JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Received 17 October 2025
Accepted 24 April 2026

Edited by A. Momose, Tohoku University, Japan

Keywords: virtual staining; 3D X-ray histology; virtual histology; biodegradable bone implants; paired CycleGAN.

Supporting information: this article has supporting information at journals.iucr.org/s

Paired CycleGAN-based virtual staining for 3D X-ray histology of bone-implant systems

Sarah C. Irvine,[a]* Christian Lucas,[b,c] Diana Krüger,[b] Bianca C. Guedert,[d] Julian Moosmann[a] and Berit Zeller-Plumhoff[b,d]*

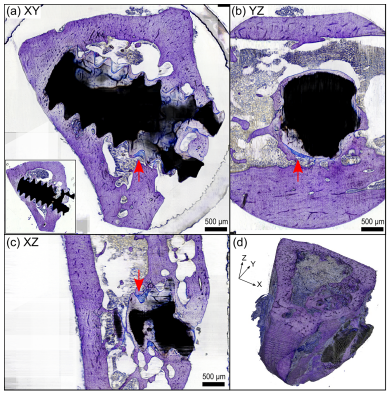
[a]Institute of Materials Physics, Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany, [b]Institute of Metallic Biomaterials, Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany, [c]Bruker Daltonics SPR, 20251 Hamburg, Germany, and [d]Data-Driven Analysis and Design of Materials, University of Rostock, 18059 Rostock, Germany. *Correspondence e-mail: sarah.irvine@hereon.de, berit.zeller-plumhoff@uni-rostock.de

Three-dimensional X-ray histology offers a non-invasive alternative to conventional 2D histology, enabling volumetric imaging of biological tissues without physical sectioning or chemical staining. However, the intrinsic grey-scale contrast of X-ray tomography limits its biochemical specificity compared with traditional histological stains. In this study, we extend deep-learning-based virtual staining to the X-ray domain via cross-modality image translation to generate artificially stained slices directly from synchrotron radiation micro-tomography (μCT) scans. Using over 50 co-registered pairs of μCT and toluidine blue-stained histology from bone-implant samples, we trained a modified CycleGAN network tailored for limited paired data. Whole-slide histology images were downsampled to the CT voxel size, with on-the-fly data augmentation for patch-based training. The model incorporates pixelwise supervision and greyscale consistency losses, enabling histologically realistic colour outputs while preserving structural detail. Results outperformed Pix2Pix and standard CycleGAN baselines across metrics of structural similarity, perceptual fidelity, and peak signal-to-noise ratio. Once trained, the model can be applied to full μCT volumes to produce virtually stained 3D datasets that enhance interpretability without additional sample preparation. This work introduces virtual staining to 3D X-ray imaging, which may provide a scalable route for chemically informative, label-free tissue characterization in biomedical research.

1. Introduction

Histological analysis is widely regarded within clinical pathology and biomedical research as the benchmark technique for characterizing tissue architecture and cellular detail. Through chemical staining of thin sections and brightfield microscopy, conventional histology provides high specificity in identifying cell types, pathological states, and tissue organization. However, this process is inherently destructive, labour-intensive, and fundamentally two-dimensional (Abraham & Levenson, 2024). While serial sectioning can be used to reconstruct three-dimensional structures, it remains time-consuming, prone to alignment artefacts, and yields anisotropic spatial resolution due to finite section thickness and non-contiguous through-plane sampling (Pichat *et al.*, 2018). As a result, conventional histology captures only limited volumetric context, posing significant challenges for studying complex 3D tissue and biomaterial interactions, particularly in whole organs or biomedical implants.

Overcoming those limitations, three-dimensional (3D) X-ray imaging techniques such as synchrotron radiation X-ray micro-tomography (SR-μCT) and high-resolution laboratory μCT present as powerful alternatives. These methods enable



label-free, non-invasive visualization of soft and mineralized tissues at micrometre-scale resolution. When applied correlatively with histology for reference, such approaches may be collectively termed '3D X-ray histology' or 'X-ray virtual histology' (Albers *et al.*, 2018; Töpperwien *et al.*, 2018; Katsamenis *et al.*, 2019). By providing volumetric datasets that retain cellular- and tissue-level detail across intact specimens, these techniques have the potential to support a transition toward 3D pathology, enabling more comprehensive characterization of morphology across entire tissue volumes (Song *et al.*, 2024). However, despite these advantages, a central limitation of 3D X-ray histology is the lack of biochemical specificity: image contrast is intrinsically greyscale, derived from differences in X-ray attenuation or phase shift, and does not replicate the cell-type or structure-specific colour-coded information provided by traditional histological stains. While X-ray contrast agents based on high-Z elements can be applied to enhance soft tissue visibility, these typically bind to tissue components differently than conventional stains, and the issue of bio-specificity remains unresolved. Some progress has been made through the development of modified histological stains that incorporate heavy metal complexes, offering improved compatibility with traditional staining protocols (Müller *et al.*, 2018; Petzold *et al.*, 2024). However, the heavy-metal related issues of toxicity, environmental safety, and clinical applicability are of some concern. As a result, computational methods that enhance the interpretability of raw X-ray volumes without the need for physical staining represent a valuable and increasingly important advancement. In one early X-ray histology study (Khimchenko *et al.*, 2016), a joint histogram-based analysis was used to map greyscale CT intensities to histological colour values, enabling stain-like colourization of 3D volumes. However, this approach has seen limited uptake in subsequent research.

Recent advances in deep learning, in particular the development of generative adversarial networks (GANs), have transformed the field of computational image synthesis, including applications of digital pathology. GANs are composed of two competing neural networks: a generator that synthesizes images and a discriminator that evaluates their realism, trained adversarially to produce outputs indistinguishable from ground-truth data (Goodfellow *et al.*, 2014). In medical imaging, GANs have been applied to tasks such as segmentation, super-resolution, denoising, artefact reduction, and, notably, cross-modality image-to-image translation, where a transformation is learned between different imaging modalities (Yi *et al.*, 2019). Examples include synthesizing computed tomography (CT) from magnetic resonance imaging (MRI) or positron emission tomography (PET) from CT, which are clinically valuable when certain modalities are limited by cost, radiation exposure or accessibility.

'Virtual staining' refers to a family of deep-learning approaches that are becoming increasingly established in computational histopathology research, whereby label-free microscopy images are transformed into the visual equivalent of chemically stained histological images, for enhanced interpretability without the need for physical staining (Latonen *et al.*, 2024). These primarily GAN-based techniques have demonstrated high-fidelity image-to-image translation, applied within various optical imaging modalities including autofluorescence (Rivenson *et al.*, 2019), photo-acoustic microscopy (Kang *et al.*, 2022) and quantitative phase imaging (QPI) (Abraham *et al.*, 2022), as well as standard brightfield microscopy (Koivukoski *et al.*, 2023). While some of these optical approaches permit limited volumetric acquisition without physical sectioning, they typically require either thin (sub-mm) specimens or optically transparent samples via tissue clearing. To date, virtual staining has not yet been extended to the X-ray domain, where the image formation mechanisms and contrast properties differ significantly. In this study, we introduce a deep-learning model for virtual staining in 3D X-ray histology, enabling direct synthesis of virtually stained 3D whole tissue volumes from raw greyscale µCT scans. This cross-modality translation aims to combine the structural integrity of X-ray imaging with histologically meaningful contrast.

For the base model architecture, we consider as reference two foundational models in image-to-image translation: Pix2Pix (Isola *et al.*, 2017) and CycleGAN (Zhu *et al.*, 2017), representing supervised and unsupervised approaches, respectively. Supervised models like Pix2Pix require spatially aligned paired datasets, which are often difficult or even impossible to obtain in medical imaging due to motion, resolution mismatch or other acquisition constraints. CycleGAN addresses this limitation through a cycle-consistency loss that enables training with unpaired data. In earlier applications of GANs in medical imaging, Pix2Pix, CycleGAN and variants thereof comprised a majority of cross-modality synthesis tasks (Yi *et al.*, 2019), although as the field has progressed there has been a shift towards increasingly task-specific GAN models tailored to address the unique challenges of modality translation and clinical relevance (Heng *et al.*, 2024). CycleGAN remains the most widely adopted unsupervised approach in virtual staining, where fully aligned datasets are rarely available (Latonen *et al.*, 2024). In the related context of virtual stain transfer, where one histochemical stain is virtually transformed into another, CycleGAN was found to outperform other architectures, including Pix2Pix, even in settings where simulated paired data were available (Zingman *et al.*, 2024).

Pix2Pix may still yield superior results when perfect alignment between modalities is available, whereas the unsupervised CycleGAN is ideal for situations with no assumed alignment. Many real-world medical imaging scenarios, however, fall between these extremes, with some degree of partial misalignment. This intermediate setting has motivated adaptations of CycleGAN for use with paired data, retaining its robust architecture while benefiting from supervision. The added stability offered by cycle-consistency constraints encourages learning of mappings that are reversible and coherent across imperfectly aligned image domains (Kaji & Kida, 2019). Several studies have explored such paired CycleGAN frameworks, including work on cone-beam CT correction and dual-energy chest X-ray imaging, where mis-

aligned data are actively corrected (Harms *et al.*, 2019; Ueda *et al.*, 2025), as well as MRI-to-CT synthesis (Lei *et al.*, 2019) and contrast-enhanced mammography (Rofena *et al.*, 2024). These studies illustrate the flexibility of CycleGAN variants in modality translation tasks where precise pixel-wise alignment is lacking but a high level of spatial correspondence is preserved. Our work adopts a similar approach, employing a paired CycleGAN framework adapted to the X-ray and histology domains, and incorporating additional supervisory loss terms to improve structural fidelity in domain mapping. We evaluate the performance of this modified CycleGAN in comparison with both the standard CycleGAN and Pix2Pix models.

To demonstrate the potential of deep-learning-based virtual staining in 3D X-ray histology, we apply our methodology to biodegradable magnesium-based implants in bone tissue. This is a challenging biomaterials system previously examined in multiple correlative studies of alloy degradation and osseointegration using a range of electron and X-ray microscopy techniques alongside conventional histology (Krüger *et al.*, 2021; Sefa *et al.*, 2023; Iskhakova *et al.*, 2024). In Krüger *et al.* (2022), comparisons of bone-implant samples at different healing intervals revealed that magnesium alloys gradually form a stable degradation layer, surrounded by newly formed non-mineralized bone. While histology confirmed the presence of these features, SR-µCT combined with a U-Net convolutional neural network (Baltruschat *et al.*, 2021) enabled tomographic segmentation, allowing 3D visualization and volume-based quantification of key parameters such as bone-implant contact and degradation rate. Although corresponding 2D histology and 3D µCT measurements were correlated, discrepancies in detail highlight this system as a strong potential candidate for 3D virtual staining. This poses the primary question of whether virtual staining can accurately infer stain-specific contrast from greyscale µCT volumes, thus enabling virtual sectioning of the implant and surrounding tissue in arbitrary orientations without additional labelling, sectioning, tissue clearing, or decalcification. For this investigation, to facilitate supervised learning, we collated and co-registered paired datasets from such prior studies, comprising upwards of 50 SR-µCT slices and their corresponding stained histological sections. An exemplary pair is shown in Fig. 1, annotated with key sample features referenced in the text.

2. Methods and materials

Collected imaging datasets from the aforementioned correlative characterization studies of bone-implant systems were reused in this computational project, pertaining to samples with explants of magnesium–gadolinium screws (Mg–5wt%Gd and Mg–10wt%Gd), titanium (Ti) or polyether-ether-kethone (PEEK) screws implanted into Sprague Dawley rat tibia with healing periods of 4, 8 or 12 weeks. The Ti and PEEK screws were used as control materials to evaluate osseointegration without degradation of the implant. Animal experiments were conducted after ethical approval by the ethical committee at the Malmo/Lund regional board for animal research, Swedish Board of Agriculture (approval number DNR M 188-15). For a comprehensive description of the sample preparation methodology which includes the initial alloy production and animal study details, refer to Krüger *et al.* (2022) and references therein.

Following the sample preparation of the bone-implant block, the methodology timeline is illustrated schematically in Fig. 2 which comprises the two imaging modalities followed by global registration to create the image pairs required for the supervised GAN inputs. Note that the histology must be performed after the tomography due to the destructive nature of the histological sectioning process. The second half of the

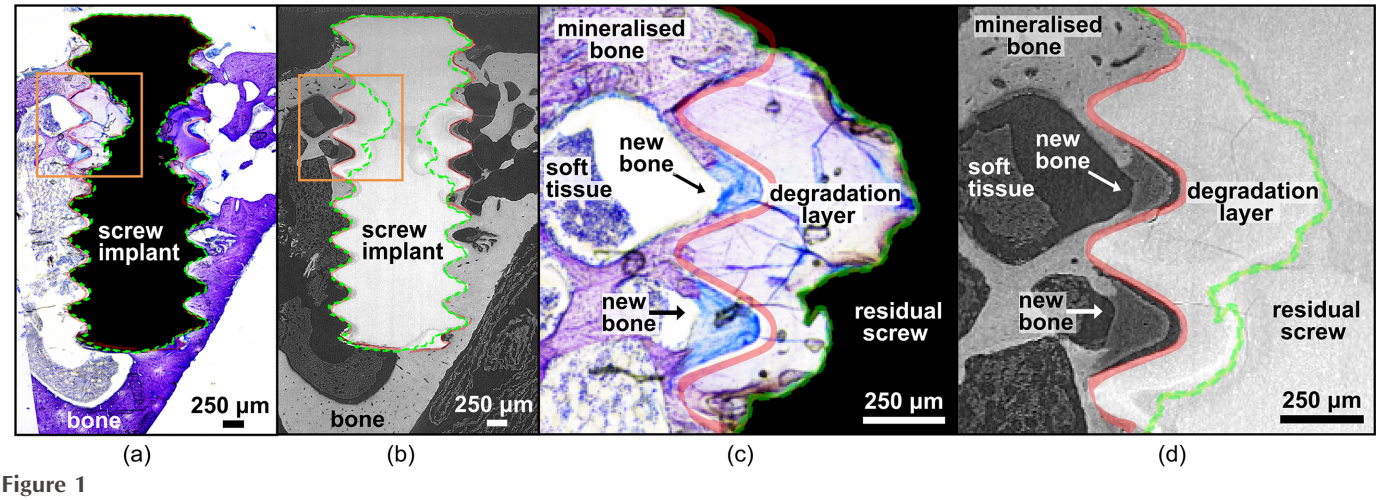


Figure 1
Illustrative example of our X-ray-histology application, with (real) paired toluidine-blue stained histology (*a*) and SR-µCT (*b*) images of an Mg–10Gd screw implant in bone extracted after eight weeks *in vivo*. In the histology, the green line marks the interface between the residual screw and degradation layer. In the µCT slice, the red line indicates the boundary between the degradation layer and surrounding bone. The residual screw appears black in histology due to total light attenuation by the metal. Regions identified as the degradation layer in histology correspond to areas in µCT with mildly differing X-ray attenuation to the residual screw. Higher magnification views (*c*, *d*) indicated by the orange boxes highlight the areas surrounding the degradation layer. Arrows in black (*c*) or white (*d*) point to regions of new (woven) bone, appearing blue in histology and exhibiting lower attenuation values in µCT due to lower mineralization content. Figure adapted from Krüger *et al.* (2022) under a CC-BY-4.0 licence.

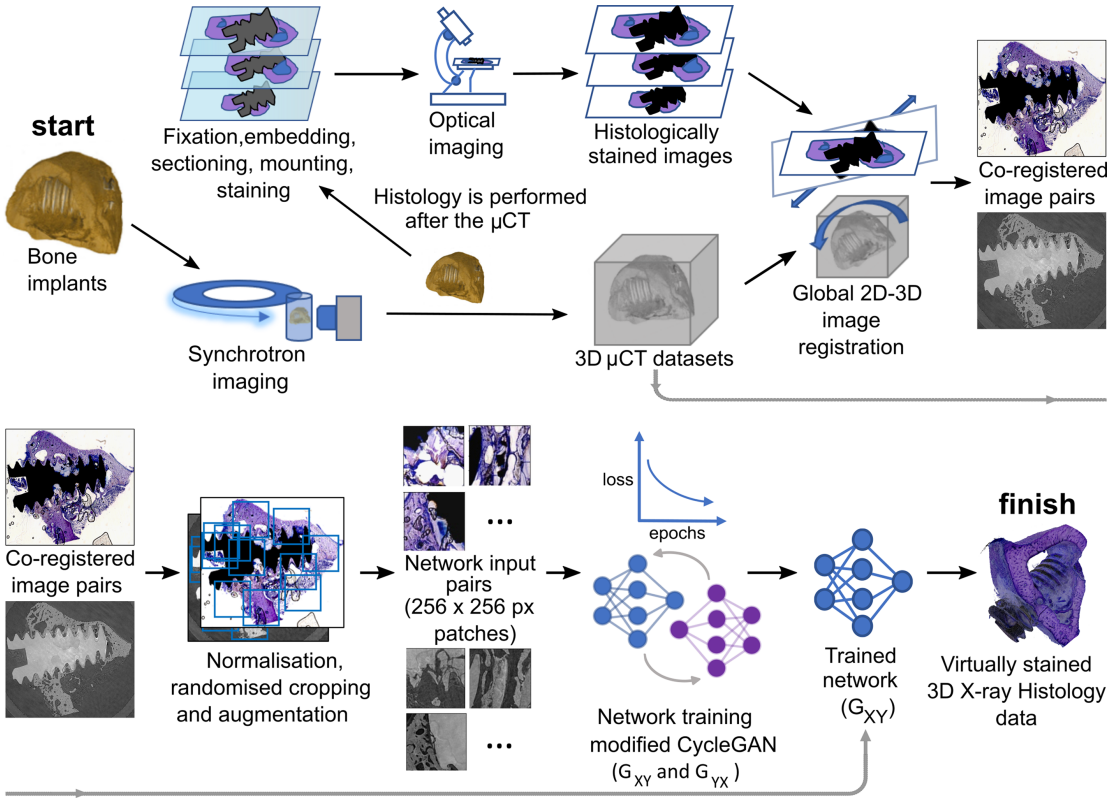


Figure 2
Schematic of the methodology for 3D virtual staining in X-ray histology of bone implants, using a paired CycleGAN network.

methodology describes the investigated models, the data and its treatment including normalization, augmentation and use of a sample correspondence mask, as well as the metrics applied for comparative 2D analysis. Finally, we demonstrate the trained model in a 3D application by processing a stack of µCT slices into a virtually stained 3D X-ray histology volume.

2.1. Sample preparation

For the SR-µCT imaging, samples consisted of bone-implant blocks comprising screw (4 mm length, 2 mm in diameter and M2 thread) and surrounding bone tissue, explanted with a trephine bur of 6 mm diameter. The bone-implant blocks were fixed in 70% ethanol for one or more days and then dehydrated in a graded series of ethanol (samples were critically point dried). The dried blocks were put into a standard Eppendorf tube (with arbitrary orientation) before being placed in the SR X-ray beam for tomographic imaging.

For the histology, which occurred after the SR-µCT imaging, explants were re-infiltrated with absolute ethanol and then embedded in methyl methacrylate resin by LLS Rowiak LaserLabSolutions GmbH (Hanover, Germany). Each sample was cut in half along the screw longitudinal axis with an Exakt saw, and then prepared for non-decalcified histology with the cutting-grinding technique *ad modum* Donath (1988). Sections of about 40 µm were obtained, mounted on glass slides and stained (Histlab, Göteborg, Sweden) with a solution of toluidine-blue/pyronine-Y. Toluidine blue solution is widely used as a staining method for undecalcified bone tissue, allowing for identification of the mineralized bone matrix, osteoids and soft tissues (Peev *et al.*, 2024). When combined with pyronin Y it results in staining of the bone tissue in various shades of purple (darker corresponding to younger bone) while the soft tissue is stained with more of a blue tint (Sarve *et al.*, 2007). New woven bone at the degradation layer interface may also be stained bright blue (Krüger *et al.*, 2022).

As a smaller secondary dataset, we also collated histology samples that were stained with hematoxylin and eosin (H&E) (LLS Rowiak, Hannover Germany). Sections were laser-cut by LLS Rowiak in a process which excludes the residual screw implant from the final mounted tissue sections (whilst retaining a portion of the screw degradation layer material). H&E is the most widely used stain across all cell types; in bone it results in a deep pink, almost red stain for the mineralized tissue and purple staining of the soft tissue. Examples of this less complete dataset are presented in Section S2 of the supporting information.

2.2. Imaging

The SR-µCT datasets were acquired over several beam times at the imaging beamline (IBL) P05, which is operated by the Helmholtz-Zentrum Hereon, at the PETRA III storage ring of the Deutsches Elektronen Synchrotron (DESY) in Hamburg, Germany (Wilde *et al.*, 2016). The primary form of contrast is attenuation-based, with minimal in-line phase contrast, acquired in full-field transmission geometry. As the beam times were part of different studies, a range of various

acquisition parameters were utilized, including X-ray energies (ranging from 25 to 45 keV to accommodate the attenuation range of the different screw alloy materials, *i.e.*, PEEK, Mg and Ti) and cameras (CCD versus CMOS) coupled to the X-ray microscope. Further details are given by Krüger *et al.* (2022). Volumes of approximately 6 mm × 6 mm × 6 mm were scanned, with effective pixel sizes of either 1.2 or 2.4 µm. Tomograms were reconstructed using a MATLAB-based framework and via *ASTRA toolbox* for tomographic back-projection. Following tomographic reconstruction all datasets were downsampled to a voxel size of 5 µm (1.5k × 1.5k pixel slices).

The collected histological images were also obtained over multiple measurement periods. All stained sections were imaged with one of two white-light optical microscopes coupled to a camera (Optical microscope 1: Nikon Eclipse Ci-L and DS-Fi3 camera controlled by NIS-Elements software, Tokyo, Japan; Optical microscope 2: Zeiss AxioCam MrC controlled by AxioVision software, Oberkochen, Germany). Images were acquired with nominal magnifications of 5×, 10× or 20× in wide field mode (for effective pixel sizes of 0.88 µm, 0.44 µm or 0.22 µm, respectively), and whole slide imaging (WSI) was achieved via guided manual stepping of the sample stage together with the inbuilt automatic image tiling function. Some sample images may be described as a partial WSI centred on the screw (approximate field of view is 2.5 mm × 4.5 mm), while others contain the full extent of sectioned screw plus bone area (roughly 6 mm × 6 mm). There is no consistent final array size with typical dimensions ranging from 6k to 13k pixels. Following registration (described in the next section), the transformed WSI images were downscaled to match the pixel size of the µCT slice pair (5 µm).

2.3. Registration

SR-µCT and histology WSI image pairs were co-registered using a semi-automated global 2D–3D registration pipeline that has been custom-written in Python, outlined by Irvine *et al.* (2024). In an iterative process based on optimizing the mutual information metric, a 3D rigid transformation is applied to the µCT dataset to find the optimal virtual plane fit to the histology image. This is followed by an affine transformation of the histology image, incorporating non-uniform scaling and local shear to compensate for minor distortions introduced during histological sample preparation after the tomography measurements. Due to the rigidity of the mineralized bone, deformation components were small, with shear parameters of <1% for the toluidine blue-stained samples and up to 3–4% for the H&E-stained samples. The pixel size of the transformed output registered histology/µCT image pair is the same as the input µCT dataset. Finally, both transformed slices were cropped to fit the minimum field of view of the pair.

2.4. Datasets

A total of 53 co-registered µCT and toluidine-blue stained histology WSI pairs were collated for this project. These comprised 38 samples containing Mg (-based implants), five containing Ti and ten containing PEEK. We performed a five-fold cross-validation with a 40/10 train/validation split rotated through five times. Each validation set was randomly selected whilst keeping a representative 7:1:2 ratio of Mg:Ti:PEEK samples. The three WSI pairs for testing (comprising two Mg and one PEEK) were kept separate and never involved in training in order to prevent data leakage.

Corresponding details of the secondary H&E-stained dataset are given in Section S2.1 of the supporting information.

2.5. Normalization

Normalization was applied to all histology and µCT image pairs to mitigate domain shift. As the data were compiled from multiple measurement sessions and datasets, they exhibit substantial variation in image quality and statistics. For both modalities, regions of interest (ROIs) of both rigid bone and air/background were sampled to define scaling bounds. With the µCT data, intensities were linearly scaled to set the mean values of bone and background (30000 and 10000, respectively, for a 16-bit greyscale image file). For the histology RGB images, a simple white-balance correction was applied via the mean sampled background value $\mu_{R,G,B}$ (such that $\mu_R = \mu_G = \mu_B$), followed by contrast-stretching. Distribution bounds were defined by the first percentile of the bone ROI intensity distribution i ($f_{min} = \min_{R,G,B}(i) - c$, clamped to 10) and the 99th percentile of the background ROI distribution j ($f_{max} = \max_{R,G,B}(j) + c'$, clamped to 255 for 8-bit images). The same constants c and c' were applied across all channels to preserve chromaticity.

2.6. Network models

2.6.1. CycleGAN and its adaptation for paired data

Our chosen model is based on the CycleGAN framework and adapted for paired data. We worked with a *PyTorch* implementation of the original CycleGAN model (Zhu *et al.*, 2017) as developed by Linder-Norén (2019). A schematic of our modified model is illustrated in Fig. 3.

The base model employs two generator networks: G_{XY}, which learns to translate images from the source domain X (comprising µCT scans) to the target domain Y (histology images), and G_{YX}, which performs the inverse mapping. Each generator is trained adversarially against a corresponding discriminator (D_Y and D_X), which learns to distinguish real from generated images.

Since in the original case no direct correspondence was assumed to exist between images in the two domains, the cycle-consistency constraint is introduced to ensure that an input image can be approximately recovered after translation and back-translation. Specifically, given an image $x \in X$, the composition $G_{YX}[G_{XY}(x)]$ should closely match the original x. This acts as a form of self-supervision, allowing the network to learn meaningful mappings even without paired data. The cycle-consistency loss is defined as

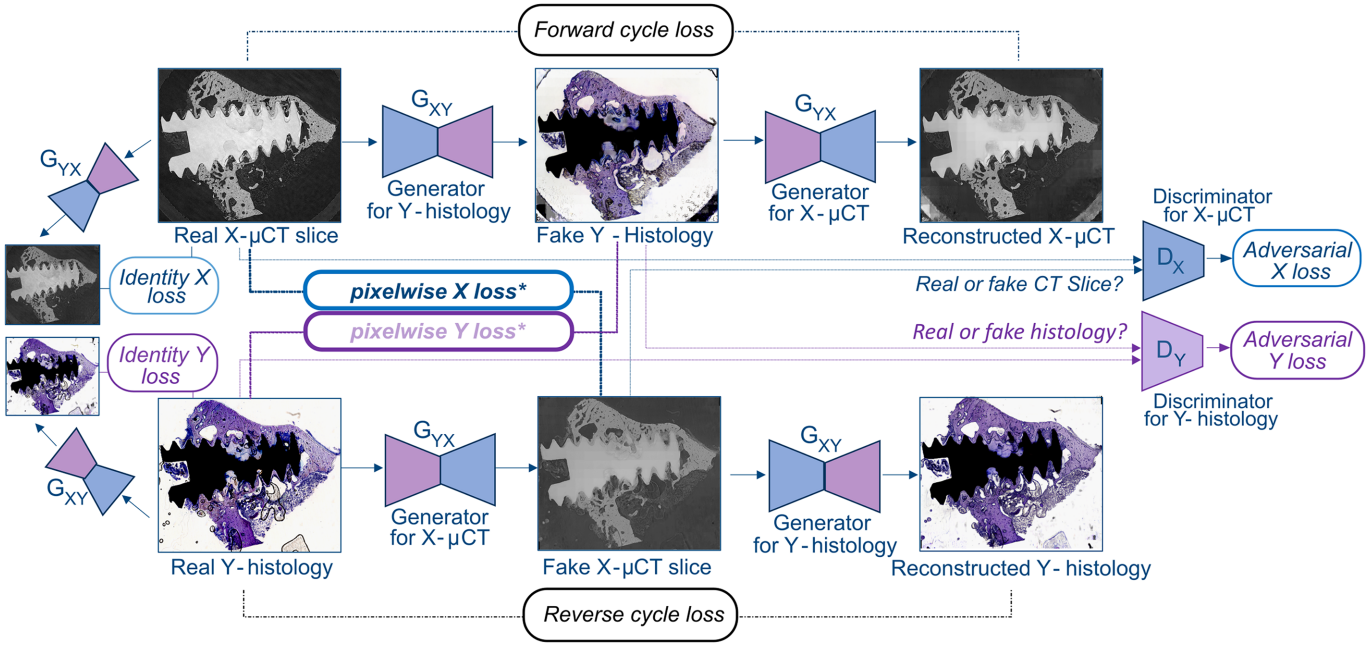


Figure 3
The modified CycleGAN for paired data, with additional pixelwise supervision loss terms that directly penalise differences between the generator output and the target image for each domain. Not shown: an additional greyscale loss term applicable to the generated µCT data only.

$$\mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX}) = \mathbb{E}_x\{\|G_{YX}[G_{XY}(x)] - x\|_1\} + \mathbb{E}_y\{\|G_{XY}[G_{YX}(y)] - y\|_1\}, \quad (1)$$

whereby an ℓ_1 (or mean absolute error) loss is applied between the original images and the reconstructed images, and $\mathbb{E}_x$ and $\mathbb{E}_y$ denote expectations over the respective data distributions.

For the adversarial loss: we use a mean squared error, or ℓ_2, loss function between discriminator predictions and real/fake labels. The generator adversarial loss is given by

$$\mathcal{L}_{\text{GAN}}(G_{XY}, G_{YX}) = \mathbb{E}_y(\{D_X[G_{YX}(y)] - 1\}^2) + \mathbb{E}_x(\{D_Y[G_{XY}(x)] - 1\}^2). \quad (2)$$

An additional identity loss encourages preservation of colour and content when mapping images already in the target domain, of the form

$$\mathcal{L}_{\text{id}}(G_{XY}, G_{YX}) = \mathbb{E}_y[\|G_{XY}(y) - y\|_1] + \mathbb{E}_x[\|G_{YX}(x) - x\|_1]. \quad (3)$$

Equations (1), (2) and (3) represent the three loss functions commonly comprising the standard CycleGAN model. The total generative loss function for the original CycleGAN model combines these three components, weighted by hyperparameters λ_{cyc} and λ_{id},

$$\mathcal{L}_{\text{total(unpaired)}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}}. \quad (4)$$

For our modified CycleGAN we adapt this model to paired data, whereby each image $x \in X$ is assumed to have a corresponding ground truth image $y \in Y$, and vice versa. To take advantage of this, we introduce a 'pixelwise supervision loss' that directly penalizes differences between the generator output $\hat{y} = G_{XY}(x)$ and the target image y, as well as $\hat{x} = G_{YX}(y)$ and the target image x. This is formulated as a combined ℓ_1 loss,

$$\mathcal{L}_{\text{px}}(G_{XY}, G_{YX}) = \mathbb{E}_{(x,y)}[\|G_{XY}(x) - y\|_1] + \mathbb{E}_{(x,y)}[\|G_{YX}(y) - x|_1]. \quad (5)$$

Lastly, we introduce a loss term which helps ensure that the generated µCT image is purely greyscale,

$$\mathcal{L}_{\text{gs}}(G_{YX}) = \mathbb{E}_y\Big[\|G^{\text{r}}_{YX}(y) - G^{\text{g}}_{YX}(y)\|_1 + \|G^{\text{r}}_{YX}(y) - G^{\text{b}}_{YX}(y)\|_1 + \|G^{\text{g}}_{YX}(y) - G^{\text{b}}_{YX}(y)\|_1\Big], \quad (6)$$

where $G^{\text{r}}_{YX}(y)$, $G^{\text{g}}_{YX}(y)$ and $G^{\text{b}}_{YX}(y)$ are the red, blue and green channels of the generated µCT image. (To present as a greyscale image, all three channel values should be equivalent.)

The total generative loss for our paired datasets combines all components, weighted by hyperparameters λ_{cyc} and λ_{id}, λ_{px} and λ_{gs},

$$\mathcal{L}_{\text{total(paired)}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{px}}\mathcal{L}_{\text{px}} + \lambda_{\text{gs}}\mathcal{L}_{\text{gs}}. \quad (7)$$

In this study, we evaluate the performance of the CycleGAN network as applied to paired data, using both the standard total generator loss [equation (4)] and our modified version [equation (7)].

2.6.2. Pix2Pix

As a further baseline comparison, we test our modified CycleGAN for paired data against the classic Pix2Pix model (Isola *et al.*, 2017), which is a conditional generative adver-

sarial network (cGAN). In this framework, the generator G learns to map an input image x (here a µCT slice) to a corresponding target image y (histological section), *i.e.* $G(x) \simeq y$. The discriminator D is explicitly conditioned on the input x and is trained to distinguish between real image pairs (x, y) and fake pairs $[x, G(x)]$.

The conditional adversarial loss between generator and discriminator can again be written as an ℓ_2 loss,

$$\mathcal{L}_{\text{cGAN}}(G) = \mathbb{E}_x\big(\{D[x, G(x)] - 1\}^2\big). \tag{8}$$

This adversarial loss is balanced in the Pix2Pix model with a pixelwise loss commonly referred to the ℓ_1 loss,

$$\mathcal{L}_{\ell_1}(G) = \mathbb{E}_{(x,y)}\big[\|G(x) - y\|_1\big]. \tag{9}$$

The total generative loss for Pix2pix then combines them with the weighted parameter λ_{ℓ_1},

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cGAN}} + \lambda_{\ell_1}\mathcal{L}_{\ell_1}. \tag{10}$$

2.6.3. Model architectures and hyperparameters

The base model architecture for both the CycleGAN model variations and Pix2Pix models tested follow the *PyTorch* implementations of Linder-Norén (2019). All models were initialized from scratch without pre-training before introducing our input data.

For the CycleGAN model variations, the generator network is based on a ResNet backbone with nine residual blocks and a starting feature width of 64. The discriminator follows an unconditional PatchGAN design with a 70 × 70 pixel receptive field. To stabilize adversarial training, an image buffer was used to store previously generated samples for the discriminator loss, following the technique described by Zhu *et al.* (2017). Training was conducted using the Adam optimiser with a learning rate of $l_r = 0.0002$, and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For our modified CycleGAN loss model for paired data given in equation (7), we achieved good results with loss factors $\lambda_{\text{cyc}} = 6$, $\lambda_{\text{id}} = 3$, $\lambda_{\text{px}} = 6$ and $\lambda_{\text{gs}} = 1$. These factors are approximately proportional to those relative to the GAN loss as optimized in the original study by Zhu *et al.* (2017). The latter two factors are effectively ablated for the comparison with the standard CycleGAN model given by equation (4). Sample correspondence masks (see Section 2.8) were applied as binary multipliers for all ℓ_1 loss terms.

For the Pix2Pix model, the standard architecture as introduced by Isola *et al.* (2017) consists of a U-Net generator and a conditional PatchGAN discriminator. Here the total loss model follows equation (10), where $\lambda_{\ell_1} = 20$ after tuning. We also applied the sample correspondence mask to the ℓ_1 pixel loss term. All other architectural and training settings were kept consistent with those used in the CycleGAN model above.

In each case models were trained on 256 × 256 pixel patches randomly sampled from 40 WSI pairs, with 100 patches per pair per epoch, a batch size of four, and a total of 500 epochs. At each epoch, new randomized data augmentations were applied on-the-fly during patch extraction. During training, model states corresponding to the minima of each individual generator loss component were saved as checkpoints for later evaluation. Each training fold required approximately 100 h on an NVIDIA Tesla V100 GPU for the CycleGAN models versus 40 h for Pix2Pix, across the fivefold cross-validation scheme.

2.7. Data augmentation

The utility of data augmentation in achieving learning invariance has been well demonstrated (Dosovitskiy *et al.*, 2014; Ronneberger *et al.*, 2015) and is particularly beneficial in medical imaging applications, where obtaining large numbers of paired datasets is often a challenge. By combining the randomized patch-based sampling with further data augmentation at each epoch we were able to simulate a substantially larger paired dataset than the initial set of WSI image pairs, thereby helping to prevent overfitting. For each training example in both histology and µCT images, a random crop position was determined for the sampled patch. We also incorporated rescaling by 90–110% (whilst maintaining patch array size), horizontal and vertical flipping, and a separate contrast and brightness jitter (up to 10%) for both the µCT and histology patches. For the latter, the chromaticity or perceived colour is not varied (RGB ratios are preserved).

2.8. Sample correspondence

Sample correspondence masks were applied to the ℓ_1 loss terms during training in order to exclude regions of the image pairs where there was known to be a mismatch between µCT and histology. These could be either in the imaging field of view or physical differences in the sample between tomographic and histological measurement arising from the sectioning process. The maps were created using a convex hull mask of the rigid bone tissue + screw sample, calculated via intensity-based segmentation. This method includes any soft tissue which is surrounded by the rigid bone material which is assumed to remain part of the sectioning, as well as a small amount of background pixels surrounding the bone. This effectively excludes from the training a majority of extraneous sample material (mainly soft tissue) in the µCT volume which did not end up in the histology section, as well as the walls of the sample holder which contained the bone-implant block. These walls can be seen in many of the CT slices. We generated convex hull masks for both CT and histology and took the minimum union, combined with a third/fourth mask generated with semi-manual selection for regions of any missing other information not accurately defined by the convex hull mask (*e.g.* histology images with missing sections of bone due to irregular sectioning). In general we did not attempt to mask out any small imperfections in the histology images, such as cracks in the optical slides, stain droplets, *etc*. The masks were also subsequently applied to our showcased results and during their comparative analysis, whereby any pixels outside of the sample correspondence mask were replaced with a value equivalent to a mean background value.

2.9. Whole slide image outputs

Within the field of virtual staining (with visible light), most methodological studies are constrained to patch-level analysis due to the gigapixel scale of WSIs at their original resolution (Liu *et al.*, 2025). Since we work with downscaled WSIs at the resolution of the input µCT data, we do not have the same memory-related constraints. Following the patch-based training of our model, we re-tiled the model outputs to obtain overlapping patch-based inference of whole slide images of the order 1k × 1k pixels. Analysis of model performance across the WSI field of view has several advantages over a patch-level analysis. This approach provides a more comprehensive assessment of output quality, as localized patches may appear visually convincing in isolation (passing the visual Turing test) but fail to reflect broader inconsistencies. In contrast, overlapping patch-based WSIs may reveal global artefacts such as tiling borders or inconsistent transitions that arise in cases of model instability. Although direct WSI inference is also possible given network scalability, we nevertheless retain patch-based inference for consistency with training and to preserve the same spatial frequency sampling.

2.10. Comparative metrics

In addition to a visual assessment of each model's performance, we also applied the following complementary metrics for a comprehensive quantitative comparison: structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS). These metrics capture different aspects of image similarity, including structural coherence, pixel-wise differences and perceptual fidelity.

We applied the metrics through patch-based (256 × 256 pixels) sampling of the inferred WSI histology images with reference to their input histology counterparts. Background patches were excluded from analysis according to the sample correspondence mask (defined by a 50% overlap or greater).

The structural similarity index measure, or SSIM (Wang *et al.*, 2004), quantifies perceptual similarity between two images based on luminance, contrast and structural components. It is defined as

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (11)$$

where μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, and C_1, C_2 are small stability constants. We used the SSIM implementation by Van der Walt *et al.* (2014).

The PSNR is a traditional metric for quantifying absolute error of the reconstruction quality (between real and generated images) on a logarithmic scale,

$$\mathrm{PSNR}(x, y) = 10\log_{10}\left[\frac{L^2}{\mathrm{MSE}(x, y)}\right], \quad (12)$$

where L is the maximum possible pixel value (*i.e.* 255 for 8-bit images) and $\mathrm{MSE}(x, y)$ is the mean squared error between the images,

$$\mathrm{MSE}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2, \quad (13)$$

where N is the number of pixels.

The learned perceptual image patch similarity, or LPIPS (Zhang *et al.*, 2018), measures perceptual similarity by comparing deep features extracted from a pretrained neural network. Unlike SSIM or PSNR, LPIPS captures high-level perceptual and colour differences, making it well suited for evaluating visual fidelity in generative tasks in the RGB space. Given feature maps $f_l(x)$ and $f_l(y)$ at layer l, the LPIPS distance is

$$\mathrm{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l}\sum_{h,w} w_l \left\| \hat{f}_l^x(h, w) - \hat{f}_l^y(h, w) \right\|_2^2, \quad (14)$$

where $\hat{f}_l$ denotes channel-wise normalized feature maps and w_l are learned weights. We used the LPIPS implementation by Detlefsen *et al.* (2022) with the SqueezeNet network type.

2.11. 3D testing

Once our chosen network model was trained, we applied the forward generative model to a stack of input SR-µCT slices for validation in 3D. These were untransformed slices, although one training pair was also produced from this dataset through the co-registration process outlined in Section 2.3. A 3D Gaussian filter with a σ of one pixel was first applied to the CT slice stack, in order to simulate the resolution reduction which occurs through interpolation steps of the co-registration process. The output stained slices were generated from the trained forward model independently as single slices. For 3D visualization these slices were loaded into Avizo3D 2024.2 (Thermo Fisher Scientific, Berlin, Germany), and the inverted mean of the RGB channels was computed as the alpha (A) channel. This enabled a volume rendering with direct RGBA mapping.

3. Results and discussion

An example of the direct training patch results of our modified CycleGAN model is shown in Fig. 4(*a*). For each 256 × 256 pixel patch region, there is the input real SR-µCT patch, and output generated histology patch, as well as the input real histology patch and corresponding output generated µCT patch. In addition, the sample correspondence map patch is also shown in the central panel. At the patch level, the similarity between input and generated images appears very high. However, it is possible to see that the patches are not perfectly aligned down to every pixel. This is partly due to the initial registration, which was performed solely on a global scale of the WSI and not subsequently re-applied to each individual patch, as well as to differences in the sample slice pair which arise from the histological sectioning process. In particular, the soft bone tissues [as shown in blue to the left of patch 3 of Fig. 4(*a*)] were observed to shift and these regions are not co-registered as accurately as the rigid bone structures. In Fig. 4(*b*), a representative WSI training result is shown, as

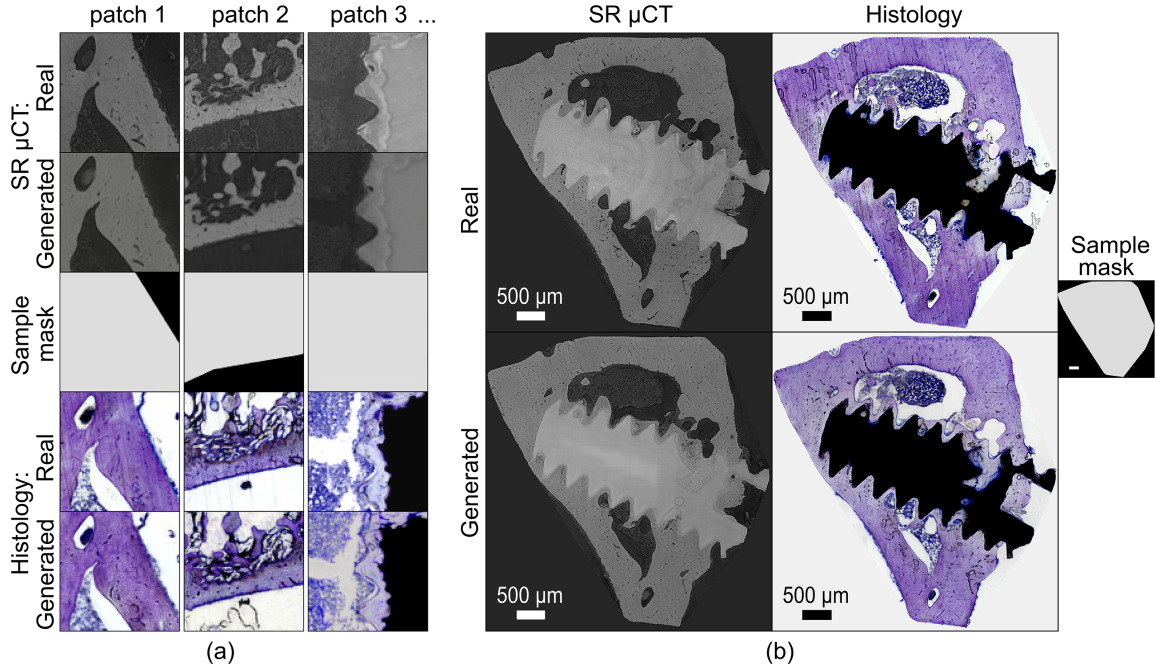


Figure 4
Examples of our modified CycleGAN model training results: (*a*) direct training patch input/outputs (256 × 256 pixels); and (*b*) example WSI result generated through overlapping patch-based inference, from one WSI sample pair included in training (958 × 1000 pixels). The displayed µCT and histology images have been masked (replacing all background pixels with the mean value) with the included sample correspondence mask.

generated through overlapping patch-based inference (stepping the 256 × 256 pixel patch across with a step size of 64 pixels). The sample correspondence mask (also shown separately) has been applied after image generation whereby pixels exterior to the sample are set to the mean background value. Examples of the inferred WSI training, validation and test results displayed without the mask are given in Fig. S1 of the supporting information. Behind the mask, the effect of overlapping patches can sometimes be observed in background regions of the generated images in the form of tiling artefacts of varying intensity steps, reflecting uncertainty in the model prediction due to exclusion from the training process. The step size may be reduced to smooth this effect; however, generally in well trained image regions there are no such artefacts. Note that, for the remainder of this manuscript, generated image results will be presented as masked overlapping patch-based inferred WSI images, or a region of interest thereof, and not the original patches directly used as model input/output.

Representative results of our modified CycleGAN model from training, validation and testing datasets are demonstrated in Fig. 5. Overall there is a good apparent level of agreement between real and generated WSIs in both domains for each dataset, although, as to be expected, there is a notable drop in perceived match accuracy when comparing against the training results. In particular, the generated WSI histology training results appear excellently matched.

In general, the forward model (SR-µCT to predicted histology) results of the CycleGAN are more likely to pass a visual Turing test than the reverse model (histology to predicted SR-µCT). Within the generated µCT slices, tiling artefacts (as mentioned above) are present in the majority of regions containing the residual screw alloy (represented in black pixels within the corresponding histology). Unlike the background regions, these screw regions were not excluded from the training process by the sample correspondence masks. However, the uncertainty in their prediction lies in the possibility of three different material components (PEEK, Mg or Ti) within the sample pool which are characterized by three different linear attenuation coefficients for SR-µCT (respectively, less than, approximately equal to, or greater than the value for dense bone tissue). Each of these are matched to the same black values of the histology input. In our preliminary studies previously reported by Irvine *et al.* (2024), we initially incorporated a fourth channel (in addition to R, G and B channels) as model input in which the known screw material could be indicated as a single value. However, this extra parameter was eventually found to contribute to a significant blurring effect over the whole generated image and we have since abandoned this in our current model. Further contributing to their synthetic appearance, the generated CT slices tend to exhibit reduced high-frequency detail compared with their real counterparts, and lack the shot noise characteristic of real X-ray images.

Other noted discrepancies between real and generated images include the many characteristic features of the real histology sample images which have not been replicated in the generated histology. These include cracks in bone, saw marks (*i.e.* one-directional striations), histology slide fractures and stain droplet spills. The same histology features were mostly reproduced in greyscale format within the generated CT slices, also distinguishing them from their real CT counterpart. All of these features were generally included in the training (unless located significantly external to the bone sample in which case

they were masked out); however, they were not able to be 'learned' as features due to a lack of correspondence. These features are not intrinsically related to the sample but are formed as part of the histological sample preparation process, and are absent from the CT acquisition. This highlights the distinction between accuracy, referring to faithful structural representation, and realism, which encompasses visual artefacts that enhance stylistic plausibility. The omission of such features reflects the model's preference for preserving structural accuracy over superficial realism.

The colour, contrast and resolution fidelity of our paired CycleGAN output results are demonstrated in more detail with the selected ROIs and corresponding RGB histogram plots and intensity profiles derived from the histology ROIs, within Fig. 5. As expected, the training results are a superior match to the validation and test results, most notably with respect to the colour fidelity of the generated histology, but also in the reproduction of the finest features. In the training data (Fig. 5), both real µCT and histology ROIs exhibit a large number of medium and small-sized pores visible down to only

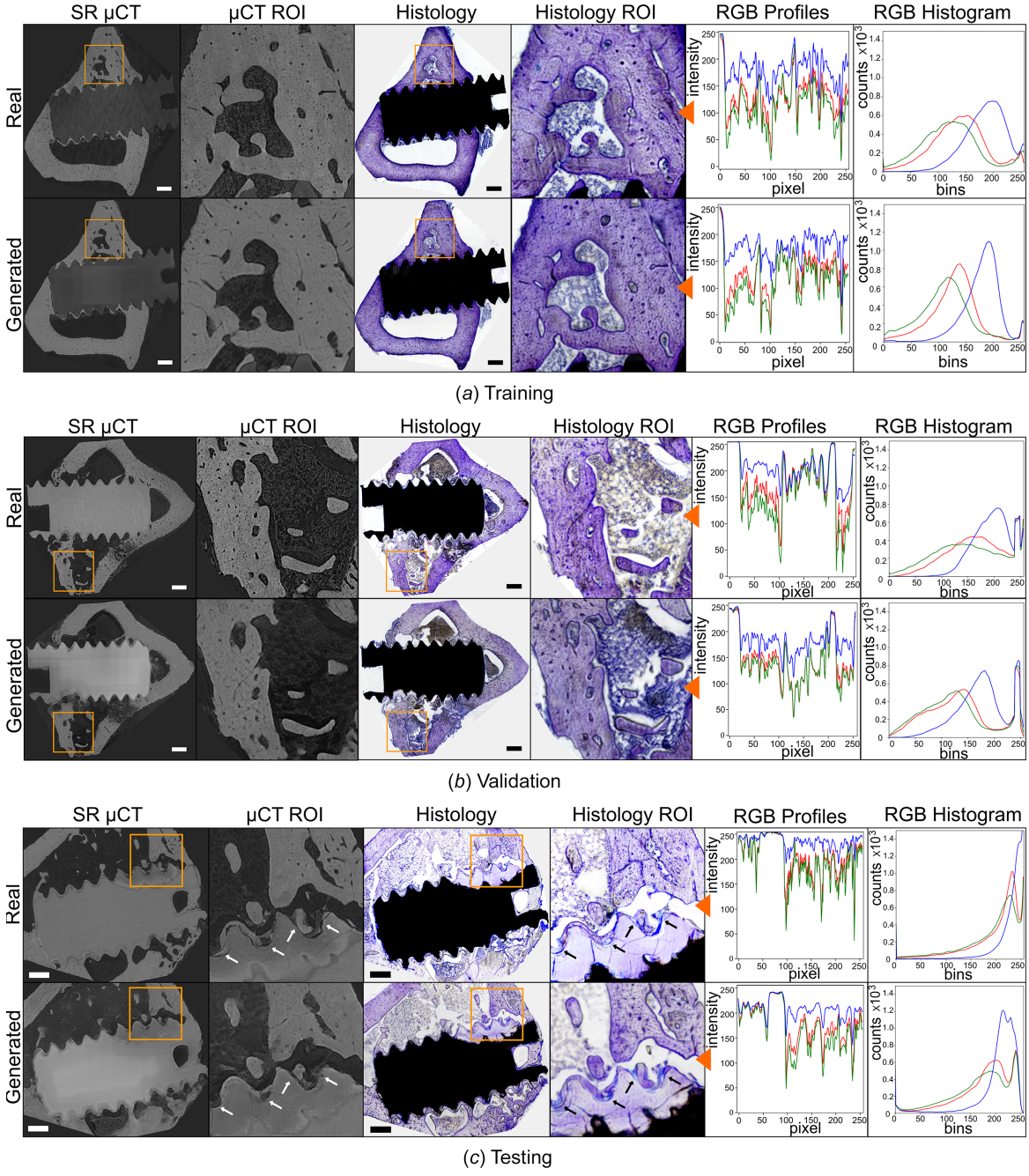


Figure 5
Example modified CycleGAN model results (with WSI output) for (*a*) training, (*b*) validation and (*c*) testing. Includes a 256 × 256 pixels region of interest (ROI) with accompanying RGB histograms and intensity profiles of real versus generated examples. Each profile is acquired horizontally across the midpoint of the ROI as indicated by the orange arrow. The scalebar in each WSI represents 500 µm. In (*c*), isolated regions of new bone growth are indicated with arrows in white (µCT) and black (histology), for both the input data and generated histology.

a few pixels in size within the dense bone structure, which correspond to the blood vessels and lacunae, respectively. These pores are well reproduced in the generated histology, demonstrating the high level of detail able to be supported by the forward model. However, some fine linear features visible in the real histology and correspondingly in the generated μCT images are absent from the generated histology. These likely represent vessels or canaliculi oriented parallel to the cutting plane, indicating a discrepancy between the thickness of the histology sections and registered virtual μCT slices which could be better matched for the model training. Accordingly, the histograms and profiles of the training set are very similar although non-identical. In the validation data set [Fig. 5(*b*)], the soft tissue in the generated histology is notably of a different colour in certain areas to the real image (with respect to blue and brown), as well as differing in intensity, although the variations could be considered realistic as they reflect the varying colours and intensities of soft tissues observed elsewhere in the histology dataset. Such differences are reflected in the histogram, and in the RGB ratios at levels attributable to the soft tissue pixels. The lacunae and other fine features are also less visible in comparison with the training example. In the test example (Fig. 5), discrepancies between real and generated histologies in the colour of both dense bone and soft tissue may also be observed. Conversely to the validation ROI above, the generated soft tissue here is skewed more towards brown than the blue of the real soft tissue. Additionally, we note that our test input μCT data are of lower resolution than the typical 50 training and validation CT examples. As a result, the generated histology also contains fewer fine features, which is visible in broadened histogram peaks and smoother line profile. However, the loss of detail is not as pronounced as one might expect based solely on the input CT resolution, particularly in the dense bone regions. In fact, while a comprehensive spatial frequency model analysis is beyond the scope of this study, preliminary findings (see Section S3) suggest that the GAN-generated histology typically gains a spatial resolution that is lower than real histology but higher than the input μCT. This places the output resolution between the two modalities and supports the qualitative observation that fine structural details may be partially preserved, despite the lower-resolution input.

As a final qualitative test, the selected testing ROI in Fig. 5 provides a valuable demonstration of the GAN generator's predictive accuracy in capturing certain key features from the original X-ray histology study, namely the Mg alloy screw degradation layer and potential regions of new (woven) bone growth (see also Fig. 1 for reference). Here we observe that the prediction for the degradation layer exhibits a boundary that is reasonably consistent with the real histology although a blurriness at the edge of the black residual screw suggests some uncertainty. Upon closer inspection within the degradation layer, we note that, while in real histology images this area is characteristically devoid of bone cells (Krüger *et al.*, 2022), inside the generated layer a faintly repeating cell-like structure is apparent. This suggests that some features may be falsely enhanced from image noise, and more representative data of degradation layers are ideally required for the model to train upon. Importantly, the generated model appears to have successfully predicted the presence of new bone, as indicated by the bright blue regions next to the degradation layer (annotated in the ROI by small black arrows) which were matched to areas of reduced density within the CT image (white arrows).

Results of our modified CycleGAN model on the secondary H&E-stained dataset are provided in Section S2. While the dataset is too small for extensive interpretation, these results are consistent with our main findings and suggest that the model may generalize across multiple stains.

3.1. Qualitative comparison of models

Here we present a comparative analysis of our modified CycleGAN model [including extra ℓ_1 pixelwise supervision and greyscale loss terms given in equation (7)] as evaluated against Pix2Pix representing the baseline model for paired data, as well as the standard CycleGAN model, which also receives paired inputs but is trained using only the standard generator loss [equation (4)]. Whilst the first two models both performed in a predictable manner, the standard CycleGAN model was observed to perform poorly and did not converge to a stable equilibrium. As the only model without direct supervisory loss terms, it was unable to overcome the strong intensity mismatch between the CT and histology domains. The forward generator attempted to create histology images with a dark background and light bone structure, whereas the reverse generator tried to create CT images with a bright background and dark bone structure. This issue is consistent with challenges previously reported when applying CycleGAN to virtual staining tasks with unlabeled microscopy images which present an inverted contrast to the desired stained target images (Bai *et al.*, 2023). To mitigate this, we adopted a simple recommended workaround (Chen *et al.*, 2021; Abraham *et al.*, 2022) and performed a second set of tests of the standard CycleGAN model whereby the input CT images were first inverted. This adjustment led to some limited improvement in both output quality and training stability. The following results are thus included for each of four model variants which includes the standard CycleGAN with both original and contrast-inverted CT inputs. For all variants we focus exclusively on the forward output (*i.e.* the generated histology), as this represents the primary objective of the application. Additionally, because Pix2Pix is composed of only one generator, it does not produce a corresponding CT output.

In Fig. 6 we qualitatively compare the inference results of all paired model variants with example validation and test outputs of a WSI with a 256 × 256 pixel ROI (the same ROI is reused from Fig. 5). Our model visibly outperforms the competing models. The Pix2Pix model was able to reproduce colour reasonably well; however, the resolution and texture of the generated histology images were noticeably inaccurate. During hyperparameter tuning an emphasis was placed on resolution, which reduced the weight of the ℓ_1 loss relative to the adversarial term. The resulting increased GAN loss

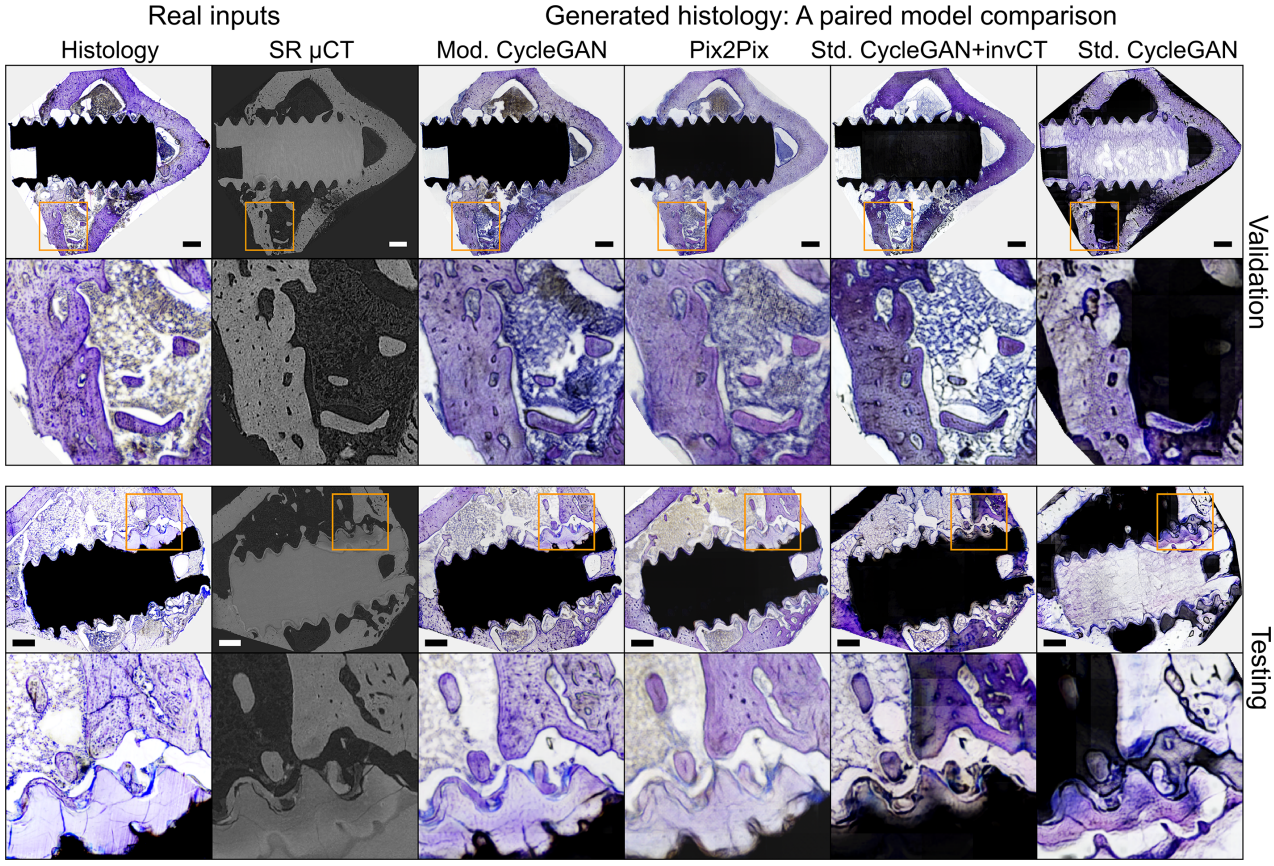


Figure 6
Inference results from each of the paired model variants tested, with WSI and 256 pixel ROI examples from validation and test data sets. 'Mod. CycleGAN' refers to the modified CycleGAN including the extra supervisory and greyscale loss terms, which is our chosen model. 'Std. CycleGAN' refers to the Standard CycleGAN model, which was trained from the same paired data sets, but without the extra loss terms. 'Std. CycleGAN + invCT' refers to a variant of the same model but trained and tested with an inverted CT input, in an attempt to combat the issue of intensity mismatch. The scalebar in each WSI represents 500 µm.

contribution yields some high-frequency artefacts and texture hallucination. The poor performance of Pix2Pix may also be attributed to its sensitivity to the imperfect alignment of paired data, particularly in finer features such as individual bone pores and in soft-tissue regions prone to deformation during histological sample preparation. As mentioned, the standard cycleGAN model (far right) performed very poorly without the additional supervisory loss terms, and using normal CT inputs. The model failed totally to replicate the appropriate histological colours with respect to bone material and background. After switching to an inverted CT input, we observed an increased performance with a more appropriate contrast range, but the model still had trouble with the black screw region of the histology images corresponding to a range of brightness values in the input CT (depending on implant material). This resulted in many regions of bone in the generated histology falsely predicted as black, as well as poor delineation of the degradation layer. Many pores and voids were also predicted with an inverted intensity. Despite a generally poor colour prediction, the standard CycleGAN without extra supervisory ℓ_1 term was able to reproduce the fine features reasonably well, and in particular the soft tissue regions were more sharply defined than in our chosen model.

Section S2 also includes a reduced model comparison on the H&E-stained dataset, centred upon performance of the modified CycleGAN versus Pix2Pix. The results indicate that the modified CycleGAN maintains an advantage over the Pix2Pix baseline, further supporting its applicability across different staining protocols.

3.2. Quantitative evaluation

Table 1 presents the quantitative measurements of all paired model variants (our modified CycleGAN model with extra loss terms, compared with the standard Pix2Pix and standard CycleGAN models, the latter applied with both normal input µCT and inverted µCT input). Measurement values for each of the metrics SSIM, PSNR and LPIPS are shown, for each of training, validation and testing samples. The values shown are the median calculated values from quite varied distributions due to the highly heterogeneous structures across the sampled WSIs. For this reason we have also included box plots of the distributions in Fig. 7 where it is possible to observe the range of values.

When evaluated quantitatively, our model consistently outranks all other variants across the three metrics, with

Table 1
Evaluation of models with median values of the metrics SSIM, LPIPS and PSNR for training (from 40 WSI samples), validation (10 WSI samples) and testing (3 WSI samples). The top values are highlighted with bold text.

	Metrics (training)			Metrics (validation)			Metrics (testing)		
Model	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑
Mod. CycleGAN (our model)	**0.63**	**0.12**	**17**	**0.60**	**0.16**	**15**	**0.59**	**0.16**	**15**
Pix2Pix	0.24	0.15	16	0.21	0.17	14	0.22	0.22	**15**
Std. CycleGAN + inv. CT input	0.27	0.25	13	0.31	0.25	13	0.33	0.25	12
Std. CycleGAN	0.15	0.44	5	0.18	0.43	5	0.20	0.43	5

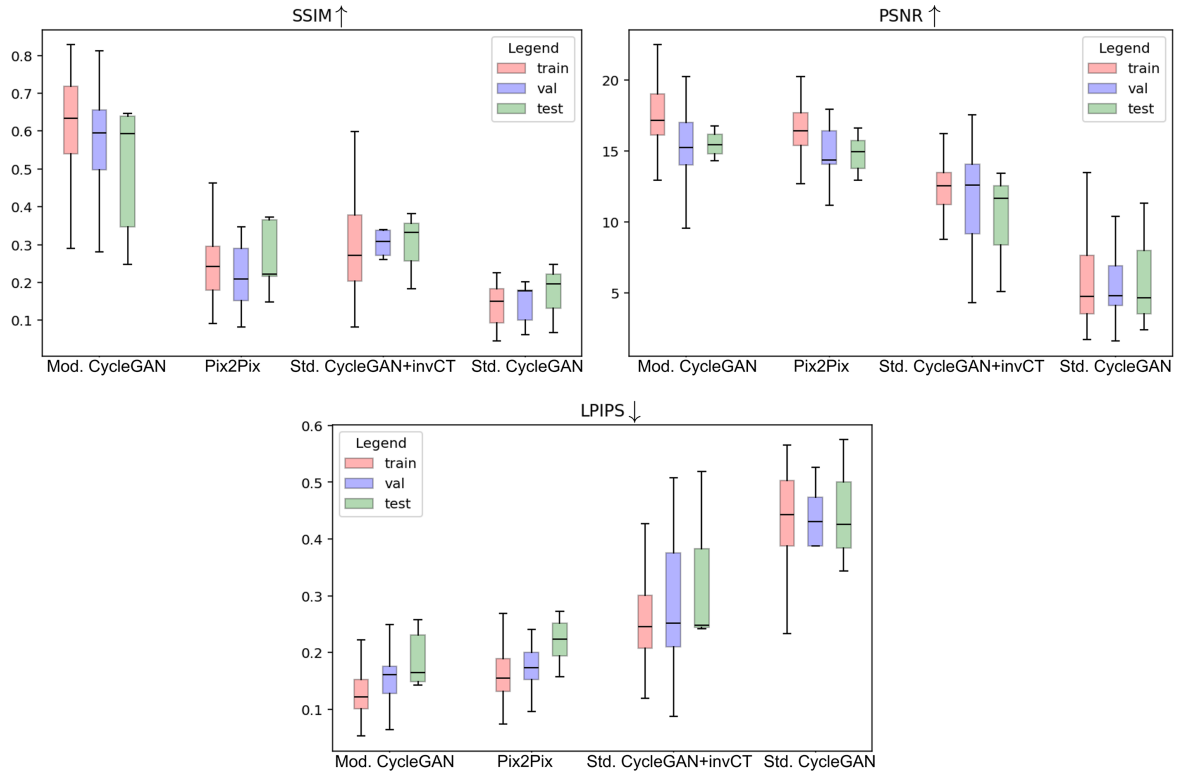


Figure 7
Box plots for a comparison of the paired model variants tested, across the three metrics of structural similarity (SSIM) metric, peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS).

particularly strong performance in the structural similarity (SSIM) metric. Among the alternative model variants, the standard CycleGAN with uninverted µCT inputs clearly performed the worst overall. Between Pix2Pix and the standard CycleGAN with inverted µCT inputs, Pix2Pix achieved higher PSNR and lower LPIPS values, indicating lower reconstruction error and a good colour match. However, it performed significantly worse on SSIM, due to the observed high-frequency artefacts. Conversely, the standard CycleGAN with inverted µCT inputs exhibited better SSIM but poor LPIPS, reflecting higher structural fidelity but poor colour consistency.

It is worth noting that while the top scores remain relatively low compared with results from virtual stain transfer, for example, they are comparatively high for other cross-modality image translation tasks. All metric values are somewhat inflated due to the inclusion of screw regions, which comprise large areas of homogeneous black pixels, although these are considered characteristic of our model's application. As a baseline, when the same metrics were computed between the input µCT slices and their co-registered histology counterparts (*i.e.* without generation), the average SSIM, LPIPS and PSNR were 0.15, 0.4 and 5, respectively. Thus, the standard CycleGAN outputs without inverted CT input are not substantially better matched to the real histology than the raw µCT inputs themselves.

3.3. 3D qualitative results from our model

Orthogonal slice views and a volume rendering of a virtually stained 3D X-ray histology dataset (featuring an Mg-based bone-implant) generated from our trained modified CycleGAN network are shown in Fig. 8. Overall, the generated volume displays strong visual consistency throughout. However, minor stripe artefacts are visible in the *YZ* and *XZ* planes [Figs. 8(*b*) and 8(*c*)], caused by slight intensity variations along the slice stack (*Z*) axis. These artefacts are more pronounced in regions containing the degradation layer,

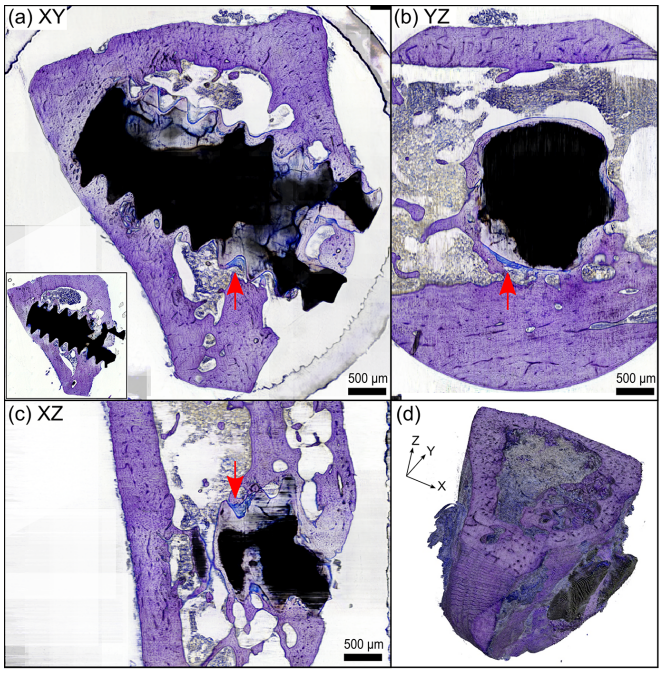


Figure 8
Orthogonal views (*a*, *b*, *c*) and volume rendering with direct RGB mapping (*d*) of a virtually stained 3D X-ray histology dataset featuring an Mg-based degradable bone-implant sample. A real histology image acquired from the same bone-implant sample is shown for reference in the bottom-left inset of (*a*). Red arrows point to a region of new bone growth adjacent to the degradation layer, distinctly coloured in a bright blue.

where the model shows increased slice-by-slice intensity variation due to uncertainty in predicting the precise interface between the residual alloy and the degradation material. Increasing the size of the training dataset would likely improve performance in this area. Alternatively, generating the volume using multiple orthogonal input planes and combining the outputs [similar to the multi-axes prediction fusing approach outlined by Baltruschat *et al.* (2021)] should help produce a smoother and more coherent result. In this initial study, we opted for a simpler approach based on independently generated slices, as this best illustrates the model's stability across the volume.

Notably, our model accurately predicts one of the key regions of interest within the original X-ray histology context, identifying multiple areas of new bone growth adjacent to the degradation layer of Mg-based implants. This newly formed bone, characterized by low mineralization, is distinctly highlighted in bright blue within the 3D virtual staining. One such region is visible across all three orthogonal planes, as indicated by the red arrows. Upon closer re-inspection of the original input 3D CT dataset, the prediction was supported by correspondingly lower greyscale values in the same region, although these features were not immediately apparent.

4. Conclusions and outlook

This work presents one of the first known demonstrations of virtual staining for 3D X-ray histology, enabled by a modified CycleGAN model adapted for paired SR-µCT and histology data of bone-implant systems. Through the integration of pixelwise supervision and greyscale loss terms, the model learns to synthesize artificially stained images from greyscale slices, bridging structural and chemical imaging modalities. Within the bone-implant context, the network captures both broad tissue organization and fine histological features, such as bone lacunae and degradation layers surrounded by new bone tissue, highlighting its potential for biomedical interpretation. Some colour mismatches, particularly in soft tissue regions, indicate limited generalizability to broader histology datasets. Addressing this may require advanced colour normalization, larger and more diverse training data, and progress on the wider challenge of colour standardization in digital pathology.

When compared against both standard CycleGAN and Pix2Pix baselines, the modified CycleGAN was shown to outperform in terms of structural similarity, perceptual fidelity and peak signal-to-noise ratio, as confirmed through both qualitative visual comparisons and quantitative metrics (SSIM, PSNR, LPIPS). In contrast to Pix2Pix, the supervised CycleGAN framework in particular demonstrated its suitability for cross-modality image pairs characterized by imperfect alignment at the pixel level, as is frequently the case in biomedical imaging applications. The standard CycleGAN without supervisory loss terms performed poorly, indicating that paired data are necessary for effective training in this specific application and domain. However, even with partial misalignment, the acquisition of paired data remains a bottleneck. Further studies could explore hybrid model architectures that incorporate unpaired data, such as the approach by Tripathy *et al.* (2019), which initially trains on paired datasets before refining with unpaired samples.

At the whole-slide image level, the method enables volumetric inference for µCT-based virtual histology at resolutions greater than 5 µm. Although this remains below the quality of conventional digital pathology slides, it is suitable for the volume sizes typically used in µCT studies. We successfully tested the trained forward model on a µCT stack, demonstrating its ability to generate virtual staining consistently across a full 3D dataset. Future work will focus on evaluating model performance at higher resolutions, closer to original histology quality, which will likely require interpolating CT data to upscale the training input. This could effectively transform the model into a super-resolution generator and enable deeper investigation into its spatial frequency response. Data augmentation strategies that incorporate artificial noise and smoothing filters could additionally improve robustness.

A further extension could include the integration of segmentation labels such as those used in 3D µCT analysis (Baltruschat *et al.*, 2021), supporting both evaluation and conditional training in future 3D applications. Quantitative evaluation could additionally include comparison with classical methods, for example the joint-histogram-based multi-class bone tissue segmentation reported by Rodgers *et al.* (2022).

Finally, to enhance clinical relevance, collaboration with medical experts will be crucial to ensure biological staining specificity and accuracy in digital pathology applications. We also plan to extend testing to other tissue types and applications, including datasets acquired with laboratory-based μCT systems, thereby broadening the model's utility and generalizability. Expanding the dataset with more samples and additional staining types would further support these goals.

5. Related literature

The following references, not cited in the main body of the paper, have been cited in the supporting information: Nieuwenhuizen *et al.* (2013); Rieger *et al.* (2024).

Acknowledgements

This computational project was carried out with the support of the Joint Laboratory Model and Data-driven Materials Characterization (JL MDMC), a cross-centre platform of the Helmholtz Association. The work was also supported through use of the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany. We thank our summer student Moral Bootbool for her contribution in the pilot testing of our model. The experimental work is acknowledged extensively within Krüger *et al.* (2022). In particular, we are grateful to Silvia Galli for the acquisition of histological images. Open access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare no conflicts of interest.

Data availability

The data that support the findings of this study are available upon reasonable request from the authors.

Funding information

Parts of this research were supported by the BMBF project 'Multi-task Deep Learning for Large-scale Multimodal Biomedical Image Analysis (MDLMA)' (project number 031L0202A), the Hereon project 'Holistic Data Analysis (HoliDAy)' of the Innovation-, Information-& Biologization-Fonds (I2B), and the ErUM-Data Verbundprojekt 'KI4D4E: Ein KI-basiertes Framework für die Visualisierung und Auswertung der massiven Datenmengen der 4D-Tomographie für Endanwender von Beamlines' (project number 05D23CG1) which is funded by the Bundesministeriums für Bildung und Forschung (BMBF).

References

Abraham, T., Costa, P. C., Filan, C. E., Robles, F. & Levenson, R. M. (2022). *Proc. SPIE*, **12136**, 121360Q.

Abraham, T. M. & Levenson, R. (2024). *Mod. Pathol.* **37**, 100443.

Albers, J., Pacilé, S., Markus, M. A., Wiart, M., Vande Velde, G., Tromba, G. & Dullin, C. (2018). *Mol. Imaging Biol.* **20**, 732–741.

Bai, B., Yang, X., Li, Y., Zhang, Y., Pillar, N. & Ozcan, A. (2023). *Light Sci. Applications* **12**, 57.

Baltruschat, I. M., Ćwieka, H., Krüger, D., Zeller-Plumhoff, B., Schlünzen, F., Willumeit-Römer, R., Moosmann, J. & Heuser, P. (2021). *Sci. Rep.* **11**, 24237.

Chen, Z., Yu, W., Wong, I. H. M. & Wong, T. T. W. (2021). *Biomed. Opt. Expr.* **12**, 5920.

Detlefsen, N., Borovec, J., Schock, J., Jha, A., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M. & Falcon, W. (2022). *J. Open Source Software* **7**, 4101.

Donath, K. (1988). *Der Präparator*, **34**(1), 197–206.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M. & Brox, T. (2014). *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, 8–13 December 2014, Montreal, Quebec, Canada, pp. 766–774.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, 8–13 December 2014, Montreal, Quebec, Canada, pp. 2672–2680.

Harms, J., Lei, Y., Wang, T., Zhang, R., Zhou, J., Tang, X., Curran, W. J., Liu, T. & Yang, X. (2019). *Med. Phys.* **46**, 3998–4009.

Heng, Y., Yinghua, M., Khan, F. G., Khan, A., Ali, F., AlZubi, A. A. & Hui, Z. (2024). *Artif. Intell. Rev.* **58**, 39.

Irvine, S. C., Lucas, C., Bootbool, M., Galli, S., Zeller-Plumhoff, B. & Moosmann, J. P. (2024). *Proc. SPIE*, **13152**, 131521Z.

Iskhakova, K., Cwieka, H., Meers, S., Helmholz, H., Davydok, A., Storm, M., Baltruschat, I. M., Galli, S., Pröfrock, D., Will, O., Gerle, M., Damm, T., Sefa, S., He, W., MacRenaris, K., Soujon, M., Beckmann, F., Moosmann, J., O'Hallaran, T., Guillory, R. J., Wieland, D. F., Zeller-Plumhoff, B. & Willumeit-Römer, R. (2024). *Bioact. Mater.* **41**, 657–671.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017). *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134.

Kaji, S. & Kida, S. (2019). *Radiol. Phys. Technol.* **12**, 235–248.

Kang, L., Li, X., Zhang, Y. & Wong, T. T. (2022). *Photoacoustics* **25**, 100308.

Katsamenis, O. L., Olding, M., Warner, J. A., Chatelet, D. S., Jones, M. G., Sgalla, G., Smit, B., Larkin, O. J., Haig, I., Richeldi, L., Sinclair, I., Lackie, P. M. & Schneider, P. (2019). *Am. J. Pathol.* **189**, 1608–1620.

Khimchenko, A., Deyhle, H., Schulz, G., Schweighauser, G., Hench, J., Chicherova, N., Bikis, C., Hieber, S. E. & Müller, B. (2016). *NeuroImage* **139**, 26–36.

Koivukoski, S., Khan, U., Ruusuvuori, P. & Latonen, L. (2023). *Lab. Invest.* **103**, 100070.

Krüger, D., Galli, S., Zeller-Plumhoff, B., Wieland, D. F., Peruzzi, N., Wiese, B., Heuser, P., Moosmann, J., Wennerberg, A. & Willumeit-Römer, R. (2022). *Bioact. Mater.* **13**, 37–52.

Krüger, D., Zeller-Plumhoff, B., Wiese, B., Yi, S., Zuber, M., Wieland, D. F., Moosmann, J. & Willumeit-Römer, R. (2021). *J. Magnesium Alloys* **9**, 2207–2222.

Latonen, L., Koivukoski, S., Khan, U. & Ruusuvuori, P. (2024). *Trends Biotechnol.* **42**, 1177–1191.

Lei, Y., Harms, J., Wang, T., Liu, Y., Shu, H., Jani, A. B., Curran, W. J., Mao, H., Liu, T. & Yang, X. (2019). *Med. Phys.* **46**, 3565–3581.

Linder-Norén, E. (2019). *PyTorch-GAN*, https://githubcom/eriklindernoren/PyTorch-GAN.

Liu, S., Liu, K., Margolis, S., Wu, W., Knezevich, S. R., Elder, D. E., Eguchi, M. M., Elmore, J. G. & Shapiro, L. G. (2025). *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, Houston, TX, USA.

Müller, M., Kimm, M. A., Ferstl, S., Allner, S., Achterhold, K., Herzen, J., Pfeiffer, F. & Busse, M. (2018). *Sci. Rep.* **8**, 17855.

Nieuwenhuizen, R. P. J., Lidke, K. A., Bates, M., Puig, D. L., Grünwald, D., Stallinga, S. & Rieger, B. (2013). *Nat. Methods*, **10**, 557–562.

Peev, S., Parushev, I. & Yotsova, R. (2024). *Appl. Sci.* **14**, 461.

Petzold, L. M., Busse, M., Mohr, H., Pellegata, N. S., Pfeiffer, F. & Herzen, J. (2024). *Proc. SPIE* **13152**, 1315205.

Pichat, J., Iglesias, J. E., Yousry, T., Ourselin, S. & Modat, M. (2018). *Med. Image Anal.* **46**, 73–105.

Rieger, B., Droste, I., Gerritsma, F., ten Brink, T. & Stallinga, S. (2024). *Opt. Express*, **32**, 21767.

Rivenson, Y., Wang, H., Wei, Z., de Haan, K., Zhang, Y., Wu, Y., Günaydın, H., Zuckerman, J. E., Chong, T., Sisk, A. E., Westbrook, L. M., Wallace, W. D. & Ozcan, A. (2019). *Nat. Biomed. Eng.* **3**, 466–477.

Rodgers, G., Sigron, G. R., Tanner, C., Hieber, S. E., Beckmann, F., Schulz, G., Scherberich, A., Jaquiéry, C., Kunz, C. & Müller, B. (2022). *Appl. Sci.* **12**, 6286.

Rofena, A., Guarrasi, V., Sarli, M., Piccolo, C. L., Sammarra, M., Zobel, B. B. & Soda, P. (2024). *Comput. Med. Imaging Graph.* **116**, 102398.

Ronneberger, O., Fischer, P. & Brox, T. (2015). *Medical Image Computing and Computer-Assisted Intervention (MICCAI2015)* edited by N. Navab, J. Hornegger, W. M. Wells & A. F. Frangi, *Lecture Notes in Computer Science Vol. 9351*, pp. 234–241. Cham: Springer International Publishing. Series Title: . https://link.springer.com/10.1007/978-3-319-24574-4_28.

Sarve, H., Johansson, C. B., Lindblad, J., Borgefors, G. & Franke Stenport, V. (2007). *12th International Conference on Computer Analysis of Images and Patterns (CAIP2007)* edited by W. G. Kropatsch, M. Kampel & A. Hanbury, *Lecture Notes in Computer Science Vol. 4673*, pp. 253–260. Berlin, Heidelberg: Springer.

Sefa, S., Espiritu, J., Ćwieka, H., Greving, I., Flenner, S., Will, O., Beuer, S., Wieland, D. F., Willumeit-Römer, R. & Zeller-Plumhoff, B. (2023). *Bioact. Mater.* **30**, 154–168.

Song, A. H., Williams, M., Williamson, D. F., Chow, S. S., Jaume, G., Gao, G., Zhang, A., Chen, B., Baras, A. S., Serafin, R., Colling, R., Downes, M. R., Farré, X., Humphrey, P., Verrill, C., True, L. D., Parwani, A. V., Liu, J. T. & Mahmood, F. (2024). *Cell* **187**, 2502–2520.e17.

Töpperwien, M., van der Meer, F., Stadelmann, C. & Salditt, T. (2018). *Proc. Natl Acad. Sci. USA* **115**, 6940–6945.

Tripathy, S., Kannala, J. & Rahtu, E. (2019). *Computer Vision – ACCV 2018* edited by C. V. Jawahar, H. Li, G. Mori & K. Schindler, pp. 51–66. Cham: Springer International Publishing.

Ueda, Y., Niu, M., Shimazaki, R., Yamazaki, A., Seki, M. & Ishida, T. (2025). *J. Digit. Imaging. Inf. Med.* **39**, 827–841.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E. & Yu, T. (2014). *PeerJ* **2**, e453.

Wilde, F., Ogurreck, M., Greving, I., Hammel, J. U., Beckmann, F., Hipp, A., Lottermoser, L., Khokhriakov, I., Lytaev, P., Dose, T., Burmester, H., Müller, M. & Schreyer, A. (2016). *AIP Conf. Proc.* **1741**, 030035.

Yi, X., Walia, E. & Babyn, P. (2019). *Med. Image Anal.* **58**, 101552.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. (2018). *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 586–595.

Zhou Wang,, Bovik, A., Sheikh, H. & Simoncelli, E. (2004). *IEEE Trans. Image Process.* **13**, 600–612.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017). *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 2242–2251.

Zingman, I., Frayle, S., Tankoyeu, I., Sukhanov, S. & Heinemann, F. (2024). *Proc. Machine Learning Res.* **227**, 1509–1525.