2-Step Sparse-View CT Reconstruction with a Domain-Specific Perceptual Network

Haoyu Wei\(^1\)*, Florian Schiffers\(^1\)*, Tobias Würfl\(^2\), Daming Shen\(^3\), Daniel Kim\(^3\), Aggelos Katsaggelos\(^1\), Oliver Cossairt\(^1\)

\(^1\)Northwestern University, Evanston, USA \(^2\)University of Erlangen-Nuremberg, Germany \(^3\)Feinberg School of Medicine, Northwestern University, Chicago, USA

florian.schiffers@northwestern.edu

Abstract

Computed tomography is widely used to examine internal structures in a non-destructive manner. To obtain high-quality reconstructions, one typically has to acquire a densely sampled trajectory to avoid angular undersampling. However, many scenarios require a sparse-view measurement leading to streak-artifacts if unaccounted for. Current methods do not make full use of the domain-specific information, and hence fail to provide reliable reconstructions for highly undersampled data.

We present a novel framework for sparse-view tomography by decoupling the reconstruction into two steps: First, we overcome its ill-posedness using a super-resolution network, SIN, trained on the sparse projections. The intermediate result allows for a closed-form tomographic reconstruction with preserved details and highly reduced streak-artifacts. Second, a refinement network, PRN, trained on the reconstructions reduces any remaining artifacts.

We further propose a light-weight variant of the perceptual-loss that enhances domain-specific information, boosting restoration accuracy. Our experiments demonstrate an improvement over current solutions by 4 dB.

1. Introduction

Since invented in 1972, Computed Tomography (CT) has quickly been recognized as an essential non-destructive modality for industrial inspection [7], medical diagnosis [30] or material sciences [43]. CT is a textbook example of computational imaging, where a physical quantity is encoded in the measured data by an image formation model and then reconstructed by inverting this forward model. In the case of X-ray CT, an image of the attenuation coefficient is reconstructed from projections acquired in a circular acquisition. Under orthographic projection, this model becomes the Radon transform. Discretized measurements of the Radon transform are called sinograms and their inversion is well-studied. The most prominent approach is a direct inversion with a closed-form solution called Filtered Back-Projection (FBP) [30].

Accurate reconstruction with FBP requires sufficient angular sampling in terms of the Crowther-criterion [6] derived from the Shannon-Nyquist sampling theorem. If violated, aliasing artifacts appear as streak-artifacts that drastically reduce diagnostic quality. This setting is called Sparse-View CT (SV-CT) and arises in many practical applications, addressing motion compensation [9] or dose-reduction [10, 48].

A number of approaches have been proposed to overcome the ill-posedness of SV-CT by incorporating prior

\*The first two authors have equal contribution.
knowledge into the reconstruction algorithm. Traditionally, Iterative Reconstruction (IR) is used to optimize a data-fidelity term coupled with appropriate priors [24, 41, 54]. For example, the popular total variation (TV) prior assumes piece-wise constant objects for CT [38]. Recently, learning-based reconstruction algorithms have attracted a lot of attention [1]. Prominent examples are convolutional neural network (CNN)-based models designed to remove artifacts in reconstruction domain [22] or unrolled networks mimicking the data-fidelity driven IR [33].

A less common approach is to interpret SV-CT as a super-resolution or inpainting problem, where the missing information is interpolated between the sparse projections. A sparse-view sinogram can be either inpainted or 1D-super-resolved. Either way, 1D-upsampling of a sinogram is more constrained than its 2D-counterpart [19] of natural images, since the sinogram data vary smoothly along the angular dimension [15]. Other typical tomographic problems with similar formulations include metal artifact removal and limited angle tomography, see Fig. 2.

Figure 2: Different problems arising in tomographic imaging and their relations to our 1D-super-resolution task, such as metal artifact removal (MAR) and limited angle tomography (LAT). SV-CT is at the intersection of inpainting and super-resolution when viewed in measurement domain.

In this work, we propose a two-step architecture that learns super-resolution in the sinogram domain and then removes the remaining artifacts in the image domain. The core idea is outlined in Fig. 1. The following summarizes our main contributions:

- A novel 2-step reconstruction approach in both measurement and image domain. It ensures maximum detail-preservation compared to single-step networks.
- A novel domain-specific, light-weight perceptual loss that outperforms the VGG-based perceptual loss [18] in terms of accuracy, memory and time efficiency.
- Loss functions and submodules tailored to sinogram-domain tasks, and an extensive ablation study that analyzes the effectiveness of each proposed module.
- Demonstration of accurate reconstructions for high compression ratios while achieving a significant performance gain (over 4 dB PSNR) compared to state-of-the-art algorithms.

2. Related Work

Learned Image Reconstruction Recent advances in the reconstruction community seek to learn better image priors from available data [33]. Two prominent strategies stand out: Postprocessing networks (PN) and Operator-based learning (OBL).

A single-step PN removes aliasing artifacts after a direct inversion with closed-form solutions such as FBP. Variants of different neural networks are proposed to learn PN models to eliminate streak artifacts, such as GoogLeNet [45], DenseNet [50] or TomoGAN [29]. An improved version was presented by Han and Ye [14], showing that networks fulfilling the frame-condition yield improved outcomes. A downside of these PN-based methods is their inability to constrain their output by relating it to the measured data.

With OBL a differentiable forward model is included into the network. Most OBL frameworks unroll IR, such as gradient descent, and learn the gradient of an image prior that is applied during each iteration. For SV-CT, Chen et al. [4] demonstrated state-of-the-art performance with OBL. A more general approach to learning a data-driven inversion of linear inverse problems are Neumann networks introduced by Gilton et al. [12].

An entirely different approach from iterative reconstruction with learned priors is to reuse an arbitrary denoising method for images as a proximal operator. This approach was pioneered by Venkatakrishnan et al. [42], referred to Plug-and-Play (P&P)-priors and can be combined with learned denoising.
**Image Super-Resolution & Inpainting** A different approach to solve SV-CT poses it as a super-resolution or inpainting problem in the measurement domain. This is advantageous because sinograms do not yet suffer from aliasing artifacts. U-Nets [37] and generative adversarial networks (GAN) [13] have demonstrated strong performance in this task by either considering the whole sinogram [8, 11, 39, 46] or local patches [25].

Recent approaches in other tomographic applications (Fig. 2) combine learning in both sinogram and image domain to leverage the advantages of both representations. For example, Zhao *et al.* [52] proposed an unsupervised sinogram inpainting network trained in both domains in a cycle for limited angle tomography. One of the first to propose adding a Radon inversion layer to directly couple measurement and reconstruction domain was Würfl *et al.* [44]. The idea of dual-domain training was borrowed by DuDoNet [28] to first inpaint missing information due to metal-artifacts, see Fig. 2 and refine the subsequent reconstruction.

**Perceptual Losses** GAN loss functions are often paired with a perceptual loss to produce visually convincing images. Perceptual loss was introduced in the seminal work by Johnson *et al.* [18] to tackle the long-standing problem that pixel-based loss-functions, such as the one using the $l_2$-norm, struggling to capture higher-level feature information. Perceptual losses are increasingly incorporated in many imaging algorithms showing superior performance while maintaining a reasonable convergence speed [49]. Yet, there is an ongoing debate about whether the natural color image features apply to images with a drastically different formation model and diverse content. In the medical imaging setting, domain-specific perceptual networks were proposed to account for the domain change [26, 34]. These methods require pre-training of an additional network which increases the complexity of the complete training pipeline, making it harder to reproduce results.

### 3. Proposed Approach

We propose a two-step framework with novel loss-functions tailored for SV-CT, shown in Fig. 3. First, we augment a sparse-view sinogram by 1D linear-interpolation and Two-Ends preprocessing, which accounts for boundary artifacts. The Sinogram Inpainting Network (SIN) generates a reliable super-resolved sinogram so that the object can be reconstructed without strong streak artifacts via FBP. While SIN successfully restores a full-view sinogram, small imperfections still lead to some less-severe, more localized artifacts. In the final refinement step, our Postprocessing Refining Network (PRN) removes such artifacts to obtain a high-quality reconstruction.

This section discusses the network architectures, introduces our novel discriminator-perceptual loss, and outlines...
Figure 4: We avoid boundary artifacts during convolutions by reusing projections on both ends of the sinogram.

the optimization procedure.

3.1. Two-Ends Sinogram Flipping

Boundary pixels have only one-sided neighborhood information. This quickly results in border blurriness detrimental for reconstruction. We present a simple two-ends method that appends angles at the two ends of sparse-view sinograms as outlined in Fig. 4.

Note that we only leverage data already present in the measurements. This is because under orthographic projection, the total intensity detected from angle $\theta$ is a flipped version of the projection at angle $\theta + \pi$. This allows us to repeat the first and last few angles of a sinogram by flipping them along the detector axis and appending them to the other end.

3.2. Network Architectures

Following previous works in learned CT-reconstruction, we use the GAN [13] framework in both steps. We first describe our adapted U-Net generators and global and local patch discriminators [16, 17] in detail. Then we introduce our SIN-4c-PRN architecture based on SIN and PRN.

Adapted U-Net Generators A U-Net is a CNN that includes an encoding-decoding process with skip connections [37]. We made the following adaptions to the U-Net modules in our framework: Four average-pooling and bilinear-upsampling layers, each followed by several deep convolutional layers. We keep skip-connections at each resolution level to ensure low-level feature consistency.

We find that resizing operations such as pooling and bilinear layers restore smooth sinusoids essential for tomographic reconstruction. In contrast, half-strided convolutions introduce severe checkerboard artifacts [32] as in our experiments.

We further replace max-pooling with average-pooling, since a differentiable operator is preferable over subdifferentials for restoration tasks [47, 53]. The preservation of gradients maximally recovers measurement-domain information. Otherwise, if the high-frequencies produced by max-pooling are misaligned, errors will be amplified by the following FBP operation.

Patch Discriminators Inspired by Isola et al. [17], our discriminators include four convolution layers that downscale the input by a factor of four and output image patches. Pixel-wise probabilities of these patches being real or fake are then calculated as adversarial losses described in Sec. 3.4. We also keep the discriminators in relatively low dimensional feature space ($\leq 256$ kernels) for ease of training without compromising performance [2].

Sinogram Inpainting Network (SIN) The proposed SIN fills in the missing projections by directly learning from sparse-view to full-view sinograms. The adapted U-Net generates super-resolved sinograms from an initial linear estimation, trained simultaneously with two patch discriminators that focus on either global or local descriptions of the generated sinograms. Specifically, the local discriminator randomly picks $1/4 \times 1/4$ local patches from super-resolved sinograms to enhance local sinusoidal details.

Postprocessing Refinement Network (PRN) PRN is connected to SIN by the FBP operator, and it includes a U-Net and a patch discriminator.

One major consideration in designing a second-step network is to prevent propagation of errors introduced by our SIN. To solve this, we introduce a set of reconstructions from multi-resolution sinograms for our PRN to learn from. We observe that SV-CT reconstructions already show sharp partial information although severely obscured by aliasing artifacts. Effective use of such information can be achieved by constructing a pool of diverse features for the network to learn from.

In particular, we downsample the SIN-inpainted sinograms along the angular axis by factors of 2 and 4. Then four different FBP reconstructions from sparse, $2\times$downsampled, $4\times$downsampled, and fully-inpainted sinograms are concatenated as a 4-channel input to the PRN.
3.3. Discriminator Perceptual Network

Perceptual losses have proven to significantly enhance the ability of networks to generate images with high perceptual quality [49] when compared to traditional pixel-wise losses. The original perceptual loss computes the $l_2$ norm of generated and target image pairs through a pre-trained Visual Geometry Group (VGG) network [18], which serves as a feature extractor that enforces high-level feature fidelity.

We propose a novel Discriminator Perceptual (DP) loss that is easy to incorporate into the formulation and especially suitable for domain-specific tasks. It interprets the first few layers of a discriminator as a feature extractor, which is trained simultaneously with the generator on the current problem, and encourages feature-level similarity, see Fig. 5. The DP-loss is computed in a similar fashion as the original perceptual-loss with VGG.

We find out that introducing the DP-loss promotes stability in our GAN training procedure. For example, we can train the discriminator until convergence at each iteration of the minimax game. An incremental strategy for training GANs with DP-loss is provided in the supplementary material for further reference.

DP loss is calculated by the $l_2$ norms of the error between outputs of each activation layer $\phi_j$ in the discriminator with respect to the input pair consisting of the generated image $\hat{y}$ and the target image $y$. While one can choose any such layer $j$ in a discriminator, we compute the averaged norm for each $\phi_j$ before the last one (i.e., we do not include the discriminator output layer):

$$L_{DP}(\hat{y}, y) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{C_j H_j W_j} \|\phi_j(\hat{y}) - \phi_j(y)\|_2^2$$

(1)

where $C_j$ is the number of channels, $H_j$, $W_j$ are height and width of the output at $\phi_j$, respectively, and $N$ is the number of activation layers we use.

3.4. Optimization

Adversarial loss In the GAN-framework [13], the generator is augmented by a discriminator that discerns real and synthesized images. During training, the generator and the discriminator compete in a min-max game that leads to the adversarial loss.

Given a corrupted image $x$ and a target image $y$, we formulate the optimization target as follows:

$$\min_G \max_D V(G, D) = \mathbb{E}_{y \sim \text{target}}[\log D(y)] + \mathbb{E}_{x \sim \text{corrupt}}[\log(1 - D(G(x))],$$

(2)

where $G$ and $D$ represent the generator and discriminator networks, respectively.

Specifically, in SIN, the input $x \in \mathbb{R}^{S \times P}$ is the pre-processed sparse-view sinogram to $S$ detector pixels and $P$ angles and target $y \in \mathbb{R}^{S \times P'}$ is a preprocessed full-view sinogram. Local discriminators compare $G(x)_{\text{patch}}$ and $y_{\text{patch}} \in \mathbb{R}^{[S/4] \times [P'/4]}$, which are small patches from the sinogram randomly chosen from $G(x)$ and $y$, respectively. In PRN, $x$ and $y \in \mathbb{R}^{S \times S}$ form the input and target reconstructions pair.

Content Loss Besides the adversarial loss, we ensure the data fidelity of the generated images by applying an $l_1$-norm based pixel-reconstruction loss:

$$L_c(\hat{y}, y) = \|\hat{y} - y\|_1,$$

(3)

where $\hat{y}$ is the generated and $y$ the target image. The content loss is calculated in both SIN and PRN. As suggested in [51], we choose the $l_1$ loss since our experiments showed slightly sharper reconstructions compared to the $l_2$ loss.

High-Frequency (HF) Loss Neural networks tend to prioritize learning of low-frequency features during training [35]. While it might be acceptable for PRN since tomographic reconstructions tend to be relatively smooth, it is of critical importance for SIN where high-frequency information is essential for correct reconstructions. This is because FBP requires ramp filtering [30] which is a high-pass filter compensating for oversampling the low-frequencies, as illustrated in Fig. 6.

To incorporate this CT-domain knowledge, we enforce our SIN to restore an accurate ramp-filtered sinogram by defining a high-frequency loss:

$$L_{HF}(\hat{y}, y) = \|\hat{y} * h - y * h\|_1,$$

(4)

where $h$ is the ramp kernel in the spatial domain, and $*$ denotes convolution. Note that the high-frequency loss is fully backpropagatable [36].

Total Objective Finally, the total objective function of the SIN generator is the weighted sum of the four losses mentioned above:

\[ L(G) = \lambda_1 L_{adv}(G) + \lambda_2 L_c(G, D) + \lambda_3 L_{HF}(G, D) + \lambda_4 L_{DP}(G, D), \]
\[ G_{SIN}^* = \arg\min_G [\lambda_1 L_{adv}(G) + \lambda_2 L_c + \lambda_3 L_{DP} + \lambda_4 L_{HF}], \]

where \( L_{adv} \) defines the adversarial loss [13] (details in the supplementary material) and \( \lambda_i \) are empirically chosen weighting parameters. Based on our experiments, we set \( \lambda_1 = 1, \lambda_2 = 50, \lambda_3 = 20, \lambda_4 = 50 \). All discriminator adversarial losses are assumed a weight equal to one.

Similarly, the total objective function for the PRN is:

\[ G_{PRN}^* = \arg\min_G [\lambda_1 L_{adv}(G) + \lambda_3 L_{DP}], \]

where all parameters \( \lambda_i \) are chosen the same as for SIN.

4. Experiments

Our experiments are performed on the open-source TCIA LDCT-and-Projection-Dataset [5, 31]. We extensively evaluate our methodology both qualitatively using samples of chest, abdomen, and head images and quantitatively using the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR).

We define a full-view target sinogam \( y^* \in \mathbb{R}^{S \times P^*} \) to have \( S = 320 \) detector pixels and \( P^* = 180 \) uniformly-distributed angles over \( \pi \). The sparse-view input includes 23 angles sampled from \( y^* \) once every eight projections, and the reconstructed image have a resolution of \( 320 \times 320 \) pixels. Due to the proposed two-ends preprocessing, we take six angles from each end of \( y^* \) to generate a 192-angle sinogram \( y \in \mathbb{R}^{S \times (P^* + 12)} \) as the ground-truth for SIN.

4.1. Datasets

Our dataset contains 5394 images resized to \( 320 \times 320 \) stemming from 68 patients taken from the TCIA dataset. Additionally, we augment each sample with a random affine transform to increase data diversity during training. Note that this transformation is performed on the reconstructions, which leads to significant variation in the sinogram domain. We then extract a set of around 500 images from the dataset patient-wise for testing. Once we choose a patient for testing, we remove all data related to this patient from the training set to avoid inter-slice correlation. Full-view sinograms are generated using parallel-beam (orthographic) projections over an angular range of 180 degree, which define the target set for SIN.

4.2. Training Details

Following the standard GAN procedure [13,16], the generator and both global and local discriminators are trained alternatively. The ADAM [21] solver with momentum parameters \( \beta_1 = 0.5 \) and \( \beta_2 = 0.999 \) are used to optimize all networks. We set the learning rate to 0.0001 and run 100 epochs or until convergence for all experiments. For further details on the training procedure, please consult the supplementary material.

4.3. Qualitative Results

We first show our results in both measurement and reconstruction domains in Fig. 7. Compared with a naive linear interpolation, our super-resolved sinograms accurately restore the smooth sinusoidal structures. It is nearly impossible to discern difference between SIN-generated and ground-truth sinograms. A subsequent reconstruction in Fig. 7d obtained by FBP shows an artifact-reduced reconstruction that resembles the target image. However, there remains a novel type of less severe artifacts. These localized artifacts can be efficiently removed by PRN in the second step. The final result of our SIN-4c-PRN preserves details such as bones and soft-tissue contrast.

Comparison to State-of-the-Art In Fig. 8, we further compare our SIN-4c-PRN model to state-of-the-art CT reconstruction approaches.

FISTA-TV [3] is a powerful iterative reconstruction algorithm that employs TV-based regularization. We use the implementation in the open-source toolkit ToMoBAR [20]. To select the optimal hyperparameters of FISTA-TV, we perform a grid-search and choose the set with the best visual results. FISTA-TV preserves edges well, but the reconstructions are over-smoothed, which is a well-known weakness of the TV-prior.

We further compare with a learning-based sinogram inpainting network that uses a conditional-GAN (cGAN)
Figure 8: Baseline results with zoom-in details. From top to bottom row: **Chest**: In our SIN-4c-PRN result, both soft tissues in the black region of the body and sharp edges of bones are maximally recovered compared to other methods. **Abdomen**: Soft tissues with low contrast, such as in the green box are hard to recover from sinograms and pose a significant challenge for our networks. **Head**: The high-frequency details in the nose and bone regions are reconstructed while others either fail or over-smooth them.

Based on an Encoding-Decoding U-Net [11] with strided-convolutions. We implement their architecture and tuned hyper-parameters to obtain the best results. Our experiments show that checkerboard artifacts still exist in generated sinograms due to the strided convolutions, see supplementary material for example images. And these artifacts are further amplified in reconstructions. Note that the original work uses much lower undersampling factors than we are aiming for in this paper, which might explain the discrepancy to their findings.

Finally, we compare against our own PyTorch implementation of Neumann networks [12]. Neumann networks are a variant of unrolled optimization that learns a CNN-based regularizer. They have proven to yield great results in solving super-resolution and MRI problems. Note that unrolling networks require a back-propagatable tomography operator, for which we use torch-radon [36]. Despite extensive tuning of hyper-parameters, our implementation fails to remove streak artifacts and over-smoothness.

### 4.4. Quantitative Results

For a quantitative evaluation we measure the PSNR and SSIM statistics within a circular region-of-interest using our test dataset. Table 1 shows that our proposed SIN-4c-PRN model significantly improves the reconstruction accuracy compared to the other state-of-the-art reconstruction methods discussed earlier. In particular, our model improves the average PSNR by over 4 dB and SSIM by 5%, with narrowed variance.

### 5. Ablation Study

This paper proposes several submodules which provide the best results when all used in conjunction. Yet, the com-
Table 1: Performance of baseline models compared with our SIN-4c-PRN model.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR (σ)</th>
<th>SSIM (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FISTA-PD-TV [3]</td>
<td>30.61 (2.67)</td>
<td>0.839 (0.036)</td>
</tr>
<tr>
<td>cGAN [11]</td>
<td>30.86 (1.92)</td>
<td>0.762 (0.043)</td>
</tr>
<tr>
<td>Neumann Network [12]</td>
<td>28.72 (2.09)</td>
<td>0.697 (0.069)</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>34.90 (2.15)</strong></td>
<td><strong>0.877 (0.029)</strong></td>
</tr>
</tbody>
</table>

Table 2: Reconstruction performance with different composition of our models.

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR (σ)</th>
<th>SSIM (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse-View</td>
<td>22.80 (1.77)</td>
<td>0.485 (0.042)</td>
</tr>
<tr>
<td>PRN</td>
<td>32.95 (1.96)</td>
<td>0.835 (0.034)</td>
</tr>
<tr>
<td>SIN</td>
<td>34.19 (2.35)</td>
<td>0.859 (0.036)</td>
</tr>
<tr>
<td>SIN-PRN</td>
<td>34.61 (2.14)</td>
<td>0.873 (0.035)</td>
</tr>
<tr>
<td>SIN-4c-PRN</td>
<td><strong>34.90 (2.15)</strong></td>
<td><strong>0.877 (0.029)</strong></td>
</tr>
</tbody>
</table>

Table 3: Reconstruction performance of different sub-modules and perceptual losses trained on single SIN model.

<table>
<thead>
<tr>
<th>Module</th>
<th>Perceptual Loss</th>
<th>PSNR (σ)</th>
<th>SSIM (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TE</td>
<td>✗</td>
<td>30.84 (1.87)</td>
<td>0.785 (0.046)</td>
</tr>
<tr>
<td>HF</td>
<td>✗</td>
<td>31.65 (2.15)</td>
<td>0.812 (0.050)</td>
</tr>
<tr>
<td>DP</td>
<td>✗</td>
<td>31.71 (2.14)</td>
<td>0.808 (0.049)</td>
</tr>
<tr>
<td>VGG</td>
<td>✗</td>
<td>33.39 (2.09)</td>
<td>0.829 (0.037)</td>
</tr>
<tr>
<td>SIN-PRN</td>
<td>✗</td>
<td>33.51 (2.02)</td>
<td>0.852 (0.036)</td>
</tr>
<tr>
<td>SIN-4c-PRN</td>
<td>✗</td>
<td><strong>34.19 (2.35)</strong></td>
<td><strong>0.859 (0.036)</strong></td>
</tr>
</tbody>
</table>

Table 3 includes comparison among non-perceptual loss, VGG16 and DP loss on sinogram data. Our DP network shows the best results on both metrics. Qualitative results and sinogram domain metrics in the supplementary material show that the inpainted sinograms with DP are considerably better than with VGG16 and lose fewer details, especially in low-contrast regions.

6. Conclusion

In this paper, we propose a two-step reconstruction framework for sparse-view CT problems. Our SR-network successfully learns the challenging task of upsampling a thin sinogram. The subsequent refinement network robustly removes the remaining artifacts. The full network outperforms state-of-the-art reconstruction algorithms. One of our core contributions is the discriminator perceptual network that enforces the network to incorporate high-level domain-specific knowledge. We further introduce task-specific sub-modules tailored to the sparse-view tomography problems such as two-ends processing and the high-frequency loss. Moreover, to test each module for effectiveness, we perform a comprehensive ablation study.

The two-step procedure showcases how to bring explainability to image reconstruction problems by tailoring the solution to the idiosyncrasies of the imaging environment. We strongly believe that our approach is not limited to CT, and we hope to inspire other imaging applications with similar ideas for pushing the limits of imaging.

References


[29] Zhengchun Liu, Tekin Bicer, Rajkumar Kettimuthu, Doga Gursoy, Francesco De Carlo, and Ian Foster. Tomogan: low-dose synchrotron x-ray tomography with generative adver-


[31] CH McCollough, B Chen, DI Holmes, X Duan, Z Yu, L Xu, S Leng, and J Fletcher. Low dose ct image and projection data [data set]. The Cancer Imaging Archive, 2020. 6


[40] Robert Tovey, Martin Benning, Christoph Brune, Marinus J Lagerwerf, Sean M Collins, Rowan K Leary, Paul A Midgley, and Carola-Bibiane Schönlieb. Directional sinogram inpainting for limited angle tomography. Inverse problems, 35(2):024004, 2019. 2


1. Introduction

This document contains the supplementary materials that were left out in our main submission since they are too detailed for and do not significantly enhance the message we want to convey. Section 2 presents an incremental strategy for training GAN with a discriminator perceptual losses. Section 3 explains in detail the adversarial loss function mentioned in the main paper. The model architecture and hyperparameters are shown in Sec. 4. Finally, more qualitative and quantitative results for ablation study are presented in Sec. 5.

We provide all code that is necessary to reproduce the experiments reported in the paper and supplementary at https://github.com/anonyr7/Sinogram-Inpainting. This git-repository contains PyTorch implementations for our proposed networks, state-of-the-art methods, our CT dataset and a sub-set of around 30 images as examples for evaluation of each model.

2. Discriminator Perceptual Loss

We present an alternative training strategy using the DP loss in Alg. 1. We show the example procedure with a super-resolution task that learns from a low-resolution (LR) image to a high-resolution (HR) image. We start by training the discriminator once in every generator iteration, then increment the discriminator iteration by one in every \( k \) generator iterations, where \( k \) is an arbitrary constant, for which we chose \( k = 10 \) in our paper. The intuition is that with more training performed, the discriminator learns better high-level features, thus benefits the generator using the DP loss.

### Algorithm 1: An Incremental Training Procedure of a naive Super-Resolution GAN with DP loss.

**Data:** LR and HR images; Generator training iteration \( n \); Discriminator increment period \( k \).

**Result:** Super-resolved (SR) images.

**while not converge or iteration \( i \leq n \) do**

- Sample a batch of LR and HR image pairs \( x \) and \( y \) from the training data;
- Generate a SR image \( G(x) \) from generator \( G \);
  **while discriminator iteration \( j \leq \lceil i/k \rceil \) do**
    - Update discriminator \( D \) with \( \mathcal{L}_{adv}(D) \);
    - \( j = j + 1 \);
  **end**
- Calculate \( \mathcal{L}_{DP} \) from \( D \) using \( G(x) \) and \( y \);
- Update \( G \) with \( \mathcal{L}_{DP} \) and \( \mathcal{L}_{adv}(G) \);
- \( i = i + 1 \);
**end**

3. Objective Function Details

3.1. Adversarial Loss

For a typical GAN network that generates images from corrupted input \( x \) to target distribution \( y \), the adversarial losses for the generator and discriminator are defined as:

\[
\mathcal{L}_{adv}(G) = \mathbb{E}_{x \sim \text{corrupt}}[\log(1 - D(G(x)))]
\]

\[
\mathcal{L}_{adv}(D) = \frac{1}{2} \left( \mathbb{E}_{y \sim \text{target}}[\log D(y)] + \mathbb{E}_{x \sim \text{corrupt}}[\log(1 - D(G(x)))] \right)
\]

Equation 1 applies to the SIN with global discriminator and PRN model. For the local discriminator of SIN, a pair of image patches \( G(x)_{\text{patch}} \) and \( y_{\text{patch}} \) randomly selected from the generated image \( G(x) \) and target image \( y \) pair are
used for calculating an additional local adversarial cost for SIN:

$$L_{\text{local adv}}(G) = \mathbb{E}_{x \sim \text{corrupt}}[\log(1 - D(G(x)_{\text{patch}}))]$$,

$$L_{\text{local adv}}(D) = \frac{1}{2}(\mathbb{E}_{y \sim \text{target}}[\log D(y_{\text{patch}})] +$$

$$\mathbb{E}_{x \sim \text{corrupt}}[\log(1 - D(G(x)_{\text{patch}}))])$$,

(2)

Therefore, the total adversarial loss for SIN is the sum of Eq. 1 and Eq. 2 with equal weights.

### 4. More Training Details

#### 4.1. Detailed Architecture

As discussed in the main paper, our generator uses an adapted U-Net architecture, which consists of four average pooling and four bilinear upsampling layers, each scales by a factor of two in both dimensions. Each of the scaling layers is followed by a double-conv block, which consists of two sets of convolution, ReLU activation and batch normalization layers. All convolution layers in the U-Net have kernel size of three and stride equal to one.

We also use a patch discriminator [2], where the first and last convolution layers have kernel size of three and stride of one, and the middle two convolutions have kernel size of four and stride of two. It also includes LeakyReLU activation and batch normalization layers. In total, the patch discriminator outputs a patch of size $1/4 \times 1/4$ than before. Please find the implementation of this architecture here: https://github.com/anonyr7/Sinogram-Inpainting/blob/master/SIN/model.py

#### 4.2. Data Augmentation

We augment our dataset by applying random affine transformation to each image. Specifically, we randomly perform the following operations: rotate between $\pm 30^\circ$, translate in two dimensions between $\pm 0.1$, scale by 0.5 to 1.1, and shear between $\pm 20^\circ$. Such augmented images are only used for training phase, in order to bring more diversity to both sinogram and reconstruction training sets.

The code for our data augmentation is found in this Jupyter notebook: https://github.com/anonyr7/Sinogram-Inpainting/blob/master/TCIA_data_preprocessing.ipynb

### 5. More Results from Ablation Study and State-of-the-Art

In this section, we show complement of the ablation study discussed in the main paper. Specifically, we present more visual results of different composition of our models, analysis of the cascaded inputs and more sinogram-domain results for our discriminator perceptual loss.

#### 5.1. Checkerboard Artifacts in Sinograms

We present more sinogram results from the cGAN model introduced by Ghani et al [1] in Fig. 3. The sinograms exhibit streak artifacts especially on the boundary areas, causing the reconstructions to have skewed streak artifacts. More examples can be found at https://github.com/anonyr7/Sinogram-Inpainting/tree/master/Toy-Dataset/CGAN_sinogram.

#### 5.2. Qualitative comparison among our models

We present the visual results of different compositions of our models in Fig. 1, corresponding to the quantitative metrics among our models in the main paper. We show that although a single PRN model generates smooth and clean images, they still suffer from unremoved streak artifacts or hallucinate wrong details that are critical in analysis of the reconstructions. Meanwhile, we also observe secondary artifacts from single SIN model. This is in correspondence with our argument that single-domain models are not enough for SV-CT reconstruction tasks, and our SIN-4c-PRN model benefits from both domain representations.

#### 5.3. Effectiveness of cascade inputs of PRN

Our proposed model SIN-4c-PRN is a concatenation of SIN and PRN models trained in two steps. In particular, we propose a four-channel input for PRN that consists of reconstructions from sinograms with different undersampling factors.

Specifically, we create an input with each channel being a FBP reconstruction from a sinogram with 23, 45, 90 and 180 angles, respectively. The 23-angle sinogram is the original sparse-view sinogram, and the intermediate angles are downsampled from the 180-angle SIN-inpainted sinogram.

In order to make fair comparisons of the proposed cascade inputs, we train two networks with the same capacity, i.e., number of learnable parameters. SIN-4x-PRN is a model with a repeated four-channel input, each channel being the same FBP reconstruction from the 180-angle SIN-inpainted sinogram. We compare the performance of SIN-4x-PRN with SIN-4c-PRN in Tab. 1. Both PSNR and SSIM metrics show that SIN-4c-PRN performs better than SIN-4x-PRN, indicating that the model learns from the cascaded inputs with different level of representations.

#### 5.4. Sinogram results for different perceptual losses

We qualitatively and quantitatively compare the sinogram results of our DP loss with the original VGG16 perceptual loss [3] trained with a SIN model. The SIN model learns 1D super-resolution from a 23-angle sparse-view sinogram to a 180-angle full-view sinogram. The models used for comparison share the same architecture, data and
Figure 1: Ablation comparison with zoom-in details. From top to bottom row: **Chest:** While PRN generates a visually clean image, it loss important details such as the bones in the blue box. In contrast, SIN-4c-PRN provides both smooth and detail-preserved reconstructions. **Abdomen:** PRN fails to remove the heavy streak artifacts in the green box area. **Head:** PRN hallucinates non-exist features on the nose and eye regions, while SIN-4c-PRN produces results close to the ground truth.

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR (σ)</th>
<th>SSIM (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIN-PRN</td>
<td>34.61 (2.14)</td>
<td>0.873 (0.035)</td>
</tr>
<tr>
<td>SIN-4x-PRN</td>
<td>34.62 (1.97)</td>
<td>0.874 (0.029)</td>
</tr>
<tr>
<td>SIN-4c-PRN</td>
<td><strong>34.90 (2.15)</strong></td>
<td><strong>0.877 (0.029)</strong></td>
</tr>
</tbody>
</table>

Table 1: Reconstruction performance comparison. SIN-4x-PRN: Same model with SIN-4c-PRN, but duplicating input reconstruction by four times. σ denotes standard deviation.

In the main paper we compared the PSNR and SSIM metrics of the FBP reconstructions of result sinograms. Here, we further provide metrics measured directly on the result sinograms in Tb. 2. We observe better metrics of DP than VGG16, despite both of them have slightly lower PSNR than without perceptual loss. This is consistent with the observation in the original work by Johnson et al. [3], because perceptual losses are optimized with a different criterion, i.e., the $l_2$-norm is minimal when one optimizes for perceptual losses.

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR (σ)</th>
<th>SSIM (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Perceptual VGG16 [3]</td>
<td>42.50 (2.54)</td>
<td>0.979 (0.005)</td>
</tr>
<tr>
<td>Our (DP)</td>
<td>39.70 (2.25)</td>
<td>0.984 (0.004)</td>
</tr>
<tr>
<td>Non-Perceptual VGG16 [3]</td>
<td>41.23 (2.83)</td>
<td>0.988 (0.005)</td>
</tr>
<tr>
<td>Our (DP)</td>
<td>34.19 (2.35)</td>
<td>0.859 (0.036)</td>
</tr>
</tbody>
</table>

Table 2: Different perceptual loss performance trained with a single SIN. ‘S’ and ‘R’ denote Sinogram and Reconstruction domain, respectively. σ denotes standard deviation.
the $l_2$ loss.

Then we show the visual results of the sinograms and their FBP reconstructions in Fig. 2. Specifically, we provide the sinogram residuals compared to the ground truth for ease of comparison. From these residuals, we find that VGG16 produce slightly more errors than the others in the sinogram domain, i.e., the training domain, possibly because sinograms are drastically different from natural RGB images where VGG16 was trained with. However, we also observe that after the FBP reconstruction, VGG16 and DP produce visually less noisy images than without perceptual losses. The reconstruction domain metrics and residuals indicate the same observation.

References


Figure 2: (Continued) Sinogram residuals and corresponding FBP reconstructions and residuals from a single SIN using no perceptual loss, VGG16 perceptual loss or our discriminator perceptual loss. The residual (error) maps are displayed in the same range, hence we omit the scale-bar for visualization reasons. Both the non-perceptual and our DP generated sinograms have fewer errors than VGG16. However, the reconstructions of Non-perceptual sinograms are slightly noiser and have more remaining streak artifacts than the others.
Figure 3: Checkerboard artifacts in the generated sinograms of the state-of-the-art learning-based approach [1] and consequent artifacts in their FBP reconstructions. **Left:** result sinograms and reconstructions. Checkerboard artifacts are especially severe in the boundary of sinograms. Please zoom in the sinogram to see more closely; **Right:** Ground Truth sinograms and reconstructions.