Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear

Jianing Wang(✉), Yiyuan Zhao, Jack H. Noble, and Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235, USA
jianing.wang@vanderbilt.edu

Abstract. We propose an approach based on a conditional generative adversarial network (cGAN) for the reduction of metal artifacts (RMA) in computed tomography (CT) ear images of cochlear implants (CIs) recipients. Our training set contains paired pre-implantation and post-implantation CTs of 90 ears. At the training phase, the cGAN learns a mapping from the artifact-affected CTs to the artifact-free CTs. At the inference phase, given new metal-artifact-affected CTs, the cGAN produces CTs in which the artifacts are removed. As a pre-processing step, we also propose a band-wise normalization method, which splits a CT image into three channels according to the intensity value of each voxel and we show that this method improves the performance of the cGAN. We test our cGAN on post-implantation CTs of 74 ears and the quality of the artifact-corrected images is evaluated quantitatively by comparing the segmentations of intra-cochlear anatomical structures, which are obtained with a previously published method, in the real pre-implantation and the artifact-corrected CTs. We show that the proposed method leads to an average surface error of 0.18 mm which is about half of what could be achieved with a previously proposed technique.

Keywords: Conditional generative adversarial networks
Metal artifact reduction · Cochlear implants

1 Introduction

Metallic implants in the human body can cause artifacts in computed tomography (CT) scans. Methods for the reduction of metal artifacts (RMA) in CTs have been investigated for nearly 40 years [1]. RMA algorithms can be roughly divided into three groups: physical effects correction, interpolation in the projection domain, and iterative reconstruction [2]. Despite these efforts, developing RMAs for dense metal implants and for multiple metallic objects in the field of view remains challenging and there is no known universal solution [1].

Conditional generative adversarial networks (cGANs) [3, 4] have emerged as a general-purpose solution to image-to-image translation problems. We propose an approach based on cGANs for RMA. At the training phase, a cGAN learns a mapping

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11070, pp. 3–11, 2018.
https://doi.org/10.1007/978-3-030-00928-1_1

from the artifact-affected CTs to the artifact-free CTs. At the inference phase, given an artifact-affected CT, the cGAN produces an artifact-corrected image. We apply our method to CT ear images of cochlear implants (CIs) recipients and get remarkable results.

Compared to the current leading traditional RMA methods, which generally necessitate the raw data from CT scanners [1], our approach is a post reconstruction processing method for which the raw data is not required. Our results also indicate that the post reconstruction processing methods, which have been considered to be ineffective [1], can in fact be effective. To the best of our knowledge, published RMA methods based on machine learning either depend on existing traditional RMA methods or require post-processing of the outputs produced by machine learning models [2, 5, 6]. Ours is unique in being able to synthesize directly an artifact-free image from an image in which artifacts are present.

2 Background

The cochlea is a spiral-shaped cavity that is part of the inner ear. CIs are surgically implanted neural prosthetic devices for treating severe-to-profound hearing loss [7] and are programmed postoperatively by audiologists to optimize outcomes. Accurately localizing the CI electrodes relative to the intra-cochlear anatomy in post-implantation CTs (Post-CTs) can help audiologists to fine-tune and customize the CI programming. This requires the accurate segmentation of the scala-tympani (ST), the scala-vestibuli (SV), and the modiolus (MOD) in these images (Fig. 1). Noble *et al*. have developed an active shape model-based method [8], which we refer to as NM, to segment ST, SV, and MOD in pre-implantation CTs (Pre-CTs). To the best of our knowledge, NM is the most accurate published automatic method for intra-cochlear anatomy segmentation in Pre-CTs. However, NM cannot be directly applied to Post-CTs due to the strong artifacts produced by the electrodes. Reda *et al*. have proposed a library-based method [9], which we refer to as RM, to segment ST, SV, and MOD in Post-CTs but it leads to segmentation errors that are substantially larger than errors obtained with pre-operative images. Here, we propose an alternative. First, we remove the artifacts from the Post-CTs using a novel RMA. Next, we apply NM to the processed images. We show that this novel approach leads to an error that is about half the error obtained with RM.

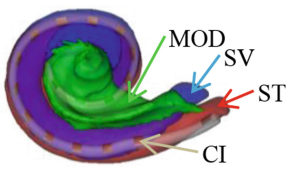

Fig. 1. An illustration of intra-cochlear anatomical structures and CI electrodes.

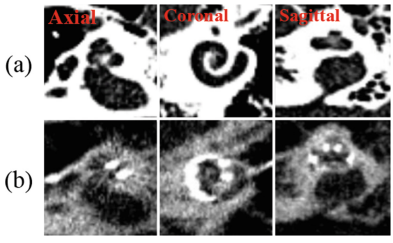

Fig. 2. Three orthogonal views of (a) the Pre-CT and (b) the Post-CT of an example ear.

3 Dataset

Our dataset consists of Pre- and Post-CT pairs of 164 ears, all these CTs have been acquired with the CIs recipients in roughly the same position. The CTs are acquired with several conventional scanners (GE BrightSpeed, LightSpeed Ultra; Siemens Sensation 16; and Philips Mx8000 IDT, iCT 128, and Brilliance 64) and a low-dose flat-panel volumetric CT scanner (Xoran Technologies xCAT® ENT). The typical voxel size is $0.25 \times 0.25 \times 0.3$ mm^3 for the conventional CTs (cCTs), and $0.4 \times 0.4 \times 0.4$ mm^3 for the low-dose CTs (lCTs). The 164 ears are randomly partitioned into a set of 90 for training and a set of 74 for testing. 82 of the 90 ears used for training have Post-CTs of type lCT and Pre-CTs of type cCT, and the remaining 8 ears have both Post-CTs and Pre-CTs of type cCT. 62 of the 74 ears for testing have Post-CTs of type lCT and Pre-CTs of type cCT, and the remaining 12 ears have both Post-CTs and Pre-CTs of type cCT.

4 Method

4.1 cGAN

In this work we rely on the cGAN framework proposed by Isola *et al*. [4]. A cGAN consists of a generator G and a discriminator D. The total loss can be expresses as

$$L = arg\ \min_G\ \max_D L_{cGAN}(G, D) + \lambda L_{L_1}(G) \tag{1}$$

wherein

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log(D(x, y))] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \tag{2}$$

is the adversarial loss and $L_{L1}(G)$ is the L1 norm loss. G is tasked with producing an artifact-corrected image from a Post-CT x and a random noise vector z. The image generated by G should not be distinguishable from the real artifact-free Pre-CT y by D, which is trained to do as well as possible to detect G's "fakes". The generated images need to be similar to y in the L1 sense.

We explore two generator architectures: (1) a U-net [10] (UNet), and (2) a network that contains two stride-2 convolutions, 9 residual blocks, and two fractionally strided convolutions with stride $\frac{1}{2}$ [11–14] (ResNet). For the discriminator, we use a 70×70 PatchGAN [4, 15] that aims to determine whether 70×70 overlapping image patches are real or fake. We run the PatchGAN convolutationally across the image, averaging all responses to provide the ultimate output of D.

4.2 Image Pre-processing

The Pre-CTs are registered to the Post-CTs using intensity-based rigid registration techniques. The registrations have been visually inspected and confirmed to be accurate. 3D patch pairs that contain the cochlea are cropped from the Pre- and Post-CTs,

i.e., paired patches contain the same cochlea; one patch with and the other without the implant (Fig. 2). These patches are then upsampled to $0.1 \times 0.1 \times 0.1$ mm^3.

We apply a band-wise intensity normalization (BWN) to the Post-CT patches that acts as a piecewise linear stretch. For each 3D patch, we calculate the 2% percentile (p_2), the 98% percentile (p_{98}) and the 99.95% percentile ($p_{99.95}$) of the intensity values. Then the patch is separated into three channels: first, we copy the whole patch into channels 1, 2, and 3; second, the intensity values in channel 1, 2, and 3 are clamped to the ranges p_2 to $(p_2 + p_{98})/2$, $(p_2 + p_{98})/2$ to p_{98}, and p_{98} to $p_{99.95}$, respectively; and finally each channel is normalized to the -1 to 1 range. As discussed later, this heuristic improves some of our results.

For each Pre-CT patch, the intensity values are also clamped to the range between the bottom 1% and the top 1% voxel values. Then the patch is normalized to the -1 to 1 range.

4.3 Evaluation

The quality of the artifact-corrected images is evaluated quantitatively by comparing the segmentations of ST, SV, and MOD obtained with NM applied to the real Pre-CTs with the results obtained when applying NM to the artifact-corrected CTs. The output of NM are surface meshes of ST, SV, and MOD that have a pre-defined number of vertices. There are 3344, 3132, and 17947 vertices on the ST, SV, and MOD surfaces, respectively, for a total of 24423 vertices. Point-to-point errors (P2PEs), computed as the Euclidean distance in millimeter, between the corresponding vertices on the meshes generated from the real Pre-CTs and the meshes generated from artifact-corrected images are calculated to quantify the quality of the artifact-corrected images.

To compare the proposed method to the state of the art, we also segment ST, SV, and MOD using RM in Post-CTs. The output of RM are surface meshes for ST, SV, and MOD that have the same anatomical correspondences as the meshes generated by NM. The P2PEs between the corresponding vertices on the meshes generated from real Pre-CTs by using NM and the meshes generated from Post-CTs by using RM are calculated and serve as baseline for comparison.

To evaluate our BWN approach, we train the cGAN with and without such pre-processing step, and we compare the results that are generated with each strategy.

5 Experiments

The PyTorch implementation of the cGAN provided by Isola *et al*. [4] is used in our experiments. Since the cGAN is a 2D network, we train our cGAN on the axial view of the CTs. Input images are 2D 3-channel images, each of those is the slice of the 3D 3-channel Post-CT patch. As the current implementation of the cGAN requires the number of input and output channels to be the same, the target images of the cGAN are 2D 3-channel images in which each channel is identical and is the patch's slice in the Pre-CT that matches the patch's slice in the Post-CT used as input. In total 14346 paired Pre- and Post-CT 2D slices are used for training. To augment the number of training pairs, each slice is resized to 256×256 pixels and then padded to 286×286

pixels. Sub-images of 256 × 256 pixels are randomly cropped at the same location in the paired Pre- and Post-CT slices during the training. Horizontal flipping of the training pairs is also applied during the training to further augment the number of training pairs. The default value of $\lambda = 100$ is used to weigh the L1 distance loss. The cGAN is trained alternatively between one stochastic gradient descent (SGD) step on D, then one step on G, using minibatch size of 1 and the Adam solver [16] with momentum 0.5. The cGAN is trained for 200 epochs in which a fixed learning rate of 0.0002 is applied in the first 100 epochs, and a learning rate that is linearly reduced to zero in the second 100 epochs. The output of D, which is recorded every 100 iterations in each epoch, represents the probability that an image is real rather than fake. D is unable to differentiate between the real and the fake images when the output of D is 0.5 [17]. For each epoch, we calculate the median of the outputs of D and the model that is saved at the epoch in which the median is the closest to 0.5 among the 200 epochs is used as our final cGAN model. At the testing phase, the cGAN processes the testing 3D Post-CTs patches slice by slice, then the artifact-corrected slices are stacked to create 3D patches. These 3D patches are resized to their original sizes and translated back to their original spatial locations in the Post-CTs. Then we use NM to segment ST, SV, and MOD in these images.

6 Results

For each testing ear, we calculate the P2PEs of the 24423 vertices, and we calculate the maximum (Max), mean, median, standard deviation (STD), and minimum (Min) of the P2PEs. Figure 3a shows the boxplots of these statistics for the 74 testing ears, wherein Baseline denotes segmenting the intra-cochlear anatomical structures in Post-CTs using RM, Unet-1 denotes using UNet as the generator of the cGAN but without the BWN, Unet-3 denotes using UNet with BWN, ResNet-1 denotes using ResNet without BWN, and ResNet-3 denotes using ResNet with BWN. The means of Max, median, and STD of the P2PEs of the five approaches are shown in Table 1.

Table 1. The means of Max, median, and STD of the P2PEs of the five approaches (mm).

	ResNet-3	ResNet-1	UNet-3	UNet-1	Baseline
Max	0.575	0.599	0.625	0.660	0.912
Median	0.191	0.211	0.214	0.225	0.409
STD	0.084	0.085	0.091	0.097	0.133

Table 1 and the plots show that the cGAN-based methods substantially reduce the P2PEs obtained with the baseline approach. The median of the baseline method is 0.366 mm, whereas the median of ResNet-3 is 0.173 mm, which is less than the half of that of the baseline method. We perform paired t-tests on the Max, median, and STD

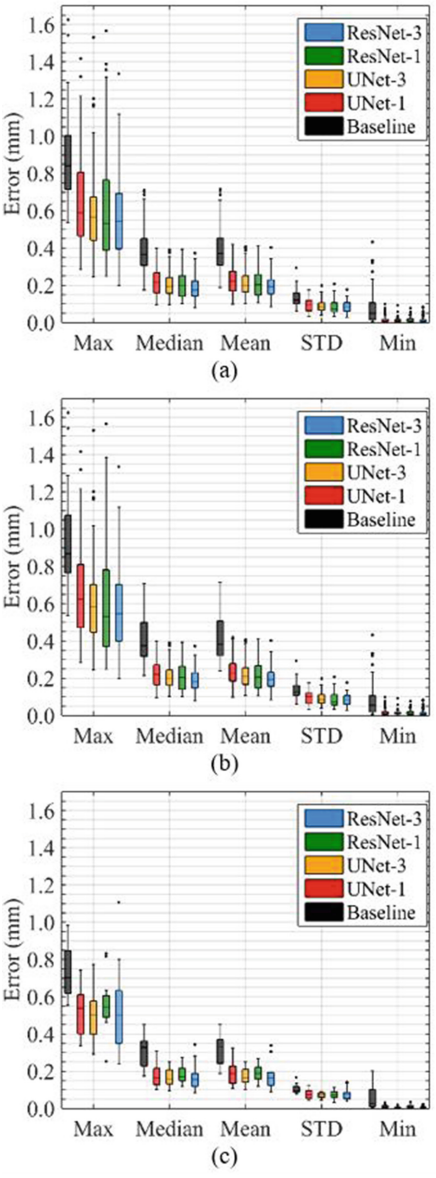

Fig. 3. Boxplot of P2PEs of (a) the 74 ears, (b) the 62 ears scanned by the conventional scanners, (c) the 12 ears scanned by the low-dose scanners.

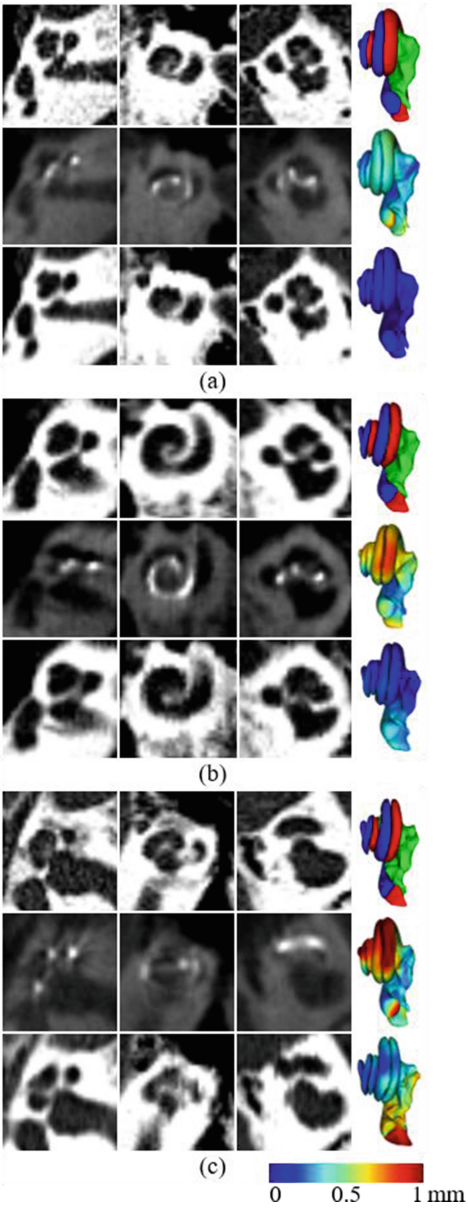

Fig. 4. Three example cases in which ResNet-3 achieves (a) best, (b) average, and (c) worst performances. The left, middle, and right column of each case show the axial, coronal, and sagittal views of the 3D patches.

values of the baseline method between cGAN-based methods, the results show that the cGAN-based methods significantly reduce the P2PEs compared to the baseline method for Max, median, and STD ($p < 0.05$). We also perform paired t-tests between the four cGAN-based approaches. ResNet-3 leads to the lowest median error among the four. Pairwise t-tests show that the difference is statistically significant ($p < 0.05$). There is a substantial but not statistically significant difference between the Max and STD of ResNet-3 and of the other three approaches ($p > 0.05$). For ResNet, applying BWN results in statistically significant lower point-to-point median errors ($p < 0.05$). There is a visible but not statistically significant difference between the medians of UNet-3 and UNet-1 ($p > 0.05$). These results show that BWN affects the architectures differently but comparing the boxplots of UNet-3 and UNet-1, and those of ResNet-1 and ResNet-3, it is apparent that the interquartile ranges of the distributions is reduced when applying BWN.

Figure 3b shows the boxplots of the 62 lCTs, which show the same trend as Fig. 3a. Figure 3c shows the boxplot of the 12 cCTs, it also shows the same trend except that the interquartile ranges of the distributions are not reduced when applying BWN. It could be that the BWN approach does not generalize well to all image types but we also note that we only have 8 lCTs in the training set and 12 lCTs in the testing set. It is thus difficult to draw hard conclusions at this point. We are acquiring more data to address this issue.

Figure 4 shows three example cases in which our proposed method (ResNet-3) achieves (a) the best, (b) an average, and (c) the worst performances, in the sense of the medians of the P2PEs. For each case, the top row shows three orthogonal views of the Pre-CT and the meshes generated when applying NM to this CT. The ST, SV, and MOD surfaces are shown in red, blue, and green, respectively. The middle row shows the Post-CT and the meshes generated when applying RM to this CT. The bottom row shows the output of cGAN and the meshes generated when applying NM to this image. The meshes in the middle and the bottom rows are color-coded with the P2PE at each vertex on the meshes. These images confirm the results presented in Fig. 3. Notably, even in the worst case, segmentation errors are on the order of 0.2 mm for a large portion of the structures. We also note that the output images of cGAN show good slice-to-slice consistency, i.e., there is no "jaggedness", in the coronal and the sagittal views, although the cGAN is a 2D network and it is trained on the axial view only.

RM has also been applied to these artifact-corrected images. The results are better than with the artifact-affected images but statistically worse than those obtained with NM on the synthetic images.

7 Summary

We have developed a method that is capable of removing metallic artifacts in CT ear images of CI recipients. This new method has been tested on a large dataset, and we show that it permits to significantly reduce the segmentation error for intra-cochlear structures in post-operative images when compared to the current state-of-the-art

approach. The technique we propose can be potentially used for RMA in CTs at other body sites. It would also be interesting to explore the possibility of applying such idea to correct various types of artifact in other types of medical images. Although the results we have obtained so far are encouraging, we believe there is room for improvement. It is possible for instance that a 3D network would be better than a 2D network. To test this, we have extended the current architecture to a 3D one. At the time of writing the results obtained with the latter are inferior to those presented herein. Training of and training set size required for a 3D network are issues we are currently addressing.

Acknowledgements. This work has been supported in parts by NIH grants R01DC014037 and R01DC014462 and by the Advanced Computing Center for Research and Education (ACCRE) of Vanderbilt University. The content is solely the responsibility of the authors and does not necessarily represent the official views of this institute.

References

1. Gjesteby, L., et al.: Metal artifact reduction in CT: where are we after four decades? IEEE Access. **4**, 5826–5849 (2016)
2. Zhang, Y., Yu, H.: Convolutional neural network based metal artifact reduction in x-ray computed tomography. arXiv:1709.01581 (2017)
3. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014)
4. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv:1611.07004 (2017)
5. Park, H.S., et al.: Machine-learning-based nonlinear decomposition of CT images for metal artifact reduction. arXiv:1708.00244 (2017)
6. Gjesteby, L., et al.: Reducing metal streak artifacts in CT images via deep learning: pilot results. In: The 14th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, vol. 14(6), pp. 611–614 (2017)
7. Cochlear Implants. National Institute on Deafness and Other Communication Disorders, No. 11-4798 (2011)
8. Noble, J.H., et al.: Automatic segmentation of intracochlear anatomy in conventional CT. IEEE Trans. Biomed. Eng. **58**(9), 2625–2632 (2011)
9. Reda, F.A., et al.: Automatic segmentation of intra-cochlear anatomy in post-implantation CT. In: Proceedings of the SPIE 8671, Medical Imaging 2013: Image-Guided Procedures, Robotic Interventions, and Modeling, 86710I (2013)
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597 (2015)
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 (2015)
13. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593 (2017)

14. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
15. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv:1609.04802, (2017)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
17. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv:1406.2661 (2014)