Comparative study of methods for generating echocardiographic images

Salomón Hernández*, Andrés Carrera†, Eduardo Romero†, Marcela Iregui* and Alexander Cerón*

*ACCEDER, *Universidad Militar Nueva Granada*, Bogotá, Colombia
alexander.ceron@unimilitar.edu.co
†*CIM@LAB*, *Universidad Nacional de Colombia*, Bogotá

***Abstract*—The realistic generation of echocardiograms using Generative Adversarial Networks (GANs) represent a significant advancement in medical simulation, clinical training, and diagnostic support. This study presents a comparative evaluation of two generative architectures —StyleGAN2-ADA and MedGAN— and two reconstruction models —VQ-GAN and Pix2Pix— under similar training conditions, aiming to assess their performance in cardiac ultrasound image synthesis, quantitative and qualitative analyses revealed that StyleGAN2-ADA achieves stable and balanced generation, effectively learning from small datasets in reduced training time while avoiding artifacts, MedGAN showed structural limitations and slower convergence, VQ-GAN produced morphologically faithful reconstructions across dimensions but failed to reproduce realistic speckle noise, Pix2Pix demonstrated strong results in supervised contour-based reconstruction tasks. StyleGAN2-ADA was the only model to reach convergence, emphasizing the need to extend training beyond 2000 epochs for full optimization, these findings establish a foundation for future research in synthetic medical image generation for educational and diagnostic purposes.**

***Index Terms*—Data augmentation, Echocardiography, GAN, StyleGan2-ADA, MedGAN, VQ-GAN, Pix2Pix**

I. Introduction

The generation and reconstruction of high-quality synthetic images have become increasingly important in biomedical applications, especially when the acquisition of real data is complex, limited, or ethically constrained, among these applications, echocardiographic imaging stands out as a non-invasive and real-time diagnostic tool in cardiology, capable of visualizing cardiac structures and monitoring pathologies with high precision [1], [2].

In biomedical research, deep neural networks have become increasingly relevant due to their ability to analyze medical images with high performance, according to Drukker et al. [3], these techniques have demonstrated a remarkable impact across numerous medical applications, ranging from image segmentation and classification to the prediction of clinical outcomes, unlike traditional methods that rely on manual feature design based on expert knowledge, deep learning enables the detection of complex representations directly from data, as highlighted by Balaji et al. [4].

This work was funded by the project "Estimation of cardiac work as an index of cardiovascular function in echocardiographic videos" with code 60946 from the call for research projects SUE Distrito Capital of 2023 and code EXT_ING 4073 from Universidad Militar Nueva Granada.

A major challenge in deep learning lies in the limited availability of high-quality medical data, which hinders the development of robust models, this limitation has motivated the search for data augmentation strategies; however, conventional geometric transformations cannot be applied to medical images, as preserving cardiac morphology is crucial, even small spatial changes, such as rotations or translations, may produce unrealistic structures or inconsistent patterns, leading to diagnostic misinterpretations. Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014 [5], have emerged as a promising alternative by enabling the synthetic expansion of datasets through highly realistic image generation, GANs employ a dual architecture, where a generator produces synthetic data and a discriminator seeks to distinguish them from real samples, through this adversarial process, the generator gradually learns to synthesize convincing and realistic examples.

Despite their success in computer vision, applying GANs to medical imaging presents unique challenges: synthetic images must not only appear realistic but also preserve essential structural and morphological patterns, therefore, the design of architectures and the choice of loss functions require careful consideration, often incorporating perceptual metrics, morphological regularization, and clinically interpretable latent encoders [6].

In this study, four GAN architectures were selected based on their conceptual diversity to evaluate performance under a controlled setting with the same dataset and metrics, the selected models include a supervised method (Pix2Pix), an unsupervised approach (StyleGAN2-ADA), and structured latent models (MedGAN and VQ-GAN), the objective is to determine which architecture achieves the most morphologically faithful synthetic echocardiograms, and thus, which is best suited for medical data augmentation tasks.

II. Methods

A. Dataset and Preprocessing

The EchoNet-Dynamic dataset from Stanford University was employed, consisting of $10,030$ echocardiographic videos in the apical four-chamber view, this dataset was chosen because it is currently the most robust, extensive, and homogeneous resource in this medical domain, while also enabling

temporal analysis of the heart across its different phases. The videos were converted to grayscale to remove bluish filter variations present in less than 5% of the set and to prioritize morphology over color, subsequently, the frames were resized from 112×112 to 128×128 pixels to optimize the performance of convolutional layers, which are more efficient with resolutions in powers of two, the frames were then normalized, and a combined mask (triangular and semi-elliptical) was applied to highlight the cardiac region and discard irrelevant information, in addition, a bilateral filter was used to reduce noise and resizing artifacts, through binarization and white-pixel counting, the moments of maximum systole and diastole were identified: frames with the highest number of white pixels corresponded to peak systole, whereas those with the lowest number indicated peak diastole, finally, four representative frames were extracted from each video: the initial frame, the final frame, the diastolic frame, and the systolic frame, this process yielded a resulting dataset of $40,120$ images, sufficiently robust for training while reducing the risk of overfitting. Furthermore, to complement Pix2Pix, an image processing algorithm (Canny) was employed to detect contours and use them as paired data with the original images.

B. GAN models

The study aimed to assess the effectiveness of four generative architectures based on Generative Adversarial Networks (GANs) for synthesizing echocardiograms with high morphological fidelity, to this end, four models—**Pix2Pix, MedGAN, VQ-GAN, and StyleGAN2-ADA**—were trained and evaluated using the preprocessed dataset of $40,120$ ecocardiograms, each architecture was adapted and optimized with a specific set of hyperparameters to maximize its performance and generative potential.

1) $StyleGAN2-ADA$ [7]
   This architecture was developed to address a key challenge: training with small datasets, the solution proposed by ADA is the introduction of an adaptive data augmentation scheme, applied dynamically to both real and generated images, but regulated to prevent distribution bias, this is achieved through so-called "leak-free" augmentations, where even if corrupted images appear, the training process compensates for such distortions while converging toward the most appropriate distribution, its main characteristics include:
   - An intermediate latent space W, obtained by mapping the initial noise vector Z through a Mapping Network, enabling disentangled control of morphological attributes.
   - A Synthesis Network that constructs images from stylized inputs at each layer, using style modulation and controlled stochastic noise.
   - An Adaptive Discriminator Augmentation (ADA) mechanism that adjusts the level of data augmentation based on discriminator overfitting, making it well-suited for small datasets.
2) $MedGAN$ [8]
   The MedGAN implementation is based on a DCGAN-type architecture, where a generator G learns to transform a latent noise vector $z \in \mathbb{R}^{nz}$ into synthetic images, while a discriminator D attempts to distinguish between real and generated images, the main adaptation of MedGAN lies in the incorporation of an adaptive training scheme controlled by two thresholds (T_1, T_2), which dynamically regulate the number of iterations assigned to G and D in each training cycle, this mechanism balances the competition between the two networks and mitigates common issues such as mode collapse or discriminator overfitting, thereby enhancing the stability of medical image generation, unlike advanced adaptive regularization mechanisms such as ADA, MedGAN relies on the alternation and dynamic adjustment between generator and discriminator updates, guided by the control thresholds (T_1, T_2).
3) $VQ-GAN$ [6]
   This architecture is powerful for compressing and reconstructing realistic images from compact codes, allowing the generation of images that capture fine textures and key morphological structures, it integrates the learning of discrete latent representations with adversarial training. Its architecture includes:
   - Training the autoencoder to learn accurate reconstruction.
   - A vector quantization (VQ) layer that assigns each position on the map to a discrete vector from a trained dictionary.
   - A decoder that reconstructs the original image from the latent code.
   - A GAN discriminator that evaluates the quality of reconstructions.
   - Combined losses including:
     - L2 reconstruction
     - Perceptual loss (VGG)
     - Adversarial loss
     - R1 regularization for training stability.
4) $Pix2Pix$ [9]
   It is a supervised conditional GAN based on the translation of input images to output images, it was implemented with a U-Net generator and a PatchGAN discriminator, ideal for contexts where there is an explicit correspondence between a source image and its target.
   - The U-Net generator consists of downsampling and upsampling blocks, with skip connections between symmetric layers.
   - The PatchGAN discriminator evaluates small local regions of the image, reducing global artifacts.
   - The loss function combined adversarial loss with an L1 reconstruction loss, balanced by a parameter $\lambda = 100$.

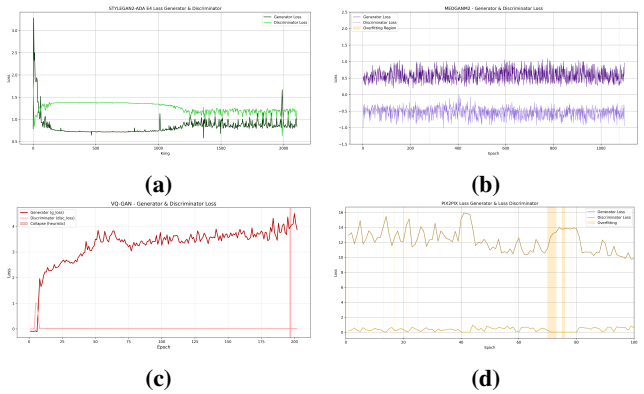


Fig. 1: Generator and discriminator loss curves (G_loss and D_loss) for the different models: (a) *StyleGAN2-ADA*, (b) *MedGAN*, (c) *VQ-GAN*, (d) *Pix2Pix*

C. Evaluation Metrics

Objective image quality metrics were used to compare the architectures:

- SSIM (Structural Similarity Index): evaluates the structural similarity between the generated image and the real image.
- FID (Fréchet Inception Distance): measures the distance between real and generated image distributions.

The metrics were recorded every 20 epochs and visualized with TensorBoard.

III. Results

Figure 1 presents the loss curves of the generator (G_loss) and the discriminator (D_loss) for the different implemented models. It can be observed that StyleGAN2-ADA exhibits a more stable training trajectory, showing some initial instability but without any signs of loss explosion or generator collapse. MedGAN displays the most adversarial training behavior among all models, maintaining a constant competition between the generator and the discriminator, which enables continuous improvement in the generated data. In VQ-GAN, the discriminator demonstrates considerable strength—although its loss does not decay to zero, it remains nearly constant—while the generator shows an increasing trend in its loss value that is not directly reflected in the resulting images. Finally, Pix2Pix exhibits a significant drop in G_loss between epochs 41 and 45 (from approximately 15.9 to 11.5), reflecting better alignment with the target image despite brief periods of overfitting.

TABLE I: Metric and loss values for each architecture up to its training total.

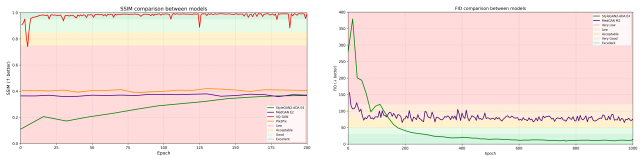
Model	Epoch max	G_loss Average	D_loss Average	SSIM max	FID min
StyleGAN2	2108	0.840	1.266	0.39	9.17
MedGAN	1095	0.584	-0.542	0.39	63.34
VQ-GAN	207	3.265	0.021	0.99	14.51
Pix2Pix	587	11.528	0.287	0.42	32.15

The ideal objective was to train all models for 2000 epochs; however, the computational demand and resource consumption of each architecture differ drastically and are extremely costly, the total training time across all models was approximately 3,630 hours, each architecture reached a different number of epochs within the same time frame (see Table I), with StyleGAN emerging as the most computationally efficient model, When analyzing the SSIM results across all models and the FID scores for generative architectures, a stagnation can be observed in GANs that generate new images from random noise. In the literature, SSIM values above 0.85 are typically considered excellent, while for FID, values below 30 are heuristically regarded as desirable for this study. However, achieving such values is not feasible in this type of imagery, as the comparison involves a batch of real images against a batch of randomly generated ones. There is no one-to-one correspondence between real and synthetic samples; therefore, it is impossible to obtain identical morphological similarities, since not all hearts or echocardiographic views are the same. Consequently, pixelwise comparison is not an appropriate approach for unconditioned GANs.

An analysis of each model reveals that VQ-GAN achieved the highest morphological fidelity, obtaining the top SSIM values (see Table I), this indicates that the images reconstructed by this model preserve the anatomical structures of the heart (see Fig. II) with superior accuracy, the discriminator losses are the lowest among all architectures relative to their corresponding generators, suggesting a strong discriminator capable of effectively evaluating reconstructed images, this behavior may indicate a certain level of training saturation, although no evident negative impact on output quality was observed, overall, VQ-GAN stands out as the most robust architecture for image reconstruction and for learning a powerful codebook, which can later be used to train a Transformer model to generate diverse new instances for the synthetic augmentation of echocardiograms.

Regarding StyleGAN2-ADA, an unsupervised model, it exhibited a balanced behavior between visual fidelity and training stability, achieving an average SSIM of 0.39, this reflects a lower performance compared to the other models. Its loss functions indicate a stable training dynamic without signs of collapse or overfitting, although it achieved the lowest SSIM value—demonstrating limited ability to reproduce images with high structural similarity to the original dataset—it obtained a considerably better average FID of 9.17, registering values close to an ideal distribution, this demonstrates a strong capacity to generate new echocardiographic images that preserve the morphological characteristics of the heart, its consistent performance suggests that StyleGAN2-ADA is a reliable and versatile model, particularly suitable for scenarios where paired data are unavailable, therefore, it represents a solid alternative for scenarios that require an architecture that ensures realism, diversity, stability, and training efficiency.

The results obtained with Pix2Pix indicate the second-highest morphological fidelity in image reconstruction, achieving an average SSIM value of 0.42 (see Figure 2), although



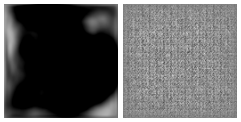
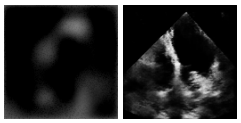
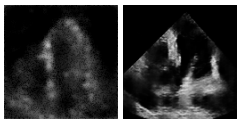
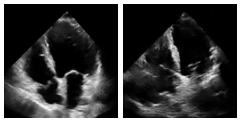
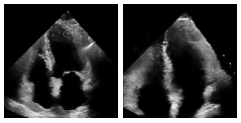
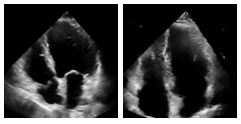
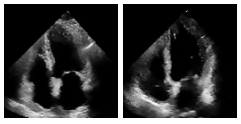
(a) **(b)**

Fig. 2: Performance metrics for the different models: (a) *SSIM* It can be seen that for less than 200 epochs the generative architectures show low levels of similarity, with VQ-GAN being the highest and improving over the course of each epoch, and with StyleGAN having the lowest values. (b) *FID* It can be seen that both generative architectures have appropriate distribution values, with StyleGAN being the model with the best generation of new hearts as training progresses, while MedGAN improves very slightly.

the generator loss at epoch 100 was relatively high (9.82), the discriminator's low loss value (0.60) suggests an effective adversarial balance, however, the model exhibits constant volatility, with the generator and discriminator continuously competing; therefore, evaluating its efficiency solely based on the final epoch may not be entirely appropriate, as a supervised model with explicit input–output correspondence (input: contour, output: echocardiogram), Pix2Pix demonstrated an excellent ability to reconstruct morphologically coherent images from structured pairs.

Finally, MedGAN exhibited highly volatile adversarial dynamics but demonstrated a consistent pattern of gradual improvement, its SSIM values position it as the second-weakest architecture, which is expected since it generates new images entirely from scratch and is therefore unable to preserve exact structural similarity, the discriminator loss is the only negative among all models; however, this behavior is explained by the fact that MedGAN employs a critic estimating the Wasserstein distance rather than a binary classifier, consequently, such loss values should not be interpreted as signs of overfitting (see Figure 1), a slight overfitting episode can be observed around epoch 278 (highlighted with a yellow band), but the model's adaptive discriminator augmentation (ADA) mechanism promptly corrects it, preventing recurrence throughout the remaining training process, although MedGAN is unable to reproduce the characteristic speckle noise typical of echocardiograms and requires many more iterations to approximate the formation of the four cardiac chambers and finer anatomical details such as the mitral structures, these results align with the qualitative trends reflected by the FID metric, within fewer than 1000 epochs, MedGAN achieved an average FID of 63.34—acceptable yet still insufficient for medical-grade applications. Despite incorporating latent encodings and perceptual regularization, its adaptation to the specific echocardiography domain appears less effective compared to its generative counterpart, StyleGAN.

TABLE II: Results of all models tested at different times

# of epoch	Generations		Reconstructions	
	StyleGAN2	MedGAN	VQ-GAN	Pix2Pix
Epoch 1				
Epoch 20				
Epoch 40				
Epoch 60				
Epoch 80				
Epoch 100				
Epoch 200				

IV. Conclusions

This study demonstrated that under a training regime limited to 1,000 epochs, StyleGAN2-ADA is the generative architecture that achieves the highest morphological fidelity, perceptual quality in echocardiogram synthesis, and synthetic diversity, these results position it as the most promising option, showing consistent, balanced, and efficient performance when trained on medium-sized datasets, in contrast, MedGAN exhibited structural limitations, reduced generalization capacity, and required more extensive training to reach an optimal point of convergence, for the reconstruction of echocardiograms, Pix2Pix showed a certain degree of effectiveness in supervised scenarios with paired data, though it raises questions about whether the use of contours is more suitable than segmentation for preserving morphological integrity, finally, VQ-GAN demonstrated exceptionally faithful reconstructions through its decoding and codebook learning process. Future work will focus on extending the variation of hyperparameters, increase the number of training epochs, and conduct a qualitative evaluation of the generated images by a medical imaging expert.

REFERENCES

[1] Z. Nie, M. Xu, Z. Wang *et al.*, "A review of application of deep learning in endoscopic image processing," *Journal of Imaging*, vol. 10, no. 11, p. 275, Nov. 2024.

[2] Y. Wang, X. Ge, H. Ma *et al.*, "Deep learning in medical ultrasound image analysis: A review," *IEEE Access*, vol. 9, pp. 54 310–54 324, 2021.

[3] K. Drukker, P. Yan, A. Sibley *et al.*, "Chapter 4 - biomedical imaging and analysis through deep learning," in *Artificial Intelligence in Medicine*, L. Xing, M. L. Giger, and J. K. Min, Eds. Academic Press, 2021, pp. 49–74.

[4] K. Balaji and K. Lavanya, "Chapter 5 - medical image analysis with deep neural networks," in *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, A. K. Sangaiah, Ed. Academic Press, 2019, pp. 75–97.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 2. MIT Press, 2014, pp. 2672–2680.

[6] M. Frid-Adar, I. Diamant, E. Klang *et al.*, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[7] T. Karras, M. Aittala, J. Hellsten *et al.*, "Training generative adversarial networks with limited data," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[8] K. Guo, J. Chen, T. Qiu *et al.*, "Medgan: An adaptive gan approach for medical image generation," *Computers in Biology and Medicine*, vol. 163, p. 107119, 2023.

[9] P. Isola, J.-Y. Zhu, T. Zhou *et al.*, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.