Improving GAN-Based CT Infilling Via Transfer Learning and Layer Freezing

Roman Fisher, MSc[1]; Tynan Stevens, PhD[1, 2, 3]

[1]Dalhousie University, Dept. of Physics & Atmospheric Science; [2]Dalhousie University, Dept. of Radiation Oncology;
[3]Nova Scotia Health, Dept. of Medical Physics

2026 JOINT AAPM | COMP MEETING
JULY 19–22 | VANCOUVER, BC

2026 JOINT AAPM | COMP MEETING
JULY 19–22 | VANCOUVER, BC

Abstract

Purpose: To assess the impact of transfer learning and layer freezing on GAN-based CT image completion.

Methods: A GAN was pretrained on three CT datasets and ImageNet, then each of the four pretrained models was fine-tuned on the same in-house CT dataset. Performance was scored on a validation set using eight metrics: precision, recall, density, coverage, FID, MSE, SSIM, and PSNR. The best-performing pretrained model was then used for layer-freezing experiments, in which nine independent layer sets from the generator and/or critics were frozen and each used to fine-tune the model. Frozen models were compared against the best unfrozen fine-tuned model on all metrics.

Results: The model pretrained on augmented in-house data outperformed all others on all eight metrics, with statistically significant improvements over the model trained solely on the original in-house dataset. Freezing the shallowest half of the critics and the generator's encoding branch significantly improved recall, FID, MSE, and SSIM over the best unfrozen model, with no significant degradations on any metric.

Conclusion: Transfer learning with an augmented dataset improves GAN-based CT infilling, and layer freezing can further boost performance on select metrics. Future work will examine impacts on radiotherapy treatment planning and extend the approach to three dimensions.

Introduction

Standard-of-care methods for correcting metal artifacts on radiotherapy treatment planning CT images, such as uniform density approximations, can be dosimetrically crude. This work seeks to develop a GAN training pipeline for CT image completion to address this issue. It has been shown that transfer learning approaches to image synthesis can mitigate overfitting, reduce data requirements, and improve semantic coherence of output [1]. This is done by training the model on a broad set of data and fine-tuning it on a smaller dataset more aligned with the purposes of the model's intended task. Layer freezing takes this approach further by fixing the weights of any given model layer to ensure certain desired model behaviours are not lost when fine-tuning.

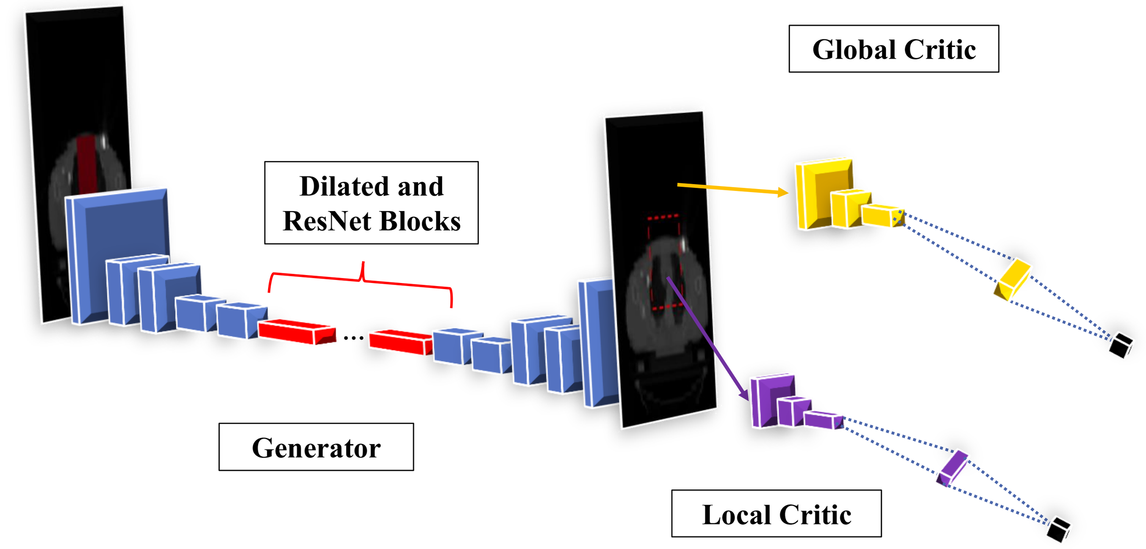

Figure 1. Schematic of the GAN used for this study.

Methods and Materials

Network architecture: Based on that of Iizuka et al. (2017) [2]. See *Figure 1* for a pictorial representation.

Pretraining: The network was used to pretrain four models for 541,000 iterations. Each model was pretrained using one of the following four datasets: ImageNet ($9.1*10^6$ samples), collated ($3.7*10^5$ head and neck samples from the Cancer Imaging Archive), in-house data ($1.5*10^4$ head and neck samples from the NS Cancer Centre archives), and augmented in-house data ($3.7*10^5$ samples synthesized by applying affine transformations to in-house data). Ninety per-cent of each of the datasets were used for training; remaining data was reserved for validation purposes. All images were 512x512 pixels in resolution.

Transfer learning experiment: Each model was fine-tuned on the in-house dataset for 50 epochs.

Transfer learning assessment: The resulting fine-tuned models were scored on a validation set of in-house data and compared on each of eight metrics: precision [3], recall [3], density [3], coverage [3], Fréchet Inception Distance (FID) [3], MSE, SSIM, and PSNR. While training was completed using amorphous 'blob' masks, metrics were scored on a centred 192x192 pixel patch with 96x96 pixels occluded by the mask region.

Layer freezing experiment: The pretrained model which scored best after transfer learning was used for this experiment. To test whether layer freezing aids in developing a CT image completion GAN, several regimes were tested. Each of the following sets of layers were frozen for a 50-epoch fine-tuning session on in-house data: the encoding branch (FG), the encoding branch and the shallowest half of the critics (FGFC), the decoding branch (BkG), the decoding branch and the shallowest half of the critics (BkGFC), the shallowest half of the generator (TG), the shallowest half of the generator and the shallowest half of the critics (TGFC), the deepest half of the generator (BG), the deepest half of the generator and the shallowest half of the critics (BGFC), and all convolutional layers in the generator and both critics. *Figure 2* is a visualization aid showing the different freezing regimes.

Layer freezing assessment: The resulting models fine-tuned on in-house data with partially fixed weights were scored on a validation set of in-house data and compared on each of eight metrics: precision, recall, density, coverage, FID, MSE, SSIM, and PSNR. Metrics were scored on a centred 192x192 pixel patch with 96x96 pixels occluded by the mask region.

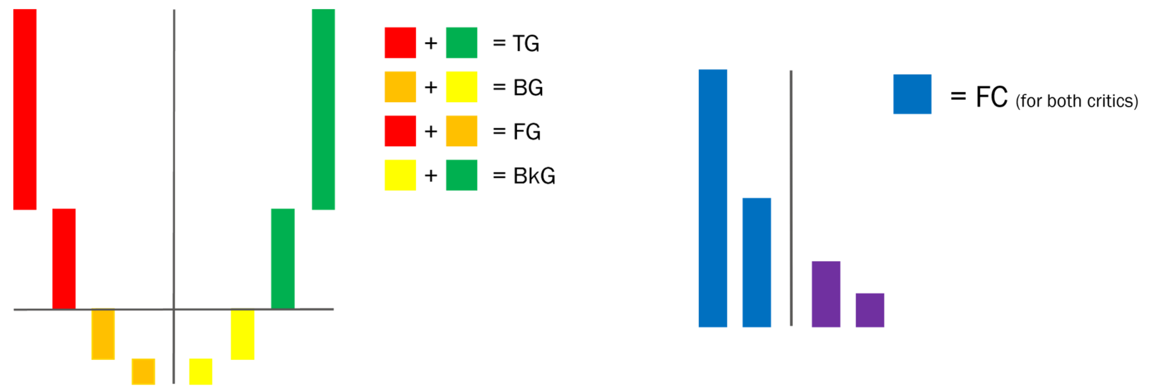

Figure 2. Schematic showing the layer freezing regimes.

Results

In *Table 1*, it can be seen that the model pretrained on augmented in-house data outperformed the other models on every metric. These improvements were statistically significant (P<0.05 for each, two tailed) with respect to the baseline model pretrained and fine-tuned on the original in-house dataset. Further improvements upon the results of the transfer learning experiment were found when freezing the encoding branch and the shallowest half of the critics (FGFC). Statistically significant improvements when employing this freezing regime were observed for SSIM, MSE, recall and FID (see *Table 2*). No statistically significant degradations were observed in this case.

Figure 3 compares the MSE score at various stages of this study and the score achieved by infilling the square region uniformly with various CT-number values. The best-scoring CT-number infill tested was -180.8HU, which achieved a value of 0.012023, an order of magnitude greater than the best performing GAN models. Qualitative samples can be found in *Figure 4*.

Evaluation Metric	Augmented In-House	In-House	Collated	ImageNet
SSIM	**0.96639** ($8*10^{-5}$)	0.95176 ($5*10^{-5}$)	0.9639 (0.0002)	0.9582 (0.0006)
MSE	**0.003489** ($1.1*10^{-5}$)	0.005101 ($9*10^{-6}$)	0.003814 ($2.8*10^{-5}$)	0.004412 ($5.7*10^{-5}$)
PSNR (dB)	**22.69 (0.04)**	21.24 (0.01)	21.68 (0.02)	22.42 (0.03)
Precision	**0.5425 (0.0024)**	0.5109 (0.0026)	0.4174 (0.0012)	0.2950 (0.0032)
Recall	**0.520 (0.002)**	0.479 (0.002)	0.413 (0.004)	0.249 (0.006)
Density	**0.3104 (0.0008)**	0.2853 (0.0015)	0.2161 (0.0009)	0.1138 (0.0018)
Coverage	**0.504 (0.001)**	0.464 (0.001)	0.391 (0.001)	0.250 (0.002)
FID	**38.34 (0.06)**	39.82 (0.04)	45.81 (0.10)	55.74 (0.36)

Table 1. Metric scores achieved by models pretrained using the indicated dataset after fine-tuning.

Evaluation Metric	UF	FGFC	% Difference	U	p
SSIM	0.96639 ($8*10^{-5}$)	**0.96721** ($3*10^{-5}$)	0.08451	0	**0.0209**
MSE	0.003489 ($1.1*10^{-5}$)	**0.003415** ($3*10^{-6}$)	2.13551	0	**0.0209**
PSNR (dB)	22.69 (0.04)	**22.78 (0.01)**	1.9731	3	0.1489
Precision	0.5425 (0.0024)	**0.5506 (0.0019)**	1.480	4	0.2428
Recall	0.520 (0.002)	**0.5342 (0.003)**	2.65	1	**0.0433**
Density	**0.3104 (0.0008)**	0.3095 (0.0010)	0.3103	7	0.9819
Coverage	0.504 (0.001)	**0.506 (0.003)**	0.478	7	0.9819
FID	38.34 (0.06)	**37.67 (0.10)**	1.737	0	**0.0209**

Table 2. Metric scores achieved by the unfrozen fine-tuned model pretrained on augment data (UF) and the best frozen model (FGFC).

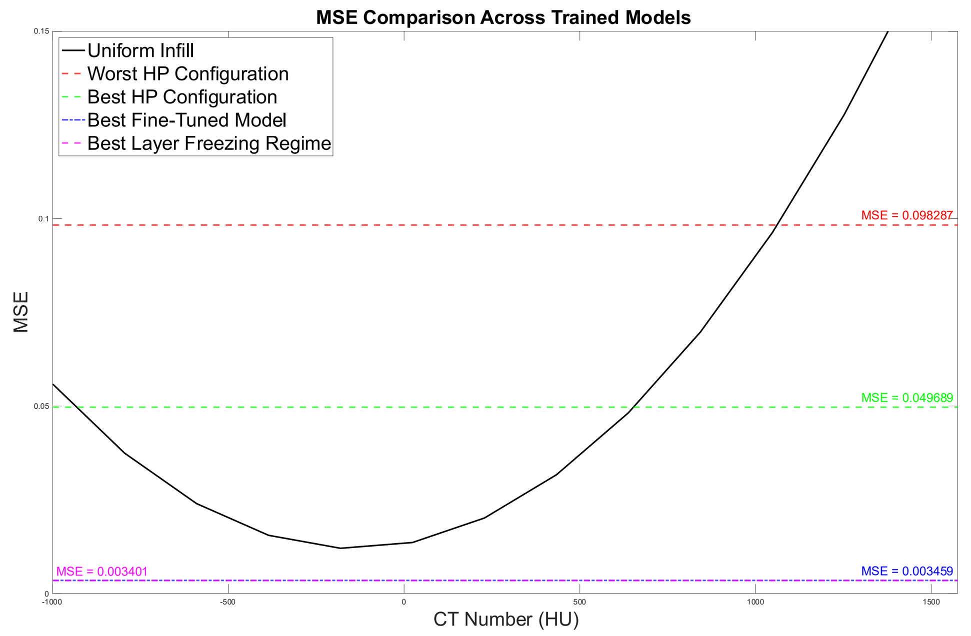

Figure 3. Comparison of MSE values across stages of the study. Reference lines showing the effect of choosing different hyperparameter configurations (from a previous study) are also provided.

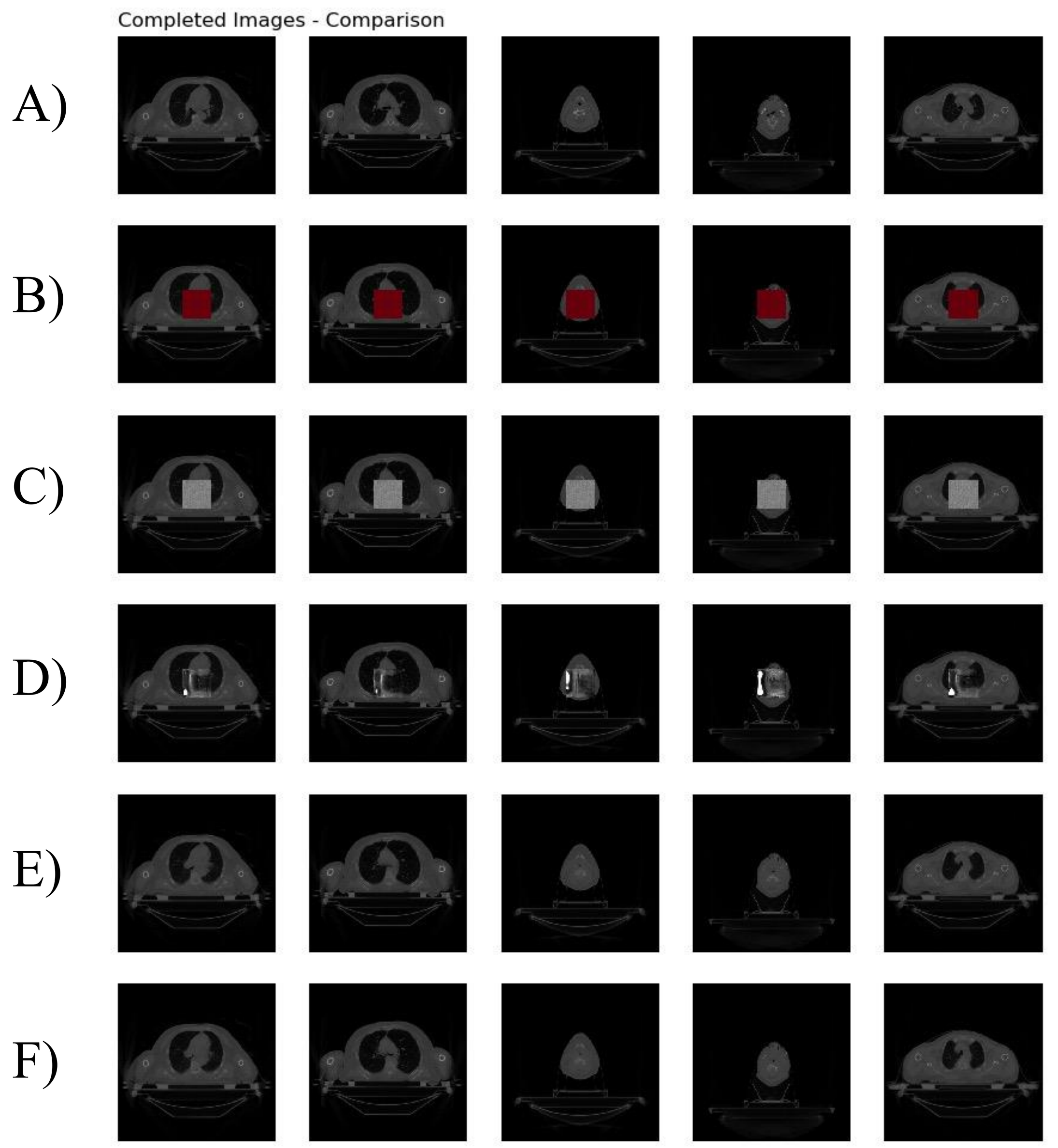

Figure 4. A qualitative performance comparison for: A) ground truth images, B) sample masked images, C) images completed by the model pretrained for 20 epochs on ImageNet data with the worst hyperparameter configuration and D) the best hyperparameter configuration, E) images completed by the fine-tuned model which was pretrained on augmented data, and F) images completed by the model fine-tuned with the best layer freezing regime.

Discussion & Conclusions

By employing a transfer learning approach, GAN-based CT infilling performance can improve without requiring the use of additional clinical data. Layer freezing shows promise in aiding to improve performance further on some measures, though with diminishing returns. By achieving a lower MSE value than uniformly infilling the mask region with a given CT-number value, GAN-based image completion shows promise as a future method of metal artifact correction. Current and future work will investigate whether performance can be improved using a three-dimensional approach. The current two-dimensional model will be compared to a 3D approach in terms of dosimetric impact of treatment plans and their effects on autocontouring techniques.

Contact

Roman Fisher, MSc
Dalhousie University
Email: Roman.Fisher@dal.ca
Phone: (902) 890-6087

DALHOUSIE UNIVERSITY

nova scotia health

Funding Acknowledgement

nova scotia health innovation hub

References

[1] Jun Lv, Li G, Tong X, Chen W, Huang J, Wang C, et al. Transfer learning enhanced generative adversarial networks for multi-channel MRI reconstruction. Computers in Biology and Medicine. 2021 July 1;134:104504–4.

[2] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion. ACM Transactions on Graphics. 2017 July 20;36(4):1–14.

[3] Borji A. Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding. 2022 Jan;215:103329.