Medical Image Analysis 58 (2019) 101553

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Metal artifact reduction for the segmentation of the intra cochlear anatomy in CT images of the ear with 3D-conditional GANs

Jianing Wang*, Jack H. Noble, Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235, USA

ARTICLE INFO

Article history:
Received 6 November 2018
Revised 14 June 2019
Accepted 3 September 2019
Available online 4 September 2019

Keywords:
3D-conditional generative adversarial networks
Cochlear implants
Segmentation, Metal artifact reduction

ABSTRACT

Cochlear implants (CIs) are surgically implanted neural prosthetic devices that are used to treat severe-to-profound hearing loss. These devices are programmed post implantation and precise knowledge of the implant position with respect to the intra cochlear anatomy (ICA) can help the programming audiologists. Over the years, we have developed algorithms that permit determining the position of implanted electrodes relative to the ICA using pre- and post-implantation CT image pairs. However, these do not extend to CI recipients for whom pre-implantation CT (Pre-CT) images are not available. This is so because post-operative images are affected by strong artifacts introduced by the metallic implant. To overcome this issue, we have proposed two methods to segment the ICA in post-implantation CT (Post-CT) images, but they lead to segmentation errors that are substantially larger than errors obtained with Pre-CT images. Recently, we have proposed an approach that uses 2D-conditional generative adversarial nets (cGANs) to synthesize pre-operative images from post-operative images. This permits to use segmentation algorithms designed to operate on Pre-CT images even when these are not available. We have shown that it substantially and significantly improves the results obtained with methods designed to operate directly on post-CT images. In this article, we expand on our earlier work by moving from a 2D architecture to a 3D architecture. We perform a large validation and comparative study that shows that the 3D architecture improves significantly the quality of the synthetic images measured by the commonly used MSSIM (Mean Structural SIMilarity index). We also show that the segmentation results obtained with the 3D architecture are better than those obtained with the 2D architecture although differences have not reached statistical significance.

1. Introduction

The cochlea is a component of the inner ear. It is a spiral-shaped cavity located inside the bony labyrinth that contains two main cavities: the scala vestibuli and the scala tympani. The modiolus is a porous bone around which the cochlea is wrapped that hosts the cochlear nerve and the spiral ganglions. These structures, which are the ones this article focuses on, are shown in Fig. 1 and will be referred together as ICA (Intra Cochlear Anatomy) structures. Cochlear implants (CIs) are surgically implanted neural prosthetic devices that are used to treat severe-to-profound hearing loss (National Institute on Deafness and Other Communication Disorders, 2011). CIs bypass the normal acoustic hearing process by replacing it with direct stimulation of neural pathways using an implanted electrode array. After implantation CIs are programmed by audiologists who adjust a processor's settings to send the appropriate signals to each of the implant's electrode. The efficacy of the CI programming is sensitive to the spatial relationship between the CI electrodes and ICA structures. Providing accurate information about the position of the contacts with respect to these structures can thus help audiologists to fine-tune and customize CI programming (Noble et al., 2013). To provide this information we have developed a number of algorithms that permit determining the position of implanted electrodes relative to the ICA using pre- and post-implantation CTs.

Pre-implantation CT (Pre-CT) images and post-implantation CT (Post-CT) images of the ear are acquired before and after the surgery, respectively. The CI electrodes are localized in the Post-CT images using the automatic methods proposed by Zhao et al. (2018, 2019). It is difficult to directly localize the ICA in the Post-CT images due to the strong artifacts produced by the metallic CI electrodes. The ICA is thus localized in the Pre-CT images and the position of the CI electrodes relative to the ICA is obtained by registering the Pre-CT images and the Post-CT images. In order to localize the ICA in the Pre-CT images, where the ICA is only partially visible, Noble et al. (2011) have developed a method, which we

* Corresponding author.
E-mail address: jianing.wang@vanderbilt.edu (J. Wang).

https://doi.org/10.1016/j.media.2019.101553
1361-8415/© 2019 Elsevier B.V. All rights reserved.

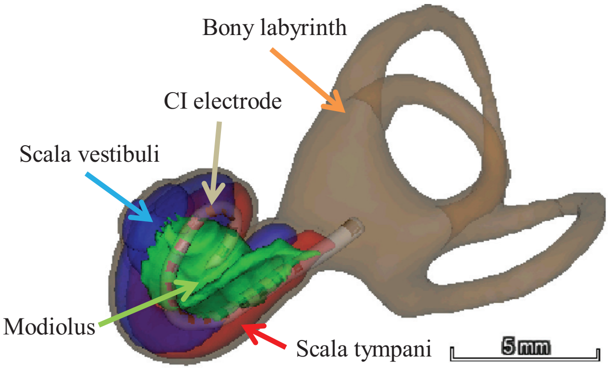

Fig. 1. An illustration of intra cochlear anatomical structures and CI electrodes.

refer to as PreCTseg for Pre-CT Segmentation. PreCTseg relies on a weighted active shape model created with high-resolution microCT scans of the cochlea acquired ex-vivo in which ICA structures are visible. The model is fitted to the partial information available in the Pre-CT images and used to estimate the position of structures not visible in these images.

This approach does not extend to CI recipients for whom a Pre-CT image is not available, which is the case for long-term recipients who were not scanned prior to surgery, for recipients for whom images cannot be retrieved, or for recipients implanted at institutions that use pre-operative MR images instead of CT images. To overcome this issue, Reda et al. (2014a, b) have proposed two methods to segment the ICA in Post-CT images. The first method, which we refer to as PostCTseg1 for Post-CT Segmentation unilateral, was developed for segmenting ICA structures in Post-CT images of CI recipients who have been implanted unilaterally (Reda et al., 2014a). PostCTseg1 relies on the intra-subject symmetry in cochlear anatomy across ears. It first segments the ICA of the contralateral normal ear and then maps the segmented structures to the implanted ear. The second method, which we refer to as PostCTseg2 for Post-CT Segmentation bilateral, was developed for segmenting ICA structures in Post-CT images of CI recipients who have been implanted bilaterally (Reda et al., 2014b). PostCTseg2 first identifies the labyrinth in the Post-CT image by mapping a labyrinth surface that is selected from a library of labyrinth surfaces, and then uses the localized labyrinth in the image as a landmark to segment the scala tympani, the scala vestibuli, and the modiolus with a standard shape model-based segmentation method.

But, while using these methods it was observed that they could at times lead to results that lacked accuracy compared to other components of the processing pipeline on which we rely to provide programming guidance to the audiologists. For instance, we can localize contacts in electrode arrays with an average accuracy better than 0.15 mm (Zhao et al., 2018, 2019) when the segmentation error of PostCTseg1 applied to the unilateral cases, which we refer to as IU for Implanted Unilaterally, included in this study is 0.26 mm. The segmentation error of PostCTseg2 applied to the bilateral cases, which we refer to as IB for Implanted Bilaterally, is larger and reaches 0.44 mm. These observations led us to explore ways to improve our segmentation accuracy.

Generative adversarial nets (GANs) (Goodfellow et al., 2014) have been applied to various computer vision tasks and have produced impressive results. In particular, conditional generative adversarial networks (cGANs) (Mirza and Osindero, 2014) have emerged as a general-purpose solution to image-to-image translation problems. Inspired by these, we proposed an alternative for localizing the ICA in the Post-CT images (Wang et al., 2018). First we train 2D-cGANs introduced by Isola et al. (2017) to synthesize artifact-free images from the Post-CT images, i.e., we train a network whose input is a 2D slice in a volume in which the artifacts are present and whose output is the corresponding synthetic artifact-free image. Once this is done for all slices, the synthetic 2D images are stacked to each other and the PreCTseg method is applied to the synthetic volume. Results obtained with the PreCTseg method on the synthesized volumes and the real pre-CT volumes can then be compared to assess the efficacy of the artifact removal method. In this earlier study performed on 74 ears we show that this approach produces segmentation errors that are about half the segmentation errors that was obtained with the methods designed to operate on the post-operative images. In this article, we increase the size of our testing dataset to 124 ears and we explore ways to improve our method further by (1) using a 3D architecture rather than a 2D architecture and (2) modifying the training objective of the 3D-cGANs, which is a sum of an adversarial loss and a L1 reconstruction loss. The quality of the artifact-corrected images is evaluated quantitatively by computing the surface error between the segmentations of the ICA obtained with PreCTseg applied to the real Pre-CT images and to the artifact-corrected CT images. We further validate our method by comparing the ICA obtained with PreCTseg applied to the artifact-corrected CTs and those which are obtained with PostCTseg1 and PostCTseg2 directly applied to the Post-CT images. Finally, as is commonly done to assess the quality of images, we compare the mean structural similarity index (MSSIM) (Wang et al., 2004) between the real images and the synthetic images obtained with the 2D and 3D architectures.

Table A in the Appendix provides a list of abbreviations used throughout this work.

2. Material and methods

2.1. Training objectives

2.1.1. Adversarial loss

Typically, GANs are implemented by a system of a generative network (G) and a discriminative network (D) that are competing with each other. G learns a mapping between a latent space and a particular data distribution of interest, while D discriminates between instances from the true data distribution and candidates produced by G. The training objective of G is to increase the error rate of D, i.e., to fool D by producing synthesized candidates that appear to come from the true data distribution (Goodfellow et al., 2014). cGANs are a special case of GANs in which both G and D are conditioned on additional information that is used to direct the data generation process. This makes cGANs suitable for image-to-image translation task, where G is conditioned on an input image and generates a corresponding output image (Mirza and Osindero, 2014; Isola et al., 2017).

For our purpose, which is to eliminate the artifacts produced by the CI, we use cGANs that are conditioned on the artifact-affected Post-CT images. G thus produces an artifact-free image $G(x)$ from a Post-CT image x, and $G(x)$ should not be distinguishable from the real artifact-free Pre-CT image y by D, which is trained to do as well as possible to detect G's “fakes”. The output of D can be interpreted as the probability of an image to be generated by G rather than a true Pre-CT image. Therefore, the training objective of D is to assign a high value to $G(x)$ and a low value to y. Conversely, the training objective of G is to fool D to assign a low value to $G(x)$ and a high value to y. Thus, the adversarial loss of the cGANs can be expressed as:

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

2.1.2. Reconstruction loss

Previous research suggests that it is beneficial to mix the adversarial loss with a more traditional reconstruction loss, such as the L1 distance between $G(x)$ and y (Isola et al., 2017), which is defined as:

$$L_{L_1}(G) = \mathbb{E}_{x,y}[\|y - G(x)\|_1] \tag{2}$$

For our ultimate purpose, which is to localize the ICA in the Post-CT images, we are more concerned about the quality of the image content in the small region that encompasses the cochlea than in the other regions in an artifact-corrected CT, therefore we assign a higher weight to the voxels inside this region when calculating the L1 loss. To do so, we first create a bounding box that encloses the cochlea. With the number of voxels inside the bounding box equal to N_{in} and the number of voxels outside of the bounding box equal to N_{out}, we assign weights to the voxels inside and outside of the bounding box that are equal to $(N_{in} + N_{out})/N_{in}$ and 1, respectively. The weighted L1 (WL1) loss can then be expressed as shown in Eq. (3):

$$L_{\mathrm{WL}_1}(G) = \mathbb{E}_{x,y}[\|W \circ (y - G(x))\|_1] \tag{3}$$

in which W is the weighting matrix and $\circ$ is the element-wise multiplication operation.

2.1.3. Total loss

The total loss can be expressed as a combination of the adversarial loss and the reconstruction loss:

$$L = \arg\min_G \max_D L_{\mathrm{cGAN}}(G, D) + \alpha L_{\mathrm{WL}_1}(G) \tag{4}$$

wherein α is the weight of the WL1 term.

2.2. Architecture of the 3D-cGANs

Fig. 2 shows the architecture of our 3D-cGANs. The generator is a 3D network which consists of 3 convolutional blocks followed by 6 ResNet blocks (He et al., 2016), and another 3 convolutional blocks (Fig. 2, the sub-network on the left). As is done in Isola et al. (2017), dropout is applied to introduce randomness into the training of the generator. The input of the generator is a 1-channel 3D Post-CT image, and the output is a 1-channel 3D synthetic Pre-CT image. The discriminator is a fully convolutional network (Fig. 2, the sub-network on the right) that maps the input, which is the concatenation of a Post-CT image and the corresponding Pre-CT image (or a Post-CT image and the synthetic Pre-CT image), to a 3D array d, where each $d_{i,\,j,\,k}$ captures whether the $(i,\,j,\,k)$-th 3D patch of the input is real or fake. The ultimate output of the discriminator is a scalar obtained by averaging d.

3. Experiments

3.1. Dataset

Our dataset consists of Post- and Pre-CT image volume pairs of 252 ears, all these CT volumes have been acquired with the CIs recipients in roughly the same position. 24 Post-CT images and all of the 252 Pre-CT images were acquired with several conventional scanners referred to as cCT scanners (GE BrightSpeed, LightSpeed Ultra; Siemens Sensation 16; and Philips Mx8000 IDT, iCT 128, and Brilliance 64). The other 228 Post-CT images were acquired with a low-dose flat-panel volumetric CT scanner referred to as lCT scanner (Xoran Technologies xCAT® ENT). The typical voxel size is and $0.25 \times 0.25 \times 0.3\,\mathrm{mm}^3$ for the cCT images, and $0.4 \times 0.4 \times 0.4\,\mathrm{mm}^3$ for the lCT images. The 252 ears are randomly partitioned into a set of 90 ears for training, 25 ears for validation, and 137 ears for testing. After random assignment, there are 13 bilateral cases for which one ear has been assigned to the training (or validation) set and the other ear has been assigned to the testing set, the 13 ears of such cases are removed from the testing set so that no image from the same patient are used for both training and testing. Details about our image set can be found in Table 1.

Table 1
The number of ears and the type of CT scanner used to acquire the images in the training, validation, and testing sets. "lCT-cCT" denotes that the ear has been scanned by the lCT scanner postoperatively and a cCT scanner preoperatively, and "cCT-cCT" denotes that the ear has been scanned by a cCT scanner postoperatively and preoperatively.

Usage	Total number of the ears		Number of Post- and Pre-CT pairs	
			lCT-cCT	cCT-cCT
Training	90		82	8
Validation	25		21	4
Testing	124	88 IB ears	78	10
		36 IU ears	34	2

The Pre-CT images are registered to the Post-CT images using intensity-based affine registration techniques (Wells et al., 1996; Maes et al., 1997). The registrations have been visually inspected and confirmed to be accurate. We apply image augmentation to the training set by rotating each image by 20 small random angles in the range of −10 and 10° about the x-, y-, and z-axis, such that 60 additional training images are created from each original image. This results in a training set that is expanded to 5490 vol.

Because in our dataset the typical voxel size of the Post-CT images is $0.4 \times 0.4 \times 0.4$ mm^3, we first resample the CTs to $0.4 \times 0.4 \times 0.4$ mm^3, so that all of the images have the same resolution. 3D patch pairs that contain the cochlea are cropped from the Pre- and Post-CT images, i.e., paired patches contain the same cochlea; one patch with and the other without the implant (Fig. 3). The size of the patches is $38.4 \times 38.4 \times 38.4$ mm^3 (96 × 96 × 96 voxels). Each patch is clamped to the range 0.1-th–99.9-th percentiles of its intensity values. Then the patches are rescaled to the −1 to 1 range.

3.2. Optimization and inference

Our PyTorch implementation of the 3D-cGANs is adapted from the 2D implementation provided by Zhu et al. (2017). α introduced in Eq. (4) is set to its default value 100. In practice, the cochlea is at the center of each 3D patch and we simply use the central 56 × 56 × 56 voxels of the 3D patch as the bounding box for calculating the weights. The 3D-cGANs are trained alternatively between one stochastic gradient descent step on the discriminator, then one step on the generator, using a minibatch size of 1 and the Adam solver (Kingma and Bai, 2015) with momentum 0.5. The 3D-cGANs are 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. At the inference phase, given an unseen Post-CT patch, the generator produces an artifact-corrected image.

The MSSIM, which is introduced in Section 3.3.2, inside the 56 × 56 × 56 bounding box of the true Pre-CT images and the artifact-corrected CTs generated by the cGANs has been used to select the number of training epochs. To do so we run inference on the validation set every 5 epochs, the MSSIM is calculated for each of the ears, and the epoch where it achieves the highest median MSSIM is selected as the optimal epoch.

3.3. Evaluation

The proposed method is compared to the published baseline methods PostCTseg1 and PostCTseg2 as well as to our previous

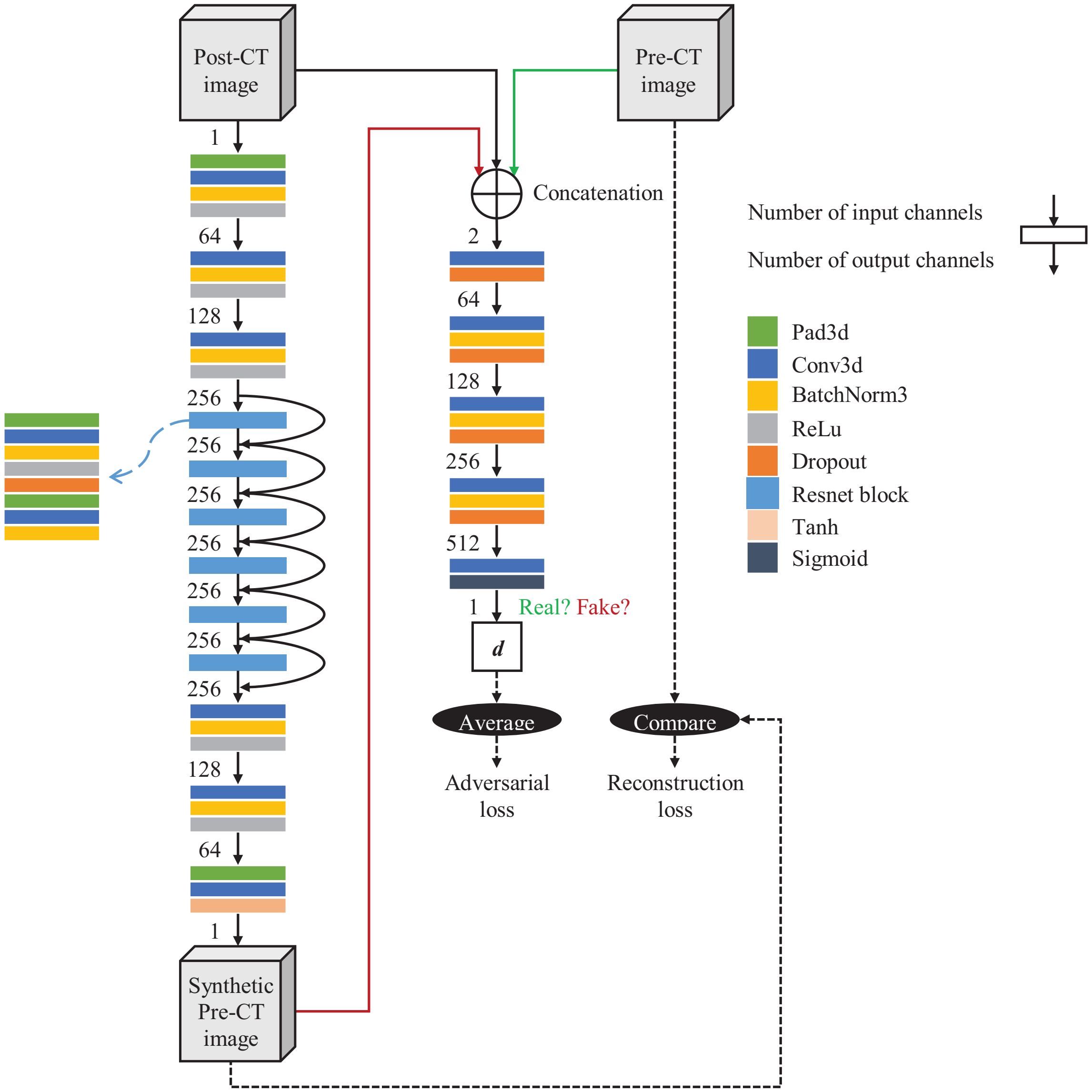

Fig. 2. An illustration of the architecture of the 3D-cGANs.

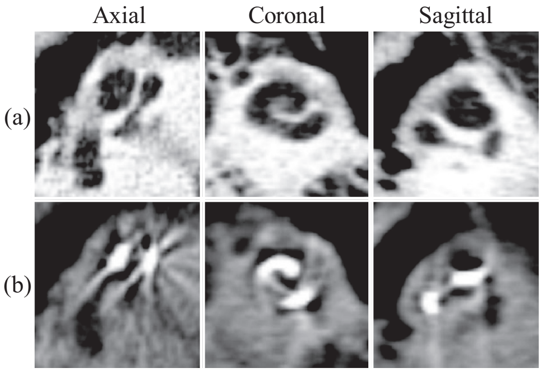

Fig. 3. Three orthogonal views of (a) the Pre-CT image and (b) the Post-CT image of an example ear.

2D-cGANs-based method. As we have done in our previous study (Wang et al., 2018), we upsample the voxel size of the CTs to 0.1 × 0.1 × 0.1 mm^3 to train and test the 2D-cGANs. This was done to improve slice-to-slice consistency. Due to memory limitations, this is not possible for the 3D-cGANs that are trained with volumes.

To evaluate the quality of the synthetic images independently from the segmentation results we compare the MSSIM between the original pre-CT images and the images produced with the 2D and 3D architectures. We also compare the performance of the 3D-cGANs trained using the weighted L1 loss and those which are trained using the original L1 loss.

3.3.1. *Point-to-point errors*

The effect of artifact reduction on segmentation accuracy is evaluated quantitatively by comparing the segmentation of the structures of interest (the scala tympani, the scala vestibuli, and the modiolus) obtained with PreCTseg applied to the real Pre-CT images with the results obtained when applying PreCTseg to the artifact-corrected CT images. Because PreCTseg is based on an active shape model approach, the outputs of PreCTseg are surface meshes of the scala tympani, the scala vestibuli, and the modiolus that have a pre-defined number of vertices, and each vertex corresponds to an anatomical location on the surface of the structures. There are 3344, 3132, and 17,947 vertices on the scala tympani, scala vestibuli, and modiolus surfaces, respectively, for a total

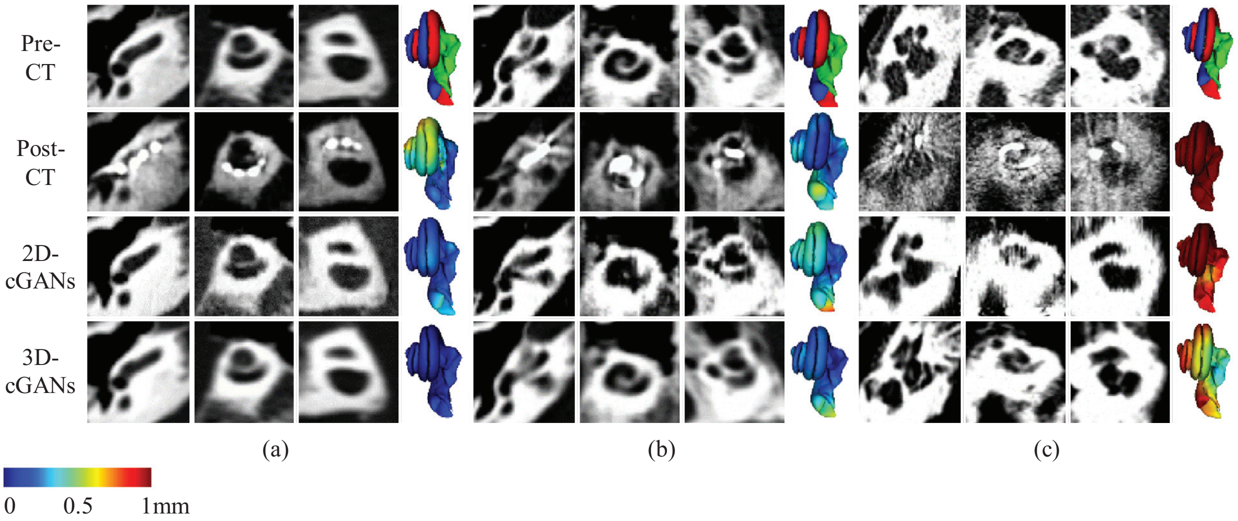

Fig. 4. Three example cases in which our proposed method leads to (a) a good, (b) average, and (c) poor results (see text for details).

of 24,423 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-CT images and the meshes generated from artifact-corrected images are calculated to quantify the quality of the artifact-corrected images.

To compare the P2PEs of the 3D-cGANs-based method to results obtained with the published baseline methods, we segment the scala tympani, the scala vestibuli, and the modiolus with PostCTseg1 and PostCTseg2 in the Post-CT images of the IU ears and the IB ears, respectively. The output of PostCTseg1 and PostCTseg2 are surface meshes for the scala tympani, the scala vestibuli, and the modiolus that have the same anatomical correspondences as the meshes generated by PreCTseg. The P2PEs between the corresponding vertices on the meshes generated with PreCTseg in the real Pre-CT images and the meshes generated with PostCTseg1 and PostCTseg2 in the Post-CT images serve as baselines for comparison.

To compare the 3D-cGANs-based method to our previous 2D-cGANs-based method, the P2PEs between the corresponding vertices on the meshes generated from the real Pre-CT images with PreCTseg and the meshes generated from artifact-corrected images generated by the 2D-cGANs with PreCTseg are also calculated.

3.3.2. Mean structural similarity

To compare the quality of the artifact-corrected images produced by our previously 2D-cGANs and those which are generated by the 3D-cGANs, we compare the MSSIM inside the 56 × 56 × 56 bounding box of the true Pre-CT images and the artifact-corrected CTs generated by the 2D- and the 3D-cGANs. The MSSIM between the true Pre-CT images and the Post-CT images serves as baseline for comparison. The MSSIM between the artifact-corrected CT image $G(x)$ and the true Pre-CT image y can be expressed as:

$$\mathrm{MSSIM}(G(x),\ y) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{SSIM}\left(g_j,\ y_j\right) \tag{5}$$

wherein SSIM(g_j,y_j) is the local structural similarity (SSIM) between g_j and y_j, which are the image contents at the j-th local window of $G(x)$ and y, and M is the number of local windows in the image. The local SSIM can be expressed as:

$$\mathrm{SSIM}\left(g_j,\ y_j\right) = \frac{\left(2\mu_{g_j}\mu_{y_j} + C_1\right)\left(2\sigma_{g_jy_j} + C_2\right)}{\left(\mu_{g_j}^2 + \mu_{y_j}^2 + C_1\right)\left(\sigma_{g_j}^2 + \sigma_{y_j}^2 + C_2\right)} \tag{6}$$

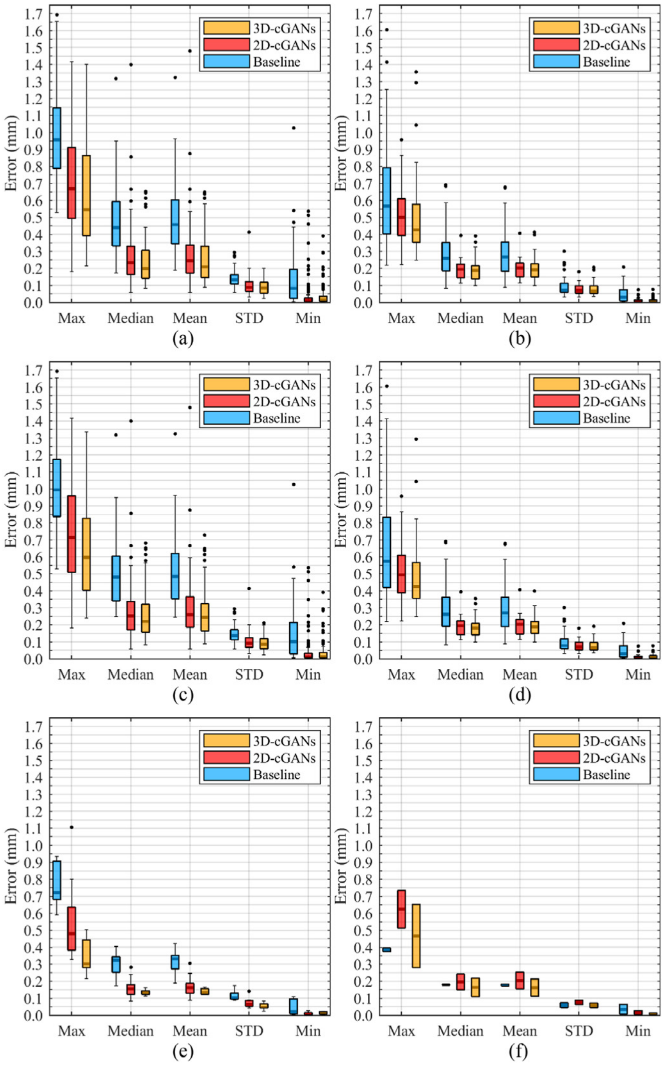

Fig. 5. Boxplot of P2PEs for (a) the 88 IB ears, (b) the 36 IU ears, (c) the 78 IB ears scanned by the lCT scanner postoperatively and the cCT scanners preoperatively, (d) the 34 IU ears scanned by the lCT scanner postoperatively and the cCT scanners preoperatively, (e) the 10 IB ears scanned by the cCT scanners postoperatively and preoperatively, (f) the 2 IU ears scanned by the cCT scanners postoperatively and preoperatively.

Table 2
The values of the MSSIM between the true Pre-CT images and the artifact-corrected CT images generated by the 2D- and the 3D-cGANs. "lCT-cCT" denotes that the ear has been scanned by the lCT scanner postoperatively and a cCT scanner preoperatively.

Image name	Baseline	2D-cGANs	3D-cGANs	Type of the Post- and Pre-CT pairs
Fig. 4a	0.771	0.891	0.971	lCT-cCT
Fig. 4b	0.499	0.780	0.931	lCT-cCT
Fig. 4c	0.348	0.473	0.552	lCT-cCT

in which μ_{g_j}, μ_{y_j}, σ_{g_j}, σ_{y_j}, and $\sigma_{g_j y_j}$ are the local means, standard deviations, and cross-covariance of g_j and y_j; C_1 and C_2 are constants to avoid instability when $\mu_{g_j}^2 + \mu_{y_j}^2$ or $\sigma_{g_j}^2 + \sigma_{y_j}^2$ are close to zero (Wang et al., 2004).

4. Results

Fig. 4 shows 3 example cases in which our proposed method leads to (a) good, (b) average, and (c) poor results. For each case, the first row shows three orthogonal views of the Pre-CT image and the meshes generated when applying PreCTseg to this CT. The scala tympani, the scala vestibuli, and the modiolus surfaces are shown in red, blue, and green, respectively. The second row shows the Post-CT image and the meshes generated when applying PostCTseg2 (or PostCTseg1) to this CT volume. The third and the last rows show the outputs of the 2D- and the 3D-cGANs and the meshes generated when applying PreCTseg to these images. The meshes from the second to the last rows are color-coded with the P2PE at each vertex on the meshes. Notably, even in the worst case, segmentation errors of the 3D-cGANs are lower than those of the baseline and of the 2D-cGANs. Note also the severity of the artifact in this case and the failure of the segmentation method designed to operate on the post-CT images. The values of the MSSIM and the type of Post- and Pre-CT pairs for these examples are listed in Table 2. In all cases the MSSIM between the original images and the synthetic images produced by the 3D networks is higher than between the original images and the synthetic images produced by the 2D networks. This is consistent with the visual appearance of the synthetic images as can be appreciated by comparing rows 3 and 4 of Fig. 4.

4.1. Point-to-point errors

For each testing ear, we calculate the P2PEs of the 24,423 vertices, and we calculate the maximum (Max), mean (Mean), median (Median), standard deviation (STD), and minimum (Min) of the P2PEs.

Figs. 5a and b show the boxplots of these statistics for the 88 IB ears and the 36 IU ears. PostCTseg2 and PostCTseg1 serve as the baseline method for the bilateral and unilateral cases, respectively. Fig. 5a shows that both the 2D- and the 3D-cGANs-based methods substantially reduce the P2PEs obtained with PostCTseg2 in the Post-CT images. The median of the baseline method is 0.439 mm, the medians of the 2D- and the 3D-cGANs-based approach are 0.233 mm and 0.198 mm, which are about half of the baseline method. We perform a Wilcoxon signed-rank test (McDonald, 2014) between the Max, Median, Mean, STD, and Min values obtained with the baseline method and the cGANs-based methods, and the resulting p-values are corrected using Holm-Bonferroni method (Holm, 1979). The results show that the cGAN-based methods significantly reduce the P2PEs compared to the baseline method ($p < 0.05$) (Table 3. 88 IB ears, row 1 and 2). We also perform a Wilcoxon signed-rank test between the 2D- and the 3D-cGANs-based approaches that shows that despite being visible

Table 3
p-Values of the two-sided and one-sided Wilcoxon signed-rank tests of the five statistics for the P2PEs of the 88 IB ears and the 36 IU ears.

Testing ears	Approaches to compare	Max		Median		Mean		STD		Min	
		Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
88 IB ears	Post-CT + PostCTseg22D-cGANs + PreCTseg	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	Post-CT + PostCTseg23D-cGANs + PreCTseg	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	2D-cGANs + PreCTseg3D-cGANs + PreCTseg	0.117	–	0.174	–	0.179	–	0.628	–	1.483	–
36 IU ears	Post-CT + PostCTseg12D-cGANs + PreCTseg	0.136	–	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.371	–	**0.002**	**< 0.001**
	Post-CT + PostCTseg13D-cGANs + PreCTseg	0.115	–	**0.002**	**0.001**	**0.002**	**0.001**	0.221	–	0.005	–
	2D-cGANs + PreCTseg3D-cGANs + PreCTseg	0.729	–	0.825	–	0.949	–	0.937	–	1.482	–

Note: Bold indicates cases that are significantly different (p-value less than 0.05). The p-values have been corrected using Holm-Bonferroni method.

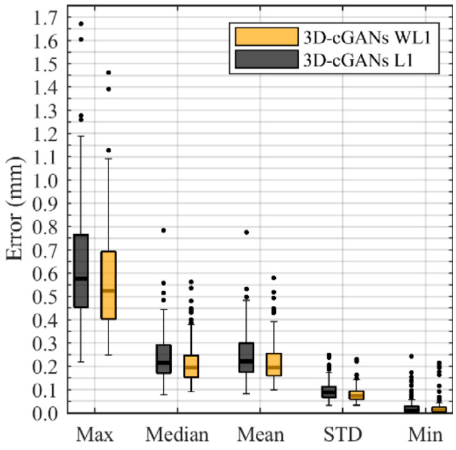

Fig. 6. Boxplot of P2PEs for the 124 testing ears. "3D-cGANs WL1" and "3D-cGANs L1" denote the results obtained with the 3D-cGANs which are trained using the weighted L1 loss and original L1 loss, respectively.

the difference between the results of the 2D- and the 3D-cGANs are not statistically significant ($p > 0.05$) (Table 3. 88 IB ears, row 3). Fig. 5b shows that both the 2D- and the 3D-cGANs-based methods reduce the P2PEs obtained with PostCTseg1 in the Post-CT images. The median of the baseline method is 0.260 mm, whereas the medians of the 2D- and the 3D-cGANs are 0.194 mm and 0.188 mm, respectively. A Wilcoxon signed-rank test shows that the cGANs-based methods significantly reduce the P2PEs compared to the baseline method for Median and Mean ($p < 0.05$) (Table 3. 36 IU ears, row 1 and 2). There is a visible but not statistically significant difference between the Max of the cGANs-based method and the baseline ($p > 0.05$) (Table 3. 36 IU ears, row 1 and 2). There is a visible but not statistically significant difference between the results of the 2D- and the 3D-cGANs ($p > 0.05$) (Table 3. 36 IU ears, row 3).

Figs. 5c and d show the boxplots of the statistics of the 78 IB ears and the 34 IU ears that have been scanned with the lCT scanner postoperatively and the cCT scanners preoperatively. Table 4 shows the results of the Wilcoxon signed-rank tests. These show the same trend as Fig. 5a and b and Table 3.

Fig. 5e and f show the boxplots of the statistics of the 10 IB ears and the 2 IU ears that have been scanned with the cCT scanners postoperatively and preoperatively. At the time of writing, we are not able to draw strong conclusions form these two plots because we only have a very limited number of such images but the trends are similar to those obtained with the other datasets.

Fig. 6 shows the boxplots of the statistics for P2PEs of the 124 testing ears processed by the 3D-cGANs that are trained using L1 and WL1. Visually, the medians of the Max, Median, and Mean error values obtained with WL1 (yellow bars) are lower than those obtained with L1 (black bars). Wilcoxon signed-rank tests reported in Table 5 show that these differences are significant for the Max, Median, Mean, and STD ($p < 0.05$).

4.2. Mean structural similarity

Fig. 7 shows boxplots of the MSSIM for the 124 testing ears. Wilcoxon signed-rank tests show that all of the cGANs-based methods achieve statistically significant higher MSSIM compared to the baseline ($p < 0.05$). Table 6 shows the p-values of the Wilcoxon signed-rank tests between the results of the 2D-cGANs and the 3D-cGANs that are trained using a different reconstruction loss. The 3D strategies achieve statistically significant higher MSSIM compared to the 2D approach ($p < 0.05$). The 3D-cGANs trained with the weighted L1 loss produce a significantly higher MSSIM than

Table 4
p-Values of the two-sided and one-sided Wilcoxon signed-rank tests of the five statistics of the P2PEs for the 78 IB ears and the 34 IU ears that have been scanned with the lCT scanner postoperatively and the cCT scanners preoperatively.

Testing ears	Approaches to compare	Max		Median		Mean		STD		Min	
		Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
78 IB ears	Post-CT + PostCTseg2 2D-cGANs + PreCTseg	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
	Post-CT + PostCTseg2 3D-cGANs + PreCTseg	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**
	2D-cGANs + PreCTseg 3D-cGANs + PreCTseg	0.136	–	0.163	–	0.172	–	0.770	–	1.110	–
34 IU ears	Post-CT + PostCTseg1 2D-cGANs + PreCTseg	0.080	–	**0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**0.285**	–	**0.002**	**< 0.001**
	Post-CT + PostCTseg1 3D-cGANs + PreCTseg	0.075	–	**0.002**	**0.001**	**0.002**	**< 0.001**	**0.285**	–	**0.006**	–
	2D-cGANs + PreCTseg 3D-cGANs + PreCTseg	0.993	–	0.590	–	0.675	–	0.857	–	1.110	–

Note: Bold indicates cases that are significantly different (p-value less than 0.05). The p-values have been corrected using the Holm-Bonferroni method.

Table 5
p-Values of the two-sided and one-sided Wilcoxon signed-rank tests of the five statistics for the P2PEs of the 124 testing ears.

Reconstruction loss to compare	Max		Median		Mean		STD		Min	
	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
WL1L1	**0.004**	**0.002**	**<0.001**	**<0.001**	**0.001**	**<0.001**	**<0.001**	**<0.001**	**0.945**	–

Note: Bold indicates cases that are significantly different (*p*-value less than 0.05). The *p*-values have been corrected using Holm-Bonferroni method.

Table 6
The *p*-values of the two-sided and one-sided Wilcoxon signed-rank tests between the MSSIM of the true Pre-CT images and the synthetic images produced by the 2D-cGANs and the 3D-cGANs trained using different reconstruction losses. “lCT-cCT” denotes that the ear has been scanned by the 1CT scanner postoperatively and a cCT scanner preoperatively, and “cCT-cCT” denotes that the ear has been scanned by a cCT scanner postoperatively and preoperatively.

Approaches to compare	Mixed (124 ears)		lCT-cCT (112 ears)		cCT-cCT (12 ears)	
	Two-sided	One-sided	Two-sided	One-sided	Two-sided	One-sided
2D-cGANs L13D-cGANs L1	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
2D-cGANs L13D-cGANs WL1	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**< 0.001**	**<0.001**
3D-cGANs L13D-cGANs WL1	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Note: Bold indicates cases that are significantly different (*p*-value less than 0.05). The *p*-values have been corrected using Holm-Bonferroni method.

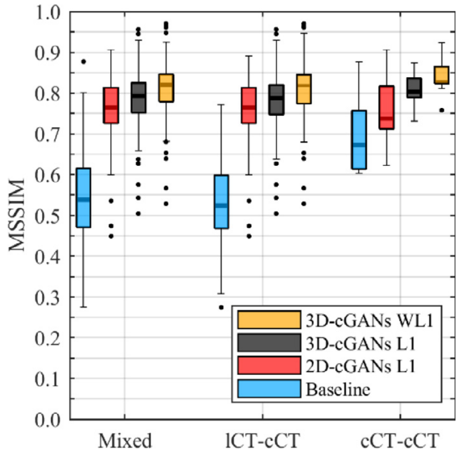

Fig. 7. Shown on the left, middle, and right are boxplots of the MSSIM for the 124 testing ears (Mixed), the 112 ears scanned by the lCT scanner postoperatively and the cCT scanners preoperatively (lCT-cCT), and the 12 ears scanned by the cCT scanners postoperatively and preoperatively (cCT-cCT). “Baseline” denotes the MSSIM between the Post-CT images and the true Pre-CT images; “2D-cGANs L1” denotes the results produced by our previous 2D-cGANs trained with the pure L1 loss; “3D-cGANs L1” and “3D-cGANs WL1” denote the results produced by the 3D-cGANs which are trained using the pure L1 loss and the weighted L1 loss, respectively.

those trained with the non-weighted L1 loss ($p < 0.05$). We also observe that the 3D-cGANs reach the optimal epoch at the 15-th training epoch when the weighted L1 loss is applied. However, they need 70 training epochs to reach the optimal epoch when the non-weighted L1 loss is applied. This suggests that using weights can accelerate the optimization of the networks.

5. Discussion and conclusion

As discussed in the recent review article by Yi et al. (2018), GANs have been extensively used to solve medical imaging related problems such as classification, detection, image synthesis, low dose CT denoising, reconstruction, registration, and segmentation. However, at the time of writing and to the best of our knowledge, GANs have not been proposed to eliminate or reduce metallic artifacts in CT images. There is also a large body of work aiming at reducing artifacts in CT images (Gjesteby et al., 2016). But, compared to the current leading methods, which generally necessitate the raw data from CT scanners, our approach is a post-reconstruction processing method for which the raw data is not required. Compared to other published machine-learning-based methods proposed for the removal of metallic artifacts that either depend on existing traditional methods or require post-processing of the outputs produced by machine learning models (Gjesteby et al., 2017; Park et al., 2017; Zhang and Yu, 2018), ours is unique in being able to synthesize directly an artifact-free image from an image in which artifacts are present. Although we have not investigated it yet, we hypothesize that our method could be applied to other types of images affected by the same type of artifacts if sufficient training images consisting of pairs of images with and without artifacts were available. We also hypothesize that if the problems are similar enough transfer learning could be used to reduce the size of the dataset needed for training.

The results we have generated show that the quality of the images produced by the 3D networks is better than that of the images produced by the 2D networks when the MSSIM is used to compare them. This is confirmed by the visual appearance of the synthetic images produces by these two architectures as shown in Fig. 4. There is also a small but not statistically significant difference in the segmentation results produced with the images generated with the 3D and the 2D networks; this difference is especially noticeable for the maximum error. The fact that the segmentation results improve only modestly when the quality of the images improves more substantially suggests that the constraints imposed by the active shape model are able to compensate for imperfections in the synthetic images. It is likely that segmentations methods that do not impose strong constraints on the shape for the ICA structures would be more sensitive to those errors. As discussed earlier, we also note that the technique we have developed to assist audiologists in programing the implant depends on the position of the contacts with respect to the anatomy (Noble et al., 2013). Any improvement in segmentation accuracy, even small, may have a positive impact on programming recommendations we provide to the audiologists. Assessing the effect of the method we use to eliminate the artifact on these recommendations, i.e., assessing whether or not recommendations would be different if the 2D or 3D version is used, is part of our plans. Finally, the methods we have developed to segment the anatomy, localize the contacts, and provide programming guidance have been integrated into an interactive software package that has been deployed to the clinic and is in routine use at our institution. Without further optimization of our current implementation of the cGANs, speed of execution for the 3D version is 1.5 s when it is 60 s for the 2D version, which is important for the integration of our methods into the clinical workflow. Overall, the study we have conducted shows that cGANS are an effective way to eliminate metallic artifacts in CT images and that the 3D version of our proposed method should be preferred over the 2D version.

Declaration of Competing Interest

None.

Acknowledgments

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 these institutes. We are grateful for the support of the NVIDIA Corporation for providing us with a GPU used to support this work and we thank the anonymous reviewers for constructive comments. Their feedback has greatly improved the quality of this article.

Appendix

Table A
The abbreviations and descriptions.

Abbreviations	Descriptions
CI	Cochlear implants
ICA	Intra cochlear anatomy
IB ears	The ears of the cochlear implants recipients who have been implanted bilaterally
IU ears	The ears of the cochlear implants recipients who have been implanted unilaterally
CT	Computed tomography
cCT	Conventional CT
lCT	Low-dose CT
Post-CT	Post-implantation CT
Pre-CT	Pre-implantation CT
PostCTseg1	A method for segmenting intra cochlear anatomy structures in post-implantation CTs of cochlear implants recipients who have been implanted unilaterally
PostCTseg2	A method for segmenting intra cochlear anatomy structures in post-implantation CTs of cochlear implants recipients who have been implanted bilaterally
PreCTseg	A method for segmenting intra cochlear anatomy structures in pre-implantation CTs of cochlear implants recipients
GANs	Generative adversarial nets
cGANs	Conditional generative adversarial networks
WL1	Weighted L1
P2PE	Point-to-point error
SSIM	Structural similarity
MSSIM	Mean structure similarity
Max	Maximum
Min	Minimum
STD	Standard deviation

References

Gjesteby, L., De Man, B., Jin, Y., Paganetti, H., Verburg, J., Giantsoudi, D., Wang, G., 2016. Metal artifact reduction in CT: where are we after four decades? IEEE Access 4, 5826–5849. doi:10.1109/Access.2016.2608621.

Gjesteby, L., Yang, Q., Xi, Y., Claus, B., Jin, Y., De Man, B., Wang, G., 2017. Reducing metal streak artifacts in CT images via deep learning: pilot results. In: 14th Int. Meet. Fully Three-Dimensional Image Reconstr. Radiol. Nucl. Med., 14(6), pp. 611–614. doi:10.12059/Fully3D.2017-11-3202009.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. NIPS URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR doi:10.1109/CVPR.2016.90.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6 (2), 65–70. doi:10.2307/4615733.

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: CVPR arXiv:1611.07004.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: ICLR arXiv:1412.6980.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Trans. Med. Imaging 16 (2), 187–198. doi:10.1109/42.563664.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. [cs, stat], 2014. arXiv:1411.1784.

National Institute on Deafness and Other Communication Disorders, 2011, NIDCD Fact Sheet: Cochlear Implants, NIH Publication No. 11–4798. National Institutes of Health, Bethesda, MD, USA.

Noble, J.H., Labadie, R.F., Majdani, O., Dawant, B.M., 2011. Automatic segmentation of intracochlear anatomy in conventional CT. IEEE Trans. Biomed. Eng. 58 (9), 2625–2632. doi:10.1109/TBME.2011.2160262.

Noble, J.H., Labadie, R.F., Gifford, R.H., Dawant, B.M., 2013. Image-guidance enables new methods for customizing cochlear implant stimulation strategies. IEEE Trans. Neural Syst. Rehabil. Eng. 21 (5), 820–829. doi:10.1109/TNSRE.2013.2253333.

Park, H.S. Lee, S.M. Kim, H.P. and Seo, J.K. Machine-learning-based nonlinear decomposition of CT images for metal artifact reduction. [physics.med-ph], 2017. arXiv:1708.00244.

Reda, F.A., McRackan, T.R., Labadie, R.F., Dawant, B.M., Noble, J.H., 2014a. Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients. Med. Image Anal. 18 (3), 605–615. doi:10.1016/j.media.2014.02.001.

Reda, F.A., Noble, J.H., Labadie, R.F., Dawant, B.M., 2014b. An artifact-robust, shape library-based algorithm for automatic segmentation of inner ear anatomy in post-cochlear-implantation CT. In: SPIE Proceedings Vol 9034, Medical Imaging 2014: Image Processing; 90342V doi:10.1117/12.2043260.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. doi:10.1109/TIP.2003.819861.

Wang, J., Zhao, Y., Noble, J.H., Dawant, B.M., 2018. Conditional generative adversarial networks for metal artifact reduction in CT images of the ear. Med. Image Comput. Comput. Assist Interv. doi:10.1007/978-3-030-00928-1_1.

Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1 (1), 35–51 doi:10.1016/S1361-8415(01)80004-9.

Yi, X. Walia, E. and Babyn, P. Generative adversarial network in medical imaging: a review. [cs.CV], 2018. arXiv:1809.07294.

Zhang, Y., Yu, H., 2018. Convolutional neural network based metal artifact reduction in x-ray computed tomography. IEEE Trans. Med. Imaging 37 (6), 1370–1381. doi:10.1109/TMI.2018.2823083.

Zhao, Y., Dawant, B.M., Labadie, R.F., Noble, J.H., 2018. Automatic localization of closely-spaced cochlear implant electrode arrays in clinical CTs. Med. Phys. 45 (11), 5030–5040. doi:10.1002/mp.13185.

Zhao, Y., Chakravorti, S., Labadie, R.F., Dawant, B.M., Noble, J.H., 2019. Automatic graph-based method for localization of cochlear implant electrode arrays in clinical CT with sub-voxel accuracy. Med. Image Anal. 52, 1–12. doi:10.1016/j.media.2018.11.005.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV arXiv:1703.10593.