Laboratoire de Traitement de l'Information Médicale
Laboratory of Medical Information Processing

UNIVERSITÉ PARIS-EST CRÉTEIL VAL DE MARNE

Laboratoire de Traitement de l'Information Médicale

Université Paris-Est Créteil
UFR de Sciences et Technologie

Robust and High-Fidelity Synthetic CT Generation: Methodological Innovations for Quality and Multicentric Data

Master 2 Optique, Image, Vision, Multimédia parcours SIM

Yaghmoracen BELMIR
Industrial supervisors: Mr. Vincent BOURBONNE, Mr. Vincent JAOUEN
Academic supervisor: Ms. Corinne LAGORRE

Academic year 2024-2025

Acknowledgement

I would like to express my sincere gratitude to my industrial supervisors, Mr. Vincent Bourbonne and Mr. Vincent Jaouen, for their invaluable guidance and continuous support throughout this project.
My deepest thanks also go to my academic supervisor, Ms. Corinne Lagorre, and to the jury members, Ms. Blanche Bapst and Mr. Emmanuel Itti, for their insightful feedback and encouragement.
I gratefully acknowledge the financial support provided by CHRU Brest, which made this research possible.
Finally, I thank my colleagues and friends for their collaboration and moral support throughout this journey.

yaghmo.

Abstract

MRI-only workflows for stereotactic brain radiotherapy aim to shorten planning and avoid additional CT acquisitions, provided that pseudo-CT (pCT) volumes are structurally faithful. We investigate the robustness of a normalized-gradient field (NGF) loss within a supervised Pix2Pix MR→CT pipeline using paired data from the TCIA GammaKnife-Hippocampal brain collection. Four trainings are compared: NGF=20, NGF=0, a schedule-switch (hereafter 'switch-off') experiment (initial 40 epochs with NGF=20, followed by two parallel continuations with NGF maintained versus dropped), and a higher-resolution variant that trains on 256×256 crops extracted from native 512×512 volumes, improving preprocessing and masking. Evaluation uses PSNR and SSIM, computed both globally and within cranial masks to limit background bias, and aggregated over each 3D case, revealing that models trained with NGF exhibit improved structural fidelity critical for radiotherapy planning.

Keywords: MRI-only radiotherapy; pseudo-CT; Generative Adversarial Networks (GANs); Pix2Pix; normalized gradient field (NGF); structural fidelity; PSNR; SSIM; GammaKnife-Hippocampal; brain SRT.

Résumé

Les workflows MRI-only pour la radiothérapie stéréotaxique cérébrale permettent de raccourcir la planification et d'éviter l'acquisition supplémentaire de scanners CT, à condition que les volumes pseudo-CT soient structurellement fidèles. Cette étude examine la robustesse de la perte « normalized gradient field » (NGF) au sein d'un pipeline Pix2Pix supervisé IRM→CT, en exploitant les paires issues de la collection TCIA GammaKnife-Hippocampal. Quatre protocoles d'entraînement sont comparés : NGF=20, NGF=0, une expérience de changement de régime (40 premières époques avec NGF=20 puis deux continuations parallèles avec NGF maintenu ou supprimé), et une variante haute résolution qui utilise des crops 256×256 extraits du natif 512×512, associée à un prétraitement et un masquage améliorés. L'évaluation s'appuie sur les métriques PSNR et SSIM, calculées à la fois globalement et restreintes au masque crânien afin de limiter le biais d'arrière-plan, et agrégées par cas en 3D, montrant que les modèles entraînés avec NGF présentent une meilleure fidélité structurelle, essentielle pour la planification en radiothérapie.

Mots-clés : radiothérapie MRI-only ; pseudo-CT ; réseaux antagonistes génératifs (GAN) ; Pix2Pix ; normalized gradient field (NGF) ; fidélité structurelle ; PSNR ; SSIM ; GammaKnife-Hippocampal ; SRT cérébrale.

Contents

List of Figures

Acronyms

AI Artificial Intelligence
cGAN Conditional Generative Adversarial Network
CHRU Centre Hospitalier Régional Universitaire
CT Computed Tomography
DVH Dose-Volume Histogram
GAN Generative Adversarial Network
GPU Graphics Processing Unit
HU Hounsfield Unit
CNNs Convolutional neural networks
MRI Magnetic Resonance Imaging
MONAI Medical Open Network for AI
NGF Normalized Gradient Field
NEC Normalized Edge Consistency
pCT Pseudo-CT
PSNR Peak Signal-to-Noise Ratio
RTSTRUCT Radiotherapy Structure Set
ROI Region of Interest
SSIM Structural Similarity Index Measure
TCIA The Cancer Imaging Archive
NIH National Institutes of Health
DICOM Digital Imaging and Communications in Medicine
NIFTI Neuroimaging Informatics Technology Initiative
U-Net U-shaped Network (CNN Architecture for Segmentation)
Pix Pixel
I2I Image to Image
I/O Input/Output

Introduction

Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are fundamental imaging modalities widely used in brain radiotherapy planning. While MRI provides exquisite soft tissue contrast facilitating precise tumor and organ delineation, CT offers calibrated Hounsfield Unit (HU) required for accurate dose calculation and treatment optimization. Traditional workflows rely on acquiring both modalities for each patient, but this dual-modality approach introduces additional costs, patient burden, and potential registration inaccuracies.

MRI-only planning workflows aim to eliminate the need for CT acquisitions by synthesizing pseudo-CT (pCT) volumes from MRI scans. Generating pCTs that are anatomically and structurally faithful to real CT is critical to ensure that dose computations and clinical outcomes remain accurate and reliable. Achieving high structural fidelity, especially at bone/air and bone/soft tissue interfaces, is a known challenge due to the modality contrast differences and the inherent complexity of the cranial anatomy.

This report investigates the application of a Normalized Gradient Field (NGF) loss as a structural regularizer within a supervised image-to-image translation framework (Pix2Pix) for MR-to-CT synthesis. The contribution lies in evaluating the impact of NGF loss on anatomical preservation, testing different resolution strategies including native-resolution cropping, and explicitly modeling background air intensities in the synthesis pipeline. Results are quantitatively assessed using metrics sensitive to structural fidelity, including global and local SSIM and skull segmentation Dice coefficients, and qualitatively examined through volumetric renderings.

The goals of this work are to: (i) quantify the benefits of NGF regularization on structural preservation; (ii) determine the effects of spatial resolution strategies on pCT quality; (iii) propose refined loss formulations enhancing the clinical applicability of pCT generation.

The report is structured as follows: Chapters 1 and 2 review related work and detail materials and methods; Chapter 3 outlines experimental design and implementation; Chapter 4 presents quantitative and qualitative results; Chapter 5 discusses findings, limitations, future directions, and concludes with practical reflections. An appendix provides supplementary figures, an accessible Artificial Intelligence (AI) overview, and algorithmic details.

Host organization: LaTIM (Laboratoire de Traitement de l'Information Médicale)

LaTIM: Mission and Positioning

The **Laboratory of Medical Information Processing (LaTIM)** is a public academic–hospital joint research unit, bringing together the University of Brest (UBO), IMT Atlantique, INSERM, and the Brest University Hospital (CHRU Brest).[1] Its mission is to advance *image-guided diagnosis and therapy* via methodological research in medical image acquisition, processing, modeling, and clinical translation, with direct pathways toward patient care through translational research and innovation in the regional university hospital. Economically, LaTIM operates as a non-profit within the French public research ecosystem, funded by competitive national (e.g., ANR) and European (e.g., Horizon Europe) grants, institutional support, and extensive hospital/industrial partnerships.

Organization and Components

Figure 1 shows the current organizational chart (organigram). LaTIM is structured into several research teams with complementary scopes:[2]

- **ACTION** (Head: D. Visvikis) – methods and systems for *image-guided interventional and therapeutic actions*; robust reconstruction/registration, quantitative imaging, treatment planning and verification.
- **IMAGINE** (Head: S. Brochard) – *image analysis and modeling* for clinical applications; segmentation, synthesis, machine learning and data integration.
- **VISION** and **CYBER-HEALTH** – complementary axes on computer vision for health, cyber-physical systems, and e-health.

[1] See the official website: [1].
[2] Team info: [2, 3].

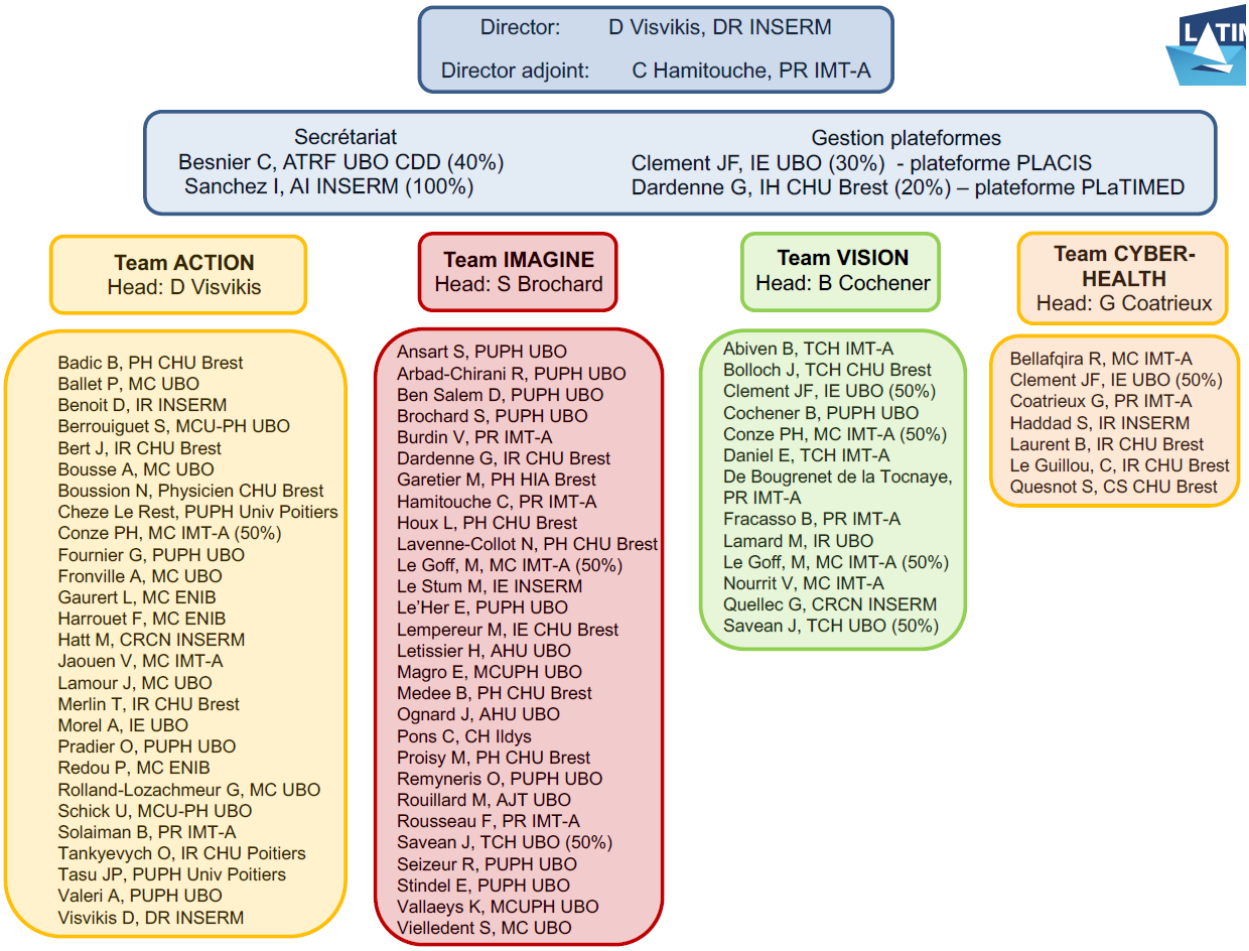


Figure 1 – LaTIM organizational chart (teams and governance).

Service hosting the internship. This internship was conducted within **Team ACTION**, in close collaboration with the **Radiotherapy Department of CHRU Brest**. Team ACTION's unique expertise lies at the interface of robust computational methods and clinical radiotherapy, focusing on multimodal image registration, synthesis pipelines, region-aware training losses for anatomical fidelity, and workflow integration for MRI-only planning. Collaboration with the clinical team enabled continuous feedback between technical solutions and clinical validation.

Affiliations and Supervision

The project received dual supervision: scientific guidance from **V. Jaouen** (Team ACTION, LaTIM) and clinical mentoring from **V. Bourbonne** (Radiotherapy, CHRU Brest). This combined mentorship ensured alignment with both leading-edge machine learning research and strict clinical requirements for safe deployment in brain radiotherapy.

Chapter 1

Background and Related Work

Magnetic resonance imaging (MRI) provides excellent soft-tissue contrast for target delineation in stereotactic radiotherapy (SRT), whereas computed tomography (CT) supplies calibrated Hounsfield Units (HU) for electron density and dose computation. The conventional workflow therefore acquires and registers both modalities. This registration step is a well-known source of geometric uncertainty in cranial SRT, and the dual-modality workflow adds logistical overhead. An *MRI-only* planning workflow seeks to remove the CT, provided that an accurate and structurally faithful pseudo-CT (pCT) can be synthesized from the MRI. This chapter reviews methods for pCT generation, the specific challenge of structural fidelity, normalized-gradient (NGF/NEC) regularization as a remedy, evaluation practices, and how our study positions itself within this landscape.

1.1 MRI-Only Planning and the Need for Structural Fidelity

Dose deposition in radiotherapy is sensitive to HU discontinuities at tissue interfaces, especially bone/air and bone/soft-tissue edges that dominate cranial anatomy. Consequently, pCT must be more than photorealistic: it must preserve *geometry* (edges, interfaces, cavity topology) with high fidelity so that the downstream dose map and dose-volume histogram (DVH) endpoints remain stable. Recent clinical validations suggest that GAN-based pCT can approach planning CT dosimetry in brain SRT when structural consistency is maintained, with very high gamma passing rates and DVH agreement [4]. Additionally, MRI-only workflows face practical adoption challenges in clinics: ensuring accurate patient positioning in dedicated RT masks[5], rigorous MRI system quality assurance for geometric precision[6], and the requirement for robust, clinically validated synthetic CT (sCT) algorithms that integrate seamlessly into radiotherapy planning software. Commercial solutions (e.g., Syngo.via RT Image Suite, Elekta Unity) are making sCT generation more accessible, but comprehensive clinical validation studies remain ongoing, and widespread adoption is still limited.

1.2 Families of MR→CT Synthesis Methods

Existing pCT approaches can be grouped into three broad families:

Atlas/registration-based Methods warp one or more CT atlases to the patient MRI and fuse them. They offer HU realism but are limited by registration errors and atlas diversity.

Segmentation/physics-based Methods segment MR into tissue classes and assign bulk HU or forward project via simplified physics. They are robust but lose fine structural detail (thin bone and sinus intricacies).

Learning-based Convolutional neural networks (CNNs) learn MR→CT mappings: (i) *supervised/paired* (e.g., Pix2Pix [7]) trained with aligned MR/CT pairs and pixel-wise losses; and (ii) *unpaired* (e.g., CycleGAN, CUT [8, 9]) trained adversarially with cycle or contrastive constraints to avoid paired data. While unpaired models are flexible, they can drift geometrically (hallucinations, local misplacements) because the adversarial objective rewards realism rather than anatomical faithfulness. GAN frameworks, in particular, have enabled significant advances in realistic image synthesis for pCT generation, supporting more robust and rapid inference while improving anatomical realism beyond conventional regression or atlas-mapping strategies[4].

1.3 Why Structural Regularization Is Needed

Even in paired settings, an $L1$ (mean absolute error) or $L2$ (mean squared error) reconstruction plus adversarial loss may not sufficiently penalize *where* structures are placed; pixel losses are dominated by large homogeneous regions, and the discriminator focuses on texture/contrast realism. Structural consistency must be explicitly encouraged. Edge-aware penalties (Sobel/LoG) help, but are sensitive to contrast and scale. A more principled option is to match *edge orientation* rather than raw gradients.

1.4 Normalized Gradient Fields (NGF) / Normalized Edge Consistency (NEC)

NGF similarity, widely used in multimodal image registration, compares the *direction* of gradients while down-weighting their magnitude and noise. For images I and J with spatial gradients $\nabla I, \nabla J$, define the normalized gradients

$$\tilde{\nabla} I = \frac{\nabla I}{\sqrt{\|\nabla I\|^2 + \epsilon^2}}, \qquad \tilde{\nabla} J = \frac{\nabla J}{\sqrt{\|\nabla J\|^2 + \epsilon^2}},$$

with sensitivity $\epsilon > 0$. A common NGF dissimilarity is

$$\mathcal{L}_{\text{NGF}}(I, J) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left(1 - \frac{\langle \nabla I(x), \nabla J(x) \rangle^2}{\left(\|\nabla I(x)\|^2 + \epsilon^2\right)\left(\|\nabla J(x)\|^2 + \epsilon^2\right)} \right),$$

i.e., one minus the (squared) cosine of the angle between gradients, stabilized by ϵ. Because orientations are contrast-invariant, the NGF penalty allows the network to change *intensity* (MR→CT

contrast) while keeping edge *geometry* anchored. Jaouen *et al.* showed that a one-sided NGF (a.k.a. Normalized Edge Consistency, NEC) combined with an adversarial objective can preserve structure in unpaired I2I without cycle constraints [10]. In our *paired* Pix2Pix setting, we use NGF as a structural regularizer with weight λ_{NGF} to stabilize anatomy in the synthesized CT.

1.5 Resolution Strategy

Resolution matters for cranial structures: naively resizing native 512×512 MR/CT to 256×256 can blur thin bone and sinus septa. Patch-based training on 256×256 *crops* from native resolution better preserves high-frequency detail, while keeping GPU memory manageable. At validation/inference, *sliding-window* prediction with overlap (and Gaussian blending) reconstructs full-field outputs from patch-trained models, mitigating edge artifacts. This strategy is standard in medical imaging and is central to our fourth experiment.

1.6 Evaluation Practices and Pitfalls

Two families of metrics are common:

Image similarity PSNR and SSIM are widely used. However, large air background inflates similarity: two volumes with identical black margins but different intracranial content can score artificially high. To reduce bias, compute metrics (i) globally, and (ii) *within a cranial mask* (possibly dilated by a few pixels to cover edge variability). When reporting per-case performance, aggregate slice-wise scores across the 3D volume (mean or median) to a single value per subject.

Task/structure metrics Dice for skull/air cavities, HU error histograms by tissue, and *dosimetric* endpoints (DVH differences; gamma analysis) reflect clinical safety more directly. Clinical studies in brain SRT report very high local/global gamma passing rates and non-significant DVH differences between pCT and planning CT when structure is preserved [4].

1.7 Positioning of This Work

The specific gaps we address are (i) a controlled, head-to-head comparison of supervised MR$\rightarrow$CT with/without NGF regularization, (ii) a *switch-off* protocol (keep vs. drop NGF mid-training from the same initialization) to isolate NGF's effect on metric drift, and (iii) the impact of *resolution strategy* (global resizing vs. native-resolution crops) under otherwise comparable pipelines. Concretely, we train Pix2Pix on paired MR/CT from the TCIA GammaKnife-Hippocampal collection in four settings: $\lambda_{\text{NGF}}=20$, $\lambda_{\text{NGF}}=0$, a branch study (20$\rightarrow$20 vs. 20$\rightarrow$0 after 40 epochs), and a high-fidelity cropped pipeline with improved preprocessing and masking. Evaluation reports PSNR/SSIM both globally and inside cranial masks, aggregated per 3D case; when available, skull

Dice is also computed. This design directly probes whether NGF stabilizes anatomy while allowing realistic CT contrast, and whether cropping improves structural fidelity compared with global resizing.

Summary

MRI-only SRT hinges on pCT volumes that are not merely realistic but *geometrically faithful*. Learning-based MR→CT translation benefits from explicit structural regularization. NGF/NEC offers a principled, contrast-invariant way to align edge orientations and has shown strong results in I2I [10]. Alongside resolution-aware training and masked evaluation, our study quantifies NGF's contribution in a supervised Pix2Pix setting and clarifies trade-offs that matter for safe clinical deployment, in line with prior clinical validations of GAN-based pCT [4].

Chapter 2

Materials and Methods

This chapter details the end-to-end pipeline used to synthesize pseudo-CT (pCT) volumes from brain MR using a supervised Pix2Pix framework regularized by a normalized-gradient (NGF) loss. The data consist of paired MR/CT from the TCIA *GammaKnife-Hippocampal* collection (390 patients), downloaded as DICOM and converted to NIfTI; at manuscript time, access is governed by the NIH controlled-access policy.[11, 12] We train in 2D on axial slices for efficiency and sample multiplicity, and validate/infer on full 3D volumes via sliding windows. Our central question is whether NGF improves anatomical fidelity of the synthesized CT—especially at tissue interfaces—without hindering contrast transfer.[10]

The generator is a 2D MONAI U-Net and the discriminator is a multi-scale PatchGAN, following Pix2Pix.[7, 13] Beyond an L_1 term and the adversarial loss, we incorporate NGF computed from spatial gradients (via `kornia`) to align local orientations between MR input and pCT output while allowing intensity changes.[14] We compare four trainings: (*i*) $\lambda_{\mathrm{NGF}}{=}20$; (*ii*) $\lambda_{\mathrm{NGF}}{=}0$; (*iii*) a switch-off study branching at epoch 40 (NGF kept vs. dropped); and (*iv*) a high-fidelity variant using centered $256{\times}256$ crops from native $512{\times}512$, mask-restricted losses, and a background term enforcing air to -1. Preprocessing standardizes geometry (affine-based axis flips) and normalizes MR/CT to $[-1, 1]$.

Evaluation includes global and mask-restricted PSNR/SSIM aggregated per 3D case, plus skull Dice for structure overlap; model selection uses the highest validation PSNR.[4] Practical choices (caching, workers, batch size 1, local GPU vs. cloud) are documented for reproducibility.

2.1 Dataset

This study uses the *GammaKnife-Hippocampal* collection hosted by The Cancer Imaging Archive (TCIA), titled "Gamma Knife MR/CT/RTSTRUCT Sets With Hippocampal Contours". The cohort comprises **390 human subjects** (head), each treated with Gamma Knife stereotactic radiosurgery for vestibular schwannoma (VS), trigeminal neuralgia (TGN), or metastatic disease. For every subject, at least one high-resolution (1 mm slice thickness) axial **T1 FLASH MR** study and a **planning CT** are provided; where available, radiotherapy planning objects (RTSTRUCT/RTPLAN/RTDOSE), DICOM registrations (REG), additional MR sequences, and follow-up MR studies are

included. All MR studies used for planning were rigidly co-registered to the CT frame in the clinic, and TCIA provides both the *aligned secondary* MR series and the corresponding DICOM registration file. A "head" ROI contour is also included to mask the stereotactic frame and inferior reconstruction artifacts.[1]

Table 2.1 – GammaKnife-Hippocampal collection (official TCIA summary).

Name	GammaKnife-Hippocampal (Image Collection)
DOI	10.7937/Q967-X166
Body site / Species	Head / Human
Subjects	**390**
Data types	MR, CT, RTSTRUCT, RTPLAN, RTDOSE, REG
Size (compressed)	≈ 312.06 GB
Release status	Complete; *Limited* access (see note below)
Last update	2022-07-22

Per-modality counts (as provided by TCIA).

- CT: **390** series
- MR: **3,868** series (includes aligned secondaries and additional sequences)
- RT objects: REG **872**, RTDOSE **928**, RTPLAN **928**
- RTSTRUCT: **931** (planning), plus **390** hippocampal research structure sets (multi-observer)

Registration and planning context. All MR images designated "Co-registered to CT" are resampled to the CT frame of reference (field of view and voxel size may differ from the MR primaries). The vendor planning system (GammaPlan) and MIM software were used clinically to perform and validate rigid MR→CT registrations; the DICOM RTREG and aligned MR secondaries are exported for research use. Follow-up imaging is available for 197 subjects (median 2 studies, range 1–13) but remains in its own frame of reference and is not co-registered to the planning CT.

Access note. Due to changes in the NIH Controlled Data Access Policy, TCIA indicates that some controlled datasets (including this collection) are temporarily unavailable for direct download via the usual portal; TCIA plans alternative access mechanisms. Researchers must also follow TCIA's Data Usage Policy and cite the dataset DOI in publications.

[1] Official collection page and summary numbers are provided by TCIA; see the Data Citation in §2.1.

Scope used in this work. We **restricted** ourselves to the clinically *co-registered T1 MR and planning CT* pairs. Series were downloaded in DICOM and converted to NIfTI for processing. Case splits were performed at the *patient level* (70%/20%/10% train/validation/test). Because our models were 2D, we trained on axial slices sampled from the 3D volumes; per volume, up to `NUM_SAMPLES_MAX=30` slices were extracted (center-cropped to the region of interest in later experiments) while preserving the MR/CT alignment. All other RT objects (RTSTRUCT/RT-PLAN/RTDOSE/REG) were not used to supervise learning but informed the dataset's clinical provenance.

This collection is the largest publicly available dataset with paired high-resolution MR/CT brain images specifically acquired for Gamma Knife stereotactic radiosurgery, including comprehensive multi-observer hippocampal contours, clinically validated rigid registrations using stereotactic frames, and full radiation therapy planning data. These unique features make it exceptionally well suited for supervised pseudo-CT learning, structural fidelity studies, and radiotherapy planning research.

2.2 Preprocessing

All experiments follow a 2D supervised MR→CT pipeline implemented with MONAI transforms and datasets [13]. Raw DICOM images were converted to NIfTI and paired via the TCIA metadata of the *GammaKnife-Hippocampal* collection [11]. The preprocessing differs slightly across experiments to study the impact of in-plane resizing vs. cropping, while keeping a common intensity normalization to $[-1, 1]$.

Pairing and format conversion

For each case, the diagnostic MR and its co-registered planning CT were identified using the collection's metadata CSV and converted to NIfTI with headers preserved. Only MR and CT were retained for training/validation; no RTSTRUCT was used in this work.

2D sampling strategy (order of operations)

All trainings are axial 2D. The *order* explicitly differs by experiment:

- **Exp. 1–3 (resize → sample):** starting from native 512×512, MR and CT are first *resized* in-plane to 256×256 (trilinear), then *axially sampled* into 2D slices/patches.
- **Exp. 4 (center-crop → sample):** to preserve native resolution, we *center-crop* a 256×256 window directly from the original 512×512 (no global downscaling), then *axially sample* slices from the cropped volume.

Training data is sampled in 2D slices with resizing or cropping applied as described, allowing efficient batch training. Validation data consists of single center-cropped axial slices (one per volume) at native 256×256 or 512×512 resolution depending on the experiment but without slice

sampling or augmentation. The test set remains fully 3D with no cropping or resizing, preserving the native $512 \times 512 \times N$ resolution for volumetric evaluation.

To avoid empty MR slices (e.g., acquisitions with clipped cranial coverage), we use a custom transform derived from `RandSpatialCropSamplesd` that *rejects* a candidate slice if the MR patch contains fewer than a specified number of supra-threshold voxels:

- `RandSpatialCropSamplesdWithMinNonZero(target=SRC, min_nonzero=1000)` keeps only MR patches with sufficient non-zero content (threshold applied after the initial MR scaling step).
- Up to `NUM_SAMPLES_MAX = 30` valid axial slices are drawn per 3D study for training.

Post-sampling orientation check (before saving)

After 2D *sampling* and *before saving* each MR/CT (and mask, when used) slice pair to disk, we enforce consistent in-plane orientation. Although MR and CT are co-registered in the source data, direct array inspection revealed occasional left–right or anterior–posterior inversions. We therefore compare the affine translation components; if they differ, we flip the CT (and mask) slice along X and/or Y to match the MR, then write using the MR affine/header when voxel sizes coincide, else fall back with a warning. This guarantees that all saved 2D samples (MR, CT, and mask) are aligned for training.

Intensity normalization (common to all experiments)

To match the generator's `tanh` output, both modalities are mapped to $[-1, 1]$:

- **MR**: percentile scaling of the input (1–99.9%) to $[0, 1]$ with clipping, then affine remap to $[-1, 1]$.
- **CT**: linear mapping of Hounsfield units from $[-1000, 3000]$ to $[-1, 1]$, so that air (≈ -1000 HU) maps near -1.

Cranial masking (Experiment 4 only)

For the high-fidelity experiment, a binary cranial mask is generated from CT using TotalSegmentator [15] and lightly dilated in-plane (2 px) to increase robustness to small misregistrations. The mask is used downstream (Section 2.4) to (i) restrict supervised losses to the head region and (ii) encourage background to the air value (approximately -1 after normalization).

Quality control

We visually inspect batches from the validation loader (MR/CT pairs, and masks in Exp. 4) to confirm alignment, intensity ranges, and the effectiveness of the non-empty slice filter. This avoids

learning from trivially empty inputs and reduces orientation-related artifacts introduced at save time.

This careful preprocessing pipeline, combined with the comprehensive clinical provenance of the GammaKnife-Hippocampal dataset, ensures rigor and consistency throughout our supervised MR-to-CT synthesis experiments.

2.3 Model

To achieve high-fidelity synthetic CT volumes that preserve anatomical consistency with MR inputs, we adopt a Conditional Generative Adversarial Network (cGAN) model in the Pix2Pix framework [7]. This approach leverages paired MR–CT data to learn a mapping conditioned on the MR input, producing outputs that are both photorealistic and anatomically faithful. The generator is a 2D U-Net [16] with residual units implemented in MONAI [13], while the discriminator employs a multi-scale PatchGAN architecture to enforce realistic local textures and structures. We extend the loss function by adding a normalized-gradient field (NGF) term computed via Kornia [14] to stabilize edge consistency between modalities.

2.3.1 Model choice: From supervised regression to conditional GANs

A plain convolutional regressor trained with an ℓ_1 or ℓ_2 loss can learn a deterministic MR→CT mapping but often results in over-smoothed outputs, particularly at bone–soft tissue interfaces. Generative adversarial networks (GANs) [17] address this by training a discriminator to penalize unrealistic high-frequency characteristics, encouraging the generator to produce sharper, more realistic textures. The Pix2Pix framework [7] specializes this idea to *paired* image-to-image translation by feeding the discriminator the conditioning MR alongside the CT, enforcing *conditional realism*. For medical applications with available alignment (such as ours), Pix2Pix is a principled and effective choice.

2.3.2 Generator architecture details

The generator $G : \mathbb{R}^{1\times H\times W} \to \mathbb{R}^{1\times H\times W}$ is a 2D U-Net implemented using MONAI's `UNet` [13], configured for single-channel input and output, and employing residual units to improve gradient flow and convergence stability:

- **Contracting path:** stacked convolutional blocks with stride-2 downsampling progressively increase channel depth while halving spatial resolution, typically doubling channels (e.g., $64 \to 128 \to 256 \to 512$).
- **Expanding path:** mirrored transpose convolution or upsample+conv blocks, with skip connections concatenating encoder features to preserve spatial details.
- **Residual units:** short skip connections within blocks to ease optimization and improve convergence; in our experiments, `num_res_units` is set to 10.

- **Activation and normalization:** LeakyReLU used within blocks; the final layer uses a `tanh` activation to output $G(A) \in [-1, 1]$, matching input intensity normalization (see Section 2.2).

Tensor interfaces. Training inputs are mini-batches of axial slices shaped $[B, 1, H, W]$, where B is batch size, H and $W = 256$ are image height and width. The generator outputs tensors of the same shape. An optional identity path $G(B)$ can be evaluated during warm-up for stability purposes (identity loss is set to zero weight by default in our main experiments).

2.3.3 Discriminator design

The discriminator D follows a multi-scale PatchGAN architecture [7], which produces a grid of local realism scores over image patches instead of a single global classification. We use two discriminators (`num_d` = 2) operating on downsampled inputs to better capture texture at multiple scales. Each discriminator consists of four convolutional layers with 128 filters in the first layer (`num_layers_d` = 4, `num_filters_d` = 128). At training, D receives concatenated conditioning and target images stacked by channels: $[A_m, B_m]$ for real pairs and $[A_m, G(A)_m]$ for synthetic pairs, where $X_m := M \odot X$ denotes masking in experiments that use cranial masks.

2.3.4 Loss components and objectives

We summarize the loss heads and their purposes, leaving detailed weights and formulas to Section 2.4:

- **Adversarial loss (cGAN).** D classifies real from fake image pairs; G is trained to fool D using the standard Pix2Pix adversarial objective [7].
- ℓ_1 **content loss.** Measures pixel-wise absolute differences between $G(A)$ and ground truth B, encouraging HU-like intensity alignment after normalization. Computed within the cranial mask M in Experiment 4.
- **Normalized Gradient Field (NGF) loss.** Computed between the gradients of $G(A)$ and the input MR A (not B), this loss promotes anatomical edge orientation consistency across modalities, enforcing structural fidelity.
- **Identity loss (optional).** $\|G(B) - B\|_1$ encourages the model to leave images in the target domain unchanged; applied optionally during early warm-up for training stability.
- **Background loss (Experiment 4 only).** Drives the non-cranial background towards air intensity (normalized to ≈ -1), complementing the masked structural losses.

2.3.5 Masking and region-specific supervision

When a cranial mask M is used (only in Experiment 4), it is broadcasted to $[B, 1, H, W]$ and applied element-wise to the synthesized output and ground truth for the adversarial, ℓ_1, and NGF losses. A 2-pixel dilation mitigates sensitivity to minor registration errors. The masked pairs are also presented to the discriminator to focus realism decisions on anatomically relevant regions.

2.3.6 Why Pix2Pix? Advantages for paired medical data

Given that our dataset consists of rigorously co-registered MR/CT pairs, we prefer the conditional Pix2Pix GAN to unpaired image-to-image methods (e.g., CycleGAN, CUT) to precisely control intensity fidelity required in radiotherapy. Pix2Pix combines a pixelwise ℓ_1 loss for intensity anchoring with the PatchGAN adversarial loss to avoid over-smoothing, while NGF regularization enforces geometric consistency especially at tissue interfaces—three non-negotiable requirements for pseudo-CT volumes intended for downstream dose calculations [4].

2.4 Training Strategy

This section details the optimization objectives, batching and scheduling choices, and the four experimental arms designed to quantify the contribution of the normalized gradient field (NGF) constraint to structural fidelity in MR→CT synthesis with Pix2Pix [7]. All training was performed on 2D axial slices sampled from 3D volumes (Sec. 2.2), with intensities normalized to $[-1, 1]$ to match the generator's `tanh` output. Data loading, caching, and patch sampling utilized MONAI [13].

2.4.1 Objectives

Let A denote the MR input, B the target CT, G the generator ($G : A \to B$), D the multi-scale PatchGAN discriminator, and $M \in \{0, 1\}^{H \times W}$ an optional cranial mask (used only in Exp. 4). For masked tensors, write $X_m := M \odot X$, and define $\bar{M} := 1 - M$. The generator loss per batch is

$$\begin{aligned} \mathcal{L}_G = \lambda_{\ell_1} \, \|G(A)_m - B_m\|_1 \\ + \lambda_{\text{NGF}} \, \mathcal{L}_{\text{NGF}}\big(G(A)_m, A_m; \alpha\big) \\ + \lambda_{\text{GAN}} \, \mathcal{L}_{\text{GAN}}\big(D(G(A)_m)\big) \\ + \lambda_{\text{bg}} \, \|\bar{M} \odot (G(A) - a_{\text{air}})\|_1. \end{aligned} \tag{2.1}$$

where $a_{\text{air}} = -1$ is the normalized air intensity level, α controls the NGF normalization smoothing parameter, and $\lambda_\star$ are scalar weights. The NGF loss enforces local edge orientation alignment, robust to contrast variations, and is implemented via finite-difference spatial gradients using Kornia [14]:

$$\mathcal{L}_{\text{NGF}}(X, Y; \alpha) = 1 - \tfrac{1}{2} \, \mathbb{E}\left[\left(\frac{\nabla_x X \cdot \nabla_x Y + \nabla_y X \cdot \nabla_y Y}{\sqrt{\|\nabla X\|^2 + \alpha^2} \, \sqrt{\|\nabla Y\|^2 + \alpha^2}} \right)^2 \right]. \tag{2.2}$$

The discriminator loss uses the standard real/fake adversarial loss applied to masked inputs in Experiment 4.

Weights and constants. Unless otherwise specified, the weights are: $\lambda_{\ell_1} = 140$, $\lambda_{\text{NGF}} \in \{0, 20\}$, $\alpha = 0.08$, $\lambda_{\text{GAN}} = 1$, and λ_{bg} is active only in Experiment 4.

2.4.2 Optimization, batching, and schedule

Training is performed for 70 epochs using the Adam optimizer (lr $= 2\times10^{-4}$, default β parameters) with a batch size of 14 and mixed shuffling. Each iteration computes losses for G and D on the same mini-batch; gradients are detached from $G(A)$ when updating D. An initial autoencoding warm-up phase of 2 epochs feeds G the source input A as the target ($B \leftarrow A$) to strengthen low-level structural preservation before adversarial learning dominates. Model checkpoints are saved based on the best validation PSNR (Sec. 2.5), with safety snapshots every 10 epochs.

Validation employs deterministic sliding-window inference with Gaussian blending on 256×256 ROIs with 0.75 overlap [13].

2.4.3 Experimental arms

Four experimental variants are designed to isolate NGF's effect and resolution strategy impact:

- **Exp. 1 (NGF=20):** Baseline Pix2Pix with $\lambda_{\text{NGF}} = 20$, resized inputs from $512 \to 256$, and global loss computation ($M \equiv 1$).
- **Exp. 2 (NGF=0):** Identical to Exp. 1 but with NGF disabled ($\lambda_{\text{NGF}} = 0$) to quantify NGF's edge-consistency contribution.
- **Exp. 3 (Switch-off study):** Training as in Exp. 1 for 40 epochs, then branched into two parallel continuations: (i) maintaining $\lambda_{\text{NGF}} = 20$; (ii) dropping to $\lambda_{\text{NGF}} = 0$. Validation PSNR/SSIM divergence post-branch investigates NGF's regularization effect.
- **Exp. 4 (Crops + masks + background loss):** Instead of global resizing, use centralized 256×256 crops from native 512×512 inputs to preserve high-frequency details. Losses (ℓ_1, NGF, GAN) are computed inside a cranial mask M (obtained from CT via TotalSegmentator [15]). A background penalty loss $\lambda_{\text{bg}} \|\bar{M} \odot (G(A) - a_{\text{air}})\|_1$ is added to encourage non-cranial regions toward air intensity (-1). The mask is dilated by 2 pixels to reduce edge sensitivity.

2.4.4 Implementation notes and environments

Models are implemented in PyTorch with MONAI components [13]. The NGF loss uses Kornia [14] for spatial gradient computations. Training was conducted on a local workstation (limited to batch size 1 due to memory constraints) and on Google Colab Pro (NVIDIA L4) for extensive experiments; caching and sliding-window validation ensured reproducibility between platforms.

2.5 Validation and Inference

We distinguish *validation during training* (2D, slice-wise) from *test-time inference* (3D, volume-wise). Validation monitors learning and selects checkpoints, while test-time inference generates full pseudo-CT (pCT) volumes for final evaluation. Notably, only **Experiment 4** applies cranial masks during training/validation to compute a *masked* PSNR; **Experiments 1–3** compute *global* PSNR without masking. The NGF *switch-off* in **Experiment 3** tracks the evolution of *global* PSNR when NGF is kept versus dropped mid-training.
Validation uses a single center-cropped 2D slice per subject at native resolution (512×512 or 256×256), whereas training used multiple resized/cropped slices per volume (§2.2). The test set remains fully 3D, preserving volumetric integrity for comprehensive assessment.

2.5.1 Validation during training (2D)

At the end of each epoch, the current generator $G_{\mathrm{MR}\to\mathrm{CT}}$ is evaluated on center-cropped validation slices with MONAI's sliding-window inference [13]:

- **Tiling and blending.** Axial tiles of 256×256 with 75% overlap and Gaussian blending, implemented via `sliding_window_inference` [13].
- **Stored panels.** MR, reference CT, and generated pCT panels are saved per validation case for qualitative review.
- **Metrics and checkpointing.**
  - **Experiments 1–3:** Compute *global* PSNR per slice, average across slices and cases to yield an epoch score; checkpoint selection optimizes this metric. Periodic model snapshots are preserved.
  - **Experiment 4:** Compute both *global* and *masked* PSNR (within the dilated cranial mask) to reduce background influence. Checkpointing remains based on *global* PSNR to ensure comparability; masked PSNR is logged for in-depth analysis.

2.5.2 Test-time inference (3D)

For each MR/CT pair in the test set, a full 3D pseudo-CT is generated with these steps:

1. **Geometry harmonization (test-only).** After DICOM→NIfTI conversion, MR and CT volumes may exhibit flipping or orientation mismatches. The test pipeline applies `ResampleToMatchd` to register MR (and the resulting pCT) onto the CT grid and affine space, ensuring voxelwise metric validity [13].
2. **Slice-wise generation.** The generator $G_{\mathrm{MR}\to\mathrm{CT}}$ is applied slice-by-slice axially, using 256×256 tiles with 75% overlap and Gaussian blending, mirroring validation.

3. **Intensity restoration.** Outputs in $[-1, 1]$ are rescaled to Hounsfield units (e.g., $[-1000, 3000]$) using inverse normalization.

4. **Optional background handling.** Voxels outside the cranial mask may be forcibly set to air intensity (-1000 HU) to suppress background artifacts before export; this does not influence training.

5. **Final metrics.** Test reporting prioritizes 3D SSIM (global and masked) and cranial Dice coefficient; PSNR is reserved for training validation and checkpointing.

Protocol summary across experiments. Validation is 2D (512×512 or 256×256 single slices) using sliding-window inference. **Experiments 1–3** employ only *global* PSNR; **Experiment 3** further tracks global PSNR evolution under NGF switch-off. **Experiment 4** combines *global* and *masked* PSNR for validation but uses global PSNR for selection. Test inference applies the common volumetric pipeline and reports 3D SSIM and Dice for all experiments.

2.6 Metrics

We report three complementary criteria: (i) peak signal-to-noise ratio (PSNR) for *training-time validation and checkpoint selection* (2D, slice-wise), (ii) structural similarity (SSIM) both *globally* and *within a cranial mask* for *test-time* 3D evaluation, and (iii) Dice overlap for the skull mask to quantify geometric consistency. Intensities are normalized to $[-1, 1]$, hence $\texttt{data\_range} = 2$ for PSNR/SSIM computations.

2.6.1 PSNR (validation only)

PSNR between a reference G and prediction P is defined as

$$\mathrm{PSNR}(G, P) \;=\; 10 \log_{10}\!\left(\frac{L^2}{\mathrm{MSE}(G, P)}\right), \qquad \mathrm{MSE}(G, P) = \frac{1}{N}\sum_{i=1}^{N}(G_i - P_i)^2, \tag{2.3}$$

where $L = 2$ is the dynamic range for $[-1, 1]$-scaled images.[2] PSNR quantifies pixel-wise fidelity but is sensitive to intensity shifts and less aligned with human perceptual judgement compared to structural metrics.

During training-time validation in **Experiment 4**, PSNR is computed locally by first applying a cranial mask that sets background voxels outside the head region to a uniform value (-1). This ensures that the PSNR loss focuses on anatomical regions and is not biased by differing background intensities in the MR, CT, and pseudo-CT volumes.

For **Experiments 1–3**, PSNR is computed *globally* across the full slice, including background. PSNR values are computed per slice, averaged across slices for per-case scores, then averaged

[2] When reporting PSNR in Hounsfield units (HU) space, L should reflect the HU range; here PSNR is computed only in normalized space during validation.

over cases for per-epoch metrics used for model selection. Experiment 3 additionally contrasts epoch-wise PSNR curves when the NGF term is kept versus switched off. Experiment 4 logs both global and masked PSNR but retains global PSNR for checkpoint selection (following common methodology, e.g., [18]).

2.6.2 SSIM (global and masked, 3D)

Structural similarity index measure (SSIM) [19] evaluates image similarity based on luminance, contrast, and structure components, making it more perceptually relevant than PSNR. It ranges from -1 to 1, with higher values indicating greater structural similarity.

The global SSIM is computed by averaging the SSIM map over all voxels in the volume. To reduce background influence and emphasize cranium fidelity, we compute a *masked* SSIM on test volumes inside a cranial mask.

Let X (predicted pCT) and Y (reference CT) be $[B, 1, D, H, W]$ tensors normalized to $[-1, 1]$, and let $M \in \{0, 1\}^{B\times1\times D\times H\times W}$ be the cranial mask, optionally dilated in-plane by 2 pixels to accommodate minor misregistrations. Using a separable 3D Gaussian window K (default size $11 \times 11 \times 11$ voxels, $\sigma = (1.5, 1.5, 1.5)$), local masked means, variances, and covariance are computed by convolution weighted by M, and pooled with mask weights to produce final masked SSIM as shown:

$$\mu_X = \frac{K * (X \odot M)}{K * M + \varepsilon}, \quad \mu_Y = \frac{K * (Y \odot M)}{K * M + \varepsilon}, \tag{2.4}$$

$$\sigma_X^2 = \frac{K * (X^2 \odot M)}{K * M + \varepsilon} - \mu_X^2, \quad \sigma_Y^2 = \frac{K * (Y^2 \odot M)}{K * M + \varepsilon} - \mu_Y^2, \tag{2.5}$$

$$\sigma_{XY} = \frac{K * (XY \odot M)}{K * M + \varepsilon} - \mu_X\,\mu_Y, \tag{2.6}$$

with a small ε for numerical stability. The local SSIM map is then

$$\mathrm{SSIM} = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}, \quad C_1 = (0.01L)^2, \quad C_2 = (0.03L)^2, \quad L = 2. \tag{2.7}$$

Final masked SSIM pools the local SSIM map weighted by the smoothed mask support

$$\mathrm{mSSIM}(X, Y; M) = \frac{\sum \mathrm{SSIM} \cdot (K * M)}{\sum K * M}, \tag{2.8}$$

averaged over the batch as detailed in Algorithm 1 (in Appendix A.3).

Implementation notes.

- The cranial mask is dilated by two pixels via max-pooling to stabilize the region of interest against registration noise.

- Global SSIM corresponds to setting $M \equiv 1$.
- Metrics are computed per volume (case) and averaged across the cohort.

2.6.3 Skull Dice (3D)

Let $\hat{S}$ and S denote the predicted and reference binary skull masks, respectively. Both skull and cranial masks used in this work are automatically generated using TotalSegmentator [15]. The Dice similarity coefficient is defined as

$$\mathrm{Dice}(\hat{S}, S) = \frac{2|\hat{S} \cap S|}{|\hat{S}| + |S|} = \frac{2\sum \hat{S}S}{\sum \hat{S} + \sum S}. \tag{2.9}$$

We report the mean Dice across test cases. This metric emphasizes the anatomical overlap of the skull, serving as a surrogate for gross geometric fidelity of the pseudo-CT relative to the reference CT.

2.6.4 Aggregation

For validation (PSNR), slice-wise scores are averaged per case, then across cases for per-epoch metrics. For test evaluation (SSIM and Dice), metrics are computed per full 3D volume and reported as cohort means (with standard deviations).

Software Unless otherwise stated, all metrics are implemented using MONAI [13] and scikit-image [20] to ensure reproducibility.

Chapter 3

Experiments

This chapter outlines the methodology and procedures used to rigorously evaluate the impact of normalized gradient field (NGF) regularization and spatial sampling strategies in supervised MR→CT synthesis using a Pix2Pix framework. Our experimental protocol was devised to isolate the effects of edge-awareness and resolution on anatomical fidelity, with all quantitative outcomes reported separately in Chapter 3.7. We detail the design of multiple training and validation arms, describe the technical implementation and training protocols, explain how models are selected and visualizations planned, and specify all compute environments and reproducibility safeguards.

3.1 Experimental design

The central aim of our experimental grid is to quantify the contribution of the normalized gradient field (NGF) loss—an edge-preserving regularization—within a supervised Pix2Pix MR→CT pipeline on paired data (section 1.7). All four experimental variants use identical dataloading and optimization protocols unless otherwise stated, enabling precise attribution of observed differences to the tested components:

- **Exp. 1 (NGF=20):** Baseline with NGF loss weight $\lambda_{\mathrm{NGF}} = 20$ throughout; tests maximum edge consistency regularization.
- **Exp. 2 (NGF=0):** Identical to Exp. 1 but with $\lambda_{\mathrm{NGF}} = 0$; quantifies effect of ablation of edge regularization.
- **Exp. 3 (switch-off):** First 40 epochs with NGF weight $= 20$ (shared initialization), then two parallel branches: one *keeps* NGF, one *drops* NGF; systematically measures impact of removing edge-consistency mid-training.
- **Exp. 4 (crop+mask+bg):** Higher-fidelity variant uses 256×256 center crops from native 512×512 volumes (instead of resizing), incorporates masked losses in the cranial ROI, and adds a background penalty driving air intensity (≈ -1) outside the mask; tests effect of higher-resolution focused supervision.

For all experiments, Exps. 1–3 preprocess by resizing in-plane to 256×256 before sampling 2D slices, while Exp. 4 uses center crops to better preserve high-frequency detail. Training uses 2D samples, validation is performed on fixed 2D center slices (one per patient), and final evaluation is conducted in 3D with sliding-window inference (section 1.7, Sections 2.5, 2.6).

3.2 Implementation details

Data I/O and transforms are managed using MONAI, utilizing `CacheDataset` or `SmartCacheDataset` for efficient local ram management as needed. On the local workstation, `SmartCacheDataset` was employed due to RAM constraints, with batch size $= 1$. On Google Colab, which allowed higher memory capacity, standard caching and batch size $= 14$ were used with multi-worker prefetch to optimize throughput.

Preprocessing exactly follows Chapter 1.7: DICOM to NIfTI conversion; MR/CT pairing via TCIA metadata; axis flips based on affine comparison (if needed); intensity scaling to $[-1, 1]$; custom 2D sampling strategies according to experiment arm; and spatial alignment.

The generator is a MONAI U-Net variant with $\tanh$ final activation; the discriminator is a multi-scale PatchGAN architecture. The normalized gradient field (NGF) regularization term is computed via spatial gradients (using Kornia) to enforce local orientation consistency, weighted by λ_{NGF}.

Optimization uses Adam (lr $= 2\times10^{-4}$). Unless otherwise specified, the generator loss is:

$$\mathcal{L}_G \;=\; \lambda_{\text{L1}}\,\|\hat{y} - y\|_1 \;+\; \lambda_{\text{NGF}}\,\mathcal{L}_{\text{NGF}}(\hat{y}, x) \;+\; \lambda_{\text{GAN}}\,\mathcal{L}_{\text{GAN}}(\hat{y}),$$

with $(\lambda_{\text{L1}}, \lambda_{\text{NGF}}, \lambda_{\text{GAN}}) = (140, 20 \text{ or } 0, 1)$ in Exps. 1–3. In Exp. 4, masked versions of $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{NGF}}$ are evaluated within the dilated head mask, and a background term drives air $\to -1$ outside the mask (full weighting and implementation details as described in Chapter 1.7).

3.3 Training protocols

Experiments 1 through 3 are trained for 70 epochs due to time and computational resource constraints. Experiment 4 training runs for a longer, variable number of epochs until convergence criteria or performance saturation warrant termination.

An initial two-epoch warm-up phase uses identity mapping (`real_B_is = SRC`) to stabilize early gradients and encourage structural preservation, after which the model is trained with the standard supervised MR→CT mapping.

For **Experiment 3**, training is initially shared for the first 40 epochs. At epoch 40, the model weights are cloned to create two parallel runs: one continuing with NGF regularization enabled ($\lambda_{\text{NGF}} = 20$), and the other with NGF disabled ($\lambda_{\text{NGF}} = 0$), enabling direct study of NGF's contribution mid-training.

Dataloading and sampling configurations per experiment are:

- **Experiments 1–3:** Resizing of 3D volumes to 256×256 in-plane, followed by sampling multiple 2D axial slices; batch size is set to 14.

- **Experiment 4:** Center-cropping of native 512×512 volumes to 256×256 patches without resizing, followed by 2D slice sampling; loss functions incorporate cranial mask awareness as described previously.

We used a batch size of 14 for all main training runs conducted on Google Colab Pro with sufficient GPU memory. Initial experiments on a local workstation with an RTX 4070 used batch size 1 due to memory limitations but were discontinued in favor of cloud training for efficiency.

3.4 Model selection (validation only)

Model selection is based on maximizing the **validation PSNR** metric computed on 2D center-cropped slices each epoch. The best-performing checkpoint corresponds to the epoch with highest average validation PSNR across the validation cohort. Note that PSNR is used solely as a *selection* criterion; final quantitative evaluation on the held-out test set employs more task-relevant metrics including global and masked SSIM as well as skull Dice (Chapter 3.7).

3.5 Visualization plan (qualitative only)

To qualitatively monitor learning progression and interpret model behavior, we implement the following visualization strategies:

- **Epoch evolution panels:** For each experiment, tri-panel visualizations juxtapose MR input, reference CT, and synthesized pseudo-CT slices at selected key epochs (e.g., epochs 10, 40, and 70) for the same validation slice. Regions of interest (ROI) zoom-ins highlight pertinent anatomical edges such as cranial bone and sinuses to assess structural fidelity.
- **Validation curves:** Per-epoch plots of validation PSNR metrics track performance evolution. For Experiment 3, PSNR curves from both training branches (NGF kept vs. dropped) are overlaid to visualize the regulatory effect of NGF. Numerical interpretation of these plots is deferred to Chapter 3.7.

3.6 Compute environment and reproducibility

Training was performed on two primary hardware environments: a local laptop equipped with an NVIDIA GeForce RTX 4070 HX GPU and 16 GB RAM, and the Google Colab Pro platform provisioned with an NVIDIA L4 GPU.

Due to limited GPU memory on the local RTX 4070 workstation, early experiments employed a batch size of 1, leveraging MONAI's `SmartCacheDataset` to mitigate memory and I/O bottlenecks. However, these runs were brief and limited by hardware constraints.

The majority of training, comprising all main experimental runs presented, was conducted on Google Colab Pro with batch size set to 14, allowing substantially faster training performance (approximately 8 minutes per epoch on Colab vs. 30 minutes locally).

Reproducibility was ensured by archiving all checkpoints, hyperparameters, and data transformation definitions. Test-time inference scripts incorporate volumetric resampling of pseudo-CT outputs onto CT reference grids to support voxelwise comparisons, as detailed in Chapter 1.7, Section 2.5.

3.7 Qualitative evolution and PSNR dynamics

To gain insights into the temporal stabilization and structural fidelity of the pseudo-CT (pCT) predictions throughout training, we present combined quantitative and qualitative visualizations. These highlight both global metric trends and subtle anatomical details that may not be fully captured by PSNR alone.

Validation PSNR comparison across experiments Figure 3.1 compares validation PSNR curves for Experiments 1, 2, and 4 on a single plot. Experiments 1 and 2 exhibit generally higher PSNR values compared to Experiment 4. This difference arises primarily because Experiment 4 incorporates a background penalty driving air voxels toward -1 (normalized scale), and applies masked losses focusing on cranial regions. PSNR's sensitivity to intensity variations in the background thus naturally yields lower global scores despite improved anatomical modeling.

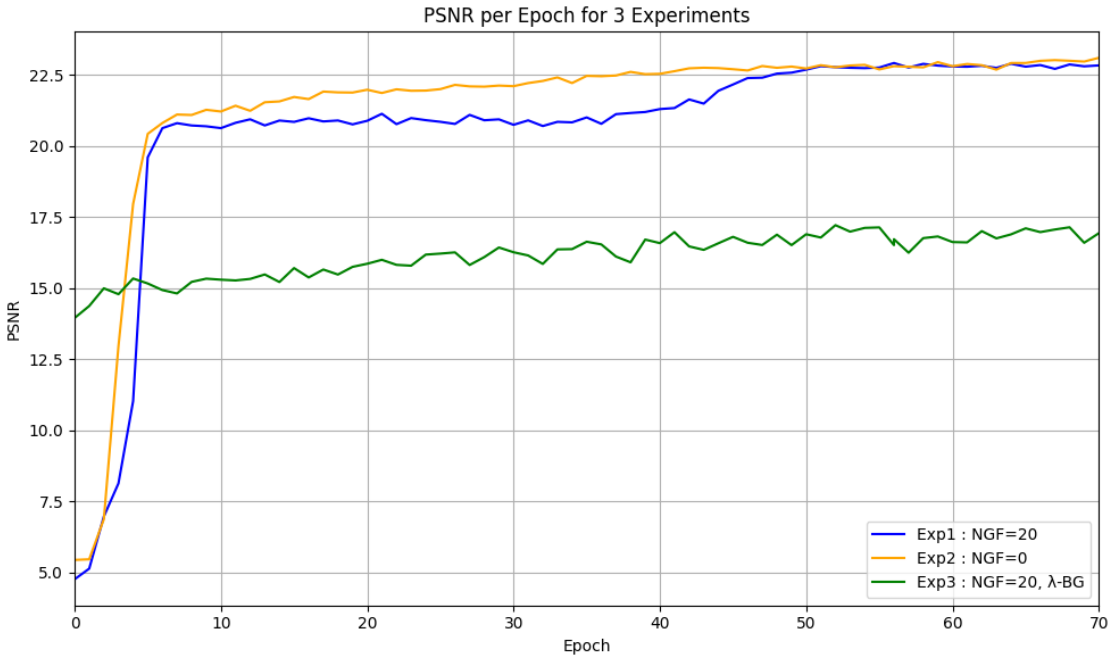
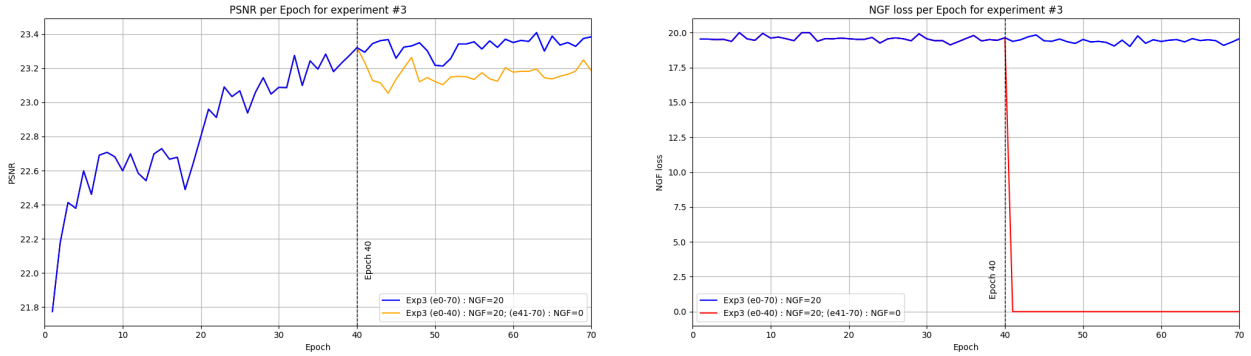


Figure 3.1 – Validation PSNR curves for Experiments 1, 2, and 4.

Experiment 4's lower PSNR results from a background penalty and masked loss focusing on cranial structures, making global intensity differences more pronounced. Despite similar PSNR, Experiments 1 and 2 differ structurally as shown in subsequent qualitative figures.

NGF switch-off study: PSNR and NGF loss dynamics Next, Figure 3.2 focuses on Experiment 3 where training branches at epoch 40 into NGF keep versus drop. The overlaid validation PSNR curves show the branch retaining NGF generally improves or maintains structural fidelity, whereas the NGF-dropped branch experiences instability and plateauing, as if restarting training. The accompanying NGF loss plot reveals the NGF-20 branch struggles to minimize this regularization loss, maintaining edge consistency as encoded in $\mathcal{L}_{\text{NGF}}$, while the NGF=0 branch's loss trivially zeros post-switch, explaining the divergent PSNR behaviors.



(a) Validation PSNR curves for NGF keep vs. drop (Exp. 3).

(b) NGF loss behavior during training (Exp. 3).

Figure 3.2 – Effect of NGF regularization toggling at epoch 40 on validation PSNR and NGF loss.

Switching off NGF causes PSNR instability and leads to a plateau, while NGF keep branch steadily improves. NGF loss reflects active enforcement in the keep branch and near zero value in the drop branch after epoch 40.

Qualitative evolution in Experiments 1 and 2 Figures 3.3 and 3.4 respectively illustrate qualitative evolution of pCT slices during validation for Experiments 1 (NGF=20) and 2 (NGF=0) across selected milestone epochs (e.g., epoch 10 and 70). Differences between epochs are subtle at full-volume scale, necessitating zoomed insets focusing on critical cranial bone and sinus edges. While both experiments exhibit reasonably high PSNR, Experiment 2 shows progressive loss in edge definition and structural fidelity — highlighting NGF's contribution beyond what global metrics capture.

Figures 3.5a, 3.5b, and 3.5c in Figure 3.5 present side-by-side comparisons of the best epoch pCT results from Experiments 1 and 2 alongside the corresponding reference CT slices. This comparison highlights the reduced edge sharpness and diminished structural fidelity in Experiment 2, despite similar PSNR scores, illustrating the contribution of the NGF loss beyond quantitative metrics.

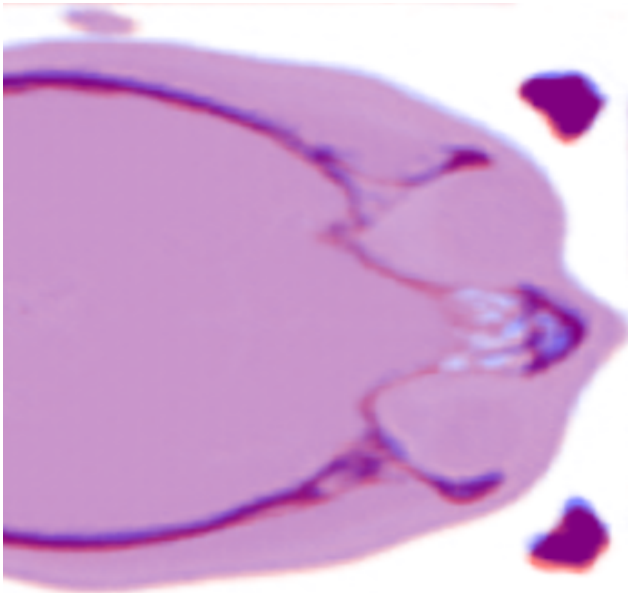


Figure 3.3 – Evolution of Exp 1 pseudo-CT predictions at epochs 10 (Blue) and 70 (Red) with zoom highlighting edge preservation.

NGF regularization in Exp 1 supports sharp edge preservation and structural fidelity through training.

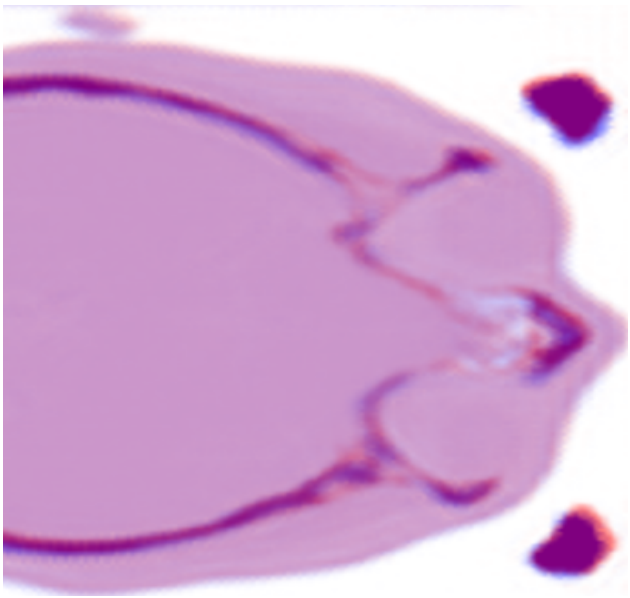
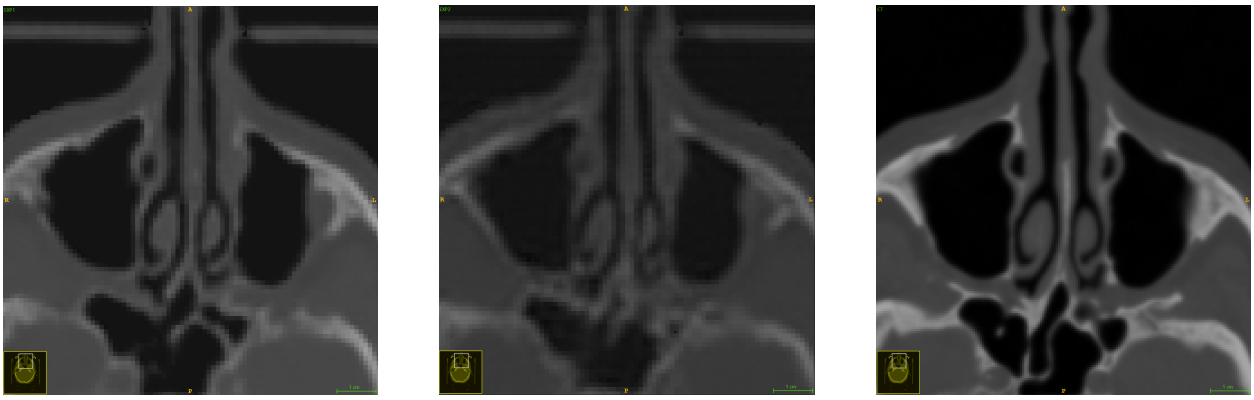


Figure 3.4 – Evolution of Exp 2 pseudo-CT predictions at epochs 10 and 70 with zoom showing gradual edge blurring.

Absence of NGF leads to progressive structural degradation and blurred edges despite similar PSNR.



(a) Exp 1 pseudo-CT (b) Exp 2 pseudo-CT (c) Reference CT

Figure 3.5 – Zoomed comparison of pseudo-CT outputs from Exp 1 and Exp 2 alongside the reference CT. Exp 2 shows diminished structural fidelity despite comparable PSNR scores.

Visual comparison reveals reduced edge sharpness and anatomical detail in Exp 2 compared to Exp 1.

(a) Exp 1 pseudo-CT (b) Exp 4 pseudo-CT (c) Reference CT

Figure 3.6 – Zoomed comparison of pseudo-CT outputs from Exp 1 and Exp 4 beside the reference CT. Exp 4 shows improved detail preservation due to cropping and masked losses.

Resolution and background handling effects: comparison of Experiments 1 and 4 Cropping and masked loss in Exp 4 preserves finer anatomical detail than resizing in Exp 1.

3D Inference visualization of background clarity in Experiment 4 Figure 3.7 shows a 3D volume rendering of the predicted pseudo-CT from Experiment 4, highlighting the cleaner background achieved by explicitly enforcing air intensities outside the cranial mask during training. This contrasts with Experiment 1's background, which typically exhibits noisier or less anatomically consistent air regions due to lack of explicit background penalty.

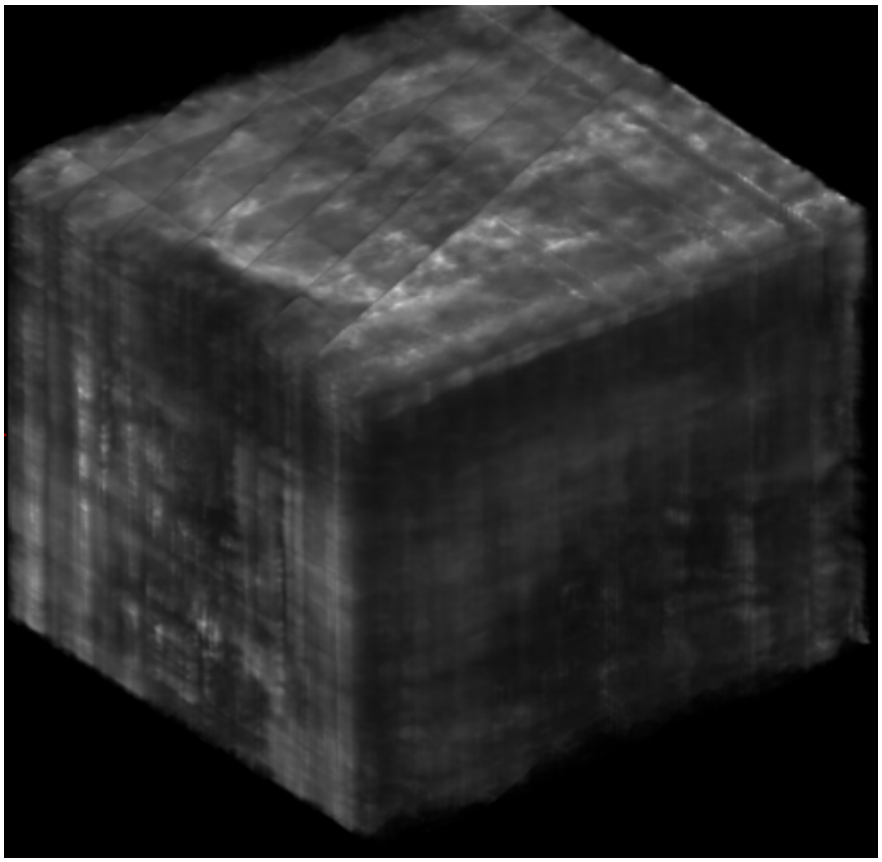


Figure 3.7 – 3D rendering of Exp 4 pseudo-CT volume highlighting clear background air intensities due to background penalty.

Explicit background penalty during training produces anatomically consistent air regions outside the cranium in 3D outputs.

Chapter 4

Results

4.1 Quantitative results

This section presents the quantitative evaluation of all experiments using global and local SSIM metrics, as well as Dice similarity coefficients for skull segmentation. The metrics were computed per-case and averaged across the test cohort to provide comprehensive performance insight.

Table 4.1 – Summary of mean global SSIM, mean local SSIM (within cranial region), and mean Dice scores across experiments.

Experiment	Mean Global SSIM	Mean Local SSIM	Mean Dice
Exp 1 (NGF=20)	0.7315	0.4131	0.8851
Exp 2 (NGF=0)	0.7134	0.4057	0.8826
Exp 4 (Crop + Mask + BG)	0.6521	0.6931	0.8660

Global SSIM Experiment 1 achieves the highest global SSIM, demonstrating superior overall structural similarity to reference CT images. The inclusion of normalized gradient field (NGF) loss appears to enhance edge preservation, as supported by the difference between Exp 1 and Exp 2. Exp 4 shows a marked reduction in global SSIM, which is expected because of the specialized masked loss and the background penalty that enforce air regions intensities, affecting global similarity metrics by amplifying background differences.

Local SSIM The local SSIM is calculated within the cranial region only, revealing another perspective on model performance. Here, Exp 4 outperforms Exp 1 and Exp 2, indicating its effectiveness in preserving structural details in the region of interest due to cropping and masked training losses. Exp 1 and Exp 2 show lower local SSIM values, reflecting their entire-image focus and absence of targeted masking.

Dice Similarity Coefficient Dice coefficients for skull segmentation mask comparison reflect the anatomical overlap quality. Exp 1 and Exp 2 show similar high Dice scores with a slight advantage

to Exp 1, consistent with the benefit of NGF regularization in enhancing spatial detail. Exp 4's Dice score is lower, which can be attributed to the increased strictness of boundary definitions under masked and background loss frameworks; despite higher local SSIM, the Dice metric's sensitivity to precise contour overlap may not fully capture qualitative improvements.

4.2 Qualitative results

Qualitative inspection confirms the quantitative insights:

- **Experiment 1** produces cleaner and sharper edges around cranial bones, consistent with its higher global SSIM and Dice scores, and demonstrates that NGF supports structural fidelity. - **Experiment 2**, while showing similar quantitative results overall, exhibits a subtle loss of edge sharpness and clarity in peripheral structures on visual inspection. - **Experiment 4**'s outputs exhibit markedly clearer separation of air and tissue regions in the background due to the explicit background penalty, with excellent detail retention within the cropped cranial ROI correlating with the highest local SSIM.

4.3 Discussion of Metrics

The results emphasize the importance of selecting appropriate metrics aligned with experimental goals and loss formulations. While global SSIM and Dice are widely used, they may underrepresent performance in specialized training approaches focusing on masked regions and background penalties. Local SSIM, sensitive to the targeted cranial region, provides more nuanced feedback, especially for approaches like Exp 4.

Chapter 5

Discussion and Perspectives

5.1 Summary of findings

Our experiments demonstrate that the normalized gradient field (NGF) regularization significantly enhances structural fidelity and boundary preservation in pseudo-CT synthesis, as evidenced by higher global SSIM and Dice metrics for experiments employing NGF (Exp 1) compared to those without (Exp 2). Furthermore, Experiment 4 highlights the clinical importance of maintaining native resolution and explicitly modeling background air regions to achieve clearer segmentations and better anatomical clarity, despite some global metric trade-offs.

Maintaining the highest possible spatial resolution and accurate edge modeling is vital for clinical applications, such as radiotherapy planning or dose calculation, where subtle anatomical details can influence treatment efficacy and safety.

5.2 Limitations

Despite promising results, several limitations must be acknowledged:

- **Computational constraints:** Limited GPU memory required compromises like smaller batch sizes and truncated training epochs, particularly on the local workstation, potentially affecting generalizability and convergence.
- **Residual background ambiguity:** The "fog" effect observed in Experiment 4's early 3D outputs indicates incomplete background suppression during the training performed so far.
- **Metric sensitivity:** Commonly used global metrics such as SSIM and Dice imperfectly capture localization errors or specialized masked training improvements, presenting interpretation challenges.
- **Dataset size and diversity:** The experiments rely on a single dataset with limited anatomical variety, which may limit robustness to varied clinical populations or imaging conditions.

Future work addressing these limitations through expanded resources, longer training, more sophisticated architectures, and diverse datasets is essential for clinical translation.

5.3 Interpretation of 3D results in Experiment 4

The volumetric pseudo-CT reconstructions from Experiment 4 exhibit a characteristic "foggy" appearance surrounding the head during early to mid-training stages. Although the ultimate goal of the background penalty is to enforce a clear air region (intensity close to -1000 HU, normalized to -1) around the cranial mask, complete suppression of residual artifacts requires extended training epochs beyond those currently completed.

Interestingly, this residual fog is spatially separated from the head by a distinct void region, creating a useful boundary that enhances anatomical delineation and segmentation feasibility. Compared to experiments lacking explicit background regularization, where structures and background often bleed or directly contact, this separation may provide a significant practical advantage for downstream segmentation tasks by reducing ambiguity near the cranial contour.

For a visual example showing the clear void between the head and surrounding fog, see Appendix Figure A.6, where this phenomenon is evident in the Exp4 pseudo-CT output.

Future training refinements and longer-term convergence are expected to further reduce this fog, sharpening the separation and resulting in cleaner volumetric predictions suitable for clinical planning and analysis.

5.4 Future perspectives and model improvements

Building on these findings, several avenues may enhance model performance and clinical applicability:

- **Attention mechanisms:** Integrating attention modules could enable the model to focus selectively on the skull and critical anatomical structures, potentially improving resolution and structural fidelity. Coupling spatial and channel attention might assist in concentrating the learning capacity on regions clinically relevant for radiotherapy or diagnosis.
- **Skull-only mask guidance:** Feeding the model auxiliary information such as binary skull masks during training could provide explicit structural priors, aiding in boundary preservation and noise reduction, especially near complex bony interfaces.
- **Exploration of novel architectures:** Investigating recently proposed architectures with demonstrated prowess in medical image synthesis—such as transformer-based models or hybrid CNN-transformer hybrids—may yield improved image quality and generalization.
- **Leveraging embedding techniques:** Incorporating learned embeddings from pre-trained networks or leveraging generative priors might enhance anatomical consistency and reduce artifacts, especially in low-contrast or ambiguous regions.

Conclusion

The journey of developing accurate and reliable pseudo-CT generation models reveals the exciting yet demanding nature of AI research in healthcare. As a final thought, aspiring researchers and practitioners venturing into this field should be prepared not only for methodological challenges but also for substantial computational resource demands. For example, the frequent refrain from this work's author has been: "Want to work with AI nowadays? Better be rich, because it really eats up resources!" While said in jest, this underscores the practical realities of training state-of-the-art models requiring powerful GPUs and ample memory far beyond many personal workstations.

Nevertheless, persistent innovation, resourcefulness, and community collaboration continue to turn these challenges into opportunities, driving us closer to impactful clinical applications and transformative healthcare technologies.

Bibliography

[1] "Latim — laboratoire de traitement de l'information médicale." https://latim.univ-brest.fr/. Accessed 2025-09-06.

[2] "Team action — latim." https://latim.univ-brest.fr/index.php/action. Accessed 2025-09-06.

[3] "Team imagine — latim." https://latim.univ-brest.fr/index.php/imagine. Accessed 2025-09-06.

[4] V. Bourbonne, V. Jaouen, C. Hognon, N. Boussion, F. Lucia, O. Pradier, J. Bert, D. Visvikis, and U. Schick, "Dosimetric validation of a GAN-based pseudo-CT generation for MRI-only stereotactic brain radiotherapy," *Cancers*, vol. 13, 2021.

[5] J. Grigo, J. Szkitsak, D. Höfler, R. Fietkau, F. Putz, and C. Bert, "A feasibility study for deep learning-based mri-only brain radiotherapy," *Frontiers in Oncology*, vol. 14, 2024.

[6] F. Putz, G. Heilemann, C. Belka, and et al., "Magnetic resonance imaging for brain stereotactic radiotherapy: A review of requirements and pitfalls," *Strahlentherapie und Onkologie*, vol. 196, no. 5, pp. 437–450, 2020.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[9] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *ECCV*, 2020.

[10] V. Jaouen and coauthors, "Normalized edge consistency for one-sided unsupervised image-to-image translation," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024. Normalized gradient field-inspired constraint for structural fidelity.

[11] E. Porter, P. Fuentes, I. Sala, Z. Siddiqui, R. Levitin, N. Myziuk, B. Squires, T. Gonzalez, P. Chen, T. Guerrero, and I. Grills, "Gamma knife mr/ct/rtstruct sets with hippocampal contours (gammaknife-hippocampal), version 1 [data set]," 2022.

[12] National Institutes of Health, "Nih controlled data access policy." https://www.cancerimagingarchive.net/nih-controlled-data-access-policy/. Accessed: 2025-09-06.

[13] The MONAI Consortium, "MONAI: Medical open network for AI." https://monai.io, 2020. Open-source framework for deep learning in healthcare.

[14] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: an open source differentiable computer vision library for pytorch," *arXiv preprint arXiv:1910.02190*, 2020.

[15] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "Totalsegmentator: Robust segmentation of 104 anatomical structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Springer, 2015.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[18] A. K. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition (ICPR)*, pp. 2366–2369, IEEE, 2010.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 2014.

Appendix A

Additional Material

A.1 Example Medical Images

This section presents representative examples of the input and output medical images used in this study. Visual inspection aids in understanding the task complexity and the model's performance qualitatively.

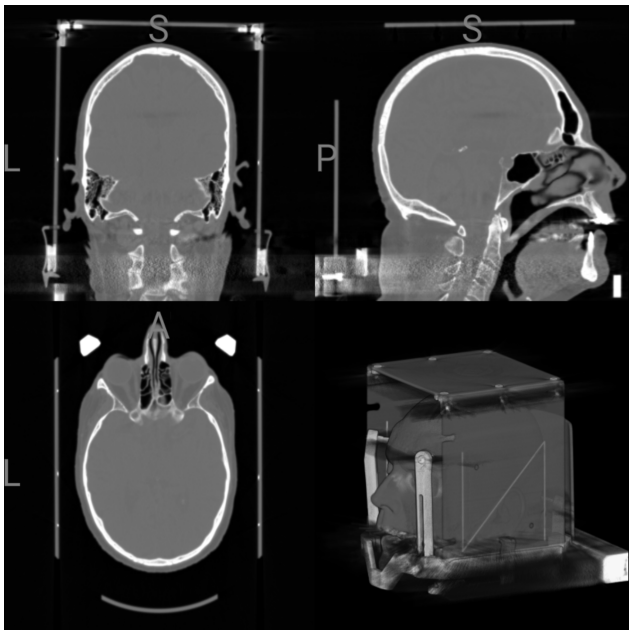


Figure A.1 – Example of a CT from the dataset.

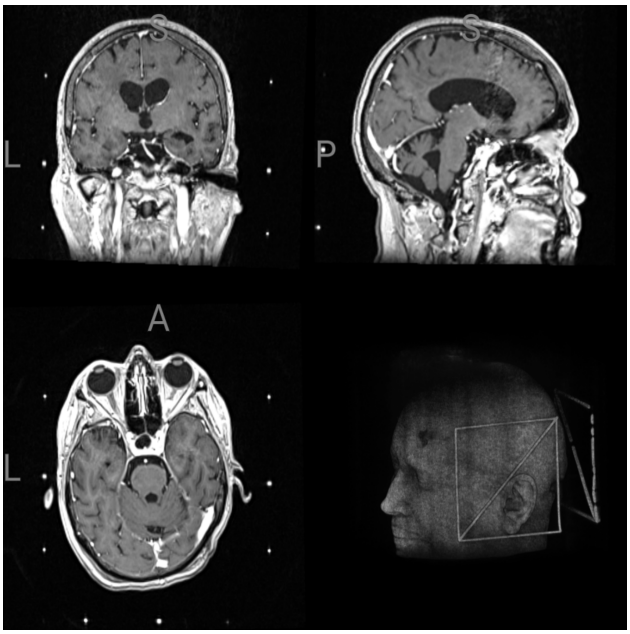


Figure A.2 – Example of an MR from the dataset.

Figure A.3 – Example of a segmented CT head used as ground truth.

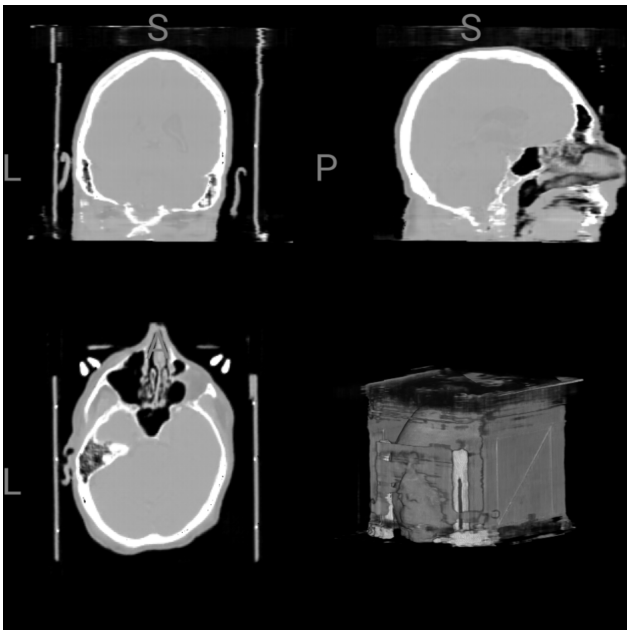


Figure A.4 – Example pseudo-CT output from Experiment 1 (NGF = 20).

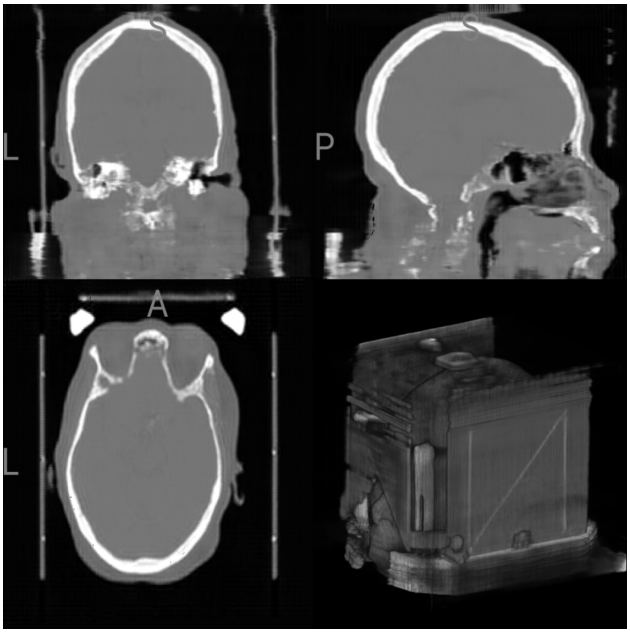


Figure A.5 – Example pseudo-CT output from Experiment 2 (NGF = 0).

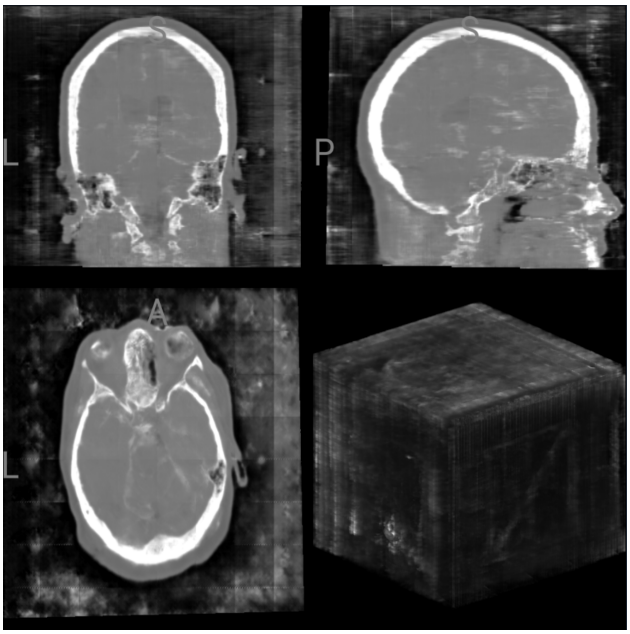


Figure A.6 – Example pseudo-CT output from Experiment 4 (cropping and masked losses).

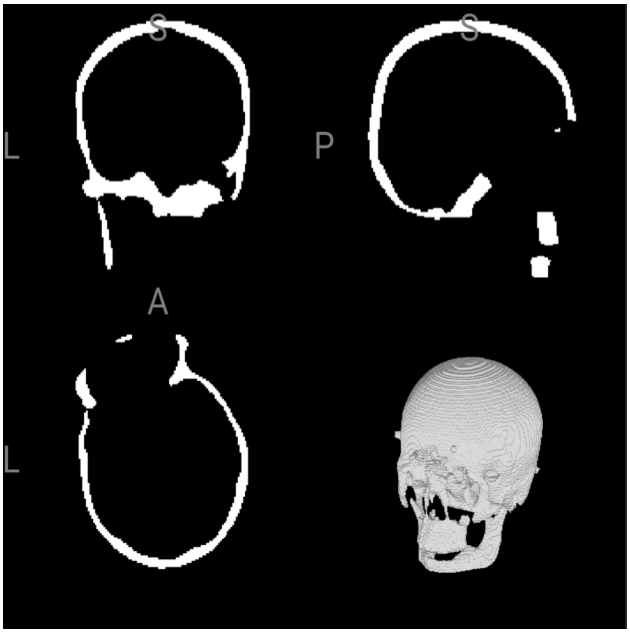


Figure A.7 – Example skull mask derived from CT images.

A.2 A Beginner's Guide to AI in Medical Imaging

Artificial Intelligence (AI) has revolutionized many fields, and medical imaging is no exception. This section offers an accessible explanation of AI concepts, from fundamental ideas to recent advances, suitable for readers with varying backgrounds.

What is AI?

At its core, AI aims to enable machines to perform tasks that typically require human intelligence. These tasks include recognizing patterns, making decisions, and learning from data. In medical imaging, AI helps automate analysis, improve diagnostic accuracy, and assist clinical workflows.

From Simple Models to Deep Learning

Early AI models used hand-crafted rules and simple statistical techniques. However, these were limited in handling complex data like medical images.

Deep learning, a subset of AI, uses artificial neural networks inspired by the human brain. These networks consist of layers of interconnected nodes (neurons) that can learn hierarchical features directly from raw data, bypassing the need for manual feature engineering.

What are Convolutional Neural Networks (CNNs)?

CNNs are specialized neural networks designed for image data. They use filters to scan input images and automatically detect edges, textures, shapes, and higher-level features. This makes CNNs especially effective for medical images, where subtle structural details matter.

Generative Adversarial Networks (GANs)

GANs consist of two neural networks—a generator that creates synthetic images, and a discriminator that distinguishes real from synthetic images. These networks are trained together in a min-max game, where the generator learns to produce increasingly realistic images to fool the discriminator.

In medical imaging, GANs can synthesize missing modalities (e.g., pseudo-CT from MR), enhance image quality, or augment scarce datasets.

Beyond CNNs and GANs

Recent advances include transformer architectures, which excel at capturing long-range relationships in data, and large language models (LLMs), known for natural language understanding but increasingly applied to multimodal medical data.

Hybrid models combining CNNs and transformers are being explored for improved accuracy and adaptability.

Practical Considerations

AI models require substantial labeled data, computational power, and careful validation. Transparency, interpretability, and compliance with medical standards are crucial for clinical deployment.

Summary

AI in medical imaging is a rapidly evolving field, blending foundational ideas with cutting-edge research. Understanding the basics helps clinicians and researchers harness AI's potential to improve patient care.

A.3 Algorithm for Masked SSIM Computation

Algorithm 1: Detailed masked SSIM in 3D with Gaussian window and ROI weighting

Input : $\mathbf{X}$ (pred), $\mathbf{Y}$ (gt), $\mathbf{M}$ (mask) $\in \mathbb{R}^{B\times 1\times D\times H\times W}$; window $w = (k_d, k_h, k_w)$; $\sigma = (s_z, s_y, s_x)$; data_range; ε

Output: $\mathrm{SSIM}_{\mathrm{masked}} \in \mathbb{R}$

Require: Shapes match: $\mathrm{shape}(\mathbf{X}) = \mathrm{shape}(\mathbf{Y}) = \mathrm{shape}(\mathbf{M})$. If inputs are 4D, unsqueeze channel to get $[B, 1, D, H, W]$.

Ensure odd window sizes and clamp to volume dims:

$k_d \leftarrow \max(3,\ \min(k_d, D)); \quad k_d \leftarrow k_d - \mathbb{1}_{k_d \text{ even}}$

$k_h \leftarrow \max(3,\ \min(k_h, H)); \quad k_h \leftarrow k_h - \mathbb{1}_{k_h \text{ even}}$

$k_w \leftarrow \max(3,\ \min(k_w, W)); \quad k_w \leftarrow k_w - \mathbb{1}_{k_w \text{ even}}$

Build separable Gaussian kernel $K \in \mathbb{R}^{1\times 1\times k_d\times k_h\times k_w}$:

$g_x[i] \propto \exp\big(-\frac{(i-\frac{k_w-1}{2})^2}{2s_x^2}\big)$; normalize $\sum_i g_x[i] = 1$;

$g_y[j] \propto \exp\big(-\frac{(j-\frac{k_h-1}{2})^2}{2s_y^2}\big)$; normalize;

$g_z[k] \propto \exp\big(-\frac{(k-\frac{k_d-1}{2})^2}{2s_z^2}\big)$; normalize;

$K \leftarrow g_z \otimes g_y \otimes g_x$ (outer products), then normalize $\sum K = 1$.

$\mathrm{pad} \leftarrow \big(\lfloor k_w/2 \rfloor,\ \lfloor k_h/2 \rfloor,\ \lfloor k_d/2 \rfloor\big)$

Windowed mask support (weights): $\mathbf{W}_m \leftarrow \mathrm{Conv3D}(\mathbf{M}, K, \mathrm{pad}) + \varepsilon$

Masked local means:

$$\boldsymbol{\mu}_X \leftarrow \frac{\mathrm{Conv3D}(\mathbf{X}\odot\mathbf{M}, K, \mathrm{pad})}{\mathbf{W}_m}, \quad \boldsymbol{\mu}_Y \leftarrow \frac{\mathrm{Conv3D}(\mathbf{Y}\odot\mathbf{M}, K, \mathrm{pad})}{\mathbf{W}_m}$$

Masked local variances and covariance:

$$\sigma_X^2 \leftarrow \frac{\mathrm{Conv3D}(\mathbf{X}^2\odot\mathbf{M}, K, \mathrm{pad})}{\mathbf{W}_m} - \boldsymbol{\mu}_X^2$$

$$\sigma_Y^2 \leftarrow \frac{\mathrm{Conv3D}(\mathbf{Y}^2\odot\mathbf{M}, K, \mathrm{pad})}{\mathbf{W}_m} - \boldsymbol{\mu}_Y^2$$

$$\sigma_{XY} \leftarrow \frac{\mathrm{Conv3D}(\mathbf{X}\odot\mathbf{Y}\odot\mathbf{M}, K, \mathrm{pad})}{\mathbf{W}_m} - \boldsymbol{\mu}_X\odot\boldsymbol{\mu}_Y$$

Stabilizers (SSIM constants): $C_1 \leftarrow (0.01\cdot \mathrm{data\_range})^2, \quad C_2 \leftarrow (0.03\cdot \mathrm{data\_range})^2$

Voxelwise masked SSIM map:

$$\mathbf{S} \leftarrow \frac{(2\boldsymbol{\mu}_X\odot\boldsymbol{\mu}_Y + C_1)\,(2\sigma_{XY} + C_2)}{(\boldsymbol{\mu}_X^2 + \boldsymbol{\mu}_Y^2 + C_1)\,(\sigma_X^2 + \sigma_Y^2 + C_2) + \varepsilon}$$

Weighted pooling over ROI (per batch element):

$s_b \leftarrow \frac{\langle \mathbf{S}_b,\ \mathbf{W}_{m,b}\rangle}{\langle \mathbf{1},\ \mathbf{W}_{m,b}\rangle}$ for $b = 1..B$ `// Inner products over` $D\times H\times W$

return $\frac{1}{B}\sum_{b=1}^{B} s_b$