# Deterministic Guided Progressive Medical Image Cross-Modal Generation Based on Deep Learning

Chujie Zhang[1], Lanfen Lin[2], and Yen-Wei Chen[1]

[1] Ritsumeikan University, Osaka, Japan
[2] Zhejiang University, Hangzhou, Zhejiang, China

**Abstract.** In this paper, we address critical challenges in medical image generation using deep learning techniques. While convolutional neural networks and generative adversarial networks (GANs) have achieved remarkable results in various image generation tasks, their application to medical imaging faces unique obstacles. These include the complexity and diversity of medical images, limitations in discriminator network structures, and the risk of model collapse and gradient vanishing in multi-scale discriminators. To overcome these issues, we propose a novel deterministic guided progressive GAN that specifically targets regions of interest (ROI) in medical images. Our approach progressively integrates adversarial generative networks, evolving from single-scale to multi-scale discriminators, to produce higher quality images. We demonstrate the efficacy of our model in generating high-precision cross-modal medical images through four comprehensive evaluation criteria, providing both quantitative and qualitative evidence of its performance compared to real images. This innovative method promises to significantly advance the field of medical image generation, potentially enhancing diagnostic accuracy and research capabilities in healthcare.

**Keywords:** Medical image synthesis · Deep learning · Progressive generative adversarial network · Multi-scale discriminators

## 1 Introduction

Medical imaging is a crucial diagnostic and research tool that provides visual representations of anatomical structures, playing a vital role in disease diagnosis and surgical planning [2]. Computed tomography (CT) and magnetic resonance imaging (MRI) are the most commonly used techniques in current clinical practice. These imaging modes offer complementary information, and their effective integration can significantly enhance medical decision-making [1]. However, obtaining paired multi-modal images is challenging, creating an increasing need for advanced multi-modal image generation techniques to support clinical diagnosis and treatment.

Medical image generation techniques have evolved from traditional machine learning methods to deep learning approaches. Earlier methods relied on explicit feature representation, such as random forests and k-nearest neighbor algorithms, optimizing feature representation through iterative methods. In recent years, convolutional neural networks, particularly generative adversarial

networks (GANs), have achieved state-of-the-art performance in various image generation tasks [4, 9–11].

Current methods often employ conditional GAN architectures with deterministic outputs, typically using L1/L2-based loss functions to learn deterministic mappings. However, these approaches do not explicitly model robustness to outliers or predictive uncertainty, leading to performance degradation when encountering unseen out-of-distribution patterns during testing [3]. While these methods can produce synthetic images of high visual quality, the content may still deviate significantly from the corresponding ground-truth values [8]. This discrepancy can lead to overconfidence or misinterpretation, potentially resulting in negative consequences, especially in medical applications. Additionally, these methods often focus on generating entire images, which can cause deformation of target regions without prior knowledge and result in poor quality generation of local target areas, manifesting as blurriness or unreasonable textures.

The discriminator in existing architectures typically uses a single-scale (Markovian) discriminator due to the significant differences in data distribution between different modalities in cross-modal medical image generation [7]. While multi-scale discriminators are common in natural image generation, their use in medical image generation often leads to mode collapse or gradient disappearance. A stronger discriminator generally produces higher quality results, and recent studies have shown that using multiple discriminator ensembles can enhance output quality [5]. However, these methods still lack guided step-by-step generation of high-quality images and fail to give special attention to areas with poorer generated results.

This paper aims to present a deep learning-based method for cross-modality medical image generation that addresses the limitations of existing techniques. Our approach utilizes a progressive adversarial generation network that gradually transitions from a single discriminator to multiple discriminators of varying complexities. This method aims to deterministically guide the generation of higher quality images, focusing on tumors or other regions of interest (ROI). Our main contributions are:

- We develop a multi-modality medical image generation technique that operates independently of paired data. This innovative method significantly improves the local generation quality of target areas, overcoming the limitations of traditional approaches that require aligned image pairs.
- We implement a progressive multi-scale discriminator approach within the adversarial generation network. This strategy guides the generator to focus on critical target regions, resulting in higher quality image production. Our method effectively mitigates common GAN issues such as gradient vanishing and mode collapse, while also enhancing the utility of generated images for downstream diagnostic and analytical tasks.
- We create a flexible and modular method that can be easily integrated with existing medical image generation algorithms. This design allows for performance enhancement without altering the original network structure,

thereby improving the quality of target area generation across various imaging modalities and frameworks.

## 2 Methods

### 2.1 Overview

Our proposed method for progressive medical image cross-modal generation utilizes a deep learning approach based on three adversarial generative modules, all built upon the CycleGAN architecture. This model is designed to generate high-quality cross-modal medical images, focusing particularly on liver and tumor regions. The process involves a staged approach, progressively incorporating more specialized discriminators to refine the generated images.

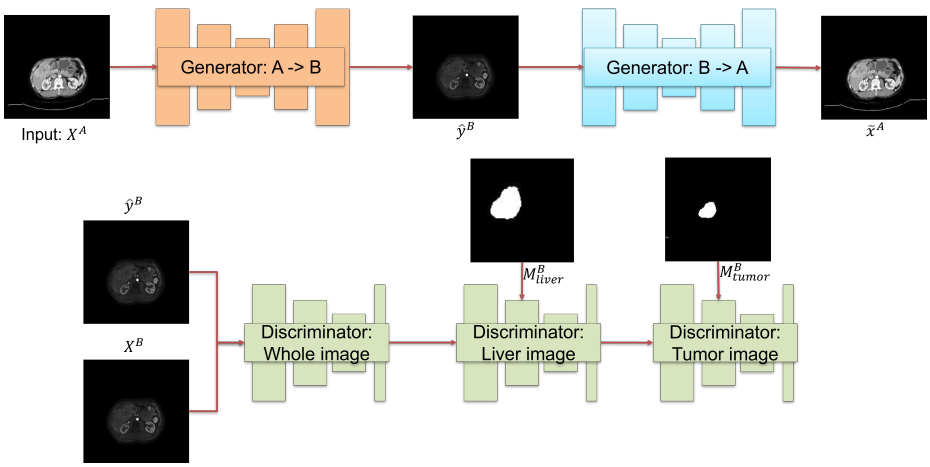


**Fig. 1:** Comprehensive Training Framework for Medical Image Cross-Modal Generation.

### 2.2 Progressive Multi-Scale GAN Architecture

The Progressive Multi-Scale GAN Architecture, as illustrated in Figure 1, consists of generators and multiple discriminators:

**Generators** Two UNet-based generators are employed:

- For CT to MR: $G_{A\to B}$ (Generator).
- For MR to CT: $G_{B\to A}$ (Generator).

where $A$ and $B$ represent CT and MR modalities respectively,

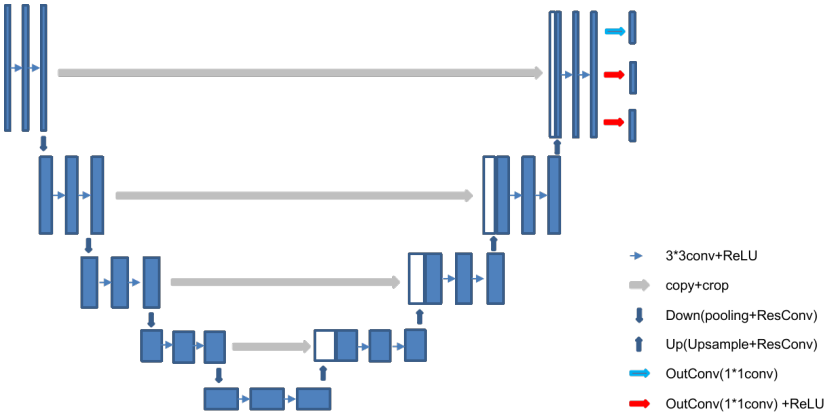


**Fig. 2:** Architecture of the Generator Network for Medical Image Cross-Modal Generation.

**Discriminators** The model incorporates three types of discriminators:

1. Whole image discriminators: $D_{whole}$.
2. Liver region multi-scale discriminators: $D_{liver}$ (two scales).
3. Tumor region multi-scale discriminators: $D_{tumor}$ (three scales).

The generator's network architecture, based on a modified UNet structure, is meticulously designed for cross-modal medical image generation, as illustrated in Figure 2. This sophisticated design incorporates four essential sub-modules: ResConv, Down, Up, and OutConv. The ResConv module implements a residual structure with 3x3 convolutions and padding of 1, enhancing feature preservation. The Down module combines max pooling with ResConv for effective feature downsampling, while the Up module employs bilinear upsampling (doubling input size) followed by ResConv, ensuring precise alignment of corner pixels between input and output tensors. The OutConv module applies a 1x1 convolution for final refinement. A distinctive feature of this generator is its tri-headed output layer, which produces not only the generated image but also scale and shape maps. These additional outputs are crucial for implementing the zero-mean generalized Gaussian distribution loss function, significantly enhancing the model's capacity to capture intricate image characteristics and improve overall generation quality, particularly in the context of medical imaging where precision is paramount.

The discriminator network in our progressive multi-scale GAN architecture employs a sophisticated Markovian design, crucial for fine-grained image analysis. This structure utilizes 4x4 convolutions with padding of 1 and stride of 2, activated by LeakyReLU (negative slope 0.2). Departing from conventional CNN-based classifiers that typically end with fully connected layers, our Markovian discriminator consists entirely of convolutional layers, producing an n x n output matrix. The final classification is determined by averaging this matrix, allowing for more nuanced spatial assessment. As illustrated in Figure 1, the model

progressively integrates specialized discriminators across three GAN modules, each focusing on increasingly specific image regions. The first module introduces a whole-image discriminator, providing a global assessment. The second module adds a liver-focused multi-scale discriminator, operating at two scales: original and half-downsampled. The third module incorporates a tumor-centric multi-scale discriminator, functioning at three scales: original, half-downsampled, and quarter-downsampled. This progressive, multi-scale approach enables increasingly refined discrimination, particularly in diagnostically crucial regions such as the liver and tumors. By gradually narrowing the focus from whole images to specific anatomical structures, our model enhances its ability to generate highly accurate and detailed cross-modal medical images, addressing the unique challenges of medical imaging tasks.

### 2.3 Loss Functions

Our model builds upon the CycleGAN framework, incorporating a sophisticated loss function that combines multiple components to ensure high-quality image generation and domain transfer. The core of our loss function is the zero-mean generalized Gaussian distribution loss, defined as:

$$L^G_{\alpha\beta}(\hat{y}^B, \alpha, \beta, x^B) = \frac{1}{K}\sum(\frac{|\hat{y}^B - x^B|}{\alpha})^\beta - \log\frac{\beta}{\alpha} + \log\Gamma(\beta^{-1}) \tag{1}$$

In this equation, $\alpha$ and $\beta$ represent the scale and shape maps generated by the generator, respectively. $\hat{y}^B$ denotes the generated MR image, while $x^B$ is the real MR image. $K$ signifies the total number of pixels in the data, and $\Gamma$ is the gamma function.

To ensure cycle consistency, we employ:

$$L^G_{cyc} = E_{x\ P_{data}(x)}[||\hat{x}^A - x^A||_1] \tag{2}$$

Our model features three adversarial modules, each with its own generator and discriminator loss:

Whole-image discrimination:

$$L^G_{adv1} = L_2(D_{whole}(\hat{y}^B), 1) \tag{3}$$

$$L^D_{adv1} = L_2(D_{whole}(\hat{y}^B), 0) + L_2(D_{whole}(x^B), 1) \tag{4}$$

Liver-focused discrimination:

$$L^G_{adv2} = L_2(D_{liver}(\hat{y}^B, M^B_{liver}), 1) \tag{5}$$

$$L^D_{adv2} = L_2(D_{liver}(\hat{y}^B, M^B_{liver}), 0) + L_2(D_{liver}(x^B, M^B_{liver}), 1) \tag{6}$$

Tumor-centric discrimination:

$$L^G_{adv3} = L_2(D_{tumor}(\hat{y}^B, M^B_{tumor}), 1) \tag{7}$$

$$L^D_{adv3} = L_2(D_{tumor}(\hat{y}^B, M^B_{tumor}), 0) + L_2(D_{tumor}(x^B, M^B_{tumor}), 1) \quad (8)$$

Here, $L^G_{adv1}$, $L^D_{adv1}$, $L^G_{adv2}$, $L^D_{adv2}$, $L^G_{adv3}$, and $L^D_{adv3}$ correspond to the generator and discriminator loss functions for each of the three adversarial modules.

The comprehensive loss function that guides our model's training is the sum of all these components:

$$L = L^G_{\alpha\beta}(\hat{y}^B, \alpha, \beta, x^B) + L^G_{cyc} + L^G_{adv1} + L^D_{adv1} + L^G_{adv2} + L^D_{adv2} + L^G_{adv3} + L^D_{adv3} \quad (9)$$

This multi-faceted loss function enables our model to generate high-fidelity cross-modal medical images while maintaining anatomical accuracy and preserving crucial diagnostic features.

## 3 Experimental Results

### 3.1 Dataset

In this paper, private data from a hospital was used, which includes magnetic resonance imaging (MRI) and computed tomography (CT) images of 305 patients, as well as corresponding tumor region masks. The dataset was divided into training, validation, and testing sets in a ratio of 6:2:2.

To train the progressive medical image cross-modality generation model, we proposed a preprocessing workflow for MR and CT images. The first step is to adjust the window width and window level. For CT images, we modified the image intensity by setting the window width and window level based on prior knowledge from doctors, in order to remove the histogram difference identified in the entire dataset. For MR images, we used the algorithm proposed by Manjón et al [12]. The second step is pixel normalization. For CT images, we directly used linear normalization to scale the pixel values between -1 and 1. For MR images, we first used the z-score algorithm and then used linear normalization to scale the pixel values between -1 and 1. The final step is data selection. To select data with tumors, we calculated the index of the slice with the largest tumor in the tumor mask data and selected four slices above and below this slice, totaling nine slices as the dataset used for each patient. In total, 2745 pairs of CT and MRI images were created.

### 3.2 Evaluation metrics

In the field of image generation, due to the limitation of human vision, the authenticity of generated images can only be subjectively evaluated. We used four different evaluation criteria to evaluate the model. The first one is based on the peak signal-to-noise ratio (PSNR) of the tumor area. This evaluation criterion is based on the characteristics of PSNR, mainly using the characteristics of PSNR to evaluate the tumor area of liver, as shown in formula 9.

$$PSNR = 10 \times \log_{10}(\frac{(2^n - 1)^2}{MSE}) \tag{10}$$

The second evaluation metric is based on the structural similarity of the tumor region, as shown in equation 10. Structural similarity is a measure of the similarity between two images. The structural similarity algorithm is mainly used to detect the similarity between two images of the same size or to detect the degree of distortion in an image. In this paper, we only calculate the structural similarity of the tumor region to validate the model.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{11}$$

In which $\sigma_x^2$ is the variance of x, $\sigma_y^2$ is the variance of y, $\mu_x$ is the mean of x, $\mu_y$ is the mean of y, $\sigma_{xy}$ is the covariance of x and y, and $c_1$ and $c_2$ are two constants.

The third one is based on learned image perceptual similarity, which was first proposed by Richard Zhang et al. in 2018 to measure the difference between two images [13]. This metric learns the inverse mapping from generated images to real ones, forcing the generator to learn the inverse mapping from fake images to real images and prioritizes the perceptual similarity between them.

The fourth one is Frechet Inception Distance (FID), which is one of the most popular metrics used to measure the feature distance between real and generated images [6]. Mathematically, Frechet Distance is used to calculate the distance between two "multivariate" normal distributions. In computer vision, especially in GAN evaluation, we use the Inception V3 model pre-trained on the Imagenet dataset. The specific algorithm details are shown in Formula 11:

$$FID = ||\mu_r - \mu_g||^2 + Tr(\sum_r + \sum_g -2(\sum_r \sum_g)^{1/2}) \tag{12}$$

Where $\mu_g$ and $\mu_r$are the means of the feature maps of the generated and real images, respectively. Tr represents the trace of the linear algebraic operation, and $\sum_g$ and $\sum_r$ represent the covariance matrices of the generated and real images, respectively.

### 3.3 Analysis

Our study focused on cross-modal generation between MR and CT images across three distinct data phases. We evaluated the performance of our trained models using four comprehensive metrics. The following analysis presents both quantitative and qualitative assessments of our results, providing a thorough examination of the model's effectiveness in generating high-quality cross-modal medical images. Our method performed MR-to-CT and CT-to-MR image translation tasks across three datasets: ART (Arterial), PV (Portal Venous), and NC (Non-Contrast). The results, evaluated using four metrics (TPSNR, TSSIM, LPIPS, and FID), are presented in Table 1. The PV dataset consistently demonstrated superior performance, while the NC dataset showed the lowest scores across most

**Table 1:** Quantitative evaluation of cross-modal medical image generation performance across three different data types (ART, PV, and NC) for MR to CT and CT to MR conversions. The performance is measured using four metrics: TPSNR (Total Peak Signal-to-Noise Ratio), TSSIM (Total Structural Similarity Index), LPIPS (Learned Perceptual Image Patch Similarity), and FID (Fréchet Inception Distance).

| Data type | | TPSNR | TSSIM | LPIPS | FID |
|---|---|---|---|---|---|
| ART | MR →CT | 33.76 | 0.85 | 0.395 | 21.75 |
| | CT→ MR | 34.64 | 0.88 | 0.410 | 20.98 |
| PV | MR →CT | 35.51 | 0.89 | 0.417 | 20.39 |
| | CT→ MR | 37.19 | 0.91 | 0.431 | 19.30 |
| NC | MR →CT | 30.91 | 0.84 | 0.366 | 24.59 |
| | CT→ MR | 37.20 | 0.85 | 0.387 | 23.67 |

metrics. This variation can be attributed to the distinct image features characteristic of each phase. The model appears to perform better when there is a greater contrast in grayscale values between the source and target images. Notably, our approach achieved high-quality results for both MR-to-CT and CT-to-MR translations. The CT-to-MR translation in the PV dataset yielded the best overall performance, with the highest TPSNR (37.19), TSSIM (0.91), LPIPS (0.431), and the lowest FID (19.30). These scores represent state-of-the-art performance in cross-modal medical image translation using unpaired datasets. Interestingly, CT-to-MR translations generally outperformed MR-to-CT translations across all datasets, as evidenced by higher TPSNR and TSSIM values. This suggests that our model may be more adept at generating MR images from CT scans than vice versa. The ART dataset showed balanced performance between MR-to-CT and CT-to-MR translations, while the NC dataset exhibited the largest performance gap between the two translation directions. These results demonstrate the effectiveness of our approach in handling various types of medical imaging data and its potential for improving cross-modal image generation in clinical applications.

Figures 3 and 4 present the qualitative results of our cross-modal image generation for MR-to-CT and CT-to-MR conversions, respectively. The images are organized by the three phases: ART (Arterial), PV (Portal Venous), and NC (Non-Contrast). To highlight the model's performance in generating tumor regions within the liver, we have marked areas of interest with red boxes (for MR-to-CT) and blue boxes (for CT-to-MR). These regions are shown enlarged alongside the original images for detailed comparison. Upon close examination, the generated images demonstrate remarkable fidelity to the ground truth in terms of both structural accuracy and spatial integrity. This is particularly evident in the enlarged tumor regions, where our model has successfully preserved fine details and subtle lesions without introducing blurriness or losing critical features. In the MR-to-CT results (Figure 3), the fake CT images closely mimic the appearance and contrast of real CT scans across all three phases. The model accurately captures the higher contrast of CT imaging, especially in bone and soft tissue differentiation. Similarly, in the CT-to-MR results (Figure 4), the generated MR images effectively replicate the characteristic soft tissue contrast

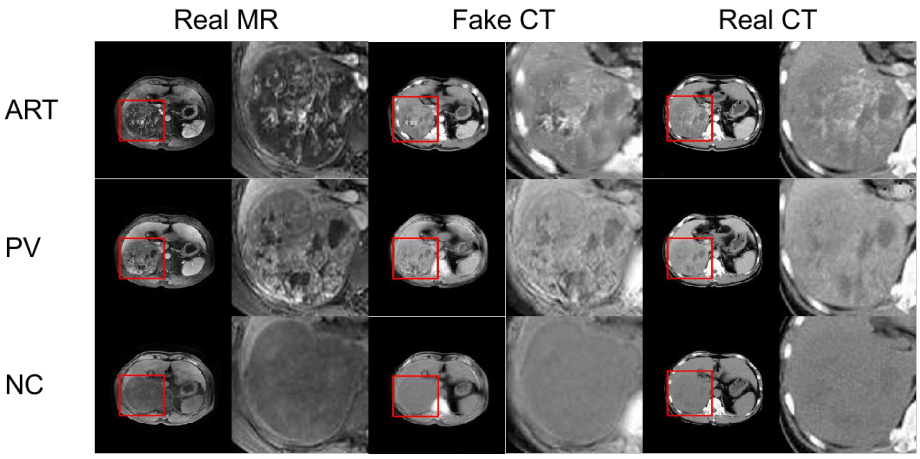


**Fig. 3:** Qualitative comparison of CT images generated from MR images across three different scan types: ART (Arterial), PV (Portal Venous), and NC (Non-Contrast). The figure shows real MR images (left column), generated "fake" CT images (middle column), and corresponding real CT images (right column) for each scan type. Red boxes highlight regions of interest to compare the quality and accuracy of the generated CT images against the real CT scans.

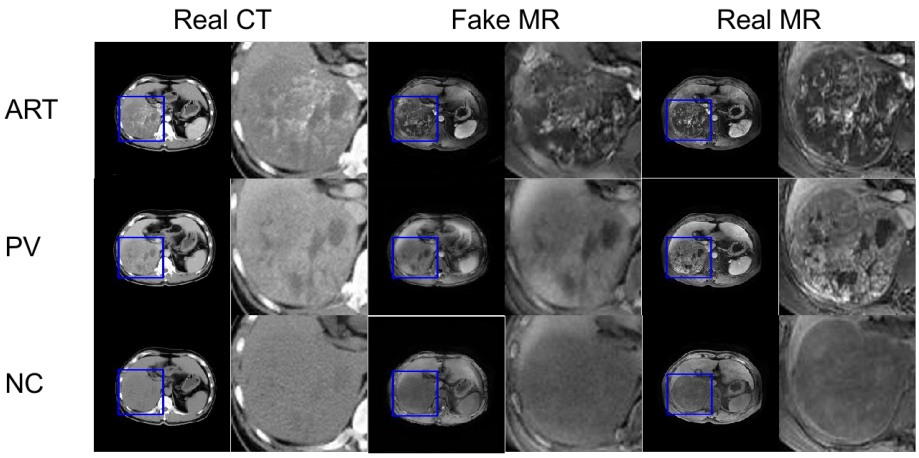


**Fig. 4:** Qualitative results of MR images generated from CT scans across three different phases: ART (Arterial), PV (Portal Venous), and NC (Non-Contrast). The figure displays real CT images (left column), synthetically generated "fake" MR images (middle column), and corresponding real MR images (right column) for each phase. Blue boxes highlight regions of interest to facilitate comparison between the generated MR images and the actual MR scans, demonstrating the effectiveness of the cross-modal generation technique.

and detail of real MR scans. The fake MR images maintain the complex textural patterns typical of MRI, particularly noticeable in the liver parenchyma and surrounding tissues. Notably, our model performs consistently well across all three phases (ART, PV, NC), adapting to the specific imaging characteristics of each. This demonstrates the robustness of our approach in handling various contrast phases commonly encountered in clinical imaging. These qualitative results corroborate our quantitative findings, showcasing the model's capability to generate high-quality, clinically relevant cross-modal medical images while preserving critical diagnostic information.

## 4 Conclusion

In this paper, we introduce a novel approach to medical image generation using a progressive generative adversarial network with multiple discriminators at different scales. Our method aims to provide deterministic guidance, focusing particularly on liver tumor regions, to generate higher quality images while overcoming common GAN training challenges. Through comprehensive qualitative and quantitative analyses, we have demonstrated the model's success in improving generation quality for target regions. The results show significant enhancements in structural accuracy, detail preservation, and overall image fidelity across various imaging modalities and contrast phases. Looking ahead, we plan to apply our method to state-of-the-art approaches to showcase its broader applicability, and we will refine the network structure of our multi-scale discriminators to better suit the specific needs of medical image generation. These future directions aim to further advance the field, ultimately contributing to improved diagnostic tools and patient care in clinical practice.

## References

1. Chen, J., Wei, J., Li, R.: Targan: target-aware generative adversarial networks for multi-modality medical image translation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 24–33. Springer (2021)
2. Ernst, P., Hille, G., Hansen, C., Tönnies, K., Rak, M.: A cnn-based framework for statistical assessment of spinal shape and curvature in whole-body mri images of large populations. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. pp. 3–11. Springer (2019)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
4. Han, X.: Mr-based synthetic ct generation using a deep convolutional neural network method. Medical physics **44**(4), 1408–1419 (2017)
5. Hardy, C., Le Merrer, E., Sericola, B.: Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. In: 2019 IEEE international parallel and distributed processing symposium (IPDPS). pp. 866–877. IEEE (2019)

6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
8. Jin, C.B., Kim, H., Liu, M., Jung, W., Joo, S., Park, E., Ahn, Y.S., Han, I.H., Lee, J.I., Cui, X.: Deep ct to mr synthesis using paired and unpaired data. Sensors **19**(10), 2361 (2019)
9. Kang, J., Kim, S., Lee, K.M.: Multi-modal/multi-scale convolutional neural network based in-loop filter design for next generation video codec. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 26–30. IEEE (2017)
10. Liu, B., Tang, R., Chen, Y., Yu, J., Guo, H., Zhang, Y.: Feature generation by convolutional neural network for click-through rate prediction. In: The World Wide Web Conference. pp. 1119–1129 (2019)
11. Liu, Z., Dou, Y., Jiang, J., Xu, J.: Automatic code generation of convolutional neural networks in fpga implementation. In: 2016 International conference on field-programmable technology (FPT). pp. 61–68. IEEE (2016)
12. Manjón, J.V.: Mri preprocessing. Imaging Biomarkers: Development and Clinical Integration pp. 53–63 (2017)
13. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)