Enhancing Radiographic Diagnosis: CycleGAN-based methods for reducing cast shadow artifacts in wrist radiographs

Stanley A Norris,[1,2] Daniel Carrion,[1] Michael Ditchfield,[1,3] Manuel Gubser,[4] Mohamed K Badawy[1,5]

[1] Monash Imaging, Monash Health, Clayton, Australia
[2] School of Science, RMIT University, Melbourne, Australia
[3] Department of Paediatrics, Monash University, Melbourne, Australia
[4] Division of Radiology and Nuclear Medicine, St. Gallen Cantonal Hospital, St. Gallen, Switzerland
[5] Department of Medical Imaging and Radiation Sciences, Monash University, Melbourne, Australia

Abstract

Objective. We extend existing techniques by using generative adversarial network (GAN) models to reduce the appearance of cast shadows in radiographs across various age groups. *Materials and Methods.* We retrospectively collected 12000 adult and pediatric wrist radiographs, evenly divided between those with and without casts. The test subset consisted of 100 radiographs with cast and 100 without cast. We extended the results from a previous study that employed CycleGAN by enhancing the model using a perceptual loss function and a self-attention layer. *Results.* The CycleGAN model which incorporates a self-attention layer and perceptual loss function delivered the best quantitative performance. This model was applied to images from 20 cases where the original reports recommended CT scanning or repeat radiographs without the cast, which were then evaluated by radiologists for qualitative assessment. The results demonstrated that the generated images could improve radiologists' diagnostic confidence, in some cases leading to more decisive reports. Where available, the reports from follow-up imaging were compared with those produced by radiologists reading AI-generated images. Every report, except one, provided identical diagnoses as those associated with follow-up imaging. The ability of radiologists to perform robust reporting with downsampled AI-enhanced images is clinically meaningful and warrants further investigation. Additionally, radiologists were unable to distinguish between AI-enhanced and unenhanced images. *Conclusion.* These findings suggest the cast suppression technique could be integrated as a tool to augment clinical workflows, with the potential benefits of reducing patient doses, improving operational efficiencies, reducing delays in diagnoses, and reducing the number of patient visits.

Keywords: generative adversarial network (GAN), artifact suppression, cast shadow, deep learning, wrist radiographs

1. Introduction

Damaged limbs often require stabilisation using splints and casts made of resin, fibreglass or plaster (1). Removing the cast before imaging is impractical during follow-up radiographs monitoring the position or healing of fractures. The presence of a cast produces undesirable artefacts obscuring the visualisation of the bone structure and encompasses the examined anatomy. At the time of writing, there has been only a single study on cast suppression methods in extremity radiographs, by Hržić et al. (2). We attempt to reproduce these results on a larger data set and modify the model by adding a perceptual loss function and self-attention layer (3,4).

The field of medical imaging has significantly advanced due to the use and application of artificial intelligence (AI) methods in an extensive range of tasks. These models can be categorised as being vision, language, and vision-language models, depending on the nature of input and output data. In the context of cast suppression, the relevant vision model is known as a generative adversarial network (GAN). GANs have effectively been used for image processing tasks such as CT denoising, artefact reduction, radiotherapy planning, intermodality image synthesis, image reconstruction, data augmentation, image registration, classification, and inversion problems (5,6). When faced with an unpaired image problem, where there is no ground truth for the model to learn from, the cycle-consistent GAN (CycleGAN) model can be used. This approach can translate an image from a source domain to a target domain without paired images, so the distribution of generated images is indistinguishable from that of target images. A cycle consistency loss is introduced so the original image can be returned from the generated one. This method performs well for textural and colour changes in images but poorly for geometric changes, and the characteristics of the training data used limit its generality (7). Since the available casted radiographs do not have paired castless images, and a cast shadow mainly presents a textural change, the CycleGAN model is well suited to this problem.

One of the difficulties in quantitatively evaluating the model in the cast suppression problem is the lack of ground truth to which results can be compared. In some studies, a limited number of paired images were available and thus used for evaluation rather than training, such as for conversion of CT scans between reconstruction kernels (8), noise reduction in low-dose CT scans (9), or generation of CT images from CBCT images (10). For many different problems, CycleGAN models are evaluated by applying quantitative metrics such as Structural Similarity Index Metric (SSIM), histogram correlation, histogram intersection, Chi-squared distance, and Hellinger distance, which compare real and generated sets of images (2,11–14). The obvious limitation of these metrics is that spatial information is lost when converting an image to a histogram; for example, a circle and square of equal area and brightness would yield equivalent histograms despite representing distinct objects. In the original cast suppression study, a qualitative evaluation of the model was also undertaken by having radiologists of varying experience rank the generated images in terms of subjective quality. One study used a GAN to create lung nodules in CT images. It tested how well radiologists could distinguish real from fake and malignant from benign nodules (15). Another study used GANs to suppress bone in chest radiographs and tested for performance changes in radiologists' ability to detect nodules (16). In what is perhaps the most compelling way of evaluating GAN models, the outputs are used as inputs to other AI models that have well-defined performance metrics. Examples include using GANs to improve segmentation and classification algorithms (17–24). A limitation in evaluating the CycleGAN method for cast suppression is a lack of precise application since there is no evidence that removing the cast shadow improves the quality or speed of extracting relevant clinical information in extremity radiographs. This presents an opportunity to bridge the gap between an exploratory study and an application that may yield tangible clinical benefits.

This study aims to extend the published CycleGAN method using an adult and pediatric dataset by incorporating a perceptual loss function and a self-attention layer. This study further evaluates the

model's performance through quantitative and qualitative analyses and investigates its potential for clinical application.

2. Materials and Methods

This study was exempt from Human Research Ethics Committee review as a retrospective quality improvement project. It was consistent with the NHMRC Ethical Considerations in Quality Assurance and Evaluation Activities (2014) guideline.

2.1 Image data acquisition

The images came from a large metropolitan hospital, including 30000 wrist radiographs taken between 2013 and 2023. We selected 10200 radiographs and divided them equally into groups of images with and without casts. First, original images were converted to 8-bit grayscale PNG files. Preprocessing included adding black pixels to images that weren't square and rescaling to 512x512 pixels by Lanczos interpolation (25). The images were then grouped into training (n=10000) and testing (n=200) subsets. Half of the test images were without cast (n=100) and used as the reference for quantitative assessment, and the other half (n=100) were with cast and used to evaluate model performance.

2.2 Model architecture and implementation

As in the Hržić et al. study, the model is trained to optimise a loss function consisting of adversarial, cycle-consistency, and identity loss. The same training parameters are used for consistency as in previous publications (2,7) (cycle-consistency loss with weight λ =10, Adam optimiser with batch size 4, and learning rate α = 0.002, which was linearly reduced for the final 100 epochs). Due to computational limitations, and since the original study showed that the U-Net 512 architecture gave the best performance, we used this generator with 9 layers each for up- and downsampling. The details of the discriminators we applied are identical to those in the original study. The CycleGAN model was developed to include a perceptual loss (PL) function and self-attention layer (SAL). The PL function compares images based on differences between high-level image feature representations rather than differences between pixels, which may improve the quality of generated images (26). This study used a VGG16 network trained on the ImageNet dataset (27). The SAL added to the generator allows the network to consider the entire input when evaluating parts of the data, which may produce more coherent and higher-quality images by allowing the generator to capture intricate and global patterns in the images (28). The model was trained using Python (v3.10), Pytorch 2.4, and NVIDIA CUDA 12 libraries, on an NVIDIA Tesla P40 GPU with 24GB of VRAM. Training losses can be found in the supplementary information.

2.3 Quantitative image analysis

Applying the CycleGAN model to an image alters its associated pixel value histogram, particularly causing an increase in high-intensity values (2). In essence, the evaluation metrics quantify the similarity between two sets of histograms, which correspond to sets of histograms from real (H1) and generated (H2) castless images. In this study, we use histogram correlation (29), histogram intersection (30,31), Chi-squared distance (32), Hellinger distance (33), and Structural Similarity (SSIM) index (34,35) as quantitative metrics to assess the model. Details about quantitative metrics can be found in the supplementary information.

2.4 Qualitative image analysis

To probe the potential for clinical application, we first identified problematic images where the original radiology report suggested repeat imaging without the cast. We searched the Picture Archiving and Communication System (PACS) for 20 such cases. The model was then applied to these images to generate AI-enhanced versions. Two radiologists with 10 and 39 years of experience were then asked to re-assess each case using the referral request or indication and the original and AI-enhanced images. The radiologists were blinded to the original reports, received all views associated with each imaging study, and could view the casted and decasted images side by side. The images were presented in a Google Form at low resolution (512×512), without the ability to adjust window and level settings. For each case, there was a field for the radiologists to write their report and a checkbox to indicate whether the AI-enhanced images improved their diagnostic confidence. A limitation of the qualitative assessment in the study by Hržić et al. is that observers ranked generated images based on perceived quality, potentially turning the test into a "beauty contest" rather than focusing on diagnostic utility (36). To address this, we compared the radiologists' new reports from our experiment against the original reports and categorised them into three outcomes:

1. The new report is essentially identical in identifying the presence or absence of fractures.
2. The new report is decisive about a fracture suspected in the original report.
3. The new report identifies a fracture not detected in the original report.

Where available, the new radiologist reports generated in this experiment were compared against reports from follow-up CT scans and radiographs without casts. Although the availability of such follow-up studies was limited, this comparison allowed us to establish a "gold standard" against which the new reports could be validated.

Each radiologist reviewed 60 image subsets via Google Forms as a separate Turing test. They were informed that the images may be all unenhanced, AI-enhanced, or a mixture of the two. Each image subset contains 15 real casted images, 15 generated castless images, 15 real castless images, and 15 generated casted images, all randomly selected. These images were unlabelled and randomly shuffled. The radiologists were allowed to zoom and pan within the image. They were also blinded to each other's evaluation and were not shown any sample images before the assessment. The radiologists were asked to classify each image as unenhanced or AI-enhanced. This test determined whether radiologists can distinguish real from generated images, implying that the model can generate high-quality outputs if radiologists cannot detect the AI model (15). The generation of high-quality images could, for example, be leveraged as input for deep learning models.

2.5 Statistical analysis

First, a Shapiro-Wilk test was performed to determine that quantitative metrics were non-normally distributed (37). Since the histograms generated by each model are produced from the same set of real images, and the same set of reference images are used for comparison, we are dealing with paired data. Therefore, a Friedman test was performed since the data is non-normal, dependent, and consists of 4 groups (38). Post hoc analysis was done using the Wilcoxon signed-rank test, which must be applied for each combination of models with six possible combinations for four groups (39). In this case, a Bonferroni adjustment was applied to the threshold p-value for significance (40). The Turing test assessments were tested for inter-rater reliability and inter-observer discrepancies via the Kappa statistic

and Chi-squared tests, respectively (41,42). The statistical analyses were performed using Python (v3.12.4).

3. Results

The dataset included 10,200 Radiographs, with an average patient age of 35 ± 28 years for those with casts and 35 ± 26 years for those without. Of these, 4,996 Radiographs were from male patients, with a mean age of 26 ± 21 years, and 5,204 Radiographs were from female patients, with a mean age of 44 ± 29 years.

The CycleGAN model, which included a perceptual loss function and self-attention layer, performed the best overall, with an SSIM of 0.5445 ± 0.1001, correlation of 0.9863 ± 0.0196, intersection of 191822 ± 28426, Chi-squared distance of 1136451 ± 2486185 and a Hellinger distance of 0.2902 ± 0.0867. The Friedman test yielded p-values of zero for all metrics. It thus implied that a Post Hoc test in the form of a Wilcoxon signed-rank test with a Bonferroni adjustment was needed to establish differences between the six comparisons between models. Most differences between models were statistically significant at the $p = 0.050$ level (Bonferroni adjusted to $p < 0.009$) in Wilcoxon signed-rank tests. The SSIMs between the CycleGAN and CycleGAN-PL ($p = 0.066$), CycleGAN and CycleGAN-PL-SAL ($p = 0.113$), and CycleGAN-PL and CycleGAN-PL-SAL ($p = 0.028$) were not significantly different. The histogram correlations between CycleGAN and CycleGAN-SAL ($p = 0.010$) and the Hellinger distances between CycleGAN-PL and CycleGAN-PL-SAL ($p = 0.049$) were not significantly different. The quantitative results are shown in Table 1 and Figure 1.

As shown in Table 2, radiologists found that the AI-enhanced images improved diagnostic confidence in their reporting for 13 and 16 of the 20 cases. In 7 of the 20 cases, the radiologists wrote reports with different conclusions. In 4 and 3 of the 20 cases, the radiologists reported more decisively and identified a fracture that was only suspected in the original report. In 2 and 0 of the 20 cases, radiologists identified a fracture not mentioned in the original report. Since cases were selected because their original reports contained suggestions for repeat imaging, we performed a PACS search for repeat scanning. Follow-up imaging was performed in 11 of the 20 cases, although one case with a follow-up radiographs had a report that suggested further CT imaging. The 10 cases which had succesful follow-up imaging are shown in Table 3. Of these 10 cases, 7 were followed up with X-rays and 3 were followed up with CT scans. As in Table 2, this shows the breakdown of outcomes the reports can be categorised into, and the number of cases where AI-enhanced images improve diagnostic confidence. Almost every report from radiologists reading AI-enhanced images had identical diagnoses as those found in reports associated with follow-up radiographs and CT imaging. In the report that differed from follow-up imaging, the radiologist identified a “nondisplaced fracture of the proximal pole of the scaphoid bone”. Although both radiologists correctly diagnosed the fractured distal radius confirmed by the follow-up CT scan, the follow-up report states that “the scaphoid is normal”.

The Turing test experiment demonstrated that radiologists were typically unable to distinguish AI-generated images from unenhanced ones. As shown in Figure 2, the radiologists classified 65% of the AI-enhanced images as unenhanced, while they correctly classified 88% of the unenhanced images. For this experiment, the Kappa statistic was 0.79 and a Chi-squared value of 0.04 ($p=0.851$), showing that the radiologists’ assessments were not significantly different.

4. Discussion

In this study, we extended the method presented by Hržić et al. for suppressing cast artefacts in wrist radiographs by incorporating a perceptual loss function and a self-attention layer into the model. We investigated the effects of these additions on the model's performance through quantitative and qualitative experiments. Additionally, our model was trained and evaluated on a dataset that required fewer preprocessing steps and was not limited to paediatric images. We quantitatively assessed the model outputs using standard image processing metrics and found that the enhanced model—including the perceptual loss function and self-attention layer—performed best overall. Although the CycleGAN-PL-SAL model showed relative improvement over the standard CycleGAN model, the quantitative metrics did not show absolute improvement compared to those reported by Hržić et al. However, these metrics are intrinsically non-reproducible because we used different training and testing images. This inability to directly compare absolute values of quantitative metrics underscores a limitation in assessing models solely through these metrics.

The first qualitative test addressed a clinically relevant question and revealed that radiologists generally found that AI-enhanced images improved diagnostic confidence. In most cases, radiologists produced reports identical to the original ones, suggesting that the model does not generate significant artifacts or hallucinations. Notably, in some instances, radiologists provided decisive diagnoses of fractures that were only suspected in the original reports, and in two cases, they identified fractures that weren't initially reported. This implies that the model may be suppressing casts effectively to uncover underlying anatomy, demonstrating that thoroughly validated cast suppression could aid in diagnosis. One radiologist commented that the model sometimes made their task easier but never harder, expressing comfort in using the tool if it was well integrated with the PACS to reassure their diagnoses. This sentiment is reinforced by the finding that radiologists mostly produced reports with diagnoses identical to those from follow-up imaging. The single report that differed may be attributed to the inherently greater diagnostic capability of CT imaging due to its tomographic spatial information. Given that all other reports provided diagnoses matching follow-up imaging, this suggests the model is not producing hallucinations but is generating reliable images for diagnosis. Furthermore, the demonstrated ability of radiologists to make reliable diagnoses using downsampled images was particularly impressive.

The second subjective test in this study demonstrated that the model produces high-quality castless images, usually indistinguishable by radiologists from unenhanced images. Such realistic image generation could be leveraged for data augmentation, addressing data scarcity issues and class imbalances in datasets used to train other models. An objective validation approach could involve using the generated images as inputs for a well-established model to see if performance improves compared to using original images. For example, the model developed by Hembroff et al. can detect the presence of casts in wrist radiographs with high accuracy (43). This model could be tested with cast-suppressed images, objectively evaluating the quality of AI-generated images. These generated images could also serve as educational tools for radiology trainees and medical students.

This study is subject to various limitations relating to data, modeling, and performance evaluation methods, which present significant challenges for clinical implementation—particularly in establishing a robust validation method. While data availability is not an issue—it would not be difficult to collect a volume of data an order of magnitude larger—the prohibitive computational costs limit the number of images used in training. Training the model on data from multiple sources and institutions at native resolution would require significantly more computational resources. A major limitation of the clinical reporting assessment was using low-resolution images that could not be windowed or leveled, preventing the experiment from reflecting routine clinical conditions. Despite strong agreement between

radiologists, discrepancies in reporting were noted, and the small sample size and number of radiologists limited the ability to perform meaningful statistical tests. The diagnostic reference standard we used demonstrated that the model can reliably generate images of sufficient diagnostic quality, which has implications for reducing cumulative population dose, improving operational efficiency, reducing delays to diagnoses, reducing patient visits, and improving radiologist workflows. However, this validation was limited in terms of sample size. Despite these limitations, the model successfully generates high-quality cast-suppressed images. Examples of the model's successes and failures are shown in Figure 3.

With access to more resources, future work should focus on training the model on a greater volume of high-resolution images from various sources and validating the results more thoroughly. This would also involve identifying a greater number of cases with follow-up imaging and involving a larger cohort of radiologists in the assessment.

5. Conclusion

We extended a GAN-based method for suppressing cast artifacts in wrist trauma radiographs by incorporating a perceptual loss function and a self-attention layer into the model architecture. This enhancement led to relative improvements over the original method, as demonstrated by quantitative metrics on a large dataset that included adult and pediatric images. Qualitative assessments revealed that the model enhanced radiologists' diagnostic confidence and that radiologists were typically unable to distinguish AI-generated images from real ones. In some cases, the availability of both original and AI-generated images led radiologists to issue more decisive diagnoses for fractures that were previously only suspected, while also identifying new fractures that had not been mentioned in the original reports. The validation of a subset of reports against follow-up reporting demonstrates that the model can generate high-quality diagnostic images. Despite several limitations, this study lays the groundwork for developing a pipeline of AI models that can be built and eventually translated to the musculoskeletal radiology clinic. This tool has demonstrated potential for reducing patient doses by avoiding repeat imaging, improving operational efficiency, reducing delays to diagnosis, reducing patient visits, and facilitating radiologist workflows.

References

1. Delft EAKV, Gelder TGV, Vries RD, Vermeulen J, Bloemers FW. Duration of Cast Immobilization in Distal Radial Fractures: A Systematic Review. J Wrist Surg. 2019 Oct;08(05):430–8.

2. Hržić F, Žužić I, Tschauner S, Štajduhar I. Cast suppression in radiographs by generative adversarial networks. J Am Med Inform Assoc. 2021;28(12):2687–94.

3. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition [Internet]. 2017 [cited 2024 Oct 15]. p. 4681–90. Available from: http://openaccess.thecvf.com/content_cvpr_2017/html/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.html

4. Tang H, Liu H, Xu D, Torr PH, Sebe N. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Trans Neural Netw Learn Syst. 2021;34(4):1972–87.

5. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. Korean J Radiol. 2024;25(3):224.

6. Hong GS, Jang M, Kyung S, Cho K, Jeong J, Lee GY, et al. Overcoming the challenges in the development and implementation of artificial intelligence in radiology: a comprehensive review of solutions beyond supervised learning. Korean J Radiol. 2023;24(11):1061.

7. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision [Internet]. 2017 [cited 2024 Oct 15]. p. 2223–32. Available from: http://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html

8. Gravina M, Marrone S, Docimo L, Santini M, Fiorelli A, Parmeggiani D, et al. Leveraging CycleGAN in Lung CT Sinogram-free Kernel Conversion. In: Sclaroff S, Distante C, Leo M, Farinella GM, Tombari F, editors. Image Analysis and Processing – ICIAP 2022 [Internet]. Cham: Springer International Publishing; 2022 [cited 2024 Oct 15]. p. 100–10. (Lecture Notes in Computer Science; vol. 13231). Available from: https://link.springer.com/10.1007/978-3-031-06427-2_9

9. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans Med Imaging. 2017;36(12):2536–45.

10. Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. Phys Med Biol. 2019;64(12):125002.

11. Tang C, Li J, Wang L, Li Z, Jiang L, Cai A, et al. Unpaired Low-Dose CT Denoising Network Based on Cycle-Consistent Generative Adversarial Network with Prior Image Information. Comput Math Methods Med. 2019 Dec 7;2019:1–11.

12. Preetha CJ, Meredig H, Brugnara G, Mahmutoglu MA, Foltyn M, Isensee F, et al. Deep-learning-based synthesis of post-contrast T1-weighted MRI for tumour response assessment in neuro-oncology: a multicentre, retrospective cohort study. Lancet Digit Health. 2021;3(12):e784–94.

13. Yao Z, Luo T, Dong Y, Jia X, Deng Y, Wu G, et al. Virtual elastography ultrasound via generative adversarial network for breast cancer diagnosis. Nat Commun. 2023;14(1):788.

14. Lei Y, Harms J, Wang T, Liu Y, Shu H, Jani AB, et al. MRI□only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med Phys. 2019 Aug;46(8):3565–81.

15. Chuquicusma MJ, Hussein S, Burt J, Bagci U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) [Internet]. IEEE; 2018 [cited 2024 Oct 15]. p. 240–4. Available from: https://ieeexplore.ieee.org/abstract/document/8363564/

16. Bae K, Oh DY, Yun ID, Jeon KN. Bone suppression on chest radiographs for pulmonary nodule detection: comparison between a generative adversarial network and dual-energy subtraction. Korean J Radiol. 2022;23(1):139.

17. Conte GM, Weston AD, Vogelsang DC, Philbrick KA, Cai JC, Barbera M, et al. Generative Adversarial Networks to Synthesize Missing T1 and FLAIR MRI Sequences for Use in a Multisequence Brain Tumor Segmentation Model. Radiology. 2021 May;299(2):313–23.

18. Lei Y, Dong X, Tian Z, Liu Y, Tian S, Wang T, et al. CT prostate segmentation based on synthetic MRI□aided deep attention fully convolution network. Med Phys. 2020 Feb;47(2):530–40.

19. Chung M, Kong ST, Park B, Chung Y, Jung KH, Seo JB. Utilizing Synthetic Nodules for Improving Nodule Detection in Chest Radiographs. J Digit Imaging. 2022 Aug;35(4):1061–8.

20. Al Khalil Y, Amirrajab S, Lorenz C, Weese J, Pluim J, Breeuwer M. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. Med Image Anal. 2023;84:102688.

21. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci Rep. 2019;9(1):16884.

22. Bargshady G, Zhou X, Barua PD, Gururajan R, Li Y, Acharya UR. Application of CycleGAN and transfer learning techniques for automated detection of COVID-19 using X-ray images. Pattern Recognit Lett. 2022;153:67–74.

23. Tmenova O, Martin R, Duong L. CycleGAN for style transfer in X-ray angiography. Int J Comput Assist Radiol Surg. 2019 Oct;14(10):1785–94.

24. Nakanishi N, Otake Y, Hiasa Y, Gu Y, Uemura K, Takao M, et al. Decomposition of musculoskeletal structures from radiographs using an improved CycleGAN framework. Sci Rep. 2023;13(1):8482.

25. Parsania PS, Virparia PV. A comparative analysis of image interpolation algorithms. Int J Adv Res Comput Commun Eng. 2016;5(1):29–34.

26. Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision – ECCV 2016 [Internet]. Cham: Springer International Publishing; 2016 [cited 2024 Oct 15]. p. 694–711. (Lecture Notes in Computer Science; vol. 9906). Available from: http://link.springer.com/10.1007/978-3-319-46475-6_43

27. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015 Dec;115(3):211–52.

28. Alami Mejjati Y, Richardt C, Tompkin J, Cosker D, Kim KI. Unsupervised attention-guided image-to-image translation. Adv Neural Inf Process Syst [Internet]. 2018 [cited 2024 Oct 15];31. Available from: https://proceedings.neurips.cc/paper/2018/hash/4e87337f366f72daa424dae11df0538c-Abstract.html

29. Marín-Reyes PA, Lorenzo-Navarro J, Castrillón-Santana M. Comparative study of histogram distance measures for re-identification [Internet]. arXiv; 2016 [cited 2024 Oct 15]. Available from: http://arxiv.org/abs/1611.08134

30. Jia W, Zhang H, He X, Wu Q. A comparison on histogram based image matching methods. In: 2006 IEEE International Conference on Video and Signal Based Surveillance [Internet]. IEEE; 2006 [cited 2024 Oct 15]. p. 97–97. Available from: https://ieeexplore.ieee.org/abstract/document/4020756/

31. de Lima JR, Boff FA, de Souza Jaccoud Filho D, Falate R. HISTOGRAM COMPARISON USING INTERSECTION METRIC APLLIED TO DIGITAL IMAGES ANALYSIS. Iberoam J Appl Comput [Internet]. 2012 [cited 2024 Oct 15];2(1). Available from: https://revistas.uepg.br/index.php/ijac/article/view/4066

32. Gagunashvili ND. Chi-square tests for comparing weighted histograms. Nucl Instrum Methods Phys Res Sect Accel Spectrometers Detect Assoc Equip. 2010;614(2):287–96.

33. Le Cam LM, Yang GL. Asymptotics in statistics: some basic concepts [Internet]. Springer Science & Business Media; 2000 [cited 2024 Oct 15]. Available from: https://books.google.com.au/books?hl=en&lr=&id=wpIrbruQJK4C&oi=fnd&pg=PR5&dq=Asymptotics+in+statistics:+some+basic+concepts&ots=BrGiXtogO5&sig=8338OXqnsD-ymmFxtSoZDD-Ko7U

34. Wang Z, Bovik AC. Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Process Mag. 2009;26(1):98–117.

35. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600–12.

36. Precht H, Hansson J, Outzen C, Hogg P, Tingberg A. Radiographers' perspectives' on Visual Grading Analysis as a scientific method to evaluate image quality. Radiography. 2019;25:S14–8.

37. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611.

38. Friedman M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. J Am Stat Assoc. 1937 Dec;32(200):675–701.

39. Wilcoxon F. Individual Comparisons by Ranking Methods. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics [Internet]. New York, NY: Springer New York; 1992 [cited 2024 Oct 15]. p. 196–202. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-1-4612-4380-9_16

40. Dunn OJ. Multiple Comparisons among Means. J Am Stat Assoc. 1961 Mar;56(293):52–64.

41. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012;22(3):276–82.

42. Greenwood PE, Nikulin MS. A guide to chi-squared testing [Internet]. Vol. 280. John Wiley & Sons; 1996 [cited 2024 Oct 15]. Available from: https://books.google.com.au/books?hl=en&lr=&id=bc8zfQSKOwIC&oi=fnd&pg=PR7&dq=A+guide+to+chi-squared+testing&ots=J_qPdy5aa4&sig=lKN4IalwUAQldrvINUSxf4H3CDg

43. Hembroff G, Klochko C, Craig J, Changarnkothapeecherikkal H, Loi RQ. Improved Automated Quality Control of Skeletal Wrist Radiographs Using Deep Multitask Learning. J Imaging Inform Med [Internet]. 2024 Aug 26 [cited 2024 Oct 15]; Available from: https://link.springer.com/10.1007/s10278-024-01220-9

Tables

Table 1: Quantitative comparison metrics with associated standard deviations for comparisons of the generated cast suppressed histograms against the reference castless test data set image histograms. The best results are given in bold.

Metric	CycleGAN	CycleGAN + PL	CycleGAN + SAL	CycleGAN + PL + SAL
SSIM	0.54 ± 0.10	0.54 ± 0.10	**0.55 ± 0.10**	0.54 ± 0.10
Histogram Correlation	0.99 ± 0.02	0.99 ± 0.02	0.99 ± 0.02	**0.99 ± 0.02**
Histogram Intersection	190 986 ± 285 926	**191 832 ± 28 590**	191 530± 28 778	191 822 ± 28 426
Chi-squared distance	1 325 700 ± 2 810 422	1 196 688 ± 2 411 085	1 229 311 ± 3 284 563	**1 136 451 ± 2 486 185**
Hellinger distance	0.30 ± 0.09	0.29 ± 0.09	0.29 ± 0.09	**0.29 ± 0.09**

Table 2: Summary results of the clinical application experiment, where diagnoses from each of 20 cases are placed into one of three categories. The far-right column indicates how many of these 20 images improve diagnostic confidence.

	Identical Report	More decisive report	New fracture identified	Improvement in diagnostic confidence
Radiologist 1	14 (70%)	4 (20%)	2 (10%)	13 (65%)
Radiologist 2	17 (85%)	3 (15%)	0 (0%)	16 (85%)

Table 3: Validation of the clinical application experiment, where follow-up imaging reports from 10 cases are compared with those written by radiologists reading AI-enhanced images. Note, these are a subset of Table 2.

	Identical Report	More decisive report	New fracture identified	Improvement in diagnostic confidence
Radiologist 1	7 (70%)	3 (30%)	0 (0%)	6 (60%)
Radiologist 2	9 (90%)	1 (10%)	0 (0%)	8 (80%)

Figures

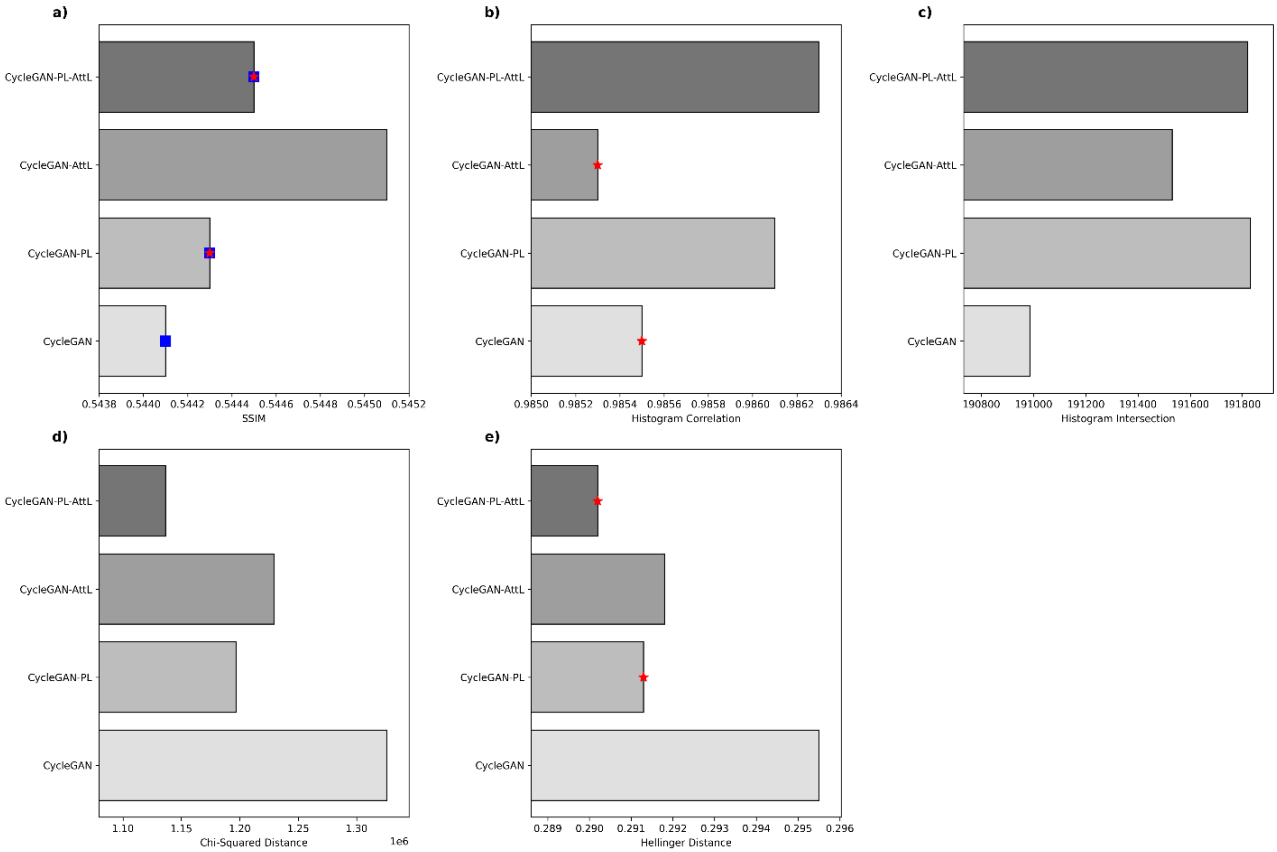

Figure 1: Bar charts displaying the quantitative metrics: a) SSIM, b) Histogram correlation and c) intersection, d) Chi-squared distance, and e) Hellinger distance. All differences between models were statistically significant at the Bonferroni-adjusted $p < 0.009$ level, apart from the 5 cases indicated by the blue squares and red stars (p-values noted in main text). Here, the bar plot labels with extensions -PL and -AttL refer to the models which include a perceptual loss function and self-attention-layer, respectively.

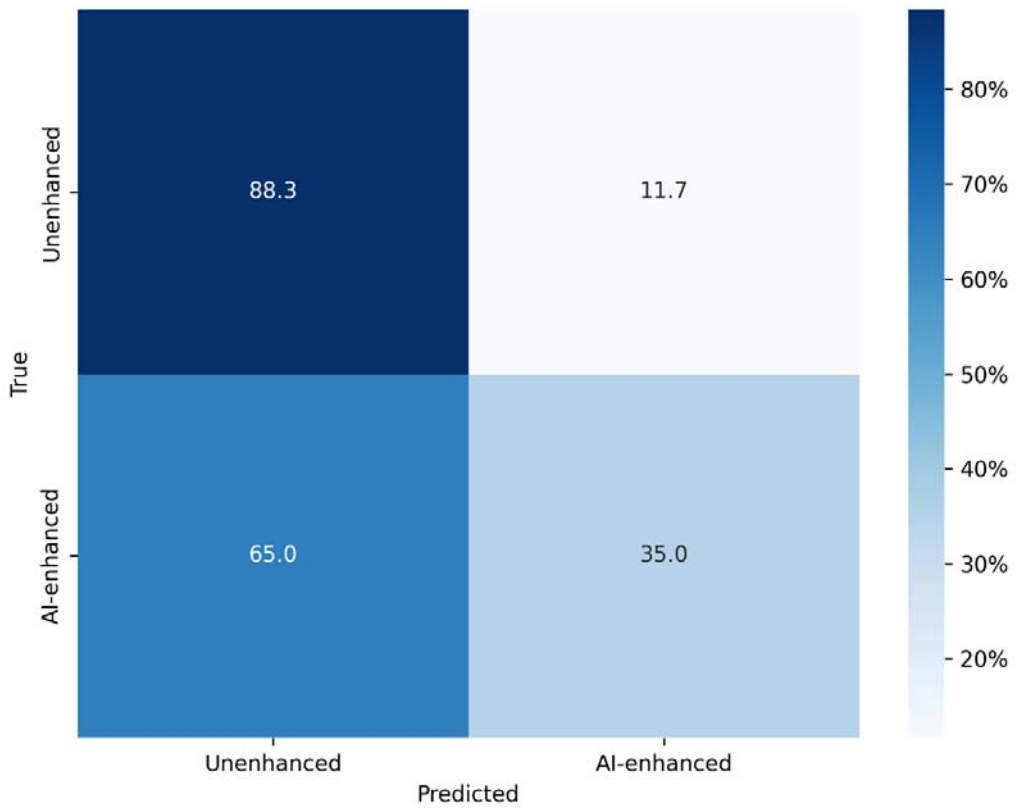

Figure 2: Confusion matrix of Turing Test results. The y-axis represents the nature of the displayed image and the x-axis represents the percentage of images radiologists classified into each category. Both radiologists evaluated 60 images each, giving a total of 120 observations.

Figure 3: Collage showing several input images sampled from the test data set (left) and the associated outputs generated by the CycleGAN model (right).