Deep Learning Methods for CT Image-Domain Metal Artifact Reduction

Lars Gjesteby¹, Qingsong Yang¹, Yan Xi¹, Hongming Shan¹, Bernhard Claus², Yannan Jin², Bruno De Man², Ge Wang¹

¹Biomedical Imaging Center, Rensselaer Polytechnic Institute, Troy, NY, USA
²Imaging, GE Global Research Center, Niskayuna, NY, USA

ABSTRACT

Artifacts resulting from metal objects have been a persistent problem in CT images over the last four decades. A common approach to overcome their effects is to replace corrupt projection data with values synthesized from an interpolation scheme or by reprojection of a prior image. State-of-the-art correction methods, such as the interpolation- and normalization-based algorithm NMAR, often do not produce clinically satisfactory results. Residual image artifacts remain in challenging cases and even new artifacts can be introduced by the interpolation scheme. Metal artifacts continue to be a major impediment, particularly in radiation and proton therapy planning as well as orthopedic imaging. A new solution to the long-standing metal artifact reduction (MAR) problem is deep learning, which has been successfully applied to medical image processing and analysis tasks. In this study, we combine a convolutional neural network (CNN) with the state-of-the-art NMAR algorithm to reduce metal streaks in critical image regions. Training data was synthesized from CT simulation scans of a phantom derived from real patient images. The CNN is able to map metal-corrupted images to artifact-free monoenergetic images to achieve additional correction on top of NMAR for improved image quality. Our results indicate that deep learning is a novel tool to address CT reconstruction challenges, and may enable more accurate tumor volume estimation for radiation therapy planning.

Keywords: Computed tomography (CT), deep learning, convolutional neural network (CNN), metal artifact reduction (MAR), proton therapy planning

1. INTRODUCTION

The long-standing problem of metal artifacts in CT images continues to hinder clinical diagnosis today. Although many MAR techniques have been implemented in the past forty years,¹ there remain important cases that challenge even the state-of-the-art algorithms, and sufficient image quality cannot be achieved. Radiation and proton therapy planning are particularly sensitive to errors in the CT images, since incorrect estimation of the treatment beam stopping power may result in under-treatment and tumor recurrence or unnecessary radiation to the surrounding healthy tissues.²⁻³

There are several classes of MAR methods, with projection completion being the most widely developed. These techniques replace corrupt sinogram data in the metal trace with data synthesized by an interpolation technique,⁴⁻⁷ reprojection from a prior image,⁸⁻¹¹ or a combination of both that involves normalization.¹²⁻¹⁴ A state-of-the-art algorithm among these is NMAR.¹³ Other classes of MAR methods include scan acquisition improvement, physics-based pre-processing, iterative reconstruction, and image post-processing. While image post-processing algorithms have had some limited success,¹⁵,¹⁶ they are more useful when combined with sinogram-domain correction.¹⁷ Since the current clinical techniques often fall short in providing requisite image quality for the most demanding applications, particularly in radiation and proton therapy planning, we are interested in developing new, improved solutions for metal artifacts.

Machine learning and deep learning have rapidly gained attention in recent years for employment on many complicated tasks with new twists.¹⁸ This technology can be applied as a novel approach for reducing metal artifacts in CT images. The ability of convolutional neural networks to extract detailed features from complex datasets makes them powerful tools for image processing and analysis.¹⁹⁻²¹ A supervised learning process can be used to train the network with labeled data/images
so that it learns how to map features between the input and the ideal output. Once trained, the network uses forward prediction to estimate an output given an input without any labeling.

Our work seeks to reduce streak artifacts in critical image regions outside the metal object by combining a CNN with the state-of-the-art NMAR method. Our preliminary results were presented earlier. In this current paper, we expand the network structure and use clinically-relevant training images from real patient data. The network learns an end-to-end mapping of patches from metal-corrupted CT images to their corresponding artifact-free ground truth. Since raw projection data is not always accessible in commercial scanners, our learning process is entirely in the image domain to demonstrate the feasibility and merits of deep networks for image-based MAR.

2. METHODS

2.1 Data generation

All training and test data for the CNN were generated using the CT simulation software CatSim (General Electric Global Research Center, Niskayuna, NY). Real data from the Visible Human Project was used to create voxelized phantoms of hip cross-sections. Two scans were simulated for each phantom based on a GE LightSpeed VCT system architecture. The phantom for the first scan contained titanium artificially added inside the femoral area of each slice to serve as the metal implant, with a maximum diameter of 50mm. A standard clinical CT protocol was used, with key parameters including a tube voltage of 120 kVp, a tube current of 300 mA, 888 detector columns, and 720 views per rotation at uniform projection angles. The 512x512 reconstructed image contained severe artifacts. For the initial correction, the image was reconstructed using the NMAR algorithm. This NMAR result later served as the input to the CNN. A second scan on each phantom without titanium was simulated with the same parameters, except that a sufficiently higher number of photons were assigned a single energy of 100 keV to generate a monoenergetic image with minimal noise and beam hardening. This served as the ground truth and target of the network. A total of 40 slices at 1mm-thickness were simulated. From the full images, approximately 50,000 patches of size 32x32 with a stride of 8 were extracted from the streak regions to form the dataset for CNN training. Five of the 40 slices were reserved for testing/evaluation and not used in the training process.

2.2 Network design and training

A CNN was defined containing five convolution layers and five deconvolution layers, all employing a small 3x3 kernel. Figure 1 illustrates the network structure with kernel size (k) and number of filters (n) denoted for each layer. The convolution layers extract features from the input patch and map them to the target patch, and the deconvolution layers use these features to build the predicted output. The first four convolution layers contain 32 filters, while the fifth contains 16 so that only the most important features are used to reconstruct the output patch via the deconvolution layers. The last deconvolution layer contains only one 3x3 filter to generate one feature map as the output. The nine hidden layers use Rectified Linear Units (ReLU) as the non-linear activation functions.

![Figure 1. Convolutional neural network containing five convolution layers and five deconvolution layers. All layers have a 3x3 kernel, denoted by k3, and the number of filters are denoted by the number following n. The first three layers contain batch normalization, and a rectified linear unit for non-linearity follows all layers except the last. The input and output patches are both 32x32. The output prediction from the extracted features is based on a mean squared error loss function.](image-url)
The Caffe framework was used for network training and inference.\textsuperscript{25} For the first three layers, batch normalization is included to speed up the initial part of training. The learning rate started at $10^{-4}$ and decreased by a factor of 0.5 after every 100,000 iterations for a total of one million training iterations. Figure 2 displays the loss curve during the test phases.

![Test Loss Curve](image_url)

**Figure 2.** Test loss curve through 200 test phases in the training process. The test phase occurred after every 5,000 of the one million training iterations.

### 3. RESULTS

To validate the network performance, three image slices from the patient data not used in training were employed. These 1mm thick slices contained small, medium, and large diameter cross-sections of the hip implants. Figure 3 shows the initial reconstructed image without correction, the NMAR-corrected image, the CNN result, and the ground truth. The initial reconstruction from the raw projection data yielded a poor image with many streaks, a dark band along the lines of the greatest attenuation, and a lot of missing data. The NMAR-corrected image served as the input to the network, and the forward prediction process estimated an output based on features learned from the ground truth. The CNN prediction significantly reduced noise streaks that remained in the NMAR result. Table 1 presents image quality metrics, structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR), to quantify this improvement.

In all cases, the CNN output has higher PSNR and SSIM scores than the NMAR result. For case 1, the metal cross-section is very small, so there are minimal artifacts in the initial reconstruction. The NMAR algorithm appears to introduce its own artifacts that could be due to poor segmentation of the metal. After passing through the CNN, many of the streaks are eliminated to achieve good similarity with the ground truth. Cases 2 and 3 have medium to large diameters of metal, so more severe streak artifacts exist in the initial reconstruction. The CNN is able to reduce residual streaks in critical regions that NMAR does not correct. There remain some areas that the CNN cannot recover for these challenging cases, and loss of texture is evident compared to the ground truth. Nonetheless, the improvement over NMAR is clear.
Figure 3. Three cases of hip CT images containing metal artifacts. Case 1, 2, and 3 correspond to a small, medium, and large metal cross-section diameter, respectively. The NMAR-corrected image is used as the input to the CNN. Display window is [-250 350] HU.
Table 1. Case-by-case image quality metrics of SSIM and PSNR calculated for the NMAR result and CNN output compared to ground truth.

<table>
<thead>
<tr>
<th>Case</th>
<th>NMAR SSIM</th>
<th>NMAR PSNR</th>
<th>CNN SSIM</th>
<th>CNN PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.526</td>
<td>25.878</td>
<td>0.779</td>
<td>27.572</td>
</tr>
<tr>
<td>2</td>
<td>0.533</td>
<td>22.878</td>
<td>0.744</td>
<td>25.361</td>
</tr>
<tr>
<td>3</td>
<td>0.523</td>
<td>21.330</td>
<td>0.700</td>
<td>22.961</td>
</tr>
</tbody>
</table>

4. DISCUSSION AND CONCLUSION

Performing this MAR study using the deep learning approach in the image domain, we have demonstrated a superiority principle of data-driven machine learning. Thanks to the availability of big data and the adaptability of artificial neural networks, any state-of-the-art traditional MAR result can be further improved using a well-trained network. In principle, learning-based MAR techniques will always outperform the best result, and could be repeatedly applied. This can be viewed as a monotonically-increasing sequence that possibly approaches the ground truth as the limiting point.

It is underlined that our current image-domain-based MAR deep learning process can be further developed in terms of an improved network architecture and refined parameters, and can include more diverse and enlarged implant types. For example, residual CNN\textsuperscript{26,27} and generative adversarial networks (GAN)\textsuperscript{28} could be adapted for MAR. The larger and the more challenging the dataset is, the better the MAR network can be trained. To enlarge our training and testing data, we may synthesize various cases using a high-quality CT simulator. At the very least, these pseudo data can be used for pre-training to facilitate fine-tuning without the need for overly many real samples. Further, it has been recognized that the common loss function based on the mean squared error may not be the best for diagnostic and therapeutic purposes,\textsuperscript{29} so exploring better task-based loss functions is an interesting topic as well.

In conclusion, we have demonstrated that deep learning can help MAR in a significant way by providing a substantial gain on the results achieved with the best existing methods. In other words, we advocate the superiority principle that machine learning can only do better for MAR (and other tomographic imaging tasks), if the techniques are appropriately developed and applied. There are still numerous opportunities for us to advance tomographic imaging, especially artifact reduction in the machine learning framework. In particular, we will report more results on machine learning based MAR results in a follow-up study.

REFERENCES