Impact of the training loss in deep learning–based CT reconstruction of bone microarchitecture

Théo Leuliet, Voichita Maxim, Françoise Peyrin, Bruno Sixou

Impact of the training loss in deep learning based CT reconstruction of bone microarchitecture

Théo Leuliet, Voichița Maxim, Françoise Peyrin, Bruno Sixou
Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621, LYON, France
e-mail: name.surname@creatis.insa-lyon.fr

Author to whom correspondence should be addressed: Théo Leuliet, email: theo.leuliet@creatis.insa-lyon.fr

Abstract

Purpose: Computed tomography (CT) is a technique of choice to image bone structure at different scales. Methods to enhance the quality of degraded reconstructions obtained from low-dose CT data have shown impressive results recently, especially in the realm of supervised deep learning. As the choice of the loss function affects the reconstruction quality, it is necessary to focus on the way neural networks evaluate the correspondence between predicted and target images during the training stage. This is even more true in the case of bone microarchitecture imaging at high spatial resolution where both the quantitative analysis of Bone Mineral Density (BMD) and bone microstructure are essential for assessing diseases such as osteoporosis. Our aim is thus to evaluate the quality of reconstruction on key metrics for diagnosis depending on the loss function that has been used for training the neural network.
Methods: We compare and analyze volumes that are reconstructed with neural networks trained with pixelwise, structural and adversarial loss functions or with a combination of them. We perform realistic simulations of various low-dose acquisitions of bone microarchitecture. Our comparative study is performed with metrics that have an interest regarding the diagnosis of bone diseases. We therefore focus on bone-specific metrics such as BV/TV, resolution, connectivity assessed with the Euler number and quantitative analysis of BMD to evaluate the quality of reconstruction obtained with networks trained with the different loss functions.
Results: We find that using L_1 norm as the pixelwise loss is the best choice compared to L_2 or no pixelwise loss since it improves resolution without deteriorating other metrics. VGG perceptual loss, especially when combined with an adversarial loss, allows to better retrieve topological and morphological parameters of bone microarchitecture compared to SSIM. This however leads to a decreased resolution performance. The adversarial loss enchances the reconstruction performance in terms of BMD distribution accuracy.
Conclusions: In order to retrieve the quantitative and structural characteristics of bone microarchitecture that are essential for post-reconstruction diagnosis, our results suggest to use L_1 norm as part of the loss function. Then, trade-offs should be made depending on the application: VGG perceptual loss improves accuracy in terms of connectivity at the cost of a deteriorated resolution, and adversarial losses help better retrieve BMD distribution while significantly increasing the training time.
Keywords: Bone structure, deep learning, low-dose micro CT, tomographic reconstruction, training loss.

I. Introduction

Tomographic reconstruction is a challenging task, not only due to the physical limitations of scanners, but also because of the need for reducing the radiation dose during the acquisition. The efficiency of analytical algorithms like the Filtered BackProjection (FBP) is suboptimal in the low-dose imaging context, due to their sensitivity to small number of projections and noisy data. Iterative methods [1,2], especially when they include a regularization term [3,4], are generally efficient to overcome the noise issue but they require a high computation time and tuning the regularization parameters for every reconstruction might be demanding.

Deep learning based methods have the potential to overcome those limits, since the associated algorithms are adaptive and fast in most cases. Especially, the reconstruction performance is enhanced by the possibility to learn from ground truth data; in other words, given acquisition data, neural networks are able to produce volumes similar to ones that could be obtained in a high-dose setting.

There are plenty of ways to design a neural network for retrieving an image given a set of low-dose projections. One can train a network in an end-to-end manner [5], performing the three following steps within a single architecture: first correcting the projections, then mapping them onto the image domain, and finally enhancing the obtained image to match the high-dose version. The benefits of including the projections within the network were demonstrated in [6], especially for sparse-view Computed Tomography (CT) data. One can instead compute the FBP from the low-dose projections to feed a network that removes noise and artifacts which remain in the obtained image: different structures can be chosen to generate the desired reconstruction, such as U-NET [7] in [8] or residual encoder-decoder networks in [9]. Architectures based on unrolled iterative algorithms have also demonstrated impressive performance on the reconstructions quality [10–13]. Moreover, using 3D networks instead of working on slices has shown to be efficient since it helps capture spatial information across slices [14,15].

In addition to the architecture of the network, the loss used to measure the prediction error during training has a major impact on the quality of reconstruction. Two networks with identical architectures can produce very different reconstructions depending on the way they have been trained. Pixelwise losses such as Mean Squared Error (MSE, or L_2 norm) and Mean Absolute Error (MAE, or L_1 norm) allow to ensure that the pixel values in the reconstructed image - representing e.g the attenuation of the studied object in Hounsfield Units (HU) - are close to the corresponding ones in the ground truth. They might however result in oversmoothed solutions - especially for MSE - leading to blurring near the edges and loss of structural information. A solution to overcome these limitations is to consider losses such as the Structural SIMilarity index (SSIM, [16]) or the VGG perceptual loss [17] that both evaluate the similarity between images in terms of structural features. Adversarial losses from Generative Adversarial Networks (GAN) [18] constitute

a third category that appears as a hybrid method between pixelwise and structural losses. They make use of a trainable network called a discriminator and allow to roughly evaluate how far the predicted images are from the target domain of high-dose images. Both pixel values and structural information are thus essential to minimize such a loss.

If structural and adversarial losses were proven to be efficient for general computer vision tasks, the application to computed tomography, i.e the ability of such losses to better retrieve structural and attenuation information on deep networks based tomographic reconstructions, is not straightforward. Combining adversarial and perceptual losses in [19] showed impressive results on the reconstruction of structural details in abdominal CT images. Also, a WGAN-VGG-MSE was used in [20] for the particular case of PET image denoising. The effect of loss functions on low-dose CT images was studied in [21] for the denoising of low dose CT images. There is however a need to perform a comprehensive comparison study with task-specific metrics that are relevant for diagnosis.

We consider the particular case of bone microarchitecture imaging, for which X-ray CT is a powerful tool [22–25]. The diagnosis of bone diseases is indeed largely correlated to the quality of the reconstructed image in terms of quantitative and structural information. Both the diagnosis of osteoporosis [26] and the prediction of mechanical failure in cancellous bone [27] depend on the quality of the reconstruction in terms of bone microstructure. The diagnosis of osteoporosis also relies on Bone Mineral Density (BMD) [28], which can be retrieved with the values of attenuation in HU, hence the need for quantitative information to be accurate in the reconstructed volume. In the low-dose context, the correct reconstruction of both bone microstructure and BMD is a very challenging task [29]. When choosing metrics to assess the performance of neural networks for reconstructing such images, it is thus necessary to take both structure and density into account.

In this work we assess the impact of the loss function in the context of the reconstruction of low dose bone microarchitecture CT images. To this aim we study combinations of pixelwise, structural and adversarial losses and evaluate the benefits and drawbacks from each of these losses considering metrics that are relevant for the diagnosis of bone diseases. To conduct the study, we consider the simplest task that consists in enhancing the quality of a FBP image obtained from low-dose projections with a deep convolutional neural network (CNN) trained on high-dose/low-dose paired images. This work is expected to give some insight on the impact of the loss function in the context of tomographic reconstruction and provide a guide in selecting the appropriate loss function when using neural networks to reconstruct bone microarchitecture.

The paper is organized as follows; in Section II. we present the different losses that we use when training

the networks along with the evaluation criteria that allow to assess the quality of bone microstructure reconstruction. In Section III. we show our experiments and results on μ-CT bone data where we simulate different realistic settings of radiation dose in the projections data, for both training and evaluation. Next in Section IV. we analyze those results and discuss the relevance of the different training losses for bones microstructure reconstruction, and finally in Section V. we conclude on the role of the training loss design when reconstructing data where both structure and density are important for medical interpretation.

II. Methods

II.A. Model

Let $y \in Y$ be the image reconstructed with FBP from low-dose projections, Y being the space of these low-dose FBP reconstructions. Let $x \in X$ be the corresponding high-dose image, where X is the target space of images obtained in the high-dose setting. The aim is to find the reconstruction operator G_θ such that

$$x = G_\theta(y) \tag{1}$$

where G_θ is a deep CNN parameterized by θ. Note that we talk about a reconstruction operator for simplicity here even though y does not correspond to a projection; our aim is to build a simple model to focus on the impact of the training loss so we proceed to the mapping between the projections domain and the image domain with the FBP, thus not inside the neural network. In what follows we consider paired high-dose and low-dose images (x, y) and a loss function L such that

$$\theta^* \in \underset{\theta}{\operatorname{argmin}} \underset{(x,y)\sim\mu}{\mathbb{E}} [L(G_\theta(y), x)] \tag{2}$$

where μ is the joint distribution of (x, y) and parameters of G_θ are trained according to the loss function L with backpropagation. In practice, empirical expectations obtained with training data are considered.

II.B. Training losses

We distinguish three types of loss functions. Pixelwise losses compare each pixel of the predicted image with the corresponding pixel in the ground truth and the average error is then considered. Structural losses compare statistics or features from the prediction and the ground truth in order to match the way human eye evaluates similarities between images. Finally, adversarial losses allow to assess whether the predicted image belongs to the distribution of ground truths or not, i.e if the network is producing an image that could be reconstructed from a high-dose acquisition in our case.

II.B.1. Pixelwise losses

A common way to compare the prediction from the generator G_θ is the mean squared error or L_2 loss

$$L_{\text{MSE}}(G_\theta(y), x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - [G_\theta(y)]_i)^2 \tag{3}$$

where n is the total number of pixels in the image and subscript i denotes pixel values of x. Another widely used loss that performs operations between pixels is the mean absolute error or L_1 loss

$$L_{\text{MAE}}(G_\theta(y), x) = \frac{1}{n}\sum_{i=1}^{n}|x_i - [G_\theta(y)]_i|. \tag{4}$$

In both cases, pixels are considered independently and outliers - for instance one pixel value $[G_\theta(y)]_i$ that is very far from x_i - are largely penalized. MSE might lead to oversmoothing in the reconstructions, but it is generally efficient to retrieve flat areas. For sharp objects, MAE is often preferred since less oversmoothing is observed in the solutions. This can be explained by the fact that MSE corresponds to a Gaussian statistic of the noise in the likelihood in a Baysesian framework, while MAE corresponds to a more sparse Laplace prior. Note that in both cases structural features are not taken into account, thus it can be expected that the generated images are accurate in terms of density but might present inconsistent anatomical objects.

II.B.2. Structural losses

To ensure the correctness of the reconstruction in terms of anatomical structure, a solution can be to train networks with loss functions that compare images with respect to aggregated statistics or features within each of them. SSIM was developed in [16] to measure the similarity of two images with respect to the structure rather than operating pixel by pixel. The corresponding loss function can be written as

$$L_{\text{SSIM}}(G_\theta(y), x) = -\frac{(2\mu_g\mu_x + c_1)(2\sigma_{gx} + c_2)}{(\mu_g^2 + \mu_x^2 + c_1)(\sigma_g^2 + \sigma_x^2 + c_2)} \tag{5}$$

where μ_g is the average of $G_\theta(y)$, μ_x is the average of x, σ_g^2 and σ_x^2 are their corresponding variance, σ_{gx} is the covariance of $G_\theta(y)$ and x, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ with L the dynamic range of the pixel values that is 1 in our case - due to rescaling - and $k_1 = 0.01$ and $k_2 = 0.03$ as we considered standard values. In practice, SSIM index is computed on sliding Gaussian windows of size 11×11 with standard deviation $\sigma = 1.5$ and the considered value is the average of the local similarities.

Perceptual losses can also be computed by comparing features within the two images. Those features can be obtained by feeding a trained neural network, and this is the idea of VGG loss in [30] that writes

$$L_{\text{VGG}}(G_\theta) = \frac{1}{n}\mathop{\mathbb{E}}_{(x,y)\sim\mu}[||VGG(G_\theta(y)) - VGG(x)||_2^2] \tag{6}$$

where VGG is the 16^{th} output of the VGG-19 model [31] that performs classification of natural images. It is shown in [17] that such a loss better suits human perception compared to pixelwise losses.

Whether it is SSIM or VGG, using such losses should allow to retrieve relevant structures in the bone reconstruction since the network specifically learns to minimize the difference in terms of structural features during the training stage. However in SSIM the pixel values are only considered with aggregated statistics and for VGG loss there is no consideration at all given to pixel values. Those losses could thus lead to networks that correctly transcribe the bone microstructure but where the BMD correspondence is missing.

II.B.3. Adversarial loss

Among the range of possible loss functions that can be used to train a neural network, adversarial losses appear as an hybrid method between structural and pixelwise losses. We consider Wasserstein GANs [32] with gradient penalty [33] - that we denote as WGAN in what follows for simplicity - as the basis for such an adversarial loss. The corresponding loss function is given by

$$\begin{aligned} L_{\text{WGAN}}(D_w, G_\theta) = &\mathop{\mathbb{E}}_{x\sim P_x}[D_w(x)] - \mathop{\mathbb{E}}_{y\sim P_y}[D_w(G_\theta(y))] \\ &- \lambda \mathop{\mathbb{E}}_{\hat{x}\sim P_{\hat{x}}}[(||\nabla_{\hat{x}} D_w(\hat{x})||_2 - 1)^2] \end{aligned} \tag{7}$$

where D_w is a neural network - called a discriminator - that is simultaneously trained to maximize L_{WGAN}, P_y and P_x are the empirical distributions of respectively low-dose FBP data and high-dose images, $\hat{x} \sim P_{\hat{x}}$ are sampled along straight lines between real high-dose images and generated ones. Finally, λ is the weighting term for the gradient penalty, that we fix to 10 which is a standard value. The aim of the discriminator is to evaluate whether the generated image belongs to the high-dose images distribution or not. Note that contrary to what is common for GANs, here the network is not stochastic. Rather, G_θ aims to be a mapping from the distribution of low-dose images onto the one of high-dose images. Here the training loss evaluates whether the generated image belongs to X, but it does not indicate whether the content corresponds to the input low-dose FBP. As a consequence, a content loss should be added to ensure that x matches its low dose version y. In practice during the training stage both the generator and the discriminator weights are updated alternately, to progressively allow the generator to produce images similar to the ones from the high-dose distribution. As the WGAN loss evaluates the quality of the generated image thanks to a probability distribution model, it is reasonable to think that both BMD and bone microstructure are taken into account in that case, and the impact of the adversarial loss should be studied accordingly.

II.C. Comparative study

We propose to combine different loss functions with weighting parameters to form a more complex cost function, with the hope to benefit from the strengths of each part. There are 31 possible combinations from the 5 losses that we presented in the previous section. As mentioned in section I., our aim is to assess the

Network	Loss function
CNN-L_1	$L_{\text{MAE}}(G)$
CNN-L_2	$L_{\text{MSE}}(G)$
CNN-SSIM	$L_{\text{SSIM}}(G)$
CNN-VGG	$L_{\text{VGG}}(G)$
CNN-SSIM-L_1	$L_{\text{SSIM}}(G) + \lambda_1 L_{\text{MAE}}(G)$
CNN-VGG-L_1	$L_{\text{VGG}}(G) + \lambda_1 L_{\text{MAE}}(G)$
CNN-VGG-L_2	$L_{\text{VGG}}(G) + \lambda_1 L_{\text{MSE}}(G)$
WGAN-L_1	$L_{\text{WGAN}}(D,G) + \lambda_1 L_{\text{MAE}}(G)$
WGAN-VGG	$L_{\text{WGAN}}(D,G) + \lambda_1 L_{\text{VGG}}(G)$
WGAN-SSIM-L_1	$L_{\text{WGAN}}(D,G) + \lambda_1 L_{\text{SSIM}}(G) + \lambda_2 L_{\text{MAE}}(G)$
WGAN-VGG-L_1	$L_{\text{WGAN}}(D,G) + \lambda_1 L_{\text{VGG}}(G) + \lambda_2 L_{\text{MAE}}(G)$
WGAN-VGG-L_2	$L_{\text{WGAN}}(D,G) + \lambda_1 L_{\text{VGG}}(G) + \lambda_2 L_{\text{MSE}}(G)$

Table 1: Tested networks and their training loss function. λ_1 and λ_2 are weighting parameters.

impact of each category of loss functions, and potentially find the most relevant one from each category. Combining losses from the same category - for instance L_1 and L_2 - is thus not interesting for us. This leaves us with 17 potential combinations, and even only 16 if we do not consider WGAN alone as we mentioned that it should be used with a content loss. For the sake of clarity, we will report in this work results for 12 of these combinations - shown in Table 1 - as they allow to answer the questions raised in this study, the 4 other combinations adding no further insights for our problem.

Table 1 highlights the potential drawback of using a complex loss function : adding weighting parameters that need to be tuned during the training stage increases the computation time for a fixed hyper-parameter optimization strategy. For instance, a grid search strategy consisting in testing n different values for each hyper-parameter requires n^2 times more trainings for WGAN-SSIM-L_1 (2 weighting parameters) compared to CNN-SSIM (no weighting parameter).

II.D. Evaluation criteria

The assessment of bone microachitecture, which is important in the context of the diagnosis of bone diseases, relies on a number of morphological and topological parameters that are extracted from the images. This is performed on images that have been segmented to differentiate between areas corresponding to bones and the rest of the image. For this, we post-process the reconstructions with Otsu segmentation [34] and compute metrics on the obtained segmented volumes.

First, the ratio between the bone volume and the total volume (BV/TV) is a key information for mechanical failure prediction [27]. BV/TV is thus not only a metric that allows to evaluate the performance of the methods, but it is also a relevant feature used for diagnosis.

Also, studying connectivity in the bone volume allows to get insight on the bone microarchitecture [35]. Connectivity can be determined in an unbiased manner by the Euler number. We evaluate it to assess the

fidelity of the reconstruction in terms of structure. In actual medical settings, this is performed considering the 3D volume but since in our study the networks are built for 2D slices, we focus on the comparison for the 2D Euler number. In the 2D case, computation of this number amounts to counting the difference between the number of objects and the number of holes that are perceived in the image obtained after segmentation. We show results considering 4 neighboring pixels for the objects counts (4-Connectivity), but similar results are observed with 8 neighboring pixels (8-Connectivity). Computation of the Euler number is performed with the *measure* module of *scikit-image* library in Python.

The ability to reconstruct thin details can be assessed with the resolution of the obtained volume. In [36], the authors introduced the Fourier Ring Correlation (FRC) as a metric to estimate the resolution of a reconstruction. The idea is to compute the correlation between an estimated 2D image f with respect to some ground truth g in the Fourier domain as

$$FRC_{f,g}(R_i) = \frac{\sum_{r\in\mathcal{C}(R_i)} |\Re(\hat{f}^*(r)\hat{g}(r))|}{\sqrt{\sum_{r\in\mathcal{C}(R_i)} |\hat{f}(r)|^2 \sum_{r\in\mathcal{C}(R_i)} |\hat{g}(r)|^2}} \tag{8}$$

where R_i is the radius of the ring $\mathcal{C}(R_i)$ in the Fourier domain within which the correlation is computed, $\hat{f}$ is the Fourier transform of f, $\hat{f}^*$ denotes the conjugate of $\hat{f}$, and $\Re$ denotes the real part. The metric aims at measuring the ability of the reconstruction to recover information at a certain frequency level. The resolution ρ of the reconstruction can then be determined as

$$\rho = \frac{1}{R_{FRC(R)\leq\tau(R)}} \tag{9}$$

where $R_{FRC(R)\leq\tau(R)}$ is the radius for which the FRC is lower than a threshold τ. This threshold may depend on the radius and in [36] it is computed as

$$\tau(R) = \frac{2}{\sqrt{\frac{N_p(R)}{2}}} \tag{10}$$

with R the radius in the Fourier domain and $N_p(R)$ the number of pixels contained within the corresponding ring.

If structure and resolution are key information for a correct diagnosis, BMD is also important to assess bones weaknesses and for the diagnosis of osteoporosis [26]. An accurate reconstruction should not only match the right structure of a bone area and have a high resolution, but it should also match the correct values of density in Hounsfield Units (HU). To this purpose, we study the flattened HU distribution of the voxels that are reconstructed for each method. Quantitative analysis of the differences in terms of voxel values can be performed by computing the Wasserstein-1 distance - see [37] - between the 1D distributions obtained when considering each voxel of the volume as one realization of a random variable. This distance

writes in one dimension

$$W_1(\phi_1, \phi_2) = \inf_{\pi \in \Gamma(\phi_1, \phi_2)} \int_{\mathbb{R}\times\mathbb{R}} |u - v| d\pi(u, v) \tag{11}$$

where ϕ_1 and ϕ_2 are the two considered 1D distributions and Γ is the set of joint distributions (ϕ_1, ϕ_2). Note here that this distance does not correspond to the one that is approximated with neural networks. The latter considers distributions over n-dimensional vectors, n being the number of pixels in an image. Rather in (11), u and v are the distributions obtained when taking n realizations of a 1-dimensional random variable, which are drawn from the distributions of the voxels taken in either the ground truth or the estimated volume. The role of such a metric is to assess that the voxel distribution correctly represents BMD values across the volume. This is useful - conditionally on the fact that the structure is correctly transcribed - since BMD analysis is a way to perform diagnosis for potential bone diseases. For instance, the structure could be well retrieved but with BMD values that are completely shifted towards higher or lower HU which would result in incorrect analysis. Also, using the Wasserstein 1-distance instead of MAE allows not to penalize reconstructions with accurate BMD but with a structure that is slighlty shifted compared to the ground truth. We compute this Wasserstein-1 distance between HU distributions with the *Stats* module of *Scipy* library in Python.

III. Experiments and results

III.A. Dataset

The ground truth data consist of volumes of human radius and tibia structures obtained on a SCANCO μ-CT 100 with a 24-μm voxel size. The training dataset is composed of ten volumes from different patients. Two volumes from two other patients are considered for evaluating the methods; the networks are not trained with those two patients data and the hyperparameters are not tuned according to these data. These two evaluation volumes have respectively a number of slices, height and width of $164 \times 882 \times 752$ and $194 \times 466 \times 372$ voxels. The ground truth training data are illustrated in Figure 1 and the volumes for evaluation are illustrated in Figure 2.

Denoting by ρ the ground truth volume, projections $p(\rho)$ were computed with the parallel Radon transform from these volumes. This was performed with ASTRA Toolbox [38] in Python. To simulate low dose data, we first consider 400 projections corresponding to approximately 50 % of the total number of projections in the high-dose setting. We consider a source intensity I_0 of 10000 photons per detector pixel, and simulate the received intensity I at each detector pixel as $I = \text{Poisson}(\frac{I_0}{K} e^{-p(\rho)})$, with K a parameter that we vary to simulate different amounts of dose, similarly to [39]. For instance, $K = 10$ corresponds to 5% of the dose, since we already consider half of the projections. Then, the noisy projections are taken as

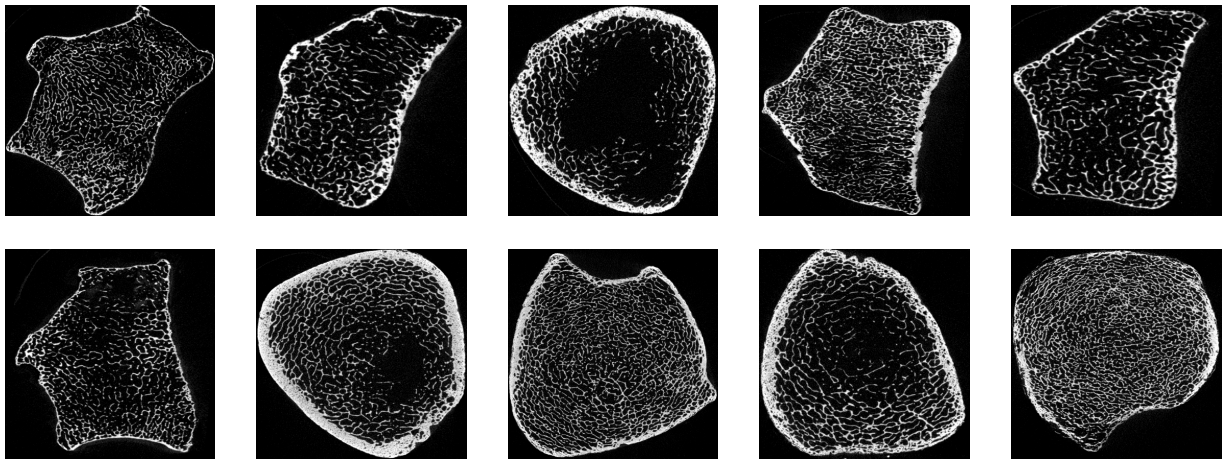


Figure 1: Volumes used for training. Each of the 10 volumes has between 152 and 248 slices, whose size ranges from 628×508 to 1068×928 voxels. Window size is $[-1000, 3000]$ HU.

$\tilde{p} = \ln \frac{I_0}{KI} + n$ with n an additive zero-mean Gaussian noise with standard deviation 1% of the first term mean value. Finally, we compute the FBP of $\tilde{p}$ with a Hann filter (cutoff 0.4) and consider it as the noisy input data of the network. In the training data, we varied K to simulate between 5% and 50% of the upper limit of the radiation dose. Also, note that data are normalized when fed into neural networks. The normalization is simply performed by dividing the images by a factor $\rho_{\max}$ which was chosen so that all data lie between 0 and 1.

III.B. Networks details

In all models, the generator is a 16-layer Convolutional Neural Network (CNN) with 128 filters in each layer, except for the last layer which has only one filter since the output is the generated image. Worse performance was observed with fewer layers. A similar deep CNN was used in [19].

For WGAN based networks, we use the same discriminator structure as in [19]. For both the discriminator and the generator, Leaky ReLU activations are used with parameter 0.3 and He initialization [40], except for the output of the discriminator that has no activation function. Zero padding is applied for every layer. Optimization is performed with Adam algorithm [41] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Training is ran on 7,000 steps, which approximately corresponds to 3 epochs. The gradient weighting parameter λ is fixed to a default value of 10 as in [33].

For a fair comparison, hyperparameters (HP) namely the kernel size, batch size, learning rate - initial and final since we use exponential decay -, number of generator updates, λ_1 and λ_2 are all optimized for every single network, on a validation set that is obtained by taking 20% of the slices from the 10 training volumes. Two stages of HP optimization are performed, the second stage allowing to zoom in the range of

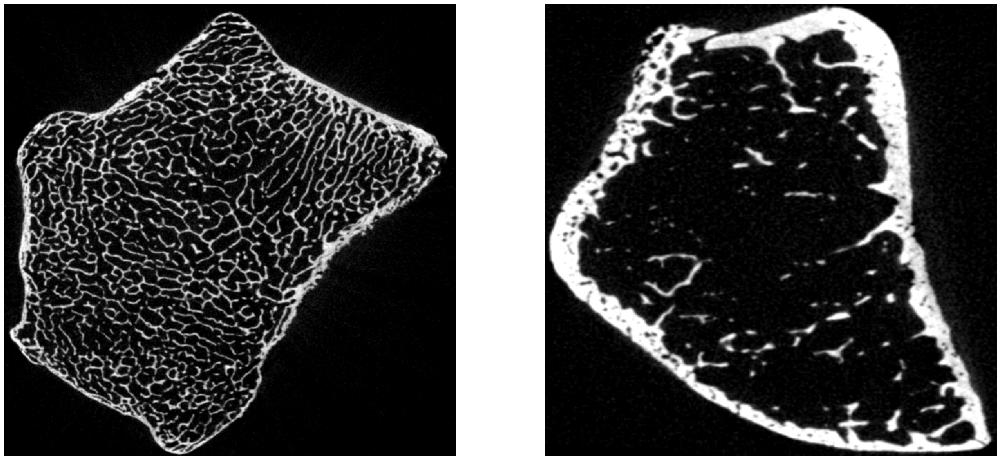


Figure 2: Volume 1 (left) and Volume 2 (right) used for the evaluation. Their number of slices, height and width are respectively $164 \times 882 \times 796$ and $194 \times 466 \times 372$ voxels. Window size is $[-1000, 3000]$ HU.

HP that gives the best validation PSNR. The same strategy is used for each network. Results show that for all networks, the optimal kernel size is 3×3, compared to 5×5 and 7×7. We find that 4 generator updates for 1 discriminator update is the best choice for WGAN based networks, as we tested ratios between 0.2 and 5 between both number of updates. The best batch size is 128, i.e the maximum size that could fit in the memory of the resources that were used for the study. The reason for these HP to be similar for every loss function is that those hyperparameters mostly depend on both the training data and the general structure of the networks which is considered as fixed. We only find differences in the optimal HPs for the learning rate, λ_1 and λ_2 since these HPs specifically depend on the loss function. The optimal values for those HPs are represented in Table 2. We tested learning rates between 10^{-8}, for which we observed very slow decrease of the loss function, and 10^{-2}, for which we observed divergence of the loss function. As for the weighting parameters we tested values between 10^{-3} and 10^{3}.

Once the hyperparameters optimal values have been found, final training is performed on 64x64 patches from all 1,992 different 2D slices for a total of 297,976 patches. Computations are performed on a NVIDIA Tesla V100 GPU. The generator has slightly more than 2×10^6 trainable parameters, the discriminator has around 18×10^6 trainable parameters, and VGG loss implies 20×10^6 extra parameters that are not trainable but that still need to fit into the memory. Training of a CNN takes approximately 2 hours per epoch, and around 10 hours for WGAN-based networks since one training step consists of 5 updates : 4 for the generator and one for the discriminator. Tests show that this difference in terms of computation time cannot be avoided since convergence of WGAN based networks still require the same number of epochs as CNNs.

Network	lr_i	lr_f	λ_1	λ_2
CNN-L_1	6×10^{-4}	8×10^{-6}	-	-
CNN-L_2	$1,5 \times 10^{-4}$	8×10^{-6}	-	-
CNN-SSIM	$1,5 \times 10^{-4}$	8×10^{-6}	-	-
CNN-VGG	10^{-3}	3×10^{-5}	-	-
CNN-SSIM-L_1	$1,5 \times 10^{-4}$	$1,5 \times 10^{-5}$	100	-
CNN-VGG-L_1	$1,5 \times 10^{-4}$	$1,5 \times 10^{-5}$	50	-
CNN-VGG-L_2	$1,5 \times 10^{-4}$	$1,5 \times 10^{-5}$	100	-
WGAN-L_1	$1,5 \times 10^{-4}$	$1,5 \times 10^{-5}$	1000	-
WGAN-VGG	$1,5 \times 10^{-4}$	8×10^{-6}	20	-
WGAN-SSIM-L_1	$1,5 \times 10^{-4}$	$1,5 \times 10^{-5}$	1	500
WGAN-VGG-L_1	$1,5 \times 10^{-4}$	8×10^{-6}	10	50
WGAN-VGG-L_2	$1,5 \times 10^{-4}$	8×10^{-6}	20	50

Table 2: Optimal hyperparameters for each method. These hyperparameters have been optimized on a validation set consisting of 20% of the slices obtained from the 10 training volumes. The learning rate decreases exponentially from lr_i to lr_f during training.

III.C. Evaluation

For evaluation, we simulate 4 configurations: 5%, 10%, 15% and 20% of the maximum dose. In what follows, only results for 10% and 20% are presented for simplicity but our conclusions take all configurations into account. Note that we control the dose amount, which is not equivalent to controlling the amount of Poisson noise since the latter depends on the density of the volume: there is more attenuation and thus more Poisson noise for more dense volumes.

Comparisons between algorithms hold as long as networks are able to correctly reconstruct images, which is no longer the case when the initial FBP is too deteriorated; in that case all networks fail to retrieve an accurate reconstruction, which is due to the limits of the reconstruction method itself and not to the choice of the loss functions. As a consequence, we decided that the quality of reconstruction for dose amounts lower than 5% was not satisfying enough to be included in our training and/or testing data.

In what follows, we study PSNR, SSIM, resolution (Resol), Wasserstein distance for the 1D distribution within the whole volume (WV) and within the bone area (WB). In the segmented reconstructions we also study DICE, BV/TV, mean absolute Euler number difference compared to the ground truth (E-N) as well as the relative absolute difference of object counts (O-C).

III.C.1. Pixelwise loss function study

Table 3 reports all tested metrics for different configurations, with the emphasis put on comparing the presence of L_1, L_2 or no pixelwise loss function in the overall cost function. Results are given for the evaluation volume 1 for 10% dose but results are similar for 20% and for the other volume. By comparing each row from every block, one can observe that for resolution, L_1 loss improves the performance compared to using no pixelwise loss or using L_2 loss. Also, using no pixelwise loss function with CNN-VGG leads to a

WGAN-VGG	PSNR	SSIM	DICE	BV/TV	E-N	O-C	Resol (μm)	WV	WB
$\emptyset$	28.91	0.811	0.848	0.140	29 ± 20	0.07 ± 0.05	98.6 ± 6.7	21.62	43.81
L_1	29.94	0.842	0.864	0.140	24 ± 17	0.07 ± 0.05	86.6 ± 6.0	13.61	19.93
L_2	29.63	0.829	0.859	0.141	23 ± 18	0.06 ± 0.04	93.2 ± 5.9	10.53	31.20
CNN-VGG	PSNR	SSIM	DICE	BV/TV	E-N	O-C	Resol (μm)	WV	WB
$\emptyset$	26.87	0.128	0.858	0.140	40 ± 28	0.20 ± 0.11	98.3 ± 4.4	332.57	581.71
L_1	30.43	0.846	0.866	0.140	37 ± 27	0.21 ± 0.05	77.9 ± 4.4	27.75	106.36
L_2	30.19	0.851	0.858	0.147	79 ± 40	0.29 ± 0.05	94.1 ± 4.6	41.55	209.64
CNN-SSIM	PSNR	SSIM	DICE	BV/TV	E-N	O-C	Resol (μm)	WV	WB
$\emptyset$	30.36	0.871	0.865	0.141	44 ± 29	0.23 ± 0.05	83.1 ± 4.1	29.45	103.6
L_1	30.35	0.859	0.865	0.139	26 ± 20	0.16 ± 0.05	77.3 ± 5.3	30.47	109.56
L_2	30.27	0.863	0.859	0.148	63 ± 37	0.28 ± 0.05	91.1 ± 5.1	39.22	215.16
CNN	PSNR	SSIM	DICE	BV/TV	E-N	O-C	Resol (μm)	WV	WB
L_1	30.43	0.848	0.866	0.140	33 ± 26	0.19 ± 0.05	78.7 ± 5.0	26.24	107.08
L_2	30.17	0.852	0.856	0.148	37 ± 31	0.20 ± 0.05	95.4 ± 5.1	44.48	240.55

Table 3: Metrics for volume 1 and 10% dose. Here we study the influence of the pixelwise loss. Bold entries in the first column indicate the part of the loss function that is fixed. BV/TV ratio for ground truth is 0.138.

	Volume 1		Volume 2	
Method	10%	20%	10%	20%
WGAN-L_1	$\mathbf{75.1 \pm 5.1}$	72.4 ± 4.3	$\mathbf{74.8 \pm 6.2}$	$\mathbf{73.5 \pm 5.1}$
WGAN-SSIM-L_1	75.4 ± 4.6	$\mathbf{72.3 \pm 4.0}$	75.3 ± 5.9	74.0 ± 5.5
CNN-SSIM-L_1	77.3 ± 5.3	73.9 ± 4.3	76.4 ± 5.3	75.3 ± 5.1
CNN-L_1	78.7 ± 5.0	75.5 ± 3.9	76.2 ± 6.7	74.7 ± 5.7
CNN-VGG-L_1	77.9 ± 4.4	75.4 ± 4.3	76.9 ± 6.4	75.6 ± 6.0
CNN-SSIM	83.1 ± 4.1	80.1 ± 4.0	79.3 ± 5.6	78.6 ± 6.2
WGAN-VGG-L_1	86.6 ± 6.0	83.1 ± 5.3	82.9 ± 7.2	79.7 ± 7.1
WGAN-VGG-L_2	93.2 ± 5.9	87.8 ± 4.9	85.2 ± 7.8	82.3 ± 7.0
CNN-VGG-L_2	94.1 ± 4.7	90.9 ± 4.7	90.4 ± 6.8	90.5 ± 6.7
CNN-L_2	95.4 ± 5.1	91.9 ± 5.2	88.8 ± 8.1	87.2 ± 7.8
WGAN-VGG	98.6 ± 6.7	93.6 ± 6.7	94.3 ± 10.8	93.1 ± 10.8
CNN-VGG	98.3 ± 4.4	95.3 ± 4.1	91.2 ± 6.7	90.3 ± 6.7

Table 4: Values of resolution in μm for each method and for the test volumes 1 and 2 considering 10% and 20% dose for both.

significant performance drop for most of the metrics. We notice that the BV/TV ratio is higher for L_2 loss. The differences between each method are slight when looking at PSNR, SSIM or DICE. This observation can also be made on the other test volumes and for different dose configurations, thus we cannot use those metrics to distinguish between the performance of each loss function. As for connectivity and metrics involving the Wasserstein-1 distance, different performance can be observed depending on the loss function, but it is not related to the pixelwise loss according to this table.

Table 4 highlights the enhancement of resolution with L_1 loss. Indeed, the most performing networks are those who have the mean absolute error as part of the loss function, with the slight exception of WGAN-VGG-L_1 that ranks behind CNN-SSIM, but this will be discussed in the next subsection dealing with the impact of the structural loss.

Figure 3 illustrates the increased performance of L_1 loss on CNN-SSIM and WGAN-VGG examples. The figure shows the evolution of FRC value with respect to the maximum frequency value for which the correlation is considered - the ring radius. Since the correlation is computed on 2D slices, we selected a

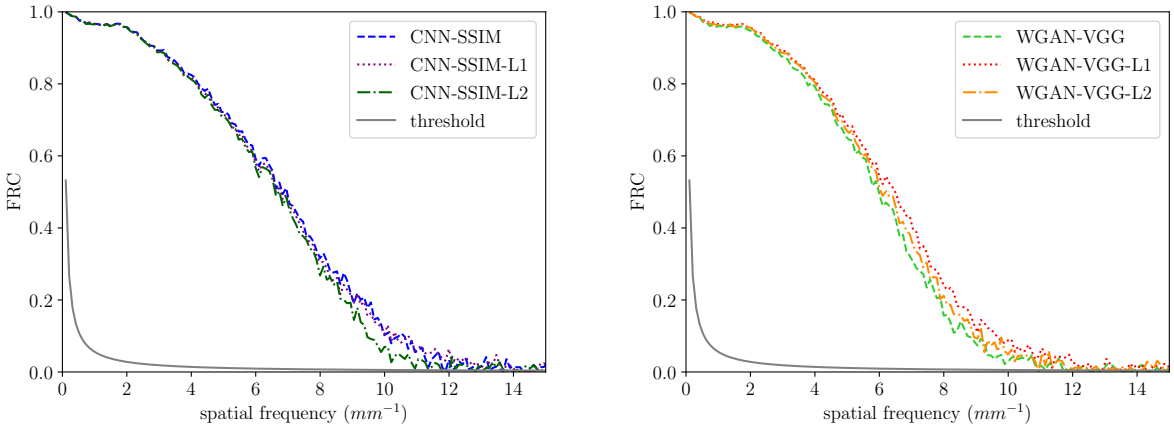


Figure 3: FRC curve on a selected slice on volume 1 for different reconstruction methods, for 10% dose. The y-axis represents the Fourier Ring Correlation value between 0 and 1, the x-axis is the radius of the ring in the Fourier domain within which the correlation is computed. The threshold to compute the resolution according to (10) is also represented.

slice on volume 1 to display the curves for the different methods on 10% dose. One can observe that for high frequencies, the L_1 curve is above the other curves. For CNN-SSIM, improvement with L_1 is observed between 9 mm^{-1} and 11 mm^{-1} (high frequencies), whereas for WGAN-VGG it is already observed with lower frequencies, around 6 mm^{-1}. In both cases, this indicates that the L_1 loss better transcribes high frequencies which allows to reconstruct thinner details.

III.C.2. Structural loss function study

By performing a similar ablation study to investigate the impact of the structural part of the loss - SSIM vs VGG - we find that VGG is more efficient when associated to WGAN and SSIM with CNN, i.e with no adversarial loss. In the same way as for pixelwise loss functions, most of the metrics do not allow to clearly distinguish between WGAN-VGG and CNN-SSIM based networks, except for resolution and connectivity related metrics. Table 5 shows the mean and standard deviation of the difference between the Euler number of both the predicted slices and the ground truth ones. As the Euler number computes the difference between the number of objects and the number of holes estimated in the image, the other column represents the relative difference for the object count only. This allows to ensure that the observed performance for the Euler number metric is not biased by the fact that both the count of holes and objects are not correct. Results clearly show that WGAN-VGG outperforms CNN-SSIM, independently from the pixelwise loss function that is potentially associated. Connectivity is thus better represented with WGAN-VGG according to our results. We also notice that using no structural loss decreases the performance in terms of connectivity : on volume 1, the error in terms of objects count is more than 10 % higher for CNN-L_1 compared to WGAN-VGG networks, and on volume 2 the Euler number mean absolute difference is between 2 and 3 times larger

	Volume 1				Volume 2			
	10%		20%		10%		20%	
Method	E-N	Obj. c	E-N	Obj. c	E-N	Obj. c	E-N	Obj. c
WGAN-VGG-L_2	$\mathbf{23 \pm 18}$	$\mathbf{0.06 \pm 0.04}$	29 ± 23	0.08 ± 0.05	$\mathbf{6 \pm 5}$	0.08 ± 0.06	$\mathbf{6 \pm 4}$	0.07 ± 0.06
WGAN-VGG-L_1	24 ± 17	0.07 ± 0.05	$\mathbf{25 \pm 17}$	$\mathbf{0.05 \pm 0.04}$	6 ± 5	0.12 ± 0.09	6 ± 5	0.09 ± 0.07
WGAN-VGG	29 ± 20	0.07 ± 0.05	34 ± 31	0.11 ± 0.06	7 ± 5	0.13 ± 0.08	6 ± 4	0.13 ± 0.08
CNN-SSIM-L_1	26 ± 20	0.16 ± 0.05	35 ± 25	0.18 ± 0.04	16 ± 8	0.08 ± 0.06	16 ± 7	0.07 ± 0.06
CNN-L_1	33 ± 26	0.19 ± 0.05	39 ± 28	0.20 ± 0.04	15 ± 7	0.08 ± 0.06	16 ± 7	0.07 ± 0.05
WGAN-L_1	39 ± 25	0.20 ± 0.05	43 ± 26	0.19 ± 0.04	10 ± 7	0.08 ± 0.06	9 ± 6	0.07 ± 0.05
CNN-VGG-L_1	37 ± 27	0.21 ± 0.05	42 ± 28	0.21 ± 0.04	15 ± 8	0.09 ± 0.06	12 ± 7	0.23 ± 0.08
CNN-L_2	37 ± 31	0.20 ± 0.05	58 ± 37	0.26 ± 0.05	11 ± 7	0.32 ± 0.09	12 ± 8	0.30 ± 0.09
CNN-SSIM	44 ± 29	0.23 ± 0.05	56 ± 31	0.24 ± 0.04	9 ± 6	0.10 ± 0.07	8 ± 6	0.09 ± 0.06
WGAN-SSIM-L_1	55 ± 26	0.22 ± 0.05	67 ± 28	0.23 ± 0.04	8 ± 6	$\mathbf{0.07 \pm 0.05}$	7 ± 5	$\mathbf{0.06 \pm 0.05}$
CNN-VGG-L_2	79 ± 40	0.29 ± 0.05	91 ± 43	0.32 ± 0.05	10 ± 7	0.25 ± 0.09	12 ± 7	0.23 ± 0.08

Table 5: Connectivity metrics : Euler number absolute difference and objects count relative difference compared to ground truth for each method and for test volumes 1 and 2 with 10% and 20% dose.

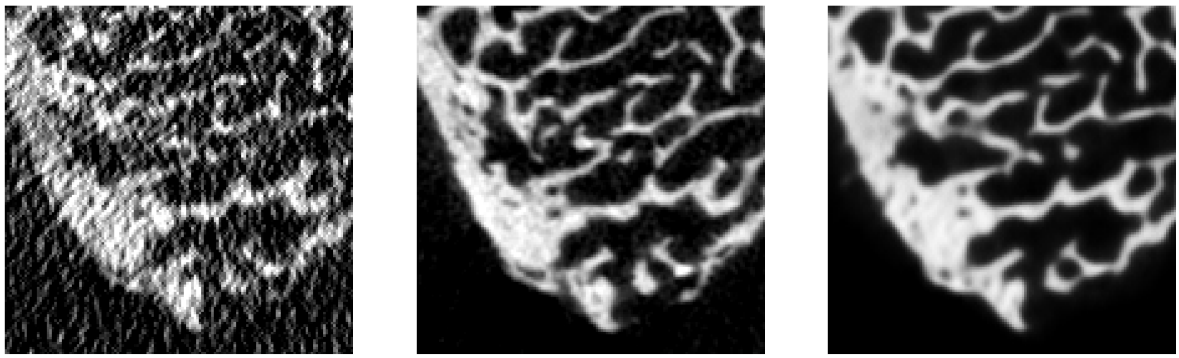
	Volume 1				Volume 2			
	10%		20%		10%		20%	
VGG	WV	WB	WV	WB	WV	WB	WV	WB
CNN	332.57	581.71	319.16	499.60	350.40	703.88	347.73	653.04
WGAN	21.62	43.81	28.97	98.68	35.27	85.67	40.59	78.73
L_1	WV	WB	WV	WB	WV	WB	WV	WB
CNN	26.24	107.08	22.04	41.43	44.20	161.46	34.55	117.62
WGAN	22.75	59.95	22.11	22.13	39.49	142.54	30.40	88.64
VGG-L_1	WV	WB	WV	WB	WV	WB	WV	WB
CNN	27.75	106.36	20.72	30.38	43.02	148.86	33.42	103.89
WGAN	13.61	19.93	23.22	65.17	43.68	126.14	39.45	75.97
SSIM-L_1	WV	WB	WV	WB	WV	WB	WV	WB
CNN	30.47	109.56	21.95	29.49	46.01	125.38	38.30	73.50
WGAN	17.55	32.62	24.50	43.50	35.18	106.09	27.17	54.11

Table 6: Wasserstein 1 distance for the 1D distributions in the entire volume (WV) and in areas considered as bone (WB) by the segmentation algorithm. Here we study the influence of the adversarial loss on those metrics. Bold entries in the first column indicate the part of the loss function that is fixed.

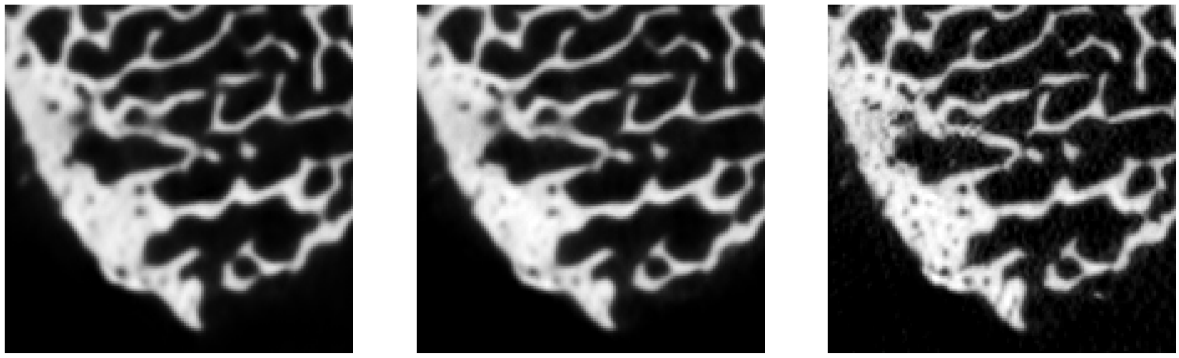
for CNN-L_1. Nevertheless, it can be observed in Table 3 or 4 that VGG loss - especially when associated with WGAN - leads to a higher value for the resolution which means a reduced ability to transcribe high frequencies.

III.C.3. Adversarial loss function study

Finally, we isolate the impact of the presence or not of an adversarial loss in the cost function. Once again, the impact is not significant for every metric. Table 6 shows improved performance of WGAN based network when focusing on the W_1 distance between the 1D distributions, especially when focusing on the distribution within the bone area, which is represented in the second column. Note that this is not straightforward since the Wasserstein distance that is used for training the networks is not the same as the one used as a metric (n-dimensional vs 1-dimensional). This result suggests that the adversarial loss helps retrieve the correct distribution of BMD in the volumes. Of course this is not useful if the structure is not correctly reconstructed, but this still means that statistics from the densities are more accurate with such a loss.



(a) FBP (b) Ground truth (c) CNN-L_1



(d) CNN-SSIM-L_1 (e) WGAN-L_1 (f) WGAN-VGG-L_1

Figure 4: ROI from volume 1 of size 140×140 voxels obtained with different methods. Window size is $[-1000, 3000]$ HU. FBP is obtained after simulation of 10% of the normal dose to obtain the projections from the ground truth. Networks are fed with this FBP as input.

Figure 4 shows reconstructions with the methods that gave the best performance when taking all metrics into account, and especially resolution, connectivity and W_1 distance.

IV. Discussion

In our study PSNR, SSIM and DICE did not allow to distinguish between pixelwise, structural and adversarial losses. It is an argument to encourage future studies to evaluate methods regarding task-specific metrics since they allow to do so according to our experiments.

Our results clearly suggest that pixelwise loss functions have a major role in the resolution that is observed in the reconstructions. We showed that the L_1 loss is the most suited one for the task of reconstructing images with the best resolution. Moreover, the choice of L_1 loss has no negative impact on all the other metrics that were tested. We also showed that L_2 loss, besides deteriorating the resolution, tends to increase the BV/TV ratio in the segmented reconstruction. This can be explained by the common tendency of such a loss to oversmooth images, which encourages the segmentation algorithm to consider bone areas wider than they really are. Our results therefore strongly suggest L_1 to be considered as part of the loss function for its ability to improve resolution performance without decreasing the quality of the reconstruction considering other metrics.

As for the structural loss, experiments show that using VGG loss alone implies a significant drop in performance for quantitative metrics. This is due to the fact that VGG network was trained to perform classification on natural images; only the structures are helpful for VGG to perform this task. The pixel intensities are not considered as relevant features for this purpose. The satisfying performance of CNN-VGG in terms of connectivity metrics, DICE and BV/TV compared to its poor performance in terms of quantitative metrics such as PSNR or Wasserstein distance is a perfect example that highlights the need for a loss function to contain elements that take both structure and pixel values into account. Following this observation, we find enhanced overall performance when VGG is associated to an adversarial loss. WGAN-VGG networks showed to significantly improve performance in terms of connectivity in the reconstructions. This however results in a decreased performance in terms of resolution, even with L_1 loss. SSIM does not present this drawback in terms of resolution, but on the other side it only shows limited improvement in terms of connectivity. The positive impact of the structural loss on connectivity metrics is thus more significant for WGAN-VGG than for CNN-SSIM, but it appears to induce an increased resolution. We suggest that depending on the application, the trade-off could be dealt with by tuning λ_1 and λ_2 accordingly while using a network like WGAN-VGG-L_1.

The last point of interest is the presence or not of an adversarial loss in the cost function. We showed in the previous point that when associated with VGG, the adversarial loss has a positive impact on the connectivity metrics. This can be understood by the fact that learning the probability distribution of high-dose reconstructions helps capture the anatomically correct shapes in the bone microstructure. We

also observe better accuracy in terms of BMD distribution. As WGAN-based networks try to learn the n-dimensional distribution of high-dose images, n being the total number of pixels, it is reasonable to think that such networks are more likely to retrieve the 1-dimensional distribution of density values since it can be induced by the knowledge of the former. Conditionnally on the fact that the structure is correctly reconstructed, this enhanced performance can be helpful for practiciancs since we mentioned that BMD values are among elements that are considered to diagnose various bone-related diseases.

Another aspect to consider when making the choice of a loss function is the computation time and memory requirement for training the networks. In our case, the reconstruction time during inference is the same for all methods since we use a similar generator for all of them. However, we mentioned in Section III. that using VGG loss increases the memory consumption. This can reduce the maximum batch size to use for training compared to other methods and potentially decrease performance, even if this is not an issue that we experienced in our tests. Also, using the adversarial loss increases the training time by a factor of 5 in our experiments, which also needs to be taken into account when considering the improvement brought by such a loss for BMD distribution accuracy. Finally, when considering a loss function composed of different parts, this adds extra hyperparameters to tune during the training phase.

The fact of increasing the training time or the number of hyperparameters might have a negative impact on the final performance of the network. Indeed, for a fixed computational budget, the number of hyperparameters that can be tested to validate the performance of the networks can be significantly reduced for complex loss functions and particularly those which require an adversarial loss. Our study does not put the emphasis on such constraints, and it is not clear whether the focus should rather be put on spending time finding optimal hyperparameters with a simpler loss function or not. Analyzing the results considering the computational cost of the loss function could therefore be the subject of another study.

Also, our hyperparameters selection was performed by choosing PSNR as the validation metric since it is common practice and we did not want results to be biased towards a stronger importance given to a specific metric. Our results however suggest that the choice of those hyperparameters should be driven by task-dependent metrics since we noticed that they do not necessarily match common evaluation metrics such as PSNR or SSIM. For practical purpose, the metrics used for optimizing the network's hyperparameters need to be carefully chosen.

An obvious limitation of our study is that the answers brought by our experiments only hold for the specific data that we are studying, i.e bone microarchitecture obtained from CT imaging; nevertheless, we believe that the conclusions obtained in this study are of interest regardless of the domain as long as one is interested in both the structural information retrieved by the network and the accuracy of the reconstructed

pixel values, whether they correspond to attenuation in the case of CT imaging or to activity for emission tomography.

Finally, our work mainly focuses on the study of the reconstruction depending on the loss used for training the network. One needs to keep in mind that the method still relies on the low-dose FBP as input, which might not be optimal since loss of information can be observed especially as the radiation dose decreases. Better results should be observed when considering a more complex architecture, that performs the reconstruction from the projection data in an end-to-end framework for instance. 3D networks and use of a U-NET generator could also be solutions to improve performance. We believe that these architectures would benefit to every training scheme without modifying the comparative results that we obtained. In any case, our results allow to understand the impact of different types of loss functions when reconstructing bone microarchitecture with deep learning based methods and remain of interest even if more complex networks are used for practical application.

V. Conclusion

The assessment of the quality of the reconstruction of bone microstructures seems to be insufficient when only considering PSNR and SSIM. Instead, relevant features such as the BV/TV ratio, Euler number, Wasserstein distance between the HU distributions of bones densities are among metrics that allow the evaluation of the retrieval of both structural details and quantitative information on the BMD. We showed that the loss function used to train a neural network has a major influence on those metrics. Pixelwise loss functions were found to improve the resolution observed in the reconstructions, with L_1 loss being the most effective in our tests. Structural loss functions play a role on the ability of networks to retrieve bone structures as shown by connectivity metrics, and VGG loss improves performance in that sense, at the cost of a deteriorated resolution. Adding an adversarial loss leads to reconstructions with more accuracy in terms of BMD. When choosing the most suited loss function for the particular task of reconstructing bone microstructure with accurate BMD values, one needs to keep in mind the trade-off between the computational cost of complex losses and the improved performance that they bring.

VI. Acknowledgments

The authors acknowledge financial support of the French National Research Agency through the ANR project LABEX PRIMES (ANR-11-IDEX-0007) of Université de Lyon. The authors thank Andrew Burghard from University of California, San Francisco, USA, for providing the experimental 3D μCT images.

VII. Conflict of Interest Statement

The authors have no relevant conflicts of interest to disclose.

References

[1] P. Gilbert, Iterative methods for the three-dimensional reconstruction of an object from projections, J. Theor. Biol. **36**, 105–117 (1972).

[2] K. Lange and R. Carson, EM reconstruction algorithms for emission and transmission tomography, J Comput Assist Tomogr **8**, 306–16 (1984).

[3] E. Sidky, J. Jørgensen, and X. Pan, Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle–Pock algorithm, Phys. Med. Biol. **57**, 3065–3091 (2012).

[4] H. Banjak et al., Evaluation of noise and blur effects with SIRT-FISTA-TV reconstruction algorithm: Application to fast environmental transmission electron tomography, Ultramicroscopy **189**, 109–123 (2018).

[5] Y. Li, K. Li, C. Zhang, J. Montoya, and G. H. Chen, Learning to Reconstruct Computed Tomography Images Directly From Sinogram Data Under A Variety of Data Acquisition Conditions, IEEE Transactions on Medical Imaging **38**, 2469–2481 (2019).

[6] K. Liang, H. Yang, and Y. Xing, Comparison of projection domain, image domain, and comprehensive deep learning for sparse-view X-ray CT image reconstruction, 2019.

[7] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by N. Navab, J. Hornegger, W. Wells, and A. Frangi, pages 234–241, Cham, 2015, Springer International Publishing.

[8] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, Deep Convolutional Neural Network for Inverse Problems in Imaging, IEEE Transactions on Image Processing **26**, 4509–4522 (2017).

[9] H. Chen et al., Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network, IEEE Trans Med Imaging **36**, 2524–2535 (2017).

[10] J. Adler and O. Öktem, Solving ill-posed inverse problems using iterative deep neural networks, Inverse Problems **33**, 124007 (2017).

[11] J. Adler and O. Öktem, Learned Primal-Dual Reconstruction, IEEE Transactions on Medical Imaging **37**, 1322–1332 (2018).

[12] H. Chen et al., LEARN: Learned Experts' Assessment-Based Reconstruction Network for Sparse-Data CT, IEEE transactions on medical imaging **37**, 1333—1347 (2018).

[13] K. Hammernik et al., Learning a Variational Network for Reconstruction of Accelerated MRI Data, Magnetic Resonance in Medicine (2017).

[14] W. Yang, H. Zhang, J. Yang, J. Wu, X. Yin, Y. Chen, H. Shu, L. Luo, G. Coatrieux, Z. Gui, and Q. Feng, Improving Low-Dose CT Image Using Residual Convolutional Network, IEEE Access **5**, 24698–24705 (2017).

[15] H. Shan, Y. Zhang, Q. Yang, U. Kruger, M. K. Kalra, L. Sun, W. Cong, and G. Wang, 3-D Convolutional Encoder-Decoder Network for Low-Dose CT via Transfer Learning From a 2-D Trained Network, IEEE Transactions on Medical Imaging **37**, 1522–1534 (2018).

[16] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing **13**, 600–612 (2004).

[17] R. Zhang, P. Isola, A. Efros, E. Shechtman, and . Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, arXiv e-prints (2018).

[18] I. Goodfellow et al., Generative Adversarial Networks, in *Advances in Neural Information Processing Systems*, 2014.

[19] Q. Yang et al., Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss, IEEE Trans Med Imaging **37**, 1348–1357 (2018).

[20] Z. Hu et al., DPIR-Net: Direct PET Image Reconstruction Based on the Wasserstein Generative Adversarial Network, IEEE Transactions on Radiation and Plasma Medical Sciences **5**, 35–43 (2021).

[21] B. Kim, M. Han, H. Shim, and J. Baek, A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images, Medical Physics **46**, 3906–3923 (2019).

[22] Y. Jiang, J. Zhao, E. Liao, R. chun Dai, X. Wu, and H. Genant, Application of micro-ct assessment of 3-d bone microstructure in preclinical and clinical studies, Journal of Bone and Mineral Metabolism **23**, 122–131 (2009).

[23] K. Engelke, C. Libanati, T. Fuerst, P. Zysset, and H. Genant, Advanced CT based in vivo methods for the assessment of bone density, structure, and strength, Curr Osteoporos Rep. **11(3)**, 246–55 (2013).

[24] Y. Li, B. Sixou, and F. Peyrin, Nonconvex Mixed TV/Cahn–Hilliard Functional for Super-Resolution/Segmentation of 3D Trabecular Bone Images, J Math Imaging Vis , 1–11 (2018).

[25] F. Peyrin and K. Engelke, *CT Imaging: Basics and New Trends*, pages 883–915, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[26] H. K. Genant, K. Engelke, and S. Prevrhal, Advanced CT bone imaging in osteoporosis, Rheumatology **47**, iv9–iv16 (2008).

[27] A. Nazarian, M. Stauber, D. Zurakowski, B. Snyder, and R. Müller, The interaction of microstructure and volume fraction in predicting failure in cancellous bone, Bone **39**, 1196–1202 (2006).

[28] L. Oei, F. Koromani, F. Rivadeneira, M. Zillikens, and E. Oei, Quantitative imaging methods in osteoporosis, Quantitative Imaging in Medicine and Surgery **6** (2016).

[29] K. Mei et al., Is multidetector CT-based bone mineral density and quantitative bone microstructure assessment at the spine still feasible using ultra-low tube current and sparse sampling?, Eur Radiol **27**, 5261–5271 (2017).

[30] J. Johnson, A. Alahi, and L. Fei-Fei, Perceptual Losses for Real-Time Style Transfer and Super-Resolution, in *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016, Springer International Publishing.

[31] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv e-prints (2014).

[32] M. Arjovsky, S. Chintala, and L. Bottou, Wasserstein GAN, arXiv e-prints (2017).

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, Improved Training of Wasserstein GANs, in *Advances in Neural Information Processing Systems*, 2017.

[34] N. Otsu, A Threshold Selection Method from Gray-Level Histograms, IEEE Trans. Syst. Man Cybern. Syst. **9**, 62–66 (1979).

[35] J. Kabel, A. Odgaard, B. van Rietbergen, and R. Huiskes, Connectivity and the elastic properties of cancellous bone, Bone **24**, 115–120 (1999).

[36] N. Banterle, K. Bui, E. Lemke, and M. Beck, Fourier ring correlation as a resolution criterion for super-resolution microscopy, Journal of Structural Biology **183**, 363–367 (2013).

[37] A.Ramdas, N.Garcia, and M.Cuturi, On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests, 2015.

[38] W. Van Aarle et al., The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography, Ultramicroscopy **157** (2015).

[39] J. Leuschner, M. Schmidt, D. Baguer, and P. Maaß, The LoDoPaB-CT Dataset: A Benchmark Dataset for Low-Dose CT Reconstruction Methods, 2020.

[40] K. He, X. Zhang, S. Ren, and J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, arXiv e-prints (2015).

[41] D. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, arXiv e-prints (2014).