





Article

DC²Anet: Generating Lumbar Spine MR Images from CT Scan Data Based on Semi-Supervised Learning

Cheng-Bin Jin ¹, Hakil Kim ^{1,*}, Mingjie Liu ¹, In Ho Han ², Jae Il Lee ², Jung Hwan Lee ², Seongsu Joo ³, Eunsik Park ³, Young Saem Ahn ⁴ and Xuenan Cui ¹

¹ School of Information and Communication Engineering, INHA University, Incheon 22212, Korea; chengbinjin@inha.edu (C.-B.J.); liumj@inha.edu (M.L.); xncui@inha.ac.kr (X.C.)

² Department of Neurosurgery, Medical Research Institute, Pusan National University Hospital, Pusan 49241, Korea; farlateral@hanmail.net (I.H.H.); medifirst@pusan.ac.kr (J.I.L.); medi98@hanmail.net (J.H.L.)

³ Team Elysium Inc., Seoul 93525, Korea; seongsu.joo@teamelysium.kr (S.J.); espark@teamelysium.kr (E.P.)

⁴ Department of Computer Engineering, INHA University, Incheon 22212, Korea; ahnsaem90@gmail.com

* Correspondence: hikim@inha.ac.kr; Tel.: +82-032-860-7385

Received: 14 May 2019; Accepted: 18 June 2019; Published: 20 June 2019



Abstract: Magnetic resonance imaging (MRI) plays a significant role in the diagnosis of lumbar disc disease. However, the use of MRI is limited because of its high cost and significant operating and processing time. More importantly, MRI is contraindicated for some patients with claustrophobia or cardiac pacemakers due to the possibility of injury. In contrast, computed tomography (CT) scans are much less expensive, are faster, and do not face the same limitations. In this paper, we propose a method for estimating lumbar spine MR images based on CT images using a novel objective function and a dual cycle-consistent adversarial network (DC²Anet) with semi-supervised learning. The objective function includes six independent loss terms to balance quantitative and qualitative losses, enabling the generation of a realistic and accurate synthetic MR image. DC²Anet is also capable of semi-supervised learning, and the network is general enough for supervised or unsupervised setups. Experimental results prove that the method is accurate, being able to construct MR images that closely approximate reference MR images, while also outperforming four other state-of-the-art methods.

Keywords: image cross-modality synthesis; lumbar spine; dual cycle-consistent adversarial network; semi-supervised learning; adversarial training

1. Introduction

Computed tomography (CT) scanning is a medical imaging technique that is widely used for diagnostic and therapeutic purposes in a variety of clinical applications. Magnetic resonance imaging (MRI) is another imaging technique that visualizes anatomical details and is used in radiology and nuclear medicine. A comparison of the strengths and weaknesses of these imaging approaches is shown in Table 1. Unlike CT scans, MRI can detect slight differences in soft tissue, ligaments, and organs, which is beneficial for diagnosis. However, MRI is not only much more expensive, but also requires more time to produce its results, meaning that patients often prefer CT scans to MRI.

Lumbar disc herniation is common among the elderly and people who sit for long periods. The use of MRI to observe the spinal cord and the disc signals of the lumbar spine is of great importance in the treatment of this condition. However, some patients with claustrophobia or cardiac pacemakers are prevented from receiving an MRI due to possible injury. Thus, the ability to generate a reliable magnetic resonance (MR) image from a CT scan for these patients is vital. This would not only

increase the diagnostic value of CT scans, but also provide additional reference information for diagnosis. Thus, in this study, we propose a synthesis method based on convolutional neural networks (CNNs) [1,2] with adversarial training [3] to construct a lumbar spine MR image from CT scan data.

Table 1. Comparison between CT scans and MRI.

	CT Scans	MRI
Principle	Uses multiple X-rays, taken at different angles to produce cross-sectional images	Uses powerful magnetic fields and radiofrequency pulses to produce detailed images
Radiation	Minimal	None
Uses	Excellent for observing bone and very good for soft tissue, especially with the use of intravenous contrast dye	Excellent for detecting very slight differences in soft tissue
Cost	Usually less expensive than MRI	Often more expensive than CT scans
Time taken	Very quick, taking only about 5 min, depending on the area being scanned	Depends on the part of the body being examined and can range from 15 min–2 h
Application	Produces a general image of an area such as internal organs, fractures, or head trauma	Produces more detailed images of soft tissue, ligaments, and organs
Benefits	Faster and can provide images of tissue, organs, and skeletal structure	Produces more detailed images
Risks	<ul style="list-style-type: none"> • Harmful for unborn babies • A very small dose of radiation • A potential reaction to the use of dyes 	<ul style="list-style-type: none"> • Possible reactions to metals due to magnets (e.g., artificial joints, eye implants, intrauterine devices, pacemakers) • Loud noises from the machine can cause hearing issues • Increase in body temperature during long MRIs • Claustrophobia

In recent years, researchers have increasingly searched for ways to replace CT scans with MRI when planning for radiation therapy [4–6]. However, CT-based MR image construction has received little attention. It is challenging to generate an MR image directly from a CT image using a linear model because it is difficult to produce high-level image domains based on low-level ones. In response to this, we propose a synthesis method based on convolutional neural networks (CNNs) [1] with adversarial training [3] to produce a lumbar spine MR image from a CT scan. In this process, the development of an objective function for the deep neural network is essential [7,8]. An objective function is a combination of loss terms that maps a real number that intuitively represents the “cost” associated with the performance of a predefined network at a certain status. The optimization process seeks to minimize this cost by updating trainable variables to determine an optimal network. Synthetic images in medical imaging need to not only be realistic, i.e., they cannot be distinguished from genuine images by human experts, but also be very similar to reference images. In this study, we propose a novel objective function that balances between two quantitative loss terms and three qualitative loss terms to construct lumbar spine MR images from CT images. A dual cycle-consistent loss is also included for semi-supervised learning that alternates between optimizing supervised and unsupervised learning in order to seek a global minimum for the optimal network.

Experimental results based on quantitative and qualitative evaluations prove the superiority of the proposed method compared with other state-of-the-art methods. The main contributions of this study are as follows:

- An objective function is proposed to balance quantitative and qualitative loss terms to construct a realistic and accurate synthetic MR image. This function consists of adversarial, dual cycle-consistent, voxel-wise, gradient difference, perceptual, and structural similarity losses. Using ablation analysis, the importance and effectiveness of each of these loss terms are investigated.

- The dual cycle-consistent adversarial network (DC²Anet) is proposed as a general synthesis system for semi-supervised learning. Due to its dual cycle-consistent structure, DC²Anet can be applied to both supervised and unsupervised learning.

This paper first summarizes previous research on the synthesis of medical images in Section 2. The proposed algorithm is outlined in Section 3, while Section 4 reports the experimental results and discussion. A conclusion is presented in Section 5.

2. Literature Review of Medical Imaging Synthesis

In medical imaging, a number of methods have been proposed for generating one image domain from another, e.g., constructing a CT image from MRI data or a positron emission tomography (PET) image from CT data. Existing methods can be divided into three categories: tissue-segmentation, learning, and atlas-based methods. *Tissue segmentation* first divides MR image voxels into different tissue classes, such as air, fat, soft tissue, and bone, and then, the segmentation classes are refined manually [5,9]. However, tissue segmentation is difficult, and its performance strongly depends on segmentation accuracy and the quality of the manual input. *Learning-based methods* extract features that represent two different domains and then construct a non-linear map between them. However, these methods depend on the quality of the feature extraction in terms of how well they can represent the different domains. Additionally, generating one image domain from another is not as simple as one-to-one mapping [10,11]. *Atlas-based methods* apply image registration to align an MR image with an atlas MR image to approximate the correspondence matrix. The matrix can then be used to warp the associated atlas CT image to generate the query CT image [12–14]. However, the performance of atlas-based methods is closely associated with the registration accuracy for the two image domains. Furthermore, it is difficult to cover pathological differences or significant anatomical variations using atlas data.

In recent years, CNNs [15,16] have demonstrated outstanding performance in various computer vision tasks. In particular, several studies have proven that CNNs are useful in medical imaging [17], such as skin cancer classification [18], X-ray organ segmentation [19], retinal vessel segmentation [20], and brain lesion detection [21]. In these applications, CNN-based medical image synthesis can be considered a form of regression in which non-linear mapping functions are stacked from one image domain to another. For example, Han [22] applied a U-Net [23] architecture consisting of an encoder network and a decoder network in which some layers were connected by skip connections to construct a synthetic CT image from an MR image. To train the deep CNN model using a limited dataset, Han [22] employed transfer learning by initializing the encoder network using a pretrained 16-layer VGG (VGG16) network [24]. The objective function of the network used voxel-wise loss only to minimize the difference between the synthetic and reference images. However, because voxel-wise loss is minimized by averaging all plausible outputs, simply minimizing this loss may produce blurry results. Additionally, the slight voxel-wise misalignment of training data may further lead to a blurry constructed image. Designing objective functions that force the CNN to operate as required, e.g., to generate sharp, realistic, and accurate synthetic images, remains an unsolved problem and generally requires both prior knowledge and experimental observations.

In image generation, generative adversarial networks (GANs) [25], which are a form of generative model [26,27], have been widely employed to produce state-of-the-art, realistic images [28,29] for applications such as GAN-based image inpainting [30] and video generation [31,32]. An adversarial loss of GAN learns to satisfy a high-level goal, such as generating an output image indistinguishable from reality. A discriminator network of GAN is used to distinguish whether an image is real or synthesized while simultaneously training a generator network to minimize the adversarial loss. For example, Bi et al. [33] presented a multi-channel GAN with an objective function consisting of adversarial loss and voxel-wise loss to generate a synthetic PET image from a CT image. Similarly, Ben-Cohen et al. [34] independently applied the advantages of a fully-convolutional network (FCN) [35] and a pixel-to-pixel (pix2pix) model [36] to synthesize a realistic PET image from a CT image.

This method also used adversarial loss and voxel-wise loss together. Extending the above method, Nie et al. [37] proposed a context-aware GAN that utilized a 3D CNN [38,39] and an auto-context model [40] to generate a CT sequence from an MR sequence. The voxel-wise loss of the 3D CNN learns both the spatial and temporal information of the sequence data. In contrast, Wolterink et al. [41] applied a cycle-consistent GAN (cycleGAN) [42] with least-squares adversarial loss [43] in which the loss term leads to the stable optimization of the network when synthesizing an MR image using a CT image. A cycle-consistent loss [42,44] produces not only a synthetic image that looks real, but also one that is similar to the input under unsupervised learning. A CT-based MR image estimation method was first proposed by Jin et al. [45]. They proposed a synthesis system referred to as MR-GAN using a dual cycle-consistent structure. Their MR-GAN is trainable with paired and unpaired data together to improve performance. In addition to dual cycle-consistent loss, their objective function includes two other loss terms: adversarial loss and the voxel-wise loss. Table 2 presents a comparison of the network architectures and objective functions of the deep-neural-network-based medical synthesis methods. Compared to the methods in Table 2, this study proposes a more general system of the cross-modality synthesis, DC²Anet, that supports both supervised and unsupervised learning and a new objective function to balance quantitative and qualitative loss terms.

Table 2. Comparison of synthesis methods based on deep neural networks. pix2pix, pixel-to-pixel.

	DCNN [22]	Multi-channel GAN [33]	Context-Aware GAN [37]	Deep MR-to-CT [41]	DiscoGAN [44]	MR-GAN [45]
Application	Brain MR to CT	Lung CT to PET	Brain and pelvic MRI to CT	Brain MR to CT	Attribute translation	Brain CT to MRI
Objective function	Voxel-wise	Adversarial [25] and voxel-wise	Adversarial [25], voxel-wise, and gradient difference [37]	Least- squares adversarial [43] and cycle- consistent [42]	Adversarial [25] and cycle- consistent [42]	Adversarial [25], voxel- wise, and dual cycle- consistent [45]
Model	Pretrained VGG16 [24] with U-Net [23]	pix2pix [36]	3D ConvNet [38,39] and auto-context model [40]	cycleGAN [42]	DiscoGAN [44]	MR-GAN [45]
No. of trainable parameters	34.9 M	54.5 M	Unknown	28.3 M	16.6 M	28.3 M
Generator	U-Net [23]	U-Net [23]	Customized	Residual Net [46–48]	Customized	Residual Net [46–48]
No. of layers in the generator	27	16	8	24	8	24
Discriminator	None	Patch GAN [36]	Customized	Path GAN [36]	Path GAN [36]	Path GAN [36]
No. of layers in the discriminator	None	5	6	5	5	5
Generation time	56.25 ms.	17.51 ms.	Unknown	48.71 ms.	10.88 ms.	47.48 ms.

The voxel-wise loss function measures the difference between the synthetic and the reference images, but it cannot reflect the perceptual difference between the two images. For example, even when two identical images have the same perceptual information, they will have very different voxel-wise loss measurements if they are offset from each other by just one pixel. Recent work has shown that high-quality synthetic images can be produced using perceptual loss based on differences between high-level feature representations extracted from a pretrained CNN [48,49]. Gatys et al. [49] conducted artistic style transfer by jointly minimizing feature reconstruction loss [50] and style reconstruction loss based on features extracted from a pretrained VGG16 network [24]. Johnson et al. [48] produced visually-pleasing results using image style transformation and single-image super-resolution, with voxel-wise loss replaced by perceptual loss. Structural similarity (SSIM) [51,52] is another qualitative measurement approach that is based on the human visual system and is used to compare local patterns of structural information that have been normalized for luminance and contrast. In our study, a similar structural loss term was proposed to retain the structural patterns of lumbar spine CT scans in the synthesis of MR images. Additionally, to balance quantitative and qualitative performance, gradient difference loss and perceptual loss were included based on adversarial, voxel-wise, and dual cycle-consistent loss.

3. Method

3.1. Converting Supervised Learning to Semi-Supervised Learning

Semi-supervised learning is a form of machine learning that makes use of a small amount of aligned data and a large volume of unaligned data. It thus represents a combination of supervised learning (which utilizes completely aligned data) and unsupervised learning (which does not include aligned data). In our study, the paired CT and MR images were aligned, meaning that the CT image and its corresponding MR image were from the same slice of the same patient, with some post-processing such as image registration for coordinate offsetting and manual correction by neuroradiologists. For image registration, we utilized the contours of the body vertebra in CT and MR images to estimate the parameters of the affine transformation to register the two images. In contrast, unaligned data included CT and MR images that were captured from different slices or even different patients.

In medical image synthesis, supervised learning can easily be converted to unsupervised learning. Supervised learning applies aligned training data where the output image corresponds to each input image. On the contrary, by disconnecting the aligned data to consist of an input and output set for training, medical synthesis becomes an unsupervised learning-based synthetic task. A semi-supervised learning framework can also be constructed to utilize both supervised and unsupervised learning together. Figure 1 illustrates the conversion from supervised learning to semi-supervised learning. The left-hand side of the figure displays supervised learning using aligned data. The squares and circles represent the image domains X and Y, respectively. The three aligned points are indicated by different colors (red, green, and blue) with parentheses. By disconnecting the aligned data and recombining the different domains, unaligned data are generated, as shown on the right-hand side of Figure 1. In this manner, the three aligned data points can be converted into six unaligned data points, with the unpaired data increasing exponentially. Semi-supervised learning is thus conducted by combining supervised learning with aligned data and unsupervised learning with unaligned data. The advantage of this approach is that supervised learning uses the averages of all plausible outputs to reduce the bias of the domain translation, while unsupervised learning focuses on the structural pattern of the two image domains, reducing the variance of the model estimation process. Additionally, semi-supervised learning can more efficiently use a limited volume of paired data by combining the three paired data points with the six unpaired data points shown in Figure 1. The proposed DC²Anet is capable of semi-supervised learning, and the network is general enough for both supervised learning and unsupervised learning.

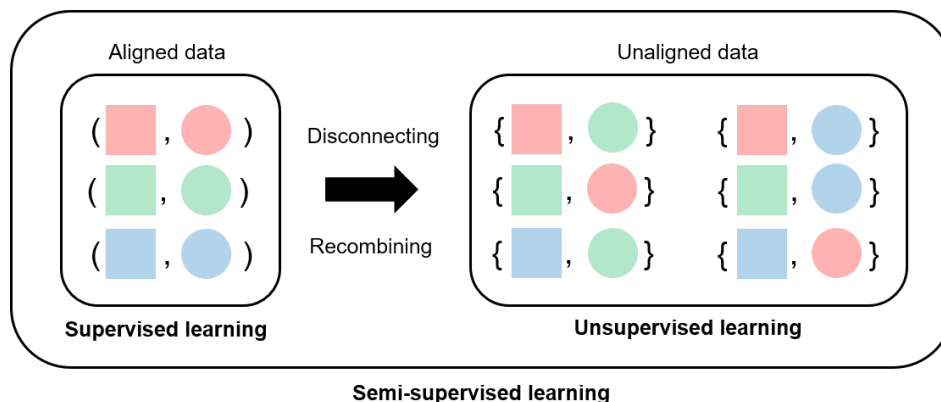


Figure 1. Illustration of the conversion of supervised learning to semi-supervised learning.

3.2. Dual Cycle-Consistent Adversarial Network

A GAN [25] is a generative model that is designed to generate synthetic samples directly from the desired data distribution without the need to model the underlying probability density function explicitly. It consists of two different networks that are trained simultaneously, with the generator network focused on image generation and the discriminator network used to distinguish between real samples and the synthetic images. The idea of using a cycle-consistent approach to regularize structural data has a long history in visual tracking [53] and structure from motion [54]. The cycle-consistent structure employed in the GAN (cycleGAN) [44] enables unsupervised learning, stitching two generator networks together head to toe so that the synthetic images can be translated into a forward cycle. In addition to the forward cycle, the cycleGAN also has a backward cycle to stabilize the training process and prevent mode collapse. The forward cycle enforces the translation from the CT domain to the MR domain, while the backward cycle moves from the MR domain to the CT domain.

The proposed DC²Anet also applies a cycle-consistent structure for its unsupervised learning setup. However, the proposed network has a dual cycle-consistent structure for the adoption of semi-supervised learning: one cycle-consistent structure for supervised learning with aligned data and the other for unsupervised learning with unaligned data. A diagram of DC²Anet is presented in Figure 2. Because the forward and backward cycle-consistent networks with aligned data or unaligned data are similar, we only illustrate a forward cycle-consistent adversarial network with unaligned learning in Figure 2a and a backward cycle-consistent with aligned learning in Figure 2b.

In the forward cycle-consistent adversarial network with unaligned learning, the Syn_{MR} network generates a synthetic MR image from a CT image, and this MR image is then used by the Syn_{CT} network to generate the original CT image in order to learn the domain structures. The input to the MR discriminator network is either a sample MR image from the real MR data or a synthetic MR image. The objective function for unaligned learning includes both cycle-consistent and adversarial loss. In the backward cycle-consistent adversarial network with aligned learning, a synthetic CT image is generated from an MR image, and this CT image is employed by the Syn_{MR} network to generate the original MR image. The CT discriminator is used to distinguish between the synthetic CT and reference CT images. In aligned learning, a reference image is matched with the synthetic image to restrain the generated structure of the output. Based on the cycle-consistent and adversarial loss, the objective function of aligned learning also considers, due to the use of the reference image, voxel-wise, gradient difference, structural, and perceptual loss within the pretrained VGG16 network [24]. The four switches are simultaneously employed to control the data flow from the reference image, and these are connected in aligned learning, but disconnected in unaligned learning. It is also important to note that the Syn_{MR} and Syn_{CT} networks utilized in the forward and backward cycles share the same weights.

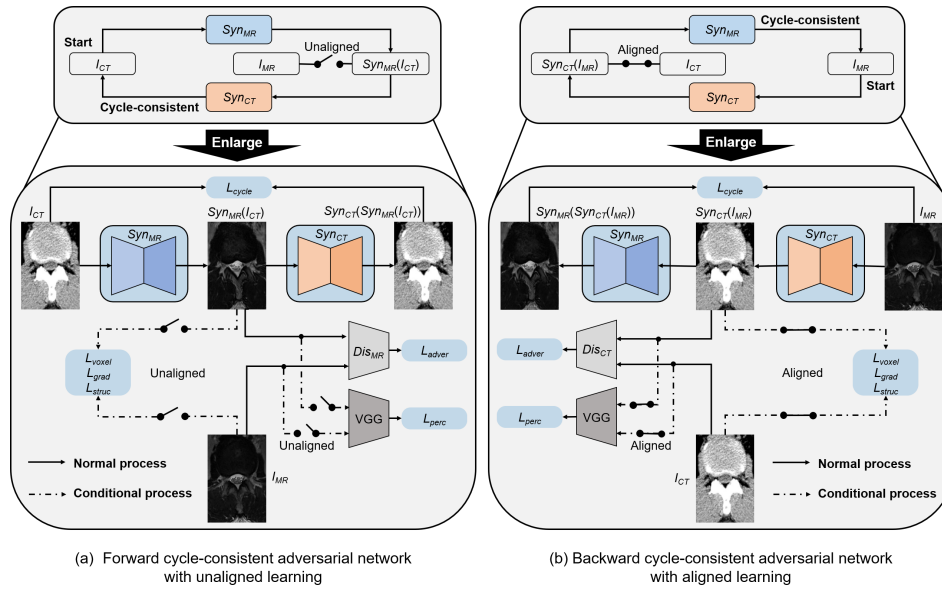


Figure 2. Diagram of the proposed dual cycle-consistent adversarial network (DC²Anet). DC²Anet consists of a forward cycle-consistent and a backward cycle-consistent network. (a) The forward cycle-consistent adversarial network with unaligned learning. (b) The backward cycle-consistent adversarial network with aligned learning.

3.3. Objective Function

Our goal is for the mapping functions between the CT image and MR image domains to be learned using the given aligned training data. As illustrated in Figure 1, aligned data (I_{CT} , I_{MR}) are converted into unaligned data I_{CT} , I_{MR} , and the aligned and unaligned data points are utilized together in semi-supervised learning. DC²Anet includes two synthesis networks, Syn_{MR} : CT \rightarrow MR and Syn_{CT} : MR \rightarrow CT, and includes two discriminator networks, Dis_{MR} and Dis_{CT} , where Dis_{MR} aims to distinguish between the reference MR image I_{MR} and the synthetic MR image $Syn_{MR}(I_{CT})$; in the same way, Dis_{CT} aims to distinguish between the reference CT image I_{CT} and the synthetic CT image $Syn_{CT}(I_{MR})$. Moreover, the four networks are different optimized objectives corresponding to the supervised and unsupervised learning due to the input aligned or unaligned data. Additionally, to measure the high-level perceptual and semantic differences between two images, the VGG16 perceptual network is employed in DC²Anet, which was pretrained on the ImageNet dataset [55]. Our objective function contains six loss terms in total: adversarial, dual cycle-consistent, voxel-wise, gradient difference, perceptual, and structural similarity. A summary of the strengths and weaknesses of each loss term is given in Table 3.

We apply adversarial loss [3] to both the supervised and unsupervised setups. The forward and backward mappings Syn_{MR} : CT \rightarrow MR and Syn_{CT} : MR \rightarrow CT and the discriminators Dis_{MR} and Dis_{CT} are expressed as follows:

$$\begin{aligned}
 L_{sup-adver}(Syn_{MR}, Dis_{MR}, Syn_{CT}, Dis_{CT}) = & \mathbb{E}_{I_{CT}, I_{MR} \sim p_{data}(I_{CT}, I_{MR})} \left[\log(Dis_{MR}(I_{CT}, I_{MR})) \right] \\
 & + \mathbb{E}_{I_{CT} \sim p_{data}(I_{CT})} \left[\log(1 - Dis_{MR}(I_{CT}, Syn_{MR}(I_{CT}))) \right] \\
 & + \mathbb{E}_{I_{MR}, I_{CT} \sim p_{data}(I_{MR}, I_{CT})} \left[\log(Dis_{CT}(I_{MR}, I_{CT})) \right] \\
 & + \mathbb{E}_{I_{MR} \sim p_{data}(I_{MR})} \left[\log(1 - Dis_{CT}(I_{MR}, Syn_{CT}(I_{MR}))) \right]
 \end{aligned} \quad (1)$$

where the first two terms are the forward adversarial loss and the last two terms are the backward adversarial loss. The network Syn_{MR} attempts to synthesize images $Syn_{MR}(I_{CT})$, which look similar to images from the MR domain, while Dis_{MR} aims not only to discriminate

between synthetic MR and reference MR images, but also to ensure it generates images from the corresponding CT images I_{CT} . For the backward adversarial loss, the synthesis network Syn_{CT} generates the reference CT images $Syn_{CT}(I_{MR})$, which look similar to images from the CT domain, while Dis_{CT} aims to distinguish between synthetic and reference CT images based on the MR images I_{MR} . The synthesis networks Syn_{MR} and Syn_{CT} attempt to minimize this objective function, while the adversarial discriminator networks Dis_{MR} and Dis_{CT} aim to maximize it, i.e., $Syn_{MR}^*, Syn_{CT}^* = \arg \min_{Syn_{MR}, Syn_{CT}} \max_{Dis_{MR}, Dis_{CT}} L_{sup-adver}(Syn_{MR}, Dis_{MR}, Syn_{CT}, Dis_{CT})$. In Equation (2), we introduce a similar form of adversarial loss for unsupervised learning for the synthesis networks and discriminators. However, the discriminators for unsupervised learning need to distinguish whether the images are real or synthetic; the source domain images are not input into the discriminators.

$$\begin{aligned}
 L_{unsup-adver}(Syn_{MR}, Dis_{MR}, Syn_{CT}, Dis_{CT}) = & \mathbb{E}_{I_{MR} \sim p_{data}}(I_{MR}) \left[\log(Dis_{MR}(I_{MR})) \right] \\
 & + \mathbb{E}_{I_{CT} \sim p_{data}}(I_{CT}) \left[\log(1 - Dis_{MR}(Syn_{MR}(I_{CT}))) \right] \\
 & + \mathbb{E}_{I_{CT} \sim p_{data}}(I_{CT}) \left[\log(Dis_{CT}(I_{CT})) \right] \\
 & + \mathbb{E}_{I_{MR} \sim p_{data}}(I_{MR}) \left[\log(1 - Dis_{CT}(Syn_{CT}(I_{MR}))) \right]
 \end{aligned} \quad (2)$$

Table 3. A summary of the strengths and weaknesses of each loss term used in DC²Anet. In the last column, the symbols ✓ and ✗ denote whether the loss term requires aligned training data or not.

Loss term	Strengths	Weaknesses	Aligned Data
Adversarial	Realistic output	Unstable training	✗
Dual cycle-consistent	Possible unsupervised learning	Heavy computational load (two synthesis and two discriminator networks)	✗
Voxel-wise	Encourage the output similar to the reference image	Tends to produce blurry output	✓
Gradient difference	Emphasizes the boundaries of the output	Sensitive to the quality of data alignment	✓
Perceptual	Preserves high-level semantic similarity	Not a fully-analyzed and task-oriented problem	✓
Structural similarity	Relaxes misalignment constraints for data alignment	Prefers low illumination	✓

In unsupervised learning, adversarial loss alone cannot guarantee that the learned synthesis network can map an input image to the desired output image. To reduce the possible mapping space between these two domains, we utilized a dual cycle-consistent structure for aligned and unaligned data. For an image I_{CT} from the CT domain, the forward cycle-consistent network should be able to bring I_{CT} back to the original image, i.e., $I_{CT} \rightarrow Syn_{MR}(I_{CT}) \rightarrow Syn_{CT}(Syn_{MR}(I_{CT})) \approx I_{CT}$. Similarly, the backward cycle-consistent network should extract an image I_{MR} from the MR domain to satisfy $I_{MR} \rightarrow Syn_{CT}(I_{MR}) \rightarrow Syn_{MR}(Syn_{CT}(I_{MR})) \approx I_{MR}$. The cycle-consistent losses are expressed as follows:

$$\begin{aligned}
 L_{sup-cycle}(Syn_{MR}, Syn_{CT}) = & \mathbb{E}_{I_{CT}, I_{MR} \sim p_{data}}(I_{CT}, I_{MR}) \left[\left\| Syn_{CT}(Syn_{MR}(I_{CT})) - I_{CT} \right\|_1 \right] \\
 & + \mathbb{E}_{I_{MR}, I_{CT} \sim p_{data}}(I_{MR}, I_{CT}) \left[\left\| Syn_{MR}(Syn_{CT}(I_{MR})) - I_{MR} \right\|_1 \right]
 \end{aligned} \quad (3)$$

$$\begin{aligned}
 L_{unsup-cycle}(Syn_{MR}, Syn_{CT}) = & \mathbb{E}_{I_{CT} \sim p_{data}}(I_{CT}) \left[\left\| Syn_{CT}(Syn_{MR}(I_{CT})) - I_{CT} \right\|_1 \right] \\
 & + \mathbb{E}_{I_{MR} \sim p_{data}}(I_{MR}) \left[\left\| Syn_{MR}(Syn_{CT}(I_{MR})) - I_{MR} \right\|_1 \right]
 \end{aligned} \quad (4)$$

where $L_{sup-cycle}$ and $L_{unsup-cycle}$ are the cycle-consistent structures for supervised and unsupervised learning, respectively. Each cycle-consistent loss has two terms: a forward cycle-consistent and a backward cycle-consistent term.

In general, adversarial loss produces visually appealing results. However, using only adversarial loss to match synthetic and reference MR images may cause the model to generate unseen structures. Voxel-wise loss helps to overcome this problem if aligned data are available. The goal of the discriminator networks remains unchanged, but the synthesis networks are tasked with not only cheating the discriminator networks, but also being similar to the reference image at an L1 distance. The voxel-wise loss of the forward and backward cycle-consistent network is defined as follows:

$$L_{voxel}(Syn_{MR}, Syn_{CT}) = \mathbb{E}_{I_{CT}, I_{MR} \sim p_{data}(I_{CT}, I_{MR})} [\|I_{MR} - Syn_{MR}(I_{CT})\|_1] + \mathbb{E}_{I_{MR}, I_{CT} \sim p_{data}(I_{MR}, I_{CT})} [\|I_{CT} - Syn_{CT}(I_{MR})\|_1] \quad (5)$$

Direct optimization of voxel-wise loss produces a suboptimal (i.e., blurry) result by minimizing the average loss for all plausible outputs. To deal with the inherently blurry results obtained from voxel-wise loss, gradient difference loss is constrained for the synthesis networks. The gradient difference loss between a synthetic and reference image is given as follows:

$$L_{grad}(Syn_{MR}, Syn_{CT}) = \mathbb{E}_{I_{CT}, I_{MR} \sim p_{data}(I_{CT}, I_{MR})} [\|\nabla(Syn_{MR}(I_{CT}))_x - \nabla(I_{MR})_x\|_1 + \|\nabla(Syn_{MR}(I_{CT}))_y - \nabla(I_{MR})_y\|_1] + \mathbb{E}_{I_{MR}, I_{CT} \sim p_{data}(I_{MR}, I_{CT})} [\|\nabla(Syn_{CT}(I_{MR}))_x - \nabla(I_{CT})_x\|_1 + \|\nabla(Syn_{CT}(I_{MR}))_y - \nabla(I_{CT})_y\|_1] \quad (6)$$

where I_{MR} in the first term and I_{CT} in the second term are the reference images in the forward and backward cycle-consistent networks, respectively, and the x - and y -direction gradients are calculated to emphasize the boundaries of the structural shape.

A pretrained VGG16 network is incorporated into the optimization of the synthesis networks to ensure perceptual similarity. We aim for the synthetic and reference images to have similar feature representations when computed by the pretrained VGG16 network ϕ . Let $\phi_j(I_{CT})$ and $\phi_j(I_{MR})$ be the activations of the j^{th} convolutional layer of the network ϕ when processing CT image I_{CT} and MR image I_{MR} , respectively. The perceptual loss is defined as follows:

$$L_{perc}(Syn_{MR}, Syn_{CT}) = \mathbb{E}_{I_{CT}, I_{MR} \sim p_{data}(I_{CT}, I_{MR})} \left[\frac{1}{K} \sum_{j=1}^K \frac{1}{H_j W_j C_j} \|\phi_j(Syn_{MR}(I_{CT})) - \phi_j(I_{MR})\|_1 \right] + \mathbb{E}_{I_{MR}, I_{CT} \sim p_{data}(I_{MR}, I_{CT})} \left[\frac{1}{K} \sum_{j=1}^K \frac{1}{H_j W_j C_j} \|\phi_j(Syn_{CT}(I_{MR})) - \phi_j(I_{CT})\|_1 \right] \quad (7)$$

where $\phi_j(Syn_{MR}(I_{CT}))$ and $\phi_j(Syn_{CT}(I_{MR}))$ are the activations of the synthetic images in the forward and backward cycles, respectively, $H_j \times W_j \times C_j$ is the shape of the activations from the j^{th} convolution layer, and K is the number of layers in the VGG16 network. By utilizing the activations of the higher layer in the VGG16 network, the synthetic images can preserve the overall spatial structure of the reference images, but not the texture and exact shape. Perceptual loss causes the synthetic images to become more perceptually similar to the reference images, but does not lead to an exact match.

The vertebra, spinal nerves, and ligaments in spinal images contain strong interdependencies. Structural similarity (SSIM) [51] is a perceptually-motivated metric that considers the human visual system and performs better in terms of visual pattern recognition than do quantitative metrics, e.g., mean-based metrics. To enhance the structural and perceptual similarity of the synthetic and reference images, structural similarity loss is expressed as follows:

$$L_{struc}(Syn_{MR}, Syn_{CT}) = \mathbb{E}_{I_{CT}, I_{MR} \sim p_{data}(I_{CT}, I_{MR})} \left[-\log \left(\max \left(0, SSIM(Syn_{MR}(I_{CT}), I_{MR}) \right) \right) \right] \\ + \mathbb{E}_{I_{MR}, I_{CT} \sim p_{data}(I_{MR}, I_{CT})} \left[-\log \left(\max \left(0, SSIM(Syn_{CT}(I_{MR}), I_{CT}) \right) \right) \right] \quad (8)$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \\ \text{s.t. } SSIM(x, y) \leq 1, \\ SSIM(x, y) = 1 \text{ if and only if } x = y \quad (9)$$

where C_1 and C_2 are constants that stabilize the division with the weak denominator and μ_x , μ_y , σ_x , σ_y , and σ_{xy} represent the mean, standard deviation, and cross-covariance of the synthetic and reference images.

The objective functions for supervised and unsupervised learning are defined as follows:

$$L_{sup}(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}) = L_{sup-adver}(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}) \\ + \lambda_{cycle} \cdot L_{sup-cycle}(Syn_{MR}, Syn_{CT}) \\ + \lambda_{voxel} \cdot L_{voxel}(Syn_{MR}, Syn_{CT}) \\ + \lambda_{grad} \cdot L_{grad}(Syn_{MR}, Syn_{CT}) \\ + \lambda_{perc} \cdot L_{perc}(Syn_{MR}, Syn_{CT}) \\ + \lambda_{struc} \cdot L_{struc}(Syn_{MR}, Syn_{CT}) \quad (10)$$

$$L_{unsup}(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}) = L_{unsup-adver}(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}) \\ + \lambda_{cycle} \cdot L_{unsup-cycle}(Syn_{MR}, Syn_{CT}) \quad (11)$$

where λ_{cycle} , λ_{voxel} , λ_{grad} , λ_{perc} , and λ_{struc} are hyper-parameters that balance the relative importance of adversarial, cycle-consistent, voxel-wise, gradient difference, perceptual, and structural similarity loss. In summary, the training objective function can be expressed mathematically as:

$$Syn_{MR}^* = \arg \min_{Syn_{MR}, Syn_{CT}} \max_{Dis_{MR}, Dis_{CT}} (L_{sup}(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}) \\ + L_{unsup}(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT})) \quad (12)$$

where Syn_{MR} and Syn_{CT} minimize the objective function, while Dis_{MR} and Dis_{CT} maximize it. During the inference process, only the Syn_{MR}^* network is used to produce a synthetic MR image from an input CT image.

3.4. Optimization of DC²Anet with Semi-Supervised Learning

DC²Anet with semi-supervised learning can be optimized in two different ways, with joint or alternating optimization:

- **Joint optimization:** For each training iteration, both the synthesis and discriminator networks are updated with regards to the objective function using supervised and unsupervised learning as defined in Equation (12). A pair of aligned data points and a pair of unaligned data points are sampled from the dataset and fed to DC²Anet to update the networks.
- **Alternating optimization:** For each training iteration, supervised and unsupervised learning for the objective function are alternated as defined in Equations (10) and (11). In this case, only the weights that correspond to the synthesis networks and the particular layers of the discriminators are updated. This form of training maintains a more stable convergence of the optimization, and it is easy to balance the synthesis and discriminator networks with Jensen-Shannon divergence [3]. However, the computational load required for alternating optimization is nearly twice as high as that of joint optimization in the training stage.

The most difficult complication of adversarial training is that one network may inevitably become more potent than the other, and this generally proved to be the discriminator network in most cases. When the discriminator network becomes too strong, the synthetic images are much easier to distinguish from the reference images. In this case, the gradients from the discriminator network approach zero. This results in no guidance for the further training of the synthesis network. To overcome this issue, alternating optimization is an effective approach for DC²Anet. DC²Anet with semi-supervised learning is described in Algorithm 1.

Algorithm 1 Mini-batch stochastic gradient descent training of DC²Anet. We used default values of $m = 1$, $n_{sup} = n_{unsup} = 1$, $\alpha = 0.0002$, $\beta_1 = 0.5$, and $\beta_2 = 0.999$.

Require: The batch size m , the number of alternative iterations between supervised learning and unsupervised learning n_{sup} and n_{unsup} , the learning rate α , and Adam hyperparameters β_1 and β_2 .

- 1: Construct unaligned data $P_{unaligned}\{I_{CT}, I_{MR}\}$ based on aligned data $P_{aligned}(I_{CT}, I_{MR})$.
 - 2: **for** number of training iterations **do**
 - 3: **for** n_{sup} steps **do**
 - 4: Sample $(I_{CT}^{(i)}, v_{MR}^{(i)})_{i=1}^m \sim P_{aligned}(I_{CT}, I_{MR})$ a batch from the aligned data
 - 5: Update the discriminator networks Dis_{MR} and Dis_{CT} by ascending their stochastic gradient:

$$L_{sup-adver}^{(i)} \leftarrow L_{sup-adver}^{(i)} \left(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}, I_{CT}^{(i)}, I_{MR}^{(i)} \right)$$

$$Dis_{MR}, Dis_{CT} \leftarrow Adam \left(\nabla_{Dis_{MR}, Dis_{CT}} \frac{1}{m} \sum_{i=1}^m -L_{sup-adver}^{(i)}, Dis_{MR}, Dis_{CT}, \alpha, \beta_1, \beta_2 \right)$$
 - 6: Update the synthesis networks Syn_{MR} and Syn_{CT} , by descending their stochastic gradient:

$$L_{sup}^{(i)} \leftarrow L_{sup}^{(i)} \left(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}, I_{CT}^{(i)}, I_{MR}^{(i)} \right)$$

$$Syn_{MR}, Syn_{CT} \leftarrow Adam \left(\nabla_{Syn_{MR}, Syn_{CT}} \frac{1}{m} \sum_{i=1}^m L_{sup}^{(i)}, Syn_{MR}, Syn_{CT}, \alpha, \beta_1, \beta_2 \right)$$
 - 7: **end for**
 - 8: **for** n_{unsup} steps **do**
 - 9: Sample $\{I_{CT}^{(i)}, I_{MR}^{(i)}\}_{i=1}^m \sim P_{unaligned}\{I_{CT}, I_{MR}\}$ a batch from the unaligned data
 - 10: Update the discriminator networks Dis_{MR} and Dis_{CT} , by ascending their stochastic gradient:

$$L_{unsup-adver}^{(i)} \leftarrow L_{unsup-adver}^{(i)} \left(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}, I_{CT}^{(i)}, I_{MR}^{(i)} \right)$$

$$Dis_{MR}, Dis_{CT} \leftarrow Adam \left(\nabla_{Dis_{MR}, Dis_{CT}} \frac{1}{m} \sum_{i=1}^m -L_{unsup-adver}^{(i)}, Dis_{MR}, Dis_{CT}, \alpha, \beta_1, \beta_2 \right)$$
 - 11: Update the synthesis networks Syn_{MR} and Syn_{CT} , by descending their stochastic gradient:

$$L_{unsup}^{(i)} \leftarrow L_{unsup}^{(i)} \left(Syn_{MR}, Syn_{CT}, Dis_{MR}, Dis_{CT}, I_{CT}^{(i)}, I_{MR}^{(i)} \right)$$

$$Syn_{MR}, Syn_{CT} \leftarrow Adam \left(\nabla_{Syn_{MR}, Syn_{CT}} \frac{1}{m} \sum_{i=1}^m L_{unsup}^{(i)}, Syn_{MR}, Syn_{CT}, \alpha, \beta_1, \beta_2 \right)$$
 - 12: **end for**
 - 13: **end for**
 - 14: **return** Syn_{MR}
-

3.5. Network Architecture

The synthesis networks Syn_{MR} and Syn_{CT} in DC²Anet adopt the same architecture as used in the network reported by Johnson et al. [48], who produced impressive results in real-time style transfer and single-image super-resolution. The network contained two stride-one convolutions at the beginning and the end, two stride-two convolutions, nine residual blocks [46,47], and two fractionally-strided convolutions with a stride of 0.5. Each residual block included two convolutions with 256 filters of a size of 3×3 and a stride of one. Instance normalization [56] and a rectified linear unit (ReLU) [57] activation function followed each convolution except in the final convolutional layer. The hyperbolic tangent (Tanh) activation function followed the final convolution to guarantee that the output was within $[-1, 1]$. A detailed description of the synthesis network is presented in Table 4.

Table 4. Model architecture of the synthesis network. Layers marked with IN indicate that the convolution layer is followed by the instance normalization layer. The ReLU activation layers are omitted.

Layer Name/Type	Output Size	Filter Size/Stride	Number of Conv. Layers	Number of Parameters
Input image	$H \times W \times 1$	None	0	0
Conv 1, IN	$H \times W \times 64$	$7 \times 7/1$	1	3200
Conv 2, IN	$H/2 \times W/2 \times 128$	$3 \times 3/2$	1	73,856
Conv 3, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	1	295,168
Residual Block 1, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 2, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 3, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 4, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 5, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 6, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 7, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 8, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Residual Block 9, IN	$H/4 \times W/4 \times 256$	$3 \times 3/2$	2	590,080
Fractional Conv 1, IN	$H/2 \times W/2 \times 128$	$3 \times 3/0.5$	1	295,040
Fractional Conv 2, IN	$H \times W \times 64$	$3 \times 3/0.5$	1	73,782
Conv 4, Tanh	$H \times W \times 1$	$7 \times 7/1$	1	3137
Total number of parameters				6,054,913
Total number of parameters if instance normalization is applied				6,065,217

For the discriminator networks Dis_{MR} and Dis_{CT} , we used a patch-based GAN (PatchGAN) [36] architecture, which aims to classify small overlapping image patches as either real or synthetic, rather than whole images. This patch-level discriminator architecture has fewer parameters than a whole-image discriminator and can emphasize detailed information in local areas. DC²Anet with semi-supervised learning has two different input flows, aligned and unaligned, with different shapes for the input data. The flow size of a volume of aligned data is $(N, H, W, 2)$. N is the batch size; H and W are the image height and width, respectively; and 2 represents a concatenation of the synthetic and input images. The flow size of a volume of unaligned data is $(N, H, W, 1)$, in which only a synthetic image can be used as input (indicated as 1). Therefore, a hybrid discriminator model was designed that consisted of two input stages, a shared stage, and two output stages. To balance the capability between synthesis and discriminator networks, the discriminator network was designed to be much

shallower than the synthesis network because generating images is much more difficult than merely distinguishing real from synthetic images. Based on the related works presented in Table 2, the number of discriminator layers was fixed at five, and the variant architectures of the hybrid discriminator are presented in Figure 3.

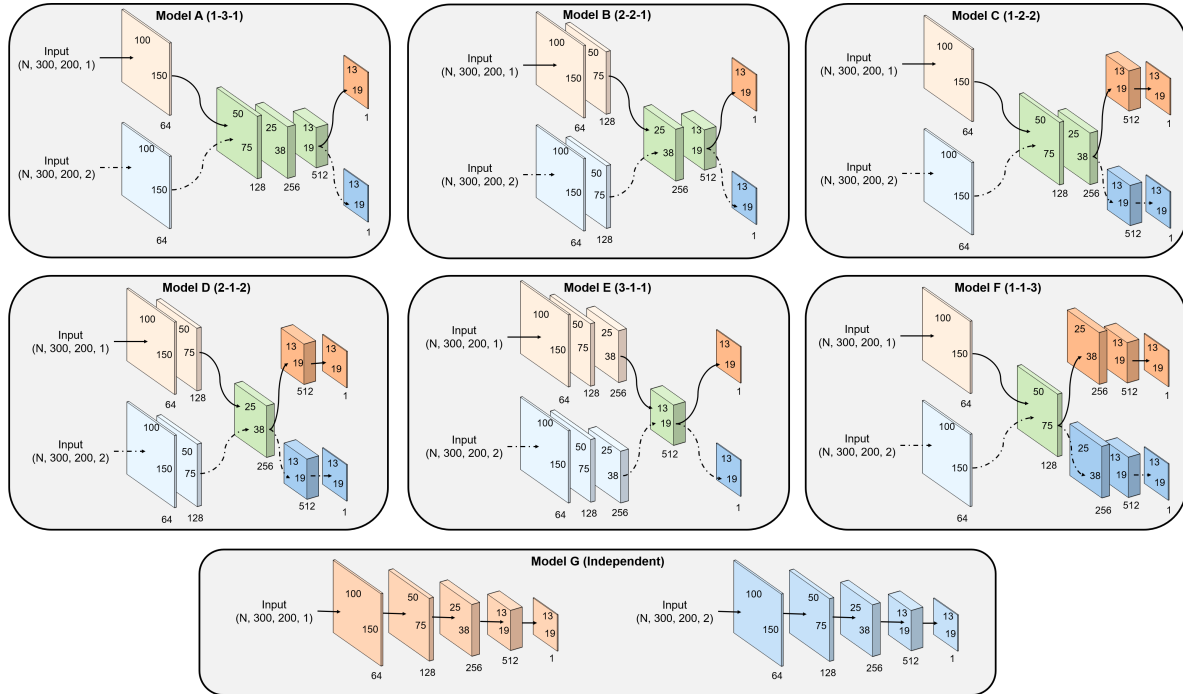


Figure 3. The variant architectures of the hybrid discriminator. Models A–F are hybrid discriminators for the aligned and unaligned data flow. Model G consists of two independent discriminators.

Models A–F represent variations of the input, shared, and output stages. Model G is the independent discriminator for unaligned and aligned data flows. Each aligned data flow consisted of a 300×200 synthetic image and a corresponding 300×200 input domain image, and each unaligned data flow had only a 300×200 synthetic image as input to the discriminator. All convolutions in the discriminator conducted 4×4 filters with a stride of 2. A leaky rectified activation (LeakyReLU) [58] followed each of the convolutions as the activation function, except for the final convolution.

4. Experimental Results and Discussion

4.1. Implementation Details

To stabilize the DC²Anet training process, we used an image pooling technique [59] that updates the discriminator networks Dis_{MR} and Dis_{CT} using a history of synthetic images rather than the ones generated by the latest synthesis networks. We maintained an image pool buffer that stored the 50 previously-synthesized images. We also conducted data augmentation using random horizontal flipping (-5 – 5 degree rotation) and the random translation of up to 15 pixels in each spatial dimension in the training images. DC²Anet was trained with mini-batch stochastic gradient descent (SGD) [60] with a mini-batch size of one. All weights were initialized from a zero-centered truncated normal distribution with a standard deviation of 0.02. All networks were trained with a learning rate of 0.0002 for the first 100,000 iterations and a linearly decaying rate that went to zero over the next 100,000 iterations. Adam is one of the most pervasive and robust optimizers used in various field [61,62]. The model was also optimized using the Adam optimizer [63] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, as suggested in [28]. For all experiments, the following empirical values were used to train the synthesis networks: $\lambda_{cycle} = 10$, $\lambda_{voxel} = 100$, $\lambda_{grad} = 100$, $\lambda_{perc} = 1$, and $\lambda_{struc} = 0.05$.

In LeakyReLU, the slope of the leak was set to 0.2. Reflection padding was used to reduce artifacts instead of zero padding in the convolution layers. The model took about 48 h to train for 200,000 iterations using a single GeForce GTX 1080Ti GPU. The code and pretrained models are available at <https://github.com/ChengBinJin/SpineC2M>.

4.2. Data Acquisition

Our lumbar spine dataset consisted of 641 patients, each with CT and MR images. The CT image was acquired helically on a GE Revolution CT scanner with a tube voltage of 120 kV, an exposure of 450 mAs, and a slice thickness of 1.00 mm. The MR image for each patient was obtained using a Siemens 3.0T Trio TIM MR scanner with T2 3D (with a repetition time of 4320 ms, an echo time of 95 ms, and a flip angle of 150°). To allow the voxel-wise comparison of the synthetic and reference MR images, the CT image was manually aligned to the MR image to produce voxel-level correspondence. After alignment, the CT and MR images from the same patient had the same image size and spacing. Because only the lumbar spine region was considered, we cropped the aligned CT and MR images to reduce the computational burden, producing a final preprocessed image size of $300 \times 200 \times 40$ (40–48 slices depending on the alignment quality) with the same voxel size ($1.00 \times 1.00 \times 1.00$ mm). We randomly separated the 641 patients into two groups: 549 patients for the training set and 92 patients for the test set. Table 5 presents a summary of our dataset, while Figure 4 displays several sample images.

Table 5. Summary of the lumbar vertebra dataset used in the experiments.

	Number of Patients	Number of Slices
Training set	549	22,428
Test set	92	4426
Total	641	26,854

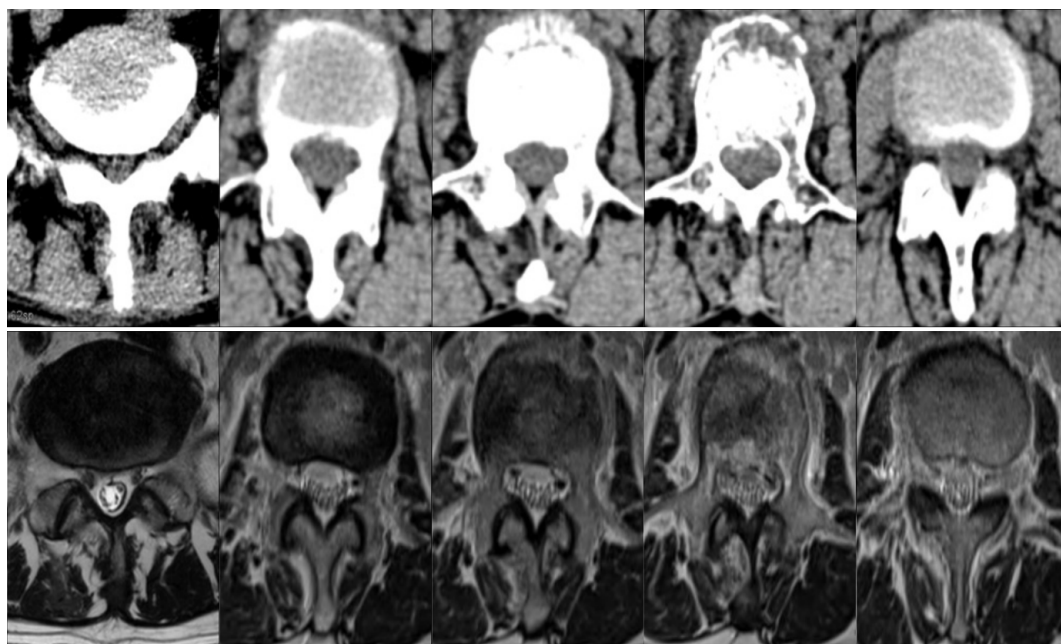


Figure 4. Sample images of axial lumbar vertebra CT (upper row) and MR images (lower row).

4.3. Evaluation Metrics

The synthesis and reference MR images were compared using the mean absolute error (MAE) and root mean squared error (RMSE), which are defined as follows:

$$MAE = \frac{1}{N} \sum_{i=0}^{N-1} \|I_{MR}(i) - Syn_{MR}(I_{CT}(i))\| \quad (13)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (I_{MR}(i) - Syn_{MR}(I_{CT}(i)))^2} \quad (14)$$

where N is the total number of image slices in the aligned voxel. MAE and RMSE measure the average distance between each pixel in the synthetic and reference MR images. In addition, the voxel-wise peak-signal-to-noise-ratio (PSNR) can also be calculated:

$$PSNR = 10 \cdot \log_{10} \left(\frac{H^2}{MSE} \right) \quad (15)$$

$$MSE = \frac{1}{N} \sum_{i=0}^{N-1} (I_{MR}(i) - Syn_{MR}(I_{CT}(i)))^2 \quad (16)$$

where H is the maximum possible intensity of the pixel and MSE is the mean square error, which represents the square of difference between I_{MR} and $Syn_{MR}(I_{CT})$. MAE, RMSE, and PSNR were based on the correct alignment of test images I_{CT} and I_{MR} .

Because of the enormous differences between two image domains, it is difficult to achieve perfect image alignment. Therefore, the structural similarity (SSIM) index and the Pearson correlation coefficient (PCC) should also be calculated for patch-wise statistical comparisons, e.g., mean, variance, and correlation. The definition of the SSIM is given in Equation (9), and the PCC is defined as follows:

$$PCC = \frac{1}{N} \sum_{i=0}^N \frac{(I_{MR}(i) - \mu_{I_{MR}(i)})(Syn_{MR}(I_{CT}(i)) - \mu_{Syn_{MR}(I_{CT}(i))})}{\sigma_{I_{MR}(i)} \sigma_{Syn_{MR}(I_{CT}(i))}} \quad (17)$$

where μ and σ are the mean and variance of the i^{th} image slice. Lower values for the MAE and RMSE are preferable, while the reverse is true for the PSNR, SSIM, and PCC.

4.4. Analysis of DC²Anet

Based on our lumbar spine dataset and the metrics described in the previous section, we quantitatively evaluated the performance of our model in generating an MR image from a CT image. In Table 6, we compare the performance of DC²Anet with supervised, unsupervised, and semi-supervised learning. Data alignment of the tuples of the corresponding images in supervised learning produced a much higher accuracy than did unsupervised learning, while semi-supervised learning with alternating optimization was better than both supervised and unsupervised learning. The joint optimization of semi-supervised learning produced substantially weaker results compared to alternating optimization. Therefore, we concluded that the alternating optimization of DC²Anet led to a more stable convergence and was critical to effective performance.

Table 6. Comparison between different learning and optimization methods. The best scores are displayed in bold. PCC, Pearson correlation coefficient.

Learning	Optimization	MAE	RMSE	PSNR	SSIM	PCC
Supervised	—	28.873	39.562	64.533	0.242	0.445
Unsupervised	—	31.297	42.205	63.976	0.227	0.420
Semi-supervised	Joint	29.048	39.801	64.502	0.244	0.439
Semi-supervised	Alternating	28.819	39.418	64.553	0.248	0.453

Table 7 presents a comparison of the performance of the variant architectures for the discriminator previously displayed in Figure 3. Models A, B, and C had a different number of convolution layers in the input and output stages, and more than one convolution layer in the shared stage. In contrast, Models D, E, and F had only one convolution layer in the shared stage with a different number of convolution layers in the input and output stages. For the two different data flows, an independent discriminator design was employed in Model G. From the experimental results, three significant observations are worth noting. First, the independent discriminator architecture (Model G) exhibited higher performance than Models D, E, and F, due to the high discriminatory capability of the independent network. Second, Models A, B, and C outperformed the other models. This is because the deep weight-sharing constraint in the shared stage can learn the joint distribution of the aligned and unaligned data. Finally, Model C, which consists of two convolutions in the shared stage and two layers in the output stage, exhibited the most effective discriminatory capability, outperforming all other models in all metrics except for SSIM.

Table 7. Analysis of the discriminator architecture. The best scores are displayed in bold.

Discriminator	MAE	RMSE	PSNR	SSIM	PCC
Model A	29.096	39.826	64.460	0.244	0.447
Model B	29.024	39.791	64.472	0.249	0.447
Model C	28.819	39.418	64.553	0.248	0.453
Model D	31.320	42.345	63.933	0.228	0.410
Model E	32.415	43.440	63.722	0.234	0.437
Model F	31.232	42.353	64.937	0.226	0.402
Model G	31.011	41.980	63.931	0.221	0.414

As demonstrated in [48,49], synthesizing an image by minimizing the perceptual loss for the early layers of the pretrained network tends to focus on low-level information, such as intensity, texture, and shape. Perceptual loss is helpful when there is misalignment in the training and test datasets. Layer selection for perceptual loss is a task-oriented problem. We considered five ReLU layers before max-pooling in the pretrained VGG16 network as in [48,49]. The performance of the different perceptual layers is summarized in Table 8. The five layers were ReLUs 1_2, 2_2, 3_3, 4_3, and 5_3, with the high-layer ReLUs always including the early layer ones. Table 8 indicates that perceptual loss defined by high layers produced more accurate output than did the early layers. We also observed that the perceptual loss from ReLU {1_2, 2_2, 3_3, 4_3} and ReLU {1_2, 2_2, 3_3, 4_3, 5_3} had minor quantitative differences.

Table 8. Evaluation of variation in perceptual layers. The best scores are displayed in bold.

Perceptual Layers	MAE	RMSE	PSNR	SSIM	PCC
ReLU {1_2}	29.147	39.858	64.459	0.241	0.442
ReLU {1_2, 2_2}	29.058	39.787	64.477	0.243	0.449
ReLU {1_2, 2_2, 3_3}	28.882	39.596	64.530	0.245	0.446
ReLU {1_2, 2_2, 3_3, 4_3}	28.825	39.383	64.574	0.241	0.448
ReLU {1_2, 2_2, 3_3, 4_3, 5_3}	28.819	39.418	64.553	0.248	0.453

Our objective function contained six independent loss terms. The experiments reported above used all of the loss terms. To investigate the strength of each loss term, we employed ablation analysis to determine how performance was affected by each loss term. We trained each network with a different objective function five times using different initialization weights and report the average of the five trials for each objective function. The evaluation results are shown in Table 9. Beginning with adversarial loss alone, each loss term was added one by one. In this process, five metrics were used to analyze the change in performance, and relative MAE improvement was also calculated. We considered

adversarial loss alone to represent 0%, and the inclusion of all loss terms (the final row) was considered to be 100% when calculating relative MAE improvement.

The synthesis results for the ablation analysis are presented in Figure 5. The performance of DC²Anet generally improved with the addition of each loss term, with voxel-wise loss the most useful in terms of relative improvement. This is because voxel-wise loss and MAE are consistent with a per-pixel mean-error-based measure. Dual cycle-consistent and gradient difference loss exhibited relative improvement of 12.19% and 6.95%, respectively, while perceptual loss and structural similarity loss had a limited effect on the improvement of performance compared to the other forms of loss. However, when all terms were used, the occurrence of unnatural features in the synthetic MR images was significantly reduced. As a result of the above results, in the remainder of the experiments, we employed DC²Anet with the following characteristics: alternating optimization of semi-supervised learning, Model C for the discriminator architecture, perceptual loss from ReLUs {1_2, 2_2, 3_3, 4_3, 5_3}, and the inclusion of all loss terms.

Table 9. Ablation analysis of the objective function. The results represent the average of five trials. The best scores are displayed in bold.

Objectives	MAE	RMSE	PSNR	SSIM	PCC	Relative MAE Improvement (%)
Adversarial alone	33.570	44.654	63.396	0.207	0.329	0.00
+ Dual cycle-consistent	32.991	44.721	63.437	0.208	0.330	12.19
+ Voxel-wise	29.276	39.989	64.402	0.237	0.450	90.38
+ Gradient difference	28.946	39.565	64.520	0.245	0.449	97.33
+ Perceptual	28.855	39.430	64.563	0.245	0.451	99.24
+ Structural similarity	28.819	39.418	64.553	0.248	0.453	100.00

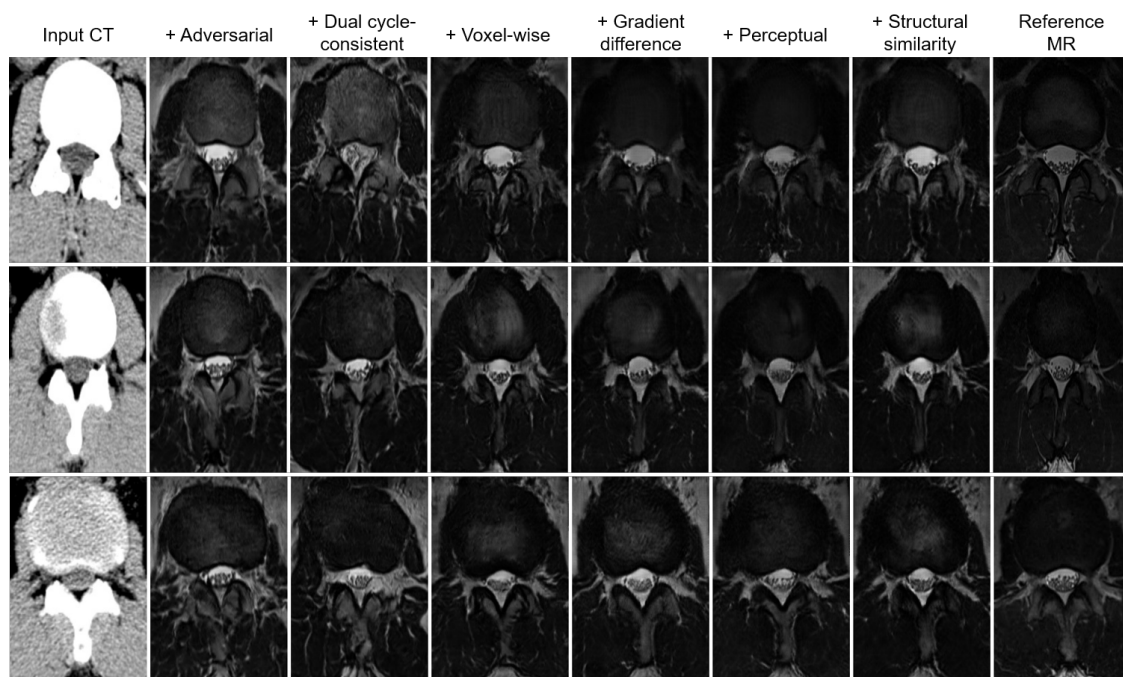


Figure 5. Ablation analysis of the proposed method. From left to right: the input CT, adversarial loss alone, the addition of dual cycle-consistent loss, the addition of voxel-wise loss, the addition of gradient difference loss, the addition of perceptual loss, the addition of structural similarity loss, and the reference MR image.

4.5. Comparison with Baselines

To compare synthetic MR images produced using different methods quantitatively, we present box plots in Figure 6 representing the MAE, RMSE, PSNR, SSIM, and PCCs resulting from the use of multi-channel GAN [33], deep MR-to-CT [41], DiscoGAN [44], MR-GAN [45], and our proposed method. The circles next to the box plots represent a single image slice from the test dataset. The top and bottom box limits were calculated from Q_{25} and Q_{75} , respectively. The green triangles and the horizontal lines denote the average and the median. The range of the box plot whiskers is given by $[Q_{25} - 1.5 \times (Q_{75} - Q_{25}), Q_{75} + 1.5 \times (Q_{75} - Q_{25})]$. Any data point that falls outside of this range is typically considered an outlier and indicated by a red cross. The averages and standard deviations displayed in Table 10 indicate that our proposed method outperformed the other methods for all measures, with the lowest MAE and RMSE and the highest PSNR, SSIM, and PCC, thus further verifying the utility of our architecture. In addition, t -tests were conducted on the results in Table 10, finding that agreement with the reference MR images was significantly lower ($p < 0.05$) for images obtained using the MR-GAN method than for the images obtained using the DC²Anet model.

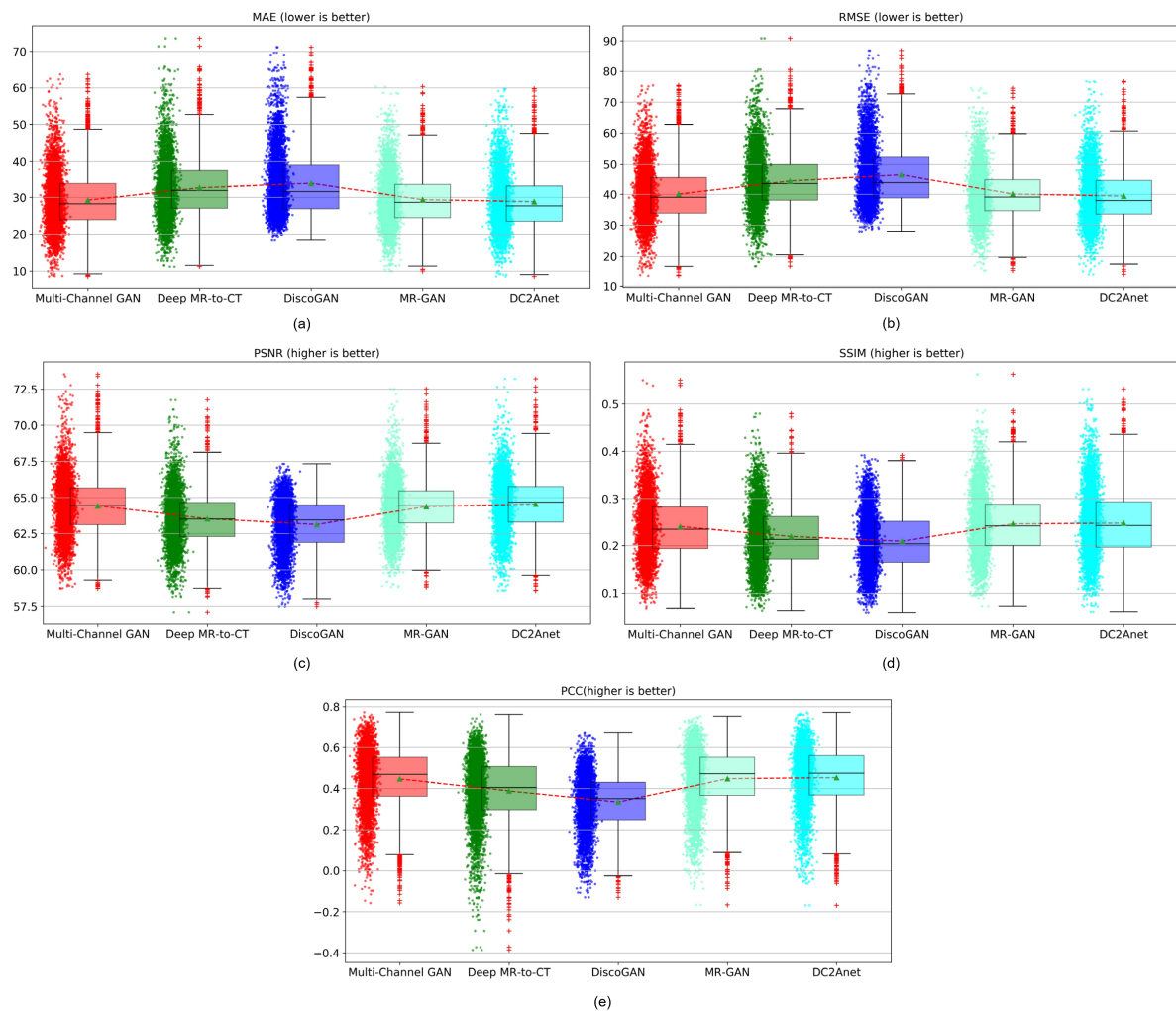


Figure 6. A comparison of the proposed approach with baseline methods based on (a) MAE, (b) RMSE, (c) PSNR, (d) SSIM, and (e) PCC metrics.

Table 10. Overall statistics for five measures of model quality: MAE, RMSE, PSNR, SSIM, and PCC. The average and standard deviation for each measure from 92 subjects in a lumbar vertebra dataset are presented for five methods. The best scores are displayed in bold.

Methods	MAE	RMSE	PSNR	SSIM	PCC
Multi-Channel GAN [33]	29.245 \pm 7.786	40.038 \pm 8.913	64.428 \pm 1.956	0.240 \pm 0.066	0.447 \pm 0.146
Deep MR-to-CT [41]	32.632 \pm 8.239	44.320 \pm 9.369	63.524 \pm 1.855	0.219 \pm 0.066	0.388 \pm 0.160
DiscoGAN [44]	33.863 \pm 9.031	46.317 \pm 9.853	63.130 \pm 1.756	0.209 \pm 0.061	0.334 \pm 0.136
MR-GAN [45]	29.276 \pm 7.218	39.989 \pm 8.167	64.402 \pm 1.769	0.237 \pm 0.066	0.450 \pm 0.144
DC ² Anet	28.819 \pm 7.655	39.418 \pm 8.660	64.553 \pm 1.890	0.248 \pm 0.072	0.453 \pm 0.146

The MAE and standard deviation for the first 20 of the 92 subjects are plotted in Figure 7, comparing DC²Anet with multi-channel GAN [33], deep MR-to-CT [41], DiscoGAN [44], and MR-GAN [45]. It can be seen that DC²Anet generated a smaller MAE than the other approaches for most of the subjects. However, for some subjects, the MR-GAN [45] approach produced smaller MAE than did DC²Anet, though the MR-GAN [45] was unstable for some subjects, such as Subject 04 and Subject 08.

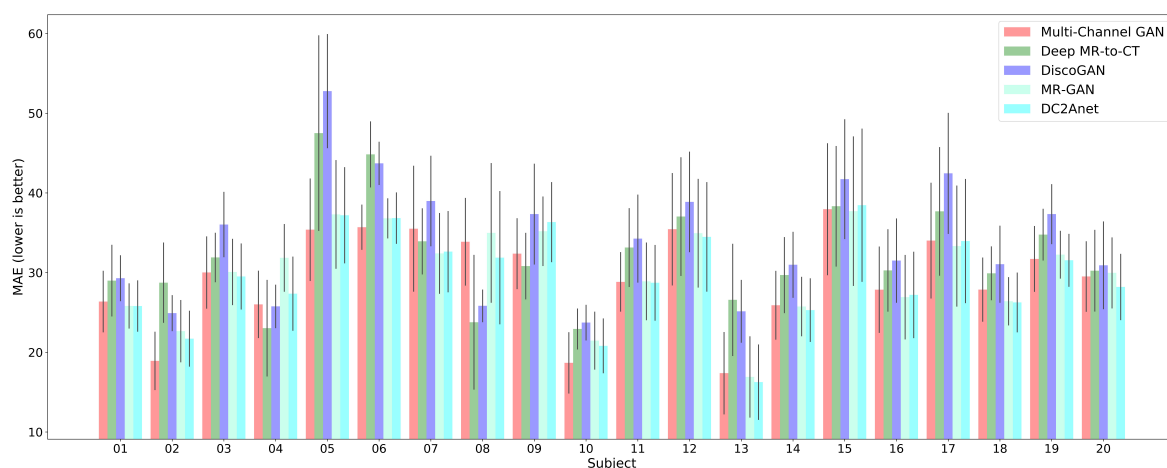


Figure 7. MAE computed for the first 20 of the 92 subjects in the test dataset.

Figure 8 presents three examples of synthetic images produced by the proposed DC²Anet method, alongside the corresponding CT and MR images. The results for multi-channel GAN [33], deep MR-to-CT [41], DiscoGAN [44], and MR-GAN [45] are also presented for comparison purposes. The spinal cord region in the central area of the image, the most important element of the image, is enlarged to evaluate the reconstruction capability of each method. DC²Anet learned to differentiate between different structures with similar intensity values in CT images, but not in MR images, such as a vertebra, fat tissue, and disc signals. DC²Anet also preserved the continuity, smoothness, and semantics of the original images in the synthetic results because our objective function with semi-supervised learning led the synthetic MR images to be similar to the reference images. In CT-based MR image generation, the accurate reconstruction of the disc signal, the degree of disc protrusion, the degree of stenosis, and the thecal sac are essential in the analysis of lumbar vertebra. We can see that the disc signal and thecal sac in the synthetic MR image obtained using the proposed DC²Anet looked more similar to the reference MR image compared to the other methods. The structures of the muscle and fat tissue had a highest similarity. However, the proposed method exhibited limitations in the reconstruction of the degree of disc protrusion and the degree of stenosis.

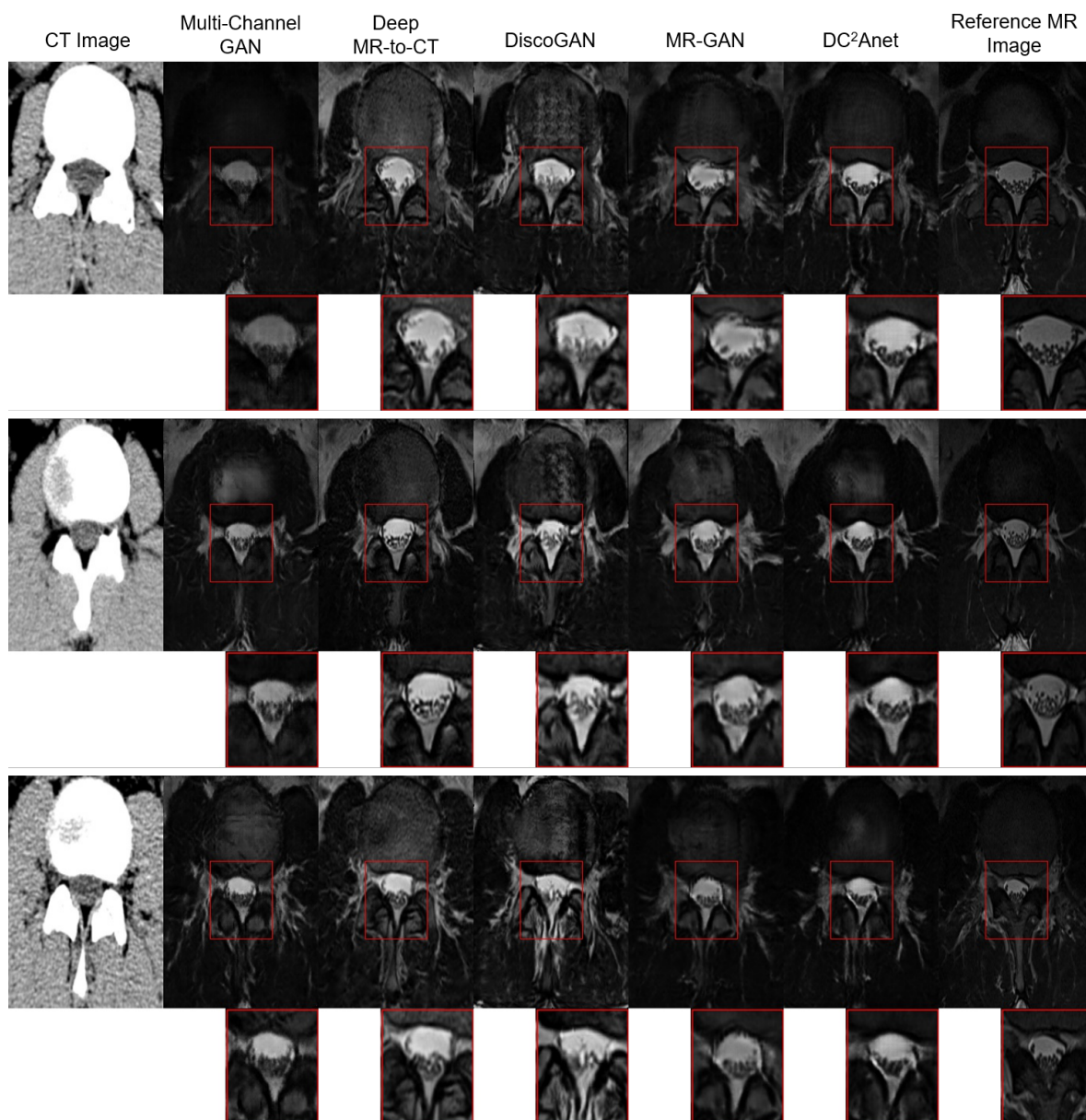


Figure 8. Qualitative comparisons of DC²Anet and baseline methods. Rows 1, 3, and 5 show CT images, synthetic images, and the reference image. Rows 2, 4, and 6 show the enlarged spinal cord of the corresponding images. From left to right: input CT image, synthesis MR images from multi-channel GAN [33], deep MR-to-CT [41], DiscoGAN [44], MR-GAN [45], the proposed DC²Anet, and the reference MR image.

5. Conclusions

In this work, we proposed an objective function and a general synthesis system, DC²Anet, that employs semi-supervised learning to generate lumbar spine MR images from single-sequence CT scans. Our objective function included six independent loss terms. Using ablation analysis, we assessed in detail the effectiveness and relative importance of each loss term. Performance was improved by adding each loss term because each had its own particular strengths and weaknesses. DC²Anet using semi-supervised learning can significantly outperform supervised and unsupervised learning approaches. To further improve the accuracy and to seek the global minimum of the objective function, alternating optimizing was much more efficient than the integrated optimization of DC²Anet. We applied our method to generate MR images from their corresponding CT images, demonstrating

that our proposed method significantly outperformed four state-of-the-art approaches, thus providing its suitability for cross-modality image synthesis. Thus, it represents a very promising method that can be employed in the diagnosis of lumbar disc conditions for patients who are prevented from receiving an MRI due to claustrophobia or the presence of a cardiac pacemaker. Future research intends to further validate the quality of the synthesis results for downstream tasks such as segmentation or classification. Extending the method to handle cross-sectional views (axial, sagittal, and coronal) and multi-sequence CT images will also be considered in future work.

Author Contributions: Conceptualization, C.-B.J. and X.C.; data curation, S.J.; investigation, S.J.; software, E.P. and Y.S.A.; supervision, H.K.; validation, I.H.H., J.I.L., and J.H.L.; writing, original draft, C.-B.J.; writing, review and editing, H.K., M.L.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, USA, 3–8 December 2012; pp. 1097–1105.
2. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [[CrossRef](#)]
3. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
4. Hsu, S.H.; Cao, Y.; Huang, K.; Feng, M.; Balter, J.M. Investigation of a method for generating synthetic CT models from MRI scans of the head and neck for radiation therapy. *Phys. Med. Biol.* **2013**, *58*, 8419. [[CrossRef](#)] [[PubMed](#)]
5. Zheng, W.; Kim, J.P.; Kadbi, M.; Movsas, B.; Chetty, I.J.; Glide-Hurst, C.K. Magnetic resonance-based automatic air segmentation for generation of synthetic computed tomography scans in the head region. *Int. J. Radiat. Oncol. Biol. Phys.* **2015**, *93*, 497–506. [[CrossRef](#)] [[PubMed](#)]
6. Kapanen, M.; Tenhunen, M. T1/T2*-weighted MRI provides clinically relevant pseudo-CT density data for the pelvic bones in MRI-only based radiotherapy treatment planning. *Acta Oncol.* **2013**, *52*, 612–618. [[CrossRef](#)] [[PubMed](#)]
7. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
8. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of Wasserstein GANs. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5767–5777.
9. Su, K.H.; Hu, L.; Stehning, C.; Helle, M.; Qian, P.; Thompson, C.L.; Pereira, G.C.; Jordan, D.W.; Herrmann, K.A.; Traugher, M.; et al. Generation of brain pseudo-CTs using an undersampled, single-acquisition UTE-mDixon pulse sequence and unsupervised clustering. *Med. Phys.* **2015**, *42*, 4974–4986. [[CrossRef](#)] [[PubMed](#)]
10. Huynh, T.; Gao, Y.; Kang, J.; Wang, L.; Zhang, P.; Lian, J.; Shen, D. Estimating CT image from MRI data using structured random forest and auto-context model. *IEEE Trans. Med. Imaging* **2016**, *35*, 174–183. [[CrossRef](#)]
11. Jog, A.; Carass, A.; Prince, J.L. Improving magnetic resonance resolution with supervised learning. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April–2 May 2014; pp. 987–990.
12. Catana, C.; van der Kouwe, A.; Benner, T.; Michel, C.J.; Hamm, M.; Fenchel, M.; Fischl, B.; Rosen, B.; Schmand, M.; Sorensen, A.G. Toward implementing an MRI-based PET attenuation-correction method for neurologic studies on the MR-PET brain prototype. *J. Nucl. Med.* **2010**, *51*, 1431–1438. [[CrossRef](#)]
13. Andreasen, D.; Van Leemput, K.; Edmund, J.M. A patch-based pseudo-CT approach for MRI-only radiotherapy in the pelvis. *Med. Phys.* **2016**, *43*, 4742–4752. [[CrossRef](#)]
14. Arabi, H.; Koutsouvelis, N.; Rouzaud, M.; Miralbell, R.; Zaidi, H. Atlas-guided generation of pseudo-CT images for MRI-only and hybrid PET-MRI-guided radiotherapy treatment planning. *Phys. Med. Biol.* **2016**, *61*, 6531. [[CrossRef](#)]

15. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
16. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
17. Tahmassebi, A.; Gandomi, A.H.; McCann, I.; Schulte, M.H.; Goudriaan, A.E.; Meyer-Baese, A. Deep learning in medical imaging: fMRI big data analysis via convolutional neural networks. In Proceedings of the PEARC'18, Pittsburgh, PA, USA, 22–26 July 2018.
18. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115. [[CrossRef](#)] [[PubMed](#)]
19. Dai, W.; Doyle, J.; Liang, X.; Zhang, H.; Dong, N.; Li, Y.; Xing, E.P. Scan: Structure correcting adversarial network for chest x-rays organ segmentation. *arXiv* **2017**, arXiv:1703.08770.
20. Son, J.; Park, S.J.; Jung, K.H. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. *arXiv* **2017**, arXiv:1706.09318.
21. Alex, V.; KP, M.S.; Chennamsetty, S.S.; Krishnamurthi, G. Generative adversarial networks for brain lesion detection. In *Medical Imaging 2017: Image Processing*; International Society for Optics and Photonics: Bellingham, WA, USA, 2017; Volume 10133, p. 101330G.
22. Han, X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med. Phys.* **2017**, *44*, 1408–1419. [[CrossRef](#)] [[PubMed](#)]
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 5–9 October 2015; Springer: Munich, Germany, pp. 234–241.
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
26. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
27. Oord, A.v.d.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv* **2016**, arXiv:1601.06759.
28. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
29. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial network. *arXiv* **2016**, arXiv:1609.03126.
30. Yeh, R.A.; Chen, C.; Yian Lim, T.; Schwing, A.G.; Hasegawa-Johnson, M.; Do, M.N. Semantic image inpainting with deep generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5485–5493.
31. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Liu, G.; Tao, A.; Kautz, J.; Catanzaro, B. Video-to-video synthesis. *arXiv* **2018**, arXiv:1808.06601.
32. Chan, C.; Ginosar, S.; Zhou, T.; Efros, A.A. Everybody dance now. *arXiv* **2018**, arXiv:1808.07371.
33. Bi, L.; Kim, J.; Kumar, A.; Feng, D.; Fulham, M. Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs). In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 43–51.
34. Ben-Cohen, A.; Klang, E.; Raskin, S.P.; Amitai, M.M.; Greenspan, H. Virtual PET images from CT data using deep convolutional networks: Initial results. In *International Workshop on Simulation and Synthesis in Medical Imaging*; Springer: Cham, Switzerland, 2017; pp. 49–57.
35. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
36. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
37. Nie, D.; Trullo, R.; Lian, J.; Petitjean, C.; Ruan, S.; Wang, Q.; Shen, D. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2017; pp. 417–425.

38. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [[CrossRef](#)] [[PubMed](#)]
39. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
40. Tu, Z.; Bai, X. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1744–1757. [[PubMed](#)]
41. Wolterink, J.M.; Dinkla, A.M.; Savenije, M.H.; Seevinck, P.R.; van den Berg, C.A.; Išgum, I. Deep MR to CT synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*; Springer: Cham, Switzerland, 2017; pp. 14–23.
42. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
43. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
44. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
45. 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* **2019**, *19*, 2361. [[CrossRef](#)] [[PubMed](#)]
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 630–645.
48. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 694–711.
49. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
50. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.
51. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
52. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [[CrossRef](#)]
53. Sundaram, N.; Brox, T.; Keutzer, K. Dense point trajectories by GPU-accelerated large displacement optical flow. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2010; pp. 438–451.
54. Zach, C.; Klopschitz, M.; Pollefeys, M. Disambiguating visual relations using loop constraints. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1426–1433.
55. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
56. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022.
57. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*; Stanford Artificial Intelligence Laboratory: Stanford, CA, USA, 2013; Volume 30, p. 3.
58. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.

59. Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; Webb, R. Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2107–2116.
60. Keskar, N.S.; Socher, R. Improving generalization performance by switching from Adam to SGD. *arXiv* **2017**, arXiv:1712.07628.
61. Tahmassebi, A. ideeple: Deep learning in a flash. In *Disruptive Technologies in Information Sciences*; International Society for Optics and Photonics: Bellingham, WA, USA, 2018; Volume 10652, p. 106520S.
62. Tahmassebi, A.; Gandomi, A.H.; Fong, S.; Meyer-Baese, A.; Foo, S.Y. Multi-stage optimization of a deep model: A case study on ground motion modeling. *PLoS ONE* **2018**, *13*, e0203829. [[CrossRef](#)] [[PubMed](#)]
63. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).