confidence interval

	$\mu\ gCNR \uparrow$	$\mu\ P_{hq} \uparrow$	*MAE LVEF* $\downarrow$	$\mu\ Dice \uparrow$
Ours (synthetic direction)	**0.62** (0.60-0.64)	**0.99** (0.99-0.99)	**4.74** (3.86-5.80)	**0.93** (0.92-0.94)
Ours (real direction)	0.56 (0.54-0.59)	0.99 (0.99-0.99)	6.72 (5.11-8.56)	0.79 (0.77-0.80)
LQ	0.53 (0.51-0.56)	0.96 (0.94-0.99)	—	—

Table 2. Quantitative results for external test set. Mean estimate and 95% confidence interval.

	$\mu\ gCNR \uparrow$	$\mu\ P_{hq} \uparrow$	*MAE LVEF* $\downarrow$	$\mu\ Dice \uparrow$
Ours (synthetic direction)	**0.60** (0.57-0.62)	**0.99** (0.99-0.99)	**6.82** (5.63-8.32)	**0.92** (0.91-0.93)
Ours (real direction)	0.51 (0.49-0.53)	0.99 (0.99-0.99)	10.8 (8.64-13.1)	0.78 (0.77-0.80)
LQ	0.48 (0.46-0.50)	0.92 (0.88-0.95)	—	—

Fig. 4. Qualitative results for internal test set. Top row depicts frames from original sonographer-labeled LQ videos. Bottom row is the HQ enhancement from our method.

Original

Enhanced

Fig. 5. Qualitative results for external test set. Top row depicts frames from original sonographer-labeled LQ videos. Bottom row is the HQ enhancement from our method.

Performance Evaluation. Tables 1 and 2 show a quantitative comparison of original LQ videos and enhanced HQ videos from our method using an edit direction computed from a synthetic dataset of paired HQ and simulated LQ images. In both internal and external experiments, we demonstrate that our enhanced HQ videos (gCNR 0.60-0.62) achieve finer contrast between the LV chamber and myocardium in comparison to the original LQ videos (gCNR 0.48-0.53). This is further supported by a comparison of the quality scores of our enhanced HQ videos (P_{hq} 0.99-0.99) and original LQ videos (P_{hq} 0.92-0.96). Moreover, the improvement in quality is consistent with the qualitative comparisons provided in Figures 4 and 5. A close observation of the frames in our enhanced HQ videos shows that our method is capable of removing unwanted noise in both the atrial and ventricular chambers of the depicted heart which is a difficult problem by conventional means. In addition, consistency metrics indicate that our approach performs reasonably well in preserving cardiac anatomy in the form of LV segmentation (Dice 0.92-0.93) and motion in the form of LVEF (MAE 4.74-6.82).
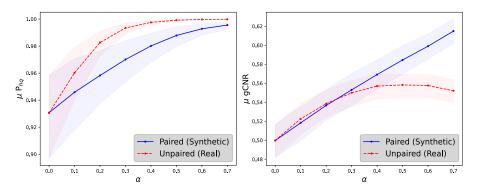
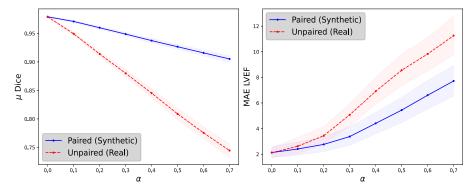**Fig. 6.** Quality metrics on validation set. Mean estimate and 95% confidence interval at different edit levels.

Fig. 7. Consistency metrics on validation set. Mean estimate and 95% confidence interval at different edit levels.

Ablation Study. We further perform an ablation study to demonstrate that the edit direction computed from a synthetic dataset of paired HQ and simulated LQ images achieves superior quality and preservation of clinical attributes than a real dataset of physician-labeled unpaired LQ and HQ images. Figure 6 shows that the quality score along both directions is comparable with similar converging behaviors. However, the gCNR begins to diverge at $\alpha = 0.3$ with deterioration in the real direction and continuous improvement in the synthetic direction, suggesting that the synthetic direction yields greater contrast between cardiac chambers and tissue. Furthermore, Figure 7 shows a drastic improvement in the preservation of LVEF and LV structure in the synthetic direction in comparison to the real direction which is crucial for clinical safety. This follows from intuition because a paired dataset functionally mitigates biases and confounding factors that can obscure the true relationship between LQ and HQ images. Considering the continuous decrease in anatomical consistency with increasing α, the optimal α for the synthetic direction is determined to be the minimum value that shows a noticeable separation in gCNR which is observed at $\alpha = 0.5$.

4   Conclusion

In this article, we introduced a scalable, data-driven approach for enhancing the quality of echo videos without the use of strong priors in the training process. We develop a deep learning approach to remove unwanted noise in cardiac chambers in B-mode echo videos which is a challenging and understudied problem that has drastic implications for both clinical and computational use. Currently, our proposed approach is limited by inconsistencies in cardiac anatomy and motion in enhanced videos which poses an issue for clinical safety. Yet, the consistency metrics of our approach are reasonable which suggests potential for further improvements. A promising avenue for future exploration involves the use of advanced generative models such as StyleGAN3 [28] and more controllable latent spaces such as StyleSpace [29] which has solved many problems associated with generative modeling in real world applications e.g. texture sticking. Nonetheless, we believe this to be a step forward in achieving controllable and scalable enhancement in medical imaging, contributing to a more equitable healthcare and research environment with implications of increasing data availability and improving diagnostics without relying on advanced equipment.

References

1. Fatemi, A., Berg, E.A., Rodriguez-Morales, A.: Studying the origin of reverberation clutter in echocardiography: In vitro experiments and in vivo demonstrations. Ultrasound in Medicine and Biology **45**(7), 1799–1813 (2019)
2. Chirillo, F., Pedrocco, A., De Leo, A., et al.: Impact of harmonic imaging on transthoracic echocardiographic identification of infective endocarditis and its complications. Heart **91**(3), 329–333 (2005)
3. Hoffman, R., Lethen, H., Marwick, T., Arnese, M., Fioretti, P., Pingitore, A., Picano, E., Buck, T., Erbel, R., Flachskampf, F., Hanrath, P.: Analysis of interinstitutional observer agreement in interpretation of Dobutamine stress echocardiograms. Journal of the American College of Cardiology **27**(2), 330–336 (1996)
4. Tenbrinck, D., Sawatzky, A., Burger, M., Haffner, W., Willems, P., Paul, M., Stypmann, J.: Impact of physical noise modeling on image segmentation in echocardiography. In: Proceedings of the 3rd Eurographics Workshop on Visual Computing in Biology and Medicine, pp. 33–40. (2012)
5. Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P.L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., Firmin, D.: DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Transactions on Medical Imaging **37**(6), 1310–1321 (2018)
6. You, C., Cong, W., Vannier, M.W., Saha, P.K., Hoffman, E.A., Wang, G., Li, G., Zhang, Y., Zhang, X., Shan, H., Li, M., Ju, S., Zhao, Z., Zhang, Z.: CT super-resolution gan constrained by the identical, residual, and Cycle Learning Ensemble (GAN-circle). IEEE Transactions on Medical Imaging **39**(1), 188–203 (2020).
7. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated MRI. Medical Image Analysis **80**, 102479 (2022)

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680. (2014)

9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4396–4405. (2019)

10. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8107–8116. (2020)

11. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics **40**(4), 1–14 (2021)

12. Tzaban, R., Mokady, R., Gal, R., Bermano, A., Cohen-Or, D.: Stitch it in time: Gan-based facial editing of Real videos. In: SIGGRAPH Asia 2022 Conference Papers, pp. 1–9. (2022)

13. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on Graphics **42**(1), 1–13 (2022)

14. Vukadinovic, M., Kwan, A.C., Li, D., Ouyang, D.: GANcMRI: Cardiac magnetic resonance video generation and physiologic guidance using latent space prompting. In: Proceedings of Machine Learning Research, pp. 594–606. (2023)

15. Stevens, T.S., Meral, F.C., Yu, J., Apostolakis, I.Z., Robert, J.-L., van Sloun, R.J.: Dehazing ultrasound using diffusion models. IEEE Transactions on Medical Imaging **43**(10), 3546–3558 (2024)

16. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 11918–11930. (2019)

17. Escobar, M., Castillo, A., Romero, A., Arbelaez, P.: UltraGAN: Ultrasound enhancement through adversarial generation. In: Burgos, N., Svoboda, D., Wolterink, J.M., Zhao, C. (eds.) SASHIMI 2020, LNCS, vol. 12417, pp. 120–130. Springer, Cham (2020)

18. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251. (2017)

19. Gow, B., Pollard, T., Greenbaum, N., Moody, B., Johnson, A., Herbst, E., Waks, J.W., Eslami, P., Chaudhari, A., Carbonati, T., Berkowitz, S., Mark, R., Horng, S.: MIMIC-IV-ECHO: Echocardiogram matched subset (version 0.1). PhysioNet (2023). https://doi.org/10.13026/ef48-v217.

20. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV (version 2.2). PhysioNet (2023). https://doi.org/10.13026/6mm1-ek67.

21. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.H., Celi, L.A., Mark, R.G.: Mimic-IV, a freely accessible electronic health record dataset. Scientific Data **10**, 1 (2023)

22. Goldberger, A., Amaral, L., Glass, L., Haussdorff, J., Ivanov, P.C., Mark, R., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23). 215–220 (2000)

23. Parihar, R., Dhiman, A., Karmali, T., R, V.: Everything is there in latent space: Attribute editing and attribute style manipulation by stylegan latent space exploration. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1828–1836. (2022)

24. Singh, P., Mukundan, R., de Ryke, R.: Synthetic models of ultrasound image formation for speckle noise simulation and analysis. In: 2017 International Conference on Signals and Systems (ICSigSys), pp. 278–284. (2017)

25. Rodriguez-Molares, A., Hoel Rindal, O.M., D'hooge, J., Masoy, S.-E., Austeng, A., Torp, H.: The generalized contrast-to-noise ratio. 2018 IEEE International Ultrasonics Symposium (IUS), 1–4 (2018)
26. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**, 252–256 (2020)
27. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatio-temporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6450–6459. (2018).
28. Karras, T., Aittala, M., Laine, S., Harkonen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 852–863. (2021)
29. Wu, Z., Lischinski, D., Shechtman, E.: StyleSpace analysis: Disentangled controls for Style-gan image generation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12858–12867. (2021)