ÉCOLE NORMALE SUPÉRIEURE PARIS-SACLAY
UNIVERSITÉ PARIS-SACLAY

MATHÉMATIQUES VISION APPRENTISSAGE
STAGE

CT to MRI Translation using CycleGAN for Medical Imaging Applications

Student :
Nassim ARIFETTE

Teacher
Maria VAKALOPOULOU

Internship supervisor :
Dima RODRIGUEZ

12 septembre 2025

Table des matières

Table des figures

Résumé

Nous étudions la traduction CT→IRM thoracique en 3D, sans paires et sans déformation géométrique, afin de transférer vers l'IRM UTE (temps d'écho ultracourt) les annotations abondantes disponibles en CT. Nous constituons des corpus CT et IRM UTE harmonisés et mettons en place une chaîne de prétraitement reproductible (unification des orientations, rééchantillonnage isotrope à 1 mm, padding multiple de K, normalisation spécifique à la modalité). Sur cette base, nous évaluons plusieurs variantes de CycleGAN 3D (ResNet-9b, U-Net++, adaptation 3D de DC-CycleGAN), avec objectifs adversariaux LSGAN/WGAN-GP et pertes structurelles.

Dans tous les cas, l'anatomie et la cohérence volumique sont bien préservées (bons scores de cycle), mais la distribution d'intensités propre à l'IRM UTE n'est pas reproduite : décalage et changement d'échelle globaux, effondrement de modes et saturations aux bornes dégradent FID/KID. WGAN-GP améliore l'inversibilité au prix d'images moins « IRM-like », et l'ajout naïf de SSIM-3D n'apporte pas de gain. Nous introduisons donc une perte d'histogramme différentiable qui aligne explicitement, pendant l'entraînement, les distributions marginales d'intensité. Cette Hist-CycleGAN améliore FID (225.96→217.89) et KID (0.1158→0.0969) tout en conservant la fidélité anatomique, malgré un biais de contraste encore perceptible.

Nos contributions sont : (i) une chaîne de prétraitement 3D reproductible pour l'apprentissage sans paires CT/IRM ; (ii) une analyse systématique des choix CycleGAN 3D en thorax ; (iii) l'identification du mésalignement des distributions d'intensité comme verrou principal ; (iv) un objectif « histogramme » qui le réduit partiellement. En perspective : histogrammes conditionnés (par régions) et distances CDF, modélisation ricienne du bruit dans la branche IRM, et validation orientée tâche (segmentation de l'arbre vasculaire sur IRM synthétique) et lectures expertes, pour obtenir un contraste cliniquement exploitable sans altérer la géométrie.

Abstract

We study unpaired, fully volumetric CT→MR translation in the thorax with a strict appearance-only (no-warping) constraint, motivated by enabling label transfer from richly annotated CT to scarce ultrashort echo-time (UTE) MR. We assemble harmonized 3D CT and UTE-MR corpora and implement a reproducible preprocessing pipeline (orientation unification, 1 mm isotropic resampling, pad-to-K, modality-specific normalization). On this foundation, we evaluate 3D CycleGAN-style systems spanning generator backbones (ResNet-9b, U-Net++ and a 3D adaptation of DC-CycleGAN), adversarial heads (LSGAN vs. WGAN-GP), and structure-aware losses.

Across settings, generators preserve anatomy and volumetric coherence (strong cycle metrics) but consistently fail to reproduce the UTE-MR intensity law, yielding global shift/scale drift, mode collapse, and boundary saturation that degrade distributional realism (FID/KID). WGAN-GP strengthens invertibility but further reduces MR-likeness; a naïve SSIM-3D addition does not close the gap. We therefore introduce a lightweight differentiable histogram loss that explicitly aligns marginal intensity distributions during training. This *Hist-CycleGAN* improves distributional scores over a strong ResNet-LSGAN baseline (e.g., FID 225.96→217.89; KID 0.1158→0.0969) while retaining anatomical fidelity, though residual contrast biases remain.

Contributions include: (i) a reproducible 3D CT/MR preprocessing pipeline for unpaired learning; (ii) a systematic analysis of 3D CycleGAN design choices for thoracic CT→UTE-MR; (iii) identification of the intensity-distribution gap as the primary bottleneck; and (iv) a histogram-aware objective that partially mitigates it. We outline next steps—region-conditioned/CDF histogram supervision, Rician noise modeling in the MR branch, and task-based validation via vascular-tree segmentation on synthetic MR and focused reader studies—to achieve clinically usable contrast while preserving geometry

Acknowledgements

I am deeply grateful to Dima Rodriguez for invaluable guidance, thoughtful feedback, and steady encouragement throughout this internship and report. I also warmly thank Georges for the many discussions and the much-appreciated company in the office over the summer. My thanks extend to the entire MRI team for their support and warm welcome, and to everyone at BioMaps for providing a friendly and stimulating research environment. I am also very grateful to Maria Vakalopoulou for her insightful advice and scientific guidance. Finally, I thank all colleagues and friends who contributed—directly or indirectly—to this work; any omissions are entirely unintentional.

1 Introduction

1.1 Context

This internship was carried out at *BioMaps* (Laboratoire d'Imagerie Biomédicale Multimodale Paris–Saclay), a joint research unit supervised by Université Paris–Saclay, CEA, CNRS, and Inserm, based at the CEA's Service Hospitalier Frédéric Joliot (Orsay). BioMaps brings together engineers, physicists, biologists, pharmacists, clinicians, and chemists to develop multimodal medical imaging methods (notably PET, MRI, and ultrasound), bridging fundamental research and potential clinical applications across oncology, neurology, and cardiopulmonary medicine.

Within BioMaps, I worked in the MRI group. The internship is embedded in the European project *V|LF–Spiro3D*, coordinated by Université Paris–Saclay and scientifically led by Xavier Maître. The consortium gathers academic, clinical, and industrial partners (e.g., Siemens Healthineers, Erasmus MC, AP–HP, Hôpital Foch, Institut Polytechnique de Paris, Tilburg University), with expertise spanning low/very–low–field MRI hardware, pulse sequence design, image reconstruction, AI methods, and clinical validation.

The project's central aim is to develop and validate *3D MR spirometry*: a non-invasive technique to map regional lung ventilation during free breathing, and to translate this technique to low and very–low magnetic fields to improve portability, cost, and accessibility compared with conventional high-field systems. Activities range from acquisition at standard and low fields, sequence development and reconstruction, and instrumentation for very-low-field systems, to AI-based processing and clinical research aimed at identifying lung biomarkers. Within this framework, my work focuses on AI for image-to-image translation from CT to MR, aligned with V|LF–Spiro3D's goals of enabling robust processing and segmentation under constrained MR data availability.

1.2 Motivation

This work addresses *CT→MR* translation in the thorax under an *unpaired, fully 3D* setting.

CT vs. MR in the chest. Computed tomography (CT) is fast, widely available, and the clinical workhorse for lung and mediastinal imaging, but it relies on ionizing radiation. Repeated follow-up in certain populations (e.g., pediatrics, young adults under surveillance, chronic lung disease) motivates alternatives with lower cumulative exposure. Magnetic resonance (MR), being non-ionizing, is attractive in principle. Moreover, ultra-short echo time (UTE) sequences improve the visibility of short-T_2 components in lung and at bone–air interfaces. However, chest MR remains technically demanding due to low proton density, rapid signal decay, and respiratory/cardiac motion, and its deployment and standardization lag behind CT.

Data imbalance and scarcity of labels. Large thoracic CT datasets exist, often with rich annotations (e.g., LUNA16 (Setio et al., 2017)). Vessel- or airway-centric CT resources are also available; in our case, we obtained PARSE2022 (Chu et al., 2025; Luo et al., 2023) (upon request), which contains vascular-tree annotations derived from CT. By contrast, open thoracic MR datasets are scarce, typically smaller, heterogeneous across sites and sequences, and—crucially for supervised learning—rarely include voxelwise ground truth for structures such as the pulmonary vascular tree. This paucity of labeled MR data is a rate-limiting factor for training segmentation or detection models directly in the MR domain.

Label transfer via modality synthesis. Many downstream tasks (e.g., vascular-tree segmentation, airway/lobe segmentation, lesion localization) require dense labels that are costly

or unavailable in MR. A pragmatic route is to *transfer* existing CT labels: synthesize MR-like images from labeled CT while *preserving anatomy*, then reuse CT labels as supervision on the synthetic MR. In this work, we aim to generate chest MR volumes (with UTE-like appearance) from labeled CT scans; the original CT vascular-tree masks then serve as labels for the synthetic MR, enabling training of MR-domain segmentation models despite the absence of a large, labeled, real-MR dataset.

Why thorax and why CT→MR? Most medical image translation studies focus on MR→CT (e.g., for attenuation correction) and on anatomies such as brain or pelvis, where paired data and standardized protocols are more common. In contrast, *thoracic* CT→MR is less explored and more challenging: motion (respiratory and cardiac), low and variable parenchymal signal, susceptibility at air–tissue interfaces, and heterogeneous UTE protocols exacerbate domain shifts. Furthermore, intensities in lung MR are not standardized to the same extent as Hounsfield units in CT, so preserving both *anatomical fidelity* and *MR contrasts* is non-trivial.

1.3 Problem statement

We address unpaired, volumetric CT→MR translation for thoracic imaging. Let $\mathcal{X}$ denote 3D chest CT volumes and $\mathcal{Y}$ denote 3D chest MR volumes with UTE-like appearance. Given two unpaired datasets $\mathcal{D}_{\mathrm{CT}} = \{x_i\}$ and $\mathcal{D}_{\mathrm{MR}} = \{y_j\}$, the objective is to learn a generator $G : \mathcal{X} \to \mathcal{Y}$ that alters *appearance* while preserving *anatomy*. In practice, $G(x)$ must share the voxel grid and geometry of x (no spatial warping), while exhibiting contrast statistics compatible with lung MR acquired using short-echo techniques. To make the mapping learnable, all volumes are consistently oriented, resampled to near-isotropic resolution, and intensity-normalized with modality-specific procedures (HU windowing/clipping for CT; bias-field correction and robust scaling for MR). The model operates on 3D patches/volumes to capture through-plane context (e.g., vessels and fissures) and to avoid slice-wise inconsistencies typical of 2D/2.5D approaches.

1.4 Objectives and hypotheses

Our overarching goal is an anatomy-preserving CT→MR mapping for the thorax that mitigates the scarcity of labeled chest MR by enabling label transfer from CT. Concretely, we (a) specify a UTE-like target appearance and an explicit no-warping constraint; (b) implement a reproducible preprocessing pipeline that harmonizes geometry and intensities across modalities; (c) design and train a 3D CycleGAN-style framework adapted to volumetric data and memory limits. Performance is assessed both in the image domain (fidelity/realism and texture analyses) and in the task domain by training a vascular-tree segmentation network directly on synthetic MR using transferred CT labels, with additional checks of robustness and generalization to available real MR.

Hypotheses. Full 3D context improves anatomical coherence across slices compared with 2D/2.5D formulations, measurable by continuity metrics (e.g., vessel centerline consistency) and downstream segmentation scores.

1.5 Contributions

We assemble thoracic CT and MR corpora suitable for unpaired 3D learning and introduce a harmonized preprocessing pipeline (orientation, near-isotropic resampling, field-of-view alignment, HU windowing for CT, bias-field correction and robust scaling for MR). We propose a volumetric CT→MR translation framework based on CycleGAN with memory-aware training on overlapping 3D patches and an explicit *no-warping* design, together with structure-aware

objectives that encourage anatomy-faithful synthesis with UTE-like appearance. We report quantitative image-level metrics alongside a task-driven evaluation in which a 3D vascular-tree segmentation network is trained on synthetic MR using transferred CT labels, and we analyze robustness through targeted ablations (architectures, losses, normalization/augmentation) and failure cases. All preprocessing, training, and evaluation steps are configured for end-to-end reproducibility.

2 Background and Related Work

2.1 CT fundamentals

Computed Tomography (CT) is a non-invasive imaging modality that uses ionizing X-ray radiation to estimate a volumetric map of tissue X-ray attenuation in the body. The physical model is well approximated by the Beer–Lambert law. Let I_0 denote the incident X-ray intensity and I the measured intensity after the beam travels along a straight path l through the patient. If $\mu(\mathbf{r})$ [cm^{-1}] is the linear attenuation coefficient at spatial location $\mathbf{r}$, then

$$I \;=\; I_0 \,\exp\Big(-\int_l \mu(\mathbf{r})\,\mathrm{d}l\Big). \tag{2.1}$$

Taking the negative logarithm linearizes the measurement and yields a line integral of μ:

$$p \;\triangleq\; -\ln\Big(\tfrac{I}{I_0}\Big) \;\approx\; \int_l \mu(\mathbf{r})\,\mathrm{d}l \;+\; \varepsilon, \tag{2.2}$$

where ε accounts for stochastic noise and modeling error. As the X-ray tube and detector array rotate around the patient, the system records such log-measurements for many angles and detector positions; the resulting collection, often written $p(\theta, s)$ for rotation angle θ and detector coordinate s, is the (discretized) Radon transform of the unknown attenuation map.

Clinical scanners acquire these data either in axial mode (slice by slice) or in helical mode (the table translates continuously during rotation, producing a spiral trajectory that accelerates volumetric coverage). From the measured projections, an image is reconstructed. The classical approach is filtered back-projection (FBP), which applies a frequency-domain ramp filter to each projection and then back-projects the filtered data over all angles to recover an estimate of μ. Contemporary systems also employ model-based iterative reconstruction (MBIR), which formulates image formation as a regularized inverse problem. In a discretized notation, if A is the system matrix (projector) and W encodes the noise statistics, a typical estimator solves

$$\min_{\mu \geq 0} \;\; \tfrac{1}{2}\,\|W^{1/2}(A\mu - p)\|_2^2 \;+\; \lambda\, R(\mu), \tag{2.3}$$

where $R(\mu)$ is a prior/regularizer (e.g., total variation) and $\lambda > 0$ balances data fidelity and regularization. Iterative methods can reduce noise and some artifacts, especially at lower radiation dose, at the cost of computation.

Reconstructed CT images are reported on the Hounsfield Unit (HU) scale, an absolute, scanner-calibrated intensity defined relative to water:

$$HU \;=\; 1000\,\frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}}. \tag{2.4}$$

This convention makes intensities comparable across scanners and protocols. Typical reference values are: air ≈ -1000 HU, water $= 0$ HU, and cortical bone $\gtrsim +1000$ HU (the exact range depends on beam energy and material). Clinical "windowing" (e.g., lung or mediastinal windows) remaps display levels for human viewing but does not alter the underlying calibrated voxel values.

CT contrast arises from two dominant interaction mechanisms: Compton scatter, which is approximately proportional to electron density, and the photoelectric effect, which scales roughly as Z^3/E^3 with atomic number Z and photon energy E. These interactions explain CT's excellent depiction of bone and its crisp visualization of aerated lung parenchyma, as well as the dependence of appearance on tube voltage (kVp) and filtration.

The noise statistics are governed by the discrete nature of photon detection. Raw detector counts are well modeled as Poisson random variables; after the logarithmic transform, the projection noise is approximately Gaussian with variance inversely related to detected counts. Subsequent analytical reconstruction introduces spatial correlation in the image-domain noise. In practice, chest CT is commonly acquired with thin slices (about 0.5–1.25 mm) in a single breath-hold; "low-dose" protocols reduce radiation exposure at the expense of lower signal-to-noise ratio and potentially more pronounced artifacts.

Several artifacts are characteristic of CT and are relevant to downstream processing. Beam hardening due to the polyenergetic X-ray spectrum leads to cupping and streak artifacts, especially near dense materials. Severe attenuation or metallic implants can produce photon starvation and metal streaks. Patient motion (respiratory or cardiac), partial-volume averaging at thin structures, and scatter further degrade image quality. Protocol choices (kVp, mA modulation, pitch, collimation) and, when indicated, the use of iodinated contrast agents are tuned to minimize these effects while answering the clinical question. Although CT uses ionizing radiation, modern protocols and dose-modulation strategies aim to keep exposure as low as reasonably achievable.

Two properties of CT are particularly important for the cross-modal translation task considered later in this thesis. First, the HU scale provides an absolute, site-independent intensity standard, which simplifies normalization and improves cross-scanner consistency compared with MRI. Second, acquisition geometry and reconstruction are highly standardized, yielding relatively stable image appearance across sites.

2.2 MRI fundamentals

Magnetic Resonance Imaging (MRI) produces images by exciting and measuring nuclear magnetization, most commonly that of hydrogen protons in water and fat. Each proton has spin and an associated magnetic moment. In a strong, static magnetic field B_0, a small excess of spins aligns with the field, yielding a net equilibrium magnetization M_0 along the z-axis of the scanner. The spins precess about B_0 at the Larmor angular frequency $\omega_0 = \gamma B_0$, where γ is the gyromagnetic ratio; for hydrogen, $\gamma/2\pi \approx 42.58$ MHz/T, so the precession frequency increases linearly with field strength. An applied radiofrequency (RF) pulse near this frequency tips the magnetization away from z by a flip angle α, creating a transverse component that induces a measurable voltage in the receive coil.

After excitation, the magnetization relaxes back toward equilibrium through two time constants described by the Bloch equations. Longitudinal (spin–lattice) recovery along z occurs with time constant T_1, often summarized by $M_z(t) = M_0\big(1 - e^{-t/T_1}\big)$. Transverse (spin–spin) decay in the xy-plane occurs with time constant T_2, giving $M_{xy}(t) = M_{xy}(0)\, e^{-t/T_2}$. In practice, microscopic field inhomogeneities and susceptibility variations introduce additional dephasing, leading to an effective constant T_2 with $T_2 < T_2$; gradient-echo sequences are sensitive to T_2, whereas spin-echo sequences refocus much of this dephasing.

Spatial encoding is achieved with linear magnetic field gradients G_x, G_y, G_z that make the precession frequency (and phase) vary with position. The received signal is thus a weighted sum of spatial Fourier components of the object magnetization. Data are sampled in the spatial-frequency domain known as *k-space*; a discrete inverse Fourier transform reconstructs the image. The sampling step Δk along a given axis sets the field of view (FOV) via $\Delta k = 1/\mathrm{FOV}$, while the largest sampled spatial frequency determines nominal resolution (e.g., $\delta x \approx 1/(2k_{x,\max})$).

Sampling trajectories may be Cartesian (line-by-line) or non-Cartesian (e.g., radial or spiral), the latter enabling very short echo times or motion-robust acquisitions at the cost of more complex reconstruction.

Contrast in MRI is controlled by the pulse sequence and timing parameters. Repetition time (TR), echo time (TE), and flip angle (α) determine the weighting between proton density (ρ), T_1, T_2, and T_2^*. For a spin-echo (SE) sequence, a common signal model is

$$S_{\mathrm{SE}} \;\propto\; \rho\left(1 - e^{-\mathrm{TR}/T_1}\right) e^{-\mathrm{TE}/T_2}, \tag{2.5}$$

so long TR and long TE emphasize T_2 differences. For a spoiled gradient-echo (GRE/SPGR) sequence with Ernst steady state, the signal can be written

$$S_{\mathrm{GRE}} \;\propto\; \rho\,\frac{\sin\alpha\,\left(1 - E_1\right)}{1 - E_1\cos\alpha}\, e^{-\mathrm{TE}/T_2^*}, \qquad E_1 = e^{-\mathrm{TR}/T_1}, \tag{2.6}$$

illustrating how short TR and an appropriate flip angle enhance T_1 weighting, while TE controls T_2^* sensitivity. Beyond these basic families, many sequence variants exist to probe specific tissue properties (e.g., diffusion-weighted imaging, inversion recovery such as FLAIR, balanced SSFP, and quantitative mapping).

In the thorax, MRI faces additional challenges. The proton density in the lungs is low and many parenchymal and bone–air interfaces exhibit extremely short apparent transverse relaxation times, with signals decaying within hundreds of microseconds. *Ultra-short echo time* (UTE) and *zero echo time* (ZTE) methods address this by driving TE to the tens of microseconds (UTE, typically with center-out radial sampling) or effectively to zero (ZTE). Respiratory and cardiac motion further degrade image quality; breath-holding, navigator echoes, respiratory/cardiac gating, parallel imaging, and compressed sensing are routinely used to mitigate these effects and reduce acquisition time.

The noise in the complex (pre-magnitude) MR signal is well modeled as Gaussian; after forming magnitude images, the distribution becomes Rician, which introduces a positive bias at low signal-to-noise ratio. With multi-coil combination, noise may be better approximated by a non-central χ distribution. In addition to stochastic noise, slowly varying receive-coil sensitivity (often denoted the B_1^- field) produces a smooth intensity shading or *bias field* across the image, while transmit field (B_1^+) inhomogeneity leads to spatially varying flip angles and thus contrast variation. Other common artifacts include susceptibility-induced distortions and signal loss (particularly in echo-planar imaging), chemical shift at fat–water boundaries, ghosting from periodic motion, and geometric distortion from B_0 inhomogeneity.

Unlike CT, MRI intensities are not reported on an absolute, inter-scanner calibrated scale; they depend on the scanner hardware, receive coil, sequence, and parameter settings. As a result, nominally similar acquisitions can differ in global intensity scaling, spatial bias fields, texture, and noise characteristics. This lack of standardization is a central consideration for learning-based cross-modal translation.

Safety in MRI is governed by the strong static field, RF power deposition, and rapidly switching gradients. Ferromagnetic objects can become projectiles in the B_0 field; careful screening for implants and devices is mandatory, with labeling that specifies *MR Safe* or *MR Conditional* operating limits. RF energy absorption is controlled via specific absorption rate (SAR) constraints to avoid tissue heating, and fast gradient switching can induce peripheral nerve stimulation; acoustic noise is also significant. Gadolinium-based contrast agents are widely used to modify T_1 and enhance vascular or lesional contrast, but their use requires clinical judgment (e.g., in severe renal impairment) and adherence to current safety guidelines.

For the CT→MRI translation task in this thesis, these properties have immediate implications. We explicitly target a well-defined MR contrast (e.g., UTE-like chest MRI), while keeping

a good structure, this will be the greatest challenge of this work.

2.3 Deep learning in medical imaging

Deep learning has substantially reshaped medical imaging. Unlike traditional pipelines built on handcrafted features, modern architectures—especially convolutional neural networks (CNNs) and their variants—learn hierarchical representations directly from data. This shift has delivered large gains in classification, detection, segmentation, registration, and image reconstruction (Litjens et al., 2017; Shen et al., 2017).

A key advantage is the ability to capture both local texture and global anatomical context. Encoder–decoder models such as U-Net (Ronneberger et al., 2015) are now standard for segmentation; task-specific backbones and losses extend to detection and registration. Generative models further broaden the scope to denoising, super-resolution, reconstruction, and cross-modality translation, including CT→MR synthesis.

Medical imaging poses domain-specific challenges. (i) *Data scale and labels:* datasets are smaller, labels are costly, and privacy limits sharing; open, harmonized multi-site data remain rare compared to natural-image corpora. (ii) *Distribution shift:* scanners, coils, sequences, and patient populations induce domain variability that hurts generalization and complicates naïve dataset merging. (iii) *Reliability:* subtle hallucinations or geometric drift—tolerable in natural images—are unacceptable in clinical settings.

Despite these hurdles, deep learning is already impactful: triage and anomaly detection in radiology; organ delineation and dose prediction in radiotherapy; longitudinal tracking in neurology/cardiology. Of particular relevance here, image-to-image translation can generate MR-like contrasts from CT when MR is scarce, enabling label transfer while keeping anatomy intact.

2.4 Image-to-image translation in medicine

Image-to-image translation learns a mapping that renders a source image with the appearance (contrast, texture statistics) of a target modality while preserving structure. Clinical constraints—no anatomical hallucinations, faithful boundaries, reproducible intensities—make the problem more stringent than in natural images.

Classical synthesis. Early approaches were registration-driven: after nonrigidly aligning a source to an atlas or exemplar with both modalities, intensity transfer or local patch regression (e.g., sparse coding, random forests) produced the target contrast. These pipelines depend on accurate registration and hand-crafted features, and tend to oversmooth or fail under protocol/pathology mismatch.

Supervised deep translation (paired). With spatially aligned pairs, encoder–decoder CNNs minimize voxelwise ℓ_1/ℓ_2 losses; conditional GANs (e.g., Pix2Pix (Isola et al., 2017)) add an adversarial term to sharpen textures while the ℓ_1 term anchors global intensities. High-quality pairs are rare in clinical data, limiting applicability.

Unpaired translation. Cycle-consistent GANs (e.g., CycleGAN (Zhu et al., 2017)) introduce two generators with (i) adversarial realism per domain and (ii) cycle consistency to approximately reconstruct the input. Identity losses discourage unnecessary changes if the input already resembles the target. While effective, pure cycle consistency can trade strict anatomical fidelity for realism; fine structures may drift without added constraints.

Structure preservation. Refinements include content–style disentanglement (UNIT/MUNIT/DRIT) to preserve shared anatomy while swapping contrast; patchwise contrastive objectives (CUT) to retain local content without explicit cycles; and structure-aware penalties (edge/gradient consistency, SSIM, segmentation- or registration-guided losses). Some works constrain deformations (invertible/diffeomorphic layers) or jointly train a downstream task to enforce task-faithfulness.

2D vs. 2.5D vs. 3D. Slice-based 2D models are memory-light but prone to through-plane inconsistency. 2.5D (multi-slice slabs) and cross-slice attention mitigate this. Fully 3D generators/discriminators enforce volumetric coherence at higher memory cost; overlapping 3D patches, mixed precision, and gradient checkpointing are common compromises in thoracic volumes.

Beyond GANs. Probabilistic translators (conditional VAEs, normalizing flows) model uncertainty (useful for one-to-many mappings like CT→multiple MR contrasts) but may blur without adversarial/perceptual terms. Diffusion models offer strong fidelity/diversity; conditioning on the source (and optionally semantics/edges) yields sharp, anatomy-consistent results at higher compute cost. Hybrid objectives (diffusion + structural losses) are emerging.

Pre-processing and evaluation. Robust preprocessing—orientation, resampling, HU windowing/clipping for CT, bias-field correction and robust normalization for MR—reduces domain shift and stabilizes training. Evaluation should combine image similarity (SSIM/PSNR/MAE), distributional realism (FID/KID with care in medical domains), and *task-based endpoints* (e.g., segmentation Dice/AP on synthetic images, dose error, reader studies). Uncertainty estimates (ensembles, MC dropout, diffusion variance) and guardrails (identity checks, residual/predict-the-contrast designs) mitigate silent failures.

2.5 GANs and CycleGAN Theory

In this part we will focus on presenting GANs and CycleGAN from a theoritical point of view such that it will guide us for the experiments later.

Standing assumptions. All distributions are defined on measurable spaces with a common base measure; equalities hold a.e. unless stated otherwise. Norms used in cycle losses are proper metrics (identity of indiscernibles).

Definition 2.1 (Pushforward (notation))**.** For a measurable map $T : \mathcal{A} \to \mathcal{B}$ and a measure μ on $\mathcal{A}$, the pushforward is $(T_{\#}\mu)(B) = \mu(T^{-1}(B))$ for measurable $B \subseteq \mathcal{B}$. We write $p_g := G_{\#}p_z$ in single-domain GANs and $G_{\#}p_X$, $F_{\#}p_Y$ for CycleGAN.

Generative Adversarial Nets: GAN

Let p_{data} be a distribution on $\mathcal{X}$ and p_z a prior on $\mathcal{Z}$. A generator $G_\theta : \mathcal{Z} \to \mathcal{X}$ induces a model distribution $p_g := G_{\#}p_z$. A discriminator $D_\phi : \mathcal{X} \to (0, 1)$ scores realness (for WGAN, use a critic $f_\phi : \mathcal{X} \to \mathbb{R}$).

Definition 2.2 (GAN minimax game)**.** The (original) GAN objective is (Goodfellow et al., 2014)

$$\min_\theta \max_\phi \; V(D_\phi, G_\theta) = \mathbb{E}_{x\sim p_{\text{data}}}\big[\log D_\phi(x)\big] + \mathbb{E}_{z\sim p_z}\big[\log\big(1 - D_\phi(G_\theta(z))\big)\big]. \tag{2.7}$$

Proposition 2.3 (Optimal discriminator)**.** *For fixed G_θ with model distribution p_g, the optimal discriminator is (Goodfellow et al., 2014)*

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}. \tag{2.8}$$

Theorem 2.4 (JS divergence equivalence)**.** *Plugging D^* into V yields (Goodfellow et al., 2014)*

$$V(D^*, G_\theta) \;=\; -\log 4 \;+\; 2\,\mathrm{JS}(p_{data} \,\|\, p_g)\,, \tag{2.9}$$

so minimizing the GAN game (ideally) minimizes the Jensen–Shannon divergence between p_{data} and p_g.

Remark 2.5 (Non-saturating generator loss). A practical alternative is the non-saturating generator loss $\mathcal{L}_G^{\mathrm{NS}}(\theta) = -\mathbb{E}_{z\sim p_z}[\log D_\phi(G_\theta(z))]$, which improves gradients while leaving the Nash equilibrium unchanged (Goodfellow et al., 2014).

Remark 2.6 (Why vanilla GANs can be unstable). If p_{data} and p_g lie on (nearly) disjoint low-dimensional manifolds, the discriminator can saturate, yielding vanishing generator gradients under JS (Arjovsky et al., 2017).

Definition 2.7 (Wasserstein-1 (Earth-Mover) distance)**.** For probability measures μ, ν on a metric space $(\mathcal{X}, d)$ with finite first moments,

$$W_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)\, \mathrm{d}\pi(x, y). \tag{2.10}$$

Theorem 2.8 (Kantorovich–Rubinstein duality)**.** *Under mild conditions,*

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}}\le 1} \Big(\mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{y\sim\nu}[f(y)]\Big). \tag{2.11}$$

(Villani, 2008)

Definition 2.9 (WGAN objective)**.** Let $f_\phi : \mathcal{X} \to \mathbb{R}$ be a 1-Lipschitz critic (no sigmoid). The WGAN problem is (Arjovsky et al., 2017)

$$\min_\theta \max_{\phi:\, \|f_\phi\|_{\mathrm{Lip}}\le 1} \mathbb{E}_{x\sim p_{\mathrm{data}}}[f_\phi(x)] - \mathbb{E}_{z\sim p_z}[f_\phi(G_\theta(z))], \tag{2.12}$$

which (approximately) minimizes $W_1(p_{\mathrm{data}}, p_g)$.

Proposition 2.10 (Gradient penalty as Lipschitz enforcement)**.** *A practical 1-Lipschitz surrogate is the WGAN-GP objective (Gulrajani et al., 2017)*

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{data}}[f_\phi(x)] + \mathbb{E}_{z\sim p_z}[f_\phi(G_\theta(z))] + \lambda\, \mathbb{E}_{\hat{x}}\big(\|\nabla_{\hat{x}} f_\phi(\hat{x})\|_2 - 1\big)^2, \tag{2.13}$$

with $\hat{x} = \epsilon x + (1-\epsilon)G_\theta(z)$, $\epsilon \sim \mathcal{U}[0, 1]$ sampled on lines between real and generated points.

Remark 2.11 (Spectral normalization). Dividing each linear/conv layer by its spectral norm enforces a global Lipschitz bound on f_ϕ and stabilizes training in practice (Miyato et al., 2018).

Definition 2.12 (Conditional GAN (cGAN))**.** Given side information y, define $G_\theta(z, y)$ and $D_\phi(x, y)$. Training is identical to the GAN game but conditioned on y (Mirza and Osindero, 2014).

Definition 2.13 (Least-Squares GAN (LSGAN))**.** Replace the discriminator's cross-entropy by a least-squares loss; with suitable label choices, this corresponds to minimizing a Pearson χ^2-type f-divergence surrogate (Mao et al., 2017).

Pointwise optimal discriminators for LSGAN and the induced Pearson-χ^2 surrogate are given in App. A.

Definition 2.14 (Distribution-level realism metrics)**. FID** (Fréchet Inception Distance): Fréchet distance between feature distributions of real vs. generated data (typically Inception features) (Heusel et al., 2017).
KID (Kernel Inception Distance): an unbiased MMD^2 estimate with a polynomial kernel, often more reliable on small sample sizes (Bińkowski et al., 2018).

Theorem 2.15 (Two-time-scale update rule (TTUR))**.** *Under stochastic-approximation assumptions, using distinct learning rates for D and G (with D updated faster) converges to a stationary local Nash equilibrium (Heusel et al., 2017).*

Cycle-Consistent Adversarial Networks : CycleGAN

Let $\mathcal{X}, \mathcal{Y}$ be two domains with distributions p_X, p_Y. Generators $G : \mathcal{X} \to \mathcal{Y}$ and $F : \mathcal{Y} \to \mathcal{X}$ induce pushforwards $G_{\#}p_X$ on $\mathcal{Y}$ and $F_{\#}p_Y$ on $\mathcal{X}$. Discriminators $D_Y : \mathcal{Y} \to (0,1)$ and $D_X : \mathcal{X} \to (0,1)$ score realness in each domain.

Definition 2.16 (CycleGAN objective (CE heads))**.** With cross-entropy (CE) adversarial heads (Zhu et al., 2017; Kim et al., 2017; Yi et al., 2017),

$$\mathcal{L}(G,F,D_X,D_Y) = \underbrace{\mathop{\mathbb{E}}_{y\sim p_Y}[\log D_Y(y)] + \mathop{\mathbb{E}}_{x\sim p_X}[\log(1-D_Y(G(x)))]}_{\mathcal{L}_{\mathrm{GAN}}(G,D_Y;X\to Y)} + \underbrace{\mathop{\mathbb{E}}_{x\sim p_X}[\log D_X(x)] + \mathop{\mathbb{E}}_{y\sim p_Y}[\log(1-D_X(F(y)))]}_{\mathcal{L}_{\mathrm{GAN}}(F,D_X;Y\to X)} + \lambda_{\mathrm{cyc}}\,\mathcal{L}_{\mathrm{cyc}}(G,F) + \lambda_{\mathrm{id}}\,\mathcal{L}_{\mathrm{id}}(G,F). \tag{2.14}$$

with

$$\mathcal{L}_{\mathrm{cyc}}(G,F) = \mathop{\mathbb{E}}_{x\sim p_X}\big[\|F(G(x))-x\|_1\big] + \mathop{\mathbb{E}}_{y\sim p_Y}\big[\|G(F(y))-y\|_1\big], \tag{2.15}$$

$$\mathcal{L}_{\mathrm{id}}(G,F) = \mathop{\mathbb{E}}_{y\sim p_Y}\big[\|G(y)-y\|_1\big] + \mathop{\mathbb{E}}_{x\sim p_X}\big[\|F(x)-x\|_1\big]. \tag{2.16}$$

Proposition 2.17 (Optimal discriminators (per domain, CE heads))**.** *For fixed G and F, the optimal discriminators are*

$$D_Y^*(y) = \frac{p_Y(y)}{p_Y(y) + (G_{\#}p_X)(y)}, \qquad D_X^*(x) = \frac{p_X(x)}{p_X(x) + (F_{\#}p_Y)(x)}. \tag{2.17}$$

Theorem 2.18 (Reduced objective under CE heads)**.** *Plugging D_Y^*, D_X^* into $\mathcal{L}$ yields*

$$\mathcal{L}(G,F,D_X^*,D_Y^*) = \big(-2\log 4\big) + 2\,\mathrm{JS}\big(p_Y \,\|\, G_{\#}p_X\big) + 2\,\mathrm{JS}\big(p_X \,\|\, F_{\#}p_Y\big) + \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}}(G,F) + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}(G,F). \tag{2.18}$$

Hence, at population level, CycleGAN jointly matches marginals in both directions while enforcing approximate invertibility via cycle loss.

Definition 2.19 (WGAN Cycle heads with gradient penalty)**.** Let $f_Y : \mathcal{Y} \to \mathbb{R}$ and $f_X : \mathcal{X} \to \mathbb{R}$ be 1-Lipschitz critics. Define the adversarial part as

$$\mathcal{L}_{\mathrm{W}}(G, F, f_X, f_Y) = - \mathop{\mathbb{E}}_{y \sim p_Y}[f_Y(y)] + \mathop{\mathbb{E}}_{x \sim p_X}[f_Y(G(x))] - \mathop{\mathbb{E}}_{x \sim p_X}[f_X(x)] + \mathop{\mathbb{E}}_{y \sim p_Y}[f_X(F(y))]. \quad (2.19)$$

With gradient penalties on segments between real and generated samples in each domain (Gulrajani et al., 2017),

$$\mathcal{L}_{\mathrm{GP}} = \mathop{\mathbb{E}}_{\hat{y}} \left(\|\nabla_{\hat{y}} f_Y(\hat{y})\|_2 - 1 \right)^2 + \mathop{\mathbb{E}}_{\hat{x}} \left(\|\nabla_{\hat{x}} f_X(\hat{x})\|_2 - 1 \right)^2, \quad (2.20)$$

and the full objective is

$$\min_{G,F} \max_{f_X, f_Y} \mathcal{L}_{\mathrm{W}}(G, F, f_X, f_Y) + \lambda_{\mathrm{gp}} \mathcal{L}_{\mathrm{GP}} + \lambda_{\mathrm{cyc}} \mathcal{L}_{\mathrm{cyc}}(G, F) + \lambda_{\mathrm{id}} \mathcal{L}_{\mathrm{id}}(G, F). \quad (2.21)$$

At population level, the adversarial part targets $W_1\big(p_Y, G_{\#}p_X\big) + W_1\big(p_X, F_{\#}p_Y\big)$ via Kantorovich–Rubinstein duality.

Remark 2.20 (Scope of divergences used in this report). In our experiments we *do not* use CE heads; we instantiate adversarial losses as LSGAN and sometimes WGAN-GP. For *WGAN(-GP)* heads, replace the JS terms above by $W_1(p_Y, G_{\#}p_X) + W_1(p_X, F_{\#}p_Y)$ (Kantorovich–Rubinstein duality Definition 2.8). For *LSGAN* heads, the adversarial terms correspond to a *Pearson* χ^2*-type* f-divergence surrogate and *do not* reduce to JS or W_1 (Mao et al., 2017).

Proposition 2.21 (Zero-cycle limit implies a.e. inverses)**.** *If* $\mathcal{L}_{\mathrm{cyc}}(G, F) = 0$ *with a metric norm (e.g.,* $\|\cdot\|_1$*), then* $F \circ G = \mathrm{id}_{\mathcal{X}}$ p_X*-a.s. and* $G \circ F = \mathrm{id}_{\mathcal{Y}}$ p_Y*-a.s. In particular, on the supports of* p_X *and* p_Y*,* G *and* F *are inverses almost everywhere.*

Remark 2.22 (Non-uniqueness / semantic ambiguity). Adversarial matching fixes only marginals; many measure-preserving bijections satisfy the GAN terms. Cycle consistency restricts to (almost) bijections but does not guarantee semantic alignment; additional priors (identity loss, geometry/contrast preservation, shared content encoders) help resolve ambiguity (Zhu et al., 2017; Liu et al., 2017; Benaim and Wolf, 2017).

Remark 2.23 (Identity loss). $\mathcal{L}_{\mathrm{id}}$ biases G to preserve content when the input is already in $\mathcal{Y}$ (and F for $\mathcal{X}$), stabilizing contrast and discouraging unnecessary changes (Zhu et al., 2017).

Remark 2.24 (PatchGAN discriminators). CycleGAN typically employs a PatchGAN (Markovian) discriminator that classifies local patches, acting as a learned texture/style prior and improving high-frequency realism (Isola et al., 2017; Zhu et al., 2017). *In our 3D setup, the effective receptive field is approximately* 96^3 *voxels (see §6.1), which encourages local MR-like appearance while limiting global warping.*

Practical takeaways for CT→MR. (i) Adversarial heads match *appearance* statistics of MR locally; (ii) Cycle+identity keep vessel/fissure geometry stable and curb drift; (iii) Capacity/regularization of 3D PatchGAN (e.g., spectral norm) balances sharpness vs. hallucinations; (iv) Even with cycles, exact identifiability is not guaranteed—structure-aware checks (edges/SSIM, round-trip error, task-based validation) remain essential.

2.6 Related Work

Multiple approaches exist for image-to-image translation in medical imaging. In our context, we aim to synthesize one imaging modality from another—specifically, magnetic resonance (MR) scans from computed tomography images. This objective situates our work within the subdomain

of image-to-image translation, a field that has gained significant interest primarily due to the development of conditional GANs (cGANs) (Mirza and Osindero, 2014).

These models extend the adversarial framework of GANs by incorporating an input condition, guiding the image generation process. Unlike traditional GANs that generate images from random noise, cGANs condition the generation on a given input image. This conditioning forces the generator to learn a mapping from the input to the desired output image, effectively enabling translation between domains. The generator in cGANs often adopts a U-shaped architecture (Ronneberger et al., 2015), designed to capture both global context and fine-grained details.

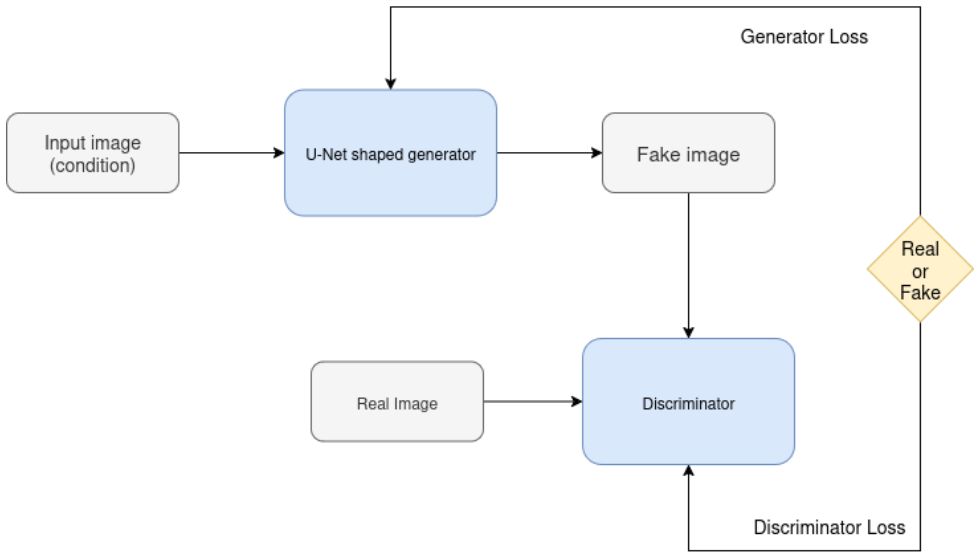


Figure 1 – Example of a paired conditional GAN (cGAN) architecture for image translation using a U-Net backbone

Conditional GANs can be categorized into two fundamental approaches: **paired** and **unpaired**. Paired methods rely on having matched image pairs—in our case, CT and MRI scans of the same patient acquired under similar conditions. Unpaired methods, which apply to our scenario, operate without such correspondence between images. This distinction is particularly valuable in medical contexts, where acquiring multiple imaging modalities for the same patient is often costly, time-consuming, or clinically unnecessary.

In this section, we present both approaches but focus primarily on unpaired methods, as these formed the core of our work during this internship.

2.6.1 Paired Approaches

Early methods for cross-modal image synthesis in medical imaging employed U-Net architectures trained with L1 loss functions to learn mappings between input and output images (Li et al., 2020). The U-Net architecture (Ronneberger et al., 2015) is an encoder-decoder structure with skip connections that captures both global context and fine-grained details, making it well-suited for medical image translation tasks. Methods utilizing this approach aimed to minimize pixel-wise differences between synthesized and target images using L1 loss:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[||y - G(x)||_1] \tag{2.22}$$

where G is the generator network, x the input image, and y the target image. However, these methods often produced blurry results due to the limitations of L1 loss in modeling high-frequency details.

The pix2pix model (Isola et al., 2017) represented a significant advancement in image-to-image translation by integrating a discriminator network into the U-Net architecture, forming a paired conditional GAN (cGAN). This model enhanced the quality of synthesized images by introducing an adversarial loss that encourages the generator to produce outputs indistinguishable from real images, adding a realism constraint and counterbalancing the tendency of L1 loss to produce blurred outputs. The adversarial loss in pix2pix is defined as:

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_x[\log(1 - D(G(x)))] \tag{2.23}$$

where D is the discriminator network that evaluates the authenticity of the generated image $G(x)$ conditioned on the input image x.

In medical imaging, pix2pix has been successfully applied to tasks such as MR-to-CT synthesis. For instance, Nie et al. (Nie et al., 2017) employed a pix2pix-based framework, and their results demonstrated that incorporating adversarial loss significantly enhanced the sharpness and structural fidelity of synthesized images compared to models using only L1 loss.

2.6.2 Unpaired Approaches

Unpaired methods tackle image-to-image translation without corresponding image pairs, making them particularly suited for medical imaging scenarios where acquiring multiple modalities is costly or where certain imaging methods carry clinical constraints. These methods enable models to learn mappings between domains using unpaired data, thereby broadening the applicability of image translation techniques.

Multiple approaches have been explored, including diffusion models, physics-informed methods, and energy-based models. One of the most established methods, and the one we employ in this work, is the CycleGAN approach.

CycleGAN (Zhu et al., 2017) introduced cycle-consistency loss in addition to adversarial losses to facilitate unpaired image-to-image translation. The architecture consists of two generators, $G : X \to Y$ and $F : Y \to X$, and two discriminators, D_Y and D_X. The generators learn bidirectional mappings between the source domain X and target domain Y.

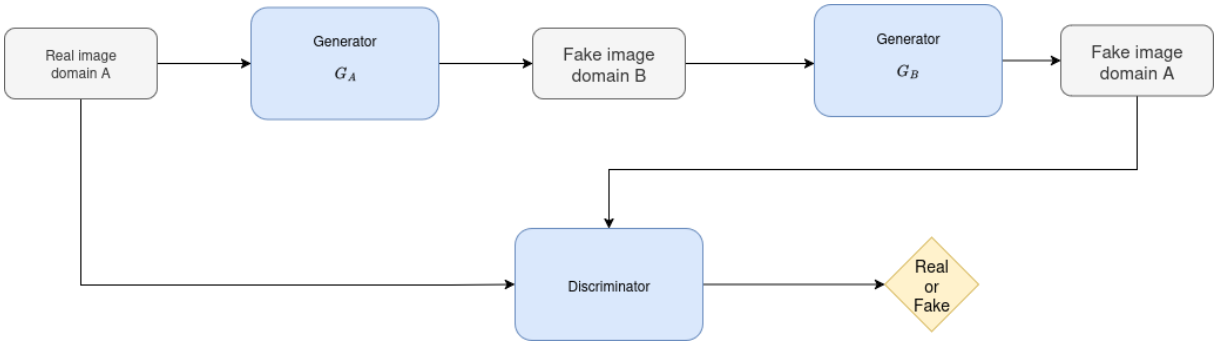


Figure 2 – CycleGAN architecture illustrating cycle consistency for unpaired image translation, showing two generators, two discriminators, and bidirectional cycle paths

The cycle-consistency loss enforces that an image translated from one domain and back should return to the original:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] + \mathbb{E}_{y \sim p_{data}(y)}[||G(F(y)) - y||_1] \tag{2.24}$$

Additionally, identity loss preserves color composition when the input already belongs to the target domain:

$$\mathcal{L}_{identity}(G, F) = \mathbb{E}_{y \sim p_{data}(y)}[||G(y) - y||_1] + \mathbb{E}_{x \sim p_{data}(x)}[||F(x) - x||_1] \tag{2.25}$$

The full CycleGAN objective combines adversarial losses for both directions with cycle-consistency and identity losses:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{id}\mathcal{L}_{identity}(G, F) \tag{2.26}$$

where λ_{cyc} and λ_{id} control the relative importance of cycle-consistency and identity preservation.

CycleGAN has found extensive application in medical imaging cross-modal translation due to its ability to learn from unpaired data. Recent extensions have incorporated structure-aware losses such as SSIM, gradient-based losses, and segmentation-guided constraints to better preserve anatomical structures. Furthermore, the choice between 2D, 2.5D, and full 3D implementations presents trade-offs between computational efficiency and volumetric consistency.

However, literature specifically addressing CT-to-MRI synthesis remains limited, with even fewer studies focusing on 3D applications, and virtually none targeting thoracic anatomy. Notable exceptions include the work of (J. Wang et al., 2023) and (Hiasa et al., 2018), who extended the approach for MR image synthesis from CT scans on brain and orthopedic images, respectively.

Our work addresses this gap by tackling 3D thoracic CT-to-MR synthesis, which presents unique challenges including memory constraints from volumetric processing, the need for anatomically consistent neighboring slice relationships, and the preservation of complex thoracic structures across modalities. These requirements necessitate careful adaptation of existing unpaired translation frameworks to handle the specific demands of thoracic imaging.

3 Dataset

3.1 CT dataset: LUNA16

The CT data are drawn from the LUng Nodule Analysis 2016 (LUNA16) challenge dataset (Setio et al., 2017). LUNA16 is derived from the publicly available LIDC-IDRI collection, consisting of thoracic CT scans acquired at multiple institutions with varying scanner models and protocols. The dataset was originally curated for lung nodule detection and analysis, but it provides full-field-of-view (FOV) chest CT volumes with voxel-level annotations of nodules. Scans are low-dose, thin-slice (typically $< 1.5\,\text{mm}$), and stored as 3D volumes with associated spacing metadata. For our purposes, the nodular annotations are not used; instead, we rely on the raw CT volumes as source modality for image translation. The advantage of LUNA16 is its standardized preprocessing and its relatively large cohort size (∼1000 scans), which provides good coverage of anatomical variability and chest pathologies.

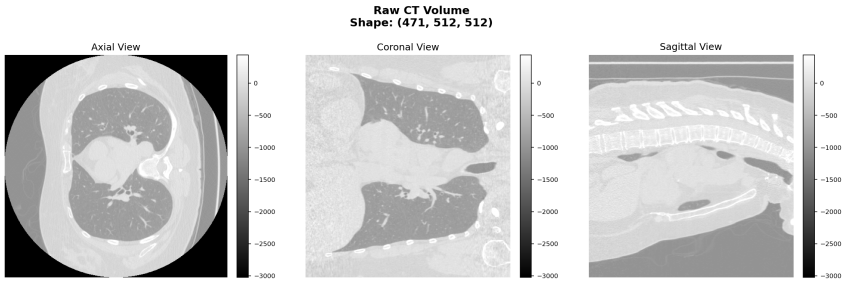


Figure 3 – Example CT slice from LUNA16 (raw intensity values in Hounsfield Units).

3.2 MRI dataset: UTE lung MRI

On the MRI side, we use ultrashort echo time (UTE) MRI acquisitions of the thorax. UTE sequences are particularly suitable for lung imaging because they capture signal from tissues with very short T_2^* relaxation times, such as lung parenchyma, which are usually invisible in conventional MRI. The raw MRI data include multiple respiratory gates (temporal phases of the breathing cycle), enabling retrospective motion analysis. In our setup, we select two gates (gate 8 and gate 15), both to artificially augment the dataset size and to ensure approximate balance with the number of CT scans. This choice also increases anatomical variability across subjects while maintaining consistent acquisition conditions. The selected MRI volumes are stored in HDF5 format with voxel spacing metadata and subsequently preprocessed following the pipeline described in Section 4.

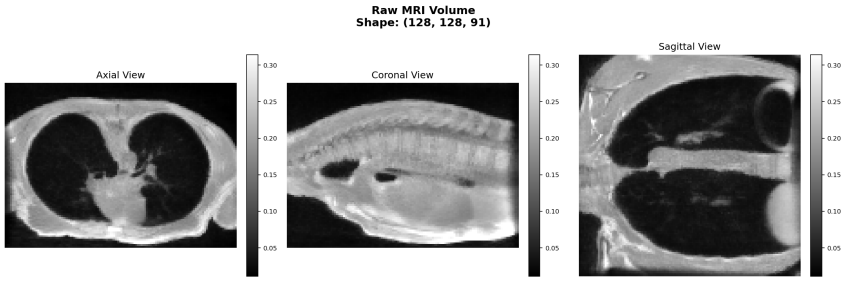


Figure 4 – Example raw UTE MRI slice before preprocessing (arbitrary intensity units).

3.3 Challenges in Thoracic CT-to-MR Translation

Cross-modal synthesis has demonstrated success in other anatomical regions, particularly brain imaging (Nie et al., 2017; Wolterink et al., 2017; Jin et al., 2019). However, applying these methods to pulmonary imaging presents unique challenges that distinguish thoracic CT-to-MR translation from other body regions.

Dataset limitations. A primary challenge is the lack of precedent and the scarcity of data specific to lung UTE MRI. To our knowledge, no prior work has attempted to synthesize lung UTE MR from CT; UTE MRI remains relatively new in thoracic applications and data remain limited (Johnson et al., 2013). Publicly available *paired* thoracic CT–MR datasets are effectively nonexistent, which hinders progress; robust datasets are essential for training deep learning models (Weiss et al., 2016).

Image quality differences. Chest CT is the clinical workhorse, offering high spatial resolution and low noise with crisp depiction of lung structures. In contrast, UTE MR volumes typically exhibit lower SNR, bias-field inhomogeneity, and motion-related blur due to respiratory dynamics (our dataset comprises 32 gated phases). Compared with CT, UTE MR commonly shows higher apparent noise/artifacts and lower spatial resolution; a translator must reproduce these modality-specific appearances (for realism) without distorting anatomy (Weiss et al., 2016).

Anatomical specificity. The thorax contains numerous fine, branching structures (vessels, airways, bronchi) that are susceptible to hallucinations or boundary drift in cross-modality synthesis. Methods must preserve geometry at high fidelity while altering only modality appearance (contrast/texture), a requirement we address via cycle/identity terms and a no-warping design (Sections 4–5).

3.4 Balance and preprocessing

We keep the number of CT and MR volumes approximately matched to limit adversarial bias. Both modalities undergo the same preprocessing (near-isotropic 1 mm resampling, symmetric padding to multiples of K, and modality-appropriate normalization), ensuring a common voxel grid and dynamic range and reducing domain shift unrelated to modality appearance (Section 4).

	Train	Val	Test
# CT patients	311	66	68
# MR patients	292	60	70

Table 1 – Dataset composition across train, validation, and test splits. Spacing is reported in millimeters.

4 Pre-processing & data pipeline

Our goal is to deliver full-FOV, geometrically consistent 3D volumes on a common voxel grid and a common intensity range suitable for GAN training, while keeping all steps reproducible. Concretely, each volume is resampled to near-isotropic 1 mm spacing, padded so that all dimensions are multiples of K (default K=16), and normalized to $[-1, 1]$ with modality-specific rules. The pipeline is implemented once and applied identically to CT and MR inputs, with the modality-dependent normalization in the final step.

Inputs & metadata. Volumes are provided as HDF5 files with a `volume` dataset and HDF5 attributes storing metadata (e.g., voxel spacing, optional origin/direction). At load time, we (i) read the 3D array in $z \times y \times x$ order, (ii) recover voxel spacing from any of several expected attribute keys (e.g., `spacing`, `spacing_zyx`), and (iii) fall back to $[1, 1, 1]$ mm with a warning if spacing is unavailable.

4.1 Orientation unification and spacing harmonization

Anatomical orientation. We reorient all images to RAS convention using the direction cosine matrix, applying axis permutations and flips as needed. We verify correct orientation by checking anatomical sidedness (e.g., heart/liver position on CT) to prevent left–right inversions that could compromise unpaired translation quality. We have to be careful to have both of our images, CT and MRI, having the same orientation.

Isotropic resampling to 1 mm. Given an input image with spacing $s = (s_x, s_y, s_z)$ and size $n = (n_x, n_y, n_z)$, we resample with trilinear interpolation to target spacing $t = (1, 1, 1)$ mm and compute the new integer size rounded to nearest voxels:

$$n'_d \;=\; \mathrm{round}\Big(\frac{n_d\, s_d}{t_d}\Big), \qquad d \in \{x, y, z\}. \tag{4.1}$$

Trilinear interpolation reduces aliasing with limited smoothing compared to nearest neighbor, while being faster and less smoothing than higher-order B-splines for full-FOV 3D resampling. We set the resampler's *default pixel value* to a robust background estimate (median for floating images, minimum for integer images) to avoid artificial bright halos at boundaries.

Why harmonize spacing? GAN discriminators and patch-based generators are sensitive to voxel size. Harmonizing to 1 mm reduces covariate shift, enables fixed patch sizes, and improves the meaning of structure-aware losses (e.g., gradient) by putting them on the same spatial scale.

It also mitigates through-plane anisotropy that would otherwise cause slice-to-slice flicker during 3D synthesis.

4.2 Padding to a multiple of K

Most 3D U-Net–style backbones downsample by factors of two at each encoder level. If there are L levels, each spatial dimension must be divisible by 2^L to avoid border artifacts and ad-hoc cropping. We therefore pad *symmetrically* so that

$$n''_d \;=\; \left\lceil \frac{n'_d}{K} \right\rceil K, \qquad K \in \{16, 32\} \text{ typ.} \tag{4.2}$$

Padding uses the volume's minimum intensity (background) to remain visually unobtrusive and to avoid creating artificial edges at the margins. Symmetric padding keeps anatomy centered and plays nicely with sliding-window inference. Padding parameters are recorded as HDF5 attributes for potential de-padding during post-processing.

4.3 Modality-specific intensity normalization

Training stability benefits greatly from mapping inputs to a compact, consistent range. We normalize *after* resampling and padding, so the stored volumes are ready to feed the network.

CT (Hounsfield Units). We clip to a thorax-appropriate HU window and linearly map to $[-1, 1]$:

$$\tilde{x} \;=\; \mathrm{clip}(x,\ \mathrm{HU}_{\min}, \mathrm{HU}_{\max}), \qquad x^{\mathrm{norm}} \;=\; 2\frac{\tilde{x} - \mathrm{HU}_{\min}}{\mathrm{HU}_{\max} - \mathrm{HU}_{\min}} - 1. \tag{4.3}$$

We use $\mathrm{HU}_{\min}$=−600, $\mathrm{HU}_{\max}$=1000 by default. Rationale: lung parenchyma (air ≈ -1000 HU) and soft tissues (≈ -100 to $+100$ HU) are preserved, while extremely high bone and metal outliers are clipped to stabilize adversarial training. This wider-than-lung-only window keeps mediastinum and ribs visible, which helps the model respect chest wall geometry. The window is a configurable parameter for sensitivity analyses.

MRI (per-volume robust scaling). Absolute MR intensities are arbitrary and vary with coil gain, sequence, and site. We therefore apply a robust per-volume affine map defined by lower/upper percentiles (p_1, p_{99}):

$$\tilde{y} \;=\; \mathrm{clip}(y,\ P_1(y), P_{99}(y)), \qquad y^{\mathrm{norm}} \;=\; 2\frac{\tilde{y} - P_1(y)}{P_{99}(y) - P_1(y)} - 1. \tag{4.4}$$

Percentile clipping (default 1–99%) suppresses rare bright spikes (e.g., fat peaks, artifacts) without handcrafting sequence-specific windows and provides stable ranges for GAN losses. Quality control statistics for our dataset are summarized in Table 2.

Modality	Clip low (%)	Clip high (%)	Median P_1	Median P_{99}	Flagged volumes
CT	–	–	–	–	0/642
MRI	1.0	99.0	85.3	892.1	2/847

Table 2 – Quality control statistics after percentile clipping.

If $P_{99} \approx P_1$ (flat image), we emit a warning, return zeros, and skip the volume during training to prevent degenerate batches.

4.4 Cropping and masking

We intentionally *preserve the full FOV* during preprocessing. Unpaired translation should act as an appearance transform without geometry changes; cropping at this stage risks leaking dataset-specific field-of-view differences (e.g., scanner couch presence/absence) into the adversarial signal and can break consistency between CT and MR distributions.

For quality control reporting, histograms are computed inside a body or lung mask to avoid background bias; however, full-FOV volumes are always stored and fed to training.

4.5 Edge cases, safeguards, and failure modes

— **Missing/ambiguous spacing.** If no usable spacing attribute is present, we default to 1 mm and log a warning. This preserves array geometry while avoiding silent mis-scaling.

— **Header/axis mistakes.** Because `GetImageFromArray` uses zyx memory order while SimpleITK spacing is xyz, we explicitly reverse the tuple when setting spacing to keep physical coordinates correct.

— **Interpolation boundary artifacts.** We set a robust default pixel value in resampling to avoid bright/negative halos at the borders.

— **Outliers and degenerate cases.** MRI robust scaling handles intensity spikes; CT windowing handles metal/bone. If $P_{99} \approx P_1$ (flat image), we log a warning, return zeros, and skip the volume during training to prevent degenerate batches.

4.6 Output format and reproducibility

To ensure complete reproducibility and facilitate debugging, our pipeline implements comprehensive metadata tracking and standardized output formats.

Each preprocessed volume is written as HDF5 with: `volume` (float32, $[-1, 1]$) and attributes documenting (i) original source file and patient/study identifiers if available, (ii) final array shape, (iii) final spacing (mm), (iv) all preprocessing parameters (target spacing, K, CT window or MRI percentiles), (v) software versions and git commit hash, and (vi) full configuration dump and timestamp. This exhaustive metadata enables exact reproduction of any preprocessing run and facilitates systematic ablation studies.

Logs printed at `INFO/DEBUG` level summarize per-volume ranges and chosen percentiles, enabling auditability and exact reruns. This logging strategy has proven invaluable for identifying problematic volumes and understanding preprocessing effects on different acquisition protocols.

4.7 Design trade-offs and rationale

The key design choices and their justification are summarized in Table 3.

4.8 Remarks on contrast and where to find visualizations

A key observation from exploratory plots is that the perceived MR contrast is closely tied to the *global intensity histogram*. We will explicitly leverage this connection in our method via a differentiable histogram term (see Section 7.7). For readability and space, all pre-processing visualizations have been moved to Appendix B.

Table 3 – Design choices for preprocessing pipeline and their rationale

Choice	**Rationale**	**Effect on training**
1 mm target spacing	Common, simple target that preserves vessel calibers and fissures while keeping memory reasonable	Enables fixed patch sizes, reduces covariate shift
Trilinear interpolation	Faster and less smoothing than higher-order B-splines for full volumes; adequate for appearance learning	Preserves edges better than nearest neighbor with minimal computational overhead
Pad-to-K with background	Guarantees compatibility with 2^L encoder strides without artificial edges	Avoids border artifacts and prevents discriminator exploitation of padding patterns
CT window $[-600, 1000]$ HU	Wider than pure lung window to preserve chest wall and mediastinal context	Stabilizes adversarial training while maintaining anatomical constraints
MRI percentile scaling	Sequence/site agnostic, robust to coil gain and occasional artifacts	Provides consistent dynamic range without requiring sequence metadata

Algorithm 1 Volume Preprocessing Pipeline

procedure PREPROCESSVOLUME(x, modality)
  Load HDF5 volume and metadata
  Reorient to RAS anatomical convention using direction cosines
  Resample to 1 mm isotropic (trilinear interpolation)
  Pad symmetrically to multiple of K with background intensity
  if modality = CT **then**
    Apply HU window $[-600, 1000]$ and normalize to $[-1, 1]$
  else if modality = MRI **then**
    Apply percentile scaling (P_1, P_{99}) to $[-1, 1]$
  end if
  Record comprehensive metadata (spacing, padding, normalization params)
  Save HDF5 with float32 array in $[-1, 1]$ range
  Log intensity ranges and percentiles for quality control
  return preprocessed volume
end procedure

5 Problem definition

Domains and data. We consider unpaired, volumetric cross-modality translation in the thorax. Let

$$\mathcal{X} \subset \mathbb{R}^{H\times W\times D} \quad \text{and} \quad \mathcal{Y} \subset \mathbb{R}^{H\times W\times D} \tag{5.1}$$

denote, respectively, the spaces of preprocessed chest *CT* and chest *MR* volumes. By preprocessing we mean a common voxel grid (near-isotropic spacing), consistent orientation, and modality-specific intensity normalization (CT in HU windowed/normalized; MR bias-field correction and robust scaling). We draw finite, *unpaired* samples

$$\mathcal{D}_{\mathrm{CT}} = \{x_i\}_{i=1}^{N_X} \overset{\text{i.i.d.}}{\sim} p_{\mathcal{X}} \quad \text{and} \quad \mathcal{D}_{\mathrm{MR}} = \{y_j\}_{j=1}^{N_Y} \overset{\text{i.i.d.}}{\sim} p_{\mathcal{Y}}, \tag{5.2}$$

with no known one-to-one correspondence between any x_i and y_j. In our setting, $\mathcal{Y}$ targets a UTE-like chest MR appearance.

Mappings (CycleGAN setting). We learn two parametric generators

$$G_\theta : \mathcal{X} \to \mathcal{Y} \qquad \text{and} \qquad F_\phi : \mathcal{Y} \to \mathcal{X}, \tag{5.3}$$

and two discriminators D_Y (for realism in $\mathcal{Y}$) and D_X (realism in $\mathcal{X}$). Training follows the CycleGAN paradigm: adversarial objectives encourage $G(x)$ to be indistinguishable from real y and $F(y)$ from real x, while cycle consistency promotes approximate invertibility,

$$F\big(G(x)\big) \approx x, \qquad G\big(F(y)\big) \approx y. \tag{5.4}$$

Identity terms discourage unnecessary changes on in-domain inputs,

$$G(y) \approx y, \qquad F(x) \approx x. \tag{5.5}$$

All networks operate on overlapping 3D patches or sub-volumes with sliding-window aggregation to respect memory limits while preserving through-plane context.

Desired properties (anatomy- and geometry-preserving appearance change). The intended solution is an *appearance* transform constrained by:

- **(P1) Anatomical consistency** For any $x \in \mathcal{X}$, $G(x)$ must preserve geometry (locations, topology, and boundaries of anatomical structures). Practically, we require that G does not introduce spatial deformations; e.g., edge/gradient structure, centerlines, and segmentation masks computed on x should be consistent on $G(x)$.
- **(P2) Target-domain realism.** $G(x)$ should match the contrast, texture, and intensity statistics of $\mathcal{Y}$ (UTE-like lung MR).
- **(P3) Cycle/identity self-consistency.** Round-trip reconstructions $F(G(x))$ and $G(F(y))$ should be close to inputs, and $G(y), F(x)$ should be near-identity.
- **(P4) Volumetric coherence.** Synthesized volumes must be consistent across slices and along vessels/fissures (no slice-wise flicker).

Learning objective. Let $\mathcal{L}^Y_{\mathrm{adv}}(G, D_Y)$ and $\mathcal{L}^X_{\mathrm{adv}}(F, D_X)$ be adversarial losses in $\mathcal{Y}$ and $\mathcal{X}$ (e.g., logistic or least-squares GAN). Cycle and identity terms read

$$\mathcal{L}_{\mathrm{cyc}} = \mathbb{E}_x\big[\|F(G(x)) - x\|_1\big] + \mathbb{E}_y\big[\|G(F(y)) - y\|_1\big], \quad \mathcal{L}_{\mathrm{id}} = \mathbb{E}_y\big[\|G(y) - y\|_1\big] + \mathbb{E}_x\big[\|F(x) - x\|_1\big]. \tag{5.6}$$

Evaluation without pairs. Because (x, y) pairs are unavailable, voxelwise ℓ_1/ℓ_2 between $G(x)$ and a true y cannot be computed. We therefore combine four complementary strategies:

- **E1 Distributional realism (domain match).** Compute patch-based FID/KID between $\{G(x_i)\}$ and $\{y_j\}$
- **E2 Anatomy preservation (self-consistency).** Round-trip error: $\|F(G(x))-x\|$ (PSNR/SSIM). Structure metrics on $(x, G(x))$ that are insensitive to contrast (edge correlation, gradient magnitude similarity). Registration-based checks with a small-regularization deformable registration from x to $G(x)$: the required deformation should be near zero if no warping is introduced.
- **E3 Task-based validation (label transfer).** When CT labels $m(x)$ exist (e.g., vascular tree), train a 3D segmenter on $(G(x), m(x))$ and evaluate on: (i) held-out synthetic MR; (ii) available labeled real MR (if any); or (iii) proxy targets (centerline continuity, lumen diameter monotonicity, branch count) on real MR. Gains here indicate that G produces *useful* MR for downstream tasks.
- **E4 Expert/qualitative assessment and robustness.** Reader studies focused on hallucinations and boundaries; stress tests across acquisition sites and protocols; ablations (architectures/losses/normalizations) to probe failure modes.

6 Methodology

Scope. We refer to Sec. 5 for the formal task, assumptions, and notation. Below we detail the model design space, architectures (including our 3D adaptation of DC-CycleGAN), training objectives, optimization protocol, inference/post-processing, and the evaluation plan used in Sec. 7.

6.1 Design space and model variants

Our approach factorizes the method into three orthogonal choices that can be recombined: *(i) generator backbone*, *(ii) adversarial head* (GAN objective), and *(iii) structural losses* (cycle, identity, and optionally SSIM-3D). This factorization lets us start from a simple, stable baseline and then introduce capacity or constraints only when they address specific failure modes of CT→UTE-MR translation under the *appearance-only, no-warping* assumption (cf. Sec. 5).

Why these models? Intuition and hypotheses.

— **Baseline: 3D ResNet + LSGAN + ℓ_1.** Residual generators are a natural starting point for *appearance remapping* tasks where content should remain intact: skip-style residual paths bias the network toward near-identity mappings when appropriate, reducing geometric drift and discouraging hallucinations. LSGAN (Mao et al., 2017) provides smoother gradients than the original logistic GAN, often yielding stable training and fewer saturated critics for medical volumes. The ℓ_1 cycle term anchors the mapping to preserve coarse anatomy.

— **Adversarial head swap: WGAN-GP (same backbone).** Wasserstein critics supply informative gradients even when real/fake supports are far apart (common across CT vs. UTE-MR), while the gradient penalty (Gulrajani et al., 2017) enforces a soft Lipschitz constraint for stability. *Hypothesis:* at equal capacity, WGAN-GP reduces texture artifacts (e.g., checkerboards) and mode collapse; the trade-off is that overly strong critics can underemphasize subtle MR contrast unless structural priors are present. Furthermore, this is intution is backed by theory 2.10.

— **Backbone swap: 3D U-Net++.** U-Net++ densifies skip connections and fuses multi-scale features, which can better preserve fine pulmonary structure (vessels, fissures) while maintaining global context. *Hypothesis:* multi-scale fusion + structure-aware losses yields higher structural fidelity in UTE; however, abundant skips can tempt *identity copying* from CT, so an identity term helps prevent under-translation.

— **Adapted DC-CycleGAN (from (J. Wang et al., 2023)).** The original DC-CycleGAN introduces *dual contrast* to improve cross-domain contrast transfer while maintaining content. We adapt its spirit to volumetric lung imaging: fully 3D convolutions, 3D PatchGAN critics, resize-conv upsampling, and a structure-aware cycle objective (Mix(ℓ_1, SSIM-3D)). *Hypothesis:* an explicit emphasis on *contrast* (via adversarial supervision + dual contrast) and *structure* (via SSIM in cycle) better matches UTE-MR intensity characteristics without warping geometry. Differences from the original design are detailed in Sec. 6.2.1.

Why 3D PatchGAN for the discriminator? Thoracic synthesis quality is judged locally (parenchyma texture, vessel walls, fissure edges). 3D PatchGAN critics operate on sub-volumes, pushing the generator to match *local* MR statistics (high-frequency contrast, micro-texture) without incentivizing global geometric changes—consistent with our no-warping constraint. Practically, patch critics reduce memory cost on 3D inputs and allow spectral normalization (Miyato et al., 2018) or gradient penalties to stabilize training.

Structural losses: role and intuition. The ℓ_1 cycle loss *anchors content* and discourages topology changes. Adding SSIM-3D within the cycle term biases reconstructions toward preserving local luminance/contrast relationships (windowed means/variances), which aligns with UTE's short-T_2^* parenchymal signal and subtle textures. Identity loss penalizes unnecessary modifications to already in-domain inputs (e.g., feeding MR to the CT→MR generator), thereby calibrating intensity shifts and reducing over-translation.

Named model variants (used throughout Sec. 7). Rather than exhaust the full Cartesian grid, we evaluate four representative systems that instantiate distinct hypotheses:

ID	Backbone	Adv. head	Cycle / Identity	Intuition / Hypothesis
M1	ResNet (9b)	LSGAN	ℓ_1 / λ_{id}=0.2	Stable, content-preserving baseline for appearance remapping.
M2	ResNet (9b)	WGAN-GP	ℓ_1 / λ_{id}=0.2	Better critic geometry ⇒ fewer artifacts, stronger realism at same capacity.
M3	U-Net++	WGAN-GP	ℓ_1+SSIM-3D / λ_{id}=0.2	Multi-scale fusion + structure-aware cycle ⇒ improved pulmonary micro-structure.
M4	DC-CycleGAN (3D, adapted)	LSGAN	ℓ_1+SSIM-3D / λ_{id}=0.2	Dual contrast targets MR contrast while preserving geometry (no warping).

Table 4 – Model variants. Components are specified in Secs. 6.2–6.3. The last column states the core hypothesis each variant tests.

Reporting convention. We refer to methods by their IDs (M1–M4) in tables and figures. Unless stated otherwise, all training protocol details (patch size, optimizer, schedule, inference tiling) are shared across variants and specified once in Sec. 7.1. Formal definitions of adversarial heads and structural losses are given in Sec. 6.3.

6.2 3D architectures

3D ResNet (content-preserving). Encoder–bottleneck–decoder with 9 residual blocks at the bottleneck; strided convolutions downsample, resize–conv upsamples (to avoid checkerboards). Inductive bias favors smooth appearance remapping with limited geometric drift; good at preserving vessels and fissures under the no-warping constraint.

3D U-Net++ (multi-scale skip fusion). Nested skip connections and dense intermediate nodes improve feature fusion across scales, which helps retain fine pulmonary structure at the cost of higher parameters and memory footprint. The dense skips are mitigated with identity loss to avoid under-translation.

6.2.1 DC-CycleGAN (3D, adapted from (J. Wang et al., 2023))

Intuition. CycleGAN enforces realism in the *target* domain and round-trip consistency, but gives no explicit signal that pushes synthesized samples *away from the source* domain. DC-CycleGAN adds a second, complementary contrast in the critic: real target images are labeled as `real`, while *both* generator outputs and randomly sampled *source-domain* images are labeled as `fake`. This *dual contrast* teaches D to carve a margin between source and target domains and nudges G to leave the source style decisively—improving cross-modality contrast transfer without requiring paired data.

Original (2D) formulation. Let $G : X \to Y$ and $F : Y \to X$ with discriminators D_Y, D_X. Besides the usual adversarial and cycle terms, DC adds the *dual contrast* (DC) losses by feeding random source samples $x' \sim p_X$ to D_Y (and $y' \sim p_Y$ to D_X) with a `fake` label:

$$\mathcal{L}_{\mathrm{DC}}(D_Y; X, Y) = \mathbb{E}_{x' \sim p_X}[\log(1 - D_Y(x'))], \quad \mathcal{L}_{\mathrm{DC}}(D_X; Y, X) = \mathbb{E}_{y' \sim p_Y}[\log(1 - D_X(y'))]. \tag{6.1}$$

The final objective couples adversarial, cycle, and DC losses (see Eqs. (4)–(12) of (J. Wang et al., 2023)):

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = {} & \mathcal{L}_{\mathrm{GAN}}(G, D_Y; X, Y) + \mathcal{L}_{\mathrm{GAN}}(F, D_X; Y, X) \\ & + \beta\big[\mathcal{L}_{\mathrm{DC}}(D_Y; X, Y) + \mathcal{L}_{\mathrm{DC}}(D_X; Y, X)\big] + \lambda\, \mathcal{L}_{\mathrm{cyc}}(G, F).\end{aligned} \tag{6.2}$$

with λ for cycle-consistency and β for DC.

Our 3D adaptation for thoracic CT↔UTE-MR (appearance-only). We port the design to volumes and to our no-warping constraint:

- **3D generators.** Encoder–bottleneck–decoder with 3D convolutions (ResNet-style residual blocks), strided convs for downsampling, resize–conv for upsampling (anti-checkerboard).
- **3D PatchGAN critics.** Local 3D receptive fields judge sub-volumes; spectral normalization on all convs for stability. The DC pathway injects random source-domain 3D patches as `fake` alongside $G(x)/F(y)$.

— **Objectives.** We keep the DC mechanism and SSIM-augmented cycle; the *adversarial head* may be LSGAN or WGAN-GP, but according the original paper we will go wtih Cross-Entropy instead (defined in Sec. 6.3). In either case, DC means "treat source-domain samples as fake for the target-domain critic"; with WGAN-GP this corresponds to giving source samples the same sign as generated samples in the critic loss (plus gradient penalty).

— **Identity.** We add $\mathcal{L}_{\text{id}}$ to limit unnecessary contrast shifts on already in-domain inputs.

Why it fits UTE-MR. UTE lung MR emphasizes very short-T_2^* parenchymal signal and subtle vessel/fissure contrast. DC explicitly discourages "CT-like" outputs by pushing them away from the source distribution, while SSIM-in-cycle preserves local luminance/contrast relationships; together they better match UTE appearance without any geometric warping.

Objective used in this work (compact). Let $\mathcal{L}_{\text{GAN}}^{(\cdot)}$ denote either LSGAN or WGAN-GP (Sec. 6.3) and $\mathcal{L}_{\text{cyc-struct}} = \mathcal{L}_{\text{cyc}} + \alpha\,[(1-\text{SSIM}_3(F(G(x)),x)) + (1-\text{SSIM}_3(G(F(y)),y))]$. Our 3D DC-CycleGAN objective is:

$$\begin{aligned}\min_{G,F}\max_{D_X,D_Y} \quad & \underbrace{\mathcal{L}_{\text{GAN}}^{(\cdot)}(G,D_Y;X,Y) + \mathcal{L}_{\text{GAN}}^{(\cdot)}(F,D_X;Y,X)}_{\text{realism in target domains}} \\ & + \beta\,\underbrace{\big[\mathcal{L}_{\text{DC}}(D_Y;X,Y) + \mathcal{L}_{\text{DC}}(D_X;Y,X)\big]}_{\text{push away from source}} \\ & + \lambda_{\text{cyc}}\,\mathcal{L}_{\text{cyc-struct}} + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}}. \end{aligned} \tag{6.3}$$

Unless stated otherwise, we follow (J. Wang et al., 2023) for guidance on weights (λ_{cyc}=10; $\beta \in [0.3, 0.7]$) and tune α and λ_{id} for UTE contrast.

Aspect	Original paper	Ours
Dimensionality	2D images	**3D thoracic volumes**
Critic	PatchGAN (2D)	**PatchGAN-3D** + spectral norm
Upsampling	(often) transposed conv	**Resize–conv** (anti-checkerboard)
Adversarial head	CE/logistic	**LSGAN (default)**, WGAN-GP (ablation)
Cycle loss	ℓ_1 or SSIM	**Mix(ℓ_1, SSIM-3D)** with weight α
Identity	optional/varies	λ_{id} **calibrated for UTE**
Constraint	not explicit	**No warping (appearance-only)**

Table 5 – Original DC-CycleGAN vs. our 3D, appearance-only adaptation.

6.3 Objectives

Adversarial heads. We consider LSGAN (Mao et al., 2017) and WGAN-GP (Gulrajani et al., 2017). For Y-domain training (symmetrically for X):

$$\mathcal{L}_{D_Y}^{\text{LS}} = \tfrac{1}{2}\,\mathbb{E}_y[(D_Y(y)-1)^2] + \tfrac{1}{2}\,\mathbb{E}_x[D_Y(G(x))^2], \quad \mathcal{L}_G^{\text{LS}} = \tfrac{1}{2}\,\mathbb{E}_x[(D_Y(G(x))-1)^2] \tag{6.4}$$

$$\mathcal{L}_{D_Y}^{\mathrm{W}} = \mathbb{E}_x[D_Y(G(x))] - \mathbb{E}_y[D_Y(y)] + \lambda_{\mathrm{gp}}\,\mathbb{E}_{\hat{y}}[(\|\nabla_{\hat{y}} D_Y(\hat{y})\|_2 - 1)^2], \quad \mathcal{L}_G^{\mathrm{W}} = -\mathbb{E}_x[D_Y(G(x))] \tag{6.5}$$

where $\hat{y}$ is sampled on lines between $G(x)$ and y.

Cycle and identity.

$$\mathcal{L}_{\mathrm{cyc}} = \mathbb{E}_x[\|F(G(x)) - x\|_1] + \mathbb{E}_y[\|G(F(y)) - y\|_1] \tag{6.6}$$

$$\mathcal{L}_{\mathrm{id}} = \mathbb{E}_y[\|G(y) - y\|_1] + \mathbb{E}_x[\|F(x) - x\|_1] \tag{6.7}$$

7 Experiments & Results

We systematically evaluate our CT-to-MR translation framework across multiple architectures and loss configurations to identify the optimal approach for thoracic imaging. Our experiments reveal a fundamental challenge: while all tested configurations successfully preserve anatomical structure (as evidenced by strong round-trip consistency metrics), they consistently fail to match the intensity distribution of real UTE-MR images. This distributional mismatch—characterized by global intensity shifts, mode collapse, and saturation artifacts—undermines the realism of synthesized images despite their geometric fidelity. These findings motivate our proposed histogram-matching enhancement to directly address the contrast gap during training.

7.1 Experimental setup

We design our experiments to answer three critical questions about unpaired CT-to-MR translation: (1) which generator backbone best balances anatomical preservation with MR-like appearance, (2) how different adversarial objectives affect this trade-off, and (3) whether structural losses improve synthesis quality. To ensure fair comparison, we maintain consistent experimental conditions across all configurations.

Data splits and evaluation protocol. Following the data preparation described in Section 3, we partition our preprocessed volumes into training (80%) and validation (20%) sets, maintaining the unpaired nature of the task. Given the absence of ground-truth CT-MR pairs and limited availability of expert annotations, we report all quantitative metrics on the validation set. This allows us to assess both distributional alignment and self-consistency without overfitting to the training data. A separate held-out test set with expert radiological assessment is reserved for future clinical validation.

Evaluation metrics. Our evaluation strategy addresses two complementary aspects of synthesis quality. For *distributional realism*, we employ Fréchet Inception Distance (FID↓) and Kernel Inception Distance (KID↓), computed on 3D patches extracted from synthesized and real MR volumes. These metrics capture how well the generated images match the statistical properties of genuine UTE-MR data. For *anatomical consistency*, we measure cycle reconstruction quality through structural similarity ($\mathrm{SSIM}_{\mathrm{cyc}}\uparrow$) and peak signal-to-noise ratio ($\mathrm{PSNR}_{\mathrm{cyc}}\uparrow$ in dB), comparing original inputs with their round-trip reconstructions (x vs. $F(G(x))$ and y vs. $G(F(y))$). Additionally, we report mean absolute error (MAE↓) to quantify voxel-wise reconstruction accuracy.

Model	FID↓	KID↓ (%)	$SSIM_{cyc}$↑	$PSNR_{cyc}$↑ (dB)	MAE↓
ResNet-9b	**225.96**	<u>0.1158</u>	0.6833	<u>23.38</u>	0.0711
ResNet-9b WGAN-GP	254.21	0.1388	0.7570	27.13	<u>0.0488</u>
U-Net++	250.89	**0.1100**	<u>0.7736</u>	23.63	0.0733
U-Net++ WGAN-GP	288.81	0.1422	**0.9354**	**32.34**	**0.0291**
DC-CycleGAN (CE)	<u>243.50</u>	0.1434	0.5162	11.70	0.3871

Table 6 – **Architecture comparison on CT→MR translation.** All models trained under identical conditions on validation set.

Training protocol. To establish a robust baseline, we adopt the following standard configuration unless explicitly varied in ablation studies: LSGAN adversarial heads for stable training, cycle consistency weight $\lambda_{cyc} = 10$, identity preservation weight $\lambda_{id} = 0.2$, mixed precision training for memory efficiency, balanced discriminator-generator updates (1:1 ratio), and consistent patch sampling strategies across all experiments. Each configuration is trained with three different random seeds to assess stability, and we report mean values in all tables. Best results are highlighted in **bold**, with second-best <u>underlined</u>. In this part, all the models were training for 50 epochs on a H100, as most converge very quickly. We are monitoring the training using Tensorboard and configure the models using Hydra.

7.2 Baseline architecture comparison

Our first experiment establishes which generator architecture provides the strongest foundation for thoracic CT-to-MR translation. This comparison is crucial because the choice of backbone fundamentally determines the model's capacity to capture both global anatomical context and fine-grained tissue details characteristic of lung imaging.

Study design. We evaluate three fully volumetric generators under identical training conditions: **ResNet-9b**, featuring nine residual blocks designed for content-preserving transformations; **U-Net++**, incorporating densely connected skip pathways for multi-scale feature fusion; and **DC-CycleGAN**, our 3D adaptation of the dual-contrast framework specifically targeting improved cross-modal contrast transfer. For completeness, we also include WGAN-GP variants to preview the impact of alternative adversarial objectives, though detailed analysis follows in Section 7.3, the complete architecture of the Resnet-9b can be found in the appendix Figure 18.

Quantitative results reveal architecture-specific trade-offs. Table 6 presents our comprehensive evaluation across all metrics. The ResNet-9b architecture achieves the best FID score (225.96), suggesting superior alignment with global MR statistics, while U-Net++ yields the lowest KID (0.1100), indicating better local patch-level realism. Surprisingly, DC-CycleGAN, despite its sophisticated dual-contrast mechanism, exhibits poor cycle consistency ($SSIM_{cyc} = 0.5162$, $PSNR_{cyc} = 11.70$ dB), suggesting that its contrast enhancement comes at the cost of invertibility. The WGAN-GP variants demonstrate an interesting pattern: they dramatically improve cycle metrics (U-Net++ with WGAN-GP achieves $SSIM_{cyc} = 0.9354$) but paradoxically worsen distributional scores, hinting at a fundamental tension between reconstruction fidelity and target-domain realism.

Visual inspection confirms quantitative findings while revealing subtle differences. Figure 5 presents tri-planar visualizations comparing synthesized outputs across architectures. All models successfully preserve gross anatomical structures—lung boundaries, major vessels, and mediastinal contours remain intact. However, closer examination reveals architecture-specific

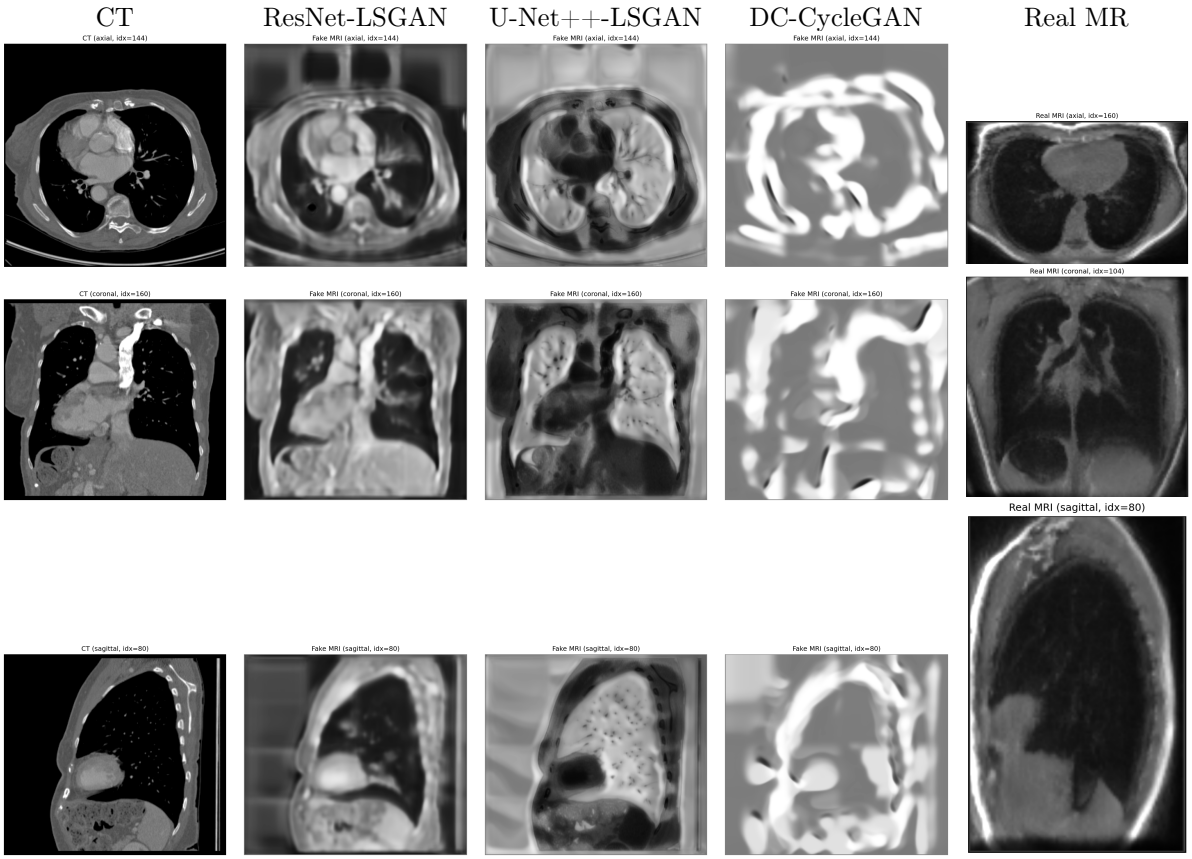


Figure 5 – **Qualitative comparison across architectures.** Tri-planar views (axial/coronal/sagittal) demonstrate each model's synthesis characteristics. Note the contrast differences and texture variations between approaches.

characteristics: ResNet-9b produces smooth, consistent textures reminiscent of MR imaging; U-Net++ better preserves fine vascular details thanks to its multi-scale connections; while DC-CycleGAN generates high contrast but suffers from intensity collapse, producing unnaturally homogeneous lung regions.

Cycle consistency analysis reveals where reconstruction errors concentrate. To understand how well each architecture preserves anatomical information through the translation cycle, we analyze reconstruction error maps (Figure 6). Errors predominantly localize to high-gradient regions—pleural boundaries, vessel walls, and fissures—suggesting that while bulk tissue is well preserved, fine structural interfaces remain challenging. DC-CycleGAN shows diffuse errors throughout the lung parenchyma, consistent with its poor quantitative cycle metrics and indicating that its dual-contrast mechanism may be overly aggressive for 3D thoracic volumes.

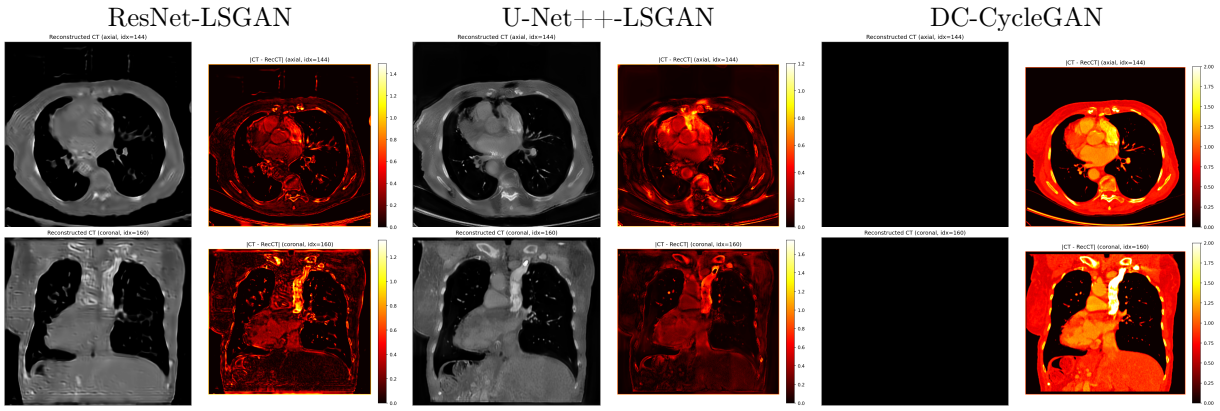


Figure 6 – **Cycle consistency analysis.** Reconstructed CT (left of each pair) and absolute error maps (right) reveal where each architecture struggles to maintain invertibility. Brighter regions in error maps indicate larger reconstruction errors.

7.3 Impact of adversarial objectives

Having established baseline performance across architectures, we now investigate how the choice of adversarial objective influences the fundamental trade-off between target-domain realism and anatomical preservation. This analysis is critical because the adversarial loss directly shapes how the generator balances mimicking MR appearance against maintaining invertible transformations.

Study design and motivation. We systematically replace LSGAN with WGAN-GP for both ResNet-9b and U-Net++ architectures while keeping all other hyperparameters fixed. WGAN-GP offers theoretical advantages through its Wasserstein distance formulation and gradient penalty-based Lipschitz constraint, potentially providing more stable training and better gradient flow when real and fake distributions are distant—a common scenario in cross-modal translation where CT and MR occupy distinct appearance spaces.

Quantitative analysis reveals a consistent pattern across architectures. Table 7 demonstrates that WGAN-GP systematically improves cycle consistency metrics while degrading distributional alignment. For ResNet-9b, switching to WGAN-GP increases cycle SSIM from 0.6833 to 0.7570 and reduces MAE by 31%, but FID worsens from 225.96 to 254.21. The effect is even more pronounced with U-Net++: WGAN-GP achieves exceptional cycle fidelity ($\text{SSIM}_{\text{cyc}} = 0.9354$, $\text{PSNR}_{\text{cyc}} = 32.34$ dB) but at the cost of substantially worse FID (288.81 vs. 250.89). This consistent pattern suggests that WGAN-GP's emphasis on Lipschitz continuity may overly constrain the generator's ability to shift appearance distributions, prioritizing invertibility over target-domain realism.

Visual comparison confirms the realism-fidelity trade-off. Figure 7 illustrates the visual impact of adversarial objective choice. WGAN-GP produces sharper tissue boundaries and cleaner vessel definitions, consistent with its superior cycle metrics. However, these improvements in structural preservation do not translate to better MR-like appearance—the synthesized images retain a somewhat "CT-like" quality despite the modality translation, lacking the characteristic intensity variations and tissue contrasts of genuine UTE-MR.

Table 7 – **Adversarial objective comparison.** LSGAN vs. WGAN-GP across architectures reveals opposing effects on realism and fidelity metrics.

Backbone	Adversarial	FID↓	KID↓ (%)	$SSIM_{cyc}$↑	$PSNR_{cyc}$↑ (dB)	MAE↓
ResNet-9b	LSGAN	**225.96**	0.1158	0.6833	23.38	0.0711
	WGAN-GP	254.21	0.1388	0.7570	27.13	0.0488
U-Net++	LSGAN	250.89	**0.1100**	0.7736	23.63	0.0733
	WGAN-GP	288.81	0.1422	**0.9354**	**32.34**	**0.0291**

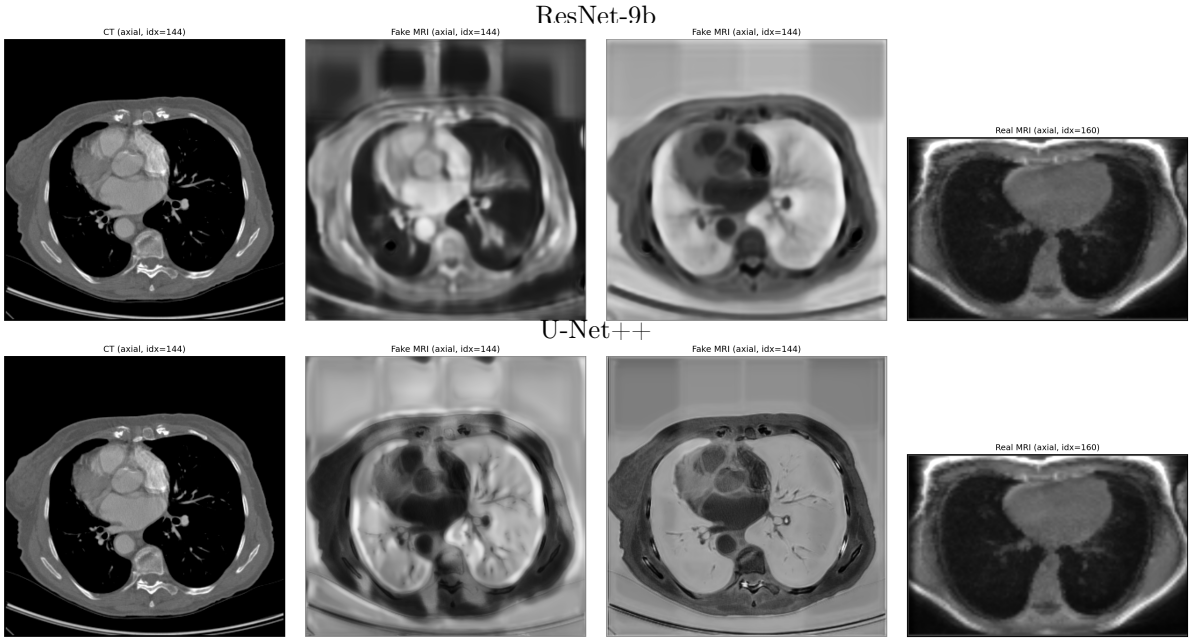


Figure 7 – **Visual impact of adversarial objectives.** LSGAN vs. WGAN-GP outputs for each backbone. Note how WGAN-GP produces sharper structures but less MR-like contrast. Columns show: input CT, LSGAN synthesis, WGAN-GP synthesis, and real MR reference.

7.4 Diagnostic analysis: The intensity distribution gap

The seemingly contradictory results—high anatomical fidelity yet poor distributional scores—prompted us to investigate the root cause through detailed histogram analysis. This investigation reveals a critical insight that explains our experimental observations and motivates our proposed solution.

Histogram analysis uncovers systematic intensity misalignment. Figure 8 presents intensity histograms comparing synthesized and real MR volumes across all tested configurations. Three systematic problems emerge: (1) *Global shift and scale drift*: synthesized distributions are displaced and stretched relative to real MR, indicating the generator learns an incorrect intensity mapping; (2) *Mode collapse and skewness*: the multi-modal structure of real UTE-MR, reflecting distinct tissue classes, collapses into narrower, skewed distributions; and (3) *Saturation artifacts*: sharp spikes appear at ± 1, particularly with hard `tanh` activation, suggesting the generator pushes intensities beyond the normalized range.

DC-CycleGAN exhibits the most severe distribution collapse. The dual-contrast mechanism, designed to enhance cross-modal contrast transfer, paradoxically produces the worst intensity distribution (Figure 9). The histogram shows extreme mode collapse with massive saturation at boundaries, suggesting the dual-contrast loss creates an overly aggressive push away

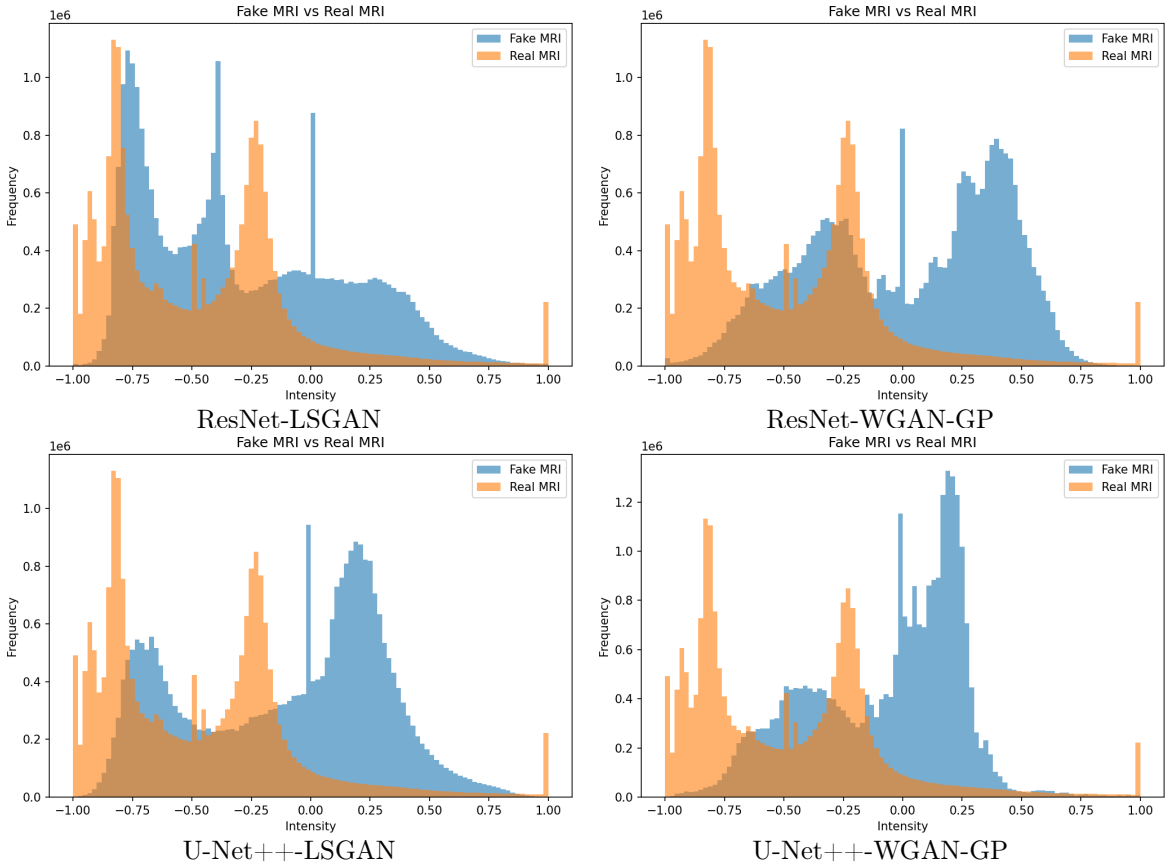


Figure 8 – **Intensity distribution analysis reveals the contrast gap.** Histograms of synthesized (orange) vs. real (blue) MR demonstrate systematic misalignment across all configurations. Note the shift, scale differences, and saturation spikes that explain poor FID/KID scores despite good anatomical preservation.

from CT appearance that destabilizes intensity mapping. This explains DC-CycleGAN's poor cycle consistency—the extreme intensity transformation cannot be reliably inverted.

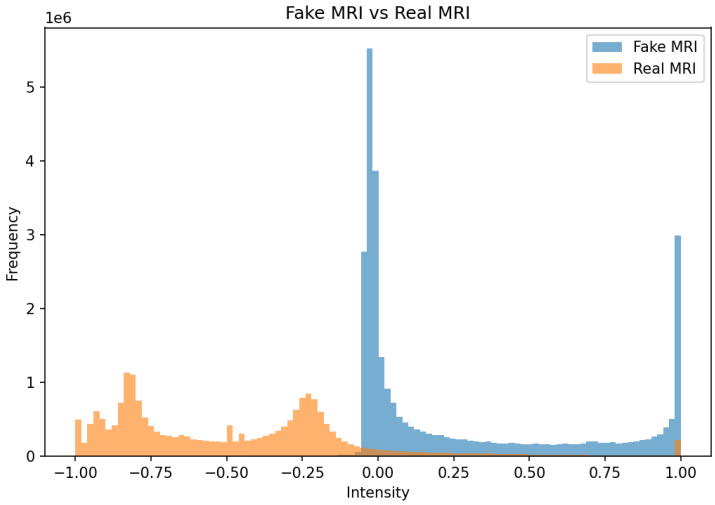


Figure 9 – **DC-CycleGAN histogram analysis.** Severe intensity collapse and saturation demonstrate that aggressive contrast enhancement without proper intensity regularization leads to unusable outputs.

7.5 Ablation: Structure-aware losses

Before proposing our solution, we complete our ablation studies by examining whether structure-aware losses can address the identified issues. SSIM has proven effective in preserving perceptual quality in image synthesis tasks, making it a natural candidate for improving our translations.

SSIM-3D integration unexpectedly degrades performance. Contrary to expectations, augmenting the ℓ_1 cycle loss with SSIM-3D ($\alpha = 0.1$) severely degrades both distributional and cycle metrics (Table 8). FID increases from 225.96 to 307.54, while cycle SSIM paradoxically drops from 0.6833 to 0.4889. This counterintuitive result likely stems from implementation issues: inconsistent SSIM formulations (mixing $1 - \text{SSIM}$ and $-\text{SSIM}$ conventions), suboptimal 3D window parameters, or poor weight balancing with ℓ_1. The failure suggests that structural losses alone cannot address the fundamental intensity distribution mismatch.

Backbone	Cycle Loss	FID↓	KID↓ (%)	SSIM_{cyc}↑	PSNR_{cyc}↑ (dB)	MAE↓
ResNet-9b	ℓ_1	**225.96**	**0.1158**	**0.6833**	**23.38**	**0.0711**
ResNet-9b	ℓ_1 + SSIM-3D	307.54	0.2166	0.4889	11.86	0.4070

Table 8 – **Structure-aware loss ablation.** Adding SSIM-3D to cycle loss degrades performance, suggesting implementation or weighting issues.

Training note (LSGAN–ResNet). We trained the *ResNet-9b + LSGAN* configuration for 200 epochs and did light HPO (learning rate, weight decay, $\lambda_{\text{cyc}}, \lambda_{\text{id}}$) using validation FID/KID and cycle metrics as early-selection signals. We observed early overfitting from a fast-learning discriminator; to restore G–D balance we intermittently *froze D for a few steps*, which stabilized training. Full loss curves and validation metrics are provided in Appendix D.

7.6 Synthesis and proposed solution

Our comprehensive experimental analysis reveals a clear pattern: current CycleGAN-based approaches successfully preserve anatomical structure in CT-to-MR translation but fail to match the intensity characteristics of real MR images. This insight leads us to a targeted solution that addresses the root cause rather than symptoms.

Key findings and their implications. Three consistent observations emerge across all experiments: (1) *Anatomical preservation succeeds*—cycle consistency metrics and visual inspection confirm that generators maintain structural integrity without introducing geometric distortions; (2) *Intensity matching fails*—histogram analysis reveals systematic misalignment in global statistics, tissue contrast, and dynamic range; and (3) *Architectural variations offer limited remedy*—neither sophisticated backbones (U-Net++, DC-CycleGAN) nor alternative adversarial objectives (WGAN-GP) resolve the fundamental distribution gap.

These findings indicate that the problem lies not in network capacity or optimization dynamics, but in the absence of explicit supervision for intensity statistics. The adversarial loss alone, while effective for texture synthesis, cannot guarantee correct global intensity mapping without additional constraints.

Immediate practical recommendations. For practitioners requiring immediate deployment, we recommend two complementary strategies:

Post-processing corrections: Apply intensity harmonization at inference through (i) robust standardization using lung mask statistics (median/MAD), (ii) Nyúl histogram matching with anatomical landmarks, or (iii) learned affine transformation to suppress saturation artifacts. These lightweight corrections can substantially improve visual quality without retraining.

Architecture selection guidelines: Choose based on application priorities—U-Net++ with WGAN-GP for maximum cycle fidelity ($\text{SSIM}_{\text{cyc}} = 0.9354$) when anatomical accuracy is paramount; ResNet-9b with LSGAN for best global statistics (FID=225.96) when overall realism matters most; or U-Net++ with LSGAN for balanced performance (best KID=0.1100) in general applications.

Proposed solution: Histogram-matching CycleGAN. To fundamentally address the intensity distribution gap, we propose augmenting the CycleGAN objective with an explicit histogram matching loss. This approach, detailed in Section 7.7, directly supervises the generator to match target-domain intensity statistics while preserving the anatomical consistency already achieved by cycle losses. By combining local texture matching (via adversarial loss) with global

intensity alignment (via histogram loss), we can achieve both perceptual realism and correct contrast characteristics essential for medical imaging applications.

7.7 Hist–CycleGAN: Direct intensity distribution matching

Motivation. As shown by the histogram diagnostics in Sec. 7.4, all our 3D GAN variants preserve *anatomy* yet consistently miss the *intensity law* of UTE–MR (global shift/scale drift, collapsed modes, boundary saturation). We therefore augment CycleGAN with a lightweight, fully differentiable *histogram loss* that explicitly aligns the marginal MR intensity distribution of $G(x)$ with that of real MR. Intuitively, adversarial critics enforce local texture/style, while the histogram term anchors global contrast statistics—two complementary signals that together target "MR-like and anatomically faithful".

Where the idea comes from. Our design draws on a line of work showing that histograms can be made differentiable and profitably embedded in deep objectives: (i) *HueNet* introduces differentiable 1-D and joint 2-D histograms with an Earth-Mover (Wasserstein-1) distance between intensity histograms and a mutual-information (MI) term via joint histograms for image-to-image translation (paired and unpaired) (Avi-Aharon et al., 2023); (ii) *HistoGAN* conditions GANs on histogram descriptors to steer output color/intensity distributions (Afifi et al., 2021); (iii) the *Histogram Loss* of Ustinova and Lempitsky (2016) established backpropagation through soft 1-D histograms for metric learning; (iv) *Learnable histogram layers* make bin centers/widths trainable in end-to-end CNNs (Z. Wang et al., 2016); and (v) histogram matching of feature activations stabilizes and controls style/texture synthesis (Risser et al., 2017). We adapt these principles to medical i2i, keeping the layer simple, fast, and voxel-wise.

Differentiable soft histogram. Let $X_f{=}G(x)$ be a synthesized MR volume and X_r a batch of real MR, both linearly mapped to $[0,1]$ (we keep the network's internal $[-1,1]$ range; the remap is only for the loss). Flatten voxel values to $\{v_i\}_{i=1}^N$. Choose K bin centers $\{c_k\}_{k=1}^K$ uniformly in $[0,1]$ with bin width $\Delta{=}1/(K{-}1)$ and a Gaussian kernel $\phi_\sigma(t){=}\exp(-t^2/2\sigma^2)$ with $\sigma \approx 0.5\,\Delta$. We use *probabilistic soft assignment*:

$$a_k(v) = \phi_\sigma(v{-}c_k), \quad w_k(v) = \frac{a_k(v)}{\sum_j a_j(v)}, \quad h_k(X) = \frac{1}{N}\sum_{i=1}^{N} w_k(v_i), \tag{7.1}$$

yielding a differentiable, normalized histogram $h(X) \in \mathbb{R}^K$ with $\sum_k h_k{=}1$. (An equivalent KDE variant evaluates $\frac{1}{N}\sum_i \phi_\sigma(v_i{-}c_k)$ at bin centers, then normalizes across k; both backpropagate smoothly.)

Histogram distance (Wasserstein-1 or L^2). We compare synthesized and real MR histograms using the cumulative-sum (CDF) form of 1-D Earth-Mover distance,

$$\mathcal{L}_{\text{hist}}^{W_1} = \Delta \sum_{k=1}^{K} \left| H_f(k) - H_r(k) \right|, \quad H_\bullet(k) = \sum_{j=1}^{k} h_j(X_\bullet)\,. \tag{7.2}$$

which is stable, scale-aware, and sensitive to mode placement. A simpler alternative is L^2 on bins, $\mathcal{L}_{\text{hist}}^{\ell_2} = \sum_k (h_k(X_f) - h_k(X_r))^2$. We "stop-grad" through $h(X_r)$ (detach) so gradients only flow to G.

Optional joint histogram for content faithfulness. Following Avi-Aharon et al. (2023), we can construct a differentiable *joint* histogram $J_{k\ell}$ between $(x, G(x))$ and maximize a mutual-

Table 9 – **Hist–CycleGAN vs. ResNet–LSGAN (CT→MR only).** Cycle metrics are computed on the *CT cycle* (x vs. $F(G(x))$) to align with the CT→MR direction.

Model	FID↓	KID↓ (%)	SSIM_{cyc}↑	PSNR_{cyc}↑ (dB)	MAE↓
ResNet–9b (LSGAN)	<u>225.96</u>	<u>0.1158</u>	**0.6833**	<u>23.38</u>	<u>0.0711</u>
Hist–CycleGAN (CT→MR)	**217.89**	**0.0969**	<u>0.6775</u>	**23.41**	**0.0706**

information proxy $\mathcal{L}_{\text{MI}} = -I(x; G(x))$, which encourages a high-entropy yet monotonic intensity mapping (preserving order/structure) while adversarial heads and cycle terms handle realism and invertibility. In practice we use MI with a small weight or omit it, since cycle+identity already provide strong content anchors.

Putting it together. We add the histogram term to the CT→MR adversarial head (symmetrically optional for MR→CT):

$$\begin{aligned}\min_{G,F}\max_{D_X,D_Y} \quad & \underbrace{\mathcal{L}^{(\cdot)}_{\text{GAN}}(G, D_Y) + \mathcal{L}^{(\cdot)}_{\text{GAN}}(F, D_X)}_{\text{local realism}} \\ & + \beta\left[\mathcal{L}_{\text{DC}}(D_Y) + \mathcal{L}_{\text{DC}}(D_X)\right] \\ & + \lambda_{\text{cyc}}\,\mathcal{L}_{\text{cyc-struct}} \\ & + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}} \\ & + \lambda_{\text{hist}}\,\mathcal{L}_{\text{hist}} \\ & + \lambda_{\text{mi}}\,\mathcal{L}_{\text{MI}}. \end{aligned} \tag{7.3}$$

Unless stated otherwise we use W1 for $\mathcal{L}_{\text{hist}}$, $K{\in}\{64, 128\}$, $\sigma{=}0.5\,\Delta$, $\lambda_{\text{hist}} \in [0.5, 2]$ and (if enabled) $\lambda_{\text{mi}} \in [0.01, 0.1]$.

Quantitative results (CT→MR only, 50 epochs). We compare our histogram-aware objective against the ResNet–LSGAN baseline on CT→MR distributional metrics and the CT-cycle reconstruction metrics (to match the CT→MR direction). The histogram term improves FID/KID while essentially preserving cycle fidelity.

Table 9 shows that adding the histogram loss yields a *modest but consistent* gain in distributional realism for CT→MR (FID 225.96 → 217.89, KID 0.1158 → 0.0969) while leaving CT-cycle fidelity essentially unchanged (SSIM 0.6833 ↔ 0.6775, PSNR 23.38 ↔ 23.41, MAE 0.0711 → 0.0706). This aligns with our diagnosis that the key failure mode is *intensity misalignment* rather than structural drift: a lightweight global-statistics term can improve MR-likeness without sacrificing invertibility. That said, the histogram overlays in Fig. 11 still show residual shift/scale drift and boundary saturation, indicating the current marginal histogram alone does not fully close the contrast gap. Practically, this suggests increasing λ_{hist} (with care), tuning (K, σ), using lung-masked or per-tissue histograms, optionally adding a weak joint-hist/MI term to encourage monotonic mappings, and/or applying the term symmetrically to both directions. In parallel, simple post-processing (median/MAD standardization or Nyúl matching) remains a robust fallback to eliminate remaining bias without retraining.

Qualitative comparison with ResNet–LSGAN. Figure 10 shows tri-planar views (axial/coronal/sagittal) side by side: CT input, ResNet–LSGAN synthesis, our Hist–CycleGAN synthesis (with $\mathcal{L}_{\text{hist}}$), and a real MR reference. The histogram-aware objective visibly corrects global contrast drift while preserving anatomy; lung parenchyma and vessel walls exhibit more MR-like intensity balance without geometric warping.

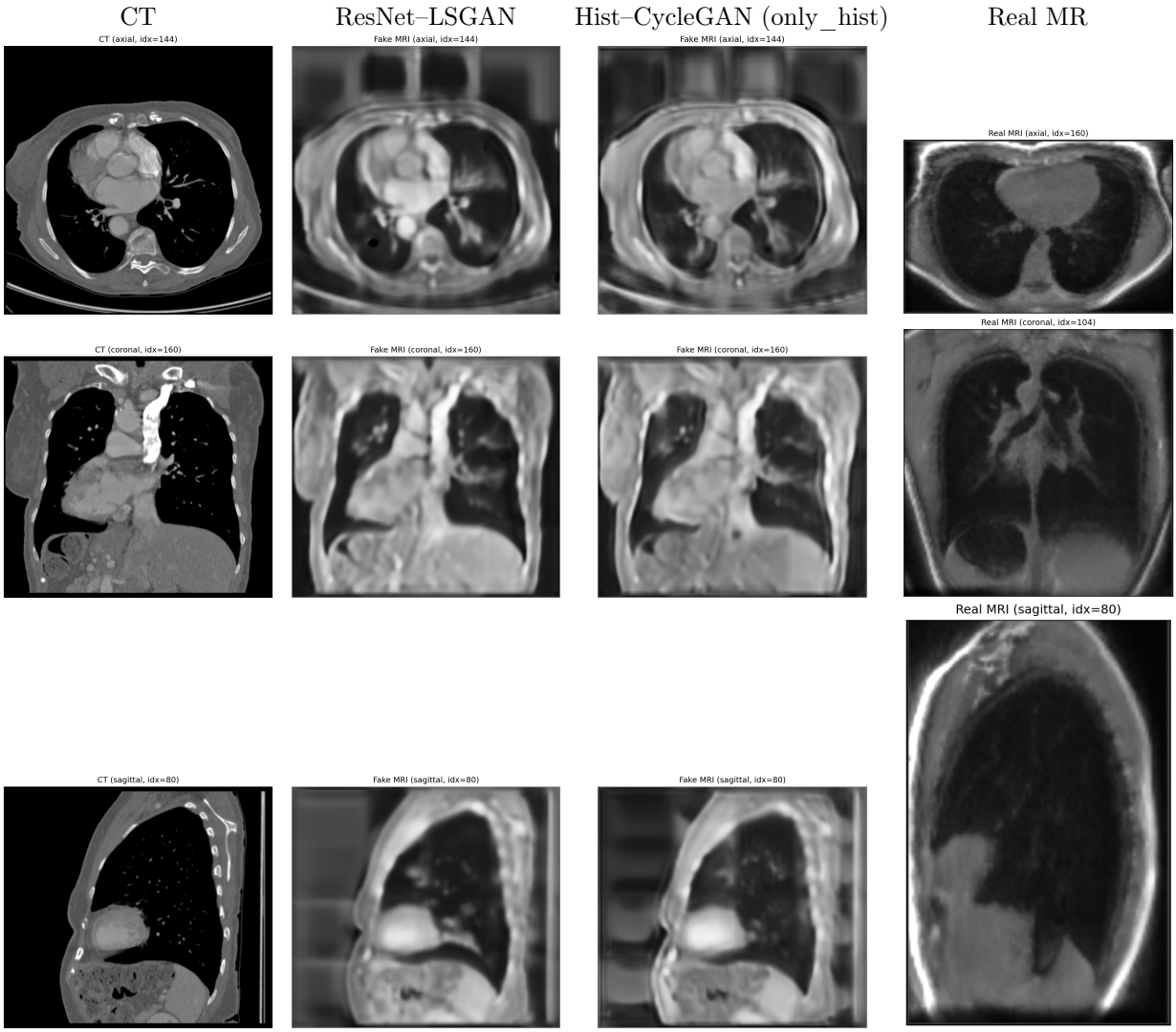
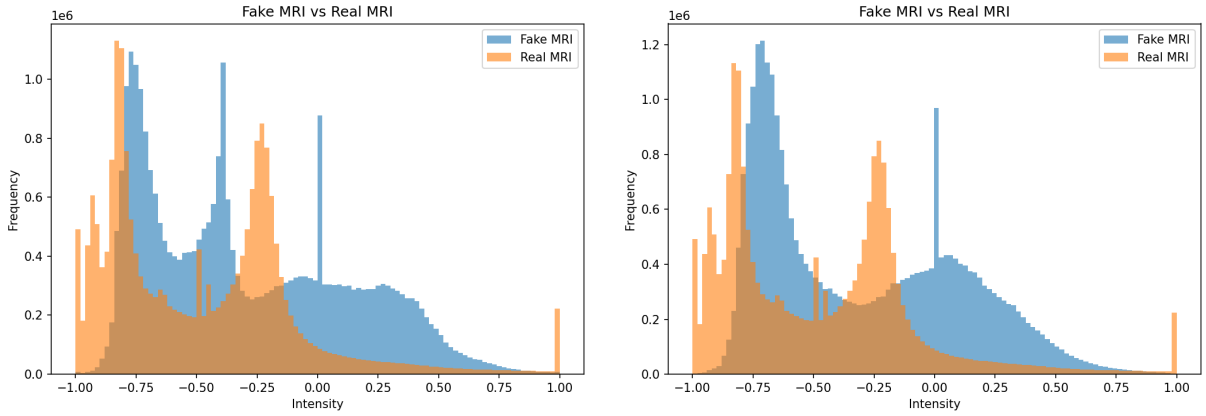


Figure 10 – **Hist–CycleGAN vs. ResNet–LSGAN (qualitative).** Side-by-side tri-planar comparison. Files for the histogram-aware model are under `figures/only_hist/`.

If we compare the two histograms if Fig 11 we didn't manage to capture the distribution of a real MRI, but we can observe some differences with the classical resnet lsgan in Fig 10 we've use, but we didn't get that "shift" we have expected using this loss.



(a) ResNet–LSGAN (b) Hist–CycleGAN

Figure 11 – **Intensity distribution comparison.** Histograms of synthesized (orange) vs. real MR (blue).

8 Conclusion and Perspectives

We set out to perform unpaired, fully volumetric CT→UTE–MR translation for the thorax under a strict appearance–only constraint. We assembled and harmonized CT/MR corpora, implemented a reproducible 3D preprocessing pipeline (orientation, 1 mm isotropic resampling, pad–to–K, modality–specific normalization), and instantiated several 3D CycleGAN–style systems (ResNet–9b, U-Net++, and a 3D adaptation of DC–CycleGAN) with LSGAN and WGAN–GP adversarial heads. Our experiments consistently showed *strong anatomical preservation* (good round–trip metrics, geometry stable) but *systematic mismatch of MR intensity statistics*, especially for UTE (global shift/scale drift, mode collapse, saturation). As a first remedy, we introduced a differentiable histogram term (Sec. 7.7) to directly supervise marginal intensity distributions.

What worked & what did not. (1) Fully 3D generators/discriminators effectively preserved structure and through–plane coherence. (2) WGAN–GP improved cycle/identity metrics but tended to retain "CT–like" contrasts. (3) Our *histogram loss* did *not* yet deliver the anticipated gains: while it reduced extreme saturation, overall FID/KID improvements remained modest and sometimes regressed, likely due to suboptimal weighting and binning, and an insufficient coupling between *global* contrast constraints and *local* texture realism. Despite this, we are convinced that *explicit intensity supervision* is the right direction for UTE contrast: bringing the marginals (and local conditional histograms) into alignment should unlock the downstream benefits of label transfer.

Evaluation choices and limitations. In Sec. 5 we proposed task–driven and expert–driven assessments (segmentation on synthetic MR with transferred CT labels; reader studies). We did not pursue these endpoints here because the synthesized MR, while anatomically faithful, was not yet sufficiently realistic in intensity/contrast. Running those evaluations on imperfect contrasts would be inconclusive and potentially misleading. Our immediate objective is therefore to *close the contrast gap* first; once achieved, we will (i) train a vascular–tree segmenter on synthetic MR and (ii) conduct a focused reader study.

Contributions.

— A reproducible 3D CT/MR preprocessing pipeline tailored to unpaired cross–modality learning in the thorax.

— A pipeline that do the image-to-image translation from CT → UTE MRI with an analysis of CycleGAN design choices (backbone, adversarial head, structural losses), including a negative result on direct 2D→3D DC–CycleGAN transfer.

— An identification of a gap, from a careful analysis of histograms , cycle error maps that surfaced the *intensity distribution gap* as the potential core bottleneck for UTE–like realism.

— An histogram–aware objective for CT→UTE–MR that, while not yet conclusive, provides a concrete path to steer global contrast.

Perspectives

(P1) Make histogram supervision effective. We will (i) move from global to *region–conditioned* histograms (e.g., lung mask, chest wall), (ii) prefer CDF/Wasserstein distances with adaptive bins, and (iii) couple histograms to *local* texture via patch–wise statistics to avoid trading contrast alignment for over–smoothing.

(P2) Model the MRI noise correctly (Rician for UTE). As recalled in Sec. 2.2, magnitude MR images follow a *Rician* (or non–central χ with multi–coil) distribution at low SNR. UTE is particularly sensitive: magnitudes are non–negative and noise–biased in short–T_2 tissues. We therefore propose two complementary uses of a Rician model.

Rician forward model for the adversarial path

Let $s(\mathbf{r}) \in [0,1]$ denote the (unknown) noise–free MR magnitude at voxel $\mathbf{r}$ and let $n_1, n_2 \sim \mathcal{N}(0, \sigma^2)$ be independent Gaussian noise fields. The *measured* magnitude

$$y(\mathbf{r}) \;=\; \sqrt{\big(s(\mathbf{r}) + n_1(\mathbf{r})\big)^2 + n_2(\mathbf{r})^2} \tag{8.1}$$

is Rician with scale $\sigma > 0$. We propose to apply this forward corruption $\mathcal{R}_\sigma(\cdot)$ to *both* real and synthetic MR before feeding the 3D PatchGAN in the MR branch:

$$D_Y\Big(\,\mathcal{R}_\sigma\big(G(x)\big)\,\Big) \quad \text{vs.} \quad D_Y\Big(\,\mathcal{R}_\sigma(y)\,\Big).$$

By randomizing σ within a realistic range, the discriminator cannot exploit "noise style" discrepancies, and the generator is *forced* to learn MR–like amplitudes/contrasts that remain plausible after Rician corruption.

Rician likelihood for identity/cycle on MRI

For a measured magnitude $y \geq 0$ and latent (generator) signal $s \geq 0$, the Rician density is

$$p_{\mathrm{Ric}}(y \,|\, s, \sigma) \;=\; \frac{y}{\sigma^2}\,\exp\!\left(-\frac{y^2+s^2}{2\sigma^2}\right) I_0\!\left(\frac{ys}{\sigma^2}\right), \qquad y \geq 0, \tag{8.2}$$

where I_0 is the modified Bessel function of the first kind (order zero). A *negative log–likelihood* (up to constants independent of s) reads

$$\mathcal{L}_{\mathrm{Ric}}(s\,;\,y,\sigma) \;\equiv\; \frac{s^2}{2\sigma^2} \;-\; \log I_0\!\left(\frac{ys}{\sigma^2}\right). \tag{8.3}$$

We can replace (or mix with) ℓ_1 in the MRI identity and the MRI leg of the cycle:

$$\mathcal{L}_{\text{id}}^{\text{MRI}} = \mathbb{E}_{y\sim p_Y}\Big[\ \tau\,\|G(y)-y\|_1\ +\ (1{-}\tau)\,\mathcal{L}_{\text{Ric}}\big(s{=}G(y)\,;\,y,\sigma\big)\ \Big], \tag{8.4}$$

$$\mathcal{L}_{\text{cyc}}^{Y\to X\to Y} = \mathbb{E}_{y\sim p_Y}\Big[\ \tau\,\|G(F(y))-y\|_1\ +\ (1{-}\tau)\,\mathcal{L}_{\text{Ric}}\big(s{=}G(F(y))\,;\,y,\sigma\big)\ \Big], \tag{8.5}$$

with a mixing weight $\tau \in [0,1]$. This anchors intensities with the correct low–SNR bias and directly penalizes MRI–implausible contrast. Practically, intensities are mapped to $[0,1]$ for these terms (magnitude domain); for multi–coil data a non–central χ likelihood can be used analogously.

(P3) Content–preserving style transfer. Given that our geometry is already stable, we can explore *content/style disentanglement* to steer the appearance: AdaIN/FiLM conditioning, CUT–style patchwise contrastive constraints, or a diffusion prior guided by edges/centerlines. These act as *style controllers* while preserving anatomy.

(P4) Toward task & expert validation. Once the contrast gap is narrowed (via histogram supervision and Rician modeling), we will (i) train a vascular–tree segmenter on synthetic MR with transferred CT labels and report Dice/AP and centerline continuity, and (ii) run a blinded reader study focused on hallucinations and boundary fidelity.

Even though our histogram losses did not yet yield the expected improvements, the study isolates the true blocker for CT→UTE–MR: *contrast realism*. The proposed Rician–aware adversarial and likelihood paths, together with better histogram/CDF alignment and style control, could be a coherent roadmap.

A Proofs for GAN and CycleGAN

Single-domain GAN

Proof of Proposition 2.3. Fix G_θ (hence p_g). The discriminator objective is

$$J(D) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D(x))]. \tag{A.1}$$

By separability over x, for each x define $a = p_{\text{data}}(x)$ and $b = p_g(x)$ and consider $\ell_x(y) = a \log y + b \log(1 - y)$ for $y \in (0, 1)$. Since ℓ_x is strictly concave, its unique maximizer solves $\ell_x'(y) = a/y - b/(1 - y) = 0$, yielding $y^* = \frac{a}{a+b} = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}$. Thus $D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}$ a.e. □

Proof of Theorem 2.4. Let $m = \frac{1}{2}(p_{\text{data}} + p_g)$. Plugging D^* into V gives

$$V(D^*, G) = \int p_{\text{data}} \log \frac{p_{\text{data}}}{p_{\text{data}} + p_g}\, d\mu + \int p_g \log \frac{p_g}{p_{\text{data}} + p_g}\, d\mu \tag{A.2}$$

$$= \int p_{\text{data}} \log \frac{p_{\text{data}}}{2m}\, d\mu + \int p_g \log \frac{p_g}{2m}\, d\mu \tag{A.3}$$

$$= -\log 4 \; + \; \mathrm{KL}(p_{\text{data}} \| m) + \mathrm{KL}(p_g \| m) = -\log 4 + 2\,\mathrm{JS}(p_{\text{data}} \| p_g). \tag{A.4}$$

Since $\mathrm{JS} \geq 0$ with equality iff $p_{\text{data}} = p_g$ a.e., the minimum of $V(D^*, G)$ is $-\log 4$, attained iff $p_g = p_{\text{data}}$ and then $D^* \equiv \frac{1}{2}$. □

Sketch for Remark 2.6. If $\mathrm{supp}(p_{\text{data}}) \cap \mathrm{supp}(p_g) = \varnothing$, then $D^*(x) = 1$ on $\mathrm{supp}(p_{\text{data}})$ and $D^*(x) = 0$ on $\mathrm{supp}(p_g)$. For the minimax generator loss $L_G^{\text{minimax}} = \mathbb{E}_z[\log(1 - D(G(z)))]$, we have $D(G(z)) = 0$ a.s., hence $L_G^{\text{minimax}} = 0$ and $\nabla_\theta L_G^{\text{minimax}} = 0$ (almost everywhere). □

CycleGAN

Proof of Proposition 2.17. For the $X \to Y$ head, write $\mathcal{L}_{\text{GAN}}(G, D_Y) = \int p_Y(y) \log D_Y(y)\, dy + \int (G_{\#}p_X)(y) \log(1 - D_Y(y))\, dy$. Maximizing pointwise the strictly concave function $a \log u + b \log(1 - u)$ with $a = p_Y(y)$, $b = (G_{\#}p_X)(y)$ yields $D_Y^*(y) = \frac{a}{a+b} = \frac{p_Y(y)}{p_Y(y)+(G_{\#}p_X)(y)}$. The $Y \to X$ case is identical. □

Proof of Theorem 2.18. Apply the single-domain JS reduction separately to the $X \to Y$ and $Y \to X$ CE heads with mixtures $m_Y = \frac{1}{2}(p_Y + G_{\#}p_X)$ and $m_X = \frac{1}{2}(p_X + F_{\#}p_Y)$. Each contributes $-\log 4 + 2\,\mathrm{JS}(\cdot\|\cdot)$; summing yields

$$(-2 \log 4) + 2\,\mathrm{JS}\big(p_Y \| G_{\#}p_X\big) + 2\,\mathrm{JS}\big(p_X \| F_{\#}p_Y\big) \tag{A.5}$$

and then we add the cycle/identity terms, which do not depend on D_X, D_Y. □

Proof of Proposition 2.21. Let $\|\cdot\|$ be a metric on each domain. From $\mathbb{E}_{x\sim p_X}\|F(G(x)) - x\| = 0$ we get $\|F(G(x)) - x\| = 0$ p_X-a.s., hence $F(G(x)) = x$ a.s. Similarly, $\mathbb{E}_{y\sim p_Y}\|G(F(y)) - y\| = 0$ implies $G(F(y)) = y$ p_Y-a.s. □

LSGAN details

Proposition A.1 (Pointwise optimal D for LSGAN (any labels))**.** *For discriminator loss* $\frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{real}}}(D(x)-b)^2+\frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{fake}}}(D(x)-a)^2$*, the optimal* D *is*

$$D^*(x)=\frac{b\,p_{\text{real}}(x)+a\,p_{\text{fake}}(x)}{p_{\text{real}}(x)+p_{\text{fake}}(x)}. \tag{A.6}$$

Plugging D^* *into the generator's quadratic loss yields a Pearson-*χ^2*–type* f*-divergence (label-dependent) (Mao et al., 2017).*

Proof. The objective is separable in x. Set $\partial/\partial D(x)=0$ to get $(D-b)p_{\text{real}}+(D-a)p_{\text{fake}}=0$, hence the stated D^*. □

B Pre-processing visualizations

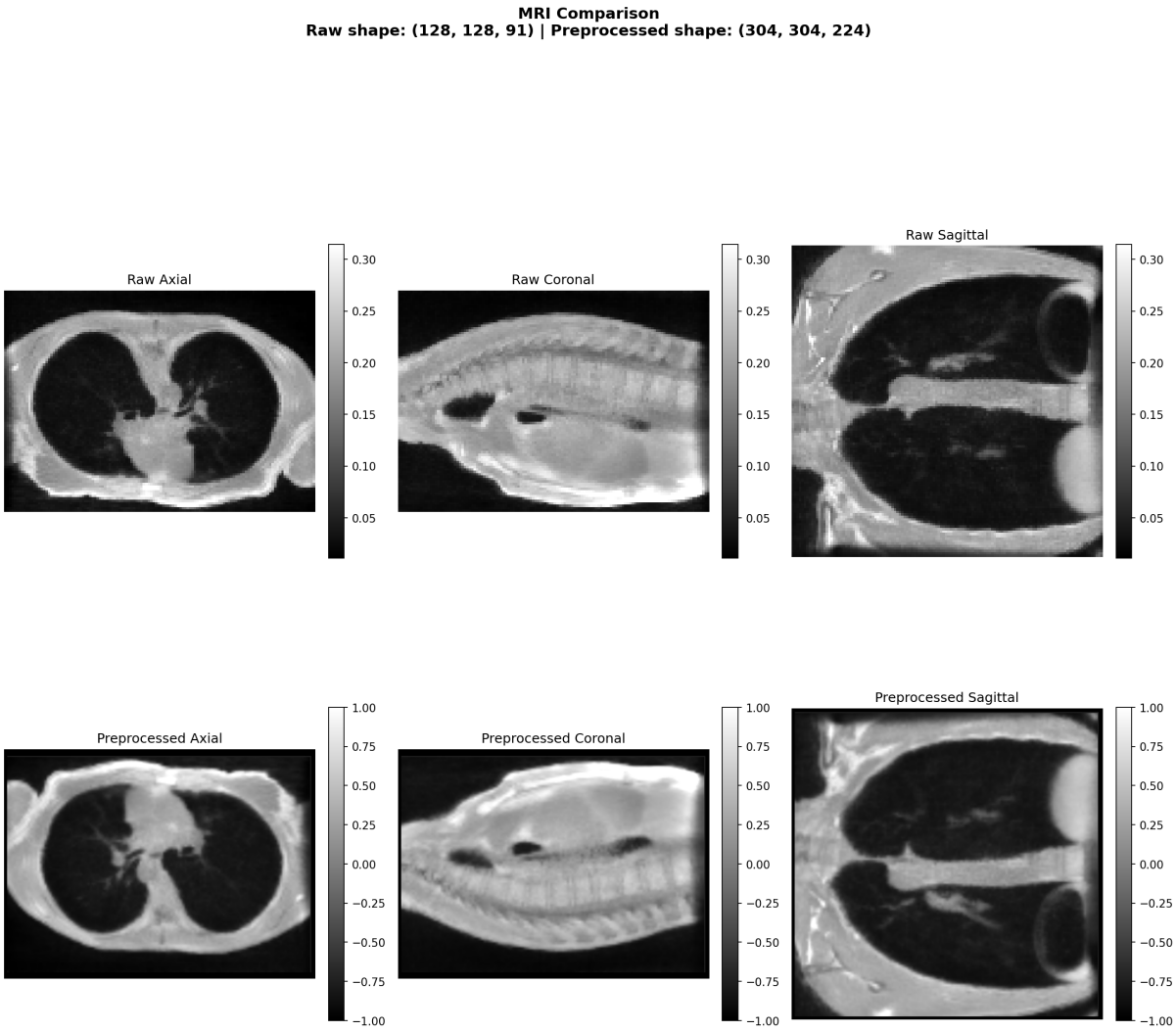


Figure 12 – MRI preprocessing comparison showing raw vs. preprocessed volumes (axial/coronal/sagittal views). Intensities scaled to $[-1,1]$ using percentile-based normalization with P_1 and P_{99} clipping.

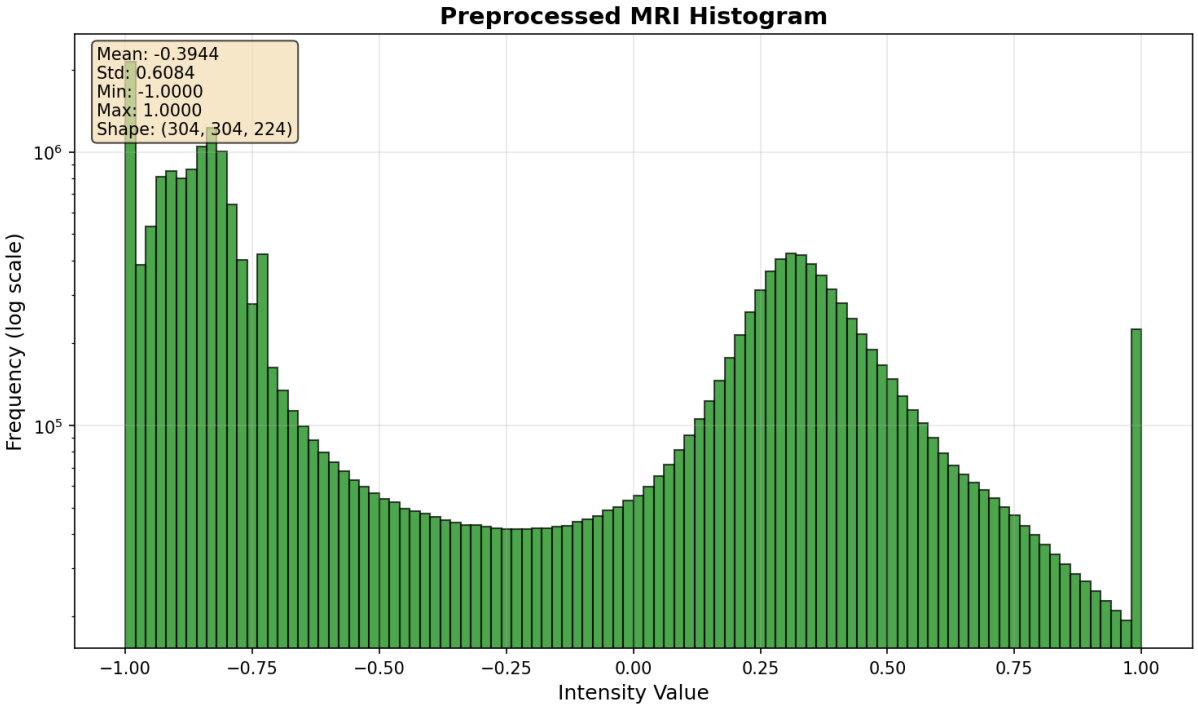


Figure 13 – MRI intensity histogram after percentile scaling (log-frequency). Vertical lines indicate P_1 and P_{99} clipping boundaries used for normalization to $[-1, 1]$.

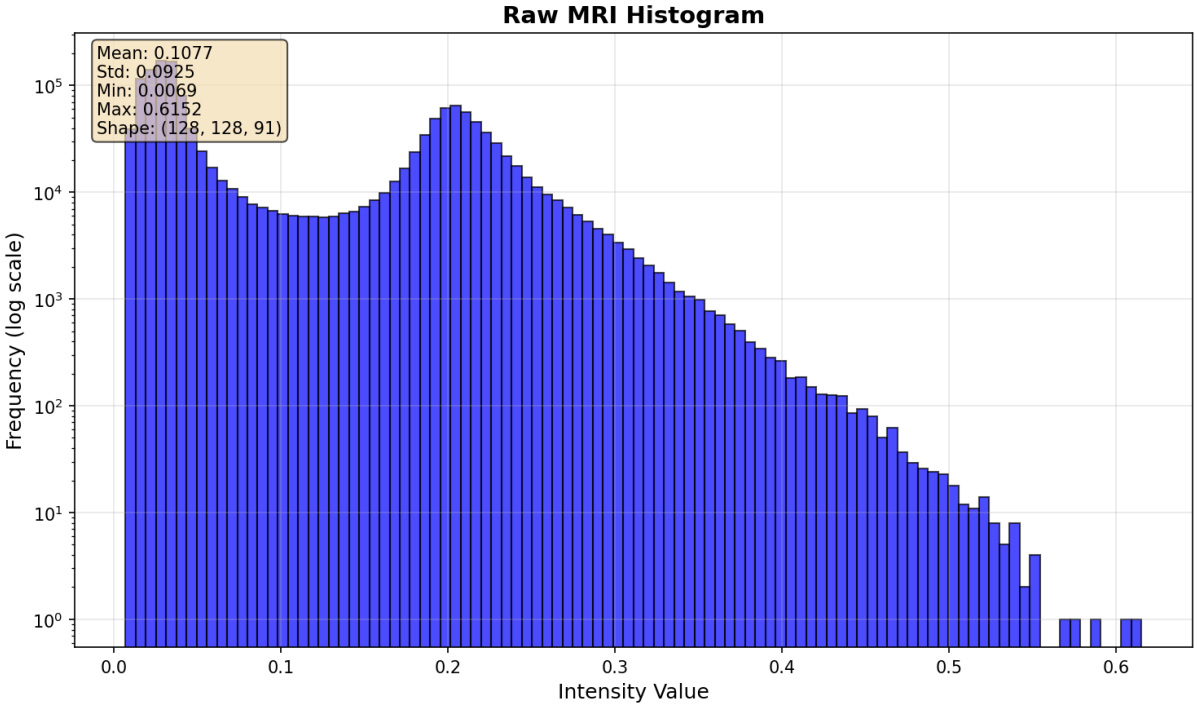


Figure 14 – MRI intensity histogram before preprocessing (log-frequency) showing original intensity distribution variability across different sequences and acquisition parameters.

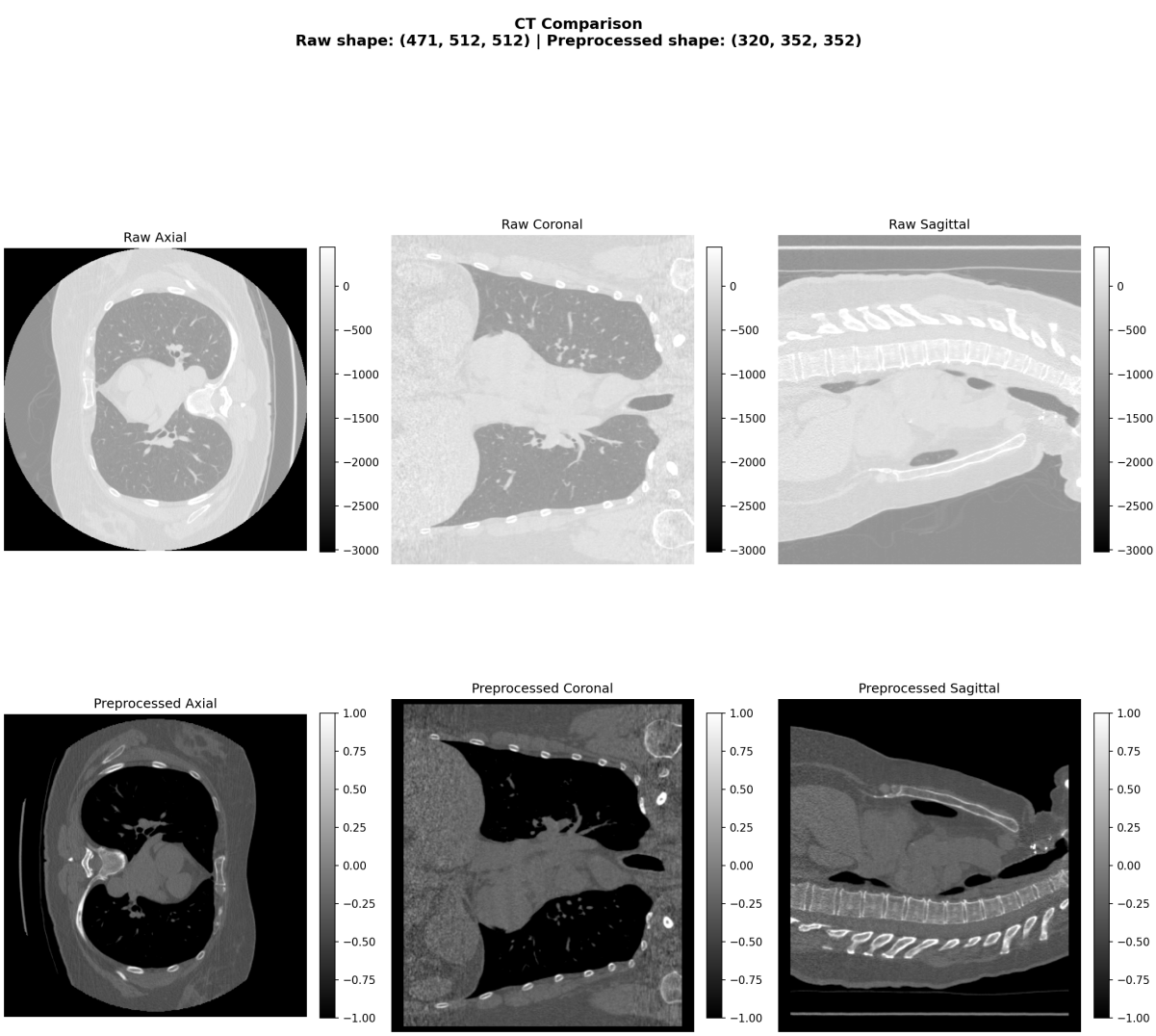


Figure 15 – CT preprocessing comparison showing raw vs. preprocessed volumes after HU windowing ($[-600, 1000]$ HU) and linear scaling to $[-1, 1]$. Note preservation of soft tissue contrast while clipping extreme bone/metal values.

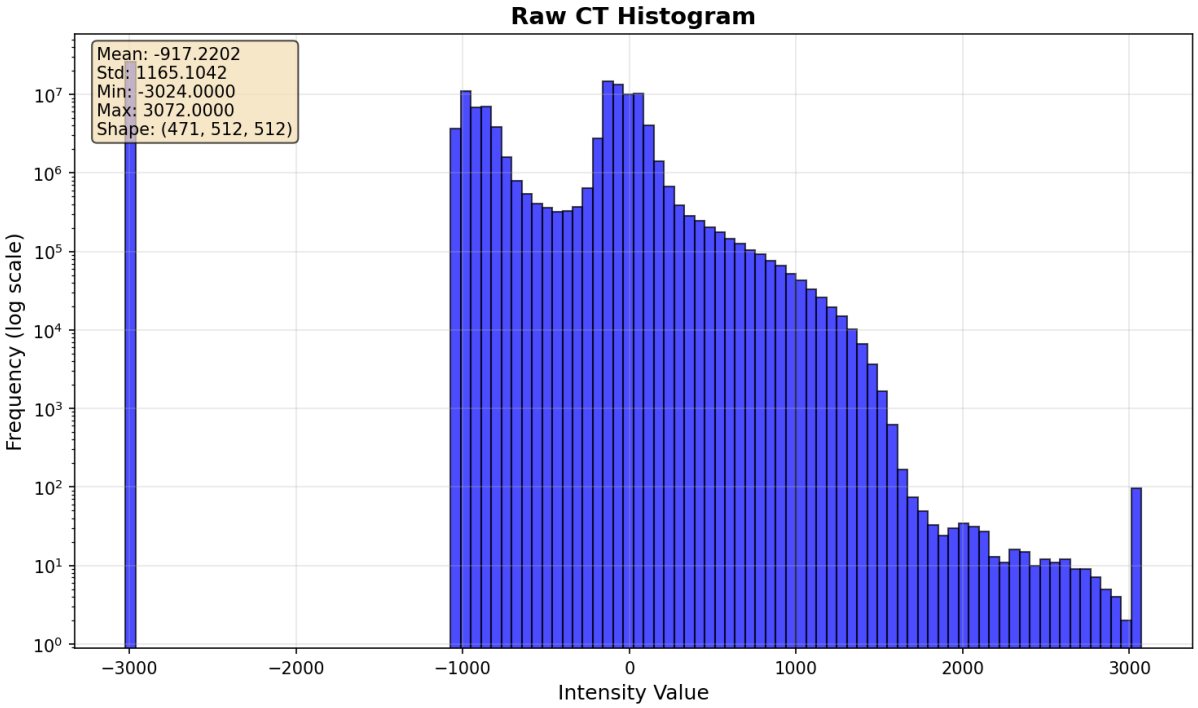


Figure 16 – CT intensity histogram before preprocessing (log-frequency) showing original HU distribution with extreme values from bone and metal artifacts.

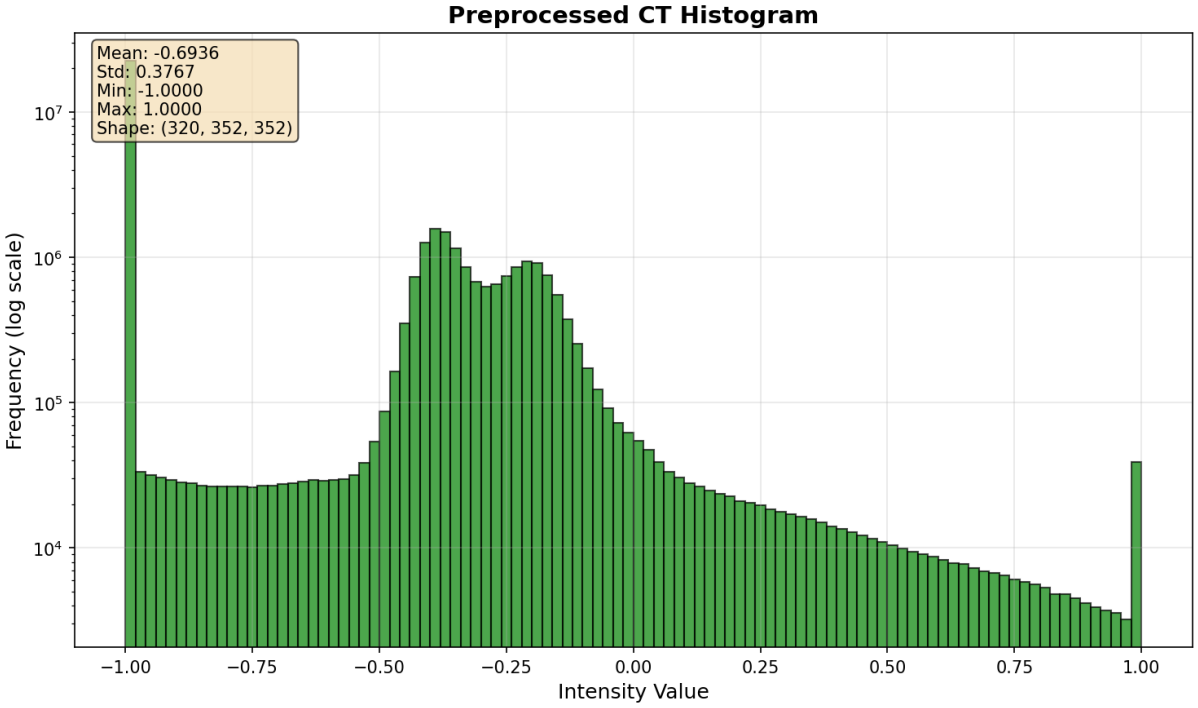


Figure 17 – CT intensity histogram after preprocessing (log-frequency). Approximately 12.3% of voxels were clipped during HU windowing, primarily extreme bone and metal values above 1000 HU.

C Architecture ResNet-9b

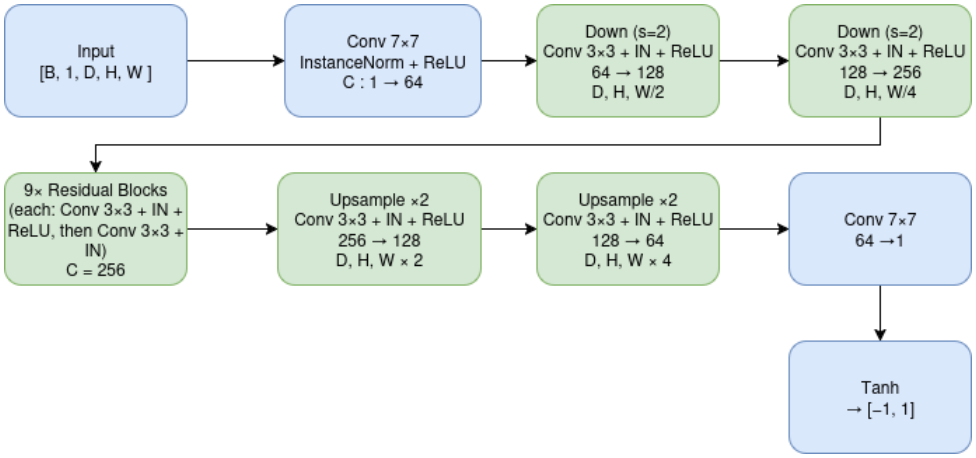


Figure 18 – Architecture of our ResNet-9b generator used for CT→MR translation.

D LSGAN–ResNet training details (200 epochs)

Train vs. Val curves (epochs)

We include the full per-loss curves:

— `cycle_ct` — Train vs. Val (epochs)

— `g_total` — Train vs. Val (epochs)

— `g_ct2mri` — Train vs. Val (epochs)

— `d_mri` — Train vs. Val (epochs)

— `d_ct` — Train vs. Val (epochs)

Validation metrics

Cycle fidelity	SSIM↑	PSNR↑ (dB)	MAE↓
CT cycle (x vs. $F(G(x))$)	0.7196	25.45	0.0668
MRI cycle (y vs. $G(F(y))$)	0.6848	24.78	0.0956
Distributional realism	CT→MRI: FID 196.53, KID 0.0758 | MRI→CT: FID 248.76, KID 0.1175		

Table 10 – LSGAN–ResNet (200 epochs): validation metrics.

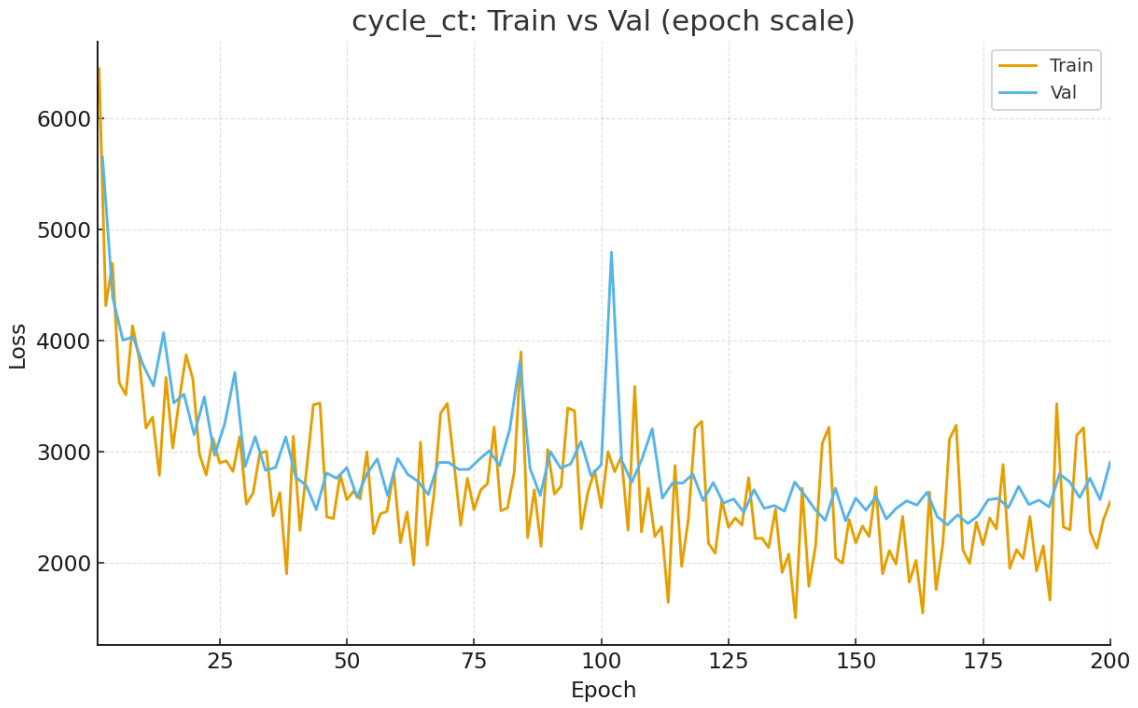


Figure 19 – `cycle_ct` — Train vs. Val (epochs).

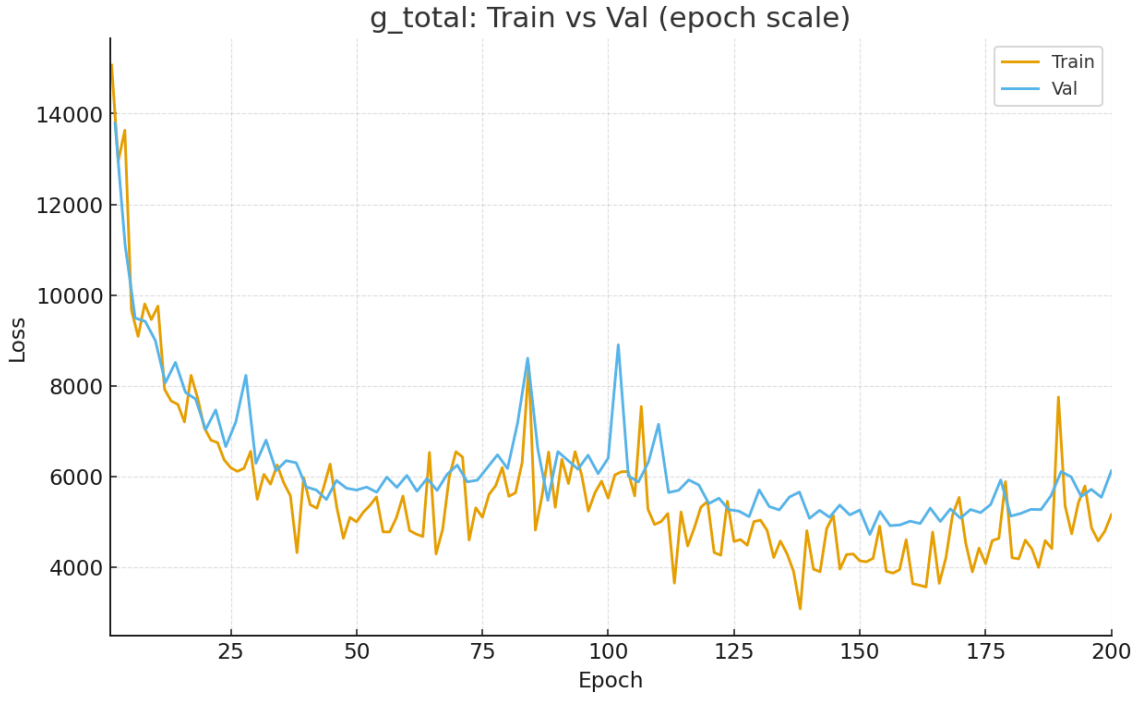


Figure 20 – `g_total` — Train vs. Val (epochs).

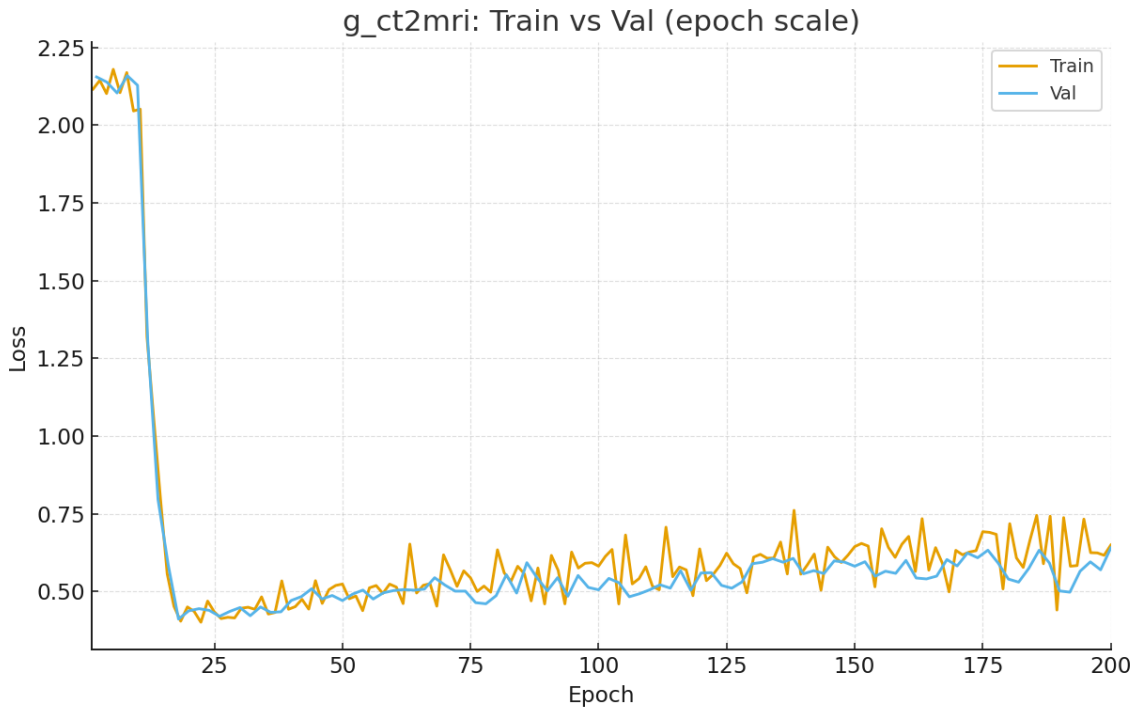


Figure 21 – `g_ct2mri` — Train vs. Val (epochs).

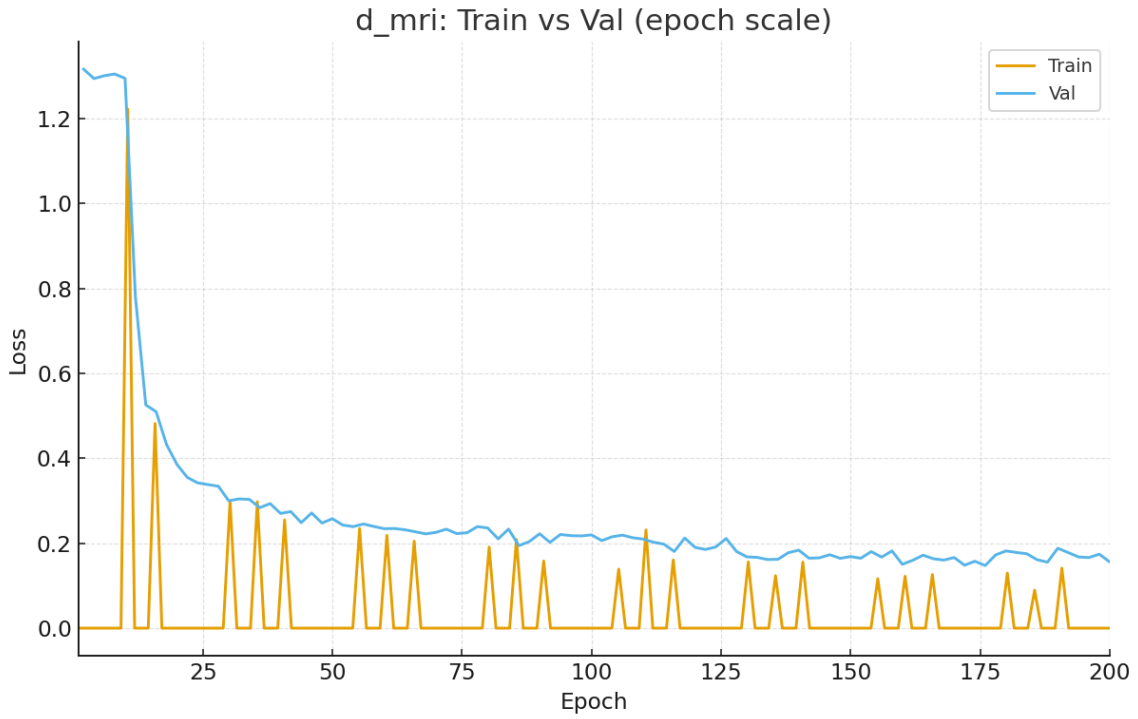


Figure 22 – `d_mri` — Train vs. Val (epochs).

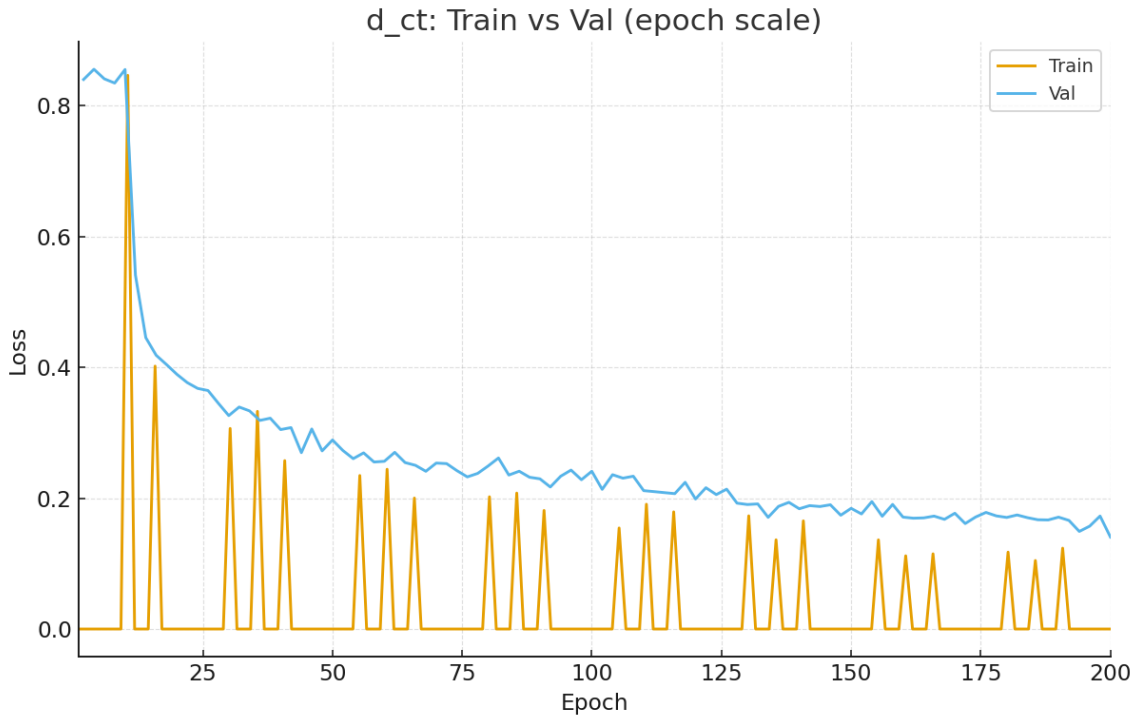


Figure 23 – `d_ct` — Train vs. Val (epochs).

References

Wang, J., Q. J. Wu, and F. Pourpanah (Sept. 2023). « DC-cycleGAN: Bidirectional CT-to-MR synthesis from unpaired data ». In: *Computerized Medical Imaging and Graphics* 108, p. 102249. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2023.102249. URL: http://dx.doi.org/10.1016/j.compmedimag.2023.102249.

Setio, A. A. A., A. Traverso, T. de Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, H. Geurts, R. van der Gugten, P.-A. Heng, B. Jansen, E. de Kaste, V. Kotov, J. Lin, J. T. Manders, N. Mastronardi, K. Messer, R. Mikhael, S. Mitra, L. Morra, S. Pal, J. Petersen, M. Prokop, M. Saletta, S. Schmidt, E. T. Scholten, H. Schulz, A. Sharma, V. Singh, M. Snoeren, L. J. Spreeuwers, M. B. Stegmann, J.-P. Thiran, E. L. Torres, R. Trapero, T. van Walsum, R. Wiemker, Y. Zhao, W. Zhu, D. Zinovev, K. Zuiderveld, and B. van Ginneken (2017). « Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge ». In: *Medical Image Analysis* 42, pp. 1–13. DOI: 10.1016/j.media.2017.06.015.

Chu, Y. et al. (2025). « Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences ». In: *Nature Communications* 16.1, p. 2262. DOI: 10.1038/s41467-025-12345-x.

Luo, G., K. Wang, J. Liu, S. Li, X. Liang, X. Li, S. Gan, W. Wang, S. Dong, W. Wang, P. Yu, E. Liu, H. Wei, N. Wang, J. Guo, H. Li, Z. Zhang, Z. Zhao, N. Gao, N. An, A. Pakzad, B. Rangelov, J. Dou, S. Tian, Z. Liu, Y. Wang, A. Sivalingam, K. Punithakumar, Z. Qiu, and X. Gao (2023). « Efficient automatic segmentation for multi-level pulmonary arteries: The PARSE challenge ». In: *arXiv preprint arXiv:2304.03708*.

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez (Dec. 2017). « A survey on deep learning in medical image analysis ». In: *Medical Image Analysis* 42, pp. 60–88. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005. URL: http://dx.doi.org/10.1016/j.media.2017.07.005.

Shen, D., G. Wu, and H.-I. Suk (2017). « Deep Learning in Medical Image Analysis ». In: *Annual Review of Biomedical Engineering* 19, pp. 221–248. DOI: 10.1146/annurev-bioeng-071516-044442. URL: https://pubmed.ncbi.nlm.nih.gov/28301734/.

Ronneberger, O., P. Fischer, and T. Brox (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. URL: https://arxiv.org/abs/1505.04597.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros (2017). « Image-to-Image Translation with Conditional Adversarial Networks ». In: *CVPR*.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros (2017). « Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks ». In: *ICCV*.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). « Generative Adversarial Nets ». In: *NeurIPS*.

Arjovsky, M., S. Chintala, and L. Bottou (2017). « Wasserstein GAN ». In: *ICML*.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville (2017). « Improved Training of Wasserstein GANs ». In: *NeurIPS*.

Miyato, T., T. Kataoka, M. Koyama, and Y. Yoshida (2018). « Spectral Normalization for Generative Adversarial Networks ». In: *ICLR*.

Mirza, M. and S. Osindero (2014). « Conditional Generative Adversarial Nets ». In: *arXiv:1411.1784*.

Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley (2017). « Least Squares Generative Adversarial Networks ». In: *ICCV*.

Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter (2017). « GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium ». In: *NeurIPS*.

Bińkowski, M., D. J. Sutherland, M. Arbel, and A. Gretton (2018). « Demystifying MMD GANs and KID ». In: *ICLR*.

Kim, T., M. Cha, H. Kim, J. K. Lee, and J. Kim (2017). « Learning to Discover Cross-Domain Relations with Generative Adversarial Networks ». In: *ICML*.

Yi, Z., H. Zhang, P. Tan, and M. Gong (2017). « DualGAN: Unsupervised Dual Learning for Image-to-Image Translation ». In: *ICCV*.

Liu, M.-Y., T. Breuel, and J. Kautz (2017). « Unsupervised Image-to-Image Translation Networks ». In: *NeurIPS*.

Benaim, S. and L. Wolf (2017). « One-Sided Unsupervised Domain Mapping ». In: *NeurIPS*.

Li, W., Y. Li, W. Qin, X. Liang, J. Xu, J. Xiong, and Y. Xie (June 2020). « Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy ». In: *Quantitative Imaging in Medicine and Surgery* 10.6, pp. 1223–1236. DOI: 10.21037/qims-19-885. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7276358/.

Nie, D., R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen (2017). « Medical Image Synthesis with Context-Aware Generative Adversarial Networks ». In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*. Vol. 10435. Lecture Notes in Computer Science. Springer, Cham, pp. 417–425. DOI: 10.1007/978-3-319-66179-7_48.

Hiasa, Y., Y. Otake, M. Takao, T. Matsuoka, K. Takashima, J. L. Prince, N. Sugano, and Y. Sato (2018). *Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size*. URL: https://arxiv.org/abs/1803.06629.

Wolterink, J. M., A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Isgum (2017). « Deep MR to CT Synthesis using Unpaired Data ». In: *arXiv preprint arXiv:1708.01155*. MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI). DOI: 10.48550/arXiv.1708.01155.

Jin, C.-B., H. Kim, M. Liu, W. Jung, S. Joo, E. Park, Y. S. Ahn, I. H. Han, J. I. Lee, and X. Cui (2019). « Deep CT to MR Synthesis Using Paired and Unpaired Data ». In: *Sensors* 19.10, p. 2361. DOI: 10.3390/s19102361.

Johnson, K. M., S. B. Fain, M. L. Schiebler, and S. Nagle (Nov. 2013). « Optimized 3D ultrashort echo time pulmonary MRI ». In: *Magnetic Resonance in Medicine* 70.5, pp. 1241–1250. DOI: 10.1002/mrm.24570.

Weiss, K., T. M. Khoshgoftaar, and D. D. Wang (May 2016). « A survey of transfer learning ». In: *Journal of Big Data* 3.1, p. 9. DOI: 10.1186/s40537-016-0043-6.

Avi-Aharon, M., A. Arbelle, and T. Riklin Raviv (2023). « Differentiable Histogram Loss Functions for Intensity-based Image-to-Image Translation ». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10, pp. 11642–11653. DOI: 10.1109/TPAMI.2023.3278287.

Afifi, M., M. A. Brubaker, and M. S. Brown (2021). « HistoGAN: Controlling Colors of GAN-Generated and Real Images via Color Histograms ». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7941–7950.

Ustinova, E. and V. Lempitsky (2016). « Learning Deep Embeddings with Histogram Loss ». In: *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*. Barcelona, Spain: Curran Associates, Inc., pp. 4170–4178.

Wang, Z., H. Li, W. Ouyang, and X. Wang (2016). « Learnable Histogram: Statistical Context Features for Deep Neural Networks ». In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Vol. 9905. Lecture Notes in Computer Science. Springer, Cham, pp. 246–262. DOI: 10.1007/978-3-319-46448-0_15.

Risser, E., P. Wilmot, and C. Barnes (2017). « Stable and Controllable Neural Texture Synthesis and Style Transfer Using Histogram Losses ». In: *arXiv preprint arXiv:1701.08893*. DOI: 10.48550/arXiv.1701.08893.