Improved MRI-based Pseudo-CT Synthesis via Segmentation Guided Attention Networks

Gurbandurdy Dovletov[1][a], Duc Duy Pham[1], Josef Pauli[1][b], Marcel Gratz[2,3][c]
and Harald H. Quick[2,3]

[1] *Intelligent Systems Group, Faculty of Engineering, University of Duisburg-Essen, Duisburg, Germany*

[2] *High-Field and Hybrid MR Imaging, University Hospital Essen, University of Duisburg-Essen, Essen, Germany*

[3] *Erwin L. Hahn Institute for MR Imaging, University of Duisburg-Essen, Essen, Germany*

Keywords: Image-to-Image Translation, Pseudo-CT, Attention Mechanism, U-Net, Generative Adversarial Network.

Abstract: In this paper, we propose 2D MRI-based pseudo-CT (pCT) generation approaches that are inspired by U-Net and generative adversarial networks (GANs) and that additionally utilize coarse bone segmentation guided attention (SGA) mechanisms for better image synthesis. We first introduce and formulate SGA and its extended version (E-SGA), then we embed them into our baseline U-Net and conditional Wasserstein GAN (cWGAN) architectures. Since manual bone annotations are expensive, we derive coarse bone segmentations from CT/pCT images via thresholding and utilize them during the training phase to guide image-to-image translation attention networks. For inference, no additional segmentations are required. The performance of the proposed methods regarding the image generation quality is evaluated on the publicly available RIRE data set. Since MR and CT image pairs in this data set are not correctly aligned with each other, we also briefly describe the applied image registration procedure. The results of our experiments are compared to baseline U-Net and conditional Wasserstein GAN implementations and demonstrate improvements for bone regions.

1 INTRODUCTION

Positron emission tomography (PET) is an imaging technique that reveals physiological and biochemical processes of tissue and organs by measuring their metabolic activity. The technique is based on the detection of radioactivity emitted in opposite directions after a small amount of a radioactive tracer is injected into a peripheral vein (Paans, 2006). When traveling through some tissue or hardware parts such as the patient table these photons lose their energy and as a consequence, only a part of them reaches detectors. In order to compensate for such physical behavior, an attenuation correction (AC) procedure is required (Ollinger and Fessler, 1997). In stand-alone PET systems, AC is performed based on a transmission scan acquired using external rod sources.

A lack of anatomical information in a pure PET initiated the development of combined PET and computed tomography (CT) systems in a single gantry, also known as PET/CT. The additional CT examination in such systems allows deriving attenuation correction maps (μ-maps) directly from Hounsfield units (HU) by scaling the acquired CT image to the energy level of a PET image (Beyer et al., 2000). Obtained in such a way μ-maps are later used for attention correction of PET scans. The main application areas of PET/CT are cardiology, neurology, and oncology. According to (Brady et al., 2008), in oncology, it is by far the most intensively used diagnostic and staging device.

Being able to provide an excellent soft tissue contrast, PET/MR combination was introduced, which moreover goes without the need of additional radiation as it is the case for PET/CT. Such systems do not provide a straightforward way to obtain μ-maps for AC by energy scaling like it is possible in stand-alone PET or PET/CT combination (Keereman et al., 2013). Unlike PET and CT, MRI uses physical mechanisms that do not allow for a direct derivation or correlation of image intensities with electron density.

Various approaches have been proposed to resolve this issue. In atlas-based methods, μ-maps are calculated based on prior registration of atlas images to a

[a] https://orcid.org/0000-0002-2401-8745
[b] https://orcid.org/0000-0003-0363-6410
[c] https://orcid.org/0000-0001-9723-5233

Dovletov, G., Pham, D., Pauli, J., Gratz, M. and Quick, H.
Improved MRI-based Pseudo-CT Synthesis via Segmentation Guided Attention Networks.
DOI: 10.5220/0010849200003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 2: BIOIMAGING*, pages 131-140
ISBN: 978-989-758-552-4; ISSN: 2184-4305

target image with subsequent utilization of achieved transformations. (Burgos et al., 2014) propose to generate synthetic CTs by locally matching target MR images to a database of MR-CT pairs and deriving the corresponding μ-maps from them. An alternative approach is a segmentation-based AC, where discrete linear attenuation coefficients (LACs) are assigned to each achieved segmentation class (Berker et al., 2012). In order to take advantage of both segmentation- and atlas-based approaches, (Paulus et al., 2015) propose to first derive μ-maps via segmenting a Dixon image into air, fat, lung, and soft tissue classes. After that, atlas-based continuous μ-maps for each of major body bones (femur, hip, spine, and skull) are derived and fused with the existing μ-maps from the previous step.

One other possible way to derive μ-maps for PET AC is to first generate a pseudo-CT (pCT) image from an MRI and then scale it to PET's energy level in the same way as it is done for PET/CT machines. This MRI-based pCT generation task, however, introduces some challenges, since the MR modality is sensitive to proton density and thus not capable of distinguishing between air and lungs. Moreover, compact bone is also indistinguishable as its signal vanishes, due to the comparably short relaxation time, before it can be acquired.

Previously proposed methods have already demonstrated the feasibility of generating pCT from MR images using deep learning methods. (Liu et al., 2017) propose a pipeline for PET attenuation correction, where an encoder-decoder like architecture is first employed to segment an MR image into three classes, namely, air, bone, and soft tissue. The obtained segmentation is later converted to a pseudo-CT image via mapping each class ID to its corresponding statistical HU value. (Nie et al., 2016) propose to use a 3D fully convolution network (FCN) for a direct non-linear mapping from MRI to CT domain. They utilize a patch-based training procedure for better preservation of the neighborhood information in predicted pseudo-CT images. Moreover, an additional image gradient difference loss term was used to retain the sharpness of synthesized CTs. In order to generate even more realistic images, (Nie et al., 2017) extend their existing architecture with an additional discriminator and employ the adversarial training approach from (Goodfellow et al., 2014). (Han, 2017) adapts and modifies the well-known U-Net (Ronneberger et al., 2015) architecture by changing the number of convolutional layers in such a way, that allows to initialize the feature extraction part of the proposed architecture from a pretrained VGG-16 model. (Wolterink et al., 2017) propose to utilize the state-of-the-art Cycle-GAN (Zhu et al., 2017) image-to-image translation model, which is the extension of GAN (Goodfellow et al., 2014) and involves the simultaneous training of two generators and two discriminator models.

There have been several attempts to improve the bone quality of generated (synthesized) pseudo-CT images with the help of additional information. (Leynes et al., 2018) propose to utilize zero-echo-time (ZTE) images additionally to fat and water maps derived from a 2-echo Dixon MRI (in-phase and out-of-phase) sequence to capture bone information, and thus, generate more accurate pseudo-CT images. The main limitations of their work, however, are the long ZTE image acquisition time (Mecheter et al., 2020) and UTE/ZTE's limited availability, since they are not standard clinical sequences. Dixon-type pulse sequences, on contrary, are increasingly popular and are offered, nowadays, by nearly every manufacturer. To this end, (Torrado-Carvajal et al., 2019) employ a U-Net architecture with four Dixon images, where fat and water maps are explicitly used as additional inputs to 2-echo Dixon MRI. Alternatively, (Qi et al., 2020) utilize a stack of images from four different sequences (T1, T2, T1C, and T1DixonC-Water) as an input to their U-Net/GAN-based networks.

This paper makes several contributions. We introduce and formulate segmentation guided attention (SGA) mechanisms that can be adapted and used for different image-to-image translation tasks. We utilize the proposed SGA and its extended version (E-SGA) for the task of MRI-based pseudo-CT generation. The proposed attention mechanisms are based on coarse bone masks, which are derived from CT/pCT images in a fully automatic manner. These masks are only used during the training phase and thus are not required during the inference. We demonstrate that these attention networks are capable of producing superior quality pCTs, especially for bone regions.

2 METHODOLOGY

In this section, we first introduce and formulate the segmentation guided attention mechanism and its extended version. After that, we introduce U-Net-based and conditional Wasserstein GAN-based attention networks for the MRI-based pseudo-CT generation task.

2.1 Segmentation Guided Attention

The main goal of the image-to-image translation task is to transform input images from a source domain

X to a target domain Y via a mapping function $G: X \rightarrow Y$. Neural networks can learn such non-linear G mapping functions and, thus, can be used to generate (predict) $\hat{y}$ from the input image x, as follows:

$$\hat{y} = G(x) \tag{1}$$

Pixel-wise loss functions, such as L1 or L2 norm, are playing an essential role while training such image-to-image translation networks. L1 loss, for example, can be mathematically formalized as follows:

$$\mathcal{L}_1(y, \hat{y}) = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} \|y_{ij} - \hat{y}_{ij}\|_1 \tag{2}$$

where $\hat{y}$ denotes the output image of size $M \times N$ predicted by a neural network and y is the corresponding ground truth (GT) image. This loss function is calculated as the mean absolute error (MAE) between GT and prediction image over all (i, j) positions. As a consequence, while training, a neural network pays attention to the entire input image.

The proposed segmentation guided attention (SGA) loss, in contrast, focuses specifically on provided regions of interest (ROI) and helps image-to-image translation networks to achieve better performance for them. We formulate the SGA loss function as follows:

$$\mathcal{L}_{\text{SGA}}(y, \hat{y}, seg^y) = \gamma(seg^y) \cdot \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} seg_{ij}^y \cdot \|y_{ij} - \hat{y}_{ij}\|_1 \tag{3}$$

where seg^y is a binary mask derived from the target domain image y. Multiplication of seg_{ij}^y with its corresponding pixel-wise difference at position (i, j) penalizes an error only for the provided area of interest. We propose to use an additional weighting function $\gamma(.)$ which is dependent on seg^y and can be calculated as follows:

$$\gamma(seg^y) = \frac{M \cdot N}{(\sum_{i=1}^{M} \sum_{j=1}^{N} seg_{ij}^y) + \varepsilon} \tag{4}$$

where ε is needed to deal with the cases of division by 0, when seg_{ij}^y is 0 at each of its positions. This ensures that $\mathcal{L}_{\text{SGA}}$ stays on approximately the same range independent of ROI mask properties.

It can be noticed that the SGA loss is a generalization of the $\mathcal{L}_1$ loss from Equation 2. When the ROI mask contains 1's at each of its positions $\mathcal{L}_{\text{SGA}}$ is approximately equal to $\mathcal{L}_1$, since $\gamma(seg^y)$ is equal to 1.

The proposed segmentation guided attention loss can be used as a separate, alternative pixel-wise function or in combination with existing ones. Moreover, an alternative SGA loss can also be formulated based on the L2 norm.

2.2 Extended-SGA

The previously proposed $\mathcal{L}_{\text{SGA}}$ loss definition can be extended to $\mathcal{L}_{\text{E-SGA}}$, which considers two segmentation masks simultaneously. In addition to seg^y, which is based on a ground truth image, a segmentation $seg^{\hat{y}}$, based on a predicted image $\hat{y}$, is utilized. Thus, the extended-SGA (E-SGA) loss is defined as follows:

$$\mathcal{L}_{\text{E-SGA}}(y, \hat{y}, seg^y, seg^{\hat{y}}) = \gamma_E(seg^y, seg^{\hat{y}}) \cdot \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} \|seg_{ij}^y \cdot y_{ij} - seg_{ij}^{\hat{y}} \cdot \hat{y}_{ij}\|_1 \tag{5}$$

where $\gamma_E(.)$ is a weighting function:

$$\gamma_E(seg^y, seg^{\hat{y}}) = \frac{2 \cdot M \cdot N}{\left(\sum_{i=1}^{M} \sum_{j=1}^{N} (seg_{ij}^y + seg_{ij}^{\hat{y}})\right) + \varepsilon} \tag{6}$$

which is dependent on seg^y and $seg^{\hat{y}}$. In contrast to SGA, minimization of the extended loss during the training not only enforces networks to generate better quality seg^y regions but also constrains them to pay attention to $seg^{\hat{y}}$ regions.

It can be noticed that $\mathcal{L}_{\text{E-SGA}}$ is equivalent to $\mathcal{L}_{\text{SGA}}$ from Equation 3 when seg^y and $seg^{\hat{y}}$ are identical.

2.3 U-Nets with Additional SGA/E-SGA

The previously introduced loss functions can be utilized for MRI-based pseudo-CT generation task. Our proposed network topologies are based on the U-Net architecture, which was originally proposed to address semantic segmentation tasks in the medical image computing domain (Ronneberger et al., 2015). It is a fully convolutional network (FCN) which consists of a contracting (encoding) path followed by an expanding (decoding) path with additional skip connections between layers of the same size, as can be seen on the left part of Figure 1 and Figure 2. While the encoding path of the network behaves similar to a traditional convolutional neural network (CNN) and learns to extract hierarchical features from an input, the decoding path is responsible for reconstructing an output image gradually by adding more details at each following resolution level.

The proposed segmentation guided U-Net$_{\text{SGA}}$ approach is schematically shown in Figure 1. By propagating an input MR image through the network we generate a pCT image. The global $\mathcal{L}_1$ loss function (see Equation 2) compares the synthesized pCT to its corresponding ground truth CT image. In U-Net$_{\text{SGA}}$ we enforce the network to pay particular attention to the bone area via utilizing the previously introduced $\mathcal{L}_{\text{SGA}}$ loss term (see Equation 3). Therefore, we first multiply both generated pCT and its

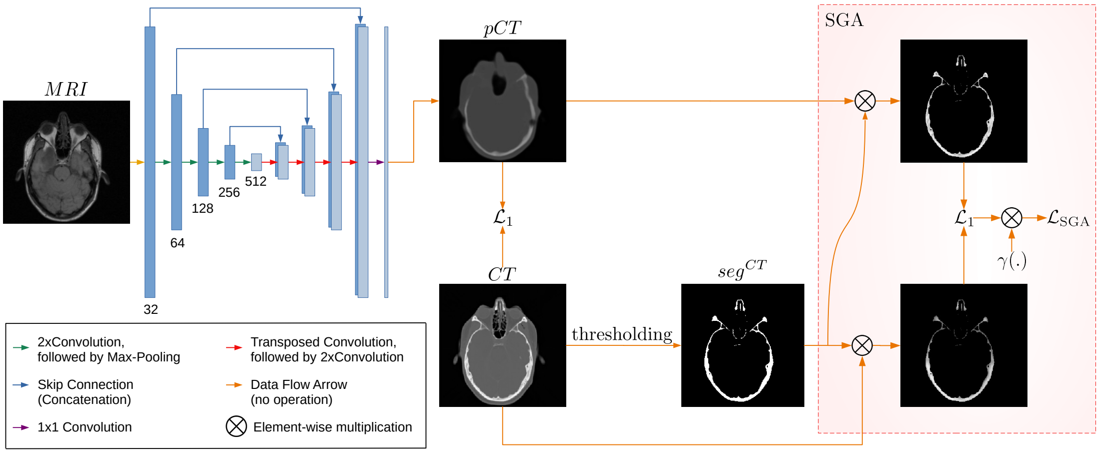

MRI
32
64
128
256
512
pCT
$\mathcal{L}_1$
CT
thresholding
seg^{CT}
SGA
$\mathcal{L}_1$
$\mathcal{L}_{\text{SGA}}$
$\gamma(.)$
2xConvolution, followed by Max-Pooling
Transposed Convolution, followed by 2xConvolution
Skip Connection (Concatenation)
Data Flow Arrow (no operation)
1x1 Convolution
⊗ Element-wise multiplication

Figure 1: The proposed segmentation guided U-Net$_{\text{SGA}}$ approach for MRI-based pseudo-CT generation. While $\mathcal{L}_1$ loss considers entire pCT and CT images, $\mathcal{L}_{\text{SGA}}$ loss concentrates exclusively on the bone area from CT images.

corresponding CT images with the binary bone mask (seg^{CT}, in Figure 1), and then calculate the difference between the resulting images. We propose to derive the required bone masks from CT via thresholding, which is described in more detail in Subsection 3.2. We extend the proposed SGA-based U-Net approach and introduce U-Net$_{\text{E-SGA}}$, which is schematically shown in Figure 2. This approach uses the additional $\mathcal{L}_{\text{E-SGA}}$ loss term (see Equation 6). In comparison to U-Net$_{\text{SGA}}$, this method requires additional seg^{pCT} bone masks from predicted pCTs. We propose to derive them in a similar manner as seg^{CT} images from CTs. Thus, the total objective for U-Net$_{\text{SGA/E-SGA}}$ consists of two loss functions and is defined as follows:

$$\mathcal{L}_{\text{Unet}_{\text{SGA/E-SGA}}} = \mathcal{L}_1 + \alpha \mathcal{L}_{\text{SGA/E-SGA}} \tag{7}$$

where α is a hyperparameter and can be used to control the relative importance of the two objectives.

2.4 Conditional Wasserstein GANs with Additional SGA/E-SGA

MRI-based pseudo-CT images can also be synthesized by using generative models, such as generative adversarial networks (GANs) (Goodfellow et al., 2014). The simplistic GAN framework consists of generator G and discriminator D networks. While the generator attempts to create data similar to those in the training set, the discriminator inspects it and tries to correctly identify whether it is real (from the training set) or fake (generated by G). Both networks are trained together by taking turns and using the adversarial loss function, which is defined as follows:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_y[\log(D(y))] + \mathbb{E}_x[\log(1 - D(G(x)))] \tag{8}$$

with x and y representing images from source and target domain, correspondingly. In other words, they are playing a minimax game against each other: while G attempts to maximize $\mathcal{L}_{\text{adv}}$ and thus fool D, the latter one tries to not be fooled by minimizing the adversarial loss.

Our proposed network topologies are based on the conditional Wasserstein GAN (cWGAN) training approach, which is an extension of Wasserstein GAN (Arjovsky et al., 2017). In contrast to traditional GANs, Wasserstein GAN can improve the stability of the learning process and can help to get rid of the mode collapse problem. This is achieved by using a different adversarial loss function, which approximates the Earth Mover's Distance, and changing the role of the discriminator from a binary classifier into a critic C, which predicts scores of how real or fake provided images are looking. For faster convergence and better preservation of structural information, MR image-based conditioning on the corresponding critic is utilized. Conditional Wasserstein GAN objective can be formalized as follows:

$$\mathcal{L}_{\text{cW}} = \mathbb{E}_{x,y}[C(x,y)] - \mathbb{E}_x[C(x, G(x))] \tag{9}$$

with x and y representing MR and CT images, correspondingly. Thus, while G tries to minimize $\mathcal{L}_{\text{cW}}$ against adversarial critic C, the latter one attempts to maximize the same objective. We use U-Net as the generator in our baseline cWGAN approach. For the critic network, CNN architecture is utilized, which is described in more detail in Subsection 3.2). Analogously to the U-Net-based extensions, we also introduce two extensions of the cWGAN. The proposed cWGAN$_{\text{SGA}}$ and cWGAN$_{\text{E-SGA}}$ approaches additionally include previously introduced segmentation guided attention loss terms while training their corresponding U-Net generator networks. For the

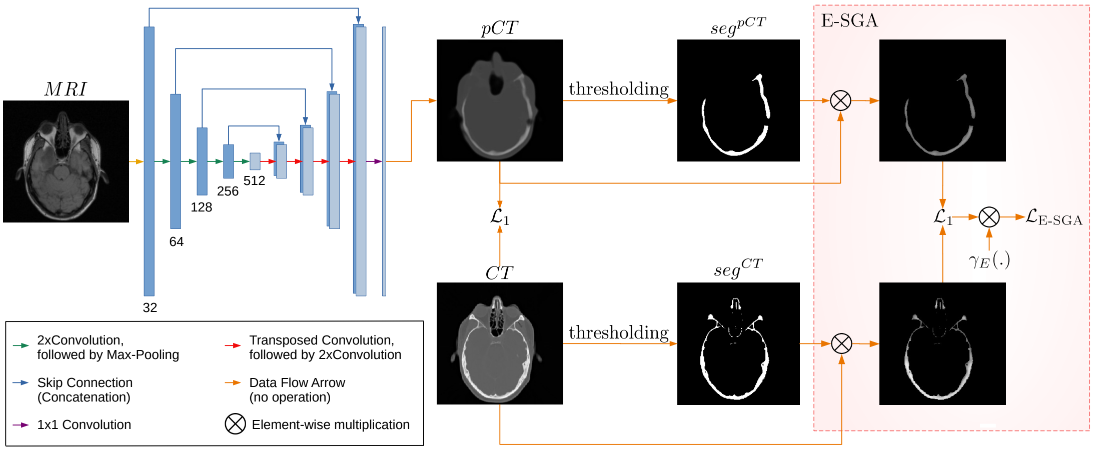

MRI
pCT
seg^pCT
E-SGA
thresholding
$\mathcal{L}_1$
$\mathcal{L}_{\text{E-SGA}}$
$\gamma_E(.)$
CT
seg^CT
thresholding
32
64
128
256
512
2xConvolution, followed by Max-Pooling
Transposed Convolution, followed by 2xConvolution
Skip Connection (Concatenation)
Data Flow Arrow (no operation)
1x1 Convolution
Element-wise multiplication

Figure 2: The proposed extended segmentation guided U-Net$_{\text{E-SGA}}$ approach for MRI-based pseudo-CT generation. While $\mathcal{L}_1$ loss considers the entire pCT and CT images, $\mathcal{L}_{\text{E-SGA}}$ loss concentrates exclusively on bone areas from both CT and pCT.

sake of shortness, we only depict cWGAN$_{\text{E-SGA}}$ approach in Figure 3. Thus, our final objectives for generators G and critics C of cWGAN$_{\text{SGA/E-SGA}}$ can be expressed as:

$$\begin{aligned}\mathcal{L}_G &= (\mathcal{L}_1 + \alpha\mathcal{L}_{\text{SGA/E-SGA}}) + \lambda\mathcal{L}_{cW}\\ \mathcal{L}_C &= \lambda\mathcal{L}_{cW}\end{aligned} \tag{10}$$

where λ denotes the weighting factor of the conditional Wasserstein GAN objective.

3 EXPERIMENTS

In this section, data set information including utilized registration procedure and following data preparation are introduced. After that, implementation details of proposed architectures and evaluation metrics are outlined.

3.1 Data

We evaluated the proposed architecture on the publicly available Retrospective Image Registration Experiment (RIRE) Project data set (West et al., 1997), that provides different modality images of the head area for 16 patients. For our experiments, we chose to use T1-weighted MR images of size 256×256 in combination with corresponding CTs of size 512×512. Both images fit into 12 bits data representation, meaning that CT's intensities range between -1024 HU and 3071 HU while MR's are ranging from 0 and 4095. Some of the CT volumes include the patient's table.

The RIRE project was originally designed to compare CT-MR and PET-MR registration techniques. Since the ground truth data is not provided, we used Mattes's mutual-information-based multi-resolution algorithm (Mattes et al., 2003) implemented in SimpleITK (Lowekamp et al., 2013; Yaniv et al., 2018) framework to register CT and MR volume pairs. During the registration procedure, CT was chosen as a fixed volume, while the corresponding MR volume was considered as a moving one. Optimization of mutual information between MR and CT volumes was done via gradient descent with a learning rate set to 0.01 value. While solving the optimization task, we utilized linear interpolation to deform the MR image. Since ground truth transformations are not available, we qualitatively inspected the registered MR-CT volume pairs and validated the utilized registration procedure.

All registered volumes were first brought to isotropic $1 \times 1 \times 1$ mm^3 voxel spacing. After that, depending on the obtained spatial resolution, cropping from the center point of the image or padding around its borders was utilized in order to achieve the same field of view. Finally, we resized the achieved MR and CT slices to the size of 256×256 pixels.

After the final visual inspection of the data set, we identified MR/CT slices that did not have valid CT/MR counterparts. These image pairs were located either on the top or at the bottom of the registered volumes when looking along the longitudinal axis. The main reason for their occurrence at these specific locations is not the low-quality registration, but rather the different fields of view of initial (not registered) MR and CT scans. Thus, we excluded them and ended up with 553 valid MR-CT image pairs in total.

3.2 Implementation Details

The experiments were conducted in a 4-fold cross-validation manner with one set used for each testing

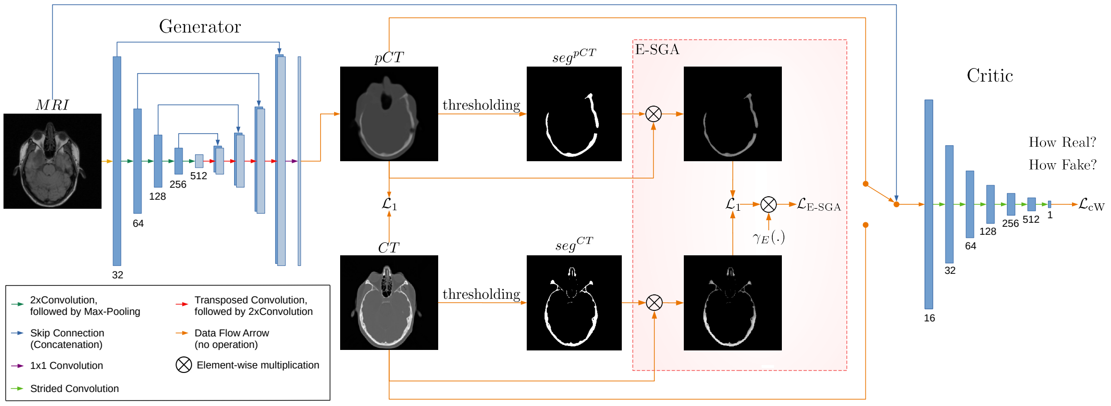

Generator
MRI
32
64
128
256
512
pCT
thresholding
seg^pCT
E-SGA
CT
thresholding
seg^CT
L_1
L_1
L_E-SGA
γ_E(.)
Critic
How Real?
How Fake?
16
32
64
128
256
512
1
L_cW
2xConvolution, followed by Max-Pooling
Transposed Convolution, followed by 2xConvolution
Skip Connection (Concatenation)
Data Flow Arrow (no operation)
1x1 Convolution
Element-wise multiplication
Strided Convolution

Figure 3: The proposed conditional Wasserstein GAN with extended segmentation guided attention (cWGAN$_{\text{E-SGA}}$) for MRI-based pseudo-CT generation.

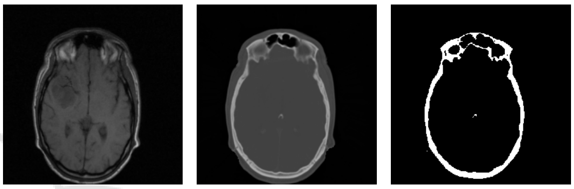
and three sets used for training and validation. During our experiments, we increased the number of training samples via data augmentation by means of random rotations (in the range of ±7.5 degrees), scaling (with factors in the range [1; 1.15]), and horizontal flipping (with the probability of 50%).

3.2.1 U-Nets

Baseline U-Net. In our baseline U-Net implementation, we started with 32 convolutional kernels in the first resolution level and we used two subsequent convolutional layers with zero padding at each resolution level. We chose to use kernels of size 5×5 at each convolutional layer and windows of size 2×2 for each following max pooling step. The number of convolutional filters was doubled at each following image resolution. In the expanding path, we used transposed convolutions with kernels of size 2×2 with a stride size of 2 pixels in both directions. At each upsample step the number of output features was reduced twice compared to the corresponding number of input channels. As a final layer, we used 1×1 convolution in order to get a single channel output image. We did not normalize MR and CT images as we did not observe any performance gain compared to the network trained with normalized inputs.

U-Nets with SGA/E-SGA. We used the same U-Net structure, as described previously, for both proposed segmentation guided approaches. For a given MRI-based pseudo-CT generation task we are particularly interested in an improved bone quality synthesis. Therefore, regions of interest required for SGA calculations should also represent bone areas.

Since manual annotation of bones from CTs is expensive and time consuming, we generated coarse bone masks by applying global threshold-based segmentation approach to ground truth CT images:



(a) MR. (b) CT. (c) Bone mask.

Figure 4: Exemplary MR, CT and coarse bone mask images for Patient 002 (slice #7).

$$seg^{y}_{ij} = \begin{cases} 1 & \text{if } y_{ij} > T \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where T is a threshold value.

We set the threshold value to 350 HU, which is approximately in the same range as stated in the literature (Buzug, 2009; Chougule et al., 2018). Figure 4 shows exemplary MR, CT and the obtained from CT coarse bone segmentation mask. In a similar manner and with the same threshold value we derived $seg^{\hat{y}}$, segmentation masks from synthesized $\hat{y}$ images, which are needed for E-SGA loss term calculations.

We set α to 1 when calculating the total loss functions for both U-Net$_{\text{SGA}}$ and U-Net$_{\text{E-SGA}}$ approaches. The intuition behind this choice is that we want to achieve a better synthesis quality for bone areas, however, we want to keep a good performance for other parts of the image as well.

Training U-Nets. We trained our U-Net models for 100 epochs with a learning rate of 0.01 using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as suggested by (Kingma and Ba, 2014). Mini-batches of size 16 were used for all our experiments.

3.2.2 Conditional WGANs

Baseline cWGAN. We used the previously introduced baseline U-Net as a generator in our baseline conditional Wasserstein GAN architecture. A hyperbolic tangent layer was added on top of it as the final activation layer. This ensured that the output intensities landed in the range between -1 and 1. In our critic architecture, we started with 16 convolutional kernels in the first resolution level. We chose to use kernels of size 4×4 with a stride of 2 pixels and 1 pixel padding in both directions, which allowed us to avoid max pooling layers, as was suggested by (Radford et al., 2015). We applied the LeakyReLU non-linear activation function after each convolutional layer. The number of filters was doubled at each following image resolution. In order to make the training procedure more stable, we applied a batch normalization at each resolution level, except for the first one. Strided convolutions were repeated until we obtained a single scalar value as an output per input image.

Input images were normalized to the range of -1 and 1 before being fed into our networks. This was achieved by applying the following min-max normalization equations for MR and CT images:

$$\mathrm{MR}_{\text{normalized}} = 2 \cdot \frac{\mathrm{MR}}{4095} - 1 \quad (12)$$

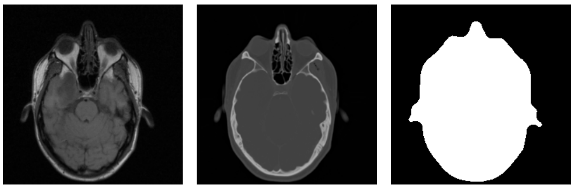
$$\mathrm{CT}_{\text{normalized}} = 2 \cdot \frac{\mathrm{CT} + 1024}{4095} - 1 \quad (13)$$

During the inference, synthesized pseudo-CT images were mapped back to the original intensity range based on the same normalization scheme, as in Equation 13.

CWGANs with SGA/E-SGA. The same baseline conditional WGAN architecture, as described previously, was utilized in our segmentation guided $\mathrm{cWGAN}_{\mathrm{SGA}}$ and $\mathrm{cWGAN}_{\mathrm{E\text{-}SGA}}$ approaches. Since their corresponding generator networks synthesize normalized pCT images, the previously utilized 350 HU threshold value was also mapped to the range between -1 and 1 using the Equation 13, which led to the value of -0.329.

With similar intuition behind as for segmentation guided U-Nets, we set α to 1 when calculating the total loss objectives of the generator networks.

Training cWGANs. We used a two times update rule for training our cWGANs as it was proposed by (Heusel et al., 2017). Our models were trained for 2000 epochs with learning rates of 0.0002 and 0.0004 for generators and discriminators, correspondingly. During the training, we utilized an additional gradient penalty as suggested by (Wu et al., 2018). Networks were optimized using the Adam optimizer and mini-batches of size 16.



(a) MR. (b) CT. (c) Head mask.

Figure 5: Exemplary MR, CT and head mask from MR images for Patient 002 (slice #2).

3.3 Evaluation Metrics

For evaluation, the obtained 2D pseudo-CT images were first stacked to build 3D volumes, and then only compared to the desired ground truth CT volumes. We chose mean squared error (MSE) and mean absolute error (MAE) as pixel-wise quality metrics.

Moreover, we calculated MSE and MAE values only for specific regions of interest, such as for bone and head areas ($\mathrm{MSE}_{\text{bone/head}}$ and $\mathrm{MAE}_{\text{bone/head}}$). Head masks were derived from MR images via Otsu's thresholding followed by subsequent morphological operations. First, a morphological opening with a disk-shaped structuring element of a 5 px radius was used to remove small artifacts from initial segmentations. Next, a closing operation (radius of 25 px) was utilized to fill the holes in nasal areas. Finally, a morphological dilation (radius of 5 px) was applied to slightly increase the total shape of the segments. An exemplary head mask is depicted in Figure 5.

To better evaluate the quality of generated pseudo-CT images we additionally calculated peak signal-to-noise ratio (PSNR) as follows:

$$PSNR(y,\hat{y}) = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE(y,\hat{y})}\right) \quad (14)$$

where MAX_I is the maximum possible intensity value. We also calculated the structural similarity index measure (SSIM) (Wang et al., 2004) between generated pCTs and corresponding CTs as follows:

$$SSIM(y,\hat{y}) = \frac{(2\mu_y\mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2)} \quad (15)$$

where $\mu_{\hat{y}}$ and μ_y denote mean values of pCT and CT images, respectively; σ_y and $\sigma_{\hat{y}}$ are corresponding variances; $\sigma_{y\hat{y}}$ is the covariance between pCT and CT image; and $C_1 = (k_1L)^2$ and $C_2 = (k_2L)^2$ are two variables to stabilize the division with weak denominators. Here, $k_1 = 0.01$, $k_2 = 0.03$, and $L = MAX_I$ represents the dynamic range of CT/pCT intensities.

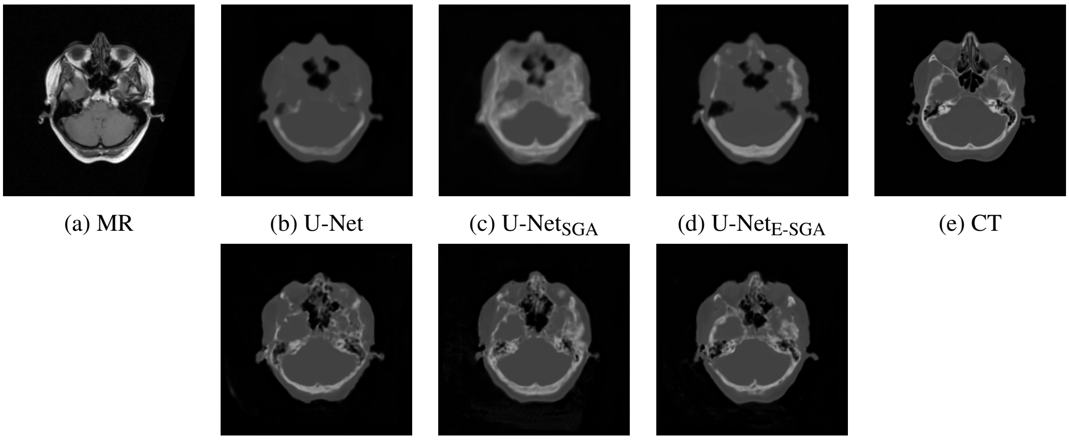
4 RESULTS

Table 1 shows averaged evaluation metrics for both baseline and segmentation guided architectures. Baseline U-Net implementation achieved MAE of 101 ± 35 HU and MSE of 69139 ± 27664 HU^2 when considering entire pseudo-CT images. Regarding the quality of image synthesis, it reached a PSNR of 24.3 ± 1.9 dB and SSIM of 79.6 ± 6.8 %. The proposed U-Net$_{\text{SGA}}$ performed significantly better than baseline U-Net when considering only the bone regions (column ”Bone Area” in Table 1). With 327 ± 46 HU and 180162 ± 56182 HU^2, over 45% reduction of MAE$_{\text{bone}}$ and around 66% reduction of MSE$_{\text{bone}}$ have been achieved. This improvement is the expected behavior since the SGA mechanism has been formulated in a such way, that it pays more attention to bone regions from ground truth images. However, U-Net$_{\text{SGA}}$’s performance is worse than baseline when considering the overall MAE and MSE metrics (columns ”Entire Image” and ”Head Area” in Table 1). This could be due to the suboptimal choice of α when calculating the total loss function of U-Net$_{\text{SGA}}$. Although U-Net$_{\text{E-SGA}}$ was not able to improve bone metrics as drastically as U-Net$_{\text{SGA}}$, values still appeared to be better than for the baseline architecture. Regarding the averaged metrics for the bone area, U-Net$_{\text{E-SGA}}$ yield 454 ± 88 HU and 345229 ± 112172 HU^2 for MAE and MSE, which is around 141 HU and 187466 HU^2 gain over the baseline U-Net implementation. In contrast to the network with SGA, U-Net$_{\text{E-SGA}}$ was able to pay attention to bones, however, was still able to retain the image quality for other parts of the image too. MSE and MAE values for entire images and head areas are in the same range as for the baseline U-Net. Exemplary pseudo-CT generation results for Patient 007 (slice #1) from baseline U-Net, U-Net$_{\text{SGA}}$ and from U-Net$_{\text{E-SGA}}$ are shown in Figure 7 (b-d). Thus, with regard to the proposed two segmentation guided mechanisms, U-Net$_{\text{E-SGA}}$ seems to yield better performance than baseline U-Net and U-Net$_{\text{SGA}}$. The same conclusion can be drawn from the difference images, which are shown in Figure 8 (first row).

The baseline conditional Wasserstein GAN performed slightly worse than the baseline U-Net when comparing the entire pCTs, however, the generated images are looking qualitatively better. As expected

Table 1: Averaged MAE, MSE, PSNR and SSIM metrics for baselines and segmentation guided networks. While MAE and MSE values (for entire images, head areas and bone areas) are given in HU and HU^2, PSNR and SSIM values are reported in dB and %, respectively.

	Entire Image		Head Area		Bone Area			
Name	MAE	MSE	MAE	MSE	MAE	MSE	PSNR	SSIM
U-Net	101±35	69139±27664	180±30	131393±38343	595±120	532695±198330	24.3±1.9	79.6±6.8
U-Net$_{\text{SGA}}$	128±34	83695±28792	257±42	192630±45684	327±46	180162±56182	23.2±1.5	77.5±6.2
U-Net$_{\text{E-SGA}}$	108±35	67528±27680	191±32	138309±37912	454±88	345229±112172	24.3±1.8	79.3±6.6
cWGAN	113±37	80507±31839	202±34	154101±42147	493±90	408417±131774	23.7±1.9	77.2±7.3
cWGAN$_{\text{SGA}}$	117±38	81997±31843	206±34	154975±40802	485±95	394859±129604	23.6±1.9	74.8±7.2
cWGAN$_{\text{E-SGA}}$	110±43	76425±35158	195±39	144978±44869	473±101	377988±133286	24.0±2.2	77.2±8.1



(a) MR (b) U-Net (c) U-Net$_{\text{SGA}}$ (d) U-Net$_{\text{E-SGA}}$ (e) CT

(f) cWGAN (g) cWGAN$_{\text{SGA}}$ (h) cWGAN$_{\text{E-SGA}}$

Figure 6: Exemplary MR, CT and MRI-based pseudo-CTs from U-Nets (b-d) and cWGANs (f-h) for Patient 007 (slice #1).

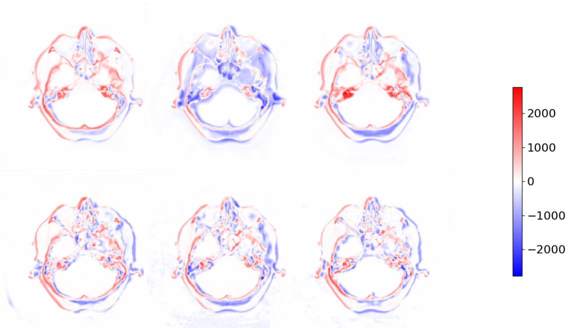


Figure 7: Differences images between CTs and corresponding pCTs for Patient 007 (slice #1). From upper left to bottom right: U-Net, U-Net$_{\text{SGA}}$, U-Net$_{\text{E-SGA}}$, cWGAN, cWGAN$_{\text{SGA}}$ and cWGAN$_{\text{E-SGA}}$. Red color represents underestimated regions, while blue one highlights overestimated areas.

and as it can be seen by comparing (b) and (f) images from Figure 7, cWGAN based approach was able to preserve more anatomical bone structures and to generate more realistic pCT images, due to the assistance of an additional critic network. While considering only ROIs, it improved MAE$_{\text{bone}}$ and MSE$_{\text{bone}}$ metrics. Thus, cWGAN yields 493 ± 90 HU and 408417 ± 131774 HU^2, which is around 102 HU and 124278 HU^2 gain over the baseline U-Net. With 485 ± 95 HU and 394859 ± 129604 HU^2 for bone regions, conditional WGAN with SGA performed slightly better than the baseline cWGAN, while its extended versions improved the metrics even further. Considering MAE$_{\text{bone}}$ and MSE$_{\text{bone}}$ metrics, cWGAN$_{\text{E-SGA}}$ yield approximately 20 HU and 30429 HU^2 improvement over the baseline cWGAN. Averaged SSIM values for cWGANs are in general slightly lower than for U-Nets, due to attempts of generative models to synthesize the patient's table. This can be confirmed by inspecting the difference images in Figure 8 (second row), where the table partially appears, although it does not exist in the corresponding ground truth image from Figure 7. The same reason lies behind the slightly higher errors for whole images (compared to the baseline U-Net and U-net$_{\text{SGA}}$). MAE$_{\text{head}}$ and MSE$_{\text{head}}$ values correlate with their corresponding metrics when considering the entire pCT images. Thus, comparing the obtained results to the desired GTs, significant qualitative and quantitative improvements for bone regions in favor of the E-SGA approach can be observed. It is worth mentioning that cWGAN$_{\text{SGA}}$ delivered slightly better overall results when compared to the corresponding baseline, whereas U-Net$_{\text{SGA}}$ was better only for bone regions. We argue that this was mainly achieved due to the contribution of the additional critic network which did not allow the corresponding generator to pay too much attention to the regions of interest.

5 CONCLUSION

In this work, we present 2D MRI-based pseudo-CT generation approaches with the additional segmentation guided attention mechanisms. We defined our total loss functions as a combination of global and local loss terms, whereas the second one enforces networks to pay particular attention to bone areas while generating pCT images. From the evaluation results, we observe that segmentation guided approaches yield improvements compared to baseline U-Net and conditional Wasserstein GAN architectures. As a result, more precise μ-maps for attenuation correction of PET image in PET/MR systems could be derived. Despite the apparent simplicity, segmentation guided attention allows networks to focus more on specific regions of interest, and as a consequence, achieve better performances for them. We believe that the proposed segmentation guidance can also be helpful when training cycle GAN-based architectures. We are currently in process of investigating this hypothesis.

REFERENCES

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Berker, Y., Franke, J., Salomon, A., Palmowski, M., Donker, H. C., Temur, Y., Mottaghy, F. M., Kuhl, C., Izquierdo-Garcia, D., Fayad, Z. A., et al. (2012). MRI-based attenuation correction for hybrid PET/MRI systems: a 4-class tissue segmentation technique using a combined ultrashort-echo-time/dixon mri sequence. *Journal of nuclear medicine*, 53(5):796–804.

Beyer, T., Townsend, D. W., Brun, T., Kinahan, P. E., Charron, M., Roddy, R., Jerin, J., Young, J., Byars, L., Nutt, R., et al. (2000). A combined PET/CT scanner for clinical oncology. *Journal of nuclear medicine*, 41(8):1369–1379.

Brady, Z., Taylor, M., Haynes, M., Whitaker, M., Mullen, A., Clews, L., Partridge, M., Hicks, R., and Trapp, J. (2008). The clinical application of PET/CT: a contemporary review. *Australasian Physics & Engineering Sciences in Medicine*, 31(2):90–109.

Burgos, N., Cardoso, M. J., Thielemans, K., Modat, M., Pedemonte, S., Dickson, J., Barnes, A., Ahmed, R., Mahoney, C. J., Schott, J. M., et al. (2014). Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies. *IEEE transactions on medical imaging*, 33(12):2332–2341.

Buzug, T. M. (2009). Computed tomography: from photon statistics to modern cone-beam CT.

Chougule, V., Mulay, A., and Ahuja, B. (2018). Clinical case study: spine modeling for minimum invasive spine surgeries (MISS) using rapid prototyping. *Bone (CT)*, 226:3071.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Han, X. (2017). MR-based synthetic CT generation using a deep convolutional neural network method. *Medical physics*, 44(4):1408–1419.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Keereman, V., Mollet, P., Berker, Y., Schulz, V., and Vandenberghe, S. (2013). Challenges and current methods for attenuation correction in PET/MR. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 26(1):81–98.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Leynes, A. P., Yang, J., Wiesinger, F., Kaushik, S. S., Shanbhag, D. D., Seo, Y., Hope, T. A., and Larson, P. E. (2018). Zero-echo-time and dixon deep pseudo-ct (ZeDD CT): direct generation of pseudo-CT images for pelvic PET/MRI attenuation correction using deep convolutional neural networks with multiparametric MRI. *Journal of Nuclear Medicine*, 59(5):852–858.

Liu, F., Jang, H., Kijowski, R., Bradshaw, T., and McMillan, A. B. (2017). Deep learning MR imaging–based attenuation correction for PET/MR imaging. *Radiology*, 286(2):676–684.

Lowekamp, B. C., Chen, D. T., Ibáñez, L., and Blezek, D. (2013). The design of SimpleITK. *Frontiers in neuroinformatics*, 7:45.

Mattes, D., Haynor, D. R., Vesselle, H., Lewellen, T. K., and Eubank, W. (2003). PET-CT image registration in the chest using free-form deformations. *IEEE transactions on medical imaging*, 22(1):120–128.

Mecheter, I., Alic, L., Abbod, M., Amira, A., and Ji, J. (2020). MR image-based attenuation correction of brain PET imaging: Review of literature on machine learning approaches for segmentation. *Journal of Digital Imaging*, 33(5):1224.

Nie, D., Cao, X., Gao, Y., Wang, L., and Shen, D. (2016). Estimating CT image from MRI data using 3D fully convolutional networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 170–178. Springer.

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., and Shen, D. (2017). Medical image synthesis with context-aware generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention*, pages 417–425. Springer.

Ollinger, J. M. and Fessler, J. A. (1997). Positron-emission tomography. *Ieee signal processing magazine*, 14(1):43–55.

Paans, A. M. (2006). Positron emission tomography.

Paulus, D. H., Quick, H. H., Geppert, C., Fenchel, M., Zhan, Y., Hermosillo, G., Faul, D., Boada, F., Friedman, K. P., and Koesters, T. (2015). Whole-body PET/MR imaging: quantitative evaluation of a novel model-based MR attenuation correction method including bone. *Journal of Nuclear Medicine*, 56(7):1061–1066.

Qi, M., Li, Y., Wu, A., Jia, Q., Li, B., Sun, W., Dai, Z., Lu, X., Zhou, L., Deng, X., et al. (2020). Multi-sequence MR image-based synthetic CT generation using a generative adversarial network for head and neck mri-only radiotherapy. *Medical physics*, 47(4):1880–1894.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Torrado-Carvajal, A., Vera-Olmos, J., Izquierdo-Garcia, D., Catalano, O. A., Morales, M. A., Margolin, J., Soricelli, A., Salvatore, M., Malpica, N., and Catana, C. (2019). Dixon-VIBE deep learning (DIVIDE) pseudo-CT synthesis for pelvis PET/MR attenuation correction. *Journal of nuclear medicine*, 60(3):429–435.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

West, J., Fitzpatrick, J. M., Wang, M. Y., Dawant, B. M., Maurer Jr, C. R., Kessler, R. M., Maciunas, R. J., Barillot, C., Lemoine, D., Collignon, A., et al. (1997). Comparison and evaluation of retrospective intermodality brain image registration techniques. *Journal of computer assisted tomography*, 21(4):554–568.

Wolterink, J. M., Dinkla, A. M., Savenije, M. H., Seevinck, P. R., van den Berg, C. A., and Išgum, I. (2017). Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer.

Wu, J., Huang, Z., Thoma, J., Acharya, D., and Van Gool, L. (2018). Wasserstein divergence for gans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 653–668.

Yaniv, Z., Lowekamp, B. C., Johnson, H. J., and Beare, R. (2018). SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of digital imaging*, 31(3):290–303.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.