

Ultrasound Image-to-Video Synthesis via Latent Dynamic Diffusion Models

Tingxiu Chen^{1*}, Yilei Shi^{1*}, Zixuan Zheng¹, Bingcong Yan¹, Jingliang Hu¹,
Xiao Xiang Zhu², and Lichao Mou¹(✉)

¹ MedAI Technology (Wuxi) Co. Ltd., Wuxi, China
lichao.mou@medimagingai.com

² Technical University of Munich, Munich, Germany

Abstract. Ultrasound video classification enables automated diagnosis and has emerged as an important research area. However, publicly available ultrasound video datasets remain scarce, hindering progress in developing effective video classification models. We propose addressing this shortage by synthesizing plausible ultrasound videos from readily available, abundant ultrasound images. To this end, we introduce a latent dynamic diffusion model (LDDM) to efficiently translate static images to dynamic sequences with realistic video characteristics. We demonstrate strong quantitative results and visually appealing synthesized videos on the BUSV benchmark. Notably, training video classification models on combinations of real and LDDM-synthesized videos substantially improves performance over using real data alone, indicating our method successfully emulates dynamics critical for discrimination. Our image-to-video approach provides an effective data augmentation solution to advance ultrasound video analysis. Code is available at https://github.com/MedAITech/U_I2V.

Keywords: image-to-video · medical ultrasound · diffusion model.

1 Introduction

Ultrasound imaging is widely used in clinical practice owing to its non-invasive, radiation-free, and real-time nature. As a primary screening tool, it enables the identification of abnormalities in major organs. However, the accurate interpretation of ultrasound images remains challenging even for experienced practitioners, as healthy and diseased areas in images frequently demonstrate indistinguishable appearances.

Deep learning has emerged as a promising approach for providing suggestions to assist clinicians in their decision making. Ultrasound image classification using deep networks has become a major research area, with state-of-the-art models achieving impressive results [7,8]. However, since individual ultrasound images provide limited views of lesions, aggregating multidimensional information across

* Equal contribution.

an entire ultrasound video is more advantageous for accurate automated diagnosis. Very recently, ultrasound video classification has become a prominent research area [3,4,5,6]. Yet, in contrast to image datasets [12,13], publicly available ultrasound video datasets [4] remain relatively scarce and limited in size, as ultrasound videos are often not recorded and stored. This poses significant challenges in training deep networks for ultrasound video classification. We consider alleviating this problem by utilizing a large number of existing labeled ultrasound images rather than collecting more ultrasound videos. Synthesizing plausible video sequences from static images is an attractive solution. By doing so, we can augment training data for video classification models and thus improve their performance.

Related work. Ultrasound image generation has been explored through physics-based simulators [1,2,14] and registration-based approaches. Motivated by the success of generative adversarial networks (GANs), recent works have focused on leveraging GANs for ultrasound simulation. For instance, [15] uses a stacked GAN model for the fast simulation of patho-realistic ultrasound images. [16] presents a sketch GAN to synthesize editable ultrasound images and introduces a progressive training strategy to generate high-resolution images. However, these methods are limited to enhancing the quantity and diversity of images, not videos.

Video-to-video translation has garnered interest in recent years. For example, [17] presents a GAN model to enhance the video quality of handheld ultrasound devices. [10] introduces a causal generative model capable of modifying cardiac ultrasound video content based on desired left ventricular ejection fraction (LVEF) values. [11] transfers motion from existing pelvic ultrasound videos to static images using key point detection and GANs. For image-to-video synthesis, [9] proposes a diffusion model to generate cardiac ultrasound sequences from ultrasound stills. However, this approach requires LVEF as an additional input, thus limiting its application. Moreover, it is computationally expensive (8 NVIDIA A100 GPUs required). In computer vision, [18] also focuses on image-to-video synthesis, but on natural images/videos. In this work, we are interested in a computationally efficient ultrasound video synthesis approach operating directly on individual ultrasound images, without requiring additional inputs.

Contributions. Ultrasound video synthesis from static images is challenging, as images lack dynamic information present in videos. To address this, we propose an approach to compensate between image and video domains for video generation. Our key contributions are as follows:

- To our knowledge, this work represents the first attempt to leverage image-to-video synthesis to augment training data for ultrasound video classification. We thoroughly investigate the feasibility of this direction.
- We propose a latent dynamic diffusion model (LDDM) to translate ultrasound images to plausible dynamic sequences in a computationally efficient manner.
- We demonstrate strong quantitative and qualitative results on the BUSV benchmark. Notably, classification performance improves substantially when

training video-based diagnostic models on combinations of real and synthetic data, indicating our generated videos well emulate real dynamics.

2 Method

Given a starting image $\mathbf{x}_0$ to serve as the initial condition, our method aims to synthesize a plausible corresponding video clip $\mathbf{v}$. The inherently under-determined nature of this problem permits the formulation of a multitude of conceivable video predictions based on $\mathbf{x}_0$. As such, comprehensively modeling video dynamics requires information beyond what is present in the initial seed image. To address this ill-posed problem, we propose to transform a noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into a dynamic representation by means of a conditional diffusion model. This approach learns to harness stochastic noise to compensate for the absence of intrinsic dynamics in the conditioning input image.

In this section, we elaborate on the two-stage framework of the proposed LDDM and elucidate the process of video synthesis for enhancing ultrasound video classification. Fig. 1 shows the overall training pipeline of LDDM. In the first stage, we learn an autoencoding framework to capture intrinsic video dynamics within a low-dimensional latent space. Specifically, an encoder $\mathcal{E}$ compresses an input video into an embedding $\mathbf{z}$, which is then decoded by $\mathcal{D}$ along with the first frame $\mathbf{x}_0$ to reconstruct the original clip. In the second stage, we model the latent space encoding dynamics using a conditional diffusion model, where $\mathbf{x}_0$ provides image context. This allows generating latent embeddings with realistic dynamics from Gaussian noise, conditioned solely on static images at test time. For synthesizing novel videos, given an ultrasound image, we first leverage the trained diffusion model to produce a latent representation $\mathbf{z}'$. Then, the trained decoder $\mathcal{D}$ transforms the simulated $\mathbf{z}'$ and the seed image into a synthetic video sequence.

2.1 Video Embedding

To acquire encoding and decoding capabilities for videos, we employ an autoencoder framework to represent video dynamics as latent embeddings. Specifically, our encoder $\mathcal{E}$ is implemented as a 3D ResNet [30] (ResNet-18) in order to capture the temporal evolution of dynamics within a video. Additionally, the decoder $\mathcal{D}$ is required to reconstruct the video from the initial frame $\mathbf{x}_0$ as well as the latent representation of the video. For $\mathcal{D}$, we adopt an architecture consisting of blocks containing AdaIN layers [29] to enable the propagation of video information across all scales of the decoder.

2.2 Conditional Diffusion Model

Our conditional diffusion model builds on denoising diffusion-based generative modeling [19,20]. Given a sample from a data distribution $\mathbf{z}_0 \sim q(\mathbf{z}_0)$, the forward process of our model produces a Markov chain $\mathbf{z}_1, \dots, \mathbf{z}_T$ by progressively

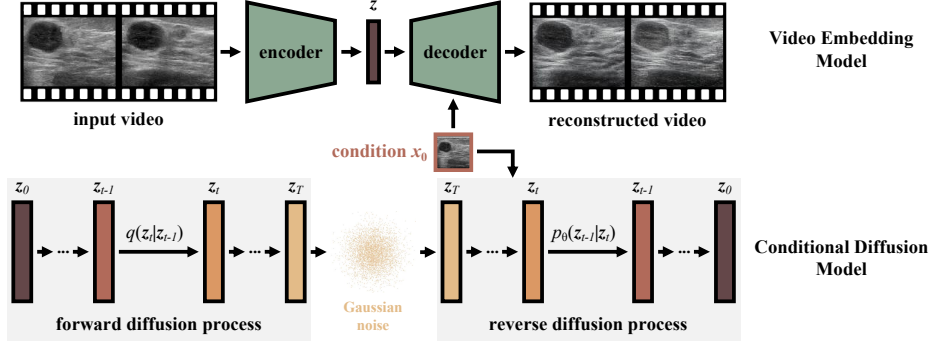


Fig. 1. Overview of the proposed LDDM framework.

adding Gaussian noise based on a variance schedule $\beta_1, \dots, \beta_T$. At time step $t \in [1, \dots, T]$, the transition probability is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t}z_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where β_t is constant. When T is sufficiently large, z_T can be well approximated by a standard Gaussian. This suggests that true posterior $q(z_{t-1}|z_t)$ can be estimated by the learned conditional distribution $p_\theta(z_{t-1}|z_t)$:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t), \sigma_t^2 \mathbf{I}), \quad (2)$$

where σ_t is also constant. Sampling then proceeds in this reverse direction, starting from $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and denoising step-by-step via $p_\theta(z_{t-1}|z_t)$ to obtain $z_0 \sim p_\theta(z_0)$.

To learn $p_\theta(z_{t-1}|z_t)$, Gaussian noise ϵ is added to z_0 to generate samples. Then, conditioned on x_0 , a denoising model ϵ_θ is trained to predict ϵ by minimizing the following mean-squared error loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon - \epsilon_\theta(\mathbf{z}, t, x_0)\|^2], \quad (3)$$

where t is uniformly sampled from $\{1, \dots, T\}$.

Given an input video $\mathbf{v}$, we first obtain its latent dynamic embedding $\mathbf{z}$ using our trained video encoder $\mathcal{E}$. The size of $\mathbf{z}$ is $[I/r, H/r, W/r, C/r]$, where r is a downsampling factor. This is significantly smaller than the original video size, thereby reducing computational requirements. This embedding abstracts away imperceptible details while retaining high-level motion and semantics. We then map $\mathbf{z}$ to a standard Gaussian variable via the forward diffusion process. The initial frame x_0 serves as a conditional input to help generate an interpretable dynamic in the latent space.

Since $\mathbf{z}$ encodes high-level semantic dynamics rather than spatial-temporal details like pixel changes, its dimensions can be much lower than pixel space. As such, the fitting objective of our diffusion model is both meaningful and low-dimensional, easing the generation process. Conditioning on x_0 allows our model

to generate interpretable latent trajectories representing high-level dynamic semantics, as $\mathbf{x}_0$ provides critical context for plausible video dynamics.

2.3 Video Synthesis from Single Images

We synthesize virtual video data using our trained LDDM conditioned on existing images. Specifically, given an image $\mathbf{x}$, we sample a Gaussian noise and gradually denoise it via the reverse diffusion process to obtain a synthesized latent embedding $\mathbf{z}'$. We then pass $\mathbf{z}'$ and $\mathbf{x}$ through the trained decoder $\mathcal{D}$ to generate a video $\mathbf{v}'$. Note that the encoder $\mathcal{E}$ is not required at this stage.

Ultrasound videos synthesized in this manner can augment various downstream tasks depending on conditional images. Here, we generate videos from breast ultrasound images to augment video classification models by balancing classes and exposing models to diverse cross-sections. The key advantage is that LDDM learns interpretable high-level dynamics from real videos which manifest in synthesized videos, enabling realistic data augmentation.

3 Experiments

3.1 Datasets

We utilize the public ultrasound video dataset BUSV [4] to validate the quality of videos generated from initial frames by our LDDM. BUSV comprises 113 malignant videos and 75 benign videos, each with a label indicating the breast lesion type. We create three data splits by randomly selecting 70%, 50%, and 30% of the videos for training, respectively, and use the remaining videos in each split for testing. This allows the comprehensive validation of our approach under different data partition protocols.

Furthermore, we leverage images in the BUSI dataset [12] to synthesize more videos for ultrasound video classification. BUSI contains 445 benign lesion images and 210 malignant lesion images. We use all BUSI images to generate video sequences via our trained LDDM.

3.2 Implement Details

We implement the proposed LDDM framework using PyTorch and conduct experiments on a single NVIDIA RTX 3090Ti GPU. During training, videos of arbitrary duration are segmented into clips of 48 frames, and we preprocess all frames to have the same dimensions. We then uniformly sample 16 frames from each clip and feed them into the encoder. We utilize the Adam optimizer with a learning rate of 1×10^{-5} for 100 iterations. For the conditional diffusion model, we also adopt the Adam optimizer with a learning rate of 0.01 and a batch size of 16 for 50 iterations. All classification models are trained for 50 epochs until convergence on a single NVIDIA RTX 3090Ti GPU.

Table 1. Quantitative evaluation of video synthesis quality. We compare our LDDM against a previous state-of-the-art approach. Both models are conditioned on the initial frame of a video to generate subsequent frames.

method	FVD ↓	DTFVD ↓	LPIPS ↓	DIV ↑
cINN [18]	58.98	0.62	0.33	2.23
LDDM	53.82	0.58	0.32	1.07

3.3 Evaluation Metrics

We employ four common metrics to evaluate the quality of synthesized videos: Fréchet video distance (FVD) [25], dynamic texture Fréchet video distance (DTFVD) [26], learned perceptual image patch similarity (LPIPS) [27], and diversity (DIV) [28] of generated videos. FVD and DTFVD measure holistic similarity between real and synthesized videos, while LPIPS computes patch-level perceptual differences. DIV quantifies variety across generated videos.

For video classification, we report accuracy and F1 score.

3.4 Results

Video Generation We compare our approach against cINN [18], a previous state-of-the-art approach that makes use of a normalizing flow model for video generation. As shown in Table 1, LDDM achieves better performance on FVD, DTFVD, and LPIPS. On the DIV metric, cINN produces a higher diversity score. However, we observe that in terms of downstream task performance, cINN underperforms compared to LDDM.

This potentially indicates a deviation from realistic motion patterns of nodules. In addition, we show example generation results in Fig. 2 and Fig. 3.

Synthesized Videos for Classification Due to the lack of ground truth videos, the quality of generated videos from images in BUSI cannot be directly evaluated using metrics such as FVD. Instead, to demonstrate the interpretability and usability of these synthesized data, we exploit downstream tasks. Specifically, we mix the generated videos with real ultrasound videos to train video classification models. We then evaluate whether the inclusion of these synthetic samples improves models’ performance on the BUSV test set.

We use four video classification models: C3D [21], TSM [22], TDN [23], and ATMNet [24]. For each model, we compare training with only real videos from BUSV against training with a mixture of real videos and synthesized samples from BUSI. We also provide baselines trained using only synthetic data. Upon thorough analysis in Table 2, we observe significant performance gains when augmenting real data with generated videos. Moreover, as the amount of real training videos decreases, the improvement from adding synthetic samples increases. This indicates that the videos generated by the proposed LDDM are

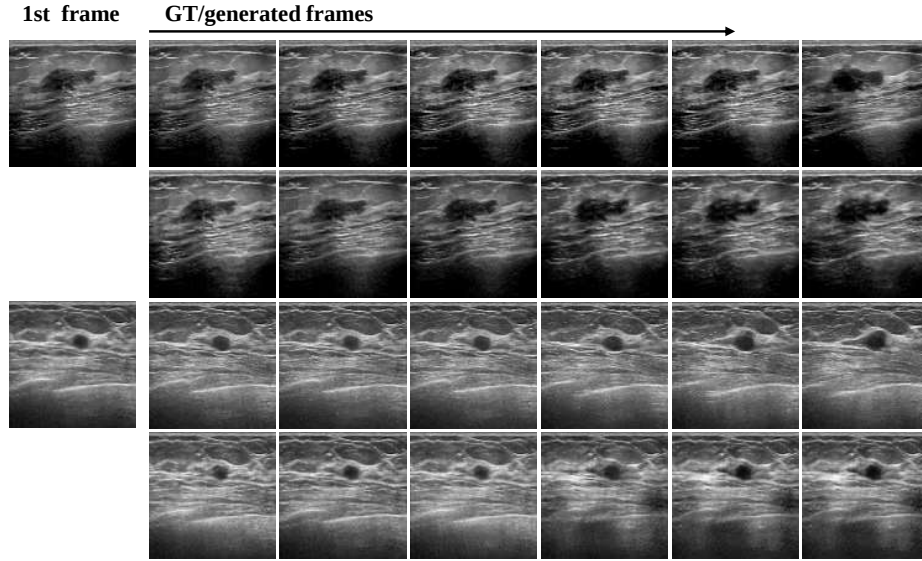


Fig. 2. Visual examples of ultrasound video generation using the proposed LDDM framework. The bottom row shows synthesized video frames generated by our model conditioned on initial frames from the BUSV dataset, which are unseen during training. The top row depicts ground truth frames for reference. As evident from the visual examples, our LDDM approach is capable of generating realistic ultrasound video sequences that are analogous to real ones.

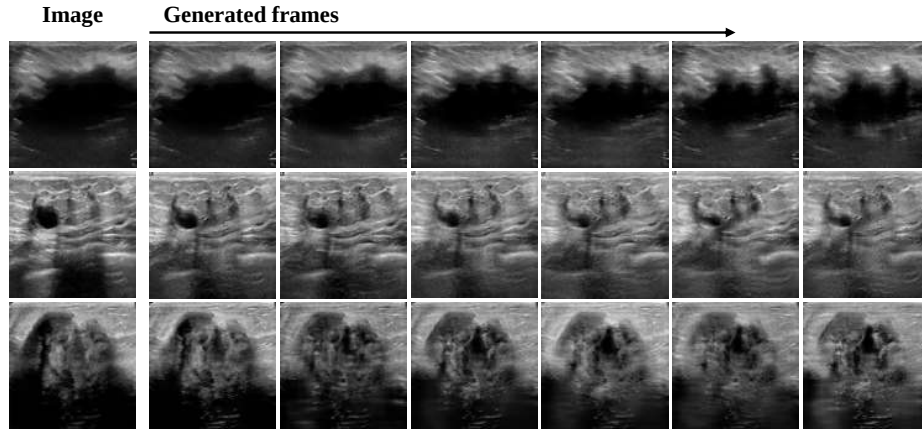


Fig. 3. Visual examples of generated ultrasound videos based on images from the BUSI dataset.

interpretable and beneficial for learning. Our method can thus serve as an effective data augmentation strategy in scenarios with limited access to real medical video data.

Table 2. Evaluation of video classification performance using models trained on different data sources: (1) real ultrasound videos only, (2) synthesized videos conditioned on static BUSI images only, and (3) a combination of real and synthetic data. We compare the downstream task performance of our proposed LDDM approach (black) against cINN (gray).

Method	real data ratio	real only		synthetic only		real+synthetic	
		F1	Acc	F1	Acc	F1	Acc
C3D	7: 3	70.96	73.21	32.89	44.64	74.15	75.00
				59.08	58.93	75.20	75.00
	5: 5	64.85	68.81	45.42	49.46	66.67	67.74
				62.58	64.52	67.92	67.74
	3: 7	60.12	64.61	44.47	49.23	64.56	65.38
				57.75	57.69	66.08	66.15
TSM	7: 3	78.91	78.57	67.86	67.86	76.03	76.79
				72.83	72.83	80.96	80.36
	5: 5	73.10	73.21	66.59	66.59	76.77	76.34
				72.30	72.30	77.52	77.42
	3: 7	72.96	73.07	57.30	57.30	71.83	72.31
				72.59	72.59	78.96	79.23
TDN	7: 3	83.03	83.93	69.21	69.21	78.62	78.57
				71.93	71.93	83.66	83.93
	5: 5	72.96	72.73	69.79	69.79	76.87	77.21
				71.93	71.93	80.40	78.66
	3: 7	63.92	65.38	69.07	69.07	72.15	72.67
				69.65	69.65	75.00	75.38
ATMNet	7: 3	72.08	73.21	69.52	69.52	73.34	75.00
				70.24	70.24	78.70	78.57
	5: 5	72.70	74.19	73.31	73.31	75.12	75.27
				77.64	77.64	75.62	76.34
	3: 7	66.91	66.92	56.19	56.19	66.91	66.92
				64.97	64.97	70.72	72.31

These compelling results illustrate that LDDM can produce realistic videos which readily transfer to downstream tasks, ultimately enhancing classification accuracy across various models. The proposed approach shows promise in addressing the scarcity of labeled video data.

4 Conclusion

Ultrasound videos provide a more comprehensive and dynamic view of anatomical structures and lesions compared to static ultrasound images. However, acquiring rich ultrasound videos poses significant challenges, as only single images are typically reported and stored in clinical practice. In this work, we propose a generative model called LDDM to synthesize ultrasound videos from ultrasound images using a diffusion model-based approach. By altering the input images, high levels of variation can be introduced when generating such synthetic videos.

We demonstrate that improvements in both video generation quality and downstream task performance verify the robustness and practicality of LDDM. An interesting area for future work is prompt-driven image-to-video synthesis, where textual prompts can control video generation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Che, H., Ramanathan, S., Foran, D.J., Noshier, J.L., Patel, V.M., Hacıhaliloglu, I.: Realistic ultrasound image synthesis for improved classification of liver disease. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 179–188. (2021)
2. Hung, A.L.Y., Galeotti, J.: Ultrasound variational style transfer to generate images beyond the observed domain. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 14–23. (2021)
3. Sun, A., Zhang, Z., Lei, M., Dai, Y., Wang, D., Wang, L.: Boosting breast ultrasound video classification by the guidance of keyframe feature centers. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 441–451. (2023)
4. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 614–623. (2022)
5. Qin, C., Cao, J., Fu, H., Anwer, R.M., Khan, F.S.: A spatial-temporal deformable attention based framework for breast lesion detection in videos. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 479–488. (2023)
6. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 497–507. (2023)
7. Huang, B., Xu, Z., Chan, S.C., Liu, Z., Wen, H., Hou, C., Huang, Q., Jiang, M., Dong, C., Zeng, J.: A style transfer-based augmentation framework for improving segmentation and classification performance across different sources in ultrasound images. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 44–53. (2023)
8. Byra, M., Klimonda, Z., Litniewski, J.: Pre-training with simulated ultrasound images for breast mass segmentation and classification. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 34–45. (2023)
9. Reynaud, H., Qiao, M., Dombrowski, M., Day, T., Razavi, R., Gomez, A., Leeson, P., Kainz, B.: Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 142–152. (2023)
10. Reynaud, H., Vlontzos, A., Dombrowski, M., Gilligan, L. C., Beqiri, A., Leeson, P., Kainz, B.: D’ARTAGNAN: Counterfactual video generation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 599–609. (2022)

11. Liang, J., Yang, X., Huang, Y., Liu, K., Zhou, X., Hu, X., Lin, Z., Luo, H., Zhang, Y., Xiong, Y.: Weakly-supervised high-fidelity ultrasound video synthesis with feature decoupling. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 310–319. (2022)
12. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data in Brief* **43**, 104863 (2020)
13. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.G.T.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. In: IEEE Winter Conference on Applications of Computer Vision, pp. 2198–2210. (2019)
14. Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A., Gomez, A.: Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation. *arXiv preprint arXiv:2305.05424* (2023)
15. Tom, F., Sheet, D.: Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. In: IEEE International Symposium on Biomedical Imaging, pp. 1174–1177. (2018)
16. Liang, J., Yang, X., Li, H., Wang, Y., Van, M.T., Dou, H., Chen, C., Fang, J., Liang, X., Mai, Z.: Synthesis and edition of ultrasound images via sketch guided progressive growing GANS. In: IEEE International Symposium on Biomedical Imaging, pp. 1793–1797. (2020)
17. Zhou, Z., Guo, Y., Wang, Y.: Handheld ultrasound video high-quality reconstruction using a low-rank representation multipathway generative adversarial network. *IEEE Transactions on Neural Networks and Learning Systems* **32**(2), 575–588 (2021)
18. Dorkenwald, M., Milbich, T., Blattmann, A., Rombach, R., Derpanis, K.G., Ommer, B.: Stochastic image-to-video synthesis using cINNs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3742–3753. (2021)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695. (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, pp. 6840–6851. (2020)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: IEEE International Conference on Computer Vision, pp. 4489–4497. (2015)
22. Lin, J., Gan, C., Han, S.: TSM: Temporal shift module for efficient video understanding. In: IEEE/CVF International Conference on Computer Vision, pp. 7083–7093. (2019)
23. Wang, L., Tong, Z., Ji, B., Wu, G.: TDN: Temporal difference networks for efficient action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1895–1904. (2021)
24. Wei, G., Zhang, Z., Lan, C., Lu, Y., Chen, Z.: Active token mixer. In: AAAI Conference on Artificial Intelligence, pp. 2759–2767. (2023)
25. Unterthiner, T., Van S.S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018)
26. Hadji, I., Wildes, R.P.: A new large scale dynamic texture dataset with application to ConvNet understanding. In: European Conference on Computer Vision, pp. 320–335. (2018)
27. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in Neural Information Processing Systems, pp. 658–666. (2016)

28. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523 (2018)
29. Tero K., Samuli L., Timo A.: A style-based generator architecture for generative adversarial networks. In : IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4401–4410. (2019)
30. Xinchun Y., Jimei Y., Kihyuk S., Honglak L.: Attribute2Image: Conditional image generation from visual attributes. In: European Conference on Computer Vision, pp.776–791. (2016)