

Anatomy-Guided Residual Motion Diffusion for Controllable 4D Cardiac MRI Synthesis

Yiheng Cao¹[0009-0001-5271-0894], Gustavo Andrade-Miranda²[0000-0002-6499-5655], Jiatian Zhang¹, Lingxiao Zhao³, and Xin Gao¹

¹ Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, Suzhou, China

caoyh@sibet.ac.cn

² Systemes Complexes et Intelligence Artificielle, IMT Mines Ales, Ales, France

³ School of Artificial Intelligence and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China

Abstract. Developing robust artificial intelligence models for 4D (3D + time) medical imaging is constrained by limited annotated data, inter-device domain shifts, and privacy restrictions. To address this, we propose a 4D controllable generative framework for anatomically consistent data augmentation. The model generates both volumetric intensity sequences and their corresponding, inherently aligned segmentation masks. A semi-supervised variational autoencoder learns a compact latent representation of anatomical volumes while jointly predicting aligned segmentation masks in a unified framework. Anatomical structure is then disentangled from temporal dynamics through a cascaded latent diffusion model (LDM). A static LDM generates subject-specific anatomy conditioned on clinical priors (diagnosis and volumes measures) and a subsequent motion LDM estimates residual latent motions, ensuring strict temporal coherence across the 4D sequence. The proposed approach was evaluated on cine cardiac MRI as a representative 4D imaging application. Experiments across multiple datasets demonstrate high controllability of static anatomy (Pearson $r > 0.8$) and strong temporal coherence (FVD = 288.08). In cross-vendor generalization experiments, augmenting training sets with synthetic 4D sequences significantly improves downstream segmentation performance. Using nnU-Net, the proposed augmentation strategy improves the average Dice score by 1.4% and reduces the Hausdorff Distance by 3.0 mm compared to training on real data alone, for the left ventricle, Dice improves by 2.8% with a 5.4 mm reduction in boundary error. Overall, this framework provides a scalable and controllable solution for 4D medical image synthesis, supporting the development of more robust models with limited annotations and cross-vendor variability. Code available on <https://github.com/cyiheng/4DCardiacMRISynthesis>.

Keywords: 4D cardiac MRI · Generation model · Latent diffusion model

1 Introduction

Cardiovascular diseases remain the leading cause of mortality worldwide and require accurate diagnosis. Cine cardiac magnetic resonance (CMR) imaging is the gold standard for assessing cardiac function, providing high-resolution spatiotemporal data on ventricular volumes and myocardial wall motion. Recent advances in deep learning (DL) have improved automated CMR analysis [3], data acquisition and sharing remain limited by privacy regulations, resulting in small multi-center datasets and domain shifts that impair generalization [8,12]. DL methods also rely on large, expertly annotated datasets, a constraint that is particularly critical for 4D (3D + time) data, where full-cycle delineation is labor-intensive and time-consuming [17]. Consequently, annotations are often restricted to end-diastolic (ED) and end-systolic (ES) phases, discarding substantial spatiotemporal information [3,17].

Generative modeling offers a potential solution through realistic data synthesis [7,14]. Medical image generation relies on both physics-based simulations and data-driven approaches, with modern deep generative models learning high-dimensional statistical distributions directly from data [6,13,19]. However, 4D cine CMR generation remains challenging due to high dimensionality and the need for physiologically plausible cardiac motion. Early GAN-based methods are primarily 2D-based, either model cine dynamics via latent trajectories with limited anatomical control or rely on label-conditioned synthesis that does not scale to volumetric 4D data without dense supervision [2,18]. In contrast, more recent 4D diffusion-based approaches improve perceptual fidelity and enable conditional generation using clinical attributes or textual prompts [15,20]. However, most existing methods combine anatomical structures and temporal dynamics in a single generative process, limiting independent controllability and long-term consistency. Additionally, many approaches either focus solely on intensity generation without paired segmentation masks or depend on densely annotated, often proprietary datasets, limiting scalability, reproducibility, and cross-institutional validation [5,20]. Despite promising visual realism, current generative models remain insufficient for producing anatomically controllable, label-aware 4D cine CMR data suitable for robust downstream learning tasks.

In this work, we bridge these gaps by proposing a robust 4D generation pipeline for cine CMR that directly generates anatomically aligned image-mask pairs, distinguishing it from prior methods that primarily focus on intensity synthesis. We simplify the 4D synthesis problem by decoupling it into static anatomical generation and residual latent motion prediction, enabling anatomically consistent and physiologically plausible cardiac dynamics. Our contribution is threefold:

- We use a semi-supervised variational autoencoder (VAE) trained on both labeled and unlabeled data to learn a shared anatomical-semantic latent representation, enabling direct synthesis of paired cardiac MR volumes and corresponding segmentation masks.
- We propose an ED-referenced residual latent motion model that represents cardiac dynamics independently of anatomy, enabling the generation of tem-

porally coherent 4D cine sequences while preserving subject-specific anatomical consistency.

- We demonstrate that adding synthetic data improves the cross-vendor generalization of cardiac segmentation models by reducing left ventricle (LV), right ventricle (RV), and myocardium (MYO) boundary errors, enabling reliable clinical quantification of ventricular volumes and ejection fraction.

2 Methods

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times T}$ denote a 4D cardiac MRI sequence with T frames. We model $p(\mathbf{x}|\mathbf{c})$ conditioned on clinical priors $\mathbf{c}$ (e.g., diagnosis, volumetric indices) through three components: (1) a semi-supervised VAE for latent compression, (2) a static latent diffusion model (LDM) for anatomical generation, and (3) a residual motion LDM for temporal dynamics (Figure 1).

2.1 Semi-supervised Latent Representation

We adopt a semi-supervised design to learn a shared latent representation from both labeled and unlabeled data. The learned latent space enables the downstream LDM to generate anatomically consistent image-segmentation pairs. The input space is first mapped to a low-dimensional latent space $\mathcal{Z}$. We use a 3D VAE-GAN with a shared encoder E and two decoders: an intensity reconstruction decoder D_{rec} and a segmentation decoder D_{seg} . Given an input volume x , the encoder produces a latent distribution $z \sim E(x)$. The training objective is a composite loss that ensures reconstruction fidelity, latent regularization, semantic alignment, and textural realism:

$$\begin{aligned} \mathcal{L}_{VAE} = & \lambda_{pixel} \mathcal{L}_{pixel}(x, D_{rec}(z)) + \lambda_{KL} \mathcal{L}_{KL}(q(z|x)||p(z)) \\ & + \mathbb{I}_{label} \cdot \lambda_{seg} \mathcal{L}_{seg}(y, D_{seg}(z)), \end{aligned} \quad (1)$$

where $\mathcal{L}_{pixel}$ includes L_1 , perceptual, and adversarial terms, $\mathcal{L}_{KL}$ regularizes the latent posterior $q(z|x)$ towards a standard Gaussian prior $p(z)$ to facilitate valid sampling. The cross-entropy $\mathcal{L}_{seg}$ is applied only when labeled data y are available ($\mathbb{I}_{label} = 1$).

2.2 Static Anatomy Generation

We train a static LDM that generates the ED latent base anatomy $\hat{z}_{ED}$ conditioned on clinical priors $\mathbf{c}$ (categorical diagnosis, slice count, and ED/ES/ejection fraction volumes). This conditioning enables controllable synthesis of clinically relevant phenotypes, facilitating targeted augmentation of disease subgroups. A Multi-Layer Perceptron (MLP) τ projects $\mathbf{c}$ to condition the LDM through cross-attention. To handle variable slice counts within a fixed latent resolution, a through-plane binary mask indicating valid slices is concatenated during training and sampling. This prevents the model from learning padded regions, ensuring center-cropped decoded volumes remain anatomically consistent without boundary artifacts.

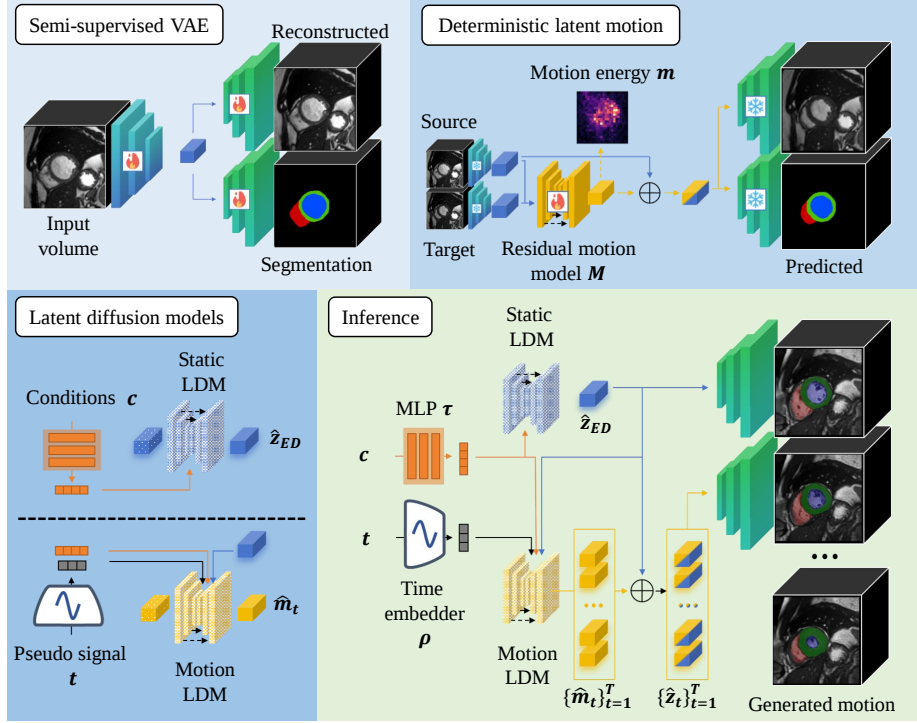


Fig. 1. Overview of the proposed framework. Blue panels: VAE compression, deterministic residual motion extraction, and cascaded LDM training. Green panel: the static LDM generates base anatomy $\hat{z}_{ED}$ from conditions $\mathbf{c}$, while the motion LDM predicts temporal residuals $\hat{m}_t$. These are aggregated ($\hat{z}_{ED} + \hat{m}_t$) and passed through the VAE decoders to produce the final 4D volumes and inherently aligned segmentation masks.

2.3 Residual Motion Dynamics Generation

We model the temporal evolution of the cardiac cycle as a residual offset relative to the ED frame. For any time frame t , we suppose that the latent representation z_t can be expressed as $z_t = z_{ED} + m_t$, where m_t represents the residual latent motion. Motion modeling is performed in two stages: deterministic residual extraction and stochastic residual generation.

Learning the Residual Latent Motion Space. A deterministic latent motion predictor M is trained to learn the residual mapping between paired frames (z_{ED}, z_t) . The network predicts: $\hat{m}_t = M(z_{ED}, z_t)$ with $\hat{z}_t = z_{ED} + \hat{m}_t$. To preserve high-frequency details and semantic integrity, the model is supervised in both latent and image space using:

$$\mathcal{L}_{motion} = \lambda_{pixel} \mathcal{L}_{pixel}(x_t, D_{rec}(\hat{z}_t)) + \mathbb{I}_{label} \cdot \lambda_{seg} \mathcal{L}_{seg} + \mathcal{L}_{reg} \quad (2)$$

The regularization term

$$\mathcal{L}_{reg} = \lambda_{sparse} |\hat{m}|_2^2 + \lambda_{identity} |M(z_{ED}, z_{ED})|_1 \quad (3)$$

encourages minimal latent motion energy and enforces an identity constraint. This ensures zero residuals for identical input frames and prevents autoencoding behavior, forcing M to learn true relative motion to preserve anatomical consistency.

Motion Diffusion Model. Once trained, the frozen M extracts residuals m_t to train a conditional motion LDM. This LDM is conditioned on (i) the static reference z_{ED} through concatenation, (ii) the clinical embeddings $\tau(\mathbf{c})$, and (iii) a sinusoidal time embeddings $\rho(t)$ representing the normalized cardiac phase.

During inference, the pipeline samples $\hat{z}_{ED}$ from the static LDM, then samples a sequence of residuals $\{\hat{m}_t\}_{t=1}^T$ from the motion LDM. These residuals are aggregated and passed through the pre-trained VAE decoders to jointly reconstruct the final 4D intensity sequence and its inherently aligned segmentation masks.

3 Experiments and Results

Datasets. Our framework was trained on a combined dataset of 956 patients, including 100 patients with annotations at ED and ES from the ACDC dataset [3] and 856 unlabeled patients from the Kaggle Data Science Bowl (DSB) dataset [1]. The semi-supervised VAE was trained using all cases for reconstruction, while segmentation supervision was provided exclusively by the 100 labeled ACDC cases. Internal evaluation was performed on 50 patients of the ACDC test set and a held-out subset of 50 patients from DSB dataset. Cross-vendor generalization was assessed using an external test sets: 134 and 157 patients from the M&Ms and M&Ms2, respectively [4,16].

Implementation details. All models were implemented in PyTorch using a single NVIDIA RTX 4090 GPU. The images were resampled at a fixed spacing of $1 \times 1 \times 10$ mm and centrally cropped to $192 \times 192 \times Z$ where $Z \in [5, 16]$ denotes the number of slices. Intensities were clipped to the $[0.5, 99.5]$ percentiles and normalized to the range $[-1, 1]$. To handle variable slice counts in Z , volumes are padded using symmetric edge replication. This preserves boundary intensities, avoiding artificial artifacts and bias in the latent space, while ensuring anatomically consistent borders without introducing unintended intensity changes during training.

The semi-supervised VAE was trained for 100,000 steps using AdamW ($lr = 10^{-4}$), with $\lambda_{seg} = 20.0$ and $\lambda_{KL} = 10^{-7}$. The reconstruction loss combined L_1 , perceptual, and adversarial terms weighted at 1.0, 0.3, and 0.1, respectively. The latent motion predictor was trained for 100,000 steps (AdamW, $lr = 2 \times 10^{-4}$) with $\lambda_{seg} = 10.0$, $\lambda_{sparse} = 0.01$, $\lambda_{identity} = 1.0$, and a reconstruction loss

weighted at 2.0, 1.0, and 0.01 (L_1 , perceptual, adversarial). Both LDM were trained for 200,000 steps using an 8-bit AdamW optimizer ($lr = 2 \times 10^{-5}$) and DDIM sampling with 50 inference steps. Static LDM training included random affine scaling and pseudo-ED sampling, with clinical volumes dynamically re-computed using frozen VAE segmentation outputs.

Evaluation metrics. We evaluate the framework based on perceptual quality, anatomical controllability, and utility for downstream segmentation task. Perceptual fidelity and temporal coherence are quantified using the Fréchet Inception Distance (FID) and Fréchet Video Distance (FVD), respectively. FID is computed on static ED volumes, while FVD is evaluated on full cine sequences. FID is computed using 10-fold bootstrapping with 2,000 generated volumes and 956 real training cases. For FVD, 1,000 synthetic 4D sequences were unpacked along the longitudinal (z) axis into temporal 2D+t video clips and compared against real training clips. FVD uncertainty was quantified via 100-iteration non-parametric bootstrap resampling at the video level. Anatomical controllability was evaluated by calculating the Pearson correlation between the clinical priors input and synthetic measurements across 20 cases per pathology using three classifier-free guidance (CFG) scales. To assess practical clinical utility, we use semantic segmentation as a representative downstream task to rigorously validate that synthetic data augmentation preserves anatomical boundaries and enhances cross-domain robustness. Segmentation models (nnU-Net [11], UNETR [10], and Swin-UNETR [9]) are trained using 900 synthetic ED/ES volume pairs and evaluated using Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD), reported as mean and standard deviation over five runs on internal (ACDC) and external (M&Ms, M&Ms2) test sets for the right ventricle (RV), myocardium (MYO), and left ventricle (LV).

Generation fidelity. The Figure 2 displays synthetic sequences for abnormal right ventricle (ARV) and hypertrophic cardiomyopathy (HCM) pathologies (full motion available in the Supplementary Material). The framework reproduces anatomical features, such as myocardial thickening in HCM, with segmentation masks that remain aligned with generated intensities across the cardiac cycle. The difference magnitude maps demonstrate substantial regional changes. Quantitative perceptual metrics support visual assessment, with a slice-wise FID of 72.21 ± 0.95 and FVD of 288.08 ± 3.63 . Due to differing experimental settings, direct comparison is not possible. TexDC [15] reported an FID of 56.13 and an FVD of 756.92 on a similar public data mix, providing qualitative context for the temporal coherence and anatomical realism achieved by our model.

Figure 3 evaluates the ability of the proposed framework to generate specific cardiac volumes based on clinical priors. Pearson correlation coefficients vary across pathologies and guidance scales, ranging from moderate to strong ($r \in [0.66, 0.94]$), with peak performance reaching $r = 0.94$. However, higher CFG values also introduce a trade-off by potentially increasing the deviation from the absolute target volume while maintaining the desired volumetric trend.

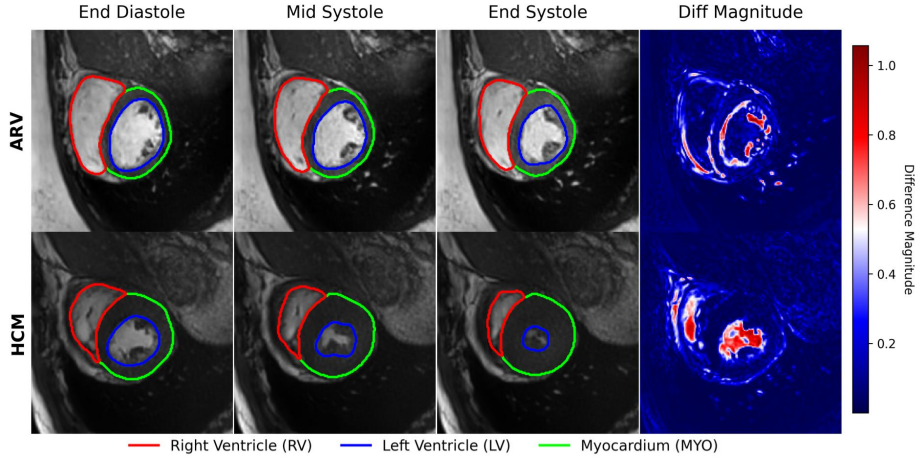


Fig. 2. Example of synthetic volumes and segmentation across pathological classes.

Segmentation evaluation. Table 1 summarizes the results using nnU-Net as segmentation backbone. On the internal ACDC dataset, the addition of synthetic data improves the LV DSC and significantly reduces HD across all structures. On the M&Ms dataset, adding synthetic data improves the LV DSC by 3.4% (from 84.8 to 88.2) and greatly reduces the LV HD by 6.1 mm (from 14.5 to 8.4). Similar gains are observed on M&Ms2, with a 4.7 mm reduction in LV boundary error. These results, validated on nearly 300 unseen multi-vendor cases, suggest that our motion LDM generates anatomically coherent dynamics that help reduce the domain gap between different scanner vendors.

We extended our evaluation to UNETR and Swin-UNETR (Table 2) to show that these improvements are not framework-specific. Vision Transformers typically require large-scale data to learn inductive biases and struggle with small datasets like ACDC. Our synthetic augmentation acts as a regularizer, providing

Table 1. Segmentation performance using nnU-Net as backbone.

Dataset	Data	DSC			HD		
		RV	MYO	LV	RV	MYO	LV
ACDC	nnU-Net	91.0$\pm$0.08	88.9$\pm$0.04	93.1 $\pm$ 0.1	4.9 $\pm$ 0.1	2.8 $\pm$ 0.09	3.3 $\pm$ 0.3
	+Synth ED	90.8 $\pm$ 0.09	88.3 $\pm$ 0.02	93.2 $\pm$ 0.05	4.1 $\pm$ 0.3	2.7 $\pm$ 0.04	3.0 $\pm$ 0.06
	+Synth ED/ES	90.9 $\pm$ 0.1	88.7 $\pm$ 0.05	93.9$\pm$0.06	4.1$\pm$0.2	2.7$\pm$0.06	2.7$\pm$0.09
M&Ms	nnU-Net	87.0 $\pm$ 0.2	81.8 $\pm$ 0.1	84.8 $\pm$ 0.7	7.6 $\pm$ 0.4	8.6 $\pm$ 0.3	14.5 $\pm$ 0.9
	+Synth ED	87.5$\pm$0.07	82.3 $\pm$ 0.1	87.5 $\pm$ 0.3	6.6 $\pm$ 0.1	6.5 $\pm$ 0.1	8.9 $\pm$ 0.5
	+Synth ED/ES	87.4 $\pm$ 0.1	82.5$\pm$0.1	88.2$\pm$0.1	6.5$\pm$0.2	6.4$\pm$0.1	8.4$\pm$0.3
M&Ms2	nnU-Net	85.9 $\pm$ 0.2	80.1 $\pm$ 0.1	86.2 $\pm$ 0.4	10.4 $\pm$ 0.8	9.0 $\pm$ 0.2	13.7 $\pm$ 1.1
	+Synth ED	86.3 $\pm$ 0.1	80.5 $\pm$ 0.1	88.0 $\pm$ 0.1	9.1 $\pm$ 0.1	6.7 $\pm$ 0.1	9.4 $\pm$ 0.2
	+Synth ED/ES	86.6$\pm$0.4	81.1$\pm$0.7	88.3$\pm$0.1	9.1$\pm$1.2	6.7$\pm$0.2	8.9$\pm$0.4

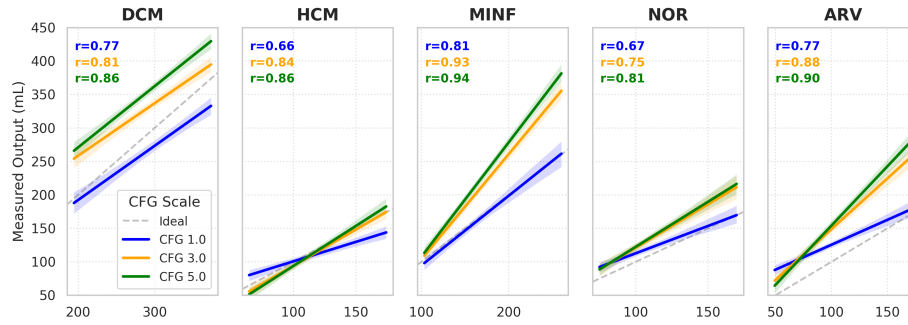


Fig. 3. Correlation between input clinical volume priors and measured synthetic volumes across different pathologies and CFG scales. Abbreviations: DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; MINF, myocardial infarction; NOR, normal; ARV, abnormal right ventricle; CFG, classifier-free guidance.

performance boosts. On the M&Ms dataset, the UNETR model achieves a 12.1% improvement in LV DSC (from 64.2 to 76.3) when trained with our synthetic data. Overall, these findings demonstrate that the proposed framework provides a promising solution for 4D data augmentation that can benefit data-hungry models under domain shift.

4 Conclusion

In this study, we developed a novel generative framework for 4D medical image synthesis, demonstrated through the explicit disentanglement of static anatomy and temporal dynamics via residual latent motion diffusion. By leveraging a semi-supervised VAE and latent motion priors, our approach generates paired 4D intensity volumes and segmentation masks. The results demonstrate that this approach significantly enhances the robustness and cross-vendor generalization of downstream tasks such as segmentation on cine CMR application. Future work will focus on extending this framework to a broader range of 4D medical imaging modalities characterized by complex spatiotemporal dynamics and annotation scarcity. Furthermore, the utility of the synthesized 4D sequences will be evaluated for additional clinical downstream tasks to further establish the universal benefits of this generative approach.

Acknowledgments. This work is supported by the Research Fund for International Young Scientists of the National Natural Science Foundation of China (grant number W2533209).

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

Table 2. Segmentation performance using UNETR and Swin-UNETR as backbone.

Dataset	Data	DSC			HD		
		RV	MYO	LV	RV	MYO	LV
ACDC	UNETR	72.5 $\pm$ 0.8	68.9 $\pm$ 0.4	83.1 $\pm$ 0.3	19.0 $\pm$ 0.7	7.8 $\pm$ 0.1	8.5 $\pm$ 0.3
	+Synth ED/ES	77.7$\pm$0.5	76.4$\pm$0.3	87.0$\pm$0.3	14.3$\pm$1.2	6.4$\pm$0.2	6.8$\pm$0.3
	Swin-UNETR	79.1 $\pm$ 0.2	75.1 $\pm$ 0.2	86.9 $\pm$ 0.2	12.7 $\pm$ 1.0	6.6 $\pm$ 0.2	7.2 $\pm$ 0.5
	+Synth ED/ES	82.7$\pm$0.5	81.2$\pm$0.5	89.8$\pm$0.3	9.5$\pm$0.2	5.1$\pm$0.2	5.5$\pm$0.5
M&Ms	UNETR	52.3 $\pm$ 4.5	50.3 $\pm$ 6.3	64.2 $\pm$ 6.4	37.0 $\pm$ 4.4	31.8 $\pm$ 4.2	35.2 $\pm$ 1.1
	+Synth ED/ES	66.5$\pm$0.8	65.9$\pm$1.2	76.3$\pm$1.6	20.2$\pm$0.7	16.0$\pm$0.5	18.9$\pm$0.6
	Swin-UNETR	56.6 $\pm$ 1.3	60.7 $\pm$ 0.7	72.4 $\pm$ 0.9	50.3 $\pm$ 2.7	30.2 $\pm$ 2.0	35.4 $\pm$ 1.8
	+Synth ED/ES	73.7$\pm$0.5	74.4$\pm$1.2	82.4$\pm$0.2	15.9$\pm$1.9	10.5$\pm$0.1	13.9$\pm$0.9
M&Ms2	UNETR	52.1 $\pm$ 2.6	55.1 $\pm$ 2.7	71.7 $\pm$ 2.8	31.5 $\pm$ 3.4	20.3 $\pm$ 2.2	22.8 $\pm$ 1.1
	+Synth ED/ES	65.9$\pm$0.9	65.6$\pm$0.4	80.1$\pm$0.4	21.1$\pm$1.0	13.0$\pm$0.2	14.6$\pm$0.8
	Swin-UNETR	52.7 $\pm$ 2.2	59.1 $\pm$ 0.5	76.3 $\pm$ 0.8	41.0 $\pm$ 2.8	19.9 $\pm$ 0.8	21.7 $\pm$ 1.2
	+Synth ED/ES	62.6$\pm$2.0	68.5$\pm$0.7	81.9$\pm$0.3	22.9$\pm$0.99	11.5$\pm$0.2	14.4$\pm$1.0

References

1. Data science bowl cardiac challenge data (Dec 2015), <https://kaggle.com/second-annual-data-science-bowl>
2. Amirrajab, S., Al Khalil, Y., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M.: Label-informed cardiac magnetic resonance image synthesis through conditional generative adversarial networks. *Computerized Medical Imaging and Graphics* **101**, 102123 (Oct 2022). <https://doi.org/10.1016/j.compmedimag.2022.102123>
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohé, M.M., Pennec, X., Sermesant, M., Isensee, F., Jäger, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Išgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transactions on Medical Imaging* **37**(11), 2514–2525 (Nov 2018). <https://doi.org/10.1109/TMI.2018.2837502>
4. Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Viladés, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavuş, E., Petersen, S.E., Escalera, S., Seguí, S., Rodríguez-Palomares, J.F., Lekadir, K.: Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Transactions on Medical Imaging* **40**(12), 3543–3554 (Dec 2021). <https://doi.org/10.1109/TMI.2021.3090082>

5. Dou, H., Huang, J., Zakeri, A., Zhou, Z., Mu, T., Duan, J., Frangi, A.F.: 4D CardioSynth: Synthesising Dynamic Virtual Heart Populations Through Spatiotemporal Disentanglement. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. pp. 3–12. Springer Nature Switzerland, Cham (2026)
6. Frangi, A.F., Tsafaris, S.A., Prince, J.L.: Simulation and Synthesis in Medical Imaging. *IEEE transactions on medical imaging* **37**(3), 673–679 (Mar 2018). <https://doi.org/10.1109/TMI.2018.2800298>
7. Gao, C., Killeen, B.D., Hu, Y., Grupp, R.B., Taylor, R.H., Armand, M., Unberath, M.: Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nature Machine Intelligence* **5**(3), 294–308 (Mar 2023). <https://doi.org/10.1038/s42256-023-00629-1>
8. Guo, B., Lu, D., Szumel, G., Gui, R., Wang, T., Konz, N., Mazurowski, M.A.: The Impact of Scanner Domain Shift on Deep Learning Performance in Medical Imaging: an Experimental Study (Oct 2024). <https://doi.org/10.48550/arXiv.2409.04368>
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images (Jan 2022). <https://doi.org/10.48550/arXiv.2201.01266>
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D Medical Image Segmentation. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 1748–1758. IEEE, Waikoloa, HI, USA (Jan 2022). <https://doi.org/10.1109/WACV51458.2022.00181>
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**(2), 203–211 (Feb 2021). <https://doi.org/10.1038/s41592-020-01008-z>
12. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2**(6), 305–311 (Jun 2020). <https://doi.org/10.1038/s42256-020-0186-1>
13. Kazerooni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacıhaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis* **88**, 102846 (Aug 2023). <https://doi.org/10.1016/j.media.2023.102846>
14. Koetzier, L.R., Wu, J., Mastrodicasa, D., Lutz, A., Chung, M., Koszek, W.A., Pratap, J., Chaudhari, A.S., Rajpurkar, P., Lungren, M.P., Willemink, M.J.: Generating Synthetic Data for Medical Imaging. *Radiology* **312**(3), e232471 (Sep 2024). <https://doi.org/10.1148/radiol.232471>
15. Liu, C., Yuan, X., Yu, Z., Wang, Y.: TexDC: Text-Driven Disease-Aware 4D Cardiac Cine MRI Images Generation. In: Cho, M., Laptev, I., Tran, D., Yao, A., Zha, H. (eds.) *Computer Vision – ACCV 2024*. pp. 191–208. Springer Nature Singapore, Singapore (2025)
16. Martín-Isla, C., Campello, V.M., Izquierdo, C., Kushibar, K., Sendra-Balcells, C., Gkontra, P., Sojoudi, A., Fulton, M.J., Arega, T.W., Punithakumar, K., Li, L., Sun, X., Al Khalil, Y., Liu, D., Jabbar, S., Queirós, S., Galati, F., Mazher, M., Gao, Z., Beetz, M., Tautz, L., Galazis, C., Varela, M., Hüllebrand, M., Grau, V., Zhuang, X., Puig, D., Zuluaga, M.A., Mohy-ud Din, H., Metaxas, D., Breeuwer, M., van der Geest, R.J., Noga, M., Bricq, S., Rentschler, M.E., Guala, A., Petersen,

- S.E., Escalera, S., Palomares, J.F.R., Lekadir, K.: Deep Learning Segmentation of the Right Ventricle in Cardiac MRI: The M&Ms Challenge. *IEEE Journal of Biomedical and Health Informatics* **27**(7), 3302–3313 (Jul 2023). <https://doi.org/10.1109/JBHI.2023.3267857>
17. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis* **63**, 101693 (Jul 2020). <https://doi.org/10.1016/j.media.2020.101693>
 18. Vukadinovic, M., Kwan, A.C., Li, D., Ouyang, D.: GANcMRI: Cardiac magnetic resonance video generation and physiologic guidance using latent space prompting. In: *Proceedings of the 3rd Machine Learning for Health Symposium*. pp. 594–606. PMLR (Dec 2023), <https://proceedings.mlr.press/v225/vukadinovic23a.html>
 19. Wang, S., Zhou, X., Li, C., Wang, S., Li, Y., Tan, T., Zheng, H.: Generative Artificial Intelligence in Medical Imaging: Foundations, Progress, and Clinical Translation. *Research* **8**, 1029 (Dec 2025). <https://doi.org/10.34133/research.1029>
 20. You, X., Zhang, M., Zhang, H., Yang, J., Navab, N.: Temporal Differential Fields for 4D Motion Modeling via Image-to-Video Synthesis. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. pp. 606–616. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-05114-1_58