Computerized Medical Imaging and Graphics 123 (2025) 102565

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

CALIMAR-GAN: An unpaired mask-guided attention network for metal artifact reduction in CT scans

Roberto Maria Scardigno, Antonio Brunetti, Pietro Maria Marvulli, Raffaele Carli, Mariagrazia Dotoli, Vitoantonio Bevilacqua *, Domenico Buongiorno

Department of Electrical and Information Engineering (DEI), Polytechnic University of Bari, Bari, 70126, Italy

ARTICLE INFO

Dataset link: http://spineweb.digitalimaginggroup.ca, https://github.com/roberto722/calimar-gan

Keywords:
Computed tomography
Metal artifact reduction
Generative adversarial network
Unpaired data

ABSTRACT

High-quality computed tomography (CT) scans are essential for accurate diagnostic and therapeutic decisions, but the presence of metal objects within the body can produce distortions that lower image quality. Deep learning (DL) approaches using image-to-image translation for metal artifact reduction (MAR) show promise over traditional methods but often introduce secondary artifacts. Additionally, most rely on paired simulated data due to limited availability of real paired clinical data, restricting evaluation on clinical scans to qualitative analysis. This work presents CALIMAR-GAN, a generative adversarial network (GAN) model that employs a guided attention mechanism and the linear interpolation algorithm to reduce artifacts using unpaired simulated and clinical data for targeted artifact reduction. Quantitative evaluations on simulated images demonstrated superior performance, achieving a PSNR of 31.7, SSIM of 0.877, and Fréchet inception distance (FID) of 22.1, outperforming state-of-the-art methods. On real clinical images, CALIMAR-GAN achieved the lowest FID (32.7), validated as a valuable complement to qualitative assessments through correlation with pixel-based metrics ($r = -0.797$ with PSNR, $p < 0.01$; $r = -0.767$ with MS-SSIM, $p < 0.01$). This work advances DL-based artifact reduction into clinical practice with high-fidelity reconstructions that enhance diagnostic accuracy and therapeutic outcomes. Code is available at https://github.com/roberto722/calimar-gan.

1. Introduction

Computed tomography (CT) has significantly advanced in resolution over the years, enhancing its ability to diagnose and perform accurate estimations for several medical tasks, such as neoplasm detection and characterization, as well as surgery and radiotherapy planning. However, the presence of metals within the body (i.e., prosthesis, screws, dental fillings, etc.) can create different kinds of artifacts (e.g., photon starvation, beam hardening, and scatter) that may lead to misinterpretation of CT scans (Gjesteby et al., 2016).

Various solutions for metal artifact reduction (MAR) have been proposed over the years (Selles et al., 2024; Boas and Fleischmann, 2012). These include, for instance, high-energy CT scans to improve the penetration depth or excluding metal objects from the CT scan's field of view. Since hardware/setup solutions are not always feasible (e.g., high energy scans increase absorbed radiation dose), MAR algorithms that process the acquired images have been developed for decades. Traditional approaches primarily operate within the sinogram domain (Kalender et al., 1987; Meyer et al., 2010, 2012), such as the linear interpolation (LI) algorithm. In contrast, standard image-based approaches have been less explored (Soltanian-Zadeh et al., 1996; Bal et al., 2005; Bevilacqua et al., 2007).

The success of deep learning (DL) in signal processing and image analysis has led to its application in MAR. Three main approaches can be identified according to the processed data domain they process: (i) sinogram-based methods, (ii) image-based methods, (iii) multi-domain-based methods. Sinogram-based methods (Yu et al., 2021; Zhu et al., 2021b; Peng et al., 2020a,b; Park et al., 2018) focus on identifying metal regions and restoring anatomical structures with a content that is consistent with the neighboring information. Image-based methods (Xu et al., 2023; Niu et al., 2022; Wang et al., 2022a; Shi et al., 2022; Kim et al., 2022; Ikuta and Zhang, 2022; Wang et al., 2021b; Nakao et al., 2020; Koike et al., 2020; Liao et al., 2020; Gjesteby et al., 2019; Zhu et al., 2019; Zhang and Yu, 2018; Huang et al., 2018), which are the most common, reduce metal artifacts by processing only the CT slices. Additionally, multi-domain (or hybrid) methods leverage information from both sinogram and image domains simultaneously (Lee et al., 2020; Yu et al., 2021b; Ketcha et al., 2021; Wang et al., 2021c, 2022b), combining data from various sources to enhance artifact reduction.

* Corresponding author.
E-mail address: vitoantonio.bevilacqua@poliba.it (V. Bevilacqua).

https://doi.org/10.1016/j.compmedimag.2025.102565
Received 20 February 2025; Received in revised form 16 April 2025; Accepted 25 April 2025
Available online 9 May 2025

Sinogram-based and hybrid techniques are trained using a supervised learning paradigm that requires image pairs: both the artifact-affected and artifact-free versions of the same CT scan, eventually along with their corresponding sinograms. This learning approach, commonly used in the so-called paired networks, is intuitive and straightforward. Nevertheless, obtaining paired images in a clinical setting is complex and often impractical, as a result, paired methods are usually trained on simulated data with virtually introduced artifacts. Furthermore, the artifact simulation processes — especially those using unrealistic or simplistic metal masks, may not fully capture the physics of artifact generation, leading to discrepancies between actual and simulated data (Liao et al., 2020). The impact of such training on clinical data (i.e., with real artifacts) is difficult to assess quantitatively and, to date, has not been sufficiently investigated. Additionally, the use of sinograms can introduce secondary artifacts due to the domain translation (i.e., forward- or back-projection) (Pan et al., 2009). This issue has been deeply explored, for instance, Arabi and Zaidi (2021) proposes an effective pipeline working either on images or sinograms. The adoption of image-based solutions offers a valid solution to the issue. In particular, the image-to-image (I2I) translation paradigm constitutes a suitable option, as it avoids projections, thus reducing the risk of corrupting the overall image quality.

The I2I translation is a widely explored technique for various problems, including denoising (Jiménez-Gaona et al., 2024; Li et al., 2022b; An et al., 2022; Gajera et al., 2021) and inpainting (Li et al., 2022a; Armanious et al., 2020). In the context of MAR, most image-based methods are fully supervised, with only limited exploration of unsupervised or semi-supervised strategies (Niu et al., 2022; Shi et al., 2022; Nakao et al., 2020; Koike et al., 2020; Liao et al., 2020). Unsupervised methods aim at addressing the domain gap between artifact-affected and artifact-free images by focusing on distinguishing features between different image distributions, rather than on the differences between paired images. However, I2I unsupervised methods face critical issues such as unintended alteration of image characteristics beyond the translation process, as well as changes in the background. To address these limitations, some approaches have been developed: Liang et al. (2017) introduces a novel approach based on an adversarial distance comparison objective for optimizing generators and discriminators, within a generative adversarial network (GAN) architecture, which helps to separate the image background from the foreground; Tang et al. (2021) proposes a CycleGAN with an attention mechanism to improve the entire cycle consistency during the forward and backward process, mitigating issues related to small receptive fields of convolution operators that may struggle global geometric or structural patterns (Gu et al., 2022).

The state-of-the-art unsupervised I2I translation solutions are mainly trained on images with simulated artifacts, using a limited set of metal masks. These methods often provide only qualitative results on real artifact images and are evaluated quantitatively on a few hundred simulated images, which may not represent the full range of real-world samples. Furthermore, assessing the impact of the artifact simulation process on the images with real artifacts is challenging. To the best of our knowledge, among all the I2I translation methods, the only work evaluating performance on real artifact data is by Xu et al. (2023), which uses different distribution-based metrics such as the sliced Wasserstein distance, the inception score, and the Fréchet inception distance (FID). Although these metrics have shown good correlation with human visual perception in different domains, their application in the medical context is controversial. These metrics are known to be biased for two main reasons: the features extractor is trained on natural images, and a minimum number of around 2000 images per distribution is required (Bischoff et al., 2024).

This work introduces CALIMAR-GAN, which uniquely integrates a mask-guided attention mechanism with a cycle-consistent GAN framework, leveraging sinogram-derived features, unpaired data, and combining simulated and real artifacts to achieve superior MAR in CT scans. By focusing domain translation on a highlighted image region, the network reduces hallucinations and blurriness compared to its base version AttentionGAN-v2 (Tang et al., 2021) and previous unsupervised methods, thereby improving output reliability — a critical aspect for clinical utility. CALIMAR-GAN is designed to address the specific demands and constraints inherent to the MAR task, particularly those associated with deep learning:

- using unpaired data can result in unintended alterations to image characteristics, such as changes in areas unaffected by the artifacts;
- relying solely on simulated artifacts can lead to a misrepresentation of the true physical properties of real artifacts;
- relying only on real clinical data may not be enough for adequately training an unsupervised network due to the lack of data.

The proposed architecture processes input data within the image domain, while features computed in the sinogram domain are used only as supplementary information once they are translated back to the image space. Furthermore, the architecture includes a Siamese network that leverages mask-like inputs to focus the translation process on the artifact (i.e., the foreground) rather than the background of the CT slice. Based on the previous limitations about the performance evaluation and the widely known data availability problems that characterize the medical scenery, the network outputs are extensively analyzed using a large simulated dataset and quantitatively assessed even on a sufficient amount of data with real artifacts. This study also aims at identifying a reliable metric to evaluate the network's performance on clinical data with real artifacts from a quantitative point of view. In particular, the potential of the FID for evaluating MAR performance on real images is explored by examining its relationship with common pixel-based metrics that detect artifacts.

The main contributions of this paper are summarized below:

- An unsupervised cycle-consistent adversarial network is proposed to effectively mitigate various types of artifacts while preserving the integrity of unaffected regions. CALIMAR-GAN employs a novel mask-guided attention approach that leverages artifact severity by feeding the non-binary difference between the corrupted CT scan and the output of a traditional MAR algorithm (i.e., LI) to the attention module. This solution helps the network focus on the most corrupted CT areas. To the best of our knowledge, this is the first study to apply the mask-guided attention approach to the MAR field.
- Given the challenges of limited data availability in medical contexts, different training strategies are examined to identify the most effective one for developing a model with strong generalization capabilities despite limited real-world data. In particular, the following three training sets are considered: (i) fully simulated artifacts data, and (ii) only real artifacts data, (iii) a mixed dataset (simulated + real artifacts data).
- A comprehensive comparison of our solution with different state-of-the-art models, both paired and unpaired, has shown the superiority of our method when processing images with both simulated and real artifacts. In addition, all the major previous works that analyze real corrupted CT scans present only qualitative evaluation results. To the best of our knowledge, this is the first work to investigate whether FID can provide meaningful insights into the effectiveness of the metal artifact reduction process in images with real artifacts. Our results show a strong correlation between the FID and other traditional pixel-based metrics on simulated dataset, indicating that it could be a robust and valuable metric for evaluating artifacts in real-world contexts.

In the authors' opinion, such contributions represent an important step forward in the clinical utility of DL-based MAR algorithms. Suboptimal MAR can have significant negative effects on diagnostic accuracy

due to distortion of anatomical structures and masked key details. As an example, in the oncology context, artifact residue can lead to obstruction of the tumor borders, leading to inaccurate tumor delineation in radiotherapy treatment planning, which could then lead to suboptimal radiation dose delivery and increase the risk of damaging healthy tissues surrounding the tumor. Similarly, in planning for surgeries, image quality can be altered in critical structures as a result of artifacts, increasing the chances of complications such as incompletely resetting a tumor or injuring adjacent organs. CALIMAR-GAN addresses these issues by selectively targeting the domain translation of the network to the areas of the image most affected by artifacts. This ultimately leads to significantly less hallucinations and blurriness, and ensures more anatomically faithful reconstructions, improving the reliability of diagnoses or therapy planning.

The remainder of the paper is organized as follows. Section 2 presents the proposed CALIMAR-GAN model. Section 3 illustrates the experimental setup, the dataset, and the compared methods. Results and discussion are presented in Section 4. Finally, Section 5 presents conclusions and future research directions.

2. Methods

In this section, we describe the architecture of our proposed solution, named CALIMAR-GAN, that is a cyclic-consistent GAN architecture based on mask-guided attention and the linear interpolation algorithm. The architecture processes CT scan images, while features coming from the sinogram domain are used as additional information once they are translated back to the image space. The proposed architecture employs a Siamese network, which considers a mask-like input to direct the I2I translation process towards the artifact. Starting from the Tang et al. (2021) original work, namely AttentionGAN-v2, the guided attention mechanism employed in CALIMAR-GAN has been optimized/specialized for the MAR task.

The subsections that follow will describe:

- the employed cycle-consistent GAN with details about the losses used in this work;
- the basis of the technique used to generate images involving the attention that is inherited from AttentionGAN-v2;
- the proposed mask-guided attention approach.

2.1. Unsupervised generation using cycle-consistent GAN

Unpaired I2I translation tasks, given the image domain X, the image domain Y, and the respective training images $x_i \in X$; $y_j \in Y$, involve the estimation of a mapping $F_{X\to Y}$ that can translate images from the first domain (i.e., the metal-affected one) to the second domain (i.e., the artifact reduced one).

Unlike the paired translation, where a direct mapping can be performed thanks to image pair availability, the unpaired translation focuses on estimating the image distribution probabilities P_X and P_Y to obtain a general mapping function such that $F_{X\to Y}(x_i)$ fall within the probability distribution of the target P_Y. The cycle consistency, introduced in CycleGAN (Zhu et al., 2020), represents one of the best-performing methods that also focuses on the reverse mapping $F_{Y\to X}(y_j)$, which has been proved to enhance the overall mapping process. The original framework consists of a generator G_θ, whose objective, given a sample x_i, is to perform the mapping $F_{X\to Y}$, and a discriminator D_θ that aims to understand if the image belongs to the probability distribution P_Y or not. In the same way, the reverse mapping $F_{Y\to X}$ is obtained by a generator G_ϕ that tries to fool the discriminator D_ϕ. The training process starts with two unpaired images from their respective distributions. The generators are trained separately and optimized through an adversarial loss which fosters a competitive dynamic between the two networks: the generator and the discriminator. In this work, the adversarial loss introduced by Mao et al. (2017) is adopted; then the loss of $[G_\theta, D_\theta]$ is given by:

$$\begin{cases} L_{G_\theta} = \frac{1}{2}\mathbb{E}_{y\sim P_Y}[(D_\theta(G_\theta(y)) - 1)^2] \\ L_{D_\theta} = \frac{1}{2}\mathbb{E}_{x\sim P_X}[(D_\theta(x) - 1)^2] \\ \quad + \frac{1}{2}\mathbb{E}_{y\sim P_Y}[D_\theta(G_\theta(y))], \end{cases} \tag{1}$$

where the generator G_θ tries to minimize L_{G_θ} while D_θ tries to maximize it. Correspondingly, $[G_\phi, D_\phi]$ are trained accordingly to:

$$\begin{cases} L_{G_\phi} = \frac{1}{2}\mathbb{E}_{x\sim P_X}[(D_\phi(G_\phi(x)) - 1)^2] \\ L_{D_\phi} = \frac{1}{2}\mathbb{E}_{y\sim P_Y}[(D_\phi(y) - 1)^2] \\ \quad + \frac{1}{2}\mathbb{E}_{x\sim P_X}[D_\phi(G_\phi(x))]. \end{cases} \tag{2}$$

To take advantage of the cycle consistency, the whole translation cycle is used such that x is primarily passed through the G_θ, then the result y has to be revamped back to its initial input, i.e., $x \to G_\theta(x) \to G_\phi(G_\theta(x)) \approx x$. The cycle consistency loss can be then expressed as:

$$\begin{aligned} L_{cycle}(G_\theta, G_\phi) = {} & \mathbb{E}_{x\sim P_X}[\| G_\phi(G_\theta(x)) - x \|_1] \\ & + \mathbb{E}_{y\sim P_Y}[\| G_\theta(G_\phi(y)) - y \|_1]. \end{aligned} \tag{3}$$

A key challenge affecting GANs' performance is their tendency to focus on unwanted parts of images, leading to undesired changes in generated content (Tang et al., 2021; Nobari et al., 2021). One of the earliest methods developed to address this issue is the introduction of a loss term L_{id} that takes into account the pixel differences of the generated output with respect to the input, as demonstrated by Li et al. (2018). However, even if the original work calculated the loss on VGG high-level features, our implementation:

$$\begin{aligned} L_{id}(G_\phi, G_\theta) = {} & \mathbb{E}_{x\sim P_X}[\| G_\theta(x) - x \|_1] \\ & + \mathbb{E}_{y\sim P_Y}[\| G_\phi(y) - y \|_1] \end{aligned} \tag{4}$$

makes use of a more explicit pixel difference, as adopted by different works (Wang and Makarenko, 2021; Zhang et al., 2018). The total loss used during our training can be formulated as:

$$\begin{cases} L_G = L_{G_\theta} + L_{G_\phi} + \lambda_{cycle} L_{cycle}(G_\theta, G_\phi) \\ \quad + \lambda_{id} L_{id}(G_\phi, G_\theta) \\ L_D = L_{D_\theta} + L_{D_\phi} \end{cases} \tag{5}$$

where λ_{cycle} and λ_{id} are tunable parameters that weights $L_{cycle}(G_\theta, G_\phi)$ and $L_{id}(G_\phi, G_\theta)$, respectively. Fig. 1 shows the overall cycle and displays how the various loss components come into play. The generators used to perform the cycle are different among them. In fact, while G_ϕ employ a slightly modified version of AttentionGAN-v2 (Tang et al., 2021), G_θ has been deeply optimized to employ a guided attention mask and exploit the information coming from the linear interpolation algorithm to address the specific MAR task.

2.2. Generation involving attention

The generator G_ϕ, shown in Fig. 2(b), is adopted from AttentionGAN-v2 and slightly modified to enhance its performance within the MAR task. Its main objective is to close the cycle providing a model that is able to impress an artifact over the cleaned image after generating a metal mask. Here, the attention mechanism is adopted to specify which zones the GAN should modify to correctly generate the metal and simulate the artifacts. In particular, the metal and the artifact generation are simultaneous. The network's concept is based on the generation of content masks, which aim to produce a modification of the original image, and attention masks, whose aim is to limit the GAN editing to only relevant zones, for example in proximity of the bones. While the encoder is in common between the

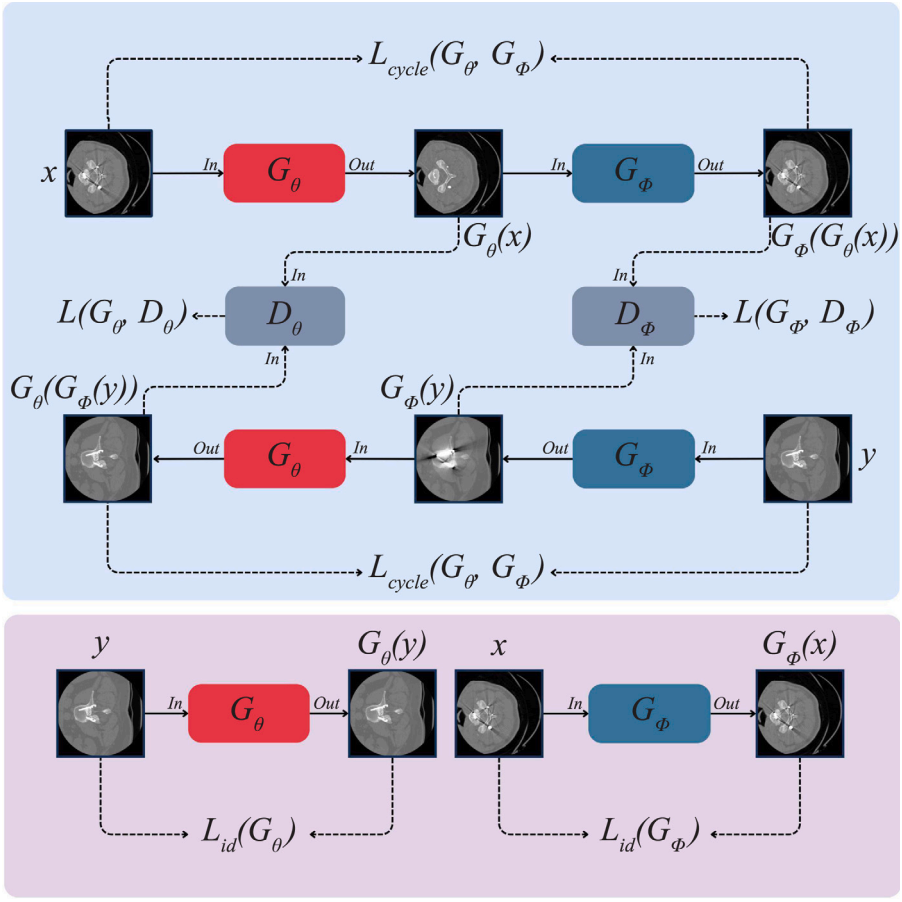

Fig. 1. Diagrams of the proposed cycle-consistent network showing the utilized losses. (A) The diagram represents the cyclic structure used in CALIMAR-GAN where L, and L_{cycle} indicate the adversarial loss between generator and discriminator, and the cycle consistency loss, respectively. (B) The diagram illustrates the identity loss L_{id} obtained from the two generators G_θ and G_ϕ. The black solid lines illustrate the network's training forward process, whereas the black dashed lines indicate the flow used for calculating the losses. x belongs to the distribution of images with artifact, while y belongs to the distribution of artifact-free images. G_θ/D_θ and G_ϕ/D_ϕ are the generators/discriminators designated for the artifact removal and creation processes, respectively. In are the images that the generators/discriminators take as input and Out are their results. The discriminators' output is a boolean value used for the loss computation.

content masks and the attention masks generators, the decoders are split. The Content-mask decoder is built to output 9 three-channel content masks $C_y^f|_{f=1}^9 \in \mathbb{R}^{H\times W\times 3}$, since one single mask could not be enough to produce a complex artifact due to the different effects that multiple metals can yield over the image. The Attention-mask decoder is constructed to output: 9 foreground attention masks $A_y^f|_{f=1}^9$ and 1 background attention mask A_y^b, where $A_y^{f,b} \in \mathbb{R}^{H\times W}$. In particular, since we use this generator to apply the artifact over a clean image, the background attention mask A_y^b will be pixel-wise multiplied with the initial input y.

The Content-mask decoder is preceded by a residual structure (He et al., 2015), where only 5 residual blocks are concatenated following Johnson et al. (2016), rather than the 9 used in the original work. This choice allows for fewer hallucinations and reduces the size of the network. Instead, the Attention mask receives its input directly from the latent space of the Siamese encoder, whereas in the original version of Attention-GAN v2, the input was the same as that of the Content-mask decoder (green dashed line in Fig. 2). In so doing, the attention-mask decoder can work on raw features where spatial information is preserved.

The output of G_ϕ is formulated as:

$$G_\phi(y) = \sum_{f=1}^{9}(C_y^f \cdot A_y^f) + y \cdot A_y^b. \tag{6}$$

2.3. Mask-guided attention for knowledge enhancement

Despite the baseline strategies adopted in AttentionGAN-v2, such as employing L_{id} loss and attention mechanisms to direct the focus of the network on relevant image zones, the artifact reduction may not be fully accomplished due to additional unwanted alterations of the image (Skandarani et al., 2023; Zhu et al., 2019a). The problem is enhanced by the limited data availability that can lead the network to misunderstand the actual differences between the image distributions X and Y. Considering how the position, shape, and material of one or more metals affect the CT slice with different artifacts, we can derive that to successfully train a GAN there will be the need for a huge amount of data to let the network understand what an artifact is (i.e., we got multiple kinds of artifacts at once), what is produced by, and which parts are the background, i.e., body, skin, organs, bone, etc... Then, we need a way to disentangle this information with a minimal amount of data, otherwise, we could fall into different problems such as change of unwanted background, limited generalizability, and the undetection of the artifacts. Thus, we propose a guided attention mechanism to obtain a network that is capable of distinguish anatomical parts from artifacts, that is generalizable, and trainable with a limited number of slices.

The number of images is an essential part of a GAN training process. Within the medical context, the limited availability of images leads to poor performance of these networks in several tasks. The question is: Does the GAN have a sufficient number of samples to disentangle

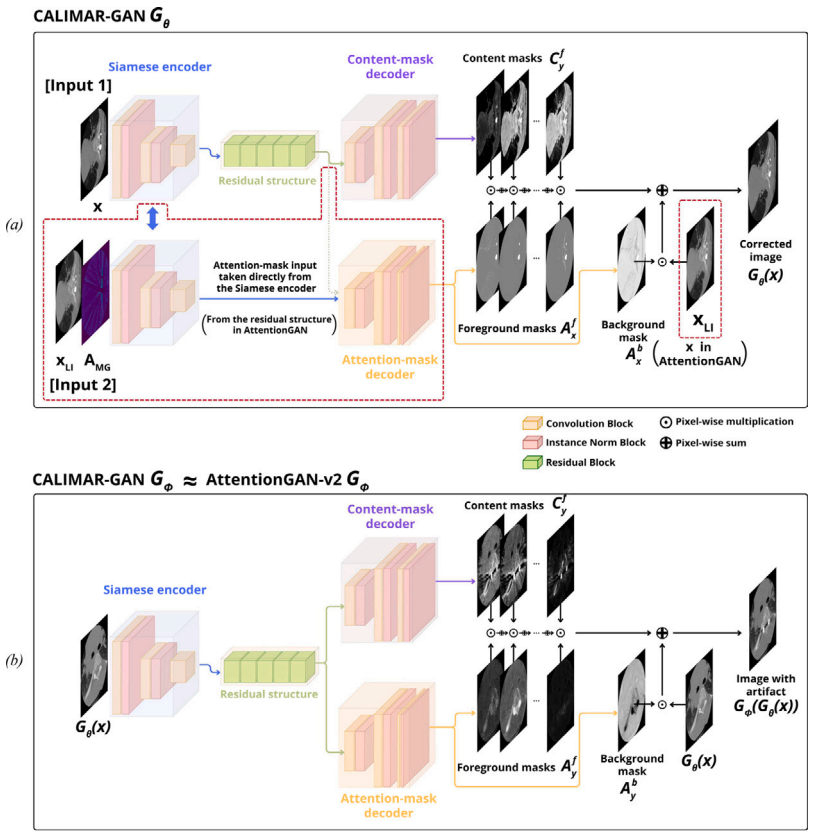

Fig. 2. *(a)* Architecture of our proposed G_θ, where red squares represent the macroscopic architectural differences from the original version proposed in AttentionGAN-v2 (Tang et al., 2021). x refers to the image with artifact from the domain X, A_y^b and A_x^b are the background attention masks, while A_y^f and A_x^f represent the foreground attention masks issued by the Attention-mask decoder; C_y^f and C_x^f summarize the content masks resulting from the Content-mask decoder; *(b)* Architecture of G_ϕ, similar to the one proposed in AttentionGAN-v2, with a residual structure of 5 blocks.

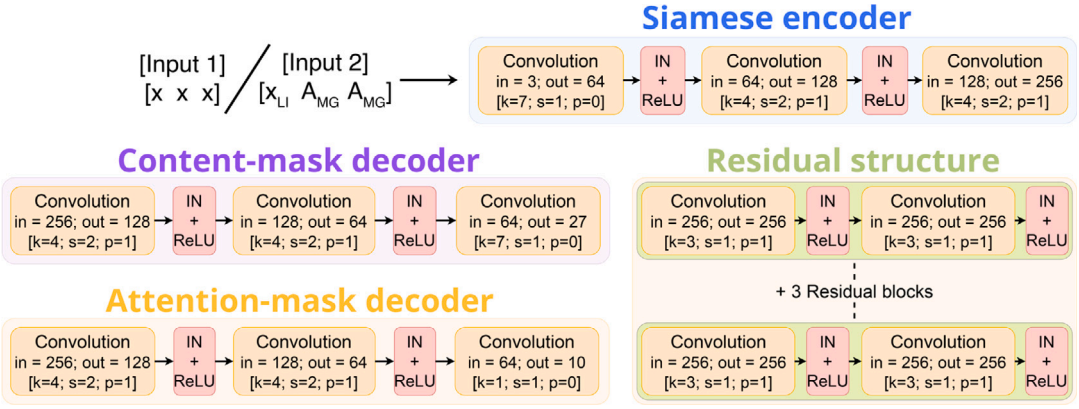

Fig. 3. Detailed scheme of the single components of CALIMAR-GAN. IN refers to the Instance Normalization layer, k, s, and p represent the kernel size, the stride, and the padding, respectively. The Siamese encoder accepts as input a tensor of 3 layers, i.e., Input 1 or Input 2.

the anatomical features from the artifact ones? It is hard to give an exact number, knowing the complexity of the human body (Karras et al., 2020). To overcome this question, we propose the adoption of a different strategy that somehow guides the network in an accelerated understanding of what an artifact is, how to remove it, and, starting from a raw result, how to improve it.

Several works, also called *mask-guided* (Luo et al., 2020), tried to exploit a mask that helps the network to produce better results since more discriminating features can be obtained by disentangling the background from the foreground (Wang et al., 2023, 2021a; Liu et al., 2023; Luo et al., 2024; Luo and Huang, 2024; Song et al., 2018; Gu et al., 2019). In the MAR context, the foreground refers to any portion of the image that has been impacted by an artifact, whereas the background represents the remaining unaffected area. Moreover, the effect of attention masks on adversarial training has also been explored: the findings demonstrated that their usage is beneficial for guiding models to focus on the region of interest (Vaishnavi et al., 2019).

The mask-guided attention approach proposed in this work consists in providing both the corrupted CT scan elaborated by the linear interpolation algorithm and the difference between the LI output and the corrupted CT as input to the attention module. As a first step, the linear interpolation algorithm based on the work of Kalender et al. (1987) is applied to the corrupted CT scan x to obtain a CT scan x_{LI}

characterized by a coarse reduction of the artifact. In detail, the extraction of the metal trace within the sinogram, which is a preliminary step for the LI, has been performed by following the procedure presented by Yu et al. (2021b), Liao et al. (2020) with a threshold value equal to 2500 Hounsfield Units (HU), since the highest HU values for biological tissues are usually associated with cortical bone, which typically ranges from +500 to +1900 HU (Chen et al., 2024; Zou et al., 2019). As a second step, the second mask-guided attention input A_{MG} is computed as the difference between x_{LI} and x following Eq. (7).

$$A_{MG} = \begin{cases} x_{LI} - x & \text{for } x_{LI} - x > 0 \\ 0 & \text{for } x_{LI} - x \leq 0 \end{cases} \tag{7}$$

Then, A_{MG}, once normalized between $[0, 1]$, is a mask containing proportional values that give the network information about not only where the artifact is located but also even the severity of the artifact. However, since A_{MG} may not entirely capture the distortions that metals produce over the image, the mask-guidance is intended to assist the attention mechanism, introduced in Section 2.2, without altering the number of parameters to train. Finally, the proposed framework takes as input three different images: x, x_{LI}, and A_{MG}.

The proposed G_θ, shown in Fig. 2(a), is then composed of a Siamese encoder, which allows the adoption of three different inputs, and two decoders. These latter works as in G_ϕ: the content masks decoder output 9 three-channels content masks $C_x^f|_{f=1}^9 \in \mathbb{R}^{H\times W\times 3}$, while the attention masks decoder output 9 background attention maps $A_x^f|_{f=1}^9$ and 1 foreground attention mask A_x^b, where $A_x^{f,b} \in \mathbb{R}^{H\times W}$. On the other hand, the Siamese encoder is trained by two inputs: the first one is the source slice x while the second one is composed of a layer of x_{LI} in the image domain and two layers of A_{MG}.

The term "guided" refers to the second input being fed into the encoder, allowing the network to understand the location of the artifact rapidly and, with the assistance of x_{LI}, pursue a preliminary reduction of the metal artifact. Nevertheless, it is widely known that linear interpolation can lead to the introduction of secondary artifacts while removing the main ones. Actually, by using the Siamese encoder, our architecture is not affected by this issue due to its ability to determine which features ought to be sourced from x_{LI} and which should be obtained from x. In fact, the Siamese encoder has already been used to distillate information coming from heterogeneous sources, preserving only the most useful ones (Wan et al., 2024).

This background attention mask A_x^b is crucial to perform a correct image-to-image translation. For instance, we want the parts not affected by the artifact to be left untouched by the GAN with only the foreground requiring to be corrected. For this reason, when reconstructing the final image, the linearly interpolated image x_{LI} will be considered as the background, multiplied by the corresponding A_x^b, and the GAN will prompt a correction over it. As mentioned before, in this way the GAN will be led to perform different operations over the single metal artifact correction on the image such as contrast and sharpness correction, leaving the linear interpolation algorithm the assignment to perform a first coarse artifact reduction. Thus, we reduce the probability of the network to consider artifacted portions of the image, i.e., beam hardening or streaks, as being normal portions. The remaining foreground attention masks are multiplied by the corresponding content masks, focusing the corrections only on the zones of the image that are still artifact-affected or do not satisfy the requirements of the target distribution. Then, the final output of the network can be formulated as:

$$G_\theta(x) = \sum_{f=1}^{9} (C_x^f \cdot A_x^f) + x_{LI} \cdot A_x^b. \tag{8}$$

The detailed algorithm steps of the CALIMAR-GAN inference are outlined in pseudo-code format in Algorithm 1.

Algorithm 1: CALIMAR-GAN pseudo-code for the inference stage

Input: x: Original Slice
x_{LI}: Linearly Interpolated Slice
A_{MG}: Guided Attention Mask
Output: $G_\theta(x)$: Artifact-reduced Slice

Processing Pipeline:
- $siameseOutput \leftarrow$ SiameseEncoder(x, x_{LI}, A_{MG})
- $residualOutput \leftarrow residualStructure(siameseOutput)$
- $contentMasks \leftarrow$ ContentMaskDecoder$(residualOutput)$
- $foregroundMasks, backgroundMask \leftarrow$ AttentionMaskDecoder$(siameseOutput)$
- $backgroundImage \leftarrow x_{LI} \times backgroundMask$
- $foregroundImage \leftarrow \sum(contentMasks \times foregroundMasks)$
- $G_\theta(x) \leftarrow backgroundImage + foregroundImage$

3. Experiment settings

In this section, all the experiments conducted to evaluate the proposed approach are reported. In particular, we describe the CALIMAR-GAN implementation details, the dataset, the compared approaches along with the common settings used to train the networks, and the evaluation metrics.

3.1. Compared approaches

The proposed approach was compared against five methods specifically designed for the MAR problem, along with baseline versions of CycleGAN (Zhu et al., 2020) and AttentionGAN-v2 (Tang et al., 2021). In particular, the comparison includes: LI, a conventional and widely used approach; CycleGAN, AttentionGAN-v2, and ADN (Liao et al., 2020), which are unpaired methods; DICDNet (Wang et al., 2022a), InDuDoNet (Wang et al., 2021a), and InDuDoNet+ (Wang et al., 2023), which are paired methods. To ensure a fair comparison, consistent training settings were applied across all networks, including: an input size equal to 256 × 256, a batch size equal to 8 (except for ADN where the value is set to 1 due to memory limitations), and sinogram dimensions of 320 × 321 (with 320 projection views uniformly spaced in 360 degrees and 321 detectors). Default values for other parameters (e.g., training epochs and learning rate) were used based on the original papers. Officially released code was used for all the compared methods except LI, which was implemented using PyTorch (Paszke, 2019), consistent with the implementations of DICDNet, InDuDoNet, and InDuDoNet+.

To ensure a fair and robust comparison between unpaired and paired methods, considering that unpaired methods require real data for training, while also ensuring a sufficient number of test data to obtain sufficiently tough metrics — each network was trained on six different data combinations. Paired methods were trained on combinations that included only simulated data in the training set, while unpaired methods used combinations of both simulated data and real data to train. In particular, the combinations were arranged so that both paired and unpaired methods were evaluated on the identical simulated and real artifact test sets, ensuring consistency across the six folds for both method types.

3.2. Dataset

CALIMAR-GAN and the other compared methods were trained and tested on a dataset built from the publicly available slices released by SpineWeb[1] of the Digital Imaging Group. In particular, the "Vertebrae

[1] http://spineweb.digitalimaginggroup.ca

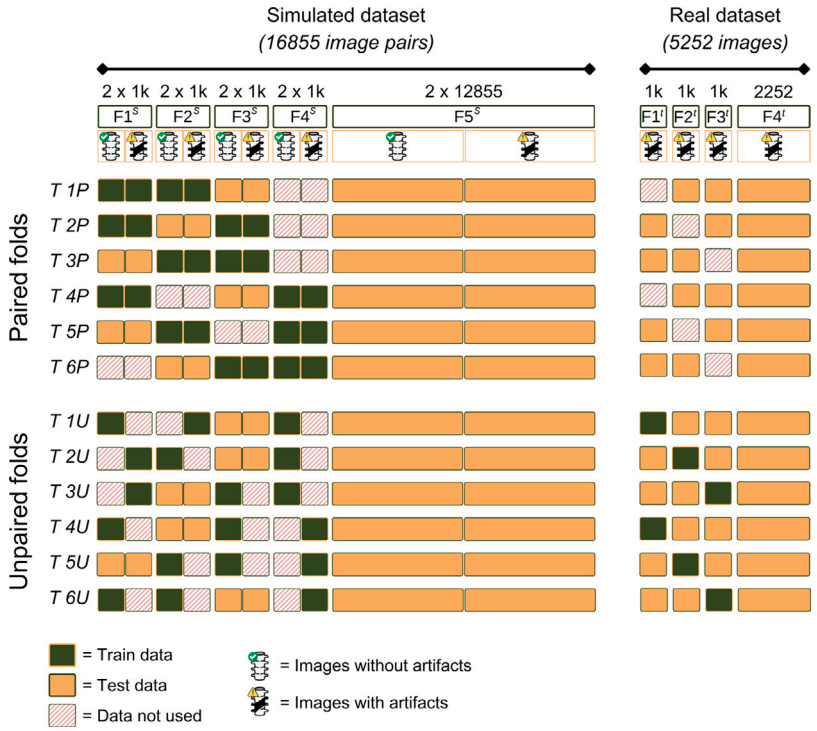

Fig. 4. Graphical representation of the dataset used to evaluate the methods. The *Simulated dataset* contains 16,855 image pairs and is divided into five folds: $F1^s$, $F2^s$, $F3^s$, and $F4^s$ contain 1000 images with simulated artifacts and 1000 corresponding artifact-free images, while $F5^s$ is composed of 12,855 images with simulated artifacts and 12,855 corresponding artifact-free images. The *Real dataset* is divided into four folds: $F1^t$, $F2^t$, and $F3^t$ contain 1000 images with real artifacts, while $F4^t$ is composed of 2252 images with real artifacts. For each fold (e.g., $T1\,P$, $T2\,P$, …, $T1\,U$, …), the green and orange squares represent data used respectively for training and testing, while the diagonal red and white striped squared correspond to data not used in any phase.

Table 1
Simulated and real data employed for training and testing, either for artifacted and artifact-free distributions, according to the method employed. The reported data correspond to one fold and are representative of all six folds.

Methods	Train set size		Test set size	
	Simulated dataset	Real dataset	Simulated dataset	Real dataset
Paired	4,000	–	13,855	4,252
Unpaired	2,000	2,000	13,855	4,252

Localization and Identification" was employed containing slices from 125 patients with various pathologies, some of which include metal implants. The dataset was well-suited for a fair comparison between paired and unpaired methods on both simulated and real artifacts. Fig. 4 reports a schematic description of the dataset organization.

Initially, slices containing metals were separated from the entire dataset following a common thresholding procedure, which identifies metal objects with Hounsfield Units (HU) greater or equal than 2500 (Yu et al., 2021b; Liao et al., 2020). All slices with metal masks larger than 400 pixel were grouped into the *Real dataset*, resulting in 5252 slices, which were split into four folds: (i) $F1^t$ - 1000 samples, (ii) $F2^t$ - 1000 samples, (iii) $F3^t$ - 1000 samples, (iv) $F4^t$ - 2252 samples.

The simulation of metal artifacts in previous studies has often been limited to simple, regular geometries (e.g., screws or circles). However, real-world scenarios involve a broader range of metal shapes and configurations, which vary significantly between cases and may include multiple metals within a single slice. This variability is illustrated in Fig. 5, where some real artifacts extracted from the dataset are shown: for example, Fig. 5(b) illustrates how a screw, perfectly represented in Fig. 5(c), may only be partially captured from the tomograph if it is not oriented perpendicularly to the detectors. Since data-driven networks strictly depend on the variety of data, a *mask database*, containing all real metals from the dataset (5252 masks), was created. The same thresholding procedure used to create the *Real dataset* was applied to obtain the *mask database*.

The remaining 16,855 slices, which do not contain any metal mask, were used to create the *Simulated dataset*, using the previously created *mask database*, in accordance with the following procedure executed on each slice:

- Random selection of a mask from the *mask database*.
- The mask is subjected to random rotations ranging from 1° to 360° and random scaling between 0.6 and 1 prior to being applied to the slices.
- The metal mask is applied specifically in the proximity of bones, using a bone selection strategy: first, a pixels intensity threshold of ≥ 500 is applied to identify bone regions; then, the centroid of the metal mask is aligned with the centroid of a randomly selected bone zone containing more than 400 pixels.
- Finally, the procedure proposed by Sakamoto et al. (2019) is used to simulate the artifacts under the assumption of using titanium metal, which is widely considered the best metal for vertebrae implants due to its exceptional biocompatibility, corrosion resistance, and mechanical properties (Fleck and Eifler, 2010; Kamachimudali et al., 2003).
- This process is repeated 25 times for each artifact-free image, generating a total of 421,375 images containing artifacts.

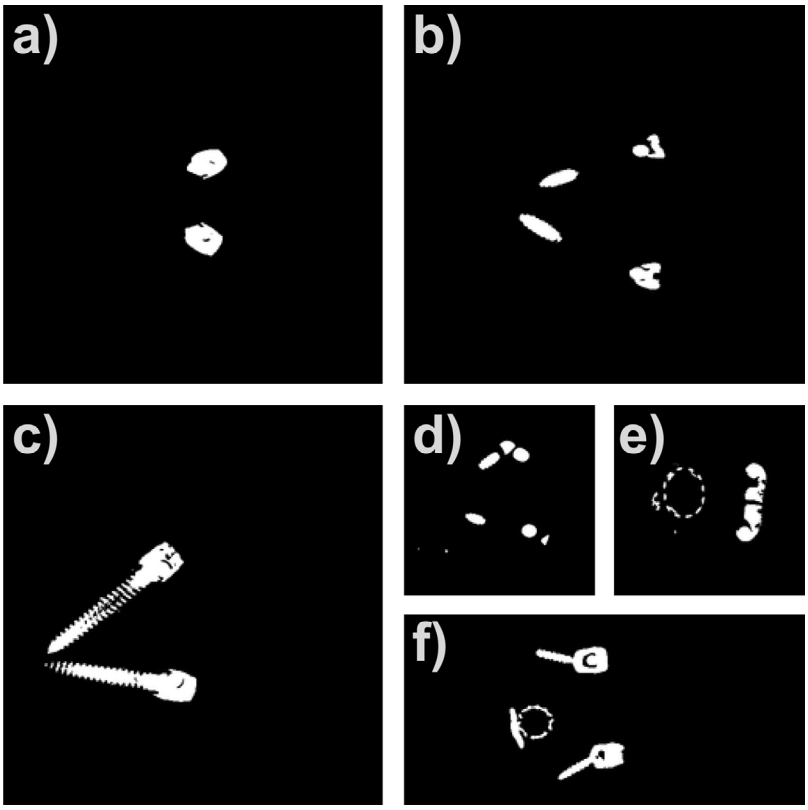

Fig. 5. Examples of the binary masks used to generate artifacts and create the *Simulated dataset*. These masks are subsequently applied during the synthetic artifact generation process over clean CT images to simulate titanium implants.

Following similar studies (Yu et al., 2021b; Wang et al., 2022a), during training, only one random mask was chosen per iteration. To provide a more seamless explanation of the dataset, the twenty-five artifact-affected slices generated from the corresponding artifact-free slice are grouped together. Consequently, when discussing the dataset, we refer to pairs of images, which consist of the artifact-free slice and the cluster of artifact-affected slices.

The *Simulated dataset* was divided into five folds: (i) $F1^s$ - 1000 pairs, (ii) $F2^s$ - 1000 pairs, (iii) $F3^s$ - 1000 pairs, (iv) $F4^s$ - 1000 pairs, (v) $F5^s$ - 12855 pairs

. To ensure a fair evaluation of the methods using consistent slices and metals, a single metal from the twenty-five available was assigned to each artifact-free image before the testing phase.

The dataset was organized into folds to enable cross-validation across networks. Specifically, six different settings were used for both paired ($T\ 1P$ to $T\ 6P$) and unpaired methods ($T\ 1U$ to $T\ 6U$), as depicted in Fig. 4. Paired methods were trained using combinations of two "1k pairs" folds from the *Simulated dataset* (i.e., from $F1^s$ to $F4^s$), whereas unpaired methods combined one "1k" samples with simulated artifacts, one "1k" samples with real artifacts (e.g. in T 1U: 1000 simulated samples from $F2^s$ and 1000 real samples from $F1^t$), and two "1k" samples without artifacts (e.g. in T 1U: 1000 samples from $F1^s$ and 1000 real samples from $F3^s$). To maintain a fair comparison with an equal quantity of slices and maximize the use of available slices, each fold attempt excluded some data from testing. Consequently, the performance evaluation used 13,855 and 4,252 slices per simulated and real test set, respectively. Furthermore, to reflect real-world data limitations, the training datasets were restricted to 2,000 slices, ensuring a sufficient number of real samples for testing. Table 1 provides a summary of the number of slices used for the training and testing phases.

3.3. Network implementation

This section outlines the CALIMAR-GAN architecture, training procedures, and parameters employed.

3.3.1. Network architecture

The generator architecture used in CALIMAR-GAN is based on CycleGAN (Zhu et al., 2020). The Siamese encoder processes a double three-channel input and produces nine three-channel content masks and ten one-channel attention masks. Specifically, the encoder receives inputs structured as follows:

$$\begin{cases} Input1 = & [x & ;x & ;x &] \\ Input2 = & [x_{LI} & ;A_{MG} & ;A_{MG} &] \end{cases}$$

As depicted in Fig. 3, both the encoder and decoders consist of three convolutional layers, each followed by an Instance Normalization (IN) layer and a ReLU activation function. The Content Mask decoder concludes with a $tanh$ activation function, while the Attention Mask decoder employs a $softmax$ activation function. A residual structure containing five residual blocks is placed after the Siamese encoder, serving as the input for the Content Mask decoder. Each residual block comprises $[Conv + IN + ReLU + Conv + IN + ReLU + SkipConnection]$, with the convolution operations (denoted as $Conv$) utilizing a 3 × 3 kernel and a stride and padding of 1.

3.3.2. Training details

The proposed CALIMAR-GAN is implemented using PyTorch. The detailed environment settings are depicted in Table 2. The Adam optimizer is employed for training with parameters $[\beta_1, \beta_2] = (0.5, 0.99)$ and a learning rate equal to 2×10^{-4}. The training uses a batch size of 8 and is conducted for 100 epochs. The same learning rate is kept for the first 60 epochs and linearly decay the rate to zero over the next 40 epochs. Weights are initialized from a Gaussian distribution $\mathcal{N}(0, 0.02)$. The

Table 2
Experimental environment settings.

	Item	Setting
Hardware	CPU	AMD Ryzen (TM) Threadripper PRO 5965WX
	GPU	NVIDIA GeForce RTX A6000
	RAM	64GB
Software	OS	Windows 11
	Language	Python version 3.9
	Framework	PyTorch version 2.2.2 + CUDA 11.8

Table 3
Training parameters for CALIMAR-GAN.

Train-parameter	Value
Epochs	100
Learning rate	2×10^{-4}
Learning decay schedule	Linear decay after 60 epochs
Image_size	256 × 256
Batch_size	8
Optimizer	Adam
λ_{cycle}	10
λ_{id}	0.8

input image size is 256 × 256 pixels. In line with protocols from similar studies (Yu et al., 2021b; Wang et al., 2022a), each training iteration involves randomly selecting one metal-free CT slice and one metal-affected CT slice. For the latter, a metal mask is randomly chosen from the twenty-five available options, as augmentation strategy, with different slices utilizing different metal masks as described in Section 3.2. For the total loss detailed in Eq. (5), λ_{cycle} is set to 10, while λ_{id} is set to 0.8. A compact view of the training parameters used in CALIMAR-GAN is provided in Table 3. The total number of model-trainable parameters is $20.6M$.

3.4. Evaluation metrics

Evaluating quality in medical imaging is a multifaceted and evolving area of research. Previous studies have demonstrated that visual perception cannot be captured by a single metric, as each metric has its own limitation (Kim et al., 2011; Johnson et al., 2011). This challenge is exacerbated when ground truth data is not available (Bischoff et al., 2024). Nonetheless, using a range of metrics that focus on different image characteristics, along with computing indices on simulated data, can help identify the most effective methods for MAR. For simulated data, several standard quantitative metrics are employed, including peak signal-to-noise ratio (PSNR), root mean squared error (RMSE), and structural similarity index measure (SSIM). Furthermore, the multi-scale structural similarity index measure (MS-SSIM) is calculated, which has been proven to be a more robust estimator than SSIM under certain conditions (Wang et al., 2003; Rouse and Hemami, 2008). These metrics are not applicable to real data due to the absence of paired artifacted and artifact-free images. To the best of our knowledge, current MAR studies generally evaluate clinical data only qualitatively, with limited exploration of the quantitative correlation between performance achieved on simulated artifact data and performance on real artifact data.

The FID is utilized for both simulated and real data. Although FID is typically more effective for large datasets (e.g., 10,000+ samples), it serves as a valuable metric for evaluating distributions in scenarios where ground truth is not available. However, given that the real dataset comprises over 5000 images, FID can still offer meaningful quantitative insights when combined with qualitative inspection. To further strengthen the effectiveness of this metric, a correlation analysis is conducted between FID and the other pixel-based metrics. Appendix A in the supplementary materials provides a complete overview of the aforementioned metrics definition and relative references.

4. Results and discussion

This section presents the results of the analysis conducted on both simulated and real artifact data. To better illustrate the variability between methods when processing different data, a summary table is provided alongside a box plot representation of the metrics discussed in Section 3.4. Additionally, three ablation studies are reported: the first focuses on the components introduced in the optimized version of G_θ, the second examines the impact of different training strategies and data selection on performance in real environments, and the third explores the network performance when varying the λ_{id}. Finally, the relationship between the FID and the pixel-based metrics is evaluated. The inference time for artifact reduction of a single slice is 12.41 ± 2.68 ms, computed over a set of 10,000 slices. This time includes both the preprocessing steps and the network inference.

4.1. Results on simulated artifact data

4.1.1. Quantitative comparison

Table 4 presents a quantitative comparison of various MAR techniques on the *Simulated dataset*, while Fig. 6 illustrates the variability between different folds for each method across the pixel-based metric. Most methods outperform the traditional LI approach, achieving higher SSIM, MS-SSIM, and PSNR and lower RMSE scores. However, CycleGAN and ADN show slightly elevated FID values, suggesting difficulties in completely removing certain types of artifacts. Paired methods generally achieve better performance than unpaired methods such as ADN and the baseline versions of CycleGAN and AttentionGAN-v2. Among the evaluated methods, CALIMAR-GAN consistently outperforms the others, demonstrating superior effectiveness in removing artifacts without distorting unaffected areas of the slice. Moreover, the boxplots reveal that CALIMAR-GAN not only obtains the highest scores across various metrics but also exhibits one of the lowest variances demonstrating its robustness.

4.1.2. Qualitative comparison

Fig. 7 presents a qualitative comparison of the results obtained by each method. To further clarify the differences between the methods and the ground truth, Fig. 8 displays visual distinctions, generated by normalizing the slices in the range $[-1; 1]$ and subtracting the ground truth from the output of each method. The baseline versions of AttentionGAN-v2 and CycleGAN, as well as ADN, attempt to remove artifacts by substituting them with gray content that is not coherent with the rest of the slice. Although these networks recognize difference between the artifact-free and the artifacted slices, they fail to fully restore the affected area. The LI method effectively reduces artifacts but often introduces secondary ones, which often are streak-like. Conversely, the paired methods, namely DICDNet, InDuDoNet, and InDuDoNet+, show a good artifact reduction. However, DICDNet tends to blur the artifacted areas, while InDuDoNet and InDuDoNet+ blur the entire image and degrade the quality of the darker areas, such as the lungs, compromising their effectiveness for certain types of slices. CALIMAR-GAN effectively preserves fine anatomical details, such as lung texture; it may occasionally introduce blurring in more complex regions, like near the spine. This effect is likely due to the network's limited ability to recover information that is heavily corrupted or missing in the input, especially in areas with high structural variability.

4.1.3. Evaluation with distribution-based metric

Evaluating experiments on real artifact data presents challenges for three main reasons: the difficulty in obtaining paired real data for pixel-based metrics, limited data availability for distribution-based metrics, and an unclear connection between distribution-based and pixel-based metrics. In this context, FID represents a reliable distribution-based metric, commonly employed in several tasks. However, its ability to recognize artifacts remains uncertain, especially given the need for

Table 4
Comparison of quantitative results on the *Simulated dataset*. Bold indicates the best performance.

	Methods	MS-SSIM ↑	SSIM ↑	PSNR ↑	RMSE ↓	FID ↓
	LI (Kalender et al., 1987)	0.946 ± 0.001	0.816 ± 0.001	24.5 ± 0.1	16.2 ± 0.1	57.8 ± 0.1
Paired methods	DICDNet (Wang et al., 2022a)	0.956 ± 0.008	0.870 ± 0.005	30.8 ± 0.9	7.8 ± 0.7	40.6 ± 7.0
	InDuDoNet (Wang et al., 2021a)	0.956 ± 0.005	0.812 ± 0.016	31.0 ± 0.5	7.4 ± 0.5	44.8 ± 11.2
	InDuDoNet+ (Wang et al., 2023)	0.955 ± 0.002	0.819 ± 0.009	31.4 ± 0.2	7.1 ± 0.2	37.7 ± 2.7
Unpaired methods	AttentionGAN-v2 (Tang et al., 2021)	0.941 ± 0.021	0.832 ± 0.022	29.4 ± 1.3	9.1 ± 1.5	44.3 ± 7.8
	ADN (Liao et al., 2020)	0.940 ± 0.001	0.843 ± 0.003	25.7 ± 0.2	13.8 ± 0.4	66.5 ± 1.1
	CycleGAN (Zhu et al., 2020)	0.936 ± 0.023	0.799 ± 0.084	28.7 ± 2.1	10.0 ± 2.3	66.7 ± 15.5
	CALIMAR-GAN	**0.963 ± 0.002**	**0.877 ± 0.003**	**31.7 ± 0.2**	**6.9 ± 0.1**	**22.1 ± 2.8**

large datasets and its focus on distribution differences rather than direct artifact detection. This section aims at demonstrating that FID can effectively capture differences between the distributions X and Y by performing a Pearson correlation analysis. Table 5 shows the correlations between FID and other metrics, including MS-SSIM, SSIM, PSNR, and RMSE. In particular, FID correlates with MS-SSIM, SSIM, PSNR, and RMSE at −0.767, 0.601, 0.797, and 0.752, respectively. All these correlations achieve a p-value less than the significance level of 0.01. The obtained results suggest that FID, which estimates the similarity of image distributions between the network output and the ground truth, is strongly related with pixel-based metrics in simulated data. This indicates that FID could be a valuable metric for evaluating artifacts in real-world contexts.

4.2. Performance evaluation on anatomical fidelity

CALIMAR-GAN was evaluated on a downstream segmentation task to assess its effectiveness in performing MAR without compromising the anatomical fidelity of structures near the artifact. We then focus on the vertebrae. The metal-free dataset VerSe (Sekuboyina et al., 2021; Löffler et al., 2020; Liebl et al., 2021) was used to train nnU-Net (Isensee et al., 2021) on the vertebrae segmentation task. Then, the trained nnU-Net is tested on the artifact-reduced slices generated by the different MAR methods. Performance is evaluated using the Dice score on the *Simulated dataset*.

Fig. 9 reports the quantitative results over the different methods and the different folds and the Dice score computed using images with artifacts as baseline. Fig. 10 shows the qualitative results of the segmentation on the outputs of the different MAR methods. Although the nnU-Net was trained on a slightly different data distribution, the segmentation results on metal-affected images prove effective in showing how CALIMAR-GAN is still able to achieve the most accurate segmentation among the different methods. It successfully restores structures compromised by artifacts, further demonstrating its potential for clinical applications.

4.3. Results on real artifact data

The performance of all MAR methods was evaluated using real data both quantitatively and qualitatively. The FID scores for the real dataset are presented in Table 6 and Fig. 11. A qualitative comparison of the methods is shown in Fig. 12, highlighting their performance on different types of artifacts. The LI approach tends to introduce secondary streaking artifacts across all images. Among paired methods, InDuDoNet and InDuDoNet+ generally blur the images, leading to a loss of details near bones and within the lungs. DICDNet partially reduces artifacts while preserving lung areas; however, it often replaces bones with soft tissues. ADN and CycleGAN effectively maintain non-corrupted areas but apply insufficient artifact reduction. AttentionGAN-v2 performs light blurring to alleviate streaking artifacts without altering artifact-free zones, yet it tends to gray out artifacts rather than restoring anatomical features. In contrast, CALIMAR-GAN excels in restoring the content affected by artifacts while also preserving the underlying anatomy. Paired methods encounter difficulties when dealing with real data, and FID confirms these qualitative findings by demonstrating their limitation. Unpaired methods, particularly CALIMAR-GAN, show improved generalization for MAR tasks in real environments and exhibit low variability across folds. As reported in Table 6, CALIMAR-GAN achieved the smallest FID value on the dataset containing real clinical metal artifacts. However, the mean value is higher than the mean value measured when testing the simulated dataset. To better understand the reason for this difference, we first analyzed the effect of the test set size. Specifically, we have computed the FID on simulated data by considering a group of randomly paired images with a size comparable to that of the real data test set (about 4,000 images). We then compared the FID distance between two groups of images: 4,000 images from the simulated dataset with no metal artifacts and the corresponding paired images processed by CALIMAR-GAN after the introduction of simulated artifacts. The computed FID is 34.7, which is comparable to the value obtained from real clinical data — thus supporting its validity as a reliable metric. CALIMAR-GAN's ability to leverage knowledge from real data during training provides a significant advantage, given that simulated environments only approximate real-world conditions. However, CALIMAR-GAN is not without limitations. It can occasionally produce secondary artifacts, particularly due to the beam hardening effect when metals are close. Moreover, the method tends to be conservative, sometimes reducing but not completely eliminating the artifact to preserve the content integrity without excessive blurring (i.e., artifacts such as streaks are diminished but not entirely removed).

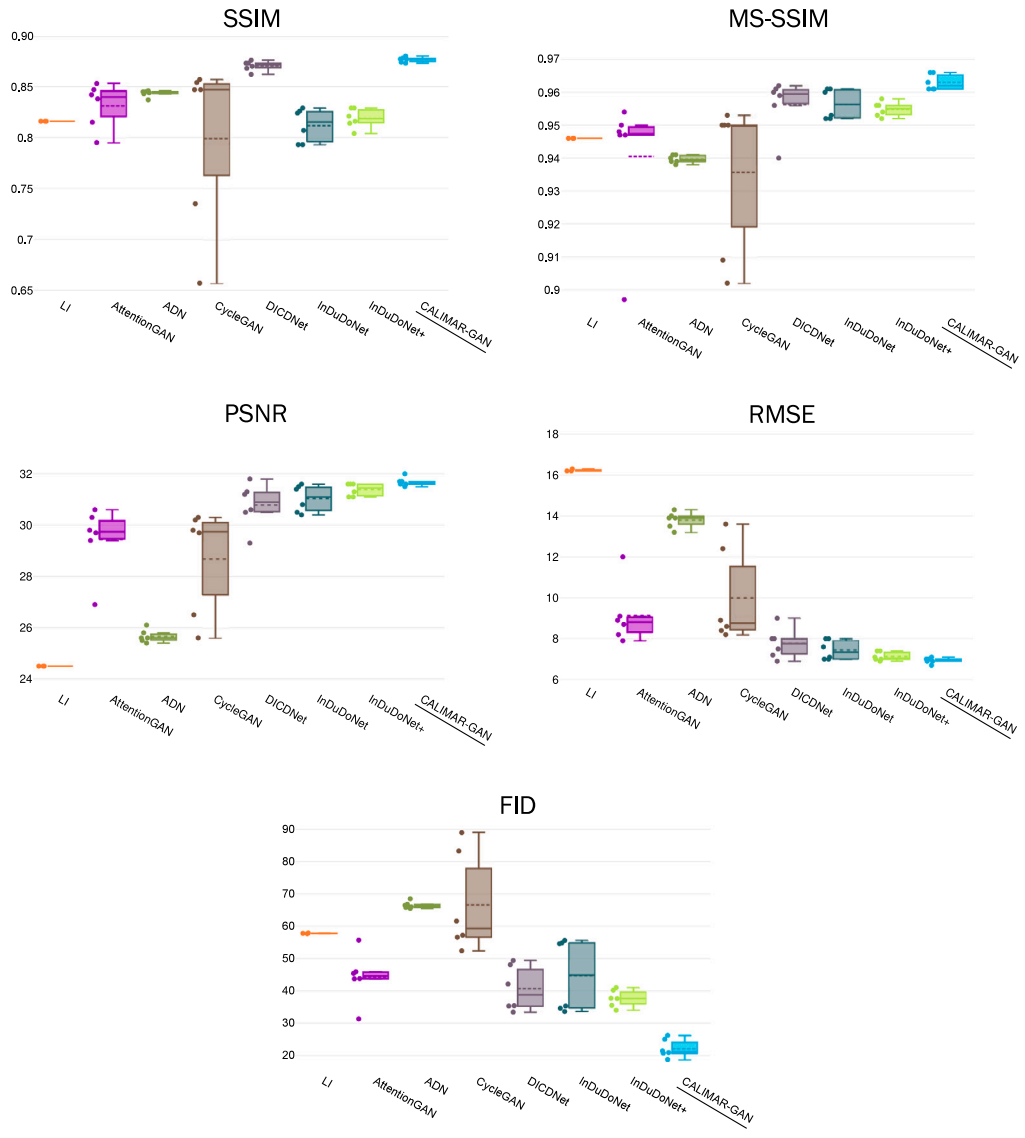

Fig. 6. Boxplots illustrating the distribution of each metric for the different methods on the *Simulated dataset*.

Table 5
Correlations between the various adopted metrics using the Pearson coefficient.

	FID ↓	MS-SSIM ↑	SSIM ↑	PSNR ↑	RMSE ↓
↓ FID	1	–	–	–	–
↑ MS-SSIM	−0.767	1	–	–	–
↑ SSIM	−0.601	0.697	1	–	–
↑ PSNR	−0.797	0.689	0.386	1	–
↓ RMSE	0.752	−0.632	−0.356	−0.997	1

Table 6
Comparison of results on the *Real dataset*. Bold indicates the best performance.

	Methods	FID ↓
	LI (Kalender et al., 1987)	85.3 ± 0.2
Paired methods	DICDNet (Wang et al., 2022a)	94.6 ± 21.9
	InDuDoNet (Wang et al., 2021a)	75.2 ± 13.1
	InDuDoNet+ (Wang et al., 2023)	66.2 ± 3.8
Unpaired methods	AttentionGAN-v2 (Tang et al., 2021)	39.8 ± 2.3
	ADN (Liao et al., 2020)	45.2 ± 4.3
	CycleGAN (Zhu et al., 2020)	46.2 ± 13.1
	CALIMAR-GAN	**32.7** ± **3.8**

4.4. Ablation studies

To further highlight the capabilities of the proposed network, three ablation studies are conducted to understand the impact of specific design/parameters choices (i.e., the use x_{LI}, and A_{MG}) and training strategies (i.e., employing 50% of simulated artifact data and 50% of real artifact data).

4.4.1. Network modules

Leveraging the baseline version of AttentionGAN-v2, CALIMAR-GAN is tested under two conditions: using x_{LI} alone as both input and background image for the network output, and incorporating both x_{LI} and A_{MG}.

The results in Table 7 demonstrate notable improvements in performance metrics, particularly in terms of SSIM and FID, when each module is utilized. This indicates that the addition of mask guidance contributes significantly to obtaining reliable and sturdy results.

4.4.2. Training strategies

Three training strategies are evaluated to determine their effectiveness: training with only simulated data, only real data, and a combination of 50% simulated data and 50% real data. As shown in

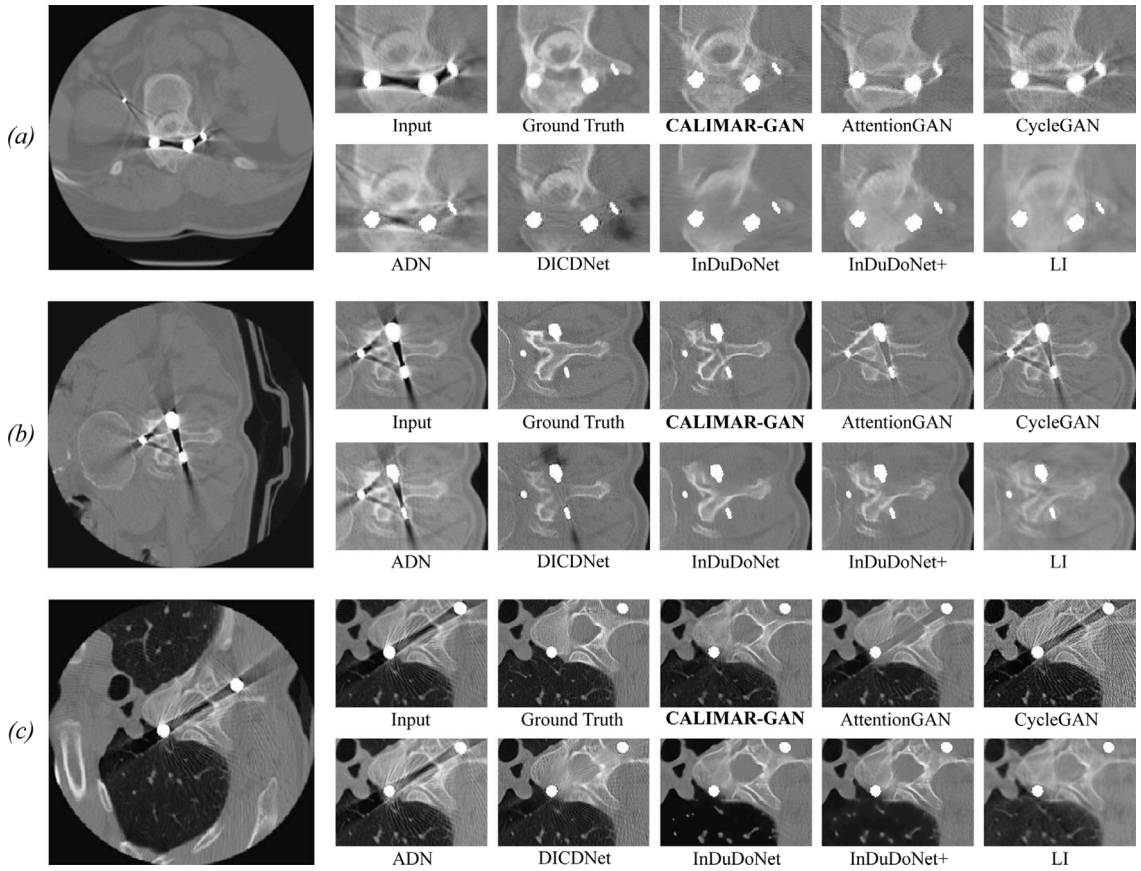

Fig. 7. Visual comparisons of the MAR methods on different simulated artifacts *(a)*, *(b)*, and *(c)*. To assess the artifact removal process more effectively, only zoomed-in portions (on the right) of the slices (on the left) are shown. Subfigure *(c)* emphasizes the influence of the methods on anatomical structures (i.e., the lungs), which are not affected by the artifacts.

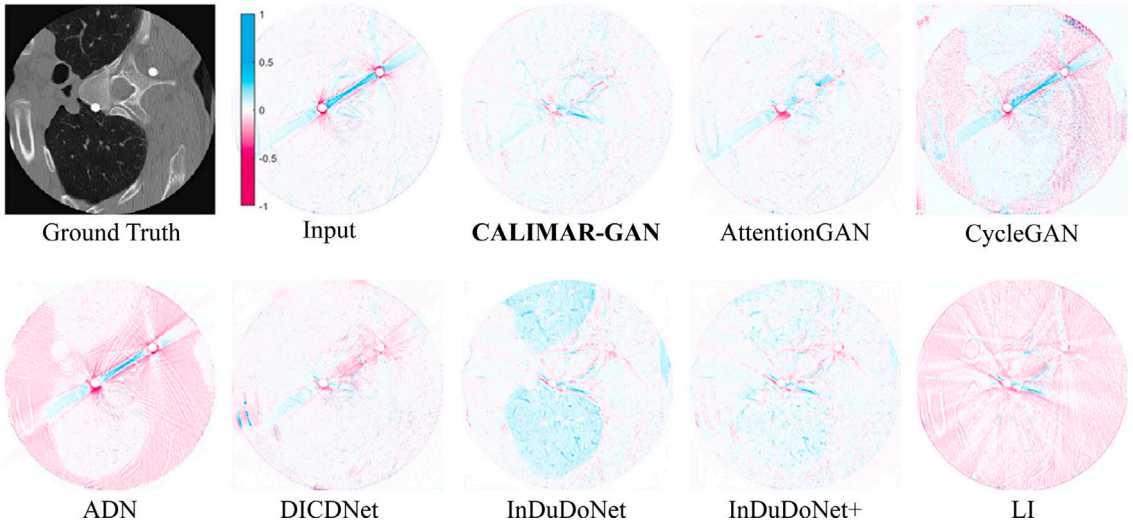

Fig. 8. Differences normalized in the range $[-1;1]$ between the ground truth and each compared method, for the input image shown in Fig. 7*c*.

Table 8, training with only real data is the best option as it achieves the lowest FID values on the *Real dataset*. However, training exclusively with simulated data results in high performance on the *Simulated dataset* but a decline in real test set performance. The composite use of simulated and real data provides a good trade-off between performance and data availability, thus addressing issues related to clinical applicability.

4.4.3. Optimal λ_{id} selection

The identity loss L_{id} helps to preserve the content similarity between the input and the artifact-reduced slice. Maintaining the artifact-free zones untouched by the network is crucial to obtain a successful MAR procedure. Therefore, an ablation study was conducted to find the optimal λ_{id} parameter that weights the L_{id} in the Eq. (5). As shown in

Table 7
Impact of different components in CALIMAR-GAN on performance, evaluated using both the *Simulated* and *Real dataset*.

Methods	*Simulated dataset*					*Real dataset*
	FID ↓	MS-SSIM ↑	SSIM ↑	PSNR ↑	RMSE ↓	FID ↓
Only attention	55.7	0.897	0.795	26.9	12.0	40.4
Attention + LI	35.5	0.965	0.868	31.7	6.9	56.5
Attention + LI + A_{MG}	**20.9**	**0.966**	**0.877**	**32.0**	**6.7**	**37.8**

Table 8
Impact of different training strategies in CALIMAR-GAN on performance evaluated using both the *Simulated* and *Real dataset*.

Data used	N. of training slices	*Simulated dataset*					*Real dataset*
		FID ↓	MS-SSIM ↑	SSIM ↑	PSNR ↑	RMSE ↓	FID ↓
Only simulated	2,000	23.1	0.966	0.877	31.8	6.89	36.7
Only real	2,000	23.1	0.937	0.828	29.5	8.79	**31.8**
Simulated + Real	2,000	**20.8**	**0.966**	**0.877**	**32.0**	**6.72**	32.6

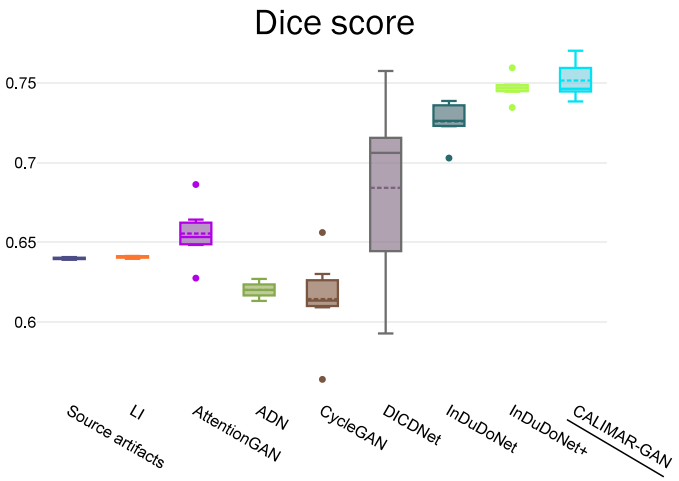

Fig. 9. Boxplot illustrating the distribution of the Dice score for the downstream segmentation task when images with artifacts or the output of different MAR methods on the *Simulated dataset* are used.

Table 9
Impact of different λ_{id} values on the CALIMAR-GAN performance evaluated using both the *Simulated* and *Real dataset*.

λ_{id} value	*Simulated dataset*					*Real dataset*
	FID ↓	MS-SSIM ↑	SSIM ↑	PSNR ↑	RMSE ↓	FID ↓
0.5	32.0	0.956	0.835	29.1	8.11	39.5
0.8	**20.8**	**0.966**	**0.877**	**32.0**	**6.72**	**32.6**
1.0	36.5	0.932	0.806	28.3	9.57	41.9

Table 9, $\lambda_{id} = 0.8$ obtains the best performance on the *Simulated dataset* as well as on the *Real dataset*.

4.5. Limitations and future works

In this subsection, the main limitations of the study, according to the authors, are reported and discussed along with some possible future works. Like all the works already presented on this topic, a limitation of our study is that CALIMAR-GAN has been tested on retrospective data. For a prospective validation, for instance, it would be useful to acquire a dataset with both pre- and post-implant CT scans to compute quantitative metrics on clinical data. However, it is not easy to build such a dataset. As a next step, the authors plan to collaborate with clinical institutions and veterinary centers to conduct prospective studies on newly acquired CT scans from both human and animal patients with metallic implants. This will ensure a more comprehensive and clinically relevant validation of the proposed approach.

While the performed experiments demonstrate the effectiveness of the method in reducing metal artifacts, its generalizability across different real-world clinical scenarios remains to be further assessed. It is well known, that the quality and characteristics of reconstructed CT images can be significantly influenced by several factors, such as different scanner hardware across manufacturers, variations in X-ray beam energy, and proprietary image reconstruction algorithms. These variations may introduce domain shifts that could affect the performance of CALIMAR-GAN on real clinical data. The future work will focus on validating the robustness of the model across multiple CT scanners and diverse acquisition protocols. This will require datasets with and without real metal artifacts acquired from different imaging systems.

Regarding CALIMAR-GAN's limitations, the beam hardening effect, which is challenging to remove in image-based methods, could be mitigated by adding a sinogram-based pre-processing reduction step. Although CALIMAR-GAN preserves fine details, such as lung texture, it may introduce blurring in more complex regions, like near the spine.

As a additional future research direction, the authors will corroborate the importance of FID metric with radiologists' judgment. Furthermore, the relation between the FID and the pixel-based metrics will be validated with a paired dataset of real artifacts. However, the FID provides a global measure of distributional similarity between processed and artifact-free images, then it is not clear the relation between the metric values and the specific artifact types, such as beam hardening or streaks. Future work could focus on developing dedicated datasets designed to isolate and analyze different types of artifacts, allowing for a more targeted evaluation.

5. Conclusion

This study has introduced a mask-guided attention network built on a cycle-consistent GAN for the metal artifact reduction (MAR) task in Computed Tomography (CT) scans. Extensive evaluations on images with simulated and real metal artifacts have demonstrated the effectiveness of the proposed method, which has been compared with various state-of-the-art methods, including paired and unpaired techniques. The proposed method exhibits robustness, showing consistent performance across multiple test folds. Ablation studies have revealed that the proposed design choices and mixed training strategy significantly enhance performance, providing notable quantitative and qualitative results. Additionally, the study has illustrated that the Fréchet inception distance (FID) serves as a valid performance indicator

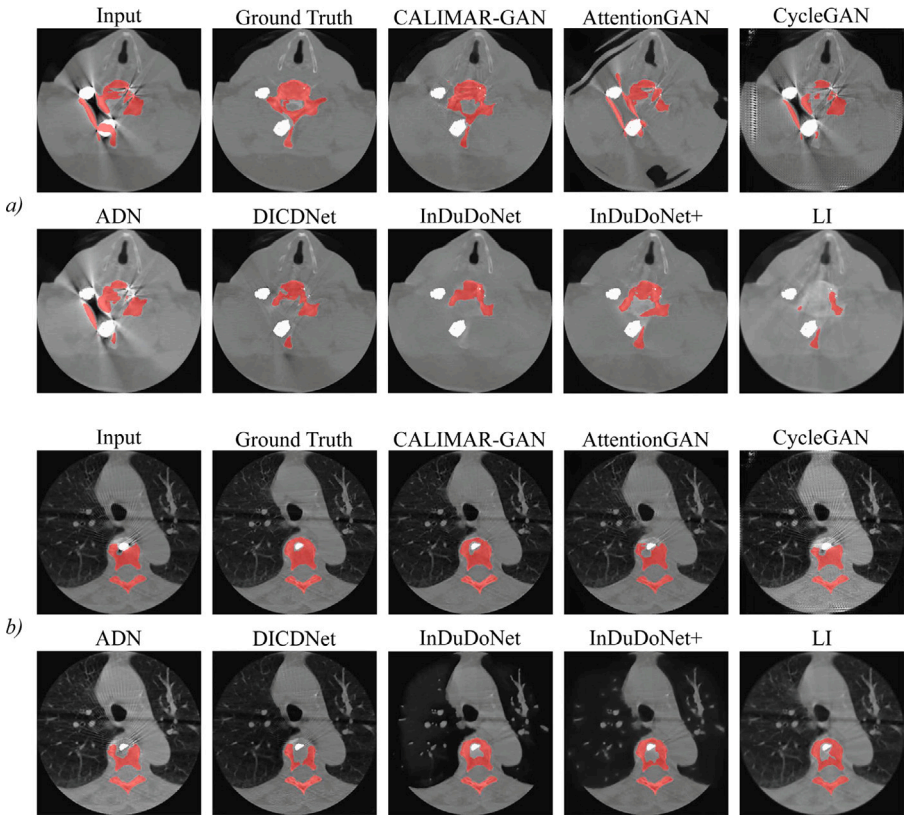

Fig. 10. Qualitative results of the downstream segmentation task on the artifacted input image, the ground truth, and the outputs obtained using different MAR methods.

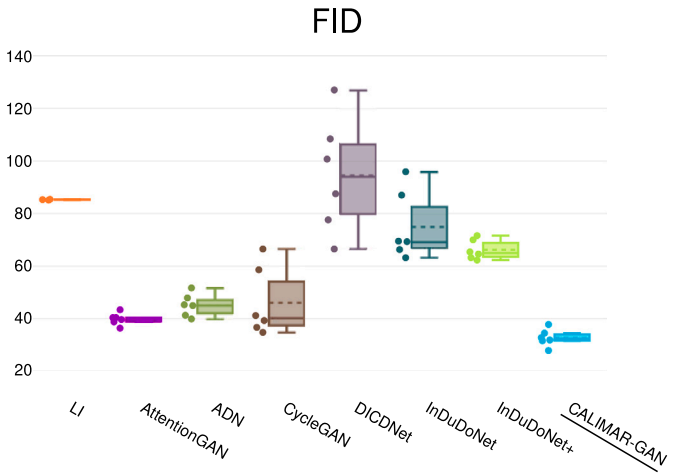

Fig. 11. Boxplot illustrating the distribution of the FID for the different methods on the *Real dataset*.

for real artifact data, complementing qualitative inspection. Quantitative evaluations on 13,855 simulated images demonstrated superior performance, achieving a PSNR of 31.7, SSIM of 0.877, and FID of 22.1, outperforming state-of-the-art methods. On 4,252 real clinical images, CALIMAR-GAN achieved the lowest FID (32.7), validated as a robust metric through correlation with pixel-based metrics ($r = -0.797$ with PSNR, $p < 0.01$; $r = -0.767$ with MS-SSIM, $p < 0.01$). This study bridges deep learning-based artifact reduction to clinical workflows through high-fidelity image reconstructions, directly improving diagnostic precision and treatment efficacy.

CRediT authorship contribution statement

Roberto Maria Scardigno: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antonio Brunetti:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Pietro Maria Marvulli:** Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Data curation. **Raffaele Carli:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Mariagrazia Dotoli:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Vitoantonio Bevilacqua:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Domenico Buongiorno:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by BRIEF - Biorobotics Research and Innovation Engineering Facilities - Missione 4, “Istruzione e Ricerca” - Componente 2, “Dalla ricerca all’impresa” - Linea di investimento 3.1, “Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione”, funded by European Union -

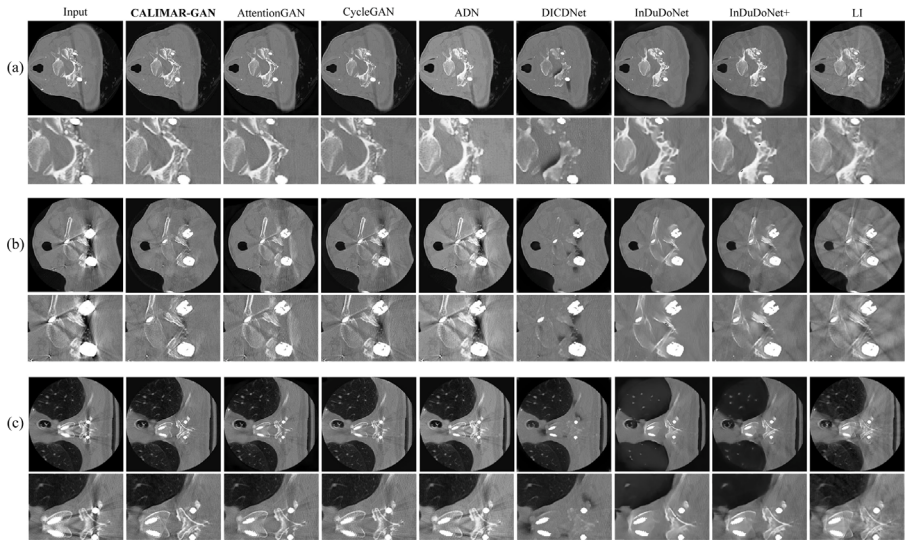

Fig. 12. Visual comparisons of methods results on different real artifacts *(a)*, *(b)*, and *(c)*. Each method is displayed with the full image and a zoomed-in view of the region near the artifact in the bottom row. Additionally, subfigure *(c)* highlights the influence of the methods on anatomical structures (i.e., the lungs) that are not affected by the real artifacts.

Next Generation EU, CUP: J13C22000400007, in part by CONTACT - CustOm-made aNTibacterical/bioActive/bioCoated prosTheses, funded by the Italian Ministry of Education, University and Research under the Programme PON R&I 2014–2020 and FSC, CUP: B99C20000300005, and in part by the European Union – Next Generation Eu - under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 3.3 - Decree No. 351 (09th April 2022) of Italian Ministry of University and Research - Concession Decree No. 2153 (28th December 2022) of the Italian Ministry of University and Research, Project code D93D22001390001, within the Italian National Programme PhD Program in Autonomous Systems (DAuSy).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2025.102565.

Data availability

The publicly available slices can be reached at http://spineweb.digitalimaginggroup.ca. The source code is available at https://github.com/roberto722/calimar-gan.

References

An, Y., Lam, H.K., Ling, S.H., 2022. Auto-denoising for EEG signals using generative adversarial network. Sensors 22 (5), 1750. http://dx.doi.org/10.3390/s22051750.

Arabi, H., Zaidi, H., 2021. Deep learning–based metal artefact reduction in PET/CT imaging. Eur. Radiol. 31 (8), 6384–6396. http://dx.doi.org/10.1007/s00330-021-07709-z.

Armanious, K., Kumar, V., Abdulatif, S., Hepp, T., Gatidis, S., Yang, B., 2020. ipA-MedGAN: Inpainting of arbitrary regions in medical imaging. In: 2020 IEEE International Conference on Image Processing. ICIP, pp. 3005–3009. http://dx.doi.org/10.1109/ICIP40778.2020.9191207, arXiv:1910.09230 [cs, eess].

Bal, M., Celik, H., Subramanyan, K., Eck, K., Spies, L., 2005. A radial adaptive filter for metal artifact reduction. In: Fitzpatrick, J.M., Reinhardt, J.M. (Eds.), In: Medical Imaging 2005: Image Processing, vol. 5747, International Society for Optics and Photonics, SPIE, pp. 2075–2082. http://dx.doi.org/10.1117/12.593095.

Bevilacqua, V., Aulenta, A., Carioggia, E., Mastronardi, G., Menolascina, F., Simeone, G., Paradiso, A., Scarpa, A., Taurino, D., 2007. Metallic artifacts removal in breast CT images for treatment planning in radiotherapy by means of supervised and unsupervised neural network algorithms. In: Huang, D.S., Heutte, L., Loog, M. (Eds.), Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1355–1363. http://dx.doi.org/10.1007/978-3-540-74171-8_138.

Bischoff, S., Darcher, A., Deistler, M., Gao, R., Gerken, F., Gloeckler, M., Haxel, L., Kapoor, J., Lappalainen, J.K., Macke, J.H., Moss, G., Pals, M., Pei, F., Rapp, R., Sağtekin, A.E., Schröder, C., Schulz, A., Stefanidi, Z., Toyota, S., Ulmer, L., Vetter, J., 2024. A practical guide to statistical distances for evaluating generative models in science. arXiv:2403.12636 [cs, stat].

Boas, F.E., Fleischmann, D., 2012. CT artifacts: causes and reduction techniques. Imaging Med. 4 (2), 229–240. http://dx.doi.org/10.2217/iim.12.13.

Chen, J., Li, Y., Zheng, H., Li, H., Wang, H., Ma, L., 2024. Hounsfield unit for assessing bone mineral density distribution within lumbar vertebrae and its clinical values. Front. Endocrinol. 15, 1398367.

Fleck, C., Eifler, D., 2010. Corrosion, fatigue and corrosion fatigue behaviour of metal implant materials, especially titanium alloys. Int. J. Fatigue 32 (6), 929–935. http://dx.doi.org/10.1016/j.ijfatigue.2009.09.009, Selected Papers of the 17th European Conference of Fracture (ECF 17), URL: https://www.sciencedirect.com/science/article/pii/S0142112309002801.

Gajera, B.V., Kapil, S.R., Ziaei, D., Mangalagiri, J., Siegel, E., Chapman, D., 2021. CT-scan denoising using a charbonnier loss generative adversarial network. IEEE Access 9, 84093–84109. http://dx.doi.org/10.1109/ACCESS.2021.3087424.

Gjesteby, L., De Man, B., Jin, Y., Paganetti, H., Verburg, J., Giantsoudi, D., Wang, G., 2016. Metal artifact reduction in CT: Where are we after four decades? IEEE Access 4, 5826–5849. http://dx.doi.org/10.1109/ACCESS.2016.2608621.

Gjesteby, L., Shan, H., Yang, Q., Xi, Y., Jin, Y., Giantsoudi, D., Paganetti, H., De Man, B., Wang, G., 2019. A dual-stream deep convolutional network for reducing metal streak artifacts in CT images. Phys. Med. Biol. 64 (23), 235003. http://dx.doi.org/10.1088/1361-6560/ab4e3e.

Gu, S., Bao, J., Yang, H., Chen, D., Wen, F., Yuan, L., 2019. Mask-guided portrait editing with conditional GANs. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Long Beach, CA, USA, pp. 3431–3440. http://dx.doi.org/10.1109/CVPR.2019.00355.

Gu, P., Zhang, Y., Wang, C., Chen, D.Z., 2022. ConvFormer: Combining CNN and transformer for medical image segmentation. arXiv:2211.08564 [cs].

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv:1512.03385 [cs].

Huang, X., Wang, J., Tang, F., Zhong, T., Zhang, Y., 2018. Metal artifact reduction on cervical CT images by deep residual learning. Biomed. Eng. Online 17 (1), 175. http://dx.doi.org/10.1186/s12938-018-0609-y.

Ikuta, M., Zhang, J., 2022. A deep recurrent neural network with FISTA optimization for CT metal artifact reduction. IEEE Trans. Comput. Imaging 8, 961–971. http://dx.doi.org/10.1109/TCI.2022.3212825.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jiménez-Gaona, Y., Rodríguez-Alvarez, M.J., Escudero, L., Sandoval, C., Lakshminarayanan, V., 2024. Ultrasound breast images denoising using generative adversarial networks (GANs). Intell. Data Anal. 1–18. http://dx.doi.org/10.3233/IDA-230631.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. arXiv:1603.08155 [cs].

Johnson, J.P., Krupinski, E.A., Yan, M., Roehrig, H., Graham, A.R., Weinstein, R.S., 2011. Using a visual discrimination model for the detection of compression artifacts in virtual pathology images. IEEE Trans. Med. Imaging 30 (2), 306–314. http://dx.doi.org/10.1109/TMI.2010.2077308.

Kalender, W.A., Hebel, R., Ebersberger, J., 1987. Reduction of CT artifacts caused by metallic implants. Radiology 164 (2), 576–577. http://dx.doi.org/10.1148/radiology.164.2.3602406.

Kamachimudali, U., Sridhar, T., Raj, B., 2003. Corrosion of bio implants. Sadhana 28, 601–637.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training generative adversarial networks with limited data. arXiv:2006.06676 [cs, stat].

Ketcha, M.D., Marrama, M., Souza, A., Uneri, A., Wu, P., Zhang, X., Helm, P.A., Siewerdsen, J.H., 2021. Sinogram + image domain neural network approach for metal artifact reduction in low-dose cone-beam computed tomography. J. Med. Imaging 8 (05), http://dx.doi.org/10.1117/1.JMI.8.5.052103.

Kim, S., Ahn, J., Kim, B., Kim, C., Baek, J., 2022. Convolutional neural network–based metal and streak artifacts reduction in dental CT images with sparse-view sampling scheme. Med. Phys. 49 (9), 6253–6277. http://dx.doi.org/10.1002/mp.15884.

Kim, B., Lee, H., Kim, K.J., Seo, J., Park, S., Shin, Y., Kim, S.H., Lee, K.H., 2011. Comparison of three image comparison methods for the visual assessment of the image fidelity of compressed computed tomography images. Med. Phys. 38 (2), 836–844. http://dx.doi.org/10.1118/1.3538925.

Koike, Y., Anetai, Y., Takegawa, H., Ohira, S., Nakamura, S., Tanigawa, N., 2020. Deep learning-based metal artifact reduction using cycle-consistent adversarial network for intensity-modulated head and neck radiation therapy treatment planning. Phys. Medica 78, 8–14. http://dx.doi.org/10.1016/j.ejmp.2020.08.018.

Lee, D., Park, C., Lim, Y., Cho, H., 2020. A metal artifact reduction method using a fully convolutional network in the sinogram and image domains for dental computed tomography. J. Digit. Imaging 33 (2), 538–546. http://dx.doi.org/10.1007/s10278-019-00297-x.

Li, L., Chen, M., Shi, H., Duan, Z., Xiong, X., 2022a. Multiscale structure and texture feature fusion for image inpainting. IEEE Access 10, 82668–82679. http://dx.doi.org/10.1109/ACCESS.2022.3196021.

Li, Z., Tian, Q., Ngamsombat, C., Cartmell, S., Conklin, J., Filho, A.L.M.G., Lo, W., Wang, G., Ying, K., Setsompop, K., Fan, Q., Bilgic, B., Cauley, S., Huang, S.Y., 2022b. High-fidelity fast volumetric brain MRI using synergistic wave-controlled aliasing in parallel imaging and a hybrid denoising generative adversarial network (HDnGAN). Med. Phys. 49 (2), 1000–1014. http://dx.doi.org/10.1002/mp.15427.

Li, M., Zuo, W., Zhang, D., 2018. Deep identity-aware transfer of facial attributes. arXiv:1610.05586 [cs].

Liang, X., Zhang, H., Xing, E.P., 2017. Generative semantic manipulation with contrasting GAN. arXiv:1708.00315 [cs].

Liao, H., Lin, W., Zhou, S.K., Luo, J., 2020. ADN: Artifact disentanglement network for unsupervised metal artifact reduction. IEEE Trans. Med. Imaging 39 (3), 634–643. http://dx.doi.org/10.1109/TMI.2019.2933425.

Liebl, H., Schinz, D., Sekuboyina, A., Malagutti, L., Löffler, M.T., Bayat, A., El Husseini, M., Tetteh, G., Grau, K., Niederreiter, E., et al., 2021. A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data. Sci. Data 8 (1), 284.

Liu, Y., Su, S., Zhu, J., Zheng, F., Gao, L., Song, J., 2023. Allowing supervision in unsupervised deformable-instances image-to-image translation. IEEE Trans. Circuits Syst. Video Technol. http://dx.doi.org/10.1109/TCSVT.2023.3343733, 1–1.

Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S., 2020. A vertebral segmentation dataset with fracture grading. Radiol.: Artif. Intell. 2 (4), e190138.

Luo, S., Huang, F., 2024. MaGAT: Mask-guided adversarial training for defending face editing GAN models from proactive defense. IEEE Signal Process. Lett. 31, 969–973. http://dx.doi.org/10.1109/LSP.2024.3380466.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J., 2020. A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans. Multimed. 22 (10), 2597–2609. http://dx.doi.org/10.1109/TMM.2019.2958756.

Luo, Y., Zhang, S., Ling, J., Lin, Z., Wang, Z., Yao, S., 2024. Mask-guided generative adversarial network for MRI-based CT synthesis. Knowl.-Based Syst. 295, 111799. http://dx.doi.org/10.1016/j.knosys.2024.111799.

Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. arXiv:1611.04076 [cs].

Meyer, E., Raupach, R., Lell, M., Schmidt, B., Kachelrieß, M., 2010. Normalized metal artifact reductionmNMAR in computed tomography. Med. Phys. 37 (10), http://dx.doi.org/10.1118/1.3484090.

Meyer, E., Raupach, R., Lell, M., Schmidt, B., Kachelrieß, M., 2012. Frequency split metal artifact reduction (FSMAR) in computed tomography. Med. Phys. 39 (4), 1904–1916. http://dx.doi.org/10.1118/1.3691902.

Nakao, M., Imanishi, K., Ueda, N., Imai, Y., Kirita, T., Matsuda, T., 2020. Regularized three-dimensional generative adversarial nets for unsupervised metal artifact reduction in head and neck CT images. IEEE Access 8, 109453–109465. http://dx.doi.org/10.1109/ACCESS.2020.3002090.

Niu, C., Cong, W., Fan, F., Shan, H., Li, M., Liang, J., Wang, G., 2022. Low-dimensional manifold-constrained disentanglement network for metal artifact reduction. IEEE Trans. Radiat. Plasma Med. Sci. 6 (6), 656–666. http://dx.doi.org/10.1109/TRPMS.2021.3122071.

Nobari, A.H., Rashad, M.F., Ahmed, F., 2021. CreativeGAN: Editing generative adversarial networks for creative design synthesis. arXiv:2103.06242 [cs].

Pan, X., Sidky, E.Y., Vannier, M., 2009. Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction? Inverse Problems 25 (12), 123009. http://dx.doi.org/10.1088/0266-5611/25/12/123009.

Park, H.S., Lee, S.M., Kim, H.P., Seo, J.K., Chung, Y.E., 2018. CT sinogram-consistency learning for metal-induced beam hardening correction. Med. Phys. 45 (12), 5376–5384. http://dx.doi.org/10.1002/mp.13199.

Paszke, A., 2019. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

Peng, C., Li, B., Li, M., Wang, H., Zhao, Z., Qiu, B., Chen, D.Z., 2020a. An irregular metal trace inpainting network for X-ray CT metal artifact reduction. Med. Phys. 47 (9), 4087–4100. http://dx.doi.org/10.1002/mp.14295.

Peng, C., Li, B., Liang, P., Zheng, J., Zhang, Y., Qiu, B., Chen, D.Z., 2020b. A cross-domain metal trace restoring network for reducing X-ray CT metal artifacts. IEEE Trans. Med. Imaging 39 (12), 3831–3842. http://dx.doi.org/10.1109/TMI.2020.3005432.

Rouse, D.M., Hemami, S.S., 2008. Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. In: Rogowitz, B.E., Pappas, T.N. (Eds.), Proc. SPIE 6806, Human Vision and Electronic Imaging XIII. SPIE, San Jose, CA, 680615. http://dx.doi.org/10.1117/12.768060.

Sakamoto, M., Hiasa, Y., Otake, Y., Takao, M., Suzuki, Y., Sugano, N., Sato, Y., 2019. Automated segmentation of hip and thigh muscles in metal artifact contaminated CT using CNN. In: Lin, F., Fujita, H., Kim, J.H. (Eds.), In: International Forum on Medical Imaging in Asia 2019, vol. 11050, International Society for Optics and Photonics, SPIE, p. 110500S. http://dx.doi.org/10.1117/12.2521440.

Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., et al., 2021. VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. Med. Image Anal. 73, 102166.

Selles, M., Van Osch, J.A., Maas, M., Boomsma, M.F., Wellenberg, R.H., 2024. Advances in metal artifact reduction in CT images: A review of traditional and novel metal artifact reduction techniques. Eur. J. Radiol. 170, 111276. http://dx.doi.org/10.1016/j.ejrad.2023.111276.

Shi, Z., Wang, N., Kong, F., Cao, H., Cao, Q., 2022. A semi-supervised learning method of latent features based on convolutional neural networks for CT metal artifact reduction. Med. Phys. 49 (6), 3845–3859. http://dx.doi.org/10.1002/mp.15633.

Skandarani, Y., Jodoin, P., Lalande, A., 2023. GANs for medical image synthesis: An empirical study. J. Imaging 9 (3), 69. http://dx.doi.org/10.3390/jimaging9030069.

Soltanian-Zadeh, H., Windham, J.P., Soltanianzadeh, J., 1996. CT artifact correction: an image-processing approach. In: Loew, M.H., Hanson, K.M. (Eds.), In: Medical Imaging 1996: Image Processing, vol. 2710, International Society for Optics and Photonics, SPIE, pp. 477–485. http://dx.doi.org/10.1117/12.237950.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2018. Mask-guided contrastive attention model for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, pp. 1179–1188. http://dx.doi.org/10.1109/CVPR.2018.00129.

Tang, H., Liu, H., Xu, D., Torr, P.H.S., Sebe, N., 2021. AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks. arXiv:1911.11897 [cs, eess].

Vaishnavi, P., Cong, T., Eykholt, K., Prakash, A., Rahmati, A., 2019. Can attention masks improve adversarial robustness? arXiv arXiv:1911.11946.

Wan, Z., Wang, Y., Yong, S., Zhang, P., Stepputtis, S., Sycara, K., Xie, Y., 2024. Sigma: Siamese mamba network for multi-modal semantic segmentation. arXiv:2404.04256 [cs].

Wang, H., Li, Y., He, N., Ma, K., Meng, D., Zheng, Y., 2022a. DICDNet: Deep interpretable convolutional dictionary network for metal artifact reduction in CT images. IEEE Trans. Med. Imaging 41 (4), 869–880. http://dx.doi.org/10.1109/TMI.2021.3127074.

Wang, H., Li, Y., Zhang, H., Chen, J., Ma, K., Meng, D., Zheng, Y., 2021a. InDuDoNet: An interpretable dual domain network for CT metal artifact reduction. arXiv:2109.05298 [eess].

Wang, H., Li, Y., Zhang, H., Meng, D., Zheng, Y., 2023. InDuDoNet+: A deep unfolding dual domain network for metal artifact reduction in CT images. Med. Image Anal. 85, 102729. http://dx.doi.org/10.1016/j.media.2022.102729.

Wang, T., Lu, Z., Yang, Z., Xia, W., Hou, M., Sun, H., Liu, Y., Chen, H., Zhou, J., Zhang, Y., 2022b. IDOL-Net: An interactive dual-domain parallel network for CT metal artifact reduction. IEEE Trans. Radiat. Plasma Med. Sci. 6 (8), 874–885. http://dx.doi.org/10.1109/TRPMS.2022.3171440.

Wang, Q., Makarenko, M., 2021. SDA-GAN: Unsupervised image translation using spectral domain attention-guided generative adversarial network. arXiv:2110.02873 [cs].

Wang, Z., Simoncelli, E., Bovik, A., 2003. Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. IEEE, Pacific Grove, CA, USA, pp. 1398–1402. http://dx.doi.org/10.1109/ACSSC.2003.1292216.

Wang, Z., Vandersteen, C., Demarcy, T., Gnansia, D., Raffaelli, C., Guevara, N., Delingette, H., 2021b. Inner-ear augmented metal artifact reduction with simulation-based 3D generative adversarial networks. Comput. Med. Imaging Graph. 93, 101990. http://dx.doi.org/10.1016/j.compmedimag.2021.101990.

Wang, T., Xia, W., Huang, Y., Sun, H., Liu, Y., Chen, H., Zhou, J., Zhang, Y., 2021c. DAN-Net: Dual-domain adaptive-scaling non-local network for CT metal artifact reduction. Phys. Med. Biol. 66 (15), 155009. http://dx.doi.org/10.1088/1361-6560/ac1156.

Xu, L., Zeng, X., Li, W., Zheng, B., 2023. MFGAN: Multi-modal feature-fusion for CT metal artifact reduction using GANs. ACM Trans. Multimed. Comput. Commun. Appl. 19 (1s), 1–17. http://dx.doi.org/10.1145/3528172.

Yu, L., Zhang, Z., Li, X., Ren, H., Zhao, W., Xing, L., 2021. Metal artifact reduction in 2D CT images with self-supervised cross-domain learning. Phys. Med. Biol. 66 (17), 175003. http://dx.doi.org/10.1088/1361-6560/ac195c, arXiv:2109.13483 [physics].

Yu, L., Zhang, Z., Li, X., Xing, L., 2021b. Deep sinogram completion with image prior for metal artifact reduction in CT images. IEEE Trans. Med. Imaging 40 (1), 228–238. http://dx.doi.org/10.1109/TMI.2020.3025064.

Zhang, T., Wiliem, A., Yang, S., Lovell, B., 2018. TV-GAN: Generative adversarial network based thermal to visible face recognition. In: 2018 International Conference on Biometrics. ICB, IEEE, Gold Coast, QLD, pp. 174–181. http://dx.doi.org/10.1109/ICB2018.2018.00035.

Zhang, Y., Yu, H., 2018. Convolutional neural network based metal artifact reduction in X-ray computed tomography. IEEE Trans. Med. Imaging 37 (6), 1370–1381. http://dx.doi.org/10.1109/TMI.2018.2823083.

Zhu, L., Han, Y., Li, L., Xi, X., Zhu, M., Yan, B., 2019. Metal artifact reduction for X-ray computed tomography using U-net in image domain. IEEE Access 7, 98743–98754. http://dx.doi.org/10.1109/ACCESS.2019.2930302.

Zhu, L., Han, Y., Xi, X., Li, L., Yan, B., 2021b. Completion of metal-damaged traces based on deep learning in sinogram domain for metal artifacts reduction in CT images. Sensors 21 (24), 8164. http://dx.doi.org/10.3390/s21248164.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2020. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593 [cs].

Zhu, J., Yang, G., Lio, P., 2019a. How can we make gan perform better in single medical image super-resolution? A lesion focused multi-scale approach. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, Venice, Italy, pp. 1669–1673. http://dx.doi.org/10.1109/ISBI.2019.8759517.

Zou, D., Li, W., Deng, C., Du, G., Xu, N., 2019. The use of CT hounsfield unit values to identify the undiagnosed spinal osteoporosis in patients with lumbar degenerative diseases. Eur. Spine J. 28, 1758–1766.

Roberto Maria Scardigno received the B.Sc. degree and the M.Sc. degree (cum laude) in Medical Systems Engineering from Polytechnic University of Bari, Italy, in 2020 and 2022, respectively. During his master thesis, he developed a Serious Game to assess the Sense of Agency in children with Cerebral Palsy. Since November 2022, he is a Ph.D. Student in Autonomous Systems at the Industrial Laboratory, Department of Electrical and Information Engineering, Polytechnic University of Bari. His research interests are focused on intelligent systems for automated diagnosis in the industrial and biomedical sectors.

Antonio Brunetti is an assistant professor in Bioengineering at the Department of Electrical and Information Engineering at the Polytechnic University of Bari. He received a Bachelor's Degree in Computer Science and Automation Engineering, and a Master's Degree (cum laude) in Computer Science Engineering, from the Polytechnic University of Bari. In February 2020, he obtained his Ph.D. in Electrical and Information Engineering from the Doctoral School of the Polytechnic University of Bari. His research activities focus on Electronic and Informatics Bioengineering, Bioinformatics, and Intelligent Frameworks, based on Image Processing, Machine Learning, and Deep Learning algorithms, to support clinical diagnosis.

Pietro Maria Marvulli received the B.Sc. and the M.Sc. degree (cum laude) in Medical Systems Engineering, in 2020 and 2023 from Polytechnic of Bari, respectively. For his master's thesis he developed an algorithm for the removal of metal artifacts from CT images. Since November 2023, he is a Ph.D. Student in Autonomous Systems at the Industrial Informatics Laboratory, DEI, Polytechnic University of Bari. His research is focused on robotic systems for minimally invasive and interventional surgery.

Raffaele Carli received his Laurea in Electronic Engineering (2002) and Ph.D. in Electrical and Information Engineering (2016) from Politecnico di Bari, Italy. He served as a Reserve Officer with the Italian Navy (2003-2004) and worked as a System and Control Engineer and Technical Manager in a space and defense company (2004–2012). Dr. Carli is now an Assistant Professor in Automatic Control at Politecnico di Bari. His research focuses on decision and control systems, and complex system modeling and optimization. He is an Associate Editor for IEEE journals and has authored over 100 international publications.

Mariagrazie Dotoli received her Laurea (1995) and Ph.D. (1999) in Electrical Engineering from Politecnico di Bari, Italy. She is a Full Professor in Automatic Control at Politecnico di Bari, which she joined in 1999. She has served as Vice Rector for Research and a member of the Academic Senate. Her research focuses on discrete event systems, manufacturing, logistics, traffic networks, smart grids, and networked systems. She is an editor for IEEE journals, has authored over 200 international publications, and holds significant organizational roles in various conferences, including serving as General Chair for the 2024 IEEE Conference on Automation Science and Engineering.

Vitoantonio Bevilacqua is Full Professor of Bioengineering at the Electrical and Information Engineering Department of Polytechnic University of Bari and the Coordinator of the Master's Degree in Medical Systems Engineering. He obtained the Laurea Degree in Electronic Engineering, the Ph.D. in Electrical Engineering from Polytechnic University of Bari. Currently he is the Head of Industrial Informatics Laboratory. Since 1996 he has been working and investigating in the field of computer vision and image processing, bioengineering, human–machine interaction based on machine learning and soft computing techniques (neural networks, evolutionary algorithms, hybrid expert systems, deep learning).

Domenico Buongiorno is an assistant professor in Bioengineering at the Department of Electrical and Information Engineering at the Polytechnic University of Bari, where he teaches Biomedical Instrumentation. He earned his Bachelor's and Master's degrees in Control System Engineering from the Polytechnic University of Bari in 2011 and 2014, respectively. In 2017, he obtained his Ph.D. in Emerging Digital Technologies from the Sant'Anna School of Advanced Studies in Pisa. Until July 2022, he worked as a postdoctoral research fellow in the field of electronic and computer bioengineering. His research activity focuses on intelligent systems for diagnosis and therapy.