cAAE-EchoGAN: A Conditional Adversarial Autoencoder Framework for Joint Echocardiographic Image and Segmentation Mask Generation

Shivani Singh, Ashutosh Kumar Singh, and Shanti Chandra
Department of Electronics and Communications Engineering
Indian Institute of Information Technology Allahabad
Prayagraj, India
{rse2024003, ashutosh_singh , shantichandra}@iiita.ac.in

***Abstract*—The field of automated medical diagnostics is largely driven by deep learning models. The performance of these models depends on the availability of large, well-annotated datasets. The acquisition of such datasets remains challenging due to privacy concerns, acquisition variability, annotation costs, and limited availability of healthcare experts. The proposed work aims to address this challenge by "$cAAE-EchoGAN$", a novel conditional adversarial autoencoder-based hybrid GAN capable of generating realistic echocardiographic images along with their corresponding multi-structure segmentation masks (LV_{Endo}, LV_{Epi}, and LA). It is guided by clinical attributes such as the cardiac phase (ED and ES) and imaging views ($2CH$ and $4CH$). The architecture uses a dual-discriminator setup (image discriminator and latent discriminator) to avoid mode collapse and for generating diversified samples. It achieves a fréchet radiomics distance (FRD) of 10.50 and ensures that the generated samples closely follow the distribution of the real echocardiographic data while maintaining the consistency of the latent space.**

***Index Terms*—Medical imaging, GAN, Fréchet Radiomics Distance, Echocardiograms, Annotated Synthetic Data.**

I. Introduction

Echocardiograms are widely used imaging modality for assessing heart function because they are safe, non-invasive, radiation-free and cost-effective [1]. It produces high-quality images of the heart structures, which are useful for the assessment of the heart's anatomical structures and functional performance. The acquisition and interpretation of echocardiograms is very challenging and is subject to observer's bias [2]. Different observers have varied interpretations, which can lead to varied clinical assessments and diagnosis. The limited availability of clinical healthcare professionals further adds to this challenge. With the increasing number of cardiac abnormalities, there is a need of automated methods for the diagnosis and downstream tasks such as view classification [3], segmentation [4], disease identification and classification [5]. However, the performances of these automated models heavily depends on a large amount of well-annotated training data. In the field of medicine, such datasets are often scarce, privacy concerns, time consuming nature of manual annotation and the limited availability of experts. Kumar et al.[6] proposed MultiResUNet for automated segmentation of left ventricles. It uses GAN for generating echocardiographic images with masks for synthetic data augmentation. However, the t-SNE visualization hinted a potential mode collapse. The Denoising Diffusion Probabilistic Model (DDPM) was used by Stojanovski et al. [7] to generate synthetic echocardiographic images by taking semantic labels as guidance. However they chose to train their model using only ED frames which limits the diversity of the generated data. Apart from heart, Fatima et al. [8] proposed a supervised autoencoder generative adversarial network (SA-GAN) for generating synthetic lung ultrasound (LUS) images. Their aim was to remove the problem of class imbalance for the severe disease patterns that are underrepresented in the dataset. It integrates autoencoder with GANs, which removes flaws like mode collapse. However, the training process is computationally expensive produces low-resolution and blurry images.

Oladokun et al. [9] compares two generative methods CycleGAN and Contrastive Unpaired Translation (CUT), for generating synthetic transesophageal echocardiography (TEE) images using 3D anatomical heart models.The study reports that CycleGAN generates more realistic images for the downstream LV segmentation task than CUT. The Fréchet inception distance (FID) for CUT was 188, which is lower than that of CycleGAN, which is contradictory to the experimental verfication and hence, they concluded that the human realism judgement is a better approach than the FID metric. As, it may not be a reliable metric for evaluating medical images like TEE. An innovative framework is proposed by Hangaragi et al. [10], which uses a convolutional autoencoder $CAESynthImgGen$ to produce synthetic CT scan images of lungs. The inherent limitation of the model was its inability to accurately reconstruct the normal class of lung CT scans, which exhibit high intra-class variance. Ashrafian et al. [11] proposed diffusion-based vision-language models to generate high-quality synthetic echocardiography images. It integrates textual prompts and semantic label maps to guide the image generation process. They concluded that the training of the diffusion model is a time-consuming and resource-intensive process. Yang et al. [12] developed a 3D generative model called CAE-ACGAN which generates a multi-contrast MR images which is often expensive or inaccessible for certain patients. It uses one-to-many cross-domain approach to synthesise the gain by merging the architectural strength of VAE and GANs. Reynaud et al. [13] proposed Echoflow, a systematic framework to generate high-quality, privacy-preserving synthetic echocardiogram images and videos. It uses an adversarial variational autoencoder (A-VAE) that defines a domain-specific latent representation. It also has a latent image flow matching (LIFM) model for generating accurate latent echocardiogram frames and a latent re-identification (ReId) module that acts as a privacy filter by rejecting anatomically similar images. They concluded that there was a clear trade-off between image quality and inference speed. Although, there is significant progress in the field of synthetic medical image generation, the existing GAN-based approaches suffer from mode collapse, limited sample diversity and insufficient latent space regularisation, which can adversely affect the quality and diversity of generated samples. Also, the recent diffusion-based models have demonstrated improved image diversity, but their iterative denoising process results in high computational complexity, long training times, and increased inference costs, limiting their practicality for large-scale dataset generation. To address these limitations, cAAE-EchoGAN is proposed. It incorporates latent-space regularisation through adversarial prior matching, promoting sample diversity and stable training. Unlike diffusion-based approaches, it generates samples in a single forward pass, significantly reducing computational overhead. Furthermore, the framework enables controlled synthesis of anatomically consistent image-mask pairs through clinical conditioning based on cardiac phase and imaging view. The proposed methodology and its distinguishing features are summarised through the following key contributions:

1) **Conditional adversarial autoencoder-based framework:** A novel architecture cAAE-EchoGAN, is proposed for generating echocardiographic images with their corresponding multi-structure segmentation mask.
2) **Condition-guided latent representation with dual discriminators:** A conditioned latent space (based on views and phases) is designed to generate diverse echocardiograms. The framework uses an image discriminator and a latent discriminator to enforce realism in the image space and regularization in the latent space, thereby improving diversity and addressing limitations of existing approaches.
3) **High-quality conditioned synthesis for downstream tasks:** It generates high-quality echocardiographic images across different conditions, with realism validated using Fréchet Radiomics Distance (FRD).

II. Proposed Methodology

This work presents a conditional generative framework for the synthesis of echocardiographic images along with their corresponding multi-structure segmentation masks. The aim is to generate high-quality and diverse dataset conditioned on clinically relevant attributes such as cardiac views and phase. The proposed framework integrates a conditional adversarial autoencoder architecture with image and latent discriminators to enforce visual realism. Furthermore, the quality of the generated images is assessed by different metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and Fréchet Radiomics Distance (FRD) [14].

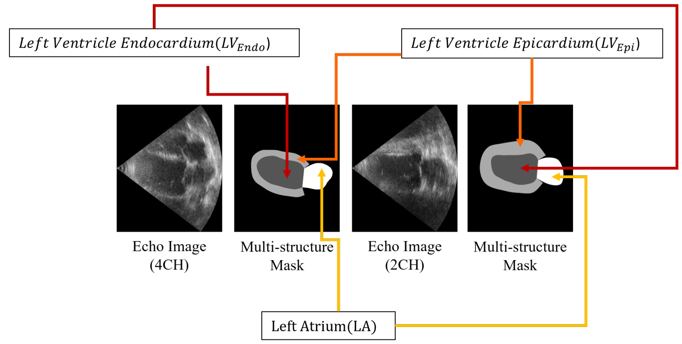

Fig. 1. Sample echocardiographic images from the CAMUS Dataset showing 4CH and 2CH views with corresponding segmentation masks for LV endocardium, LV epicardium, and Left atrium.

A. Acquisition of Dataset and it's Pre-processing

The CAMUS dataset [15] is used for the study in such a way that all the echocardiographic images and their multistructure masks as shown in Fig.1, are saved in paired form for the training of the model. The data consists of apical two-chamber (2CH) and four-chamber (4CH) views acquired at two key cardiac phases, namely end-diastole (ED) and end-systole (ES), representing distinct physiological states of the cardiac cycle. A condition vector 'z' is constructed to encode the cardiac phases and imaging views to generate diverse and medically reliable images. A total of 500 patients' images were used for this work. Out of 500 patients, 400 patients data was used for training and rest for testing. This patient-wise split ensures that images from the same patient do not appear in both training and testing sets.

B. Model Architecture and Training

The proposed framework is based on a Conditional Generative Adversarial Autoencoder (cAAE) designed to generate paired echocardiographic images and segmentation masks. It comprises of four major components: **Encoder:** The encoder receives the concatenated echocardiographic image, one-hot encoded segmentation mask ($M_{oh} \in R^{4\times256\times256}$), and condition vector ($c \in R^4$). It consists of four convolutional blocks with kernel size 4×4 and stride 2, that extracts hierarchical features and maps them into a compact latent representation $z_{enc} \in R^{256}$. To improve sample diversity and enable stochastic generation, Gaussian noise is injected into the latent space prior to decoding.

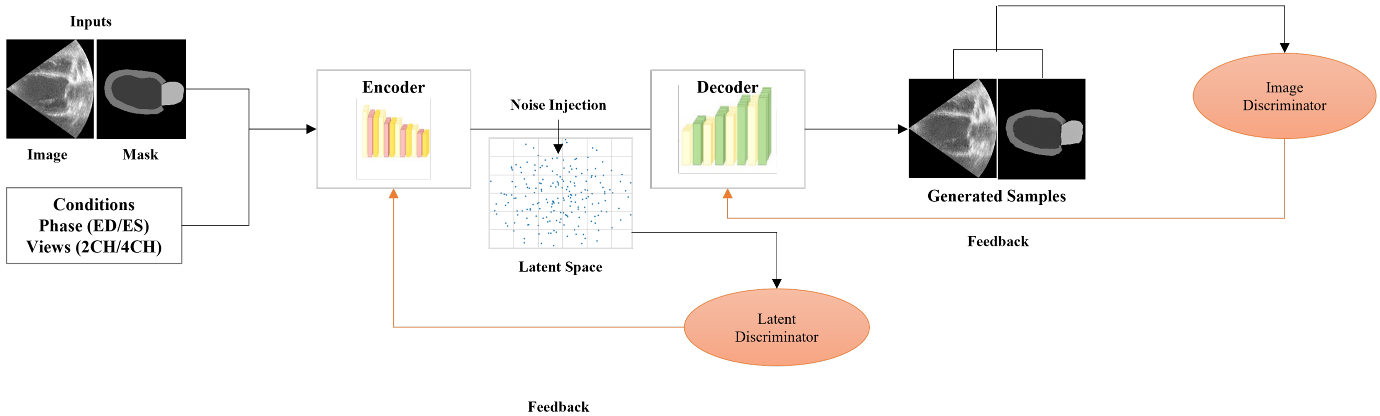

Fig. 2. Architectural setup of the proposed cAAE-EchoGAN

Decoder: The decoder reconstructs image-mask pairs from the noise-perturbed latent representation conditioned on the clinical attributes. The latent vector is concatenated with the condition vector and projected through a fully connected layer into a high-dimensional feature map. Subsequently, a sequence of transposed convolutional layers with batch normalization and ReLU activations progressively upsamples the features from 16×16 to 256×256. Finally, separate output heads generate the grayscale echocardiographic image-mask pairs.
Latent Discriminator: The latent discriminator consists of two fully connected layers that classify whether a latent vector originates from the encoder or a gaussian prior distribution. By enforcing adversarial prior matching in the 256-dimensional latent space, it regularizes the latent representations, promotes smooth latent interpolation, and improves sample diversity while reducing mode collapse.
Image Discriminator: The image discriminator receives the concatenated image-mask pairs as a 5-channel input. It comprises four convolutional layers with leaky ReLU activations and progressively extracts multi-scale features to distinguish real and fake pairs, thereby enforcing image realism, anatomical consistency, and distributional fidelity.
Further, the model is trained for 200 epochs using hybrid loss function taking the combination of reconstruction, segmentation, and adversarial losses to ensure accurate, realistic, and diverse generation of echocardiographic images and masks. The overall generator loss is defined as:

$$\mathcal{L}_G = 0.5\,\mathcal{L}_{adv}^{img} + 3\,\mathcal{L}_{rec} + 5\,\mathcal{L}_{seg} + 0.1\,\mathcal{L}_{adv}^{lat}. \tag{1}$$

where $\mathcal{L}_{adv}^{img}$ denotes the image adversarial loss, $\mathcal{L}_{rec}$ represents the image reconstruction loss, $\mathcal{L}_{seg}$ corresponds to the segmentation loss, and $\mathcal{L}_{adv}^{lat}$ denotes the latent-space adversarial loss. Since the proposed framework jointly generates an echocardiographic image and its corresponding anatomical mask, the image adversarial loss is computed on the generated image-mask pair. The reconstruction loss is computed using the L1 norm between the real and synthesized echocardiographic images as shown:

$$\mathcal{L}_{rec} = \|I_{real} - I_{syn}\|_1 \tag{2}$$

Furthermore, the latent discriminator regularises the latent space by encouraging the encoded representations to match the target prior distribution. The image discriminator is trained to distinguish real image-mask pairs from synthesized pairs.

$$\mathcal{L}_D = \mathcal{L}_D^{img} + \mathcal{L}_D^{lat}. \tag{3}$$

The model uses the Adam optimizer with carefully selected hyperparameters, including a learning rate of 2×10^{-4} for the generator and 1×10^{-5} for the discriminators, along with momentum parameters and batch size of 8 for stable convergence. The performance of the model is is evaluated using multiple quantitative metrics and its computational complexity. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM),Jensen-Shannon Divergence (JSD) and most importantly the Fréchet Radiomics Distance (FRD).It measures the distance between the radiomic feature distributions extracted from real and synthetic images as mentioned in Eq.4.Unlike standard perceptual metrics based on natural images, FRD utilizes clinically interpretable and anatomically relevant radiomic features. It strongly correlates with expert radiologist assessments of image quality, and maintains stability even on small datasets.[14]

$$FRD(X_r, X_g) = \|\mu_r - \mu_g\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}\right) \tag{4}$$

Where, X_r and X_g are the set of real and generated images. The term (μ_r, Σ_r) and (μ_g, Σ_g) represent the empirical mean and covariance of the radiomic features extracted from real and generated images, respectively. It is a dimensionless metric as it operates on normalized radiomic feature distributions. Lower FRD values indicate a closer alignment between the statistical properties of real and generated images in the radiomic feature space.

III. Results and Discussion

This section presents the qualitative and quantitative analysis of the proposed cAAE-EchoGAN framework. The training curve of the generator and discriminator losses is shown in Fig 3, trained over 200 epochs. It presents that the hybrid loss function effectively mitigates common issues such as mode collapse and unstable convergence. The fluctuations in the curve between 20 to 60 epochs shows the adversarial training setup between the generator and discriminator, where both networks try to outperform each other. After this, the loss gradually stabilises which indicates the convergence of the adversarial setup and balanced training.

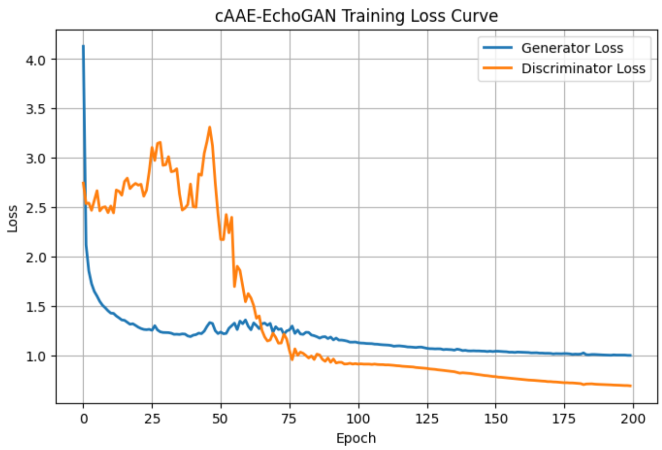

Fig. 3. Generator and discriminator loss curves over training epochs

This stable convergence is further reflected in the T-distributed stochastic neighbour embedding (t-SNE) plot shown in Fig.4. It shows that the real and synthetic markers for any given condition exhibit deep, continuous spatial overlap rather than forming isolated clusters. This confirms that the synthetic cohorts closely replicate the true clinical variance without a significant distributional shift. The visual boundaries between distinct anatomical configurations remain clear, confirms that the cAAE-EchoGAN effectively captures the underlying anatomical and conditional features while maintaining the diversity. Quantitatively, the Table I represents the evaluation metrics for the conditional generated images. For both the phases and across both the views the model demonstrates high dice scores of around 0.97. The high DICE scores demonstrate accurate preservation of cardiac structures, whereas the moderate SSIM values (0.57–0.59) indicate that the model generates realistic texture variations rather than memorizing training samples, thereby producing anatomically consistent yet diverse echocardiographic images essential for downstream clinical applications. Also, slightly higher SSIM and DICE scores in the ES phase compared to the ED phase suggest that the model may capture the more compact cardiac structures during systole more effectively than the diastolic phase. This is due to the fact that echocardiographic images contain speckle noise. And, during diastole, the heart expands, producing larger and more heterogeneous structures that are inherently more challenging to reconstruct accurately. The PSNR ranges from 20-21dB which supported by the qualitative analysis of the generated images for different conditions in Fig. 5.

Apart from the standard image metrics, FRD was calculated to quantitatively assess the similarity of radiomic features between real and synthetic echocardiographic images. It evaluates the feature distributions extracted from real and generated images, capturing not only pixel-level fidelity but also the higher-order statistical patterns relevant to clinical imaging. For the training batches the FRD was 5.7571 and the average FRD on the test set is 10.50. The model effectively generated high-quality samples for the different cardiac views and phases, preserving both anatomical fidelity and condition-specific features.

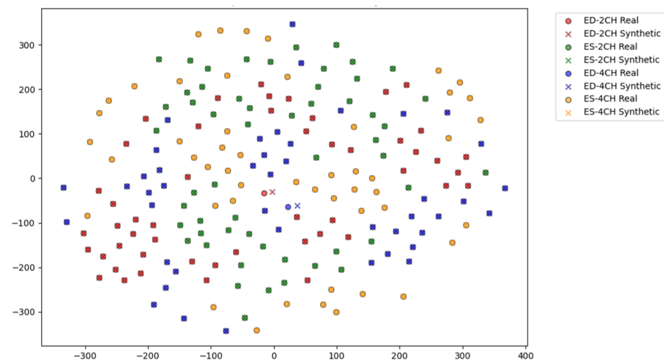

Fig. 4. t-SNE visualization of real and synthetic samples across different cardiac conditions, highlighting distribution similarity and class overlap.

TABLE I
PERFORMANCE METRICS FOR DIFFERENT CONDITIONS

Condition	**PSNR**	**SSIM**	**DICE**	**MAE**	**JSD** ($\times 10^{-3}$)
2CH (ED)	20.6732	0.5809	0.8809	0.0488	32.91
2CH (ES)	20.8090	0.5744	0.9706	0.0482	35.30
4CH (ED)	20.2046	0.5813	0.9669	0.0521	22.46
4CH (ES)	20.7547	0.5871	0.9738	0.0485	21.36

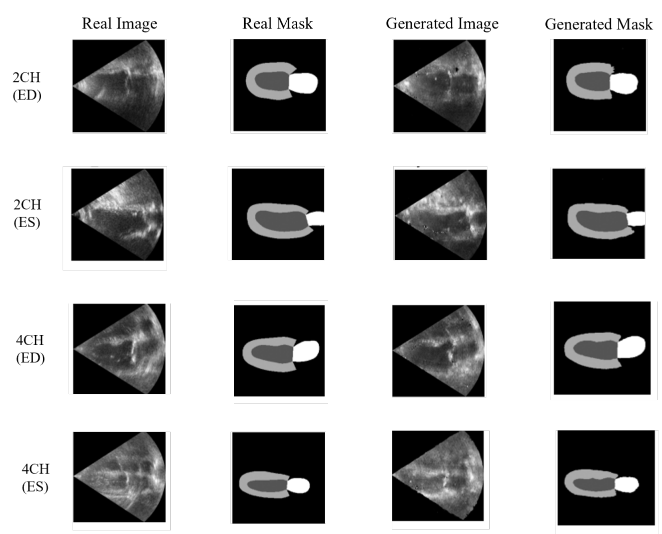

Fig. 5. Generated images with masks for different anatomical conditions

TABLE II
COMPUTATIONAL COMPLEXITY AND INFERENCE EFFICIENCY ANALYSIS OF THE PROPOSED CAAE-ECHOGAN FRAMEWORK.

Metric	Value
Hardware Platform	Kaggle (NVIDIA Tesla T4 $\times$ 2)
Total Parameters	73.99 M
Total FLOPs	19.56 GFLOPs
Inference Time/Image	8.51 ms
Frames Per Second (FPS)	117.46
Peak GPU Memory Usage	3.14 GB

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS

Model framework	PSNR	SSIM	FRD	JSD
MUNIT[14]	-	-	25.7	-
GcGAN[14]	-	-	12.1	-
MaskGAN[14]	-	-	27.8	-
SpeckleGAN [16]	-	0.431	-	33.2
DDPM+SPADE [17]	-	0.56	-	-
Free-Echo [18]	20.43	0.48	-	-
cAAE-EchoGAN	20.61	0.580	10.50	28.02

IV. CONCLUSION AND FUTURE SCOPE

This work proposed a novel conditional adversarial autoencoder-based framework $cAAE - EchoGAN$,for generating high-quality echocardiographic images across different cardiac views and phases with their synthetic anatomical mask. The key strength of the model lies in its dual-discriminator architecture.The image discriminator, which ensures image realism at the pixel level, while the second discriminator operates in the latent space, which regularises the latent space to promote diversity and stable training. The model generates anatomically consistent and condition-aware image-mask pairs by informed clinical conditioning based on cardiac phase and imaging view. The generated samples follow the distribution of real echocardiographic data, highlighting the potential of the proposed framework for synthetic data generation and data augmentation in cardiac imaging applications, supported by experimental results. Although the model was evaluated on a single echocardiographic dataset, its generalisation across different datasets with diverse cardiac pathologies needs to be investigated. Future work will focus on extending the framework to different datasets and conducting a comprehensive study of comparison with recent state-of-the-art generative models to further assess the trade-off between image quality, sample diversity, and computational efficiency. Overall, the proposed "$cAAE - EchoGAN$" provides an effective and computationally efficient solution for condition-aware echocardiographic image and mask synthesis, with strong potential to support the development of robust AI-driven cardiac imaging and diagnostic systems.

REFERENCES

[1] H. Tanaka, "Evolving role of echocardiography across the heart failure spectrum: from prevention to advanced management," *Heart Failure Reviews*, vol. 31, no. 6, 2026, doi: 10.1007/s10741-025-10584-1.

[2] G. Zamzmi, L.-Y. Hsu, W. Li et al., "Harnessing machine intelligence in automatic echocardiogram analysis: Current status, limitations, and future directions," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 181–201, 2021, doi: 10.1109/RBME.2020.2988295.

[3] H. Peng, Z. Chen, X. Tan, X. Fan, L. Xu, S. Li, et al., "A high-performance lightweight network for real-time view recognition and quality control in transthoracic echocardiography," *Expert Systems with Applications*, 2026, doi: 10.1016/j.eswa.2026.131777.

[4] S. Singh, A. Jain, and S. Chandra, "Selective denoising by NLM filtering for enhanced deep learning-based echocardiographic segmentation," *Biomedical Signal Processing and Control*, vol. 118, Art. no. 109647, 2026, doi: 10.1016/j.bspc.2026.109647.

[5] V. Pandey, U. K. Lilhore, and R. Walia, "A systematic review on cardiovascular disease detection and classification," *Biomedical Signal Processing and Control*, vol. 102, Art. no. 107329, 2025, doi: 10.1016/j.bspc.2024.107329.

[6] V. Kumar, N. M. Sharma, P. K. Mahapatra, N. Dogra, L. Maurya, F. Ahmad, et al., "Enhancing left ventricular segmentation in echocardiograms through GAN-based synthetic data augmentation and MultiResUNet architecture," *Diagnostics*, vol. 15, no. 6, Art. no. 663, 2025, doi: 10.3390/diagnostics15060663.

[7] D. Stojanovski, U. Hermida, P. Lamata, A. Beqiri, and A. Gomez, "Echo From Noise: Synthetic Ultrasound Image Generation Using Diffusion Models for Real Image Segmentation," in *Simplifying Medical Ultrasound (ASMUS 2023)*, B. Kainz, A. Noble, J. Schnabel, B. Khanal, J. P. Müller, and T. Day, Eds., Lecture Notes in Computer Science, vol. 14337. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-44521-7_4.

[8] N. Fatima *et al.*, "Synthetic Lung Ultrasound Data Generation Using Autoencoder With Generative Adversarial Network," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 72, no. 5, pp. 624–635, May 2025, doi: 10.1109/TUFFC.2025.3555447.

[9] E. Oladokun, M. Abdulkareem, J. Šprem, and V. Grau, "Transesophageal echocardiography generation using anatomical models," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Cham, Switzerland: Springer Nature Switzerland, Oct. 2023, pp. 43–52.

[10] S. Hangaragi, N. Neelima, V. Venugopal, S. Ganguly, J. Mudi, and J. H. Choi, "CAE SynthImgGen: Revolutionizing cancer diagnosis with convolutional autoencoder-based synthetic image generation," *Alexandria Engineering Journal*, vol. 115, pp. 343–354, 2025, doi: 10.1016/j.aej.2024.11.117.

[11] P. Ashrafian, M. Yazdani, M. Heidari, D. Shahriari, and I. Hacihaliloglu, "Vision-Language Synthetic Data Enhances Echocardiography Downstream Tasks," arXiv:2403.19880, 2024.

[12] H. Yang, X. Lu, S. H. Wang, et al., "Synthesizing multi-contrast MR images via novel 3D conditional variational auto-encoding GAN," *Mobile Networks and Applications*, vol. 26, pp. 415–424, 2021, doi: 10.1007/s11036-020-01678-1.

[13] H. Reynaud, A. Gomez, P. Leeson, Q. Meng, and B. Kainz, "EchoFlow: A Foundation Model for Cardiac Ultrasound Image and Video Generation," arXiv:2503.22357, 2025.

[14] N. Konz, R. Osuala, P. Verma, Y. Chen, H. Gu, H. Dong, et al., "Fréchet radiomic distance (FRD): A versatile metric for comparing medical imaging datasets," *Medical Image Analysis*, 2026, Art. no. 103943.

[15] S. Leclerc, E Smistad, J.Pedrosa "Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, Sept. 2019, doi: 10.1109/TMI.2019.2900516.

[16] L. Bargsten and A. Schlaefer, "SpeckleGAN: A generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 9, pp. 1427–1436, 2020, doi: 10.1007/s11548-020-02215-7.

[17] N. Van Phi, T. M. Duc, P. H. Hieu, and T. Q. Long, "Echocardiography video synthesis from end diastolic semantic map via diffusion model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 13461–13465.

[18] V. P. Nguyen, T. N. Luong Ha, H. H. Pham, and Q. L. Tran, "Training-free condition video diffusion models for single frame spatial-semantic echocardiogram synthesis," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2024)*, Cham, Switzerland: Springer Nature Switzerland, Oct. 2024, pp. 670–680.