Metal artifact correction in cone beam computed tomography using synthetic X-ray data

Harshit Agrawal, Ari Hietanen, and Simo Särkkä, Senior Member, IEEE

Abstract—Metal artifact correction is a challenging problem in cone beam computed tomography (CBCT) scanning. Metal implants inserted into the anatomy cause severe artifacts in reconstructed images. Widely used inpainting-based metal artifact reduction (MAR) methods require segmentation of metal traces in the projections as a first step which is a challenging task. One approach is to use a deep learning method to segment metals in the projections. However, the success of deep learning methods is limited by the availability of realistic training data. It is challenging and time-consuming to get reliable ground truth annotations due to unclear implant boundary and large number of projections. We propose to use X-ray simulations to generate synthetic metal segmentation training dataset from clinical CBCT scans. We compare the effect of simulations with different number of photons and also compare several training strategies to augment the available data. We compare our model’s performance on real clinical scans with conventional threshold-based MAR and a recent deep learning method. We show that simulations with relatively small number of photons are suitable for the metal segmentation task and that training the deep learning model with full size and cropped projections together improves the robustness of the model. We show substantial improvement in the image quality affected by severe motion, voxel size under-sampling, and out-of-FOV metals. Our method can be easily implemented into the existing projection-based MAR pipeline to get improved image quality. This method can provide a novel paradigm to accurately segment metals in CBCT projections.

Index Terms—CBCT, deep learning, synthetic data, X-ray simulation, metal artifact reduction, metal segmentation.

I. INTRODUCTION

ONE beam computed tomography (CBCT) acquires 2D X-ray projections of an object to reconstruct 3D images. Presence of high-density metallic implant in the object may corrupt the projection data and cause severe artifacts such as blooming, streaking, and shading [1] in the reconstructed images. Several projection-based metal artifact removal (MAR) methods have been proposed in literature [2]–[7]. A crucial step in these methods is the segmentation of corrupt metal traces in the projections. First, the metals are segmented in the 3D image-domain by a threshold. Then the segmented metals are forward-projected to obtain the metal trace in the projections. The data inside the metal trace is discarded and interpolated from the nearby projection data. The corrected projections are back-projected (reconstruction) to generate corrected 3D images.

Several deep learning-based MAR methods have also been proposed recently. They either work in projection-domain, image-domain, or both. Projection-domain methods [8]–[12] inpaint the metal trace in the projection (CBCT) or sinogram (CT) using deep learning architectures. All of these methods rely on the availability of segmented metal traces. Image-domain methods, such as [13], [14] directly translate from artifact to artifact-free images. Image-domain methods can effectively reduce metal artifacts in some cases, however, their effect can be limited by the presence of scatter, large and varying shapes of metals, multiple metal implants, and motion artifacts [9]. While inpainting-based methods deal better with these problems, they still require accurate metal segmentation in the projection [15]. In [16]–[18], it was proposed to incorporate both image and sinogram-domain deep learning. These methods used filtered back projection (FBP) and forward projection (FP) layers to propagate loss from the image-domain to sinogram-domain. These methods were implemented for CT images (slice-wise reconstruction) though full volume CBCT reconstruction is not feasible due to its large memory requirements [19].

It was demonstrated in [20] that a precise segmentation of metal traces can enhance image quality, and on the other hand, less accurate segmentation may cause additional artifacts or even remove anatomical details. As forward projection of the image-domain metal is used to obtain the metal traces in the projection, it has several problems. When metal lies outside of the field-of-view (FOV), image-domain segmentation methods cannot segment metals in the projections [21]. This may cause image in-homogeneity in the center and also among the different zones of the FOV [22], [23]. Also, in the presence of motion artifacts [24]–[26], it is difficult to obtain reliable forward projection of the image-domain segmented metals and extra steps might be needed [27], [28]. Moreover, the forward projections of the metals need to account for the partially covered pixels. This can be done by reconstructing the voxels with smaller size or by taking the voxel diameter into account [29] at the cost of increased complexity and processing time.
While several deep learning methods have been investigated for MAR, only a few have been proposed to segment directly projection-domain metals in CBCT scans, owing to the unavailability of accurate metal labels [30], [31]. A U-Net [32] architecture was used in [30] to segment metals in the dental CT projections. The authors mentioned that their network was trained and tested on a very small dental CT dataset (five patients for training and four patients for testing) due to the time consuming process of creating ground truth labels. Pairs of metal and metal free projections from cadaver CBCT scans were used in [31] to train a U-Net architecture. Acquiring data from cadaver images is a rare process and the data might still lack the quantity and variety of images. Their method also used a consistency condition check to reduce false positives and generate more consistent segmentation of metals in the projections. The consistency check involved extra steps to reconstruct larger volumes and needed to calculate accurate thresholds for the metals in the reconstruction. Furthermore, we found that consistency check did not work well for motion-affected scans.

Monte Carlo simulations have been used to generate training data for deep learning-based metal in-painting and scatter correction methods [10], [17], [33], but the use of simulations in metal segmentation training have not been investigated yet. Motivated by the importance and need of accurate training data for metal segmentation, we used Monte Carlo simulations to generate metal-corrupted CBCT projections and corresponding metal labels for network training. For the tasks of in-painting and scatter correction, simulations with low noise levels are needed [10], [17], [33]. As the noise in the simulated projections decreases with the increasing number of photons per detector pixel [33], running the simulations with the required large number of photons is computationally expensive. In contrast, we use simulations with only 300 to 1400 photons per detector pixel (noisy) for metal segmentation training and compared with 5000 photons per detector pixels (clean). Noisy simulations are faster to compute and add realistic noise to the dataset.

We then trained a modified version of U-Net only on the simulated data to segment metal traces in the projections. In addition, we used a simple strategy to augment our simulated dataset further by training the network on full size and cropped projections together. Our simulated dataset included multiple anatomies such as knee, wrists, ankle, palm, or foot. We demonstrate the robustness of our metal segmentation model on 10 metal-affected and 6 non-metal clinical datasets.

Our main contributions are as following:

1) We propose a new training approach for metal segmentation in CBCT projection that uses simulated dataset (with different noise levels) obtained from real metal-affected cases.
2) We propose a simple data augmentation strategy by training the network with crops and full size images. This strategy effectively reduces false positives in 6 unseen clinical CBCT scans without metals.
3) We demonstrate the metal artifact reduction on 10 unseen clinical CBCT scans affected by metals. Results show good performance in challenging cases such as motion-affected and out-of-FOV metals, and they are robust to voxel size changes.

II. METHODS

In this section, we explain our pipeline to generate the training data, the simulation process, and the network architecture used for segmentation of metals in detail.

A. Training-Data Generation Pipeline

The pipeline to generate pairs of metal corrupted projections and corresponding ground truth metal traces is shown in Fig. 1. The pipeline had two main parts. In the first part (shown in the colored box), segmentation of metals in 3D and initially corrected volume were obtained. In the second part, a simulation process was used to generate training data pairs.

First, the object was reconstructed from real CBCT projections. Then we segmented the image-domain metals in the 3D reconstructed images using a global threshold. The segmented metals were not smooth and complete due to the artifacts remaining in the reconstruction. To smoothen and complete the metal boundaries, we applied median filtering and binary dilation to the segmented metals in 3D. These metals were simulated to get only-metal projections. Binary metal traces were obtained by applying a threshold to simulated metal projections. The binarized metal traces constituted ground truth for the model training.

To obtain the metal corrupted projections, we first created metal free reconstructions. The metals were put back into the metal free reconstructions and the metal-inserted volume was used to simulate metal-corrupted projections. An example of original projection, a simulated projection and corresponding ground truth metal trace is shown in Fig. 2.

B. Simulation Process

We used Monte Carlo simulations to generate accurate metal segmentation training dataset. The implementation was similar to the one described in [35]. We simulated path of X-ray photons using reconstructions obtained from 10 metal-affected clinical CBCT scans. The clinical volume set was used to represent realistic variations in the metal shapes and sizes. All reconstructions in this work were based on a modification of the algorithm given by Feldkamp, Davis and Kress (FDK) [34]. The reconstruction volume to be simulated was first segmented into four different tissue materials by applying thresholds on Hounsfield Units (HU) values. The volume was segmented into air, soft tissue, bone, or metal as follows:

\[
M_{x,y,z} = \begin{cases} 
\text{air,} & \text{if } V_{x,y,z} < t_1, \\
\text{soft tissue,} & \text{if } t_1 \leq V_{x,y,z} < t_2, \\
\text{bone,} & \text{if } t_2 \leq V_{x,y,z} < t_3, \\
\text{metal,} & \text{if } V_{x,y,z} \geq t_3,
\end{cases}
\]  

where \(V_{x,y,z}\) is the voxel value in Hounsfield Units (HU) at a location \(x, y, z\) in the reconstructed volume \(V\), \(t_1\) is 500, \(t_2\) is 1500, \(t_3\) is 4400, and \(M_{x,y,z}\) is the corresponding material obtained from thresholds.
Fig. 1. Pipeline for generating training pairs. The colored box illustrates conventional method for metal artifact reduction. The metal corrupted original 2D projections were first reconstructed using FDK [34]. The metal was segmented and refined. The metals are put back in the initial corrected volume. Then metal and metal inserted volumes were simulated using Monte Carlo simulations. The ground truth binary traces were obtained by applying a threshold to the simulated metal masks.

We simulated the actual geometries used during the clinical acquisitions to keep the complexity of the dataset realistic. Each X-ray photon was started from a source location given by the scanner geometry. The initial energy of the photon was randomly sampled from a poly-chromatic energy distribution. The path of the photon was simulated along a straight line from the X-ray source location to the detector. When the X-ray photon passes through the object volume, the photon interacts with the object’s material according to object’s density and photon energy. Three types of X-ray interactions were simulated, namely, photoelectric absorption, Compton scattering, and Rayleigh scattering based on [36]. The material composition obtained from the segmentation in (1) is used to model interactions of the X-ray photons.

Multiple projection views were simulated by rotating source and detector around the object as per the calibrated acquisition geometry of the scan. This way we acquired multiple 2D projection views from the same object according to the realistic geometry. A simulation was also done for each projection view without the object in the field-of-view to get 2D projections, called flat field. Projections were normalized and linearized, according to

\[ p_{\text{sim}} = -\log \left( \frac{S}{F} \right), \]

where \( S \) is the simulated projection and \( F \) is the flat field.

C. Network Architecture

We used a modified U-Net architecture for metal segmentation training. The architecture is shown in Fig. 3. It consisted of 4 downsampling (encoder) and 4 upsampling (decoder) blocks. Each downsampling block had two convolutional layers (kernel size = 3 × 3, stride = 1 × 1, zero padding), each followed by a rectified linear unit (ReLU) and an instance normalization layer [37]. In the first block, network had 16 channels. The number of channels was doubled in each block. After every two convolutional layers, a 2 × 2 max-pooling layer was applied to downsample features. Each upsampling block had a bilinear upsampling layer followed by a convolutional layer (kernel size = 3 × 3, stride = 1 × 1, zero padding). The features from encoder blocks were concatenated to the features of decoder blocks to allow flow of high resolution information from earlier layers. Output of each upsampling block was padded to match the dimensions of the skip connection from the encoder side. This padding scheme ensured the size of input and output remain same and varying sizes of input images could be given to the network during training. The last layer in the network was a 1 × 1 convolutional layer without any activation. A sigmoid function was applied after the last layer of the network to convert the output to a probability map. To obtain metal segmentation during inference, pixels with probability equal to or above 0.5 were classified as metal and pixels below the probability of 0.5 were classified as background.

III. EXPERIMENTS

In this section we describe the dataset used for model training and evaluation, the implementation details of the model, and the comparison methods to evaluate our model’s performance.
A. Dataset

We used unidentifiable and pseudonymized clinical dataset acquired by Planned Verity® scanner (Planned Oy, Helsinki, Finland) with patient permission to use in research and development studies. This study was performed in line with the principles of the Declaration of Helsinki. The dataset had 26 clinical scans taken at different extremity anatomies and was collected over multiple years, which makes the dataset diverse and realistic. Out of 26 scans, 20 were metal-affected and 6 were without any metal. Out of 20 metal-affected scans, we included 10 scans for simulated training data generation. Remaining 10 metal-affected and 6 non-metal scans were used for evaluation of the trained models. The description of training and testing dataset is given in Table I. Each scan had multiple projection views, ranging from 300 to 450. The tube voltage varied from 90 kV to 96 kV. All the scans had isotropic spatial resolution, with voxel spacing of 0.4 mm or 0.2 mm. Different anatomy locations were scanned, including, knee, wrist, foot, ankle, palm, and forearm. The size of each projection was $740 \times 948$. From the training set of 10 scans in Table I we created 10 reconstructions. Two more reconstruction were created by combining metals from multiple reconstructions. These reconstructions were used to generate 3450 projections using simulation procedure described in Section II-B. As we did not want to compromise the quality of segmentation near metal boundaries, we did not downsample projections for training as in [30]. To increase the data variability, we created four crops of size $370 \times 474$ from each projection. The crop was included in the training only if the metal size was at least greater than 100 pixels in the projection. Thus we had 9305 crops for training. From the test set of 10 metal-affected scans in Table I we created metal labels manually. Since it was time consuming to segment metals in all of the projections, we manually segmented metals in 10 projections from each of the 10 scans. So, in total we gathered 100 pairs of metal-corrupted projections and corresponding ground truth metal labels from the clinical dataset. For the remaining test set of 6 metal-free scans, we did not need ground truth labels. Those scans had a total of 2400 metal-free projections.

| Train | Test |
|-------|------|
| Train | Test |
| kV | voxel | views | anatomy | kV | voxel | views | anatomy |
| set1 | 96 | 0.4 | 400 | knee | 92 | 0.2 | 300 | foot |
| set2 | - | 0.2 | 300 | wrist | 96 | 0.2 | 400 | ankle |
| set3 | - | 0.2 | 300 | ankle | 96 | 0.2 | 400 | ankle |
| set4 | 96 | 0.2 | 300 | knee | 96 | 0.2 | 300 | leg |
| set5 | 96 | 0.2 | 450 | foot | 90 | 0.2 | 300 | forearm |
| set6 | 90 | 0.2 | 300 | wrist | 90 | 0.4 | 400 | knee |
| set7 | 90 | 0.2 | 300 | wrist | 96 | 0.2 | 400 | foot |
| set8 | 90 | 0.2 | 300 | wrist | 96 | 0.6 | 400 | leg |
| set9 | 96 | 0.2 | 300 | ankle | 92 | 0.2 | 300 | knee |
| set10 | 90 | 0.2 | 300 | palm | 90 | 0.2 | 450 | wrist |
B. Implementation Details

Our network was implemented using PyTorch [38] deep learning library and was trained on a single GeForce RTX 2080 Ti Rev. A (11GB), with a batch size of 4. We used Adam optimizer \( (\beta_1 = 0.9, \beta_2 = 0.999) \) [39] to optimize the network parameters to minimize loss between the network output and ground truth. The initial learning rate was set to \( e^{-4} \) and reduced after each epoch in logarithmic steps down to \( e^{-6} \). Learning rate was reduced during first 25 epochs and was fixed to \( e^{-6} \) after that. As [40] showed that the stacked data augmentations are effective for model’s generalization in the medical segmentation tasks, we applied a number of stacked data augmentations during training. Each input projection was normalized by its maximum value before augmentation. We used horizontal and vertical flips, image shift, rescale, rotation, Gaussian noise, multiplicative noise, elastic transformation, and mask dropout augmentations. The augmentations were applied on-the-fly during training with probability of 0.2 for each augmentation. Therefore, each epoch had slightly different set of inputs. To avoid over-fitting, the training was stopped if the loss did not decrease for five epochs (early stopping). The model parameters and data augmentations were initialized with random seed 2060. To account for stochastic training process, we trained each model two more times with random seeds 12060 and 22060.

We used binary cross entropy loss for training the neural network, defined as

\[
l = \frac{1}{N \times w \times h} \sum_{i=0}^{N \times w \times h} y_i \log(\sigma(x_i)) + (1-y_i) \log(1-\sigma(x_i)),
\]

where \( N \) is the batch size, \( w, h \) are width and height of the 2D projection image, \( x_i \) is the network output at pixel location \( i \), \( \sigma \) is sigmoid function, and \( y_i \) is the corresponding ground truth pixel.

C. Evaluation

We compared U-Net-based models trained on noisy and clean simulations. For each of noisy and clean simulations training data, the model was trained on full size, crops, and their combinations. For forward projection-based conventional MAR, we segmented the metals in 3D images using connected component region growing segmentation similar to [3]. Metal seeds were started at the voxels with intensities of more than 8000 HU. The metal region was grown starting from the seeds, by checking the intensities of the connected pixels in the neighborhood. If the neighboring pixel had an intensity more than 4000 HU, it was included in the metal region.

The segmented metals were forward projected to create metal traces in the 2D projections. For the patch-based training (Noisy (crops) and Clean (crops)), we followed [31] and used random crops of size \( 512 \times 512 \) during the training. The inference was done on full size projections. For a fair comparison, we used same U-Net network with same parameters for all the compared methods. For the comparison with consistency check, we followed [31] to reconstruct 3D binary metal from the initial segmentation of Noisy (full+crops) model and forward project 3D binary metal to get updated metal traces.

The quantitative segmentation performance of the compared methods was evaluated using Dice Similarity Coefficient (DSC), Intersection over union (IOU) and False positive rate (FPR) [41]. These metrics are defined as

\[
DSC = \frac{2TP}{2TP + FP + FN}, \quad \text{(4)}
\]

\[
IOU = \frac{TP}{TP + FP + FN}, \quad \text{(5)}
\]

\[
FPR = \frac{FP}{TN}, \quad \text{(6)}
\]

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Metal traces found from the compared methods were inpainted from the nearby pixels using a 2D interpolation method. The inpainted projections were reconstructed using a modification of FDK [34] algorithm. The 3D metals obtained from region growing-based method were put back in the reconstructions of conventional MAR and U-Net-based methods.

| Method          | mean DSC \( \uparrow \) | mean IOU \( \uparrow \) |
|-----------------|-------------------------|-------------------------|
| Conventional MAR| 90.5 ± 2.0              | 81.4 ± 3.2              |
| Noisy (full)    | 93.7 ± 0.9              | 88.0 ± 1.6              |
| Noisy (crops)   | 94.4 ± 0.7              | 89.6 ± 1.2              |
| Noisy (full+crops) | 94.8 ± 0.6            | 90.2 ± 1.1              |
| Consistency check| 91.0 ± 3.4              | 85.0 ± 4.8              |
| Clean (full)    | 93.0 ± 1.0              | 87.4 ± 1.6              |
| Clean (crops)   | 92.2 ± 1.6              | 86.5 ± 2.1              |
| Clean (full+crops) | 93.8 ± 0.8            | 88.6 ± 1.4              |

IV. RESULTS

In this section we compare the quantitative metal segmentation performance of U-Net-based models with the forward projection-based conventional MAR and consistency check. Then we show the qualitative impact of metal trace segmentation on the reconstructed images.

A. Quantitative Analysis

1) Clinical Data with Metals: As performance measures, we calculated the mean DSC and mean IOU for 10 test scans. Table II shows the segmentation performance of the conventional MAR, six U-Net-based models, and the consistency check-based segmentation applied on the model Noisy (full+crops). Compared to the conventional MAR, all the six U-Net-based models had higher mean DSC and mean IOU. This shows that the deep learning methods segmented metals more accurately than the forward projection-based conventional MAR. The model Noisy (full+crops) gave highest mean DSC=94.8 (SE=0.6) and highest mean IOU= 90.2 (SE=1.1). Furthermore, when consistency check [31] was applied on the best model...
which was less than the mean FPR given by Noisy (crops), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similarly, the model trained on simulations with 5000 photons per pixel (Clean (full+crops)) had least mean FPR i.e., $0.85 \times 10^{-3}$ (SE$=0.36 \times 10^{-3}$) and the mean FPR given by Noisy (full), i.e., $14 \times 10^{-3}$ (SE$=5 \times 10^{-3}$). Similar analysis on the metal-free projections showed that the models trained on the full size and cropped projections together were more robust to the false positives compared to the models trained on only full size or only cropped projections.

![Fig. 4. Boxplot visualization of DCE and IOU scores for conventional MAR, U-Net models and consistency check. The rectangles contain data within the first and the third quartiles. The endpoints of the lower and upper whiskers are selected as the first quartile - 1.5 times the interquartile range (IQR) and third quartile + 1.5 IQR, respectively. The medians are visualized as red lines and the means as green triangles. The x-axis label shows the name of the method. The outliers are shown as blue triangles. The outliers are the points that are outside the interval defined by the whiskers.](image)

Noisy (full+crops), it affected the metal segmentation performance adversely and reduced the mean DSC=91.0 (SE=3.4) and mean IOU=85.0 (SE=4.8). The DSC and IOU scores decreased due to the challenging cases of metals which is explained further in the qualitative analysis in Section [V-B]. Among the three models trained on clean simulations, Clean (full+crops) had the highest mean DSC=93.8 (SE=0.8) and mean IOU=88.6 (SD=1.4), which is similar to the finding that the Noisy (full+crops) had the best mean DSC and mean IOU among the three models trained on noisy simulations.

We also present the DSC and IOU results in Fig. 4. The box plots of DSC and IOU of conventional MAR had most spread over values. The medians of DSC and IOU of Noisy (full + crops) model were close to the medians of Noisy (crops). Boxplots of Noisy (full+crops) did not have any outliers as opposed to one outlier in Noisy (crops). The boxplots of Clean (crops) and Clean (full+crops) had more outliers than Noisy (crops) and Noisy (full+crops). Also the boxplots of consistency check had more vertical spread and showed one outlier in comparison to the boxplots of Noisy (full + crops).

2) Clinical Data without metals: When comparing the segmentation performance of the models on the projections with the presence of the metals, it is also important to know how the models behave when the CBCT projections do not contain any metal. This comparison might show the model’s generalization. This comparison does not require ground truth metal traces. Table III shows the FPR of the 6 models calculated for metal-free projections from 6 clinical scans. The mean FPR for the Noisy (full+crops) was $0.51 \times 10^{-3}$ (SE$=0.16 \times 10^{-3}$) which was less than the mean FPR given by Noisy (crops). Overall, this analysis on the metal-free projections showed that the models trained on the full size and cropped projections together were more robust to the false positives compared to the models trained on only full size or only cropped projections.

| Method       | FPR ↓ |
|--------------|-------|
| Noisy (full) | 14 ± 5|
| Noisy (crops)| 0.85 ± 0.36|
| Noisy (full+crops) | 0.51 ± 0.16|
| Clean (full) | 23 ± 5|
| Clean (crops) | 8 ± 2|
| Clean (full+crops) | 0.44 ± 0.16|

3) Comparison of different training runs: The comparison of DSC scores on test data for six methods obtained from three independent training runs of each method is shown in the Table [IV]. The mean DSC score for each run is shown in the columns named, seed1 (2060), seed2 (12060), and seed3 (22060). The last column named delta is the difference between maximum and minimum DSC scores across three runs. The model Noisy (full+crops) trained on the noisy full size and cropped projections had the best mean DSC scores in each independent run, i.e., 94.8 (SE=0.6), 94.5 (SE=0.7), and 94.5 (SE=0.7). The delta for the model Noisy (full+crops) was 0.3, which was less than the delta of model Noisy (full), i.e., 2.3 and the delta of model Noisy (crops), i.e., 1.9. Similarly, the delta for the model Clean (full+crops) trained on clean full size and crops was 0.4, which was less than the delta of model trained on Clean (full), i.e., 2.1 and the delta of the model Clean (crops), i.e., 2.4. So, it can be deduced that the models trained on full size and crops were giving more stable mean DSC scores across three independent runs. Also the model Noisy (full+crops) had slightly smaller delta (0.3) than the delta of the model Clean (full+crops), i.e., 0.4. These results showed that the model trained on noisy full size and cropped projections was most robust model in terms of DSC scores.

B. Qualitative Analysis

For the qualitative analysis, we show the effect of accurate metal segmentation on the corresponding reconstructions. For all qualitative visualizations we compared Noisy (full+crops) reconstructions with the reconstructions from conventional MAR, consistency check step [19] applied on the segmentation from Noisy (full+crops), and Clean (full+crops). We analyzed improvements in the image quality in complex cases such as the motion affected and out-of-fov metal cases.
Fig. 5. Test scan 4 with large motion and complex metal implants. Segmentation of metals in a projection is shown in the first row. The second row shows two axial image slices from the reconstructions. The third row is showing the zoomed region of interest from the corresponding second row axial slices. Our method Noisy (full+crops) segmented metals in the projection accurately and reduced most of the artifacts in the reconstructed axial slice.

Fig. 6. Test scan 9 with complex metal shape. First row shows a linearized projection and segmentation of metal by several methods. On the right of the projection zoomed region of interests are shown. Second and third row show axial image slices from the reconstruction on the left and the zoomed region of interest from corresponding axial slices are shown on the right. Conventional MAR over-segmented metals and caused loss of some details in the axial slices. Consistency check under-segmented metals in the projection which caused some artifacts in the axial slices. Noisy (full+crops) and Clean (full+crops) segmented metals well which reduced most of the artifacts and preserved details.

### TABLE IV

| Method       | seed 1      | seed 2      | seed 3      | delta |
|--------------|-------------|-------------|-------------|-------|
| Noisy (full) | 93.7 ± 0.9  | 92 ± 1.4    | 94.2 ± 0.8  | 2.2   |
| Noisy (crops)| 94.4 ± 0.7  | 92.5 ± 1.6  | 92.5 ± 1.3  | 1.9   |
| Noisy (full+crops) | 94.8 ± 0.6 | 94.5 ± 0.7  | 94.5 ± 0.7  | 0.3   |
| Clean (full) | 93.0 ± 1.0  | 90.9 ± 2.2  | 91.9 ± 1.4  | 2.1   |
| Clean (crops)| 92.2 ± 1.6  | 91.2 ± 2.0  | 93.4 ± 1.0  | 2.4   |
| Clean (full+crops) | 93.8 ± 0.8 | 93.4 ± 0.9  | 93.4 ± 1.1  | 0.4   |

1) Motion affected volume: The qualitative results on the challenging motion case are shown in the Fig. 5. The segmentation of metals and two axial images from the reconstructions are visualized. The uncorrected images contained large artifacts from the presence of the large metals as well as the motion during the scan. The conventional MAR method over-segmented the metal trace in the projection. The conventional method reduced some artifacts in the reconstructed images but still there were many remaining streaks and darkening around the metal. The over-segmentation of metals in conventional MAR reduced bone details in the axial image as well. Noisy (full+crops) model segmented metal trace in projections most
Fig. 7. Test scan 6 containing out-of-fov metal. First row shows segmentation of metals in the projection. The second row shows an axial slice from the reconstruction. The third row depicts zoomed region of interest from the second row axial slice. When reconstruction is done with a smaller grid size, the forward projection of the metals misses out-of-fov metals. This causes artifacts in conventional MAR and consistency check. When reconstruction is done with larger grid, the segmentation of metal is better. Noisy (full+crops) and Clean (full+crops) can segment metals directly in the projections. Because of better segmentation, both Noisy (full+crops) and Clean (full+crops) preserve details better near the metal and reduce most artifacts.

Fig. 8. Test scan 8 with 0.6 mm voxel size. First row shows segmentation of metals in the projection. The second row shows an axial slice from the reconstruction. The third row depicts zoomed region of interest from the axial slice. When reconstruction is done with small voxel sizes, the forward projection of the metal is affected by the sampling artifacts. Neural network-based method are not affected by sampling artifacts, however, Clean (full+crops) missed some metals but Noisy (full+crops) segmented metal trace best.

accurately. Most of the artifacts were reduced and the bones near the metal were clearly visible in the axial images. Because of the large motion, the consistency check method missed some of the blurred metals in the reconstructions and the forward projection of these metals removed most of the metal trace in the projections. This caused the artifacts to reappear in the images and affected image quality badly. The last column of the Fig. 5 shows that the model Clean (full+crops) missed
3) **Out-of-FOV metal**: It is shown in Fig. 7 that when metal was in the out-of-FOV region, forward projections of image-domain metal failed to segment metal traces in projection domain, which, in turn, caused artifacts in the reconstructions. The original volume size for the reconstruction of the scan was $801 \times 801 \times 601$. To account for out-of-FOV metals in the forward projections, image reconstruction needed to be done on a larger size grid. When reconstruction was done on a volume size of $801 \times 801 \times 661$, the forward projection included all of the metals in the projections. Noisy (full+crops) and Clean (full+crops) models segmented the metal traces directly in the projections without the need of large size reconstruction. It is also evident from the figure that if the metals were segmented well in the projections, most of the artifacts were removed from the reconstructed images. In this case, the out-of-FOV metal was relatively close to the edge. In some cases where the out-of-FOV metal is far, the reconstruction of extra large volume might not get metals segmented. In such situation, our method will still segment metals in the projections.

4) **Effect of voxel size**: The effect of small voxel sizes on the forward projections of metals is shown in Fig. 8. The images were reconstructed with a voxel size of 0.6 mm. The forward projection did not take into account the voxel size and sampling artifacts occurred in the metal projections of the conventional MAR and consistency check. As a result, additional artifact can be seen in the reconstruction of conventional MAR and consistency check. The model trained on the clean simulations missed many metals in the projections which caused artifacts in the reconstructed images. The model trained on noisy simulations segmented metals well which did not cause any additional artifacts. This visualization showed, that our method did not need to account for the voxel size dependent correction as it segmented metals directly in the projections.

**V. Conclusion**

In this paper, we proposed to use noisy Monte Carlo simulations to train a U-Net architecture for the segmentation of metals in CBCT projections. We experimentally demonstrated that synthetic data could substitute real data for metal segmentation training. We showed that the model trained with noisy simulations outperformed the model trained on time-consuming cleaner simulations for the metal segmentation task. We also showed that adding the crops to the full size projections during the training helped to get more robust metal segmentation on the real clinical test scans.

The model trained on full size projections and crops, significantly reduced the false positives in the non-metal projections. Although, we trained our model on limited simulations, we have shown the applicability of the model on a diversified real dataset acquired from multiple anatomies and scanner sites. We have discussed noticeable improvements in the reconstruction quality of motion-affected and out-of-fov metal scans. In our future work, we will consider to test the application of our method on more diversified and larger dataset including dental scans.

**References**

[1] R. Schulze, U. Heil, D. Groß, D. D. Bruellmann, E. Dranischikow, U. Schwanecke, and E. Schoemer, “Artefacts in CBCT: a review,” Dentomaxillofac. Radiol., vol. 40, no. 5, pp. 265–273, 2011.

[2] W. A. Kalender, R. Hebel, and J. Ebersberger, “Reduction of CT artifacts caused by metallic implants,” Radiol., vol. 164, no. 2, pp. 576–577, 1987.

[3] E. Meyer, R. Raupach, M. Lell, B. Schmidt, and M. Kachelrieß, “Normalized metal artifact reduction (NMAR) in computed tomography,” Med. Phys., vol. 37, no. 10, pp. 5482–5493, 2010.

[4] J. Wang, S. Wang, Y. Chen, J. Wu, J.-L. Coatixues, and L. Luo, “Metal artifact reduction in CT using fusion based prior image,” Med. Phys., vol. 40, no. 8, p. 081903, 2013.

[5] M. Meilingner, C. Schmidgundst, O. Schütz, and E. W. Lang, “Metal artifact reduction in cone beam computed tomography using forward projected reconstruction information,” Z. Med. Phys., vol. 21, no. 3, pp. 174–182, 2011.

[6] X. Duan, L. Zhang, Y. Xiao, J. Cheng, Z. Chen, and Y. Xing, “Metal artifact reduction in CT images by sinogram TV inpainting,” in Proc. 2008 IEEE Nucl. Sci. Symp. Conf. Rec. IEEE, 2008, pp. 4175–4177.

[7] C. Peng, B. Qiu, M. Li, Y. Guan, C. Zhang, Z. Wu, and J. Zheng, “Gaussian diffusion sinogram inpainting for X-ray CT metal artifact reduction,” Biomed. Eng. Online, vol. 16, no. 1, pp. 1–17, 2017.

[8] W. Liao, W.-A. Lin, Z. Huo, L. Vogelsang, W. J. Sehnert, S. K. Zhou, and J. Luo, “Generative mask pyramid network for CT/CBCT metal artifact reduction with joint projection-sinogram correction,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), Springer, 2019, pp. 77–85.

[9] M. U. Ghani and W. C. Karl, “Fast enhanced CT metal artifact reduction using data domain deep learning,” IEEE Trans. Comput. Imag., vol. 6, pp. 181–193, 2019.

[10] C. Peng, B. Li, P. Liang, J. Zheng, Y. Zhang, B. Qiu, and D. Z. Chen, “A cross-domain metal trace restoring network for reducing X-ray CT metal artifact artifacts,” Med. Eng. Phys., vol. 39, no. 12, pp. 3831–3842, 2020.

[11] H. Agrawal, A. Hietanen, and S. Särkkä, “Metal artifact reduction in cone-beam extremity imaging using gated convolutions,” in Proc. 2021 IEEE 18th Int. Symp. Biomed. Imag. (ISBI), IEEE, 2021, pp. 1087–1090.

[12] J. Ahn, Y. Choi, and J. Baek, “Convolutional neural network based metal artifact reduction method in dental CT image,” in Proc. SPIE, vol. 11595, 2021, pp. 715 – 720.

[13] L. Gjestey, Q. Yang, Y. Xi, H. Shan, B. Claus, Y. Jin, B. D. Man, and G. Wang, “Deep learning methods for CT image-domain metal artifact reduction,” in Proc. SPIE, vol. 10391, 2017, pp. 147 – 152.

[14] K. Liang, L. Zhang, H. Yang, Y. Yang, Z. Chen, and Y. Xing, “Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning,” Med. Phys., vol. 46, no. 12, pp. e823–e834, 2019.

[15] D. J. Moseley, J. H. Sweersden, and D. A. Jaffray, “High-contrast object localization and removal in cone-beam CT,” in Proc. SPIE, vol. 5745, 2005, pp. 40 – 50.
