Controllable latent diffusion model to evaluate the performance of cardiac segmentation methods

Romain Deleat-besson[1] *, Celia Goujat[1], Olivier Bernard[1], Pierre Croisille[1,3], Magalie Viallon[1,3], and Nicolas Duchateau[1,2]

[1] Universite Claude Bernard Lyon 1, INSA-Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France
[2] Institut Universitaire de France (IUF), Paris, France
[3] Department of Radiology, CHU Saint-Etienne, UJM Saint-Etienne, France

Abstract. In medical imaging, the evaluation of segmentation methods remains confined to a limited set of metrics (e.g. Dice coefficient and Hausdorff distance) and annotated datasets with restricted size and diversity. Besides, segmentation is often a preliminary step for extracting relevant biomarkers, accentuating the need to redirect evaluation efforts towards this objective. To address this, we propose an original methodology to evaluate segmentation methods, based on the generation of realistic synthetic images with explicitly controlled biomarker values. Image synthesis is based on *Stable Diffusion*, conditioned by either a 1D vector (clinical attributes or latent representation) or a 2D feature map (latent representation). We demonstrate the relevance of this approach in the context of myocardial lesions observed in cardiac late Gadolinium enhancement MR images, controlling the image synthesis with segmentation masks or infarct-related attributes, among which size and transmurality. We evaluate it on two datasets of 3557 and 932 pairs of 2D images and segmentation masks, the second dataset being for testing only. Our conditioning not only leads to very realistic synthetic images but also brings varying levels of task complexity, a must-have to better assess the readiness of segmentation methods.

Keywords: Latent Diffusion · Segmentation · Evaluation · Cardiac MRI

1 Introduction

Public datasets from specialised challenges offer valuable insights into the performance of state-of-the-art machine learning methods. However, they often suffer from design issues involving the data and the evaluation metrics [13], lowering the enthusiasm about the reported results. In cardiac MRI segmentation, these drawbacks are pronounced when quantifying complex structures, especially when anatomical contours are of moderate quality. This is particularly critical in myocardial infarct segmentation, which is typically performed on late gadolinium

* Corresponding author: romain.deleat@creatis.insa-lyon.fr

enhancement (LGE) MRI [2]. In LGE images, myocardial infarcts appear as hyperintense (white) regions within the hypointense (black) healthy myocardium. Lesions can substantially vary from small areas that are difficult to detect to larger ones that may be misidentified as the ventricular cavity. Additionally, image contrast depends on the acquisition parameters, leading to cases with highly variable and complex appearances. Given these limitations, the question arises of how segmentation quality can be evaluated effectively beyond traditional metrics such as the Dice coefficient and the Hausdorff distance (HD).

Among the plethora of deep learning segmentation methods, nnU-Net [8] emerged with excellent performance across a wide range of segmentation tasks [1]. It also stood out in recent challenges specific to myocardial infarct segmentation, such as MYOSAIQ[4] and EMIDEC [12]. However, although these reports reveal limitations of the segmentation both on the myocardium and the lesions inside (which can be subtle and of potentially complex shape), anatomical accuracy was not explicitly considered in the evaluation. A recent survey that evaluated nnU-Net on a broad range of cardiac MR images from public challenges [5] also underscored the need for better datasets (larger and more varied, mainly) to assess the actual relevance of this method for clinical practice.

On many applications, among which cardiac MRI, diffusion models [6] are a very promising alternative to GANs to synthesize highly realistic medical images [14]. They push forward the generation of synthetic datasets while overcoming two major challenges: real data sharing restrictions, and data scarcity for training or evaluation. The state-of-the-art in cardiac MR image generation is *Stable Diffusion* [20], which accelerates the diffusion process by taking advantage of a latent representation. Moreover, Latent Diffusion Models (LDM) can be conditioned on various inputs, such as textual descriptions or spatial layouts, enabling precise control over the generated images. This conditional capability is very advantageous for medical imaging, where the synthesis of anatomical structures or pathological features can be explicitly guided. However, conditioning over geometrical attributes [3], a posteriori [18] or via textual input [21] can be difficult and current solutions have limited clinical usefulness.

In this work, we propose a complete and effective method based on conditioned latent diffusion to generate synthetic images based on clinically relevant attributes, which goes beyond these limitations. It allows the controlled generation of specific populations, an asset we specifically exploit to revisit the evaluation of segmentation methods. A total of 3557 pairs of 2D images and segmentation masks from the M1-M12 subset of the MYOSAIQ challenge were used (3163 pairs for training), while 932 pairs from the D8 subset were reserved for testing. With this, segmentation remains a preliminary step for extracting relevant biomarkers, and its evaluation is effectively reoriented towards the relevance of the extracted biomarkers, beyond the use of the Dice coefficient and Hausdorff Distance. Our main contributions are three-fold:

[4] https://www.creatis.insa-lyon.fr/Challenge/myosaiq/platform.html

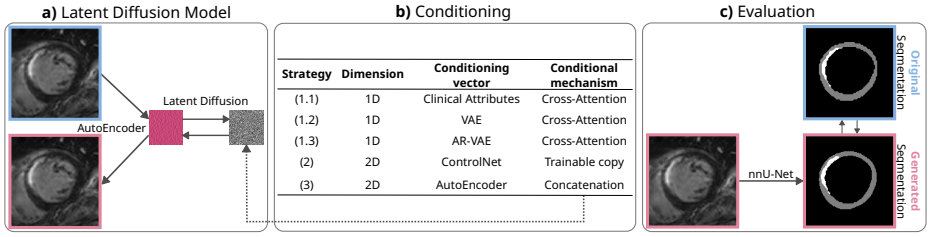

Fig. 1. Overview of the pipeline we propose to evaluate segmentation methods (c). Realistic images are generated by latent diffusion (a), conditioned by different strategies that consider segmentation masks or clinical attributes (b).

- We design a complete pipeline (see overview in Fig.1) based on conditioned latent diffusion models to generate task-oriented datasets for the evaluation of segmentation methods, with a focus on clinical attributes,
- We specifically compare conditioning strategies to reach physiologically relevant image generation,
- We thoroughly evaluate this methodology both qualitatively and quantitatively on a large variety of image configurations.

2 Methods

2.1 Background knowledge

Realistic synthetic images are generated by latent diffusion (Fig.1a). Diffusion models are a class of generative models that synthesize data by gradually transforming a simple noise distribution into structured samples through a learned denoising process (DDPM [6]). They consist of two main stages: forward diffusion and reverse denoising. The forward process progressively corrupts a real data point $\mathbf{x}_0$ by adding Gaussian noise, $\mathbf{x}_t$ being the noisy image at timestep t. The reverse process, performed by a neural network of parameters θ, aims to iteratively remove noise from the final corrupted state $\mathbf{x}_T$ and reconstruct the original data distribution characterized by its mean and covariance.

Here, we specifically rely on latent diffusion [20], which extends standard diffusion, operating in a learned latent space as opposed to the raw image space. The diffusion process is applied in this latent space, producing $\mathbf{z}_T$ given an encoder E that maps images to a compressed latent representation $\mathbf{z} = E(\mathbf{x})$. The reverse process then reconstructs $\mathbf{z}$, which is subsequently decoded into an image by a decoder D. This approach reduces memory requirements and enhances training efficiency, making it particularly effective for conditioning.

2.2 Conditioning Latent Diffusion Models

The generation of physiologically relevant synthetic images is achieved by properly conditioning the latent diffusion. We rely on the *Stable Diffusion* frame-

work [20], which minimizes $\mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{z},t,\mathbf{c})\|^2\right]$, where $\mathbf{c}$ is the conditioning vector, $\boldsymbol{\epsilon}$ is the real noise sampled from a standard normal distribution, and $\boldsymbol{\epsilon}_\theta$ is the noise predicted by the neural network, parameterized by θ.

In our study, we compare three conditioning strategies (see Fig.1b), each being tailored to our application: (1) incorporating a 1D vector into the cross-attention mechanism, (2) with a 2D representation from the ControlNet method, or (3) by concatenating a 2D feature map with the latent noise $\mathbf{z}_T$.

In strategy 1, the model is either conditioned with three clinically relevant scalar attributes and one geometrical scalar (strategy 1.1), or a latent vector from a representation learning method trained on segmentation maps (strategies 1.2 and 1.3). The clinical attributes (see details in Sec.2.3) consist of transmurality, infarct size, and endocardial surface length, and the geometrical scalar stands for the slice position along the sagittal axis. The latent vector is obtained from a Variational Autoencoder (VAE / strategy 1.2) [11] or an Attribute-based Regularized VAE (AR-VAE / strategy 1.3) [16]. The AR-VAE introduces explicit regularization in its latent space, ensuring that specific latent dimensions are correlated with the selected attributes. The cross-attention mechanism from *Stable Diffusion* is used to condition the LDM by modulating the key and value matrices of the query-key-value attention with the chosen 1D vector.

In strategy 2, we employ the ControlNet method [23], a recent plug-and-play approach that enables fine-tuning of the LDM with multi-conditional inputs. Here, only 2D feature maps derived from the segmentation maps are used to condition the model.

Finally, strategy 3 consists of incorporating a 2D latent representation of segmentation maps (extracted from the AutoEncoder of *Stable Diffusion*) and concatenating it with the latent noise representation $\mathbf{z}_T$ [4].

2.3 Evaluation of segmentation methods

We use the following three infarct characteristics and one geometrical image attribute to condition the diffusion of strategy 1.1:

- *Infarct size*: The proportion of pixels classified as infarct.
- *Transmurality*: The extent of infarct from endocardium to epicardium, estimated as the mean of the segmentation map over sections of 10°, and averaged across the sections where infarct is present.
- *Endocardial surface length (ESL)*: The proportion of pixels near the endocardium (radial coordinates lower than 25%) where infarct is present.
- *Sagittal axis*: Slice position along the sagittal axis, ranging from the apex to the base of the heart (excluding slices with myocardial openings).

In addition, a larger variety of generated images is obtained by artificially rotating the segmentation map used in strategies 2 and 3.

Given the relevance of the generated images with respect to the conditioning on these attributes or the segmentation masks (see details in Sec.3.3), we propose

to revisit the evaluation of segmentation methods, by directly comparing the infarct attributes to the ones extracted from a segmentation of the synthetic images by a given segmentation model (Fig.1c). Agreement with the original attributes is quantified by the linear regression coefficient r^2. We also report standard metrics (Dice coefficient and HD), evaluated on both subsets (M1-M12 and D8) for strategies 2 and 3, which are based on segmentation masks.

Finally, we report the performance of two clinical experts at labelling real and synthetic images among randomly selected samples, to better evaluate the realism of the generated images.

3 Experiments and Results

3.1 Data and preprocessing

We used the data from the public challenge MYOSAIQ[4], which was designed to evaluate automatic segmentation methods for quantifying myocardial infarction lesions across different phases of disease progression. Data consisted of pairs of LGE MRI images and their respective segmentation masks from the same cohort at 1 and 12 months after the infarct (M1 and M12 subsets), and another cohort at 8 days after the infarct (D8 subset). We divided the data as:

- *Training set:* 85% ($N = 172$) of patients from the M1 and M12 subsets, resulting in 3163 slices.
- *Validation set:* 5% ($N = 10$) of patients from the M1 and M12 subsets, resulting in 139 slices.
- *Testing set:* 10% ($N = 21$) of patients from the M1 and M12 subsets + $N = 121$ patients from the D8 subset (exluding slices with microvascular obstruction), resulting in 255 + 932 slices.

3.2 Implementation details

For the *Stable Diffusion* and conditioning model, we used the implementation provided by MONAI [17]. For the segmentation method, we used the nnU-Net framework to automatically configure a 2D U-Net architecture based on the input data characteristics. The model was trained with a batch size of 50, optimizing a combined cross-entropy and Dice loss function using the SGD optimizer, a PolynomialLR scheduler, and a learning rate of 0.01. The architectures of both autoencoders (*Stable Diffusion* and the conditioning model) were set with the *KL-reg* variant and a MSE loss, similar to the original paper, except using a single-channel latent representation and processing input images of 128×128 pixels that were downsampled to 32×32 pixels. For the LDM, the architecture consisted of the following components: 1 residual block, 2 downsampling blocks with 32 and 64 channels, and an attention block with 64 head channels in the final layer. The model was trained with the following hyperparameters: a batch size of 32, a learning rate of $2e-5$ using the Adam optimizer, 1000 timesteps for the denoising process, a linear beta schedule, a scale factor of 1, and a total of $5k$ epochs ($500k$ steps). All models were trained on a NVIDIA RTX A600 48GB.

Table 1. Comparison of different conditioning strategies for the LDM: with 1D vectors (the clinical attributes or the latent vectors from a VAE or AR-VAE) and with 2D feature maps (from ControlNet or the AutoEncoder from *Stable Diffusion*). Baseline stands for comparison with the original segmentations. Bold stands for the best results.

	Method	**Data**	**Linear regression (r^2) ↑**			**FID ↓**	
			Transmu	ESL[5]	Infarct size	Real/Real	Real/Synth
	Baseline train	Real	0.98	1.00	1.00	-	-
	Baseline val	Real	0.87	0.92	0.93	-	-
	Baseline test	Real	0.72	0.84	0.80	-	-
(1.1)	Attributes	Synth	0.38	0.61	0.59	29.8	83.4
(1.2)	VAE	Synth	0.26	0.38	0.46	29.8	81.7
(1.3)	AR-VAE	Synth	0.35	0.44	0.49	29.8	91.8
(2)	ControlNet	Synth	0.64	0.78	0.80	29.8	**79.1**
(3)	AutoEncoder	Synth	**0.77**	**0.90**	**0.90**	29.8	86.7

3.3 Evaluation of segmentation methods

Evaluation on synthetic images We first applied the nnU-Net to the synthetic images generated by our method, and compared the clinical attributes extracted from the nnU-Net segmentation to those used for conditioning. Table 1 summarizes the linear regression coefficients obtained on the three infarct attributes for the different conditioning strategies. It also reports the Fréchet Inception Distance (FID), which is in line with the literature [15,21]. Our approach provides a more expressive representation of the segmentation data, while being independent of textual input conditioning. This property explains its capacity to better follow the conditioning attributes.

Since the synthetic images generated using ControlNet and our method effectively adhere to the segmentation conditioning, we can compare both the original and generated segmentations. This enables the evaluation of segmentation methods using conventional metrics, as shown in Tab.2.

Figure 2 illustrates this with a synthetic image (generated with strategy 3, which corresponds to the best results in Tab.2) along with its corresponding segmentations. It highlights the ability of our approach to maintain conditioning fidelity and its potential for evaluating the quality of segmentation methods.

Comparison with the original segmentations (baseline) For baseline comparisons, we applied the nnU-Net to the images from the training, validation, and testing sets (M1-M12 subset), and compared the infarct attributes extracted from the segmentation output to those obtained from the ground truth segmentations (Tab.1, first three rows). On the testing set, the linear regression coefficients were 0.72, 0.84, 0.80 for transmurality, endocardial surface length, and infarct size, respectively. The Dice coefficient was 0.84 ± 0.03 and 0.65 ± 0.18 for the myocardium and infarct respectively, aligning with the performance reported by participants of the challenge [19,22]. These results indicate that, while

[5] ESL: Endocardial Surface Length, see definition in Sec.3.1.

Table 2. Evaluation of ControlNet and AutoEncoder conditioning on synthetic images using conventional metrics (for strategies 2 and 3, which involve a segmentation mask). Bold stands for the best results on the M1-M12 dataset, for which baseline is available.

	Method	**Subsets**	**Dice** ↑		**HD** ↓	
			Myocardium	Infarct size	Myocardium	Infarct size
	Baseline test	M1-M12	0.84 ± 0.03	0.65 ± 0.18	5.1 ± 1.9	19.6 ± 14.1
(2)	ControlNet	D8	0.72 ± 0.10	0.52 ± 0.24	12.4 ± 8.5	16.5 ± 13.9
(2)	ControlNet	M1-M12	0.76 ± 0.11	0.56 ± 0.26	7.8 ± 5.9	14.2 ± 14.8
(3)	AutoEncoder	D8	0.78 ± 0.08	0.63 ± 0.20	11.0 ± 8.6	13.2 ± 13.5
(3)	AutoEncoder	M1-M12	$\mathbf{0.80 \pm 0.08}$	$\mathbf{0.61 \pm 0.21}$	$\mathbf{7.1 \pm 5.2}$	$\mathbf{13.2 \pm 13.7}$

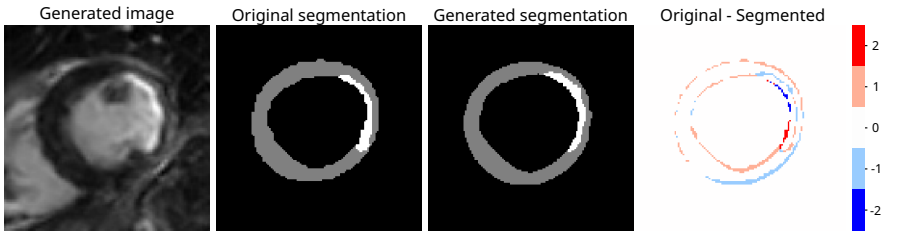

Fig. 2. Difference between the original segmentation used for the conditioning with strategy 3 and the generated segmentation from the nnU-net model.

there is still room for improvement in terms of overall segmentation accuracy, nnU-Net demonstrates strong performance in preserving clinical attributes.

Qualitative assessment of synthetic images realism Figure 3 shows representative images generated with strategy 3, from segmentation masks covering a broad range of clinical attributes. Such masks were identified beforehand as the ones for which a given characteristic is the closest to a desired value (e.g. transmurality $\approx$ 50%). The generated images are anatomically coherent and accurately reflect the clinical attributes, underscoring the effectiveness of our conditioning. In addition, the incorporation of segmentation maps from other datasets (e.g. the D8 subset) or the application of transformations such as rotation can further enhance the diversity of the generated infarct patterns.

We systematically checked that there was no overfitting to real samples across all conditioning methods, by identifying the closest training image for each generated image and qualitatively assessing their distinctiveness.

Finally, two clinical MRI experts labeled synthetic images (generated with strategy 3) on a dataset of 100 shuffled real and synthetic images, and another dataset of 100 pairs of real and synthetic images generated from the same segmentation mask. Their respective accuracies were 68% and 88% (expert 1) and 45% and 67% (expert 2), highlighting the realism of the synthetic images.

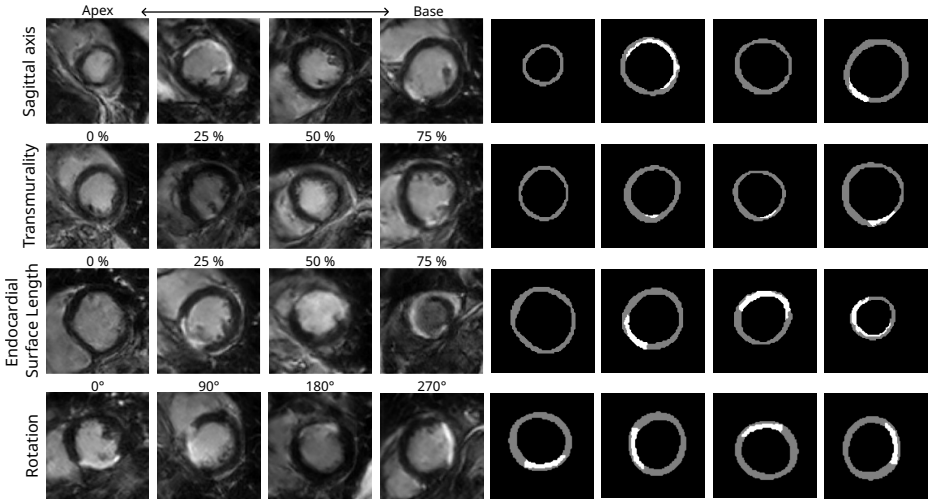

Fig. 3. Representative samples generated with strategy 3, based on segmentation maps covering a geometric and two clinical attributes, and an artificial transformation.

4 Discussion

We proposed a pipeline to generate very realistic synthetic images conditioned on clinically relevant attributes, which we exploit to evaluate segmentation models in a novel manner. Extensive experiments showed that our method produces anatomically coherent synthetic images that accurately reflect clinical attributes, enabling the assessment of segmentation methods across varying levels of task complexity. These contributions have strong potential to mitigate challenges related to real data scarcity and sharing restrictions in medical imaging.

While our method effectively conditions the generated images on clinically relevant attributes, a notable limitation is the lack of texture variability. This can be attributed to the strong influence of conditioning, which constrains the flexibility of the generated images. Specific techniques have been proposed to mitigate this [7], which we will integrate in future work. Furthermore, the limited variability in segmentation maps remains a challenge. This could be addressed by generating synthetic segmentations, either through a user interface or a neural network, guided by clinical attributes. Such an approach, which is beyond the scope of this work, would provide a more diverse generation of cardiac LGE images, and better robustness of the synthetic dataset.

Another limitation of our approach is the size of the challenge dataset, as diffusion models have been shown to exhibit data memorization when trained on limited data [10]. We have ensured that our model does not memorize the training data by computing the MSE between each synthetic image and all training slices, identifying the closest one, and comparing them qualitatively. However, a more diverse population with varying textures would enhance generalization and improve the diversity of the generated images. To address this, future work

will explore the integration of additional datasets or MRI sequences to improve the robustness and generalizability of the model.

Finally, we have demonstrated that clinical attributes can serve as a valuable metric for assessing the performance of a segmentation method. Further analysis of the correlations between Dice scores from various segmentation methods and their associated clinical metrics could strengthen the validation of these models. It would also be relevant to investigate uncertainty methods [9] providing insights into the calibration of the segmentation method while offering reliability measures for both the segmentation maps and the computed clinical attributes.

Code and data availability The data we used come from a public challenge (MYOSAIQ[4]), with restricted access provided by the organizers. The code for our proposed pipeline is publicly available[6].

Acknowledgments. The authors acknowledge the partial support from the LABEX PRIMES of Université de Lyon (ANR-11-LABX-0063), the French ANR (MIC-MAC project, ANR-19-CE45-0005), the Fédération Française de Cardiologie (MI-MIX project, Allocation René Foudon), the Université Lyon 1 (AAP SENS 2022), and the Institut Universitaire de France. The authors are grateful to the H2P platform[7] (CREATIS Lyon, France) for the data storage and management support.

Disclosure of Interests. MV and PC have research agreements with Siemens, Olea Medical, and Circle Cardiovascular Imaging. They did not influence the contents of this work, which was not sponsored. The other authors have no relationships to disclose.

References

1. Antonelli, M., Reinke, A., Bakas, S., et al.: The medical segmentation decathlon. Nat Commun **13**, 4128 (2022)
2. Bulluck, H., Dharmakumar, R., Arai, A., et al.: Cardiovascular magnetic resonance in acute st-segment-elevation myocardial infarction: Recent advances, controversies, and future directions. Circulation **137**, 1949–64 (2018)
3. Daum, D and Osuala, R and Riess, A and others: On differentially private 3D medical image synthesis with controllable latent diffusion models. Proc. MICCAI-DGM4MICCAI workshop, LNCS **15224**, 139–49 (2024)
4. Dorjsembe, Z and Pao, HK and Odonchimed, S and others: Conditional diffusion models for semantic 3D brain MRI synthesis. IEEE J Biomed Health **28**, 4084–93 (2024)
5. Gunawardhana, M and Xu, F and Zhao, J: How good nnU-Net for segmenting cardiac MRI: a comprehensive evaluation. arXiv preprint (2024)
6. Ho, J and Jain, A and Abbeel, P: Denoising diffusion probabilistic models. Proc. NeurIPS **33**, 6840–6851 (2020)
7. Ho, J. and Salimans, T.: Classifier-free diffusion guidance. Proc. NeurIPS Workshop on Deep Generative Models and Downstream Applications (2021)

[6] https://github.com/creatis-myriad/cLDM_project
[7] https://humanheart-project.creatis.insa-lyon.fr/

8. Isensee, F and, Jaeger, PF and Kohl, SAA : nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods **18**, 203–11 (2020)
9. Judge, T., Bernard, O., Cho Kim, W., et al.: Uncertainty propagation for echocardiography clinical metric estimation via contour sampling. arXiv preprint (2025)
10. Kadkhodaie, Z and Guth, F and Simoncelli, EP and others: Generalization in diffusion models arises from geometry-adaptive harmonic representation. arXiv preprint (2023)
11. Kingma, DP and Welling, M: Auto-encoding variational Bayes. Proc. ICLR (2014)
12. Lalande, A., Chen, Z., Pommier, T., et al.: Deep learning methods for automatic evaluation of delayed enhancement-MRI. The results of the EMIDEC challenge. Med Image Anal **79**, 102428 (2022)
13. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: recommendations for image analysis validation. Nat Methods **21**, 195–212 (2024)
14. Müller-Franzes, G., Niehues, J., Khader, F., et al.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Sci Rep **26**, 12098 (2023)
15. Osuala, R and Skorupko, G and Lazrak, N and others: medigan: a Python library of pretrained generative models for medical image synthesis. J Med Imaging **10**, 061403 (2023)
16. Pati, A., Alexander, L.: Attribute-based regularization of latent spaces for variational auto-encoders. Neural Comput Appl **33**, 4429–44 (2021)
17. Pinaya, WH and Graham, MS and Kerfoot, E and others: Generative AI for medical imaging: Extending the MONAI framework. arXiv preprint (2023)
18. Pinaya, WH and Tudosiu, PD and Dafflon, J and others: Brain imaging generation with latent diffusion models. Proc. MICCAI-DGM4MICCAI workshop, LNCS **13609**, 117–26 (2022)
19. Qayyum, A and Razzak, I and Mazher, M and others: Unsupervised unpaired multiple fusion adaptation aided with self-attention generative adversarial network for scar tissues segmentation framework. Inf Fusion **106**, 102226 (2024)
20. Rombach, R and Blattmann, A and Lorenz, D and others: High-resolution image synthesis with latent diffusion models. Proc. CVPR pp. 10684–95 (2022)
21. Skorupko, G and Osuala, R and Szafranowska, Z and others: Debiasing cardiac imaging with controlled latent diffusion models. arXiv preprint (2024)
22. Thaler, F and Gsell, MA and Plank, G and others: CaRe-CNN: cascading refinement CNN for myocardial infarct segmentation with microvascular obstructions. arXiv preprint (2023)
23. Zhang, L and Rao, A and Agrawala, M: Adding conditional control to text-to-image diffusion models. Proc. ICCV pp. 3836–47 (2023)