Tackling data scarcity: Synthetic tumour and mask generation to improve image segmentation

Félix Quinton, Benoit Presles, Romain Popoff, François Godard, Olivier Chevallier, Julie Pellegrinelli, Jean-Marc Vrigneaud, Jean-Louis Alberini, Fabrice Meriaudeau

Artificial Intelligence In Medicine 173 (2026) 103348

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Tackling data scarcity: Synthetic tumour and mask generation to improve image segmentation

Félix Quinton [a], Benoit Presles [a], Romain Popoff [b], François Godard [b], Olivier Chevallier [d], Julie Pellegrinelli [c], Jean-Marc Vrigneaud [b], Jean-Louis Alberini [b], Fabrice Meriaudeau [a,*]

[a] Université Bourgogne Europe, CNRS, ICMUB UMR 6302, Dijon, 21000, France
[b] Université Bourgogne Europe, Centre Georges-François Leclerc, Unicancer, service de médecine nucléaire, Institut de Chimie Moléculaire, UMR CNRS 6302, Dijon, 21000, France
[c] Service de radiologie diagnostique et interventionnelle, Centre Georges-François Leclerc, Dijon, 21000, France
[d] Service de radiologie et imagerie médicale diagnostique et therapeutique, Centre Hospitalier Universitaire, Dijon, 21000, France

ARTICLE INFO

Keywords:
Deep learning
Segmentation
Image generation
MRI
Tumour synthesis

ABSTRACT

Given the increasing data requirements of deep learning models and the scarcity of medical imaging data, new data augmentation techniques are receiving particular attention. This paper explores the subfield of tumour synthesis within medical image generation, focusing on the development of synthetic tumours in MR images. This study introduces a novel tumour generation method using diffusion models, designed to inpaint visually convincing 3D synthetic liver tumours into real MRI volumes while generating the corresponding masks using simplex deformation. This approach has been employed successfully to inpaint images with 1 000 synthetic tumours. Furthermore, it has shown significant performance improvements when applied in image segmentation tasks. In particular, our method improved the Dice coefficient by 6.7 points on the ATLAS test set without relying on external data. When combined with a pseudo-annotated external dataset, the improvement increased to 10 points. This study not only demonstrates the ability to segment tumours but also paves the way for various synthetic data-based applications in medical imaging.

1. Introduction

1.1. Deep learning in medical imaging

In the past decade, numerous automatic methods particularly those based on deep learning have been developed to perform tasks such as classification, segmentation, and registration. Deep learning frameworks have significantly impacted computer vision. Convolutional neural networks (CNNs) [1–4] and, more recently, transformer models [5,6], originally introduced for natural language processing, have marked substantial progress. Transformers, in particular, have shown strong performance on large datasets, outperforming state-of-the-art techniques.

These advances have naturally extended to medical imaging, particularly in image segmentation [7–10]. Tumour segmentation has emerged as a major focus area, where deep learning methods have demonstrated significant potential [11,12]. However, the performance of these models is often limited by the scarcity of large annotated datasets, which remains a critical challenge in medical imaging.

1.2. Strategies for dealing with limited data

To address data scarcity, several strategies are being explored, including data augmentation [13], image generation techniques [14], unsupervised learning [15], semi-supervised learning [16], and self-supervised learning [17] (SSL). In recent years, SSL has emerged as one of the most popular and effective paradigms for mitigating the lack of annotations, particularly through contrastive representation learning. Several notable studies have demonstrated its potential for image-level tasks, including voxel-wise representation distillation [18], anatomical-aware contrastive distillation [19], and anatomy-guided self-supervision [20]. These approaches exploit latent anatomical structure within unlabelled data to enhance segmentation performance under limited annotation budgets.

Beyond contrastive and SSL-based methods, generative modelling has become the second major strategy for overcoming data scarcity. Many studies have demonstrated the ability of generative models to produce visually convincing images. These models fall into several categories. Energy-based models [21,22] define an energy landscape

* Corresponding author.
E-mail address: fabrice.meriaudeau@ube.fr (F. Meriaudeau).

https://doi.org/10.1016/j.artmed.2025.103348
Received 1 August 2025; Received in revised form 21 December 2025; Accepted 29 December 2025
Available online 31 December 2025

over data and generate samples by minimising this energy. Variational autoencoders (VAEs) [23–25] use probabilistic encodings to reconstruct data through latent variables. Flow-based models [26,27] rely on invertible transformations to map data to latent spaces. Generative adversarial networks (GANs) [28–32] employ a generator and a discriminator in a minimax game to produce realistic samples. More recently, denoising diffusion probabilistic models (DDPMs) and their variants [33–36] have expanded the range of tools available for generative modelling. By expanding training datasets with synthetic data, these methods have shown strong potential in segmentation tasks [37–40].

1.3. Tumour synthesis for data augmentation

Tumour synthesis is a generative technique designed to inpaint synthetic tumours into medical images. This strategy is especially valuable, as tumours are less frequent and harder to identify and annotate than organs. Additionally, deep learning models often struggle to fully exploit external datasets due to significant domain shifts caused by differences in population, equipment, sequences, protocols, and operators. In this context, tumour synthesis offers a promising way to enrich datasets by inserting synthetic tumours into both healthy and pathological organs.

Several approaches such as SelfMix [41], LesionMix [42], and CarveMix [43] propose strategies to manipulate tumours within images, either by modifying their appearance and shape or by transferring them from one image to another.

In contrast, other methods aim to generate synthetic tumours from scratch, rather than manipulating existing ones within real images.

In 2D imaging, a number of studies have investigated the use of GANs for tumour synthesis. Shin et al. [44] developed a method to synthesise colonic polyps in colonoscopy videos using conditional GANs to improve detection. Horvath et al. [45] employed conditional GANs to generate tumour cells in fluorescence microscopy images, transforming images of normal cells into ones containing synthetic tumour cells.

In 3D CT imaging, several studies have focused on generating synthetic tumours without relying on deep learning models. For the pancreas, Li et al. [46] proposed an approach using deformable ellipsoidal shapes combined with expert-defined parametrised textures to aid early tumour detection. Similarly, Wei et al. [47] applied local texture synthesis to enhance pancreatic images and improve segmentation accuracy.

For the liver, most studies have focused on the generation of liver tumours in CT images using the LiTS dataset [48]. Hu et al. [49] used manual heuristics to create realistic tumours and corresponding masks from deformed ellipsoids. Based on this work, Hu et al. [50] demonstrated that training segmentation models in small sets of synthetic tumours and their masks improves the delineation of liver tumours while reducing the dependency on manual annotations.

Zhang et al. [51] proposed an unsupervised liver tumour segmentation approach based on pseudo anomaly synthesis. Their method integrates a random-shaped anomaly generation module and a two-stage training strategy within the DRAEM architecture [52] to detect liver lesions as anomalies, achieving competitive performance on the LiTS dataset.

Using GANs, Wu et al. [53] proposed FreeTumor, a method for tumour synthesis and segmentation. Their pipeline involves training a segmentation model as discriminator, generating tumours with a GAN using both labelled and unlabelled data, and filtering low-quality samples via a segmentation-based discriminator.

More recently, diffusion models and their variants have been explored for tumour synthesis. Chen et al. [54] introduced DiffTumor, a latent diffusion model (LDM) designed to enhance small tumour segmentation by leveraging a dataset combining liver, pancreas, and kidney tumours, which often exhibit similar appearances on CT imaging. Peng et al. [55] extended this work by analysing the influence of tumour size and boundary clarity on segmentation. Chan et al. [56] modified DiffTumor to generate tumours in PET/CT image pairs. Li et al. [57] proposed TextoMorph, a 3D diffusion model conditioned on textual descriptions from radiology reports, improving the realism and diversity of synthetic tumours through attributes such as texture, margins, and pathological type.

Finally, some teams have focused on alternative tasks using synthetic tumours. Fei et al. [58] explored the domain shift between real and synthetic tumours in liver segmentation and its impact on model generalisation. Hu et al. [59] examined how incorporating synthetic tumours into the validation set affects segmentation performance.

1.4. Challenges in MRI tumour synthesis

Synthesising realistic liver tumours on magnetic resonance imaging (MRI) is inherently more challenging than on CT. On CT, tumours appear relatively homogeneous and often present as areas of low contrast, similar to noise, and—despite the normalisation steps commonly applied in deep-learning pipelines—the underlying Hounsfield Unit scale provides a physically meaningful and standardised reference across scanners. In contrast, MRI does not rely on a unified intensity scale, and its signal intensities lack a direct physical interpretation. As a result, tumour appearance varies markedly depending on acquisition protocols, sequence type, scanner manufacturer and patient-specific parameters, leading to substantial variability in contrast and texture. This pronounced heterogeneity makes realistic tumour synthesis considerably more difficult in MRI and partly explains why only a limited number of studies have addressed this problem.

Using the BraTS dataset [60], Wolleb et al. [61] developed a 2D diffusion-based method to simulate brain tumour growth in MR images. Similarly, Rouzrokh et al. [62] proposed a multitask diffusion model to generate 2D brain tumours. Kebaili et al. [63] introduced a 2D latent diffusion model for generating 3D brain MRIs with segmentation masks, using a 2D autoencoder and positional encoders to process volumes slice by slice for memory efficiency.

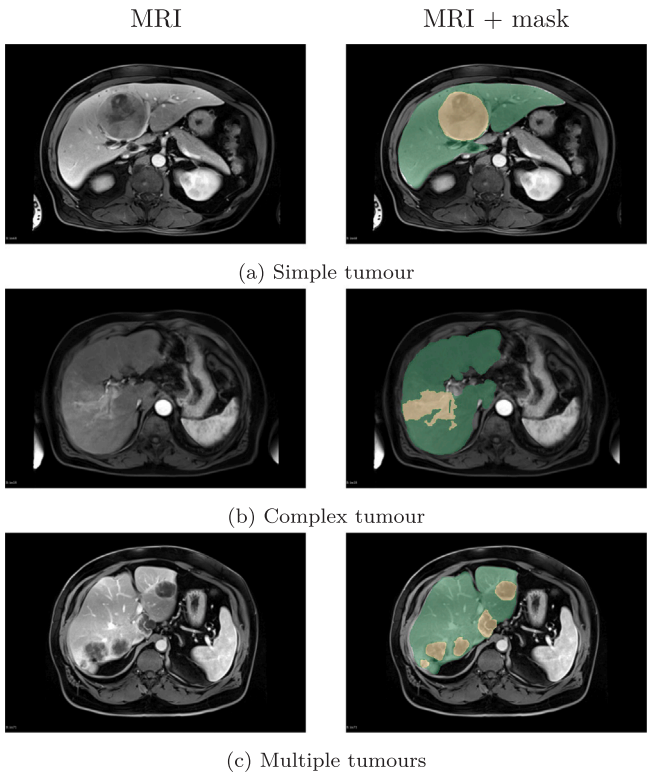
More recently, Durrer et al. [64] presented a pseudo-3D denoising diffusion model for MRI inpainting. Their method processes stacks of adjacent slices using a combination of 2D and 1D convolutions to produce consistent 3D outputs while remaining computationally efficient. Trained on BraTS and evaluated on downstream tasks, their model outperformed both standard 2D and full 3D alternatives, offering a strong trade-off between realism and computational cost.

Due to the substantial biological and morphological heterogeneity of hepatocellular carcinoma (HCC), reliable tumour characterisation on MRI remains particularly challenging. HCC comprises a broad spectrum of sub-types (macrotrabecular-massive, steatohepatitic, clear cell, scirrhous, and others), each exhibiting distinct growth patterns, vascularisation profiles, cellular compositions and radiological appearances [65,66]. This diversity, combined with the limited availability of large, annotated MRI datasets, has so far prevented studies from addressing the generation of HCC tumours in this modality. The work presented in this article therefore focuses on the generation of liver tumours and their corresponding masks in 3D MRI volumes, with the aim of improving tumour segmentation in this challenging context.

Our approach consists of two main steps: generating a binary tumour mask from deformed ellipsoids, and embedding synthetic tumours into MRI using a DDPM. The complete implementation is publicly available at https://gitlab.in2p3.fr/iftim/public-projects/tackling-data-scarcity.

The key contributions of the present study are outlined below:

- we introduce mask-generator method that produces realistic synthetic tumour masks by deforming ellipsoids using elastic or simplex deformation, allowing for considerable diversity in shape and size;
- we propose a diffusion-based inpainting model capable of synthesising high-quality tumours within real MRI scans;



(a) Simple tumour

(b) Complex tumour

(c) Multiple tumours

Fig. 1. Axial CE-MRI slices from three different patients of the ATLAS dataset with corresponding liver and tumours masks. Liver appears in green and tumour in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- we conducted an extensive analysis of the impact of our method on image segmentation. The improvement increases further when external data with pseudo-masks is included.

2. Material and methods

2.1. Dataset

In this work, two datasets were used for data generation and segmentation: ATLAS [67] and LLD-MMRI [68].

The ATLAS dataset consists of 90 T1-weighted contrast-enhanced (CE) MRIs obtained from 90 different patients diagnosed with HCC. A single image is available for each patient, corresponding to either the arterial, portal, or delayed phase after injection of the gadolinium-based contrast agent. The ATLAS dataset is composed of thorax and abdominal scans. In addition to the CE-MRI images, masks of the liver and tumours are provided. These masks were manually delineated by an experienced MRI radiologist and used as ground truth. Some CE-MRI images and their corresponding masks are shown in Fig. 1 . The 90 images and associated masks from the ATLAS dataset were split into training, validation, and test sets, consisting of 48, 12, and 30 images, respectively.

The LLD-MMRI dataset consists of MRI from 394 patients with seven different types of lesions. Within this group, 100 patients exhibit HCC. Each patient comes with a series of eight MRI scans representing different phases. It includes T1-weighted CE-MRIs in arterial, portal, and delayed phases similar to the ATLAS dataset. Therefore, 300 images (3 images × 100 patients) from the LLD-MMRI dataset were used. Unlike the ATLAS dataset, which includes pixel-level annotation, the LLD-MRI dataset provides the coordinates of the bounding box for each tumour present in the axial slice (2D bounding box). To counter this, we generated pseudo-masks in a manner similar to the teacher/student approach described in [69]. We first trained a 3D nnU-Net segmentation model [70] using the ATLAS dataset. During training, each ATLAS MRI volume was provided together with the corresponding tumour bounding box as an additional input channel, while the ground-truth tumour segmentation served as the supervisory signal. In this setting, the bounding box does not define the target; rather, it guides the network toward the approximate tumour location, which it must refine into a precise segmentation based on the annotated ATLAS masks. Once trained and validated on ATLAS, the nnU-Net model was applied to the LLD-MMRI dataset in inference: each MRI volume and its associated bounding box were fed into the network to generate refined tumour pseudo-masks. To maintain consistency with ATLAS, a liver mask was simultaneously predicted without relying on bounding boxes. Examples of the original bounding boxes and the resulting pseudo-masks can be seen in Fig. 2. These pseudo-labelled images were used exclusively to train the generative model.

A min–max normalisation was applied to all images. The images were resampled to a uniform resolution of 1.5 mm×1.5 mm×3.0 mm. This allows full tumours to be loaded into our 3D DDPM using a patch size of 128 × 96 × 64 voxels.

2.2. Mask and tumour synthesis

To synthesise realistic tumours and their corresponding masks, we propose a two-step method: first, tumour shapes are generated by applying deformation to an ellipsoid. This shape can then be embedded into a patient's organ mask. In the second step, an embedding model is trained to produce synthetic tumours within a specified volume. Fig. 3, illustrates the complete process that enables the creation of a new tumour along with its corresponding binary mask.

2.2.1. Mask synthesis

To generate binary masks with tumour-like shapes, inspired by Hu et al. [50], we start by generating random ellipsoidal shapes. The ellipsoid semi-axes (a, b, c) are sampled independently from a normal distribution $\mathcal{N}(\mu, \sigma)$, with a mean $\mu = 27$ mm and a standard deviation $\sigma = 4.5$ mm, ensuring that the radii along each axis fall within a realistic range of $[13.5, 40.5]$ mm for a secondary tumour 99.7% of the time.

The ellipsoid is then deformed using elastic deformation, a process based on the application of a displacement field. Various types of displacement fields can be used to produce these deformations, including deterministic fields or random perturbations [4,71]. Here, we consider a random displacement field generated from a normal distribution.

While elastic deformations are commonly employed to perform ellipsoid deformation, we propose to use simplex deformation instead [72]. Simplex noise is an improved variant of Perlin noise [73]. This type of noise is particularly well-suited to generate organic and irregular shapes, making it relevant for modelling the heterogeneous contours of tumours.

The details of tumour mask deformation using elastic or simplex deformation are provided in Appendix A.

2.2.2. Tumour synthesis

In order to perform tumour generation within MRI volumes, we choose to use the standard DDPM introduced by Ho et al. [33]. DDPMs operate in two main stages: a forward process that gradually corrupts the input data with Gaussian noise, and a reverse process where a model is trained to denoise this signal step-by-step and reconstruct realistic samples.

Let x_0 be our input data sampled from a real data distribution $q(x)$, i.e. $x_0 \sim q(x)$. In the forward stage, x_0 is progressively noised over T steps to obtain $x_T \sim \mathcal{N}(0, I)$. To perform inpainting, we apply a binary mask m to ensure that only the masked region is corrupted by noise, while the unmasked parts remain unchanged throughout the process.

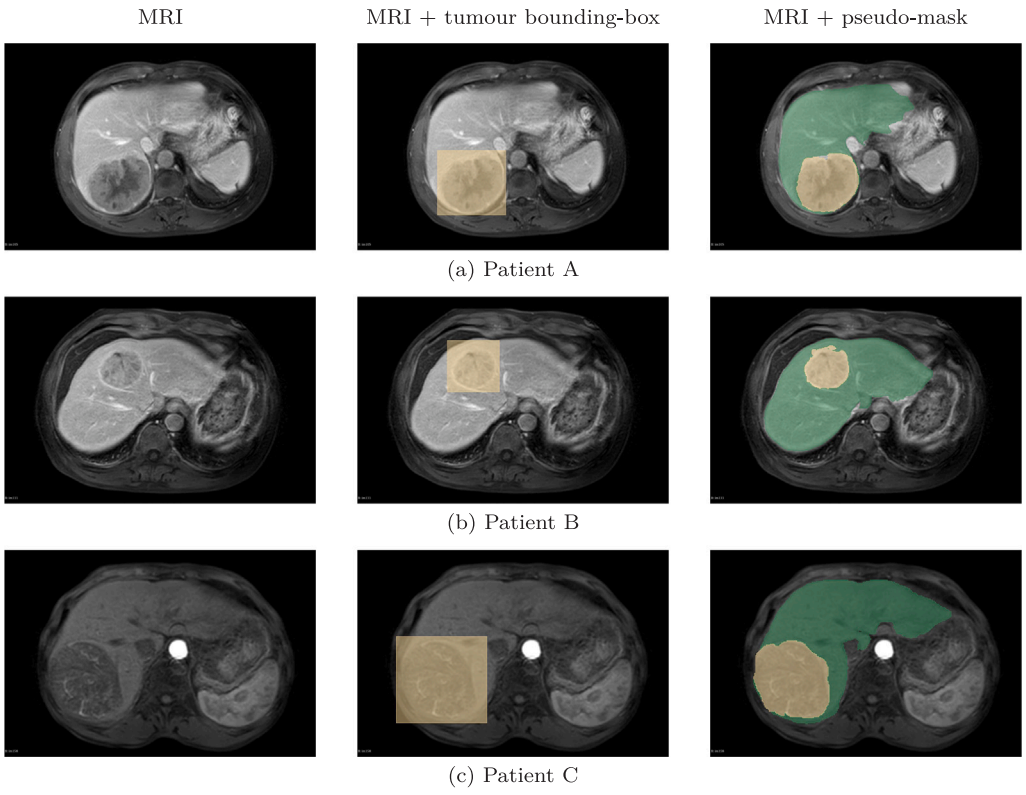


Fig. 2. Axial slices from three patients of the LLD-MMRI dataset, with MRI only, MRI + tumour bounding boxes and MRI + pseudo-masks for liver and tumours. The liver is shown in green and the tumour in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

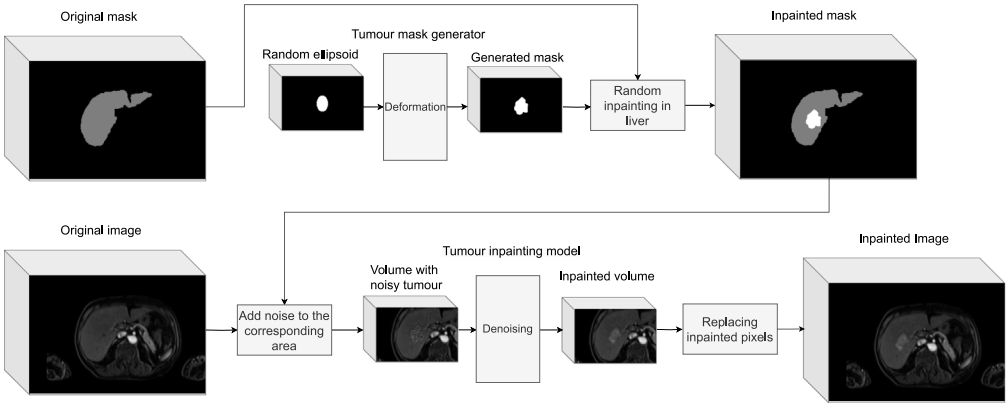


Fig. 3. The complete generation process comprises two steps. Initially, a new mask is synthesised from random ellipsoids deformed using elastic or Simplex deformations and positioned randomly within the healthy liver. Subsequently, noise is introduced to the new tumour location on the corresponding MRI image, and the tumour inpainting model is employed to generate a new tumour at this site. Finally, the inpainted area is integrated into the original image. In this figure, the mask of the liver is in grey and the tumour in white.

At each time step t, the noised image x_t is computed directly from x_0 as:

$$x_t = (1 - m) \odot x_0 + m \odot (\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon), \tag{1}$$

where $\overline{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative product of the noise schedule, $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, and $\odot$ denotes the Hadamard product. This formulation ensures that noise is only applied within the masked region defined by m, preserving the rest of the image.

In the reverse process, the model learns to denoise x_t and approximate the true posterior $q(x_{t-1}|x_t)$. As in the forward process, the reconstruction is conditioned on the mask. Sampling is performed iteratively as:

$$x_{t-1} = (1 - m) \odot x_0 + m \odot (\mu_\theta(x_t, t) + \sqrt{\Sigma_\theta(x_t, t)}\ z), \quad z \sim \mathcal{N}(0, I), \tag{2}$$

The model is trained to predict the noise added during the forward process. The loss is computed only on the masked area:

$$\mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,1), t}\left[\|m \odot (\epsilon - \epsilon_\theta(x_t, t))\|_2^2\right]. \tag{3}$$

The noising and denoising processes are run for 1 000 steps during training. At inference time, we reduce the number of denoising steps to 250 using a DDIM scheduler [74], which allows faster sampling.

As our objective is to generate synthetic tumours, noise is applied exclusively to tumour voxels at each training step, thereby creating a

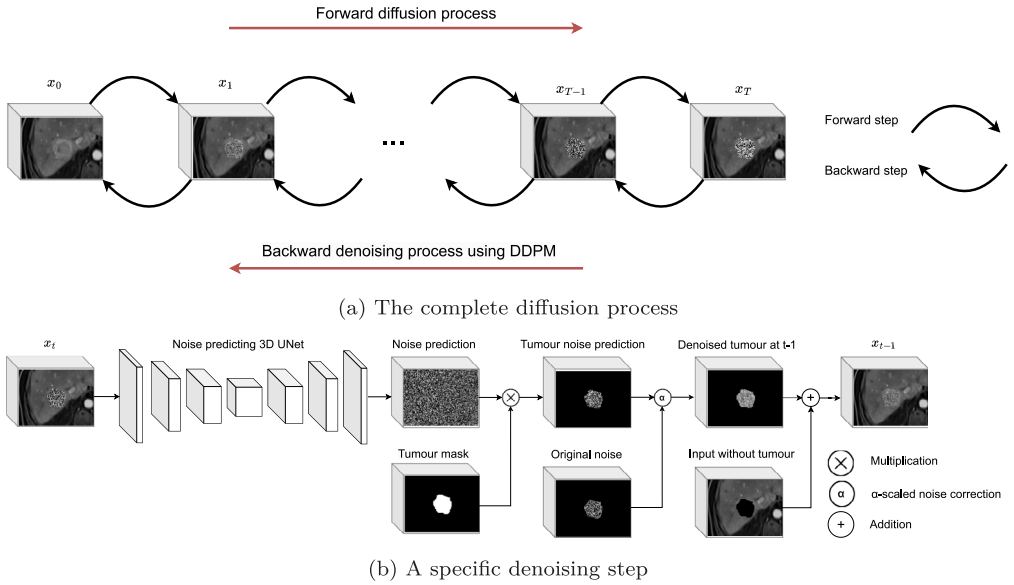


Fig. 4. (a) The complete diffusion process involves cropping 3D patches containing tumoural regions, which are progressively noised during the forward diffusion process from x_0, the input image, to x_T, representing pure noise. This process is reversed during the backward denoising phase using a DDPM. (b) A specific denoising step occurs at timestep t, where the input image x_t is passed through a 3D U-Net that predicts the noise component associated with the tumour region. An α-scaled noise correction is then applied, the predicted noise is appropriately scaled according to the diffusion coefficients before being removed from x_t yielding a partially denoised image. Only the corrected tumour region is retained, while the non-tumour regions are restored to their original appearance.

Table 1
Model configuration: patch size, number of channels, number of attention head channels, number of residual blocks, number of parameters, and additional training settings.

Patch size	N Channels	N attention heads channels	N residual blocks	N Parameters	Beta schedule
$128 \times 96 \times 64$	128, 256, 512, 1024	0, 0, 0, 768	3	8.62×10^8	scaled linear beta

distinction between known and unknown regions. The known region, corresponding to healthy liver tissue or background, provides context for the embedding process. In contrast, the unknown region, composed of a mixture of tumour tissue and noise, is progressively reconstructed by the model. This process is illustrated in Fig. 4.

2.3. Experiments

In this study, two DDPM models were trained: one on the ATLAS dataset, and the ATLAS and LLD-MMRI datasets. The input patch size was set to 128 × 96 × 64 voxels, corresponding to the maximum GPU capacity, with a batch size of 1. Random flipping and rotation data augmentation techniques were applied during training to improve model's generalisation and robustness. The model was trained using a mean square error (MSE) loss function and the AdamW optimiser, with a learning rate of $5{\times}10^{-5}$ and a weight decay of $1{\times}10^{-5}$. Each model was trained for 5 000 epochs, with validation based on MSE, measuring the difference between predicted and actual noise. The model configuration is presented in Table 1.

To compare tumour generation using DDPM with other generative approaches, we also trained an inpainting GAN to synthesise tumours. We adapted the ipA-MedGAN model proposed by Armanious et al. [75] to 3D. This model has been specifically designed for medical image inpainting, enabling the reconstruction of arbitrarily shaped masked regions without the need for prior localisation. It relies on a multi-scale architecture using cascaded MultiRes-UNets as the generator and two discriminators with different receptive fields. One is global and the other local, enforcing both contextual consistency and fine structural details. Since this is an inpainting GAN, the objective was to reconstruct the tumour regions that were masked during training. As with the DDPM-based approach, the ATLAS training set was used as input, and the generator was trained to inpaint tumour regions, while the discriminator learned to distinguish between real and inpainted tumour areas giving the full image as context. The training was guided by a combination of adversarial, L1, and perceptual losses [75].

2.4. Model selection

The selection of the generative models is based on the analysis of the quality of the latent representation of the synthetic tumours they produce. To this end, we used the nnU-Net segmentation network, previously trained on the ATLAS dataset, to extract latent representations from the validation set images. The tumours in these images were then replaced with synthetic tumours generated by the generative models under evaluation. The modified images were processed in the same way to obtain their corresponding latent representations. This procedure was repeated at different validation steps. A one-class support vector machine (SVM), trained on latent representations of images with real tumours, was used to assess how closely the latent spaces of the synthetic tumours matched those of the real ones [76]. The overall score for each generative model was calculated as the average of the scores across all modified images. The results of this selection are available in Appendix B.

2.5. Turing test

To evaluate the quality of the synthetic data, a visual Turing test was conducted on the synthetic tumours. This task was performed by two radiologists and one nuclear medicine physician, each possessing between four and six years of experience. They were instructed to distinguish real tumours from synthetic ones.

A total of 30 images were presented: 15 original MRIs containing real tumours, and 15 MRIs in which the original tumours had been replaced by synthetic tumours using the tumour inpainting model. The

experts were asked to review the MRIs and indicate, for each volume, whether the tumour was real or synthetic.

2.6. Inference process

Once the model has been trained, it can be used to generate a new pair of synthetic tumour and mask. Initially, an ellipsoid is generated and deformed (using elastic or simplex deformation) to create a tumour mask. This mask is then randomly inserted into a healthy region of the liver, based on the original liver mask of a specific patient. Gaussian noise is subsequently introduced into the region where the tumour has been placed on the corresponding MRI. Finally, the tumour embedding model generates a synthetic tumour within this region.

Although such synthetic tumours can be embedded in images of healthy patients, we focused on MR images of patients already presenting tumours, as no public dataset provides recent, high-quality MRI scans of healthy livers. Consequently, we generate secondary tumours.

Using the generative model trained on the ATLAS dataset only, 1 000 synthetic tumours were inserted into healthy liver regions of images from the ATLAS training set. Similarly, with the model trained on both the ATLAS and LLD-MMRI datasets, 1 000 tumours were synthesised, 500 on each of the ATLAS and LLD-MMRI training sets. Since both datasets contain fewer images than the number of synthetic tumours generated, the same image appears multiple times, with a different synthetic tumour embedded each time. This process was performed twice: once using masks generated by elastic deformations, and once using simplex deformations.

2.7. Segmentation model

To evaluate the impact of our tumour generation method on segmentation, a segmentation network was trained to delineate liver and tumour using either real images only or a combination of real and synthetic images. As shown by Quinton et al. [77], the nnFormer model introduced by Zhou et al. [78] achieved very promising segmentation results on the ATLAS dataset. In our experiments, nnFormer was trained on 3D image patches of size $224 \times 224 \times 64$, using the Adam optimiser with a learning rate of 10^{-4} for 3 500 epochs. Standard data augmentation techniques were applied during training, including random rotations, flipping, zooming, intensity scaling, Gaussian noise and blurring, contrast adjustment, and gamma transformations. All segmentation models were trained using real data (from ATLAS and, where applicable, LLD-MMRI). The key difference between configurations lies in the number of images with synthetic tumours included during training, and in the method used to generate them.

2.7.1. Impact of the number of synthetic images

First, we evaluated the effect of increasing the number of synthetic tumours on segmentation. A baseline model was trained using real images only from the ATLAS dataset. Then, several models were trained using the same real images combined with an increasing number of synthetic samples: 50, 100, 200, 500, and 1 000 synthetic tumours. All of the synthetic tumours were generated using our DDPM model trained on the ATLAS dataset only. To ensure consistency, the training set of a model trained with a greater number of synthetic images always included the synthetic data used by the model with fewer samples.

2.7.2. Comparison with GAN-based augmentation

To compare our DDPM-based generation strategy with other generative approaches, we trained a model using 500 synthetic tumours generated via the IpA-MedGAN method. The GAN model was trained under the same conditions as the DDPM and also relies on real tumour data during training, as it performs inpainting over noisy tumour regions.

2.7.3. Impact of standard data augmentation

We also evaluated the standalone impact of tumour generation, independently of standard data augmentation. Two models were trained without any additional augmentation: a baseline using only real ATLAS images, and a model including 500 synthetic tumours. This allows us to assess whether tumour synthesis alone can lead to significant gains in segmentation performance.

2.7.4. Impact of external datasets

To assess the influence of external data, we trained models using a combination of the ATLAS and LLD-MMRI datasets. One configuration used the real data only, while another added 500 synthetic tumours to each dataset (i.e. 1 000 synthetic tumours in total). Since LLD-MMRI contains more images, we applied an oversampling strategy to balance both datasets during training.

2.7.5. Simplex vs elastic mask deformation

Lastly, we evaluated the impact of the mask deformation method employed during the generation of synthetic tumours. To do this, we trained every model with synthetic tumours with masks generated either with elastic deformations or simplex deformations.

2.8. Evaluation metrics

To perform a quantitative analysis of the segmentation quality on the test set, the following metrics were selected: Dice coefficient, 5 mm surface Dice (SD), precision, recall, Hausdorff distance (HD) and 95th Percentile Hausdorff Distance (HD95) [79]. To assess the statistical significance of the results, a one-tailed paired t-test was applied to the Dice scores between the baseline model and the most efficient models trained with the proposed method.

2.9. Implementation details

All generative models were trained on Nvidia A100 GPUs with 80 GB of memory, while all segmentation models were trained on Nvidia Tesla V100 GPUs with 32 GB of memory. The experiments were conducted using the PyTorch library version 1.11.0 and the Nvidia CUDA toolkit version 11.3.1. Data processing was carried out using the medical open network for artificial intelligence (MONAI) framework version 1.2.0 [80].

3. Results

3.1. Image generation

3.1.1. Mask generation

As illustrated in Fig. 5, the type of noise used to perform elastic deformations impacts the final shape. While both elastic and simplex deformations applied to ellipsoids can produce realistic tumour shapes, simplex deformation allows for larger distortions without fragmenting the structure, thus better preserving the cohesion and overall integrity of the ellipsoid.

3.1.2. Tumour synthesis

Examples of synthetic tumours with masks deformed using elastic deformation and simplex deformation are illustrated in Fig. 6.

Fig. 7 illustrates cases of synthetic tumours inpainted using GANs and DDPM. When comparing those methods, a clear difference in the realism of the synthetic tumours can be observed. In GAN-based tumours, the textures lack granularity and intensity variation compared to the surrounding hepatic parenchyma. This results in a smooth, almost plastic-like appearance that looks unrealistic and poorly integrated with the native tissue. In contrast, the DDPM-based tumours demonstrate a much more realistic inpainting. The synthetic textures show finer granularity and smoother intensity transitions, matching

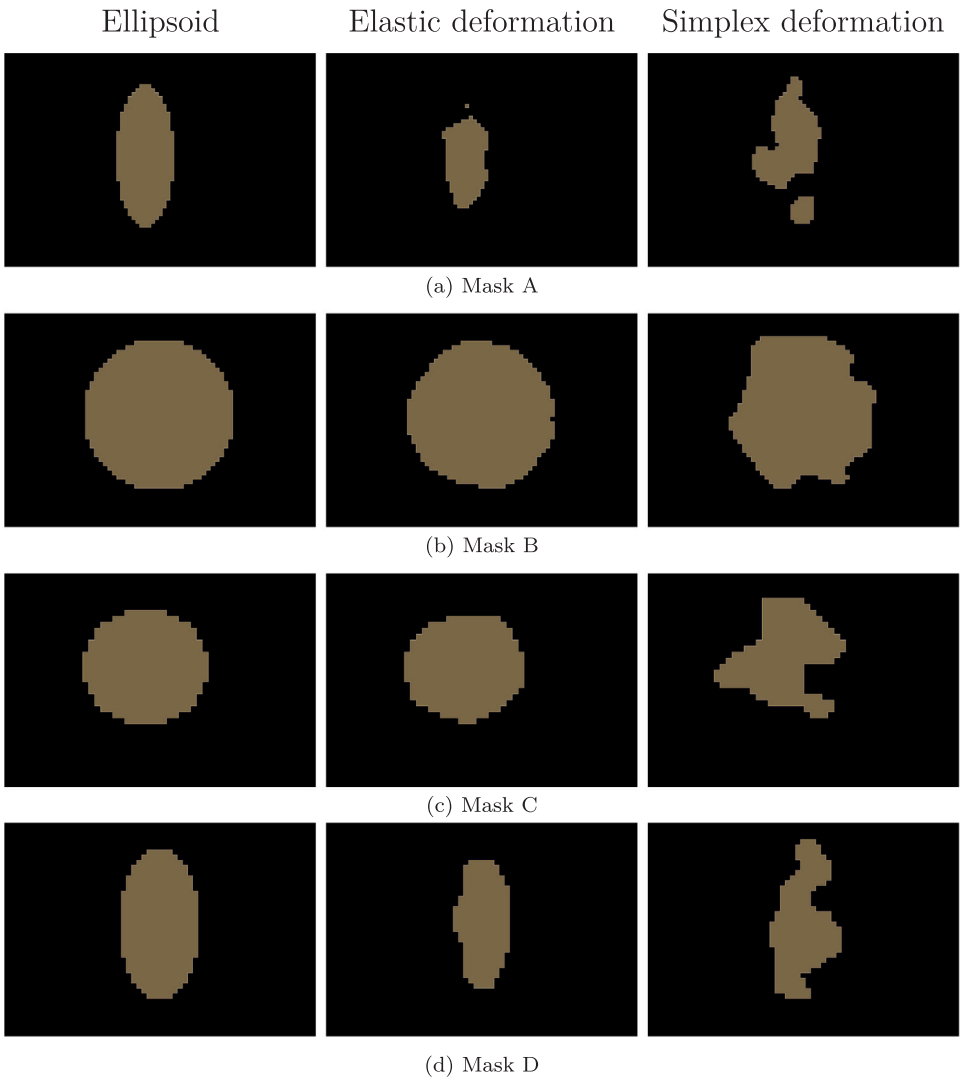


Fig. 5. Examples of four binary masks (a, b, c and d) generated from ellipsoids deformed using elastic deformation or simplex deformation. The same ellipsoid was used to produce the masks shown in each row.

the characteristics of adjacent hepatic tissue. As a result, the inpainted areas blend more naturally into the anatomical context.

The results of the visual Turing tests performed on DDPM based tumours are summarised in Table 2. All experts correctly identified the real tumours, with an average true positive rate of 95.5%, demonstrating their ability to accurately recognise genuine tumours. In contrast, the classification of synthetic tumours revealed significant variability in performance. The nuclear medicine specialist achieved a success rate of 100%, whereas the radiologists correctly classified synthetic tumours in fewer than 50% of cases. The inter-observer agreement was poor, with a Fleiss' Kappa coefficient of $\kappa = -0.04$, indicating a clear disagreement among the specialists. Upon retrospective inspection, we observed that most slices of synthetic tumours appeared realistic; however, one or two slices (typically at the extremities) occasionally displayed abnormal contrast, which could serve as subtle cues. A clinician particularly attentive to such details may therefore find it easier to identify synthetic lesions.

Finally, to evaluate the behaviour of the model with respect to different contrast injection phases, we visually assessed the DDPM output when applied to the same anatomical region across the arterial, portal, and delayed phases of MRI acquisition. As illustrated in Fig. 8, synthetic tumours were inpainted into the same location in a patient volume from the LLD-MMRI dataset, across all three phases. The generated tumours exhibit intensity profiles and textures that are consistent with the surrounding tissue contrast in each phase.

Table 2
True positive rates obtained during the visual Turing test conducted by experts for the classification of real and synthetic tumours.

Expert	Real tumours (%)	Synthetic tumours (%)
Radiologist 1	93.3	46.7
Radiologist 2	93.3	46.7
Nuclear medicine physician	100	100
Average	95.5 ± 4	64.5 ± 31

3.2. Tumour segmentation

3.2.1. Impact of the quantity of synthetic tumours (ATLAS only)

The performance of the segmentation model on the test dataset is summarised in Table 3. For models trained exclusively on the ATLAS dataset, the baseline model consistently ranked last across all evaluated metrics. Statistical analysis revealed that most models exhibited a significant improvement over the baseline, especially in terms of Dice score, SD, and recall. In particular, Dice scores ranged from 66.0% for the baseline to 72.7% for the nnFormer ATLAS simplex 500 model, representing an improvement of 6.7 points. Furthermore, models trained

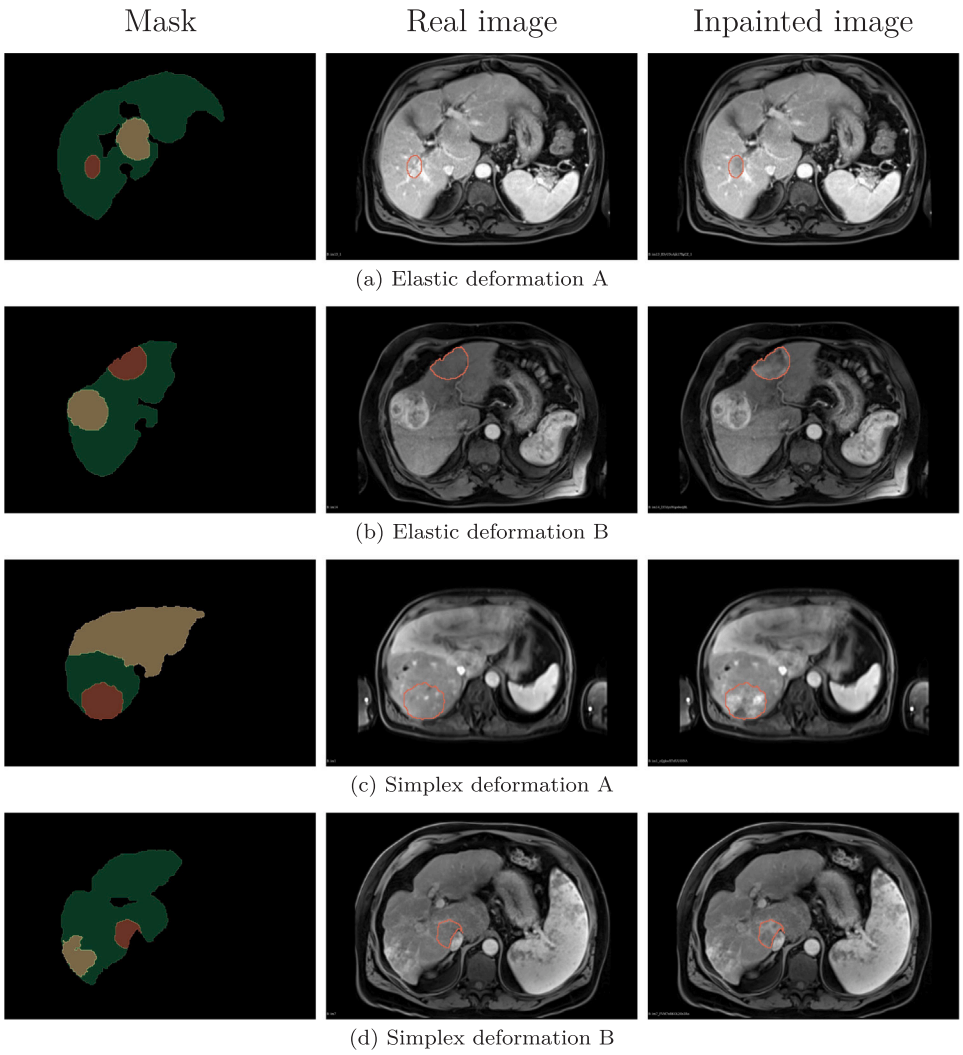


Fig. 6. Axial slices of masks and MRIs from four different patients, including the mask with an added synthetic tumour mask, the original image, and the image with an embedded synthetic tumour. In the masks, the liver is shown in green, real tumours in yellow, and synthetic tumours in red. The first two rows (a and b) correspond to masks generated from ellipsoids deformed using elastic deformation, and the last two rows (c and d) with simplex deformation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with synthetic tumours exhibited lower standard deviations than the baseline, especially for distance-based metrics, suggesting more precise and consistent segmentations.

3.2.2. Comparison of DDPM vs GAN-based generation

Using GANs to generate synthetic tumours leads to improvements in three to four out of six metrics compared to the baseline. However, these gains remain below those achieved by DDPMs. Both nnFormer ATLAS elastic 500 and nnFormer ATLAS simplex 500 outperform their GAN-based counterparts, nnFormer ATLAS GAN elastic 500 and nnFormer ATLAS GAN simplex 500, in all metrics, with Dice score increases of 2.1 and 4.9 points, respectively. A statistically significant difference was consistently observed between DDPM-based and GAN-based models on precision and HD metrics. Furthermore, significant differences were also found in recall and HD95 for the best-performing models.

3.2.3. Impact of external datasets (ATLAS + LLD-MMRI)

Both the nnFormer ATLAS + LLD-MMRI simplex 1000 and nnFormer ATLAS + LLD-MMRI elastic 1000 models showed statistically significant improvements on the Dice, SD, and recall metrics compared to the nnFormer baseline and nnFormer ATLAS + LLD-MMRI baseline models. In particular, the nnFormer ATLAS + LLD-MMRI simplex 1000 model, augmented with 1 000 additional synthetic images, outperformed all other models. It ranked first in four out of six evaluated metrics and achieved a Dice score of 76.0% compared to 70.0% for the nnFormer ATLAS + LLD-MMRI baseline. In addition, a reduction in standard deviations was observed for all metrics, indicating a more stable and robust performance.

3.2.4. Simplex vs elastic deformation

Although no real performance difference was observed between the use of elastic deformation and simplex deformation to produce the masks when training on the ATLAS dataset alone, a performance gap emerged in favour of simplex deformation when training was conducted on both ATLAS and LLD-MMRI. The nnFormer ATLAS + LLD-MMRI simplex 1000 model outperformed its elastic counterpart in five out of six measured metrics, with a 3.0 point advantage in Dice score. However, no statistically significant difference were measured.

3.2.5. Effect without standard data augmentation

Finally, the evaluation of the proposed method without standard data augmentation strategies demonstrated improved performance when generating masks using simplex deformation. In particular, this type of deformation resulted in the greatest performance gain compared

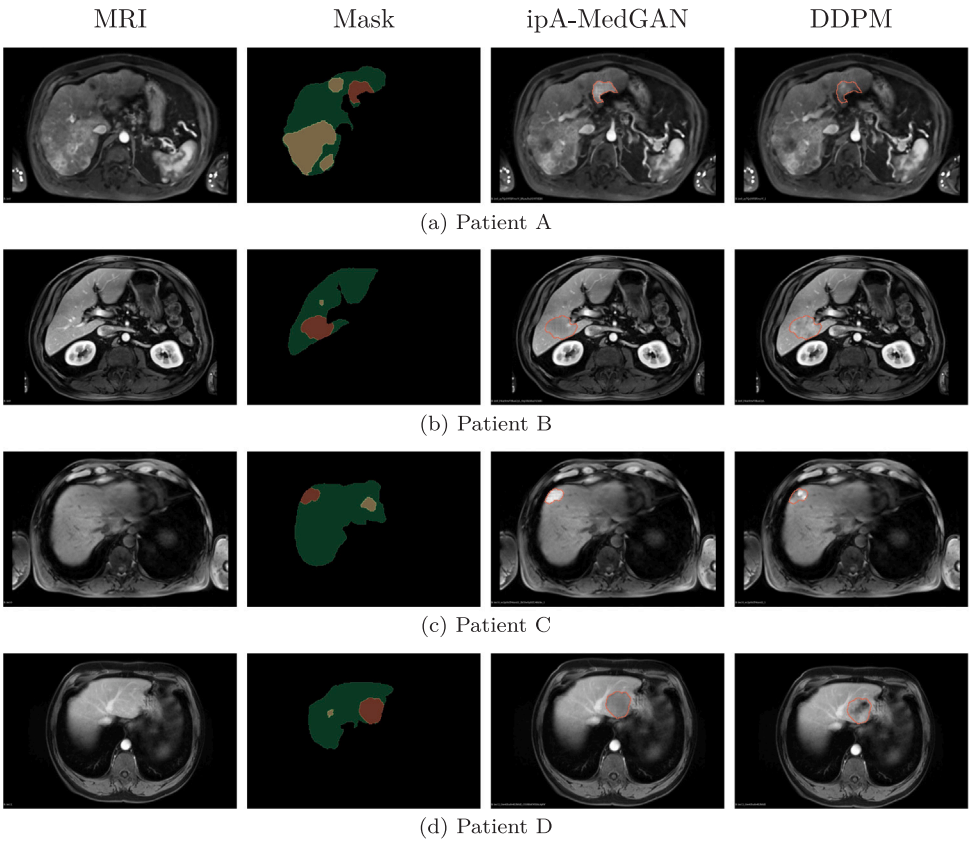


Fig. 7. Qualitative evaluation of generation performances between our DDPM based method and the ipA-MedGAN based method. Liver is shown in green, real tumours are in yellow and synthetic tumours in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

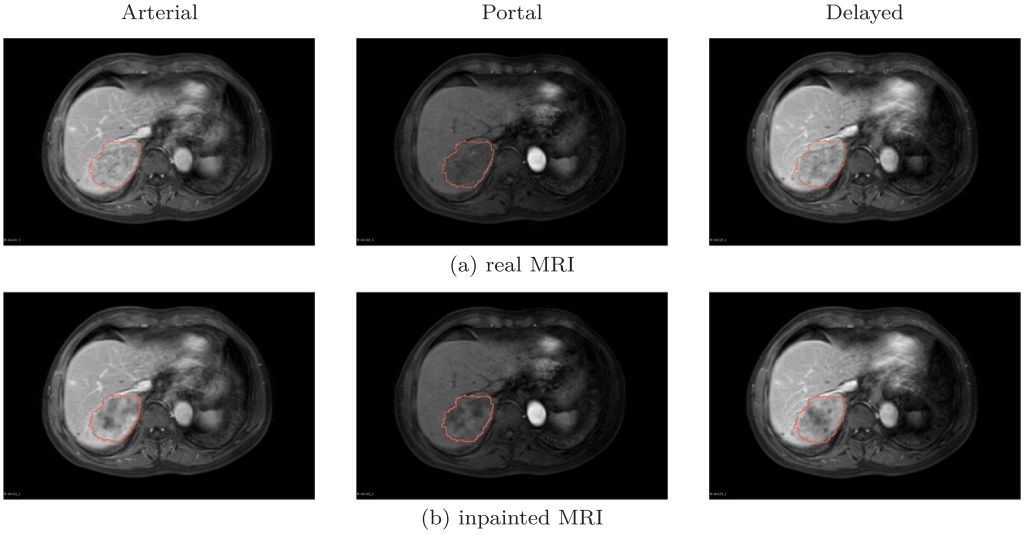


Fig. 8. Axial MRI slices from a patient in the LLD-MMRI dataset at three acquisition phases: Arterial, Portal, and Delayed. For each phase, images are shown before and after inpainting of the tumour region using the DDPM model. The inpainted area is outlined in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Quantitative comparison of segmentation performance on the ATLAS test set. For each model, baseline indicates that no synthetic images were used, elastic and simplex indicate that tumour masks were generated using elastic deformations or simplex deformation, the number specifies how many synthetic images were used during training, GAN indicates that the synthetic tumours were produced using the ipA-MedGAN model, ATLAS + LLD-MMRI indicates that the model was trained on both the ATLAS and LLD-MMRI datasets, and no aug denotes that no traditional data augmentation was used.

Model	Dice ↑	5 mm SD ↑	Precision ↑	Recall ↑	HD (mm) ↓	HD95 (mm) ↓
nnFormer ATLAS baseline	66.0 ± 25.9	61.3 ± 24.7	77.8 ± 26.9	61.6 ± 26.2	65.9 ± 67.8	40.4 ± 72.1
nnFormer ATLAS GAN elastic 500	69.7 ± 27.1	67.8 ± 27.2	81.6 ± 28.7	63.9 ± 27.2	75.8 ± 83.8	48.2 ± 87.7
nnFormer ATLAS GAN simplex 500	67.8 ± 27.6	64.9 ± 26.8	71.8 ± 30.8	67.9 ± 28.0	98.8 ± 71.6	54.1 ± 80.4
nnFormer ATLAS elastic 50	68.3 ± 26.5	65.8 ± 25.8	82.0 ± 24.5	63.1 ± 27.8	56.6 ± 36.0	27.9 ± 32.7
nnFormer ATLAS elastic 100	72.3 ± 23.9	70.9 ± 24.5	84.2 ± 21.0	68.1 ± 25.2	66.2 ± 60.5	31.8 ± 63.4
nnFormer ATLAS elastic 500	72.1 ± 21.6	71.8 ± 20.5	**85.9 ± 14.4**	67.4 ± 23.7	64.8 ± 39.5	28.8 ± 33.6
nnFormer ATLAS elastic 1000	72.3 ± 25.0	70.7 ± 25.0	81.8 ± 22.1	70.0 ± 25.5	62.0 ± 44.7	35.8 ± 42.1
nnFormer ATLAS simplex 50	70.0 ± 23.2	66.9 ± 22.4	81.4 ± 24.8	64.1 ± 24.3	65.2 ± 58.3	**26.7 ± 47.6**
nnFormer ATLAS simplex 100	71.3 ± 23.9	70.2 ± 23.4	79.5 ± 25.9	67.9 ± 25.0	58.0 ± 51.4	30.9 ± 51.6
nnFormer ATLAS simplex 500	72.7 ± 18.8	70.3 ± 19.3	80.8 ± 16.2	70.9 ± 22.7	66.5 ± 38.2	36.6 ± 39.0
nnFormer ATLAS simplex 1000	72.3 ± 23.8	71.5 ± 24.5	83.5 ± 20.2	68.3 ± 25.1	65.9 ± 43.3	39.6 ± 45.3
nnFormer ATLAS + LLD-MMRI baseline	70.0 ± 27.0	67.4 ± 27.3	81.0 ± 24.7	67.4 ± 28.4	62.1 ± 60.1	36.2 ± 61.3
nnFormer ATLAS + LLD-MMRI elastic 1000	73.0 ± 25.9	72.2 ± 26.6	83.1 ± 19.0	71.8 ± 28.1	63.8 ± 48.4	33.3 ± 43.2
nnFormer ATLAS + LLD-MMRI simplex 1000	**76.0 ± 20.5**	**75.5 ± 21.2**	82.7 ± 19.4	**73.7 ± 24.0**	**55.4 ± 44.6**	29.9 ± 38.0
nnFormer ATLAS no aug baseline	42.9 ± 22.2	38.1 ± 16.1	65.3 ± 31.6	35.1 ± 19.2	107.6 ± 39.6	51.5 ± 34.4
nnFormer ATLAS no aug elastic 500	53.4 ± 20.1	47.2 ± 18.2	61.0 ± 26.9	56.5 ± 21.6	124.0 ± 44.9	63.6 ± 39.4
nnFormer ATLAS no aug simplex 500	57.0 ± 21.0	49.0 ± 19.8	62.3 ± 26.3	62.4 ± 22.6	116.5 ± 43.7	64.9 ± 40.9

to the baseline, surpassing even the improvements achieved with elastic deformations. The model outperformed the baseline by 14.1 points in Dice score, 10.9 points in surface Dice, and 27.2 points in recall, all with statistically significant differences. Thanks to the increase in data, the model becomes bolder and detects more tumour zones. On the other hand, we note a drop in accuracy and a deterioration in distance metrics, reflecting the appearance of errors at the periphery of segmentations. Overall, this approach delivers more complete segmentations, at the cost of a slight increase in false positives.

3.3. Comparison of the best-performing models

The differences in Dice scores on the test set between nnFormer baseline, nnFormer ATLAS simplex 500, and nnFormer ATLAS + LLD-MMRI simplex 1000, as well as the distribution of their performance, are illustrated in Fig. 9. These models were selected because they represent key milestones in our evaluation: the baseline without synthetic data, the best-performing model trained only on ATLAS with synthetic augmentation (nnFormer ATLAS simplex 500), and the best-performing overall model trained on both datasets with synthetic data (nnFormer ATLAS + LLD-MMRI 1000). A qualitative comparison of segmentation results obtained by these models on the ATLAS test set is also provided in Fig. 10.

4. Discussion

4.1. Generation

In this study, a DDPM was trained to insert tumours into MRI scans, generating more realistic results than the GAN-based approach. The internal contrast variations produced by the DDPM contributed to a more heterogeneous and lifelike appearance. Moreover, as illustrated in Fig. 8, the model managed to produce phase-consistent synthetic tumours. In future work, explicitly incorporating the contrast phase as a conditioning input to the model could further improve consistency and realism across temporal sequences.

Nevertheless, limitations persist. Experts highlighted occasional inconsistencies within synthetic volumes, such as isolated slices appearing less convincing due to sharper boundaries, blurrier textures, or exaggerated contrast. Additionally, certain anatomical details, such as visible vascular structures present in real tumours, were absent. The complex, infiltrative nature of HCC remains challenging to replicate faithfully, indicating that further refinement is needed to enhance anatomical accuracy and variability in synthetic tumour generation.

Further limitations emerged when conducting supplementary tests on external datasets, such as the CHAOS [81] dataset. These experiments did not yield significant improvements, which may be explained by the substantial domain shift between the datasets. Nevertheless, extending the study to additional datasets would be valuable, as demonstrated by Chen et al. [54], who explored a variety of tumour types across multiple organs to increase data diversity instead of focussing only on HCC-CE MRIs. Adopting a similar strategy could enrich the variability of training examples and strengthen the robustness of future synthetic data generation pipelines. In the long term, a sufficiently consistent model trained across a wide range of datasets may also mitigate the need to retrain a network from scratch for each new task, thereby improving scalability and applicability.

In parallel, recent advances in generative modelling also point to potential methodological extensions. While this study relies on DDPMs, several more recent and computationally efficient diffusion-based architectures have demonstrated strong performance for conditional image synthesis and editing tasks. Notably, LDMs [36], as popularised by Stable Diffusion, operate in a compressed latent space, enabling higher-resolution synthesis with reduced computational cost. Interestingly, Chen et al. [54] have already explored a related direction by leveraging a latent diffusion framework trained on multi-organ tumour data. ControlNet [82] further extends this framework by introducing explicit structural conditioning, which could be particularly relevant for tumour inpainting guided by masks or anatomical priors. Other promising alternatives include score-based diffusion models [83], classifier-free guidance [84], and diffusion transformers [85], which have shown improved scalability and representation capacity. Integrating such architectures into the proposed pipeline represents a natural and promising direction for future work, with the potential to further improve realism, controllability, and generalisation.

Finally, our method also provides practical advantages through its controllable mask generation process. In particular, the use of simplex noise allows the creation of binary tumour masks with irregular, realistic shapes that are both diverse and reproducible. These masks can be easily parameterised to simulate tumours of varying complexity and size. Additionally, since the masks are generated in standard image formats, they remain fully compatible with image viewers, enabling potential interactive use by clinicians or researchers wishing to customise or inspect tumour placement.

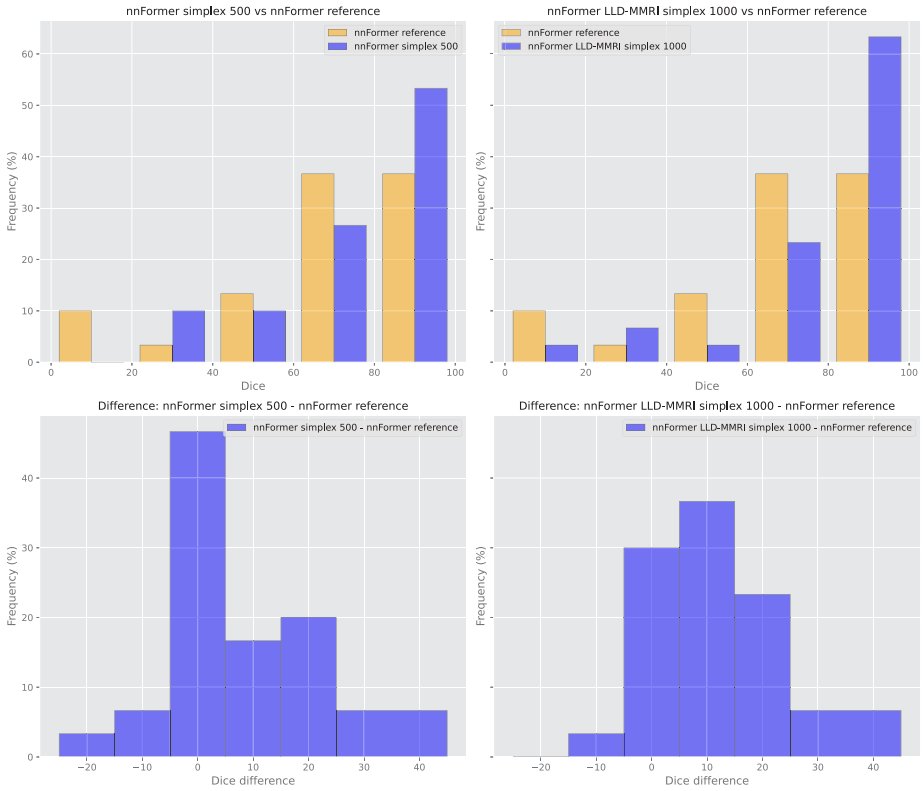


Fig. 9. Histograms illustrating performance differences in Dice scores between nnFormer ATLAS baseline, nnFormer ATLAS simplex 500, and nnFormer ATLAS + LLD-MMRI simplex 1000 on the ATLAS test set. The top two histograms show the performance distribution for each model. The bottom two histograms show the per-image difference.

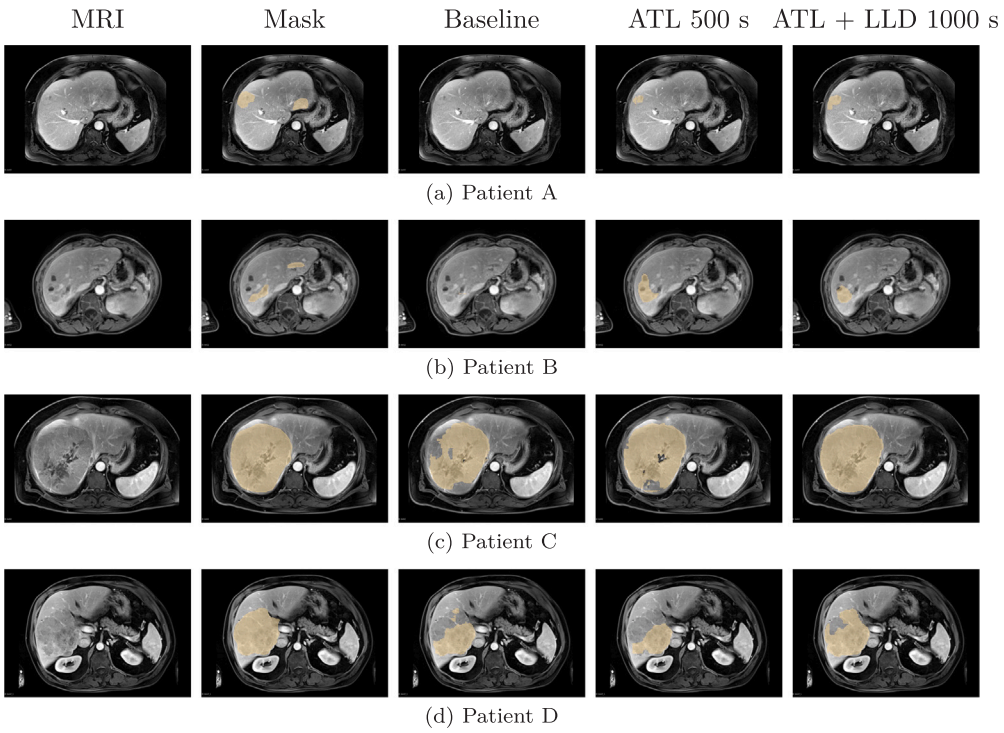


Fig. 10. Qualitative evaluation of tumour segmentation performance on the ATLAS test set. Tumours are shown in yellow. Baseline refers to nnFormer ATLAS baseline, ATL 500 s refers to nnFormer ATLAS simplex 500, and ATL + LLD 1000 s refers to nnFormer ATLAS + LLD-MMRI simplex 1000. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Tumour segmentation

By comparing the performance of the nnFormer ATLAS baseline model trained without synthetic data with that of models trained using synthetic tumours on the ATLAS dataset, it is striking that the models trained with synthetic images consistently outperform the nnFormer ATLAS baseline model. Furthermore, as shown in Table 3, the performance of the nnFormer ATLAS simplex 500 model exceeds that of the nnFormer ATLAS LLD-MMRI baseline model in three out of six evaluation metrics: Dice, 5 mm SD, and recall. This suggests that, beyond improving the baseline method, our approach can even surpass models trained with external datasets.

Finally, when combined with the external LLD-MMRI dataset, the nnFormer ATLAS + LLD-MMRI simplex 1000 model achieves a Dice score of 76.0% (Table 3), representing a notable increase of 10.0 points over the nnFormer ATLAS baseline and 6.0 points over the nnFormer ALTAS + LLD-MMRI baseline.

Our method not only outperforms models trained without external data; it also enhances segmentation performance when used in conjunction with external datasets. Distance metrics show a significant reduction and increased stability.

Analysing the segmentation results of the nnFormer ATLAS simplex 500 model on a per-image basis, we observe that, segmentation performance slightly decreases in 23% of cases, generally when the tumour was already well segmented. In 43% of cases, a slight improvement is noted, with gains between 0 and 10 Dice points. Finally, in the remaining 34%, a substantial improvement is observed, with gains ranging from 10 to 28 Dice points.

When external data are included, these effects are amplified. For the nnFormer ATLAS + LLD-MMRI simplex 1000 model, performance slightly decreases in only 7% of cases already well segmented. Modest gains (< 10 Dice points) are observed in 47% of cases, while major gains, ranging from 10 to 36 Dice points, are recorded in 46% of the cases. In particular, the number of cases where the model achieved a Dice score above 80% increased from 11 to 19 out of 30 images. Interestingly, on the few images in the test set presenting multiple tumours, similar performance gains were observed as for single-tumour cases, even though the average number of tumours per patient in the training data increased from 2.3 to 3.3 due to the addition of one synthetic tumour per case. While this change alters the case-level lesion count distribution, it does not negatively affect model behaviour. Notably, the model does not tend to hallucinate non-existent secondary lesions on the test set, even in cases where there is only one real tumour, indicating that the number of tumours present during training does not significantly influence the inference process.

It is also relevant to examine the precision and recall metrics. Although the baseline model shows high precision but low recall, the proposed method improves recall without compromising precision. As a result, predicted tumours are not only better segmented, but their sizes are also more representative of reality.

The performance gains are most pronounced in two scenarios: first, when the baseline model fails to detect certain tumours, as illustrated in Figs. 10(a) and 10(b); and second, when tumours are only partially detected, with missing portions, as shown in Figs. 10(c) and 10(d).

A model trained with our method appears to achieve the greatest improvements in the most complex cases, successfully identifying more distinct tumour regions where baseline models fail.

5. Conclusion

In this study, we proposed a novel generative data augmentation method to improve 3D automatic tumour segmentation, by inpainting synthetic tumours into real medical images. Synthetic binary tumour masks were created by deforming ellipsoids with elastic and simplex deformation and inserted into liver masks. A DDPM model was then used to generate realistic 3D tumours within the corresponding area in the MRI volumes. By enabling the generation of numerous modified images, this approach significantly enhanced segmentation performance. Furthermore, its effectiveness remained consistent when combined with external datasets containing pseudo-masks.

Our method, which enables the generation of large numbers of synthetic image instances without human intervention, is particularly relevant in the context of data-hungry segmentation models and could become a new standard for tumour segmentation. Although our technique shows good results, improvements remain possible, particularly regarding the boundaries of the generated structures.

CRediT authorship contribution statement

Félix Quinton: Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Benoit Presles:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition. **Romain Popoff:** Writing – review & editing, Resources, Data curation. **François Godard:** Validation, Resources. **Olivier Chevallier:** Validation, Resources, Data curation. **Julie Pellegrinelli:** Validation, Resources, Data curation. **Jean-Marc Vrigneaud:** Writing – review & editing, Resources, Data curation. **Jean-Louis Alberini:** Writing – review & editing, Validation, Supervision, Project administration. **Fabrice Meriaudeau:** Writing – review & editing, Validation, Supervision, Methodology.

Ethics statement

This study did not involve any prospective experimentation on human or animal subjects. All medical images used in this work were obtained from publicly available and fully anonymised datasets (ATLAS and LLD-MMRI). As such, ethical approval and informed consent were not required. The use of these datasets complies with relevant institutional and national guidelines. No private or identifiable patient data were used or disclosed in this study.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT 4o in order to improve the clarity and readability of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding sources

This work was supported by the Agence Nationale de la Recherche, France, grant number ANR-21-CE45-0002; the HPC resources of IDRIS under the allocation 2023-AD010313884 made by GENCI.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Details of tumour mask deformation using displacement fields

This appendix provides a detailed explanation of the mathematical formulations underlying the generation of tumour-like masks through elastic and simplex-based deformations.

A.1. Elastic deformation using Gaussian noise

The displacement field is constructed in two steps. For each spatial direction, Gaussian noise $g(p)$ is sampled from a standard normal distribution $\mathcal{N}(0,1)$. This noise is then smoothed using a Gaussian filter $\mathcal{G}_\sigma$, where the standard deviation σ controls both the magnitude and spatial coherence of the displacements. The displacement $d(p)$ at a voxel location p is given by:

$$d(p) = \mathcal{G}_\sigma(g(p)). \tag{A.1}$$

Once the displacement field is generated, it is applied to the voxel coordinates. For an initial voxel position $p = (x, y, z)$, the new position after deformation is:

$$p' = (x + d_x, y + d_y, z + d_z), \tag{A.2}$$

where d_x, d_y, and d_z are the displacement components along the x, y, and z axes, respectively.

Since the new voxel positions generally do not correspond to integer grid indices, interpolation is required to reconstruct the deformed image. For binary masks, nearest-neighbour interpolation is typically employed to preserve discrete structures.

A.2. Simplex noise-based deformation

Simplex noise is an enhanced variant of Perlin noise [73]. Simplex noise is particularly effective for generating organic and irregular deformations, making it suitable for modelling heterogeneous tumour contours.

Unlike Perlin noise, which relies on Cartesian grids, simplex noise subdivides space into a network of tetrahedra. Each point belongs to a specific tetrahedron whose vertices are used for interpolation.

For a point $p = (x, y, z)$, the coordinates are first projected into simplex space using the constant:

$$F_n = \frac{\sqrt{n+1} - 1}{n}. \tag{A.3}$$

In three-dimensional space ($n = 3$), we compute:

$$s_p = (x + y + z) \cdot F_3, \tag{A.4}$$

and determine the indices of the simplex containing point p:

$$i = \lfloor x + s_p \rfloor, \quad j = \lfloor y + s_p \rfloor, \quad k = \lfloor z + s_p \rfloor. \tag{A.5}$$

The coordinates are then unprojected using:

$$G_n = \frac{1 - \frac{1}{\sqrt{n+1}}}{n}, \tag{A.6}$$

which allows calculation of:

$$t = (i + j + k) \cdot G_3, \tag{A.7}$$

followed by the relative distances between p and the simplex vertices:

$$x_0 = x - (i - t), \quad y_0 = y - (j - t), \quad z_0 = z - (k - t). \tag{A.8}$$

Each simplex vertex is associated with a gradient vector, and contributions are weighted using the influence function r_s:

$$r_s = 0.6 - (x_0^2 + y_0^2 + z_0^2), \tag{A.9}$$

where $s \in S$, and S is the set of vertices of the simplex containing p. If $r_s > 0$, the contribution from vertex s is:

$$\text{contribution}_s = r_s^4 \cdot (g_s \cdot (x_0, y_0, z_0)), \tag{A.10}$$

where g_s is the gradient vector associated with vertex s, chosen pseudo-randomly via a permutation table to ensure coherent, non-repetitive noise.

Table B.4
Evaluation of the generative performance of DDPM models trained for different numbers of epochs, using SVM distance and FID as evaluation metrics. Models achieving the lowest values were selected for synthetic tumour generation in the downstream segmentation experiments. Two separate models were trained: one on the ATLAS dataset only, and another on the combined ATLAS and LLD-MMRI datasets. The numerical suffix in each model name indicates the training epoch at which the evaluation was performed, while "best" denotes the epoch corresponding to the best validation performance.

Model	Average SVM difference ↓	FID ↓
DDPM ATLAS 2000	**0.001**	**0.002**
DDPM ATLAS 3000	0.037	0.032
DDPM ATLAS 4000	0.017	0.009
DDPM ATLAS 5000	0.038	0.042
DDPM ATLAS best	0.027	0.017
DDPM ATLAS + LLD-MMRI 2000	0.020	0.006
DDPM ATLAS + LLD-MMRI 3000	0.004	0.002
DDPM ATLAS + LLD-MMRI 4000	0.041	0.019
DDPM ATLAS + LLD-MMRI 5000	0.021	0.012
DDPM ATLAS + LLD-MMRI best	**0.001**	**0.001**

The final simplex noise value at point p is obtained by summing the weighted contributions:

$$\text{Simplex}(p) = \sum_{s \in S} \text{contribution}_s. \tag{A.11}$$

To generate a displacement field using simplex noise, the noise is applied directly to the spatial coordinates (x, y, z), scaled by a factor λ to control spatial frequency, and shifted by an offset to ensure independence between displacement fields d_x, d_y, and d_z. For example, the d_x component is computed as:

$$d_x(x, y, z) = \sigma \cdot \text{Simplex}(x \cdot \lambda + \text{offset}_x, y \cdot \lambda + \text{offset}_y, z \cdot \lambda + \text{offset}_z), \tag{A.12}$$

where σ controls the displacement amplitude. The d_y and d_z components are computed similarly. The final displacement is applied as:

$$p' = (x + d_x(x, y, z), y + d_y(x, y, z), z + d_z(x, y, z)). \tag{A.13}$$

Appendix B. Generative model selection

To identify the best-performing DDPM models for synthetic tumour generation, we conducted a quantitative evaluation on the validation set. Specifically, for each trained model, all real tumours in the validation images were replaced with synthetic tumours generated using the corresponding DDPM. We then assessed the similarity between the synthetic and real tumours using two complementary metrics: the Fréchet Inception Distance (FID), computed using a feature extractor pre-trained on medical images, and the average distance to a one-class SVM fitted on real tumour latent representations.

These evaluations were performed independently for the models trained on the ATLAS dataset and those trained on both the ATLAS and LLD-MMRI datasets. As shown in Table B.4, the model trained on ATLAS with 2 000 epochs (DDPM ATLAS 2000) and the model trained on ATLAS + LLD-MMRI with at the best validation epoch (DDPM ATLAS + LLD-MMRI best) obtained the lowest SVM distances and FID scores, indicating a better match with the distribution of real tumours. These two models were thus selected for subsequent experiments involving synthetic image generation.

References

[1] Zhang W, Tanida J, Itoh K, Ichioka Y. Shift invariant pattern recognition neural network and its optical architecture. In: Proceedings of annual conference of the Japan society of applied physics. 1988.

[2] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989;1(4):541–51.

[3] Waibel A, Hanazawa T, Hinton G, Shikano K, Lang KJ. Phoneme recognition using time-delay neural networks. In: Backpropagation. Psychology Press; 2013, p. 35–61.

[4] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015, p. 234–41.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems, vol. 30, 2017.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, arXiv preprint arXiv:2010.11929.

[7] Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: A survey. IET Image Process 2022;16(5):1243–67.

[8] Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustainability 2021;13(3):1224.

[9] Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, Karimijafarbigloo S, Cohen JP, Adeli E, Merhof D. Medical image segmentation review: The success of U-net. 2022, ArXiv Preprint. URL: http://arxiv.org/abs/2211.14830.

[10] Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, Silva B, Sena J, Shickel B, Bihorac A, et al. Transformers and large language models in healthcare: A review. Artif Intell Med 2024;154:102900.

[11] Survarachakan S, Prasad PJR, Naseem R, de Frutos JP, Kumar RP, Langø T, Cheikh FA, Elle OJ, Lindseth F. Deep learning for image-based liver analysis—A comprehensive review focusing on malignant lesions. Artif Intell Med 2022;130:102331.

[12] Lakshmipriya B, Pottakkat B, Ramkumar G. Deep learning techniques in liver tumour diagnosis using CT and MR imaging-A systematic review. Artif Intell Med 2023;141:102557.

[13] Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: A systematic literature review. Comput Biol Med 2023;152:106391.

[14] Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med Image Anal 2020;63:101693.

[15] Raza K, Singh NK. A tour of unsupervised deep learning for medical image analysis. Curr Med Imaging 2021;17(9):1059–77.

[16] Jiao R, Zhang Y, Ding L, Xue B, Zhang J, Cai R, Jin C. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Comput Biol Med 2023;107840.

[17] Shurrab S, Duwairi R. Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Comput Sci 2022;8:e1045.

[18] You C, Zhou Y, Zhao R, Staib L, Duncan JS. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans Med Imaging 2022;41(9):2228–37.

[19] You C, Dai W, Min Y, Staib L, Duncan JS. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In: International conference on information processing in medical imaging. Springer; 2023, p. 641–53.

[20] You C, Dai W, Liu F, Min Y, Dvornek NC, Li X, Clifton DA, Staib L, Duncan JS. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. IEEE Trans Pattern Anal Mach Intell 2024.

[21] LeCun Y, Chopra S, Hadsell R, Ranzato M, Huang F. A tutorial on energy-based learning. Predict Struct Data 2006;1.

[22] Du Y, Mordatch I. Implicit generation and modeling with energy based models. In: Advances in Neural Information Processing Systems, vol. 32, 2019.

[23] Kingma DP, Welling M. Auto-encoding variational bayes. 2013, arXiv preprint arXiv:1312.6114.

[24] Kingma DP, Mohamed S, Jimenez Rezende D, Welling M. Semi-supervised learning with deep generative models. In: Advances in neural information processing systems, vol. 27, 2014.

[25] Kingma DP, Welling M, et al. An introduction to variational autoencoders. Found Trends Mach Learn 2019;12(4):307–92.

[26] Kobyzev I, Prince SJD, Brubaker MA. Normalizing flows: An introduction and review of current methods. IEEE Trans Pattern Anal Mach Intell 2020;43(11):3964–79.

[27] Papamakarios G, Nalisnick E, Rezende DJ, Mohamed S, Lakshminarayanan B. Normalizing flows for probabilistic modeling and inference. J Mach Learn Res 2021;22(57):1–64.

[28] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun ACM 2014;63(11):139–44. http://dx.doi.org/10.1145/3422622, URL: https://arxiv.org/abs/1406.2661v1.

[29] Jordon J, Yoon J, Van Der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In: International conference on learning representations. 2018.

[30] Yoon J, Jarrett D, der Schaar M. Time-series generative adversarial networks. In: Advances in neural information processing systems, vol. 32, 2019.

[31] Kazeminia S, Baur C, Kuijper A, Van Ginneken B, Navab N, Albarqouni S, Mukhopadhyay A. GANs for medical image analysis. Artif Intell Med 2020;109:101938.

[32] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Commun ACM 2020;63(11):139–44.

[33] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Advances in neural information processing systems, vol. 33, 2020, p. 6840–51.

[34] Vahdat A, Kreis K, Kautz J. Score-based generative modeling in latent space. In: Advances in neural information processing systems, vol. 34, 2021, p. 11287–302.

[35] Song Y, Shen L, Xing L, Ermon S. Solving inverse problems in medical imaging with score-based generative models. 2021, arXiv preprint arXiv:2111.08005.

[36] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 10684–95.

[37] Fernandez V, Pinaya WHL, Borges P, Tudosiu P-D, Graham MS, Vercauteren T, Cardoso MJ. Can segmentation models be trained with fully synthetically generated data? In: International workshop on simulation and synthesis in medical imaging. Springer; 2022, p. 79–90.

[38] Kim B, Oh Y, Ye JC. Diffusion adversarial representation learning for self-supervised vessel segmentation. 2022, arXiv preprint arXiv:2209.14566.

[39] Wolleb J, Sandkühler R, Bieder F, Valmaggia P, Cattin PC. Diffusion models for implicit image segmentation ensembles. In: International conference on medical imaging with deep learning. PMLR; 2022, p. 1336–48.

[40] Xun S, Li D, Zhu H, Chen M, Wang J, Li J, Chen M, Wu B, Zhang H, Chai X, et al. Generative adversarial networks in medical image segmentation: A review. Comput Biol Med 2022;140:105063.

[41] Zhu Q, Wang Y, Yin L, Yang J, Liao F, Li S. Selfmix: a self-adaptive data augmentation method for lesion segmentation. In: International conference on medical image computing and computer-assisted intervention. 2022, p. 683–92.

[42] Basaran BD, Zhang W, Qiao M, Kainz B, Matthews PM, Bai W. LesionMix: A lesion-level data augmentation method for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2023, p. 73–83.

[43] Zhang X, Liu C, Ou N, Zeng X, Zhuo Z, Duan Y, Xiong X, Yu Y, Liu Z, Liu Y, Ye C. CarveMix: A simple data augmentation method for brain lesion segmentation. NeuroImage 2023;271. http://dx.doi.org/10.1016/j.neuroimage.2023.120041.

[44] Shin Y, Qadir HA, Balasingham I. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access 2018;6:56007–17.

[45] Horvath I, Paetzold J, Schoppe O, Al-Maskari R, Ezhov I, Shit S, Li H, Ertürk A, Menze B. Metgan: Generative tumour inpainting and modality synthesis in light sheet microscopy. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022, p. 227–37.

[46] Li B, Chou Y-C, Sun S, Qiao H, Yuille A, Zhou Z. Early detection and localization of pancreatic cancer by label-free tumor synthesis. 2023, arXiv preprint arXiv:2308.03008.

[47] Wei Z, Chen Y, Guan Q, Hu H, Zhou Q, Li Z, Xu X, Frangi A, Chen F. Pancreatic image augmentation based on local region texture synthesis for tumor segmentation. In: International conference on artificial neural networks, vol. 1, 2023, p. 419–31.

[48] Bilic P, Christ P, Li HB, Vorontsov E, Ben-Cohen A, Kaissis G, Szeskin A, Jacobs C, Mamani GEH, Chartrand G, et al. The liver tumor segmentation benchmark (lits). Med Image Anal 2023;84:102680.

[49] Hu Q, Xiao J, Chen Y, Sun S, Chen J-N, Yuille A, Zhou Z. Synthetic tumors make ai segment tumors better. 2022, arXiv preprint arXiv:2210.14845.

[50] Hu Q, Chen Y, Xiao J, Sun S, Chen J, Yuille AL, Zhou Z. Label-free liver tumor segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 7422–32.

[51] Zhang Z, Deng H, Li X. Unsupervised liver tumor segmentation with pseudo anomaly synthesis. In: International workshop on simulation and synthesis in medical imaging. Springer; 2023, p. 86–96.

[52] Zavrtanik V, Kristan M, Skočaj D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 8330–9.

[53] Wu L, Zhuang J, Ni X, Chen H. FreeTumor: Advance tumor segmentation via large-scale tumor synthesis. 2024, arXiv preprint arXiv:2406.01264.

[54] Chen Q, Chen X, Song H, Xiong Z, Yuille A, Wei C, Zhou Z. Towards generalizable tumor synthesis. 2024, arXiv preprint arXiv:2402.19470.

[55] Peng L, Zhang Z, Durak G, Miller FH, Medetalibeyoglu A, Wallace MB, Bagci U. Optimizing synthetic data for enhanced pancreatic tumor segmentation. In: International workshop on personalized incremental learning in medicine. Springer; 2024, p. 35–44.

[56] Lennon Chan LY, Li C, Yuan Y. AutoPET challenge: Tumour synthesis for data augmentation. 2024, p. arXiv–2409, arXiv e-Prints.

[57] Li X, Shuai Y, Liu C, Chen Q, Wu Q, Guo P, Yang D, Zhao C, Bassi PRAS, Xu D, et al. Text-driven tumor synthesis. 2024, arXiv preprint arXiv:2412.18589.

[58] Fei L, Mang Y, Andy J. M, Terry C-FY, Grace L-HW, Pong CY. Learning from synthetic CT images via test-time training for liver tumor segmentation. IEEE Trans Med Imaging 2022;41(9):2510–20.

[59] Hu Q, Yuille A, Zhou Z. Synthetic data as validation. 2023, arXiv preprint arXiv:2310.16052.

[60] Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, Farahani K, Kalpathy-Cramer J, Kitamura FC, Pati S, Prevedello LM, Rudie JD, Sako C, Shinohara RT, Bergquist T, Chai R, Eddy J, Elliott J, Reade W, Schaffter T, Yu T, Zheng J, Moawad AW, Coelho LO, McDonnell O, Miller E, Moron FE, Oswood MC, Shih RY, Siakallis L, Bronstein Y, Mason JR, Miller AF, Choudhary G, Agarwal A, Besada CH, Derakhshan JJ, Diogo MC, Do-Dai DD, Farage L, Go JL, Hadi M, Hill VB, Iv M, Joyner D, Lincoln C, Lotan E, Miyakoshi A, Sanchez-Montano M, Nath J, Nguyen XV, Nicolas-Jilwan M, Jimenez JO, Ozturk K, Petrovic BD, Shah C, Shah LM, Sharma M, Simsek O, Singh AK, Soman S, Statsevych V, Weinberg BD, Young RJ, Ikuta I, Agarwal AK, Cambron SC, Silbergleit R, Dusoi A, Postma AA, Letourneau-Guillon L, Perez-Carrillo GJG, Saha A, Soni N, Zaharchuk G, Zohrabian VM, Chen Y, Cekic MM, Rahman A, Small JE, Sethi V, Davatzikos C, Mongan J, Hess C, Cha S, Villanueva-Meyer J, Freymann JB, Kirby JS, Wiestler B, Crivellaro P, Colen RR, Kotrotsou A, Marcus D, Milchenko M, Nazeri A, Fathallah-Shaykh H, Wiest R, Jakab A, Weber M-A, Mahajan A, Menze B, Flanders AE, Bakas S. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. 2021, ArXiv Preprint. URL: http://arxiv.org/abs/2107.02314.

[61] Wolleb J, Sandkühler R, Bieder F, Cattin PC. The swiss army knife for image-to-image translation: Multi-task diffusion models. 2022, arXiv preprint arXiv:2204.02641.

[62] Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vahdati S, Erickson BJ. Multitask brain tumor inpainting with diffusion models: A methodological report. 2022, arXiv preprint arXiv:2210.12113.

[63] Kebaili A, Lapuyade-Lahorgue J, Vera P, Ruan S. 3D MRI synthesis with slice-based latent diffusion models: Improving tumor segmentation tasks in data-scarce regimes. 2024, URL: http://arxiv.org/abs/2406.05421.

[64] Durrer A, Wolleb J, Bieder F, Friedrich P, Melie-Garcia L, Ocampo Pineda MA, Bercea CI, Hamamci IE, Wiestler B, Piraud M, et al. Denoising diffusion models for 3d healthy brain tissue inpainting. In: MICCAI workshop on deep generative models. Springer; 2024, p. 87–97.

[65] Chung W, Kim H. The histopathological and molecular heterogeneity of hepatocellular carcinoma: a narrative review. Ewha Med J 2024;47(4).

[66] Jia X, Jiang H, Ye Z, Wei H, Chen J, Qu Y, Sirlin CB, Song B, Wang Y. Imaging for molecular and pathological subtyping of hepatocellular carcinoma—a critical appraisal and future directions. Eur Radiol 2025;1–20.

[67] Quinton F, Popoff R, Presles B, Leclerc S, Meriaudeau F, Nodari G, Lopez O, Pellegrinelli J, Chevallier O, Ginhac D, Vrigneaud JM, Alberini JL. A tumour and liver automatic segmentation (ATLAS) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. Data 2023;8(5). http://dx.doi.org/10.3390/data8050079.

[68] Lou M, Liu X, Zhang Y, Yu Y, Zhou H-Y. Liver lesion diagnosis challenge on multi-phase MRI (LLD-MMRI2023). 2023, URL: https://github.com/LMMMEng/LLD-MMRI2023.

[69] Fredriksen V, Sevle SOM, Pedersen A, Langø T, Kiss G, Lindseth F. Teacher-student approach for lung tumor segmentation from mixed-supervised datasets. Plos One 2022;17(4):e0266147.

[70] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2021;18(2):203–11.

[71] Ogden RW. Non-linear elastic deformations. Courier Corporation; 1997.

[72] Perlin K. Improving noise. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. 2002, p. 681–2.

[73] Perlin K. An image synthesizer. ACM Siggraph Comput Graph 1985;19(3):287–96.

[74] Song J, Meng C, Ermon S. Denoising diffusion implicit models. 2020, arXiv preprint arXiv:2010.02502.

[75] Armanious K, Kumar V, Abdulatif S, Hepp T, Gatidis S, Yang B. ipA-MedGAN: Inpainting of arbitrarily regions in medical modalities. In: ArXiv. IEEE; 2019, p. 3005–9.

[76] Pinon N, Trombetta R, Lartizien C. One-Class SVM on siamese neural network latent space for unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities. In: Medical imaging with deep learning. PMLR; 2024, p. 1783–97.

[77] Quinton F, Presles B, Leclerc S, Nodari G, Lopez O, Chevallier O, Pellegrinelli J, Vrigneaud J-M, Popoff R, Meriaudeau F, et al. Navigating the nuances: comparative analysis and hyperparameter optimisation of neural architectures on contrast-enhanced MRI for liver and liver tumour segmentation. Sci Rep 2024;14(1):3522.

[78] Zhou H-Y, Guo J, Zhang Y, Han X, Yu L, Wang L, Yu Y. Nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Trans Image Process 2023.

[79] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:1–28.

[80] Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, Murrey B, Myronenko A, Zhao C, Yang D. MONAI: An open-source framework for deep learning in healthcare. 2022, arXiv preprint arXiv:2211.02701.

[81] Kavur AE, Gezer NS, Barış M, Aslan S, Conze P-H, Groza V, Pham DD, Chatterjee S, Ernst P, Özkan S, Baydar B, Lachinov D, Han S, Pauli J, Isensee F, Perkonigg M, Sathish R, Rajan R, Sheet D, Dovletov G, Speck O, Nürnberger A, Maier-Hein KH, Bozdağı Akar G, Ünal G, Dicle O, Selver MA. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med Image Anal 2021;69:101950.

[82] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 3836–47.

[83] Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, Poole B. Score-based generative modeling through stochastic differential equations. 2020, arXiv preprint arXiv:2011.13456.

[84] Ho J, Salimans T. Classifier-free diffusion guidance. 2022, arXiv preprint arXiv:2207.12598.

[85] Peebles W, Xie S. Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 4195–205.