Standardizing Obstetric Ultrasound Segmentation Using Unpaired Domain Translation Techniques

Emilio J. Ochoa*, Arthur Masiukiewicz*, Cristina Orihuela†, Maria Helguera*, and Benjamin Castaneda*
*Department of Biomedical Engineering, University of Rochester, New York, USA
†Laboratorio de Imágenes Médicas, Pontificia Universidad Católica del Perú, Lima, Perú

***Abstract*—Prenatal ultrasound is essential for fetal monitoring, yet access in low-resource settings remains limited by the shortage of trained personnel and variability in imaging equipment. Volume Sweep Imaging (VSI) enables acquisition by non-experts and provides diagnostically useful cine-loops for physician review. However, when extending VSI interpretation to artificial intelligence (AI) pipelines, cross-scanner variability introduces domain shift—scanner-dependent differences in contrast, speckle, and resolution that degrade model generalizability. In this study, we evaluate unpaired domain translation methods to standardize obstetric VSI across two scanners (Butterfly iQ+ and Mindray DP10). We implemented CycleGAN, denoising diffusion GANs, and a sequential (CG→Diff) approach. Performance was assessed with distribution similarity (PSNR, MI, BC, MAE) and structural similarity (SSIM, LNCC, CSS) metrics, alongside qualitative segmentation analysis. Results show that the hybrid method achieved the most balanced performance, improving PSNR (21.69) and LNCC (0.59) while preserving anatomical structures. These findings highlight the potential of adversarial-diffusion pipelines to mitigate domain shift and enable scalable AI-assisted obstetric ultrasound in low-resource environments.**

***Index Terms*—Obstetric ultrasound, Volume Sweep Imaging, domain adaptation, CycleGAN, diffusion models, medical image translation, low-resource settings, fetal imaging.**

I. Introduction

Prenatal ultrasound imaging plays a pivotal role in monitoring fetal growth, identifying complications, and informing delivery planning. The World Health Organization recommends at least one ultrasound examination per trimester to ensure appropriate fetal assessment [1]. However, in many low- and middle-income countries (LMICs), especially rural and remote regions, this essential service is hindered by the scarcity of trained sonographers, limited access to advanced equipment, and infrastructural constraints. Volume Sweep Imaging (VSI) has emerged as a promising solution to overcome acquisition-related barriers in low-resource settings. VSI is a standardized, operator-independent ultrasound protocol in which non-expert health workers perform systematic sweeps of the transducer over predefined anatomical landmarks (fig. 1). This method has demonstrated the ability to produce diagnostically useful cine-loops without requiring the operator to have detailed anatomical knowledge, thereby reducing dependence on specialized personnel [2], [3].

Artificial intelligence (AI) methods, particularly convolutional neural networks (CNNs) and attention-based architectures, have shown promise in automating VSI ultrasound interpretation [4]–[6]. Nevertheless, a persistent challenge in deploying such models at scale is a *domain shift* problem: ultrasound images vary substantially across different scanners, acquisition protocols, and imaging environments. Variations in gray-level distribution, contrast, resolution, and speckle pattern can degrade the performance of AI models when applied to data from unseen devices or sites, even if the anatomical content is similar.

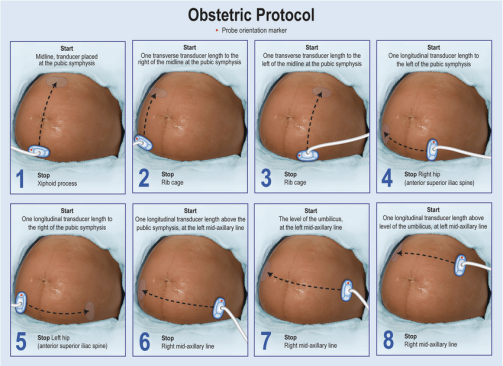

Fig. 1. Volume Sweep Imaging obstetric protocol.

Domain shift is especially problematic in multi-center or multi-device VSI-OB deployments, where models trained on one scanner (or site) often generalize poorly to others. Collecting large, fully labeled datasets from every device is impractical, given the resource constraints and the high cost of expert annotation in ultrasound imaging.

Unsupervised domain transformation offers a practical pathway to mitigate these discrepancies. By translating images from multiple source domains into a common target domain representation, downstream segmentation or classification models can operate on standardized inputs, improving robustness and generalizability without requiring additional labels from each source. In this work, we evaluate the potential of unsupervised domain transformation to standardize obstetric VSI images across 2 different scanners, with the goal of enabling segmentation models trained on target-domain annotations to operate reliably on heterogeneous VSI acquisitions. This approach aims to facilitate scalable AI-assisted fetal assessment in low-resource settings by mitigating device and site-specific variability without requiring additional labeled data.

II. Methods

A. Dataset

The dataset used in this study consisted of cine-loops acquired using two different ultrasound scanners: the Butterfly iQ+(source domain) and the Mindray DP10 (target domain) [2], [3]. In total, we collected 7200 paired frames, corresponding to 187 patients in the source domain and 127 patients in the target domain.

1) Acquisition details: VSI videos were acquired by non-expert operators following standardized sweep instructions. Each video was cropped to include only the ultrasound field of view, removing on-screen annotations and surrounding margins. Since the number of frames per video varied, we harmonized the dataset by selecting the minimum number of frames across all videos. For videos with more frames than this minimum, frames were randomly sampled to avoid bias toward longer acquisitions. This procedure ensured that all patients contributed an equal number of frames, mitigating imbalances across the dataset.

2) Preprocessing: To avoid data leakage, we adopted a strict patient-wise split, ensuring that no patient appeared simultaneously in training, validation, or test sets. Patients were divided into three groups with proportions of 70% for training, 15% for validation, and 15% for testing. This resulted in balanced partitions across both source and target domains.

Frames were resized to 256×256 pixels and normalized before training. No data augmentation was applied due to the moderate size of the dataset and the risk of introducing artificial domain variations.

3) CSV generation: To facilitate reproducibility, patient-level splits were encoded into CSV files containing paired low- and high-quality frame paths. In total, the final splits consisted of 5090 training pairs, 1051 validation pairs, and 1060 testing pairs.

B. Models

1) Cycle GAN: In this application, we utilized the CycleGAN (CG) to learn a bidirectional translation between ultrasound frames imaged in a target domain x and source domain y. A pair of two generators and two discriminators is used, such that the first generator-discriminator set can learn a translation of images from the source domain to target domain and the second learns the reverse. [7], [8]

2) Denoising Diffusion GAN: In previous literature, diffusion models have followed the principle that in the limit of an infinitesimal step size, the diffusion and reversal of the diffusion process have an identical functional form. In other words, both the diffusion process where noise is added to the data, and the denoising process can both be approximated with a Gaussian distribution.

Our Denoising Diffusion GAN (DDGAN) follows the architecture proposed by NVIDIA, which breaks the Gaussian reverse assumption, and assumes that it can be modeled with a multimodal distribution. This diffusion process is completed in a few large steps, which are modeled by a conditional GAN.

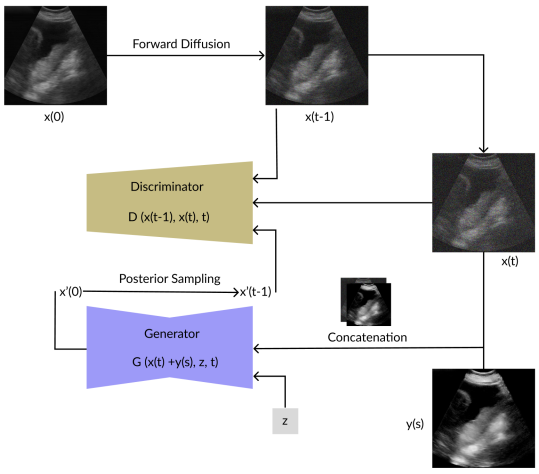

Fig. 2. The denoising diffusion GAN training process. **x(0)** is a target-domain image at time step 0, where **(t)** denotes the image at time step t in the forward diffusion process. **y(s)** denotes a source-domain image synthesized from target-domain image x(0) using the CycleGAN. **z** denotes the latent noise that is added into the generator. **x'(0)** denotes the generator's guess of a non-noisy image corresponding to x(0) with **x'(t-1)** created via posterior sampling.

Starting from a target-domain image $x_{(0)}$, a random time-step "t" is sampled. This time-step is used to obtain a noisy pair of images $x_{(t-1)}$ and $x_{(t)}$ through the process of forward diffusion. These pairs are passed to the discriminator D as examples of target-domain images at the noise level at time-step t. [9], [10]

To apply the principle of the DDGAN to ultrasound imagery, the CycleGAN is used to synthesize a source-domain image from the original target-domain image; $y_{(s)}$. $x_{(t)}$ and $y_{(s)}$ are then concatenated along the z-axis to create a 2-channel tensor containing both images. This is done so that the model learns to preserve the structure of the image; minimizing hallucination in the large reverse diffusion steps.

This tensor is passed to the generator, G, along with the time embedding at t, and latent noise z. The generator creates a plausible clean image $x'_{(0)}$ that could have been produced from $x_{(t)}$. Posterior sampling is then completed to create a candidate $x'_{(t-1)}$ state from the $x'_{(0)}$ image. The discriminator, D, then receives the $x_{(t)}$ and $x_{(t-1)}$ pair along with the $x_{(t)}$ and $x'_{(t)}$ pair to learn to distinguish between the two.

C. Training setup

CycleGAN and two diffusion-based models were trained under two configurations: (t=250, e=125) and (t=170, e=50), where t denotes diffusion timesteps and e the number of epochs. In addition to these single-model baselines, we also evaluate sequential CG→Diffusion pipeline, where outputs of the CycleGAN translator are refined by a diffusion stage. This combined setup aims to retain anatomical fidelity while improving distribution alignment.

D. Evaluation protocol

Performance was assessed with **distribution similarity metrics** (PSNR (↑), MI (↑), BC (↑), MAE (↓)) and **structural**

TABLE I
TRAINING HYPERPARAMETERS FOR CYCLEGAN AND DIFFUSION-BASED METHODS.

	CycleGAN	Diffusion	Diff+CG
Timesteps t	–	250 / 170	250 / 170
Epochs e	125	125 / 50	125 / 50
Batch size	128	128	128
Optimizer	Adam	Adam + EMA	Adam + EMA
Learning rate	2×10^{-4}	1×10^{-4}	1×10^{-4}

TABLE II
LOSS FUNCTIONS USED FOR CYCLEGAN, DIFFUSION, AND CYCLEGAN→DIFFUSION TRAINING.

Model	Loss components
CycleGAN	Adversarial loss (LSGAN) - to learn domain, Cycle-consistency loss (L_1) - for maintaining anatomy, Identity loss (L_1).
Diffusion (DDGAN)	Conditional adversarial loss (GAN on pairs over large steps); with R1 Regularization
M2O-Diffgan	Combination of CycleGAN losses + Diffusion losses; Ultrasound-specific content loss (MSE + LPIPS loss) for structure/detail preservation.

similarity metrics (SSIM (↑), LNCC (↑), CSS (↑)) based on M20-DiffGAN approach [10]. CSS is a non-standard metric computed as the average of the *contrast* and *structure* terms of SSIM, excluding luminance; higher values indicate better structural preservation. Arrows indicate the preferred direction: ↑ higher is better, ↓ lower is better.

III. RESULTS

A. Quantitative Results

Table III and Fig. 3 present the performance of CycleGAN, Diffusion, and CG→Diffusion. The hybrid method achieved the highest PSNR (21.69) and LNCC (0.59), with improved CSS (0.56–0.57) compared to Diffusion (0.44). CycleGAN obtained the best MI (0.75) and slightly higher SSIM (0.62), while MAE was lowest for both CycleGAN and CG→Diffusion (0.06). Overall, CG→Diffusion provided the best balance between structural fidelity and intensity alignment. The most consistent improvements were observed with the $(t = 170, e = 50)$ configuration, which yielded the highest PSNR and stable structural scores, indicating that fewer diffusion steps with moderate training epochs were more effective than longer schedules.

B. Qualitative results

Since manual ground truth is unavailable for translated images, we provide a qualitative assessment using a segmentation model trained on the target domain only [6]. Figure 4 shows that CG→Diffusion produces translations where fetal structures are well-preserved and segmentation masks closely follow anatomical boundaries.

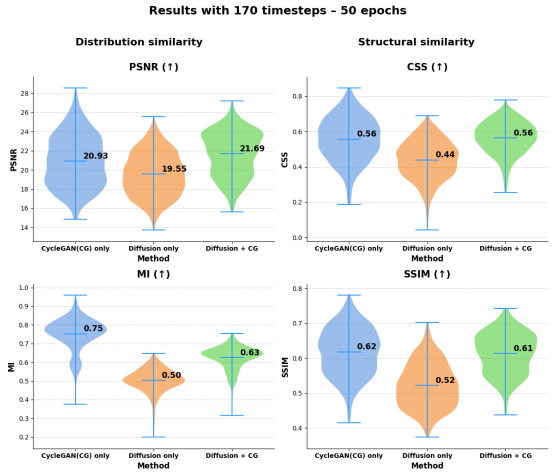

Fig. 3. Violin plots comparing translation methods using **170 timesteps** and **50 epochs**. Left column: *Distribution similarity* (PSNR (↑), MI (↑)). Right column: *Structural similarity* (CSS (↑), SSIM (↑)). Blue markers indicate min/mean/max with the numeric mean annotated. Methods: CycleGAN(CG) only, Diffusion only, and Diffusion + CG.

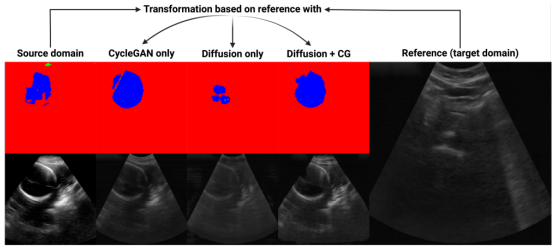

Fig. 4. Qualitative evaluation of domain translation. Top: segmentation overlays produced by a target-domain model. Bottom: corresponding B-mode frames. Translations include CycleGAN only, Diffusion only, and Diffusion+CG, with the reference target image shown on the right. Diffusion+CG achieves the most anatomically consistent results.

IV. DISCUSSION

The results demonstrate that unpaired domain translation can substantially mitigate device-specific variability in obstetric VSI data. Among the methods evaluated, the hybrid CG→Diffusion approach consistently provided the most balanced performance, combining high PSNR and LNCC with competitive SSIM and low MAE. This suggests that the sequential use of adversarial translation followed by diffusion refinement is effective in aligning intensity distributions while preserving structural fidelity of fetal anatomy.

A notable observation is that CycleGAN achieved the highest MI and SSIM values, indicating strong global distribution alignment and perceptual similarity. However, this often came at the cost of structural distortions, as reflected by lower CSS and LNCC scores. Diffusion models, by contrast, tended to oversmooth images, which reduced noise but degraded anatomical detail.

From the perspective of downstream clinical utility, structural preservation is critical. Segmentation models trained on target-domain labels rely on consistent representation of anatomical boundaries, and the improvements in CSS and

TABLE III

AVERAGE ± STANDARD DEVIATION OF METRICS FOR **CYCLEGAN**, **DIFFUSION**, AND **CYCLEGAN→DIFFUSION**. GROUPED HEADERS INDICATE THE TWO PARAMETER SETTINGS USED FOR DIFFUSION-BASED METHODS: **t** = TIMESTEPS, **e** = EPOCHS. ARROWS DENOTE PREFERRED DIRECTION (↑ HIGHER IS BETTER; ↓ LOWER IS BETTER). BEST VALUES PER METRIC ARE HIGHLIGHTED IN BOLD.

Metric	**CycleGAN**	**Diffusion**		**CycleGAN→Diffusion**	
		(t=250, e=125)	**(t=170, e=50)**	**(t=250, e=125)**	**(t=170, e=50)**
PSNR (↑)	20.93 ± 2.70	18.59 ± 2.23	19.55 ± 2.23	21.57 ± 2.19	**21.69 ± 2.32**
CSS (↑)	0.56 ± 0.12	0.44 ± 0.11	0.44 ± 0.11	**0.57 ± 0.09**	0.56 ± 0.09
MI (↑)	**0.75 ± 0.09**	0.52 ± 0.05	0.50 ± 0.06	0.63 ± 0.06	0.63 ± 0.06
SSIM (↑)	**0.62 ± 0.07**	0.49 ± 0.06	0.52 ± 0.06	0.60 ± 0.06	0.61 ± 0.06
LNCC (↑)	0.56 ± 0.11	0.43 ± 0.10	0.43 ± 0.11	**0.59 ± 0.08**	**0.59 ± 0.08**
BC (↑)	**0.53 ± 0.11**	0.49 ± 0.10	0.46 ± 0.09	0.43 ± 0.09	0.41 ± 0.09
MAE (↓)	**0.06 ± 0.02**	0.08 ± 0.02	0.07 ± 0.02	**0.06 ± 0.01**	**0.06 ± 0.02**

LNCC with CG→Diffusion directly support this requirement. The comparison of diffusion schedules highlights the role of hyperparameters in shaping performance. The configuration with 70 timesteps and 50 epochs was superior versus the longer training schedule of 250 timesteps and 125 epochs in terms of PSNR and LNCC. This suggests that excessive diffusion steps may introduce over-smoothing, reducing structural fidelity despite better distribution alignment. Several limitations should be acknowledged. First, our qualitative analysis relied on applying a segmentation model trained on the target domain to translated images. While this provides indirect evidence of anatomical preservation, a more robust approach would involve generating ground-truth segmentations for the source domain as well, enabling direct evaluation of segmentation accuracy before and after translation.

Second, our pipeline implements CG→Diffusion as a sequential cascade rather than a joint training strategy. While effective, this design may not fully exploit the potential synergy between adversarial and diffusion objectives. Parallel or unified architectures, such as M2O-DiffGAN [10], could offer better integration and improved stability.

Third, although we included LPIPS as a perceptual loss, it was originally trained on ImageNet [11] and may not accurately capture structural or textural fidelity in ultrasound images. Domain-specific perceptual similarity models tailored for ultrasound could provide more meaningful evaluations [12].

Finally, the study was limited to two devices; broader validation across multiple scanners and acquisition environments is necessary to confirm the scalability of the approach.

V. CONCLUSSION

Our findings highlight the feasibility of unsupervised domain translation to harmonize VSI acquisitions. By mitigating domain shift without requiring new annotations, hybrid adversarial-diffusion models could enable more robust and scalable deployment of AI-based obstetric ultrasound in low-resource settings. Future work should expand validation to multi-device datasets, incorporate domain-specific perceptual metrics, and explore unified adversarial-diffusion training strategies for greater stability and fidelity.

REFERENCES

[1] W. H. Organization, *WHO antenatal care recommendations for a positive pregnancy experience. Nutritional interventions update: Multiple micronutrient supplements during pregnancy*. World Health Organization, 2020.

[2] M. Toscano, T. J. Marini, K. Drennan, T. M. Baran, J. Kan, B. Garra, A. M. Dozier, R. L. Ortega, R. A. Quinn, Y. T. Zhao *et al.*, "Testing telediagnostic obstetric ultrasound in peru: a new horizon in expanding access to prenatal ultrasound," *BMC Pregnancy and Childbirth*, vol. 21, no. 1, p. 328, 2021.

[3] M. Erlick, T. Marini, K. Drennan, A. Dozier, B. Castaneda, T. Baran, and M. Toscano, "Assessment of a brief standardized obstetric ultrasound training program for individuals without prior ultrasound experience," *Ultrasound quarterly*, vol. 39, no. 3, pp. 124–128, 2023.

[4] E. J. Ochoa, S. E. Romero, T. J. Marini, A. O'Connell, G. Brennan, J. Kan, S. Meng, Y. Zhao, T. Baran, and B. Castaneda, "A comparison between deep learning architectures for the assessment of breast tumor segmentation using vsi ultrasound protocol," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

[5] D. Khaledyan, T. J. Marini, A. O'Connell, S. Meng, J. Kan, G. Brennan, Y. Zhao, T. M. Baran, and K. J. Parker, "Watunet: a deep neural network for segmentation of volumetric sweep imaging ultrasound," *Machine Learning: Science and Technology*, vol. 5, no. 1, p. 015042, 2024.

[6] J. Arroyo, T. J. Marini, A. C. Saavedra, M. Toscano, T. M. Baran, K. Drennan, A. Dozier, Y. T. Zhao, M. Egoavil, L. Tamayo *et al.*, "No sonographer, no radiologist: New system for automatic prenatal detection of fetal biometry, fetal presentation, and placental location," *PloS one*, vol. 17, no. 2, p. e0262107, 2022.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[8] L. Huang, Z. Zhou, Y. Guo, and Y. Wang, "A stability-enhanced cyclegan for effective domain transformation of unpaired ultrasound images," *Biomedical Signal Processing and Control*, vol. 77, p. 103831, 2022.

[9] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.

[10] L. Huang, J. Zhou, J. Jiao, S. Zhou, C. Chang, Y. Wang, and Y. Guo, "Standardization of ultrasound images across various centers: M2o-diffgan bridging the gaps among unpaired multi-domain ultrasound images," *Medical Image Analysis*, vol. 95, p. 103187, 2024.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[12] Z. Zhou, Y. Guo, and Y. Wang, "Handheld ultrasound video high-quality reconstruction using a low-rank representation multipathway generative adversarial network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 575–588, 2020.