Combining conditional GAN with VGG perceptual loss for bones CT image reconstruction

Théo Leuliet, Voichiţa Maxim, Françoise Peyrin, Bruno Sixou

Combining conditional GAN with VGG perceptual loss for bones CT image reconstruction

Théo Leuliet, Voichiţa 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
{theo.leuliet,voichita.maxim,francoise.peyrin,bruno.sixou}@creatis.insa-lyon.fr

Abstract Reducing both the radiation dose to patients and the reconstruction time is key for X-ray computed tomography. The imaging of bone microarchitecture at high spatial resolution is all the more challenging as noisy data can severely deteriorate structural details. Deep Learning based algorithms are efficient for post-processing poor-quality reconstructions obtained with Filtered BackProjection, though MSE-trained networks hardly capture the structural information relevant for bones. Instead, conditional GANs allow to generate very realistic volumes that correspond to their corrupted FBP. Moreover, perceptual losses are efficient to capture key features for the human eye. In this work we combine both concepts within a new framework called CWGAN-VGG that is designed for the reconstruction of bones at high spatial resolution, with an emphasis put on the preservation of their structural information. We show on simulated low-dose CT bones data that our CWGAN-VGG outperforms state-of-the-art methods that involve GANs and/or perceptual losses in terms of PSNR and other metrics.

1 Introduction

Bone microstructure study with X-ray computed tomography (CT) is a challenging task due to the complexity of the underlying structures [1], [2]. Physical limitations of scanners and the need for reducing the patient's radiation dose may lead to noisy data, which need to be corrected to help practitioners get relevant parameters. When the number of projections is sufficiently large with a reasonable amount of noise, analytical algorithms like the Filtered Back-Projection (FBP) can offer satisfying results. When this is no longer the case, iterative methods [3] [4] can be considered. A major drawback of such algorithms is the reconstruction time and the need for tuning parameters for every reconstruction.

Deep Learning based algorithms have the potential to enhance the quality of images by learning patterns from ground-truth data, while significantly reducing the reconstruction time compared to iterative algorithms. A solution is to use neural networks to improve a poor-quality analytically obtained reconstruction [5] [6], e.g the FBP obtained from low-dose projections.

A critical point to address is the way the network should be trained. A Mean Squared Error (MSE) loss between predicted images and the corresponding ground-truths as in [7] might lead to slight oversmoothing that deteriorates some important structural details and thus affects the study of bone microarchitecture.

Instead, the use of a generative adversarial network (GAN) [8] allows to capture the probability distribution of the ground-truth images. In [5], such a network is trained with the Wasserstein distance along with a perceptual loss that compares the network output against the ground truth in a feature space designed to match the human eye perception, thus preserving key structural information. The resulting WGAN-VGG achieves impressive results on noise removal and artifacts correction. A similar architecture was proposed in [9] for underwater image restoration.

Nevertheless in both cases, the Wasserstein distance might be low even if the output does not correspond to the FBP it has been generated from. This is not the case when considering a conditional GAN [10] where the discriminator also takes the conditional information as an input. Such a framework was proposed in [11] and for medical image reconstruction in [6]. In [12], authors propose a conditional Wasserstein GAN (CWGAN) to capture the probability distribution of some volume conditionally to the FBP obtained from low-dose projections.

Combining such a conditional GAN with a perceptual loss has never been performed, though it seems to be perfectly adapted to bone microarchitecture imaging in order to capture their structural information. We then propose the CWGAN-VGG network that learns a probability distribution conditionally to the FBP obtained from low-dose projections, with a perceptual loss that is added to the generator loss function in order to preserve bone microstructure information.

In section 2, we present our CWGAN-VGG algorithm. In section 3 we detail the numerical experiments that we performed on simulated low-dose projections of μCT bone data. In section 4 we discuss the impact of both the conditioning and the perceptual loss on the quality of the reconstructions.

2 CWGAN-VGG framework

2.1 Conditional GAN and perceptual loss

Let y be the FBP reconstructed volume from low-dose projections and x the reference volume. We recall the CWGAN introduced in [12], where the aim is to approximate the posterior distribution $\pi(x|y)$ with a parametrized generator $G_\theta(y)$. Knowing this distribution allows to generate a number of

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 μCT data.

volumes that can be responsible for data y. To approximate such a posterior distribution, the objective is to find θ^* that minimizes $d(G_\theta(y), \pi(x|y))$, with d some distance between probability distributions. A now commonly used method to improve neural networks convergence is to consider the Wasserstein distance [13]. Denoting the probability distribution associated to the generator $G_\theta(y)$ by $P_\theta(y)$ - remember that y is the condition here -, the dual characterization of this distance writes

$$W(\pi(x|y), P_\theta(y)) = \sup_{||f||_L \leq 1} \mathop{\mathrm{E}}_{x\sim\pi(x|y)}[f(x)] - \mathop{\mathrm{E}}_{v\sim P_\theta(y)}[f(v)] \quad (1)$$

where the supremum is taken over all the 1-Lipschitz functions. Since it is not feasible to cover the entire space of these functions, they are parametrized with a neural network D_w called a discriminator, with parameters w. Also, the generator $G_\theta(y)$ takes as input realizations z drawn from a simple probability distribution η. We ensure the Lipschitz condition by adding a gradient penalty term to the distance function as in [14]. Minimizing the Wasserstein distance approximated by neural networks finally gives the optimization problem

$$\begin{aligned}\theta^* \in \mathop{\mathrm{argmin}}_\theta \sup_w L_{\mathrm{CWGAN}}(D,G) = &\mathop{\mathrm{E}}_{(x,y)\sim\mu}[D_w(x,y)] \\ &- \mathop{\mathrm{E}}_{\substack{z\sim\eta \\ y\sim P_y}}[D_w(G_\theta(z,y),y)] \\ &+ \lambda \mathop{\mathrm{E}}_{\hat{x}\sim P_{\hat{x}}}[(||\nabla_{\hat{x}} D_w(\hat{x},\hat{y})||_2 - 1)^2]\end{aligned} \quad (2)$$

where μ is the joint distribution of (x,y) corresponding to paired low-dose FBPs and high-dose ground truths, P_y the unknown distribution of low-dose FBP data, $\hat{x} \sim P_{\hat{x}}$ are sampled along straight lines between samples from both $\pi(x|y)$ and the generated distribution $P_\theta(y)$, $\hat{y}$ are sampled along straight lines between the corresponding FBPs, and λ is the weighting term for the gradient penalty. During training, expectations are replaced by their empirical counterpart obtained with paired data.

The resulting network generates stochastic samples conditionnally to the FBP of low-dose projections, according to a probability distribution that approximates the true distribution $\pi(x|y)$. Also, one can take η as a Dirac distribution. In that case, the network is deterministic and generates a single output from the low-dose FBP. We call this network Det-CWGAN.

In [5], authors propose the WGAN-VGG framework which consists in training a generator by adding a perceptual loss to an unconditioned WGAN objective function, in order to better fit human perception of images, given as

$$L_{\mathrm{VGG}}(G_\theta) = \frac{1}{n} \mathop{\mathrm{E}}_{(x,y)\sim\mu}[||VGG(G_\theta(y)) - VGG(x)||_F^2] \quad (3)$$

where n is the total number of voxels and VGG is the 16^{th} output of the pre-trained VGG-19 model [15], $||.||_F$ is the Frobenius norm, and in their case G_θ only takes y as input.

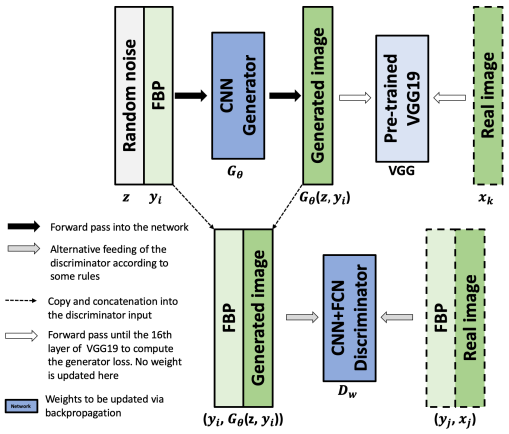

Figure 1: Scheme of the proposed CWGAN-VGG model. The FBP data is taken as input of the generator G_θ. Both the FBP and the generated image are concatenated to produce the input of the discriminator. The network is trained according to (4).

It is shown in [16] that such a loss better suits human perception compared to pixel-wise based losses. In this framework, the output is deterministic and the discriminator is not conditioned on the FBP input, which amounts to taking η as a Dirac distribution and $x \sim \pi(x)$ instead of $x \sim \pi(x|y)$ in (1).

2.2 Proposed architecture

In this work, we make use of the FBP computed from the acquired low-dose projections, to learn a conditional probability, by adding this FBP as an input to the discriminator. The benefits of conditioning the discriminator were already shown in [11] for natural images. Though authors used pixel-wise based additional losses, we propose to use the VGG perceptual loss since retrieving structural information on data is of major importance in bone microarchitecture imaging. Thus we propose the CWGAN-VGG framework that is trained as

$$\min_\theta \max_w L_{\mathrm{CWGAN}}(D_w, G_\theta) + \lambda_1 L_{\mathrm{VGG}}(G_\theta) \quad (4)$$

with λ_1 a weighting parameter. The scheme of the resulting network is presented in Fig. 1. In WGAN-VGG, the discriminator is not fed with the low-dose FBP, which results in a different paradigm compared to conditional GANs; the distribution that is learned is $\pi(x)$, and the generator is a mapping between the space of low-dose FBPs and the space of high-dose images. In conditional GANs, the low-dose FBP y is the conditional data and the generator is a mapping between the latent space Z, where samples z are drawn from η, and the space of high-dose images. To our knowledge this is the first time that the CWGAN-VGG architecture is proposed.

Moreover, both [11] and [12] pointed out the difficulties of CWGAN to generate stochasticity, as the network tends to ignore the input noise. Thus in our tests, we also implemented a deterministic CWGAN-VGG (Det-CWGAN-VGG) that only learns a Dirac distribution, for comparison.

3 Numerical Experiments

3.1 Materials and methods

The ground-truth data consist of human bone volumes reconstructed from acquisitions of radius and shin structures obtained on a SCANCO μ-CT 100 with a 24-μm resolution. We create 180 2D projections of these volumes - corresponding to a low-dose acquisition - with ASTRA Toolbox [17] in Python. To simulate counting noise, random values are drawn from a Poisson law with mean the projections pixels. To simulate detectors noise, we then add a zero-mean Gaussian noise with standard deviation $\sigma = 0.8\%$ of the mean value in the projections. We then take the FBP - with Hann filter - as the input of the neural networks.

The dataset is composed of 13 volumes from different patients, 3 of which are only taken for evaluation. These 3 volumes have respectively a number of slices, height and width of $164 \times 882 \times 752$, $194 \times 466 \times 372$ and $180 \times 824 \times 702$.

The trained networks are first evaluated with the Peak Signal to Noise Ratio (PSNR) and the Structural SIMilarity index (SSIM). Then, we post-process the reconstructed volumes with Otsu segmentation [18], and we compute the DICE index between the segmented reconstructed volumes and the segmented ground-truth data. Also we compute the ratio between the segmented bone volume and the total volume (BV/TV) that we compare with the one of ground-truth data. These metrics help better reflect the capability of the networks to preserve bone microstructure information.

Since CWGAN and CWGAN-VGG produce stochastic outputs, we average each voxel of 10 generated outputs to produce the volume for evaluation. Note that in our tests, increasing this number does not improve the performance.

Training is performed on 64x64 patches from 1,992 different 2D slices for a total of 297,976 patches, 20% of which are used for validation. The evaluation is performed by averaging metrics on the 3 test volumes.

The generator is a 16-layer Convolutional Neural Network (CNN) with 128 3x3 filters in each layer, except for the last layer which has only one since the output is the generated image. We used the same discriminator structure as in [5]. For both the discriminator and the generator, Leaky ReLU activations are used with parameter 0.3 and He initialization [19], except for the output of the discriminator that has no activation function. Optimization is performed with Adam algorithm [20] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is 10^{-6} - except for WGAN-VGG where it is 10^{-5} -, with a batch size of 128 and 7,000 epochs. We took $\lambda_1 = 10$ for all the algorithms that include a VGG loss. For one update of the discriminator, we update the generator 4 times.

For a fair comparison, the kernel size, batch size, learning rate, number of generator updates and λ_1 have all been optimized for every single network, on the validation set. Computations are performed on a NVIDIA Tesla V100 GPU, and training of one network takes approximately 30 hours.

	PSNR	SSIM	DICE	BV/TV
FBP	15.96	0.491	0.880	0.2317
Det-CWGAN	23.05	0.634	0.928	0.1951
CWGAN	25.41	0.739	0.939	0.2043
WGAN-VGG	25.10	0.739	**0.952**	0.2154
Det-CWGAN-VGG	25.20	0.696	0.948	0.2096
CWGAN-VGG	**26.00**	**0.753**	0.951	**0.2091**

Table 1: Metrics computed on the 3 test volumes. PSNR, DICE and BV/TV were computed by stacking the 3 - potentially segmented - volumes, SSIM is the average value of the metric computed on each of them. The ground-truth BV/TV is 0.2077.

3.2 Results

Reconstructions of one of the three testing volumes are shown in Fig 2, along with a region of interest. Note that the 2 other testing volumes as well as the 10 training volumes all have a significantly different shape, which attests for the ability of the networks to generalize. In the second row of Fig 2, we notice that Det-CWGAN is the only one that fails to recover some continuous structure of the bone, which is a key feature for bones imaging. However, for the others there is no clear indication that one reconstruction outperforms the others.

Table 1 allows to better distinguish between the obtained reconstructions. Results show that CWGAN-VGG performs the best in terms of PSNR, SSIM and BV/TV ratio, which is an important metric for bone microarchitecture. The algorithm presents a DICE index that is only slightly inferior to the one of WGAN-VGG, which outperforms the other algorithms for this metric. We also note that both CWGAN and CWGAN-VGG perform better than their deterministic version for the tested metrics.

4 Discussions

Our results suggest that using a perceptual loss for training our generator as in [5] allows the network to produce volumes that are closer to the real ones, in terms of pixel-wise metrics and structural-specific evaluation methods. Indeed, CWGAN-VGG outperforms CWGAN, whether it is on the deterministic or stochastic version.

We also argued that conditioning the discriminator would produce outputs that better match the FBP they are conditioned on. This is the case in our tests, where CWGAN-VGG produced better results than WGAN-VGG for 3 out of the 4 tested metrics, and the DICE index for both methods is very close.

We also find that it is less optimal for the network to learn a Dirac conditional distribution. Indeed, the strategy of averaging several stochastic outputs gave a significant improvement compared to using a deterministic network for both the CWGAN and CWGAN-VGG networks. Along with improvements on those metrics, the non-deterministic outputs that CWGAN and CWGAN-VGG produce might be very

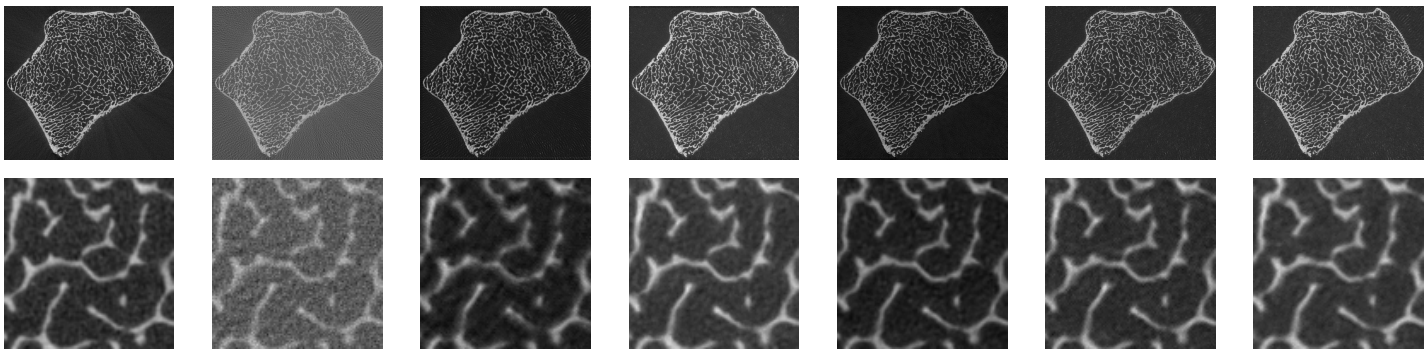

Figure 2: Entire slice (first row) and zoom on this slice (second row) of the bone volume reconstructed with different architectures, with pixel intensities between 0 and 1

useful for practitioners in order to get a level of confidence for specific regions of interest in the reconstruction.
In order to fully show the potential of CWGAN-VGG, work is under progress to train and test it on different noise configurations to get a more robust model and evaluate it on more realistic data for even more metrics.

5 Conclusion

We proposed a new framework called CWGAN-VGG for the task of enhancing the quality of a FBP acquired from low-dose projections. It combines both the ability of GANs to learn conditional probabilities and the preservation of key structural information provided by perceptual losses. We showed the benefits provided by both conditioning the discriminator with the low-dose FBP and adding a perceptual loss to train the generator. We also showed the improvement on the evaluated metrics when using a non-deterministic network. Our resulting architecture thus outperformed state-of-the-art ones that rely on similar methods, for PSNR and other metrics, on CT bones data.

References

[1] 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* (2018), pp. 1–11. DOI: 10.1007/s10851-018-0858-1.

[2] F. Peyrin and K. Engelke. "CT Imaging: Basics and New Trends". *Handbook of Particle Detection and Imaging*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 883–915. DOI: 10.1007/978-3-642-13271-1_36.

[3] P. Gilbert. "Iterative methods for the three-dimensional reconstruction of an object from projections". *J. Theor. Biol.* 36.1 (1972).

[4] E. Sidky et al. "Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle–Pock algorithm". *Phys. Med. Biol.* 57.10 (2012), 3065–3091. DOI: 10.1088/0031-9155/57/10/3065.

[5] Q. Yang et al. "Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss". *IEEE Trans Med Imaging* 37.6 (2018), pp. 1348–1357. DOI: 10.1109/TMI.2018.2827462.

[6] Y. Wang et al. "3D conditional generative adversarial networks for high-quality PET image estimation at low dose". *NeuroImage* 174 (2018), pp. 550 –562. DOI: https://doi.org/10.1016/j.neuroimage.2018.03.045.

[7] H. Chen et al. "Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network". *IEEE Trans Med Imaging* 36.12 (2017), 2524–2535. DOI: 10.1109/tmi.2017.2715284.

[8] I. Goodfellow et al. *Advances in Neural Information Processing Systems*. 2014.

[9] X. Yu, Y. Qu, and M. Hong. "Underwater-GAN: Underwater Image Restoration via Conditional Generative Adversarial Network". *Pattern Recognition and Information Forensics*. Cham: Springer International Publishing, 2019, pp. 66–75. DOI: 10.1007/978-3-030-05792-3_7.

[10] M. Mirza and S. Osindero. "Conditional Generative Adversarial Nets". *arXiv e-prints*, arXiv:1411.1784 (Nov. 2014).

[11] P. Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". *CoRR* abs/1611.07004 (2016).

[12] J. Adler and O. Öktem. "Deep Bayesian Inversion". *arXiv e-prints*, arXiv:1811.05910 (Nov. 2018).

[13] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein GAN". *arXiv e-prints*, arXiv:1701.07875 (2017).

[14] I. Gulrajani et al. "Improved Training of Wasserstein GANs". *Advances in Neural Information Processing Systems*. 2017.

[15] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". *arXiv e-prints*, arXiv:1409.1556 (Sept. 2014).

[16] R. Zhang et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". *arXiv e-prints*, arXiv:1801.03924 (Jan. 2018).

[17] W. Van Aarle et al. "The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography". *Ultramicroscopy* 157 (2015). DOI: https://doi.org/10.1016/j.ultramic.2015.05.002.

[18] N. Otsu. "A Threshold Selection Method from Gray-Level Histograms". *IEEE Trans. Syst. Man Cybern. Syst.* 9.1 (1979).

[19] K. He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". *arXiv e-prints*, arXiv:1502.01852 (Feb. 2015).

[20] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". *arXiv e-prints*, arXiv:1412.6980 (Dec. 2014).