Master Thesis

Predicting Tumor Hypoxia Map from FDG-PET/CT Images using GANs

C.S. Rao

Master Thesis DKE-21-24

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Artificial Intelligence
at the Department of Data Science and Knowledge Engineering
of the Maastricht University

Thesis Committee:

Dr. A. Briassouli
Dr. E. Hortal Quesada

Maastricht University
Faculty of Science and Engineering
Department of Data Science and Knowledge Engineering

July 9, 2021

Abstract

Tumor hypoxia is characterized by an insufficient oxygen concentration in certain localized regions of a tumor. Hypoxic cancer cells develop higher therapy resistance and greater migratory capability, thereby reducing the effectiveness of radiotherapy treatment. Positron Emission Tomography (PET) imaging has recently become a major focus of research for its use as an *in vivo* hypoxia detection tool. Hypoxia PET provides an oxygen concentration map that is overlaid on a Computed Tomography (CT) scan to enhance contrast around hypoxic regions, thereby enabling their localization. It can, therefore, be integrated into radiotherapy workflow to account for tumor hypoxia and adjust the radiation dosage accordingly. HX4-PET is an instance of hypoxia PET imaging that uses the newly developed radiotracer [^{18}F]HX4. Despite its great potential in improving cancer treatment, HX4-PET imaging is expensive and is currently only limited to clinical trials. This work investigates a GAN-based computational alternative to hypoxia imaging. We apply image translation GANs to synthesize whole HX4-PET images from the more readily available FDG-PET and CT modalities, exploring both paired and unpaired translation approaches. Using the paired Pix2Pix as a reference method, unpaired translation with CycleGAN is studied and compared. We argue that naively applying the default CycleGAN system to our use case is a flawed strategy because the invertibility assumption of CycleGAN is seriously violated here, and propose a design modification to CycleGAN training for circumventing this issue and optimizing the system for our purpose. We perform an extensive evaluation of the GANs by first testing on a simulated translation task, followed by comprehensively evaluating on an appropriate 3D medical image dataset. We, additionally, perform a set of clinically relevant downstream tasks on the synthetic HX4-PET images to determine their clinical value. Our experiments show that the modified CycleGAN attains high image-level performance, close to that of Pix2Pix, and although our synthetic HX4-PET images may not yet meet the clinical standard, the results suggest that unpaired translation approaches could be more suitable for the task due to their immunity to noise induced in the training data by spatial misalignments.

Contents

Chapter 1

Introduction

1.1 Problem Statement

Tumor hypoxia is a condition wherein certain regions within the tumor volume are deprived of oxygen, and is caused when the tumor's blood supply system cannot keep up with the proliferation rate of its cells [1]. Hypoxic cells develop certain alterations in their functioning mechanisms that make them resistant to conventional modes of cancer treatment, including radiotherapy and chemotherapy, and increase metastatic spread [2], thereby resulting in poor prognosis [3]. In radiotherapy, it would be of great value to detect these hypoxic sub-volumes during treatment planning so that the radiation dose can be adjusted accordingly to be more effective, for example, by selectively delivering higher doses to the hypoxic regions – a technique known as *radiotherapy boosting*. Among various non-invasive methods investigated for hypoxia detection, Positron Emission Tomography (PET) based hypoxia imaging has been the most preferred one [4]. PET is a molecular imaging technique that involves the administration of a specific radioactive pharmaceutical contrast agent, known as *radiotracer*, to the patient, followed by tracking it via scanning in order to provide information about the status of a particular biochemical process. Hypoxia PET imaging provides, essentially, an oxygen concentration heatmap that can be overlaid on a Computer Tomography (CT) scan to provide contrast around the anatomical regions that are hypoxic. A newly developed hypoxia tracer denoted as [^{18}F]HX4 has been shown to be highly stable and effective for detecting regions of tumor hypoxia [5], and the PET image acquired using this tracer is denoted as HX4-PET. HX4-PET imaging technique, however, is not widely available, is expensive, and has only been used in research and clinical trials so far. It would be of great benefit both to patients and to the care provider if entire HX4-PET images could be synthesized from more readily available imaging modalities. FDG-PET, which indicates metabolism levels, and CT modalities have been shown to capture features capable of reliably predicting hypoxic regions in the tumor [6, 7] which could enable the use of computational (machine-learning)

models to potentially substitute HX4-PET acquisition.

Generative adversarial networks (GANs) have recently become a subject of investigation as a general solution to image-to-image translation problems. The Pix2pix and CycleGAN frameworks [8, 9] are among the earliest of the systems developed in this direction targeting paired and unpaired variants of the problem, respectively. While paired translation tasks supply spatially aligned pairs of input and target images in the training data to directly learn an input-target mapping, unpaired translation tasks do not provide paired data usually because it is either impossible, unrealistically difficult, or unacceptably risky to obtain such data. In the space of medical imaging, GAN-based techniques have been widely explored for translating images of one modality into another [10]. However, since medical images are quantitative in nature that represent some underlying physical or biochemical properties of the human body, and since these images are used for clinical decision making, the GAN-generated synthetic images must be representative of the patient's condition and must conform to high quality standards. Evaluation of synthetic medical images is a difficult problem and is, in many cases, application-specific.

The problems tackled by this thesis project are two-fold – (1) Synthesis of full HX4-PET images from FDG-PET and CT scans of the patients using image translation GANs, and (2) Evaluation of different aspects of the synthetic HX4-PET images using general image quality metrics and via application-specific image analysis.

1.2 Significance of the Thesis

This work formulates the hypoxia prediction problem as an image-to-image translation task where entire 3D HX4-PET images are to be synthesized from FDG-PET and CT. We investigate both the paired and unpaired GAN approaches, specifically Pix2Pix and CycleGAN, as potential solutions to this problem. CycleGAN system utilizes cycle-consistency to learn an image translation model in the absence of paired training data by simultaneously learning an input-target mapping as well as its inverse. However, since the input in our use case is multimodal, applying CycleGAN naively to this problem may not be optimal. CycleGAN assumes the existence of a one-to-one mapping between the input and target domains, which is violated here as it is practically impossible to precisely reconstruct both FDG-PET and CT back from a single HX4-PET. Therefore, in order to sidestep this issue, a design improvement in the CycleGAN training is proposed for this specific use case. To our best knowledge, ours is the first work that investigates GAN-based cross-modality medical image translation for the hypoxia PET synthesis application.

To address the problem of evaluating synthetic HX4-PET images, we use six different metrics that measure different properties of an image, including voxel-wise difference, fidelity of the perceived structure and global intensity statistics. Since each metric has its unique strengths and drawbacks, we aim at using metrics of different types to complement each other and reach a consensus on the

models' assessment. Based on these metrics, we empirically show that the modified CycleGAN achieved remarkable improvement over the default CycleGAN, and we simultaneously validate the assessment of these metrics by performing a systematic visual inspection of the 3D synthetic images. Additionally, we investigate the applicability of these image quality metrics in tracking the stability and convergence of the CycleGAN models during training.

Finally, we perform a clinically relevant analysis of the synthetic HX4-PET images to determine their clinical value. Application-specific downstream tasks are identified from hypoxia imaging literature and applied to quantify hypoxia locally within the tumors. This is of great importance and of more interest to clinicians as a benchmark to gauge the potential of the models as an alternative solution to HX4-PET imaging to be integrated into radiotherapy practice in future. Although our quantitative results on tumor hypoxia quantification indicate no significant performance difference among the different methods, our qualitative analysis highlights the drawbacks of paired GAN approaches and suggests that unpaired methods might be more suitable for the task, given sufficient data.

1.3 Research Questions

The research questions posed initially at the time of drafting the thesis plan were slightly updated, without changing their overall theme, later when the work on the project actually began. The primary reason is that the exact specifications of the project were not provided earlier, rather only a vague formulation of the problem was given – synthesis of Contrast-Enhanced Computed Tomography (CECT) images from CT using GANs. "Contrast-enhanced" CT is an umbrella term that includes a variety of techniques used to selectively provide contrast enhancement to certain regions in a CT scan, ranging from using iodine-based contrast agents that improve visibility of organs like liver and pancreas by raising the radiodensity of blood to acquiring PET scans that utilize radiotracers to highlight sites of certain physiological processes like metabolism and hypoxia. Therefore, one of the main changes in the research questions was replacing the term "CECT" with "HX4-PET". Reasons for other question-specific changes are provided further in 1.3.2.

1.3.1 Original

1. What measures need to be considered during evaluation to more accurately quantify the similarity of synthetic CECT images with clinically acquired ones in an unpaired image-to-image translation scenario?

2. How reliable are deep-learning based downstream analysis approaches, such as image classification or segmentation, for evaluating synthetic CECT images?

3. How can such application-specific validation methods be applied to assess the stability of CycleGAN training for this task and detect issues like mode collapse?

1.3.2 Updated

1. What measures need to be considered during evaluation to accurately quantify the similarity of synthetic HX4-PET images with the clinically acquired scans?

2. How can application-specific downstream image analysis aid in evaluating the utility of synthetic HX4-PET images for tumor hypoxia measurement?

3. To what extent can image quality metrics provide information on convergence of the GAN models during training?

All question-specific changes were made to make the questions more sensible and meaningful. In Question 1, the phrase "in an unpaired image-to-image translation scenario" was removed because in the medical image dataset used in this project, the ground-truth is available, especially for the validation images, and the evaluation can be performed by merely comparing the predictions with the ground-truth, whether or not the training was performed in an unpaired manner. In Question 2, the phrase "deep-learning based downstream analysis" was replaced with "application-specific downstream analysis" because simpler thresholding based hypoxia classification and segmentation techniques are sufficient for image analysis. Since PET images are quantitative in nature, standardized intensity thresholds are already established for the downstream tasks and have been used in previous HX4-PET related clinical studies [11, 6, 5]. Using deep-learning to perform these downstream tasks would not only be unnecessary and less reproducible, but also infeasible due to the extremely small size of our dataset. This question now focuses on the tumor region and the value that synthetic HX4-PET image can provide in a clinical scenario in predicting hypoxia patterns within this region, instead of focusing on the overall quality and perfectness of the full image. This is reasonable because the tumor region is already delineated prior to HX4-PET acquisition. In Question 3, the phrase "application-specific validation methods" was replaced with "image quality metrics" because otherwise, the question would be invalid. This reason being that "application-specific methods" implies hypoxia analysis methods (which is addressed in Question 2), which are not suitable for measuring the overall quality of the full synthetic HX4-PET images since the tumor size is much smaller than the size of the full image. To decide whether or not a GAN model is converging during training, one needs to check the overall quality of the its generated samples. Therefore, this question can now be viewed as an extension of Question 1 and as being complementary to Question 2.

Chapter 2

Related Work

2.1 Predicting Hypoxia from FDG-PET and CT

In a hypoxia PET image, the tumor is the sole object of interest whose exact location and shape is usually known before the hypoxia image acquisition. Zegers et al. [12] show the presence of significant correlation among tumor-level parameters – tumor size, overall tumor metabolism (indicated by FDG-PET uptake) and overall tumor hypoxia (indicated by HX4-PET uptake). They additionally perform sub-volume level (i.e. voxel-wise) analysis between FDG-PET and HX4-PET to evaluate the spatial similarity of both tracers' uptake patterns within the tumor, and observe a reasonable correlation between them. However, they also note that this correlation is not significant likely due to the involvement of genetic properties of the tumor, and conclude that HX4-PET imaging does indeed provide information complementary to FDG-PET. Even et al. [6] investigate the possibility of predicting hypoxia patterns in the tumor using other indirect markers of hypoxia including anatomy (CT), metabolism (FDG-PET), and blood perfusion parameters (tumor blood flow and blood volume maps obtained from Dynamic CT). They argue that a voxel-wise regression between the inputs and the HX4-PET ground truth would be negatively affected by registration-related imperfections, and instead use a supervoxel based approach for robustness. They train random-forest-based regression models on simple features – median, standard deviation and entropy – derived from the supervoxels in each input image modality to infer the level of hypoxia in the corresponding supervoxel regions. The models were shown to reliably predict the spatial distribution of tumor hypoxia, and the features from CT and FDG-PET were found to be the most informative. Sanduleanu et al. [7] conduct a large-scale radiomic study using data from six different medical centers. Radiomic features comprise a large set of standardized quantitative features extracted from medical images developed with the aim of supporting high-throughput automated image analysis in radiation oncology [13]. The authors hypothesize that the combined radiomic features derived from both CT and FDG-PET modalities

can predict tumor hypoxia status more effectively compared to either of these modalities alone, and build and validate multiple random-forest-based tumor classification models. They conclude with the finding that the radiomic models using both modalities did indeed classify the tumors as hypoxic or non-hypoxic with high accuracy.

In this work, we utilize CT and FDG-PET information and aim at predicting hypoxia patterns in the entire body region covered by the input images, instead of focusing on just the tumor locality. In order to synthesize full HX4-PET images accurately, the HX4 tracer uptake in the human body must be effectively modeled as a function of CT-derived anatomical features and FDG tracer uptake. Then, inferring the HX4 uptake within the tumor becomes a special case of inferring it at any given region in the body, and this special case is characterized by the presence of abnormal structures in CT and an immense amount of activity in FDG-PET.

2.2 Image-to-Image Translation GANs

Generative adversarial networks (GANs) [14] are a class of deep generative models comprised of two networks – the generator and the discriminator – that are jointly trained in a competitive setting. The generator's task is to synthesize samples (often images) that are indistinguishable from samples from the true data distribution, whereas the task of the discriminator is to accurately distinguish between real and synthetic samples. Over the training period, the generator learns to produce increasingly realistic samples in response to the discriminator's feedback, which itself also improves. The model eventually converges to an equilibrium state where the generator ideally models the real data distribution, and as a result, the best the discriminator can do is random guessing. While the user has no control over the semantics of the images generated by the GAN model, the development of Conditional GAN (cGAN) [15] has allowed using simple semantic information, such as class labels, to train a conditional generative model. During inference, this enables the user to query the generator to synthesize images belonging to a given class.

Based on cGAN, Isola et al. [8] introduce the Pix2Pix system as a general-purpose framework for solving image-to-image translation problems. Any computer vision or graphics task that takes as input an image and expects another image as its output having the same size and containing the same semantic information can be formulated as image-to-image translation. For example, semantic segmentation of photographs, synthesis of photographs from semantic label map or edge map, and colorization of monochrome photos. In such problems, the input and the output images are viewed as different renderings of the same underlying scene. The Pix2Pix generator, conditioned on the entire input image, produces an output image that is directly compared to the ground truth via a pixel-wise L1 loss. Additionally, as a consequence of the adversarial training, the Pix2Pix discriminator model serves as a "structured" loss function that is learned from the data itself. This essentially rules out the need

to handcraft custom loss functions for each image translation task. Pix2Pix, however, requires pixel-wise aligned input-target image pairs, i.e. *paired* data, for training. A classic example task where paired data doesn't exist is horse-to-zebra translation [9]. Based on ideas similar to Pix2Pix, Zhu et al. present the CycleGAN system [9] for learning an image translation model from unpaired data. CycleGAN consists of four deep neural networks, which compose two generator-discriminator pairs. Given two sets of images, each containing images of one *domain*, the CycleGAN model learns a forward mapping between the input and the target domains as well as its inverse. This pair of mappings is used to implement a cycle-consistency mechanism to encourage the preservation of high-level structure (i.e. the "content") across an input image and its translated counterpart in the target domain. Since there is no direct way to ensure content preservation between input and output images due to the absence of paired data, this strategy serves as an indirect means to achieve this.

The Pix2Pix and CycleGAN systems have been widely applied in medical imaging research to various cross-modality image translation use cases [10] with slight custom modifications in network architecture and loss functions, although we limit our discussion to works that focus on synthesizing contrast maps from CT images. Chandrashekar et al. [16] show that sufficient contrast exists between blood and soft tissue to differentiate them in CT angiograms acquired without an iodine-based contrast agent, and train a 2D CycleGAN model to selectively enhance the existing contrast and accurately simulate contrast agent's effect in the CT images. Haubold et al. [17] instead investigate into reducing the amount of required iodine-based contrast agent dose from full dose to as low as 50%, and apply a 2D Pix2Pix-based post-processing step to provide the remaining contrast to the CT. Motivated by the drawback of multimodal FDG-PET/CT-based lesion detection systems requiring FDG-PET acquisition, Ben-Cohen et al. [18] investigate synthesizing FDG-PET images from just CT scans. The authors use a two-step approach combining a fully convolutional regression network with a Pix2Pix conditional GAN. They argue that CT-derived synthetic FDG-PET can potentially substitute a clinically acquired FDG-PET scan for improving tumor visibility, although they do not provide any physiological basis for their image translation problem. Bi et al. [19], as opposed to other studies which focus on the direct clinical application of synthetic images, aim at synthesizing realistic FDG-PET samples for providing *data augmentation* to train auxiliary deep-learning-based systems, for example, malignancy detection models. In order to preserve the position and shape of the tumor across translation, they utilize explicitly the tumor annotation together with the CT image in a two-channel Pix2Pix model to generate the corresponding FDG-PET.

Based on our literature survey, we found no prior work that investigated GAN-based cross-modality translation for synthesizing hypoxia PET from other modalities. We explore both the paired and unpaired image translation approaches, using Pix2Pix and CycleGAN systems, respectively, since they are among the most straightforward and widely applied image translation methods.

2.3 Evaluation of Synthetic Medical Images

Evaluation of GAN-generated synthetic images is a difficult problem even when the ground truth reference images are available. In quantitative medical images, pixel-wise differences such as mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) can be useful in only partially assessing image quality since these metrics do not account for the image statistics and local structure. Perceptual metrics like structural similarity index (SSIM) have, therefore, been widely applied to assess the synthetic images based on the similarity of their perceived local structure with that of the ground truth [10]. Image registration literature provides entropy-based metrics for image matching, such as normalized mutual information (NMI) [20], that can measure structural alignment even across different image modalities. Mutual information has been applied as an image quality metric in works concerning translation across Magnetic Resonance (MR) image sequences [21, 22].

In many cases, the end goal of the synthetic images is to potentially substitute the clinically acquired scans, and therefore, image evaluation must be application-specific and clinically relevant. It is, therefore, common to perform downstream tasks using synthetic images to measure their clinical value. Haubold et al. [17] include in their evaluation a manual "pathological consistency" testing procedure where synthetic image slices are compared to their corresponding ground truth slices by trained radiologists to check whether the synthetic image preserves the patient's existing pathology and doesn't insert pathology that is non-existent in the ground truth. Ben-Cohen et al. [18] evaluate their synthetic FDG-PET images by using them, in combination with CT, to detect lesions in the image and measuring the detection performance.

We perform an evaluation on our synthetic HX4-PET images via both general image quality metrics and application-specific downstream image analysis. Because each image quality metric has its own strengths and weaknesses, we use a multitude of them to effectively quantify the fidelity of the synthetic images. Additionally, a systematic visual inspection is performed to identify common failure modes of the image translation GANs used. As downstream tasks, we perform binary classification of the tumors as hypoxic or non-hypoxic and segmentation of the 3D hypoxic regions within the tumor. HX4-PET images are quantitative in nature and have established procedures for hypoxia quantification [11]. The hypoxia quantification tasks are, therefore, performed using simple thresholding-based methods with standard threshold values derived from the relevant clinical literature [11, 6].

Chapter 3

Methodology

3.1 Data

The medical imaging dataset used in this work, hereafter referred to as the *Maastro Lung HX4 dataset*, is a combination of two different collection of scans originally acquired during two clinical trials at Maastro Clinic (Maastro Clinic, Maastricht, The Netherlands), and was previously also used by Even et al. [6] in their hypoxia prediction study. The dataset includes 3D scans of 34 Non-Small Cell Lung Cancer (NSCLC) patients in total, of which 15 belong to the *PET-Boost* trial (registration number NCT01024829) [23] and 19 to the *Nitroglycerin* trial (registration number NCT01210378) [6]. The request for the usage of this data in this project was reviewed and approved by the institutional review board (IRB).

3.1.1 Original Scans

In the original dataset, each individual scan is stored as a separate DICOM series [1] which contains the image intensity data, the spatial information of the image, and the image acquisition details. The first set of images for each patient includes the FDG-PET/CT images, whose CT component is the radiotherapy treatment *planning* CT (pCT). During the treatment planning process, while the pCT scan shows the precise anatomy of the patient which is crucial for annotating various regions-of-interest (ROI), FDG-PET improves tumor visibility by providing high contrast to the malignancy region. Both images were acquired using a single hybrid scanner and are, hence, spatially aligned with each other by default. The 3D field-of-view (FOV) of each image includes the patient's chest, although the FOV of pCT is different than that of FDG-PET. Spatial resolution of all FDG-PET images is $4\times4\times3$ mm^3 and that of the pCT is $0.97\times0.97\times3$ mm^3, specified in the $x\times y\times z$ format. Additionally, annotations

[1] DICOM is an international standard for storing and managing medical images: `https://www.dicomstandard.org/`

of certain structures relevant to the treatment planning process (i.e. ROIs), for example, the primary tumor, nearby organs-at-risk, and the patient's body, are provided for each patient in the DICOM *RTstruct* format [2]. These structures were delineated by a radiation oncologist on the pCT. Figure 3.1 visualizes the FDG-PET and pCT scans of a sample patient.

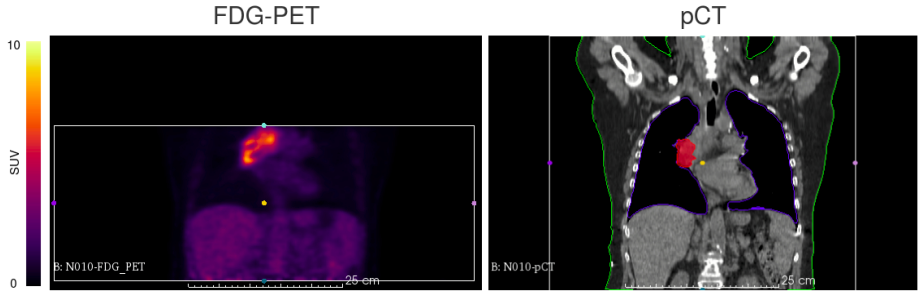

Figure 3.1: Single coronal (x-z plane) slice of FDG-PET and pCT images. White bounding boxes show their respective FOVs. Radiotherapy structure annotations were provided for pCT of which three are shown here – the patient's body (green), the lungs (dark blue) and the primary tumor (red). The FDG tracer has high specificity to metabolism, resulting in high contrast near the tumor.

In PET image acquisition, a CT scan is usually acquired in the same session whether or not this scan has a separate role, for example, in treatment planning. The first reason being that the CT modality provides the anatomical context for using the PET image as a contrast map. Second, the material density information captured in the CT scan is used to apply density correction to the recorded PET intensities. The second set of images in the dataset, therefore, includes for each patient an HX4-PET image as well as its accompanying CT scan, both of which are spatially aligned with each other. This CT was acquired using a low radiation dose and is hereafter referred to as *ldCT*. For each patient, the HX4-PET/ldCT scans were acquired on a different day than their corresponding FDG-PET/pCT scans. Both HX4-PET and ldCT images cover approximately the same FOV of the patient's chest. However, the common FOV covered by the HX4-PET/ldCT couple is more focused on the tumor and is smaller in the axial direction compared to that of the FDG-PET/pCT couple. The HX4-PET images have a resolution of $4\times4\times4$ mm^3 and the resolution of ldCT is $1.17\times1.17\times4$ mm^3. Figure 3.2 shows a visualization of sample HX4-PET and ldCT images.

Voxel intensity values of both pCT and ldCT images are expressed in Hounsfield Units (HU), which relate to the radiodensity of materials. In FDG-PET, the units represent levels of radioactivity emitted from the FDG tracer in different parts of the body as measured by the scanner and are represented in terms of Becquerel per milliliter (Bq/mL, abbreviated as BQML). In HX4-PET, however, intensities are expressed in a different unit called *counts* (abbreviated as

[2]Radiotherapy structure set (RTstruct) is the DICOM-specified format for storing radiotherapy-related annotations: `https://dicom.innolitics.com/ciods/rt-structure-set`

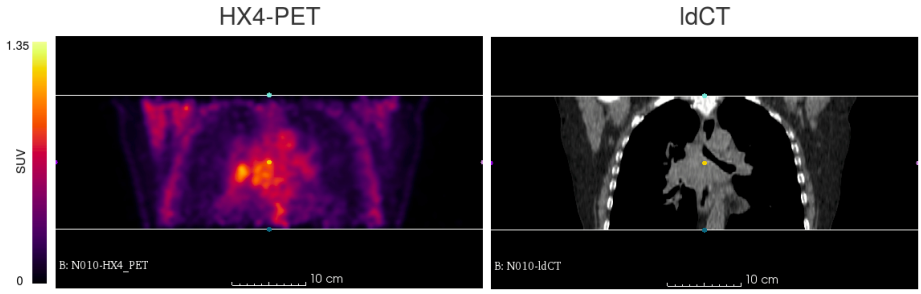

Figure 3.2: Single coronal slice of HX4-PET and ldCT images. Both had similar FOVs. No annotations were provided for the ldCT.

CNTS). This unit of PET representation is specific to the manufacturer and the model of the PET/CT scanner (Philips Gemini TF64) that was used for HX4-PET acquisition. Before utilizing the FDG-PET and HX4-PET images for any purpose, their intensities must be converted to a standard and clinically relevant intensity representation – the Standardized Uptake Value (SUV). SUV calculation is discussed in further detail in 3.1.3.

3.1.2 Image Registration

Note that the two PET/CT scan pairs – FDG-PET/pCT and HX4-PET/ldCT – were acquired on different days for each patient, and because of the differences in patient positioning, they do not spatially align with each other. Having a version of HX4-PET that is aligned with FDG-PET/pCT is essential to serve as the ground truth image for evaluating the GANs as well as for paired training of Pix2Pix. Even et al. [6] describe a registration procedure for achieving this alignment. They use a two-step approach where the ldCT is first subjected to rigid alignment followed by non-rigid elastic registration over the pCT. The resulting transformation parameters are then applied to the HX4-PET image.

In addition to the originally acquired scans, the dataset also already includes the registered HX4-PET and ldCT images, and therefore this registration step wasn't needed to be performed. These images are supplied in the *.mhd/.raw* format. The intensity values of the registered HX4-PET are expressed in SUVs, supposedly calculated before the registration process.

To distinguish the registered HX4-PET from its unregistered counterpart, the two are denoted as *HX4-PET-reg* and *HX4-PET-unreg*, respectively. The term "HX4-PET" is hereafter used to denote the imaging modality in a general sense and not particularly referring to the registered or unregistered versions of the image. FDG-PET and pCT modalities would be used as inputs to the GAN models and HX4-PET-reg as the ground truth wherever required. Since HX4-PET-reg contains artificial deformations and is likely to contain registration errors as well, the HX4-PET-unreg images instead would be used in the unpaired CycleGAN training. The *unregistered* ldCT, *ldCT-unreg*, would play a crucial

role in our modified CycleGAN system, described further in 3.2.2. However, the *registered* ldCT is not required anymore beyond this point since its sole purpose was to aid in the registration of HX4-PET, and is thus discarded.

3.1.3 Data Preparation Procedure

The dataset needed to be prepared by converting it into a form that could be used for GAN training. The data preparation process can be broken down into two phases, which are elaborated in the following paragraphs.

Loading Images and Converting to Appropriate Representation

The first phase of data preparation involves loading the various images to the computer memory and converting them to a suitable representation. The *PyDicom* library [24] is used for reading the DICOM series along with their metadata, and the *SimpleITK* library [25] for the programmatic representation and manipulation of the loaded images. In certain specialized processing tasks, such as PET SUV calculation and RTstruct conversion, parts of code from this [3] public repository are used to ensure the correctness of the implementation.

- *SUV Calculation:* First, the FDG-PET and HX4-PET-unreg intensity values are converted into the SUV scale. When the intensity units are expressed as BQML, as for all FDG-PET images here, the patient's body weight as well as imaging parameters including the decay time and the total administered dose of the radiotracer are required to calculate the SUV. Equation 3.1 shows the conversion formula.

$$SUV = \frac{BQML * 1000 * Weight}{Dose * 2^{-t/\tau}} \tag{3.1}$$

  where $\frac{1}{2^{-t/\tau}}$ is the decay correction factor at time of acquisition t, given the half-life τ of the radiotracer. The values for the patient's weight, administered dose and decay parameters are stored in the DICOM image series as metadata.

  HX4-PET-unreg images, however, are represented in terms of the non-standard CNTS units here, for reasons discussed earlier in 3.1.1. CNTS-to-SUV conversion is performed using the formula implemented here [4].

- *Conversion of RTstruct Contours to Binary Masks:* RTstruct is the format defined in the DICOM standard for storing radiotherapy-related annotations (known as "radiotherapy structure sets") as a set of contour points. The RTstruct file is stored separately from the scans. Since the

[3] HECKTOR 2020 Challenge repository: `https://github.com/voreille/hecktor`. It contains utility code created as part of the HECKTOR PET/CT segmentation challenge [26] by its respective authors.

[4] Implementation of CNTS-to-SUV conversion from the HECKTOR source code: `https://bit.ly/3wbtGgW`

pCT is used for annotating these contours during radiotherapy treatment planning, the physical coordinate system of the pCT scan is the reference frame for the contour points. Of all the different objects annotated in the pCT scan, we select only the contours for the primary tumor and the patient's body, and convert them to binary masks. Each of these resulting masks has the same 3D size and resolution as the pCT image. The tumor mask would be needed during the clinical evaluation of the hypoxia maps predicted by the models, whereas the body mask would be required during model training to mask away and remove out-of-body objects that are visible in the scans, for instance, the scanner table. The algorithm used for RTstruct-to-binary-mask conversion can be found here [5].

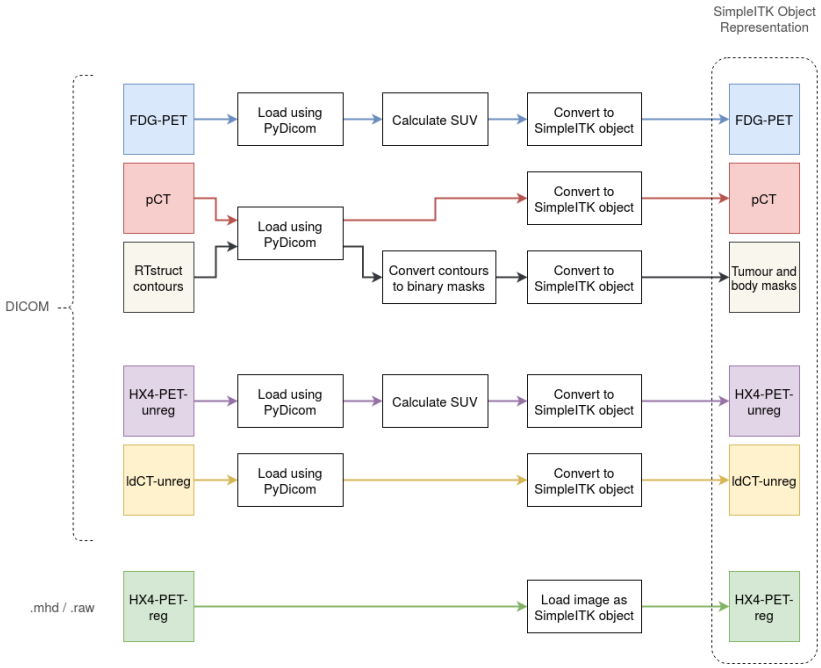

Figure 3.3: Overview of the first phase of data processing. Images are loaded from storage and converted into a suitable representation for further processing.

Following this, the FDG-PET, HX4-PET-unreg, pCT and the tumor and body masks are converted to SimpleITK objects in Python. The ldCT-unreg is also loaded using PyDicom and converted to this representation. HX4-PET-reg, stored in *.mhd/.raw* format containing pre-computed SUV values, is directly loaded into a SimpleITK object. Figure 3.3 shows a visual overview of these steps.

[5] Implementation of RTstruct-to-binary mask conversion from the HECKTOR source code: `https://bit.ly/2RnNI9t`

Cropping and Resampling

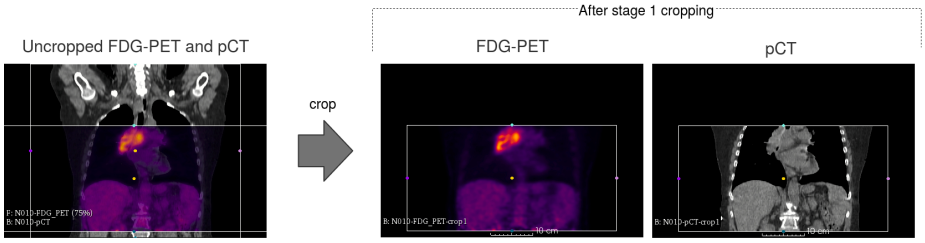

Figure 3.4: Stage 1 cropping of FDG-PET, pCT and the tumor and body masks (not shown here) to a common FOV. Uncropped FDG-PET and pCT are overlaid to show the overlap of their FOVs.

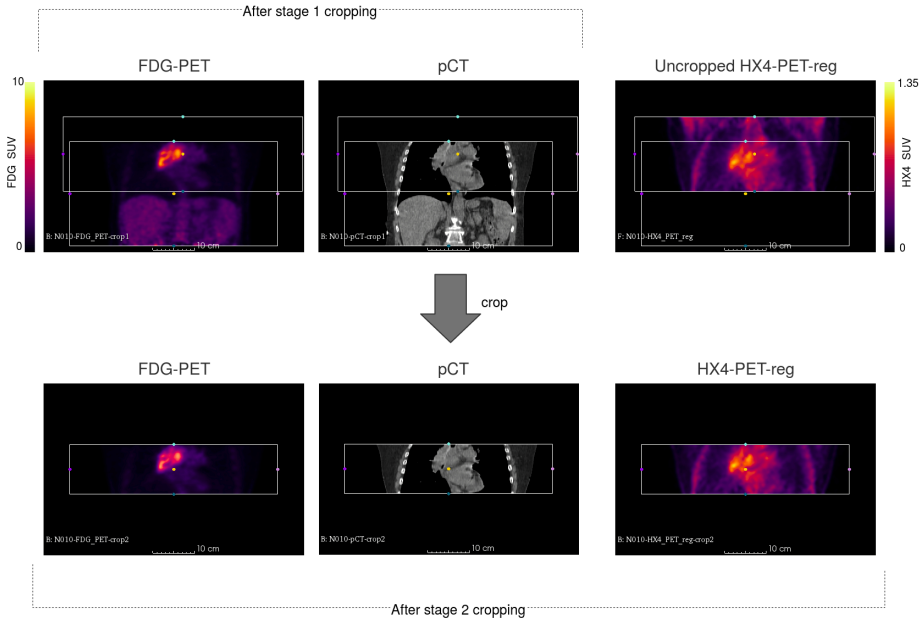

Figure 3.5: Stage 2 cropping of the input images (FDG-PET and pCT), the masks (tumor and body, not shown here) and the ground truth (HX4-PET-reg) to a common FOV. All FOVs are displayed to show their overlap.

The second phase of data processing involves cropping and resampling the loaded images. A standard resampling resolution of $1\times1\times3$ mm^3 is used. Cropping is performed in a progressive manner as described in the following steps:

1. *Cropping and resampling FDG-PET, pCT and masks:* As shown earlier in 3.1.1, the FOVs of FDG-PET and pCT are different. Both images must

be cropped such that each of them would contain the same FOV of the patient. This is accomplished by determining the 3D volumetric intersection of their FOVs and then cropping the images to this 3D bounding box. Resampling is performed immediately following this. The tumor and body masks share the same FOV with pCT, and are, therefore, cropped in the same way as pCT. This step corresponds to the *Stage 1* cropping of FDG-PET, pCT and the masks. Figure 3.4 illustrates this step.

2. *Cropping and resampling HX4-PET-unreg and ldCT-unreg:* The same process of cropping to common FOV followed by resampling is performed on HX4-PET-unreg and ldCT-unreg images. This corresponds to the *Stage 1* cropping of HX4-PET-unreg and ldCT-unreg.

3. *Resampling HX4-PET-reg:* The HX4-PET-reg image is then only resampled to the standard resolution.

4. *Cropping the input images, the masks and the ground truth image to common FOV:* The two input images (FDG-PET and pCT), the two masks, and the ground truth image (HX4-PET-reg) are then cropped to the common FOV shared across all of them. The is required because although HX4-PET-reg is spatially aligned with the input images and shares the same physical reference frame with them (due to registration), its FOV is different. This step corresponds to the *Stage 2* cropping, and is visualized in Figure 3.5.

Finally, all the processed images are saved as *.nrrd* files. The NRRD format allows the storage of N-dimensional raster images along with their spatial information, including voxel spacing (resolution) and physical coordinates of the origin, in a single file. Figure 3.6 shows an overview of the second phase of data processing discussed thus far. The dataset is then split into a training set and a validation set. The training set is comprised of data from the *PET-Boost* clinical trial (15 patients), and the validation set consists of data from the *Nitroglycerin* trial (19 patients).

3.2 GAN Systems

This section describes the image translation GAN systems used in our study. Unless explicitly mentioned otherwise, domain A is used throughout this section to denote the domain of valid multimodal FDG-PET/pCT images, and B to denote the domain of HX4-PET images. Synthetic HX4-PET is referred to as *HX4-PET-syn*.

3.2.1 Pix2Pix

The Pix2Pix framework provides a straightforward image translation method, provided that the domain A and domain B images in the training dataset are paired, i.e. they are spatially aligned and have pixel-wise correspondence. Given

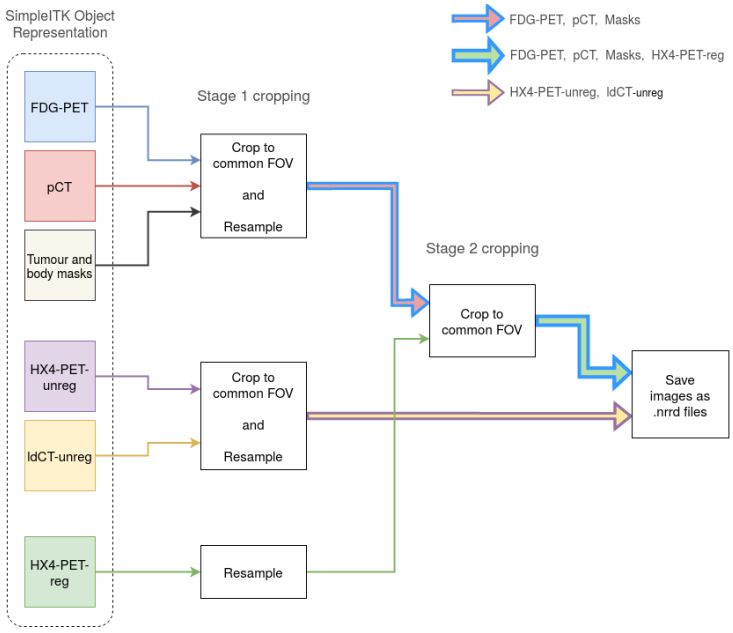

Figure 3.6: Overview of the second phase of data processing.

a task of translating a domain A image to its equivalent domain B image, it is first assumed that an $A \rightarrow B$ mapping exists. The generator models this relationship between the domains, and the discriminator models the conditional probability that an image belongs to domain B, given its domain A counterpart. During training, the conditional discriminator evaluates the generated domain B images, and its adversarial feedback is used to improve the generator's performance. Pix2Pix additionally includes an element-wise loss to directly enforce similarity between the synthetic and the ground truth domain B images. For a generator G and a discriminator D, Equations 3.2 show the components of each of their loss functions.

$$\begin{aligned} L(G) &= L_{adv}(G) + \lambda_{elem} L_{elem}(G) \\ L(D) &= L_{adv}(D) \end{aligned} \tag{3.2}$$

The generator loss has two terms – an adversarial loss $L_{adv}(G)$ and an element-wise loss $L_{elem}(G)$. The discriminator loss is a single adversarial loss $L_{adv}(D)$ which is the adversarial counterpart of $L_{adv}(G)$. For the generator, $L_{adv}(G)$ can be viewed as a "structured" loss that is *learned* from the data itself as a result of the adversarial setting and can produce images with sharp and fine structures, whereas $L_{elem}(G)$ emphasizes on the fidelity of larger and lower-frequency components of the images. The hyperparameter λ_{elem} can be used to adjust the balance of importance between the two terms depending on the specific application.

We use the least-squares version of the adversarial loss, instead of the original cross-entropy loss, since it is known to improve training stability [27]. Similar to the original Pix2Pix framework [8], we implement $L_{elem}(G)$ as voxel-wise L1 loss. Equations 3.3 show the expansion of each of these loss terms.

$$\begin{aligned} L_{adv}(G) &= E_{x\sim p_{data}(x)}[(1 - D(x, G(x)))^2] \\ L_{adv}(D) &= E_{x\sim p_{data}(x)}[D(x, G(x))^2] + E_{x,y\sim p_{data}(x,y)}[(1 - D(x, y))^2] \\ L_{elem}(G) &= E_{x,y\sim p_{data}(x,y)}[||y - G(x)||_1] \end{aligned} \tag{3.3}$$

where x and y are random variables representing paired samples from domain A and B, respectively. During training, G and D are updated in an alternating fashion. For a given D, G learns to synthesize images that D would wrongly classify as real data samples, with value 1 (optimizing for $L_{adv}(G)$), while simultaneously aiming for voxel-wise fidelity of its output with the ground truth (optimizing for $L_{elem}(G)$). Whereas, for a given G, D learns to correctly classify the synthetic samples as 0 and the real samples as 1 (thereby optimizing for $L_{adv}(D)$).

We apply Pix2Pix to our translation problem in a 3D manner on the volumetric medical images. Given the FDG-PET and pCT images as input and the corresponding HX4-PET-reg images as the target, the model is trained to generate HX4-PET-syn images representative of the target domain. Figure 3.7 shows a schematic of the generator and discriminator training phases.

3.2.2 CycleGAN

CycleGAN addresses unpaired image translation problems where A-B image pairs are not provided in the training set. It is first assumed that there exists some relationship between domain A and domain B and that this relationship is bijective in nature. In other words, for an $A \rightarrow B$ transformation, an inverse $B \rightarrow A$ transformation exists. CycleGAN aims at modeling these two mappings simultaneously by incorporating cycle-consistency with the principal idea being that translating an image to another domain and then back to its original domain should yield the same image. The GAN system consists of two generators – G_{AB} and G_{BA} – that implement the $A \rightarrow B$ and $B \rightarrow A$ mappings, respectively. Two discriminators D_B and D_A are adversarially coupled with the respective two generators. Note that, unlike Pix2Pix, discriminators in CycleGAN are not conditioned on input images as this is not possible with unpaired training data. Equations 3.4 show the generator and discriminator loss terms.

$$\begin{aligned} L(G_{AB}, G_{BA}) &= L_{adv}(G_{AB}) + L_{adv}(G_{BA}) + \lambda_{cyc} L_{cyc}(G_{AB}, G_{BA}) \\ L(D_B, D_A) &= L_{adv}(D_B) + L_{adv}(D_A) \end{aligned} \tag{3.4}$$

The adversarial loss term $L_{adv}(G_{AB})$ of G_{AB} is coupled with $L_{adv}(D_B)$ of D_B and encourages G_{AB} to produce images resembling domain B samples. Similarly, $L_{adv}(G_{BA})$ and $L_{adv}(D_A)$ are counterparts of each other and drive G_{BA}

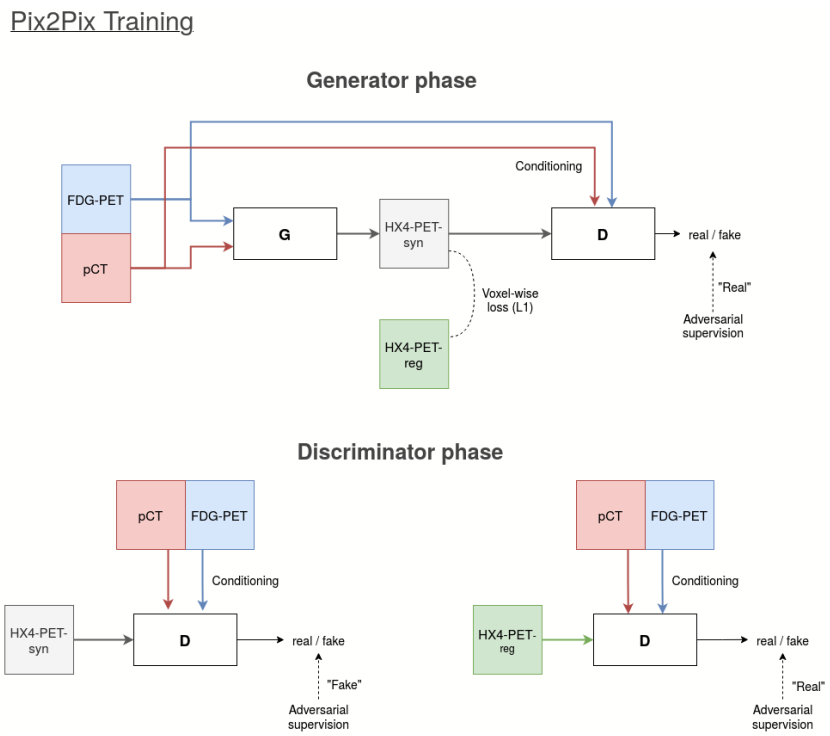

Figure 3.7: Pix2Pix training phases. G and D are updated in an alternating manner. In practice, FDG-PET and pCT images are stacked channel-wise and given as input to G. And the discriminator conditioning is implemented by stacking channel-wise the HX4-PET-reg or HX4-PET-syn with FDG-PET and pCT.

to produce images indistinguishable from domain A. It would be possible for a generator to satisfy its adversarial constraints by translating its input image from its source domain to just *any* image that would seem like a sample from its target domain, whether or not the translated image preserves the contents of the input image. The cycle-consistency term $L_{cyc}(G_{AB}, G_{BA})$, therefore, acts as an indirect content-preservation criterion that constrains G_{AB} and G_{BA} to preserve the high-level structure across translation, thereby complementing the adversarial losses. The hyperparameter λ_{cyc} is used to adjust the relative importance given to cycle-consistency loss.

Let $x \sim p_{data}(x)$ represent the random variable associated with domain A following a distribution $p_{data}(x)$, and let $y \sim p_{data}(y)$ be the random variable associated with the domain B distribution. Equations 3.5 show each of the loss

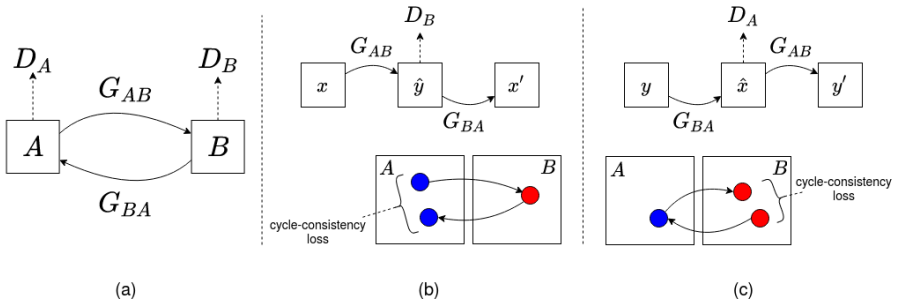

Figure 3.8: Simplified representation of CycleGAN training, figure adapted from the original paper [9] with notations changed. (a) G_{AB} models $A \rightarrow B$ mapping and G_{BA} its inverse. D_B and D_A are domain B and A discriminators, respectively. (b) In the $A \rightarrow B \rightarrow A$ cycle, a domain A image x is translated to a representation $\hat{y} = G_{AB}(x)$ in domain B, which is evaluated by D_B. It is then used to obtain a reconstruction $x' = G_{BA}(G_{AB}(x))$ of the original image x which is encouraged by the cycle-consistency to be close to x. (c) The $B \rightarrow A \rightarrow B$ cycle.

terms in expanded form.

$$
\begin{aligned}
L_{adv}(G_{AB}) &= E_{x \sim p_{data}(x)}[(1 - D_B(G_{AB}(x)))^2] \\
L_{adv}(D_B) &= E_{x \sim p_{data}(x)}[D_B(G_{AB}(x))^2] + E_{y \sim p_{data}(y)}[(1 - D_B(y))^2] \\
\\
L_{adv}(G_{BA}) &= E_{y \sim p_{data}(y)}[(1 - D_A(G_{BA}(y)))^2] \\
L_{adv}(D_A) &= E_{y \sim p_{data}(y)}[D_A(G_{BA}(y))^2] + E_{x \sim p_{data}(x)}[(1 - D_A(x))^2] \\
\\
L_{cyc}(G_{AB}, G_{BA}) &= E_{x \sim p_{data}(x)}[||x - G_{BA}(G_{AB}(x))||_1] \\
&\quad + E_{y \sim p_{data}(y)}[||y - G_{AB}(G_{BA}(y))||_1]
\end{aligned}
\tag{3.5}
$$

Cycle-consistency loss $L_{cyc}(G_{AB}, G_{BA})$ encourages a reconstructed image to be as close as possible to its original image in L1 sense, and is applied to both $A \rightarrow B \rightarrow A$ and $B \rightarrow A \rightarrow B$ cycles. Figure 3.8 shows a conceptual visualization of CycleGAN training.

CycleGAN-naive: Default CycleGAN applied to HX4-PET synthesis problem

We investigate CycleGAN for unpaired translation of FDG-PET/pCT to HX4-PET. The term "unpaired" can be used at two levels here – the patient level and the voxel level. At the patient level, our dataset includes all three image modalities for each patient. The training data used for the CycleGAN is patient-level unpaired, meaning the A-B image correspondence information is not used, and both domain A and domain B samples are independently shuffled while training. At the voxel level, the registered HX4-PET-reg images are spatially

aligned input-target paired data, whereas the unregistered images – HX4-PET-unreg – are devoid of any artificial deformations and imperfections that are related to registration. We, therefore, use HX4-PET-unreg images as the set of domain B samples which results in our CycleGAN training data being unpaired at the voxel level as well. The default CycleGAN framework naively applied to our use case is henceforth referred to as *CycleGAN-naive*. Figure 3.9 shows a schematic of the generator and discriminator training phases of CycleGAN-naive.

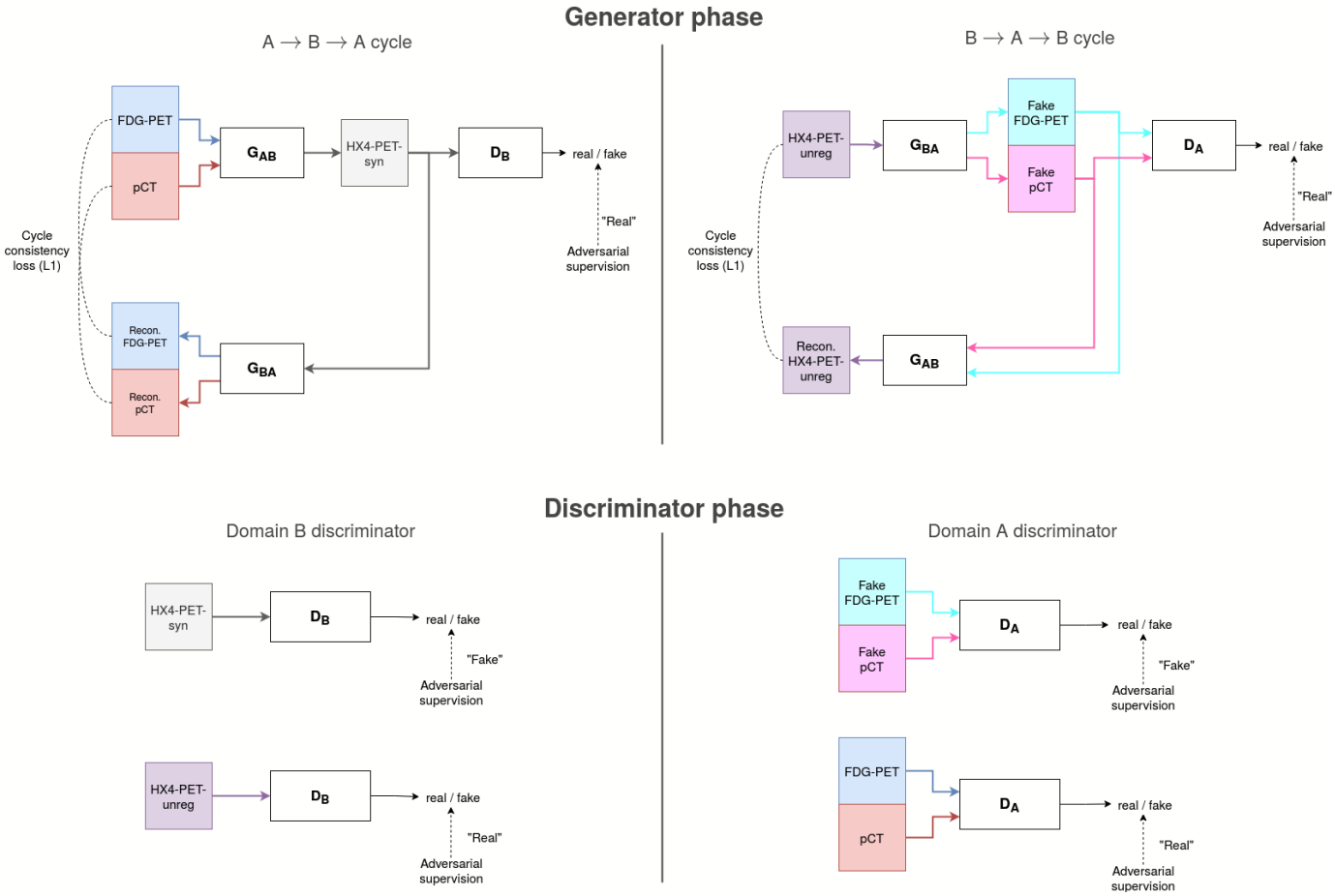

Figure 3.9: CycleGAN-naive training phases. The generators and discriminators are updated in an alternating manner.

We argue that CycleGAN-naive has several drawbacks and is, therefore, not an optimal implementation of CycleGAN for our task. In the $A \rightarrow B \rightarrow A$ cycle, G_{AB} takes as input FDG-PET and pCT to output HX4-PET-syn, and G_{BA}

uses HX4-PET-syn to reconstruct FDG-PET and pCT. The learning task of G_{AB} is realistic since hypoxia patterns can be reliably predicted from the inputs, and has a physiological basis (see Section 2.1). However, the learning task of G_{BA} is not reasonable. It is not possible to precisely predict a patient's anatomy (CT) as well as their metabolism levels (FDG-PET) from just the hypoxia information; the CT modality is of more concern to our discussion. A CT image contains highly detailed structural information with sharp and fine features, whereas PET images are functional heatmaps having blurred and diffused characteristics. G_{BA} cannot accurately rebuild pCT from HX4-PET-syn unless G_{AB} encodes the CT information into the latter in the form of noise. But then, the HX4-PET-syn image would not resemble a clinically acquired HX4-PET scan. Therefore, cycle-consistency constraint conflicts with the domain B adversarial criterion instead of complementing it. Now looking at the $B \rightarrow A \rightarrow B$ cycle, the task of G_{BA} is to compute realistic FDG-PET and pCT when given only the real HX4-PET-unreg image, which it can perform, especially for pCT, by hallucinating the patient's anatomy. However, in that case, G_{AB} would be given this fake information to precisely reconstruct the HX4-PET-unreg. Because the task of G_{AB} is to correctly predict hypoxia based on given CT (and FDG-PET) information, it is unreasonable to provide it with fake anatomical information and expect it to predict hypoxia patterns from it that match the real HX4-PET-unreg. In this cycle as well, contradictions across the learning objectives exist. The fundamental assumption made by CycleGAN regarding the existence of a bijective relationship between domains A and B does not hold in our data, and the resulting images are likely to be highly inaccurate and not representative of the patient's condition. Based on these reasons, we believe that CycleGAN, in its default form, cannot tackle this translation problem and that a more intelligent strategy is required.

CycleGAN-balanced: An improved design

We propose a custom design improvement to CycleGAN training to make the system more optimized for our use case. We utilize the fact that an HX4-PET scan is always acquired together with a CT scan which, in our dataset, is a low-dose CT (ldCT). For this system, the training process is unpaired similar to CycleGAN-naive, and the ldCT-unreg images are included in the training data accompanying their corresponding HX4-PET-unreg images.

We begin by slightly altering what the domains A and B denote here – let A be the domain of FDG-PET images and B of HX4-PET images. The CT images – pCT and ldCT – are then considered to be "supporting" images that provide *anatomical context* to the generators. Note that pCT has spatial correspondence only with FDG-PET, and ldCT-unreg is aligned only with HX4-PET-unreg. Now, the task of each generator is to take as input a PET image from its source domain and compute its target domain PET image, while being supplied anatomical context using the *available* CT image (i.e the one that is aligned with the PET from the generator's source domain). To elaborate on this, consider the two cycles separately:

- *$A \rightarrow B \rightarrow A$ cycle*: Here, pCT is the available CT image which accompanies the FDG-PET. In $A \rightarrow B$ translation, G_{AB} uses the two images and computes HX4-PET-syn. In $B \rightarrow A$ translation, G_{BA} now *reuses* the pCT as a support image along with the HX4-PET-syn (note that the two are aligned) to reconstruct FDG-PET. The pCT provides the same anatomical context for FDG-PET reconstruction as it did for HX4-PET-syn prediction. The cycle-consistency loss is computed between the FDG-PET and its reconstructed version. If the HX4-PET-syn represents a plausible hypoxia prediction, then the FDG-PET can be recovered from it up to a great extent since both have an underlying physiological relationship. Therefore, the adversarial loss would drive G_{AB} to produce realistic HX4-PET-syn image and the cycle-consistency loss would encourage this image to be representative of the patient's physiology.

- *$B \rightarrow A \rightarrow B$ cycle*: In this case, ldCT-unreg is the available CT image that is coupled with the HX4-PET-unreg. In $B \rightarrow A$ translation, G_{BA} uses these two images to compute a synthetic FDG-PET. In $A \rightarrow B$ translation, G_{AB} reuses the ldCT-unreg and together with the synthetic FDG-PET (note that the ldCT-unreg and the synthetic FDG-PET are aligned), it reconstructs HX4-PET-unreg. Cycle-consistency is applied between the original and reconstructed HX4-PET-unreg. The same ldCT-unreg image provides anatomical context for FDG-PET synthesis and for HX4-PET-unreg reconstruction, and if the synthetic FDG-PET represents plausible metabolism patterns, then HX4-PET-unreg can be largely recovered. In this cycle as well, the adversarial and cycle-consistency objectives would behave in a mutually reinforcing manner.

The introduction of CT as an *input* to both the generators creates a balance between the complexities of the $A \rightarrow B$ and $B \rightarrow A$ mappings since none of them involves *predicting* fake CT information, while still using completely A-B unpaired training data. Unlike CycleGAN-naive, the $A \rightarrow B$ and $B \rightarrow A$ relations are not anymore inverses of each other. We term this modified design *CycleGAN-balanced*. Figure 3.10 shows a schematic of the training phases of its generators and discriminators.

Low-dose CT scans are characterized by a higher amount of acquisition-related noise compared to routine- or high-dose CT scans (such as planning CT). An additional, possibly desirable, consequence of exposing the two generators to both pCT and ldCT images during training is that the model is likely to learn effectively the relevant anatomical features while disregarding any pCT-specific or ldCT-specific noise signatures. That is, they are likely to generalize over both types of CT images. This is merely a property of CycleGAN-balanced that we make a note of, although we do not investigate this further.

CycleGAN-balanced Training

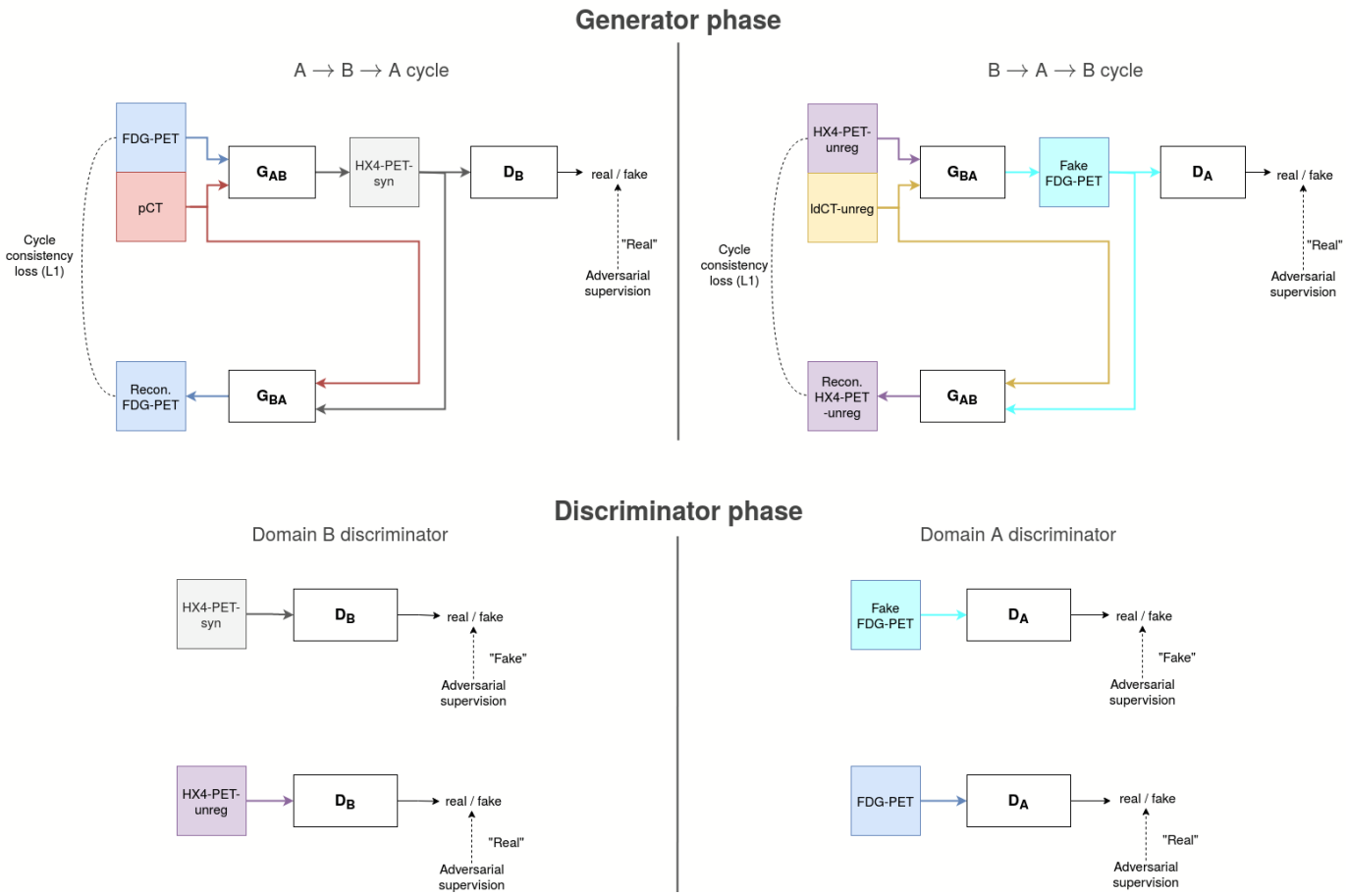

Figure 3.10: CycleGAN-balanced training phases. $A \rightarrow B$ and $B \rightarrow A$ translation tasks are balanced. The CT images provide anatomical context to the generators. In $A \rightarrow B \rightarrow A$ cycle, the pCT is used as an input by G_{AB} and then reused by G_{BA}. In the $B \rightarrow A \rightarrow B$ cycle, the ldCT-unreg is the common CT image used by both the generators.

3.3 Network Architecture

The Pix2Pix, CycleGAN-naive, and CycleGAN-balanced systems share the same generator and discriminator network architectures for the sake of fair comparison. Our design choices of the architectures are based on the specifications of the original Pix2Pix and CycleGAN frameworks [8, 9] and involve the 3D equivalent of the 2D spatial operations.

3.3.1 Generator

Each generator network is based on the 3D U-Net architecture [28]. 3D U-Net is a fully convolutional network that takes as input a 3D volumetric image of arbitrary size and can produce an output image of the same size as the input. It consists of an *analysis* path (or encoder) where hierarchical features are progressively extracted from the input image and a *synthesis* path (or decoder) that progressively constructs the output image from the features. A desirable property of the U-Net is its skip connections that transport features of multiple abstraction levels directly into the synthesis path, enabling it to utilize both high- and low-level features in synthesizing the output. In contrast to a U-Net, a traditional encoder-decoder network without any skip-connections must depend on its highest-level features (i.e. features from its "bottleneck"). In many image translation problems, the input and target domain images could share a large amount of low-level information. The U-Net is known to produce better quality images in such cases compared to the encoder-decoder architecture [8].

Loosely based on the notation used in [8], we use *Enc-Ck* to represent an encoder block composed of Convolution-InstanceNorm-LeakyReLU layers where the convolution operation uses k 3D kernels of size $4\times4\times4$ voxels, applied with a stride $2\times2\times2$ and zero padding of 1 at each side of the input along each axis. The Leaky ReLU has a negative side slope of 0.2. For the decoder, let *Dec-Tk* represent a block comprising of TransposedConvolution-InstanceNorm-ReLU layers where the 3D upsampling transposed convolution uses k kernels of size $4\times4\times4$, stride $2\times2\times2$, and input padding of 1 at each side along each axis. Under this configuration, the encoder downscales the image representation by a factor of 2 after each block, and the decoder upscales it by the same factor. Our U-Net architecture is then specified in Table 3.1.

Encoding Level	**Encoder block**	**Decoder block**
1	Enc-C64	Dec-T1
2	Enc-C128	Dec-T64
3	Enc-C256	Dec-T128
4	Enc-C512	Dec-T256
5	Enc-C512	Dec-T512

Table 3.1: The first encoder block Enc-C64 accepts a two-channel input image and the encoder path processes the image from encoding level 1 to 5. The first decoder block Dec-T512 is in correspondence with the last encoder block Enc-C512, and the decoder path performs computation starting from encoding level 5 to 1, ultimately producing a single channel output image from Dec-T1. An exception to the U-Net's input and output channels appears in CycleGAN-naive, where the generator G_{BA} requires a single channel input and produces a two-channel output.

There are two exceptions to the aforementioned block notations. The first encoder block Enc-C64 and the last decoder block Dec-T1 do not use instance normalization. And, the latter uses the *tanh* activation function to output an image with intensity values bounded within the range (-1, 1).

3.3.2 Discriminator

We use the 3D PatchGAN architecture for our discriminators. The network is essentially a convolutional classifier that, instead of computing a single validity score for an image (0 if fake, 1 if real), evaluates independent scores for fixed-size local regions (or patches) of the image. It, therefore, encourages the synthetic images to not only resemble the target domain at the global level but also for its local high-frequency structures to reflect the characteristics of the domain. PatchGAN discriminator of a fixed architecture can take an input image of arbitrary size and would produce an output map whose size depends on the input image size, while the size of *receptive field* of each output unit on the input image is constant. The receptive field size can be increased by increasing the depth of the network.

Let *Ck* denote a block composed on Convolution-InstanceNorm-LeakyReLU where the 3D convolution layer uses k kernels of size $4\times4\times4$ and input padding of 1 at each side along each axis, and the Leaky ReLU has a negative side slope of 0.2. The exact discriminator architecture is then shown in Table 3.2. Under this configuration, the receptive field of each output unit is $70\times70\times70$ voxels on the input image.

Block number	Discriminator block
1	C64
2	C128
3	C256
4	C512
5	C1

Table 3.2: The first discriminator block C64 takes as input a single channel image (except of D_A of CycleGAN-naive which requires two-channel input). The last block C1 always outputs a single-channel validity score map. Blocks 1, 2 and 3 use a stride of $2\times2\times2$ resulting in a downscaling of the image representation by a factor of 2, whereas blocks 4 and 5 uses a stride of $1\times1\times1$.

As exceptions to the notation, instance normalization is not used in the first and the last blocks. Also, because we use the least-squares adversarial loss, the last block does not include an activation function. The final convolution layer's output map is considered the discriminator's final output.

3.4 Model Training

3.4.1 Overview of the Data Requirements

Figure 3.11 provides a visual overview of the data requirements of each GAN system to clarify which image type is used where, much of which is covered in detail earlier in Section 3.2 for each individual GAN. The FDG-PET and pCT images comprise the model input. HX4-PET-reg is the ground truth hypoxia map, and images from the validation set are used to evaluate and track

model generalizability during the training process. HX4-PET-reg images from the training set are used to train the paired Pix2Pix model. The unregistered image couples HX4-PET-unreg and ldCT-unreg are required only in the training set, of which only the HX4-PET-unreg is involved in CycleGAN-naive training, whereas both are used in CycleGAN-balanced training.

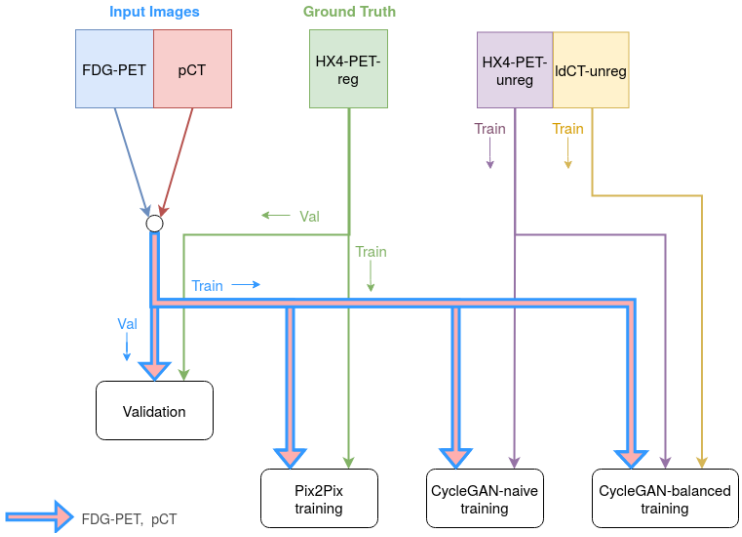

Figure 3.11: Overview of data requirements of the GAN systems.

3.4.2 Training Pipeline

Data pre-processing and normalization

Depending on the GAN, the required training images and the body mask are first fetched from their *.nrrd* form into the computer memory. The body mask is applied to pCT, FDG-PET, and HX4-PET-reg (if using), and the out-of-the-body region intensities are set to zero signal value – i.e. 0 SUV (zero activity) in the PET images and -1024 HU (value for air) in pCT. These body masks were obtained from the pCT (see 3.1.3), and therefore, cannot be applied to the unregistered images. When performing unpaired training with HX4-PET-unreg and ldCT-unreg, the latter is used to automatically generate their own body mask via HU thresholding and morphological operations. This new body mask is then applied to HX4-PET-unreg (and ldCT-unreg, if using). Body masking is then followed by normalization. CT intensities are clipped and limited to the HU range [-1000, 2000] and then linearly rescaled to range [-1, 1]. FDG-PET is clipped into the SUV range [0, 15] followed by rescaling to [-1, 1]. In the case of HX4-PET images, following previous related work [6], the SUV values are first divided by the mean SUV of the patient's aorta region. The resulting intensity representation ($\frac{SUV}{SUV_{aorta-mean}}$ ratio) has a special meaning when

used in the context of the tumor, wherein it is referred to as the tumor-to-background ratio (TBR). Here, the aorta is considered the "background" with respect to the tumor. This is a clinically useful intensity representation that has pre-defined standard thresholds [11] and can, hence, be used for hypoxia quantification during image evaluation (described further in Section 3.5). We perform this conversion step at the beginning of the training pipeline, instead of during the evaluation stage, because the aorta annotation was not provided in the dataset which makes it impossible to compute the $SUV_{aorta-mean}$ value directly from the HX4-PET images. These values were instead provided for each patient directly in a separate file, and we had to depend on them for SUV $\leftrightarrow$ TBR conversion throughout the training and evaluation stages. Following the conversion of the body masked HX4-PET images into this intensity unit, the intensities are clipped to the range [0, 3] and then rescaled to [-1, 1]. The aforementioned clipping ranges for FDG-PET SUV and HX4-PET $\frac{SUV}{SUV_{aorta-mean}}$ are used because they covered most of the intensity variation in their respective images, especially in the tumor area.

Patch-based training

We use a patch-based training approach to cope with the high memory requirements of the 3D GANs and with the small size of the training set. For each pre-processed image, a single 3D patch is extracted from the image and is used to feed the model. The body mask is used to limit patch sampling to the body region to avoid extracting patches from the background. Patches from the input images (FDG-PET and pCT) are sampled randomly with a uniform random distribution over the body region. Then, depending on whether the training is paired or unpaired, a suitable *patch propagation* strategy is used to obtain a corresponding patch from the given HX4-PET image.

Let f_A be the location of the focal point of the patch sampled from the input images. During paired training, a corresponding focal point f_B in HX4-PET-reg is then calculated simply as $f_B = f_A$, since this image is spatially aligned to the input images. Figure 3.12 shows a simplified visualization of the paired patch sampling process.

In case of unpaired training, a more sensible patch propagation strategy is used instead of sampling patches randomly and independently from the input images and the target domain's unregistered images. The idea is to create a form of "weak pairing" between the two sets of patches like so – the anatomical region covered in the input image patches and the HX4-PET-unreg patch should be roughly similar, even though the the input and the target domain images correspond to two different patients. We exploit the fact that all scans have the same patient orientation and a similar FOV. Given the focal point f_A of the patch sampled from the input images, an equivalent focal point f_B is estimated on the HX4-PET-unreg as shown in Equation 3.6.

$$f_B = \frac{f_A}{size_A} size_B \tag{3.6}$$

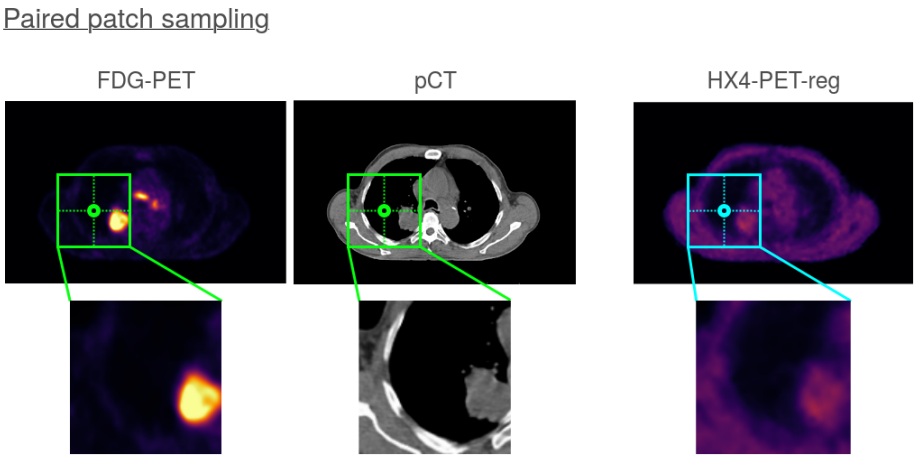

Figure 3.12: Using uniform random sampling in the body region, focal point f_A coordinate (green circles) is first sampled from FDG-PET (and pCT). A corresponding point f_B (cyan circle) is then propagated to HX4-PET-reg such that $f_B = f_A$. Since all three images are spatially aligned, the resulting patches (green and cyan bounding boxes) are aligned as well and cover the same exact anatomical region.

f_B is calculated such that its relative position with respect to the volume size $size_B$ of the HX4-PET-unreg image is same as the relative position of f_A with respect to the size $size_A$ of the input images. This is then followed by defining a *local sampling region* of size $l_x \times l_y \times l_z$ around f_B and then updating f_B by randomly sampling from this local sampling region. Let this new focal point be denoted as f'_B. Again, the condition is applied on f'_B that it must lie within the patient's body region and not in the image background. This additional stochasticity is introduced in the process to account for the variation in the body position and anatomy across different patients. When using ldCT-unreg (i.e. during CycleGAN-balanced training), it shares the same f'_B as HX4-PET-unreg. Figure 3.13 shows a visual representation of the unpaired patch sampling process.

The local sampling region is specified as a fraction of the volume size of HX4-PET-unreg image, and in our experiments, we use size 35% $\times$ 35% $\times$ 60%. This relatively large size is set to provide just sufficient freedom for a set of valid patches to be sampled from the full images. A similar, yet much simpler, 2D slice-based unpaired patch sampling was previously suggested by Yang et al. [29], which produced better results compared to randomly and independently selecting slices for training. We use a fixed patch size of 128×128×32 voxels in both paired and unpaired cases.

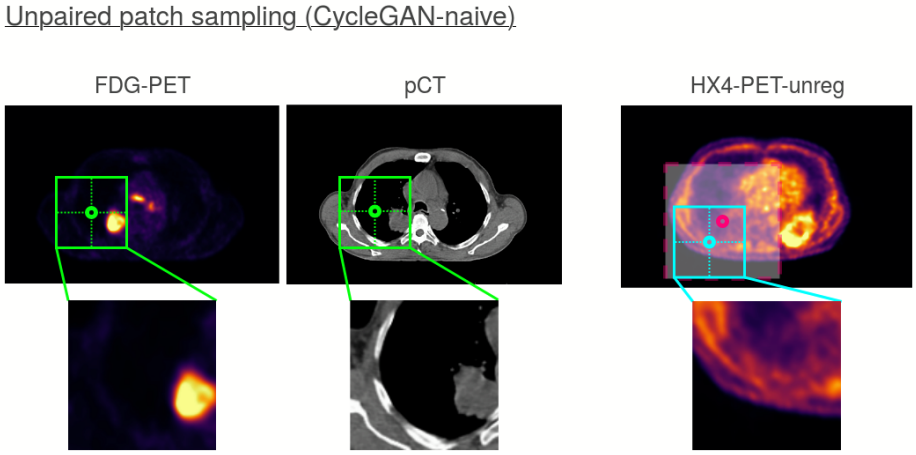

Figure 3.13: Given a focal point f_A (green circles) in the input images, a point f_B (pink circle) is propagated to the HX4-PET-unreg image and a local sampling region (pink translucent bounding box) is defined around it. A new point f'_B (cyan circle) is then sampled from this local region and is used to define the HX4-PET-unreg patch (cyan bounding box). Both sets of patches cover similar anatomical regions, despite belonging to images of two different patients.

Patch-based inference

A patch-based inference is used during the model validation phase using a *sliding window* scheme [6]. The same patch size as used in training is used here as well. Given a pair of full-size input images, overlapping patches are sequentially extracted and fed into the GAN model independently. Their corresponding predicted HX4-PET-syn patches are then stitched together into the full-size HX4-PET-syn image. Following this, various image similarity metrics (discussed further in 3.5.1) are computed between this HX4-PET-syn image and its ground truth HX4-PET-reg. We rescale the predicted and the ground truth images back to their prior intensity scale ($\frac{SUV}{SUV_{aorta-mean}}$ ratio) before computing the metrics. The HX4-PET-syn images are then converted back into SUV intensity representation and saved as NRRD files for later inspection.

3.5 Evaluation Methods

The validation set used to track model generalizability is the same data used for final evaluation. We perform an evaluation on the predicted HX4-PET-syn from two different perspectives. In the first, the objective is to evaluate the overall success of the models in learning the input-target mapping, and the evaluation is performed by assessing the fidelity of the full synthetic samples produced by the models with respect to the ground truth. The second perspective regards

[6] Sliding window inferer: `https://docs.monai.io/en/latest/inferers.html`

the clinical value of the synthetic hypoxia maps as the most important factor since the clinical utility of the models is the end goal. In a clinical scenario, the tumor is the sole region-of-interest (ROI) whose hypoxia distribution needs to be measured, and therefore, the evaluation is limited to this locality and consists of relevant application-specific downstream tasks. All metrics are applied on images with intensities expressed in the aorta-normalized SUV ($\frac{SUV}{SUV_{aorta-mean}}$) scale which is clipped to the interval [0, 3].

3.5.1 Image Quality Metrics

Isola et al. [8] note that since Pix2Pix (and this would apply to CycleGAN by extension) learns a structured loss via its discriminator, using simple pixel-wise difference metrics such as mean-squared error is not an effective evaluation strategy, especially for computer graphics tasks involving natural images. Such metrics do not take into account the image statistics and texture that the model has learned. However, in the case of quantitative medical images where intensities have some standard scale, element-wise differences can be a meaningful choice and could at least partially be informative about the image quality. We use the following three voxel-wise metrics:

1. Mean-Squared Error (MSE): MSE measures the average square of the distances of predicted intensities from the true values. It is more sensitive to large errors. The MSE values are represented in the squared scale of the image intensities and are in the range [0, 9] in our case. Voxel-wise similar images have low MSE.

2. Mean-Absolute Error (MAE): This metric measures absolute deviations from the true value and is therefore more robust to outliers as compared to MSE. MAE is reported on the same scale as the image intensities.

3. Peak Signal-to-Noise Ratio (PSNR): PSNR is a measure related to signal power and is specified in decibels. It is related to MSE in that the "noise" part in PSNR corresponds to MSE measured with respect to the "peak signal" which corresponds to the maximum intensity value in the reference image. A higher PSNR value signifies higher quality image in a voxel-wise sense.

We use three additional metrics that measure the statistical properties of a given image in order to complement the ones based on voxel-wise difference. Like the previous three metrics, each of these is a full-reference metric requiring a ground truth reference image. They are given as follows:

1. Structural Similarity Index Measurement (SSIM): SSIM [30] is a perceptual metric that focuses on gauging the similarity of perceived local structure between the sample and reference images. It has a fixed scale with a range [-1, 1], -1 being the worst score and 1 being the best which is observed when both images are identical.

2. Normalized Mutual Information (NMI): NMI [20] is a variant of the mutual information (MI) metric commonly used in multimodal image registration as an image similarity measure. It is an information theoretic quantity which, in intuitive terms, measures the certainty of predicting intensity values in a region of a sample image knowing the intensities of the reference image in the same region. MI indicates the amount of information shared between the two images, and is maximum (1 in case of NMI) when the images are perfectly correlated and is minimum (0 in case of NMI) when they have no correlation. Given a sample synthetic image $\hat{\mathbf{y}}$ and its corresponding ground truth $\mathbf{y}$, NMI is calculated as:

$$NMI(\mathbf{y}, \hat{\mathbf{y}}) = \frac{H(\mathbf{y}) + H(\hat{\mathbf{y}})}{H(\mathbf{y}, \hat{\mathbf{y}})} \tag{3.7}$$

where $H(\mathbf{y})$ and $H(\hat{\mathbf{y}})$ are marginal entropy values for $\mathbf{y}$ and $\hat{\mathbf{y}}$, and $H(\mathbf{y}, \hat{\mathbf{y}})$ is their joint entropy.

3. Image histogram distance: Histogram distance between two images measures the overlap of their intensity distributions, and has been used as an image similarity measure in applications such as image retrieval. Two identical images would have identical histograms and therefore the distance between the histogram vectors would be zero. However, a drawback of this measure applied on global image histograms is that it is unaware of any structures in the images. As an image quality metric, we specifically compute the χ^2 distance between the global histograms of a synthetic image and its ground truth as follows:

$$\chi^2(\mathbf{h}, \hat{\mathbf{h}}) = \frac{1}{2} \sum_b \frac{[\mathbf{h}_b - \hat{\mathbf{h}}_b]^2}{\mathbf{h}_b + \hat{\mathbf{h}}_b} \tag{3.8}$$

where $\mathbf{h}$ are $\hat{\mathbf{h}}$ are histograms of the ground truth and synthetic images, respectively, and b denotes a histogram bin.

The idea behind using a set of six metrics for image quality assessment is that each of these metrics has its unique strengths and drawbacks and that we hope of them to complement each other, thereby comprehensively capturing the notion of image quality. The same metrics are used to track model generalizability during training as well. During the evaluation of the final fully trained models, this set of metrics is used to derive the performance ranking of the models. One could think of the models as candidates in an election and the metrics as the voters, each of which provides a ranking of the models. In this view, the global performance ranking is then derived using the Borda voting scheme.

3.5.2 Downstream Image Analysis

As clinical analysis of the HX4-PET-syn images, we quantify the predicted tumor hypoxia via two downstream tasks – hypoxic tumor classification and

hypoxic region segmentation. The hypoxic tumor classification task involves classifying a tumor as hypoxic or non-hypoxic and thus yields a relatively "weak" yet robust measure of tumor hypoxia. Knowing the hypoxic status of a tumor during radiotherapy, clinicians can choose to adjust the overall radiation dose to the tumor by increasing the dosage if the tumor is hypoxic. On the other hand, the task of segmenting precisely the exact hypoxic region inside the tumor yields a "stronger", although less robust, hypoxia measurement. When the exact 3D spatial pattern of the high-hypoxia area is known, the radiation beam can be shaped such that a higher radiation dose is delivered to this area. In addition to the two downstream tasks, the MSE and SSIM metrics are computed locally inside the primary tumor to measure the similarity of the predicted tumor hypoxia distribution with the ground truth.

3.6 A Simulated Problem: Depth Estimation from Multimodal Input

Pix2Pix and CycleGAN are general frameworks capable of being applied to a wide range of image translation problems without much modification. On the other hand, the altered design of the CycleGAN-balanced system is meant to be problem-specific and optimized for our use case. Either way, testing the different GAN approaches on medical imaging data is difficult for several reasons – (1) training GANs on 3D images is time-consuming, (2) special visualization software is required to inspect the predicted 3D images, and (3) certain failure modes and artifacts in the medical images can be hard to detect for an untrained observer. Therefore, it is extremely difficult to identify any existing methodological flaws or implementation bugs that may propagate into the results, thereby leading to misinterpretation. To rapidly and effectively prototype, debug, and test the GANs, we construct a simple 2D image-to-image translation task inspired by computer vision literature, that is also representative of the challenges of our medical image translation task. This simulated computer vision task involves estimation of depth from a multimodal input comprised of color photographs and surface normals, and the dataset we use is the ClearGrasp dataset of transparent objects [31].

ClearGrasp is a publicly available dataset [7] for benchmarking deep-learning-based monocular depth estimation (and depth completion) methods in challenging environments where transparent objects are the objects of interest. It contains 50,000 training samples and 532 validation samples representing five different transparent objects. Each sample is a set of synthetic 2D images consisting of different spatially aligned renderings, including a color photograph, a depthmap, a surface normal map, segmentation masks, and object outlines. Of main interest to us are the first three since they essentially correspond to different modalities that contain complementary information. Figure 3.14 visualizes a sample set of these images.

[7] ClearGrasp website: `https://sites.google.com/view/cleargrasp/synthetic-dataset`

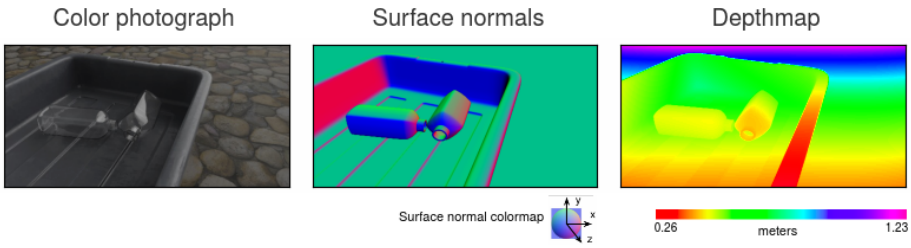

Figure 3.14: Transparent objects are relatively difficult to delineate in the color photograph since their surfaces are not distinctly visible. Furthermore, estimating depth precisely from merely one monocular color photo is a challenging problem by itself. However, given explicitly the information on the surfaces in the scene, estimation of depth of transparent objects becomes much more approachable.

We argue that this depth estimation task can simulate up to a certain extent the challenges of the HX4-PET prediction task. Following is an analysis of the involved modalities and the analogy between the two translation tasks:

1. Color photograph: This modality captures the optical properties of a scene including the shades of colors on various objects and the shadows and reflections cast by the objects. Using these features, one can differentiate one object from the other and can perceive their 3D structure. However, transparent objects are difficult to delineate when relying solely on optical information due to the refractive property of their surfaces. The role of a photograph in this depth estimation task can be viewed as being analogous to the role of pCT in hypoxia map prediction problem, since both the modalities supply rich structural information regarding the overall scene but fail to be informative about one specific region – transparent objects' structure in case of photograph image and tumor biology in case of pCT.

2. Surface normals: Surface normal images encode information about 3D surface angles into a 2D image via a colormap. In the ClearGrasp dataset, surface normal renderings explicitly supply precise surface data, even for transparent objects. Therefore, this modality provides an additional layer of information that complements photographic images. Surface normals, however, fail to differentiate between parallel surfaces even when the surfaces are at different depths from the observer. In contrast, a photograph can be useful in such a situation depending on the lighting conditions. Both modalities are, therefore, mutually complementary. Surface normal image is then analogous to FDG-PET in the sense that both are informative about certain regions in the scene which their respective structural image modalities (color photograph and pCT) fail to capture, although each of them on their own is not sufficient for their respective task.

3. Depthmap: Depthmaps are single-channel 2D images whose pixel intensities represent the distance of the scene elements from the camera and are

the target image modality in the simulated problem task. Depthmaps have a mathematical relationship with surface normals since both are "quantitative" images whose intensities represent related physical properties. In fact, a surface normal image can be directly computed from a given depthmap using gradients of the depth values [8], although the *vice versa* is not possible. For our purpose, the depthmap serves an analogous role to HX4-PET, and in the latter, tumor hypoxia is closely related to metabolism – a relationship that is physiological in nature.

Using the equivalent 2D architecture of their 3D generator and discriminator networks, we apply the same three GAN systems – Pix2Pix, CycleGAN-naive, and CycleGAN-balanced – on this simulated problem task to obtain a prior estimation of their relative performance. In CycleGAN-balanced training, to serve a role that is analogous to ldCT-unreg in the HX4-PET synthesis task, photographic images corrupted with added Gaussian noise are used as the structural context images for their corresponding depthmaps.

A small-scale dataset for this simulated task is curated as a subset of the larger ClearGrasp set. Out of the five types of transparent objects whose images the original dataset contains, we limit our data to one object type which has a relatively simple structure – the "square plastic bottle". From the dataset, we select 25,000 training samples and 100 validation samples corresponding to this object to compose the subset. Each sample consists of a color photograph, a surface normal image, and a depthmap.

[8] More information on computing surface normals from depthmap: `https://bit.ly/3mQS588`.

Chapter 4

Experiments

In experiments 4.1 and 4.2, we consider Pix2Pix as an "upperbound" image translation method, and measure the degree to which the performances of CycleGAN-naive and CycleGAN-balanced approach this upperbound. Pix2Pix is expected to perform better due to its supervised training. We used the Python package *ganslate* [32], our *PyTorch*-based GAN framework for image-to-image translation, to build the complete pipelines for these experiments. In conjunction with this, we used the *Weights and Biases* (*WandB*) online experiment tracking tool [33] to record the training losses, validation metrics, and intermediate outputs. Additional project-specific code, including scripts, utility code, and Jupyter notebooks with step-by-step documentation for dataset preparation and image evaluation, is available in this [1] public repository.

4.1 Experiment 1: Testing the GANs on Depth Estimation Problem

This experiment concerns with testing the GAN systems on the simulated problem of depth estimation from multimodal input comprising of photographic and surface normal images.

4.1.1 Experiment Setup

Training configuration

The curated dataset for the depth estimation task is a subset of the original ClearGrasp dataset (see 3.6). During training, images of all three modalities are first rescaled to size 512×256 and normalized to [-1, 1] range. In Pix2Pix, λ_{elem} value of 100 was observed to produce good results, and λ_{elem} value of 10 is used in both CycleGANs. Similar to the optimizer configuration used in the original Pix2Pix and CycleGAN papers [8, 9], we use Adam optimizer with

[1] `https://github.com/Maastro-CDS-Imaging-Group/HX4-PET-translation`

moment parameter settings β_1=0.5 and β_2=0.999. An initial learning rate of 0.0002 is used for the generators and 0.0001 for the discriminators. We train each model with a batch size of 1 on full images for 50 epochs. The learning rates are fixed for the first 25 epochs and then linearly decayed to 0 over the next 25 epochs. Model validation is performed after each epoch. The training was run on an NVIDIA Tesla P100 SXM2 (16 GB) GPU hardware provided as part of the RWTH Compute Cluster [2], and the training time for Pix2Pix, CycleGAN-naive, and CycleGAN-balanced models was approximately 2 hours, 4 hours, and 5 hours respectively.

Evaluation settings

We use pixel-wise mean-squared error (MSE) between the synthetic and ground truth depthmaps as the evaluation metric. Before computing the MSE, the image intensities are converted back to their original depth representation in meters.

4.1.2 Result and Analysis

Table 4.1 reports the MSE values for each model's predictions on the validation set. Pix2Pix performs the best by a large margin and is also the most robust among the three, as expected, setting the performance upperbound among the three models. CycleGAN-balanced achieves better results as compared to CycleGAN-naive.

Method	MSE
Pix2Pix	**0.004** ± 0.005
CycleGAN-naive	0.192 ± 0.253
CycleGAN-balanced	*0.147* ± 0.170

Table 4.1: Mean-squared error (expressed in meters^2) over the validation subset.

Figure 4.1 shows qualitative results with samples that are representative of the MSE measurements. In general, Pix2Pix predicts the depthmaps accurately with minimal artifacts. The predictions of CycleGAN-naive contain heavy artifacts, including horizontal edges that are observed recurring across multiple images at the same position (a sign of mode collapse) and hallucinated spurious objects. The model doesn't learn the concept of depth. In case of CycleGAN-balanced, the predicted depthmaps contain a greatly reduced amount of artifacts. The model appears to have learned to infer depth, although only in a limited manner – the depth of larger non-transparent objects (such as the tray) is estimated better compared to that of the transparent objects.

To give a concrete example of the conceptual issues with CycleGAN-naive (described earlier in 3.2.2), the outputs produced by both its generators are

[2]RWTH Compute Cluster documentation: `https://help.itc.rwth-aachen.de/service/rhr4fjjutttf/`

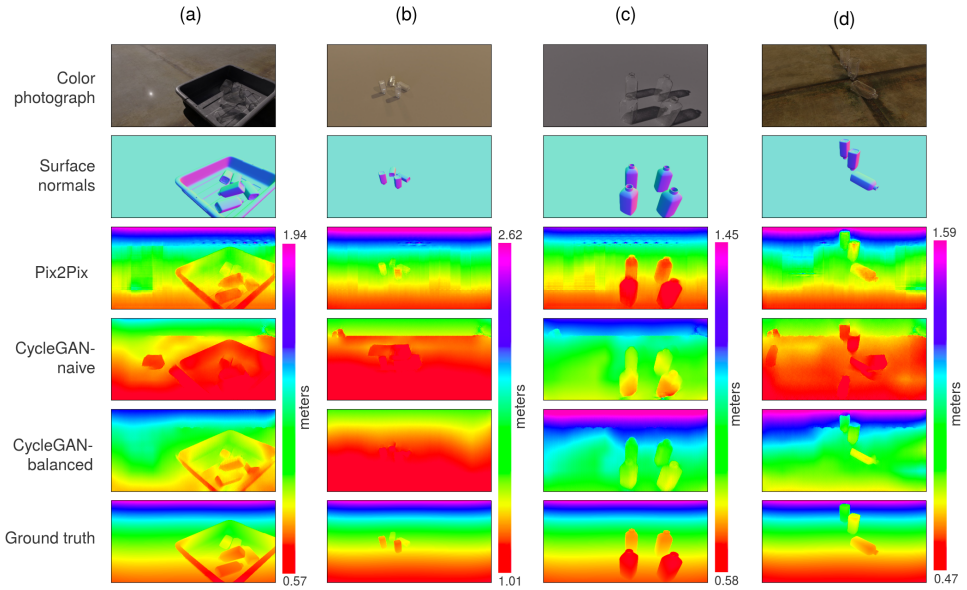

Figure 4.1: Selected validation samples chosen to highlight the challenges in input data — (a) a nearby tray containing transparent objects and a reflection on the floor, (b) distant objects, (c) objects casting strong shadows, and (d) a distinct cross-shaped pattern on the floor. Pix2Pix model estimated the depth of the transparent objects, other objects and the floor very close to the ground truth. CycleGAN-naive produced depthmaps with heavy artifacts, which were absent in case of CycleGAN-balanced, notably in (a) and (d).

visualized and compared with the generator outputs of CycleGAN-balanced. Figures 4.2 and 4.3 show the outputs of the fully-trained CycleGAN-naive and CycleGAN-balanced models, respectively, given an equivalent set of unpaired input images. In Figure 4.2, the generator G_{AB} of CycleGAN-naive in the $A \rightarrow B \rightarrow A$ cycle produces an incorrect depthmap containing spurious objects from the given photographic and surface normal images. G_{BA} then not only uses this flawed depthmap to accurately reconstruct the objects in the original photograph and surface normal image, but also manages to perfectly recover the colors in the photograph. In order to satisfy cycle-consistency, G_{AB} encodes the information about the transparent objects' shape as well as the photograph's optical information into the depthmaps, which G_{BA} uses to almost completely recover the input images, especially the color photograph. The model learns to bypass the content-preservation requirement which cycle-consistency is intended to indirectly encourage. In $B \rightarrow A \rightarrow B$ cycle, the generator G_{BA}, which in previous cycle relied on information encoded as noise in the fake depthmap images, now fails to predict the "correct" photograph and surface normals given a real (and noise-free) depthmap, thereby producing highly noisy images completely devoid of any meaningful features. However, G_{AB} still reconstructs the

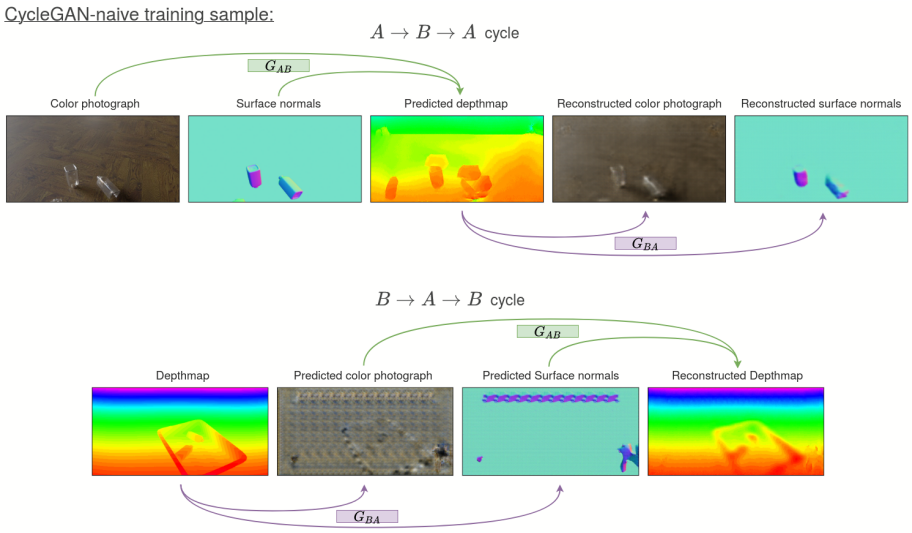

Figure 4.2: Single unpaired training sample and generator outputs of the fully-trained CycleGAN-naive model. In $A \rightarrow B \rightarrow A$ cycle, G_{AB} predicted depthmap with hallucinated structures and yet G_{BA} was able to recover back the photograph and surface normals from it. Similarly, in the $B \rightarrow A \rightarrow B$ cycle, the same G_{BA} failed to compute the "correct" photograph and surface normals from a real depthmap, and yet despite the noisy images it produced, G_{AB} somehow reconstructed the depthmap from them. Here, the training problem is ill-posed and cycle-consistency fails as a result.

depthmap from these noisy images, indicating that the G_{BA} must have encoded relevant information, possibly as noise, into its output. Cycle-consistency is bypassed in this case as well. The problem arises because although the task of learning G_{AB} might be solvable, learning its inverse (G_{BA}) isn't, since such a unique inverse mapping doesn't exist. The learning problem as formulated for the CycleGAN-naive model is ill-posed, in the sense that there is no solution to it, and therefore, the training would fail inevitably.

On the other hand, the outputs of CycleGAN-balanced generators shown in Figure 4.3 highlight the system's improved capability of learning the relevant mappings. Here, the main participant modalities of the translation task are the surface normal image and the depthmap. Neither of the generator tasks involves synthesizing the color photographs from the less information-rich modalities. Using the given photographic images always as an input component for the generators provides them shared contextual information within each cycle. The unpaired learning problem for CycleGAN-balanced is better formulated and solvable, and hence the training process can yield a suitable translation model.

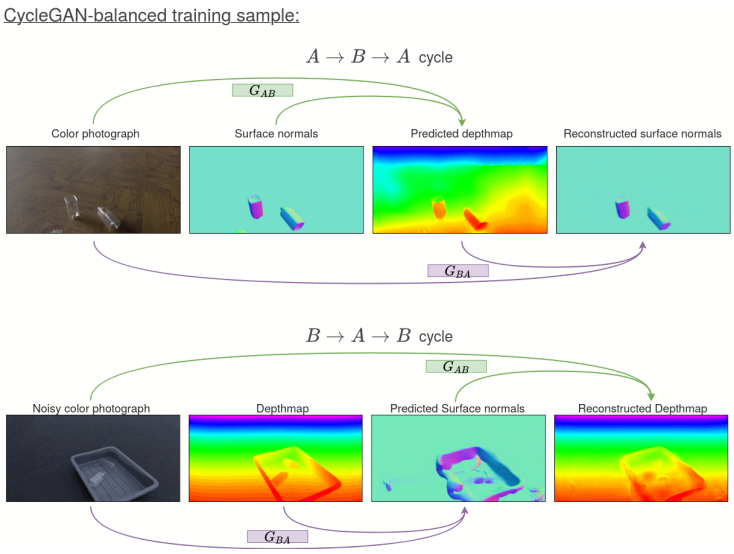

Figure 4.3: Single unpaired training sample and generator outputs of the fully-trained CycleGAN-balanced model. In both cycles, the translation task is mainly across surface normals and depthmaps, while the available photograph serves as a common contextual information for prediction and reconstruction. In $\hat{c}$ycle, the predicted surface normal image has a deformed structure. However, the reconstructed depthmap reflects this distortion and thus the cycle-consistency constraint can act on correcting it. Here, the training problem is well defined and solvable.

4.2 Experiment 2: Image Quality Metrics For Synthetic HX4-PET Assessment

This experiment focuses on evaluation of the GAN systems for the HX4-PET synthesis task. Global quality of the HX4-PET-syn images is assessed using general metrics of image quality and similarity. Additionally, the applicability of these metrics in tracking model convergence during training is explored.

4.2.1 Experiment Setup

Training configuration

The dataset used in this experiment is the Maastro Lung HX4-PET dataset after undergoing the preparation steps described in 3.1.3. The patch-based training pipeline described in 3.4.2 is used for all model training. Because a the models are trained on sets of patches extracted from the images using stochastic strategies, the notion of an "epoch" doesn't exist here. Instead, the number of iterations is used as a measure of training duration, where an iteration is composed of sampling a set of image patches from the full images, performing the required forward passes through the networks, and updating the weights

of all the networks in the GAN system. We train each GAN model for 60,000 iterations. The scaling factor λ_{elem} for the element-wise loss in Pix2Pix is set to 10, and a λ_{cyc} of 10 is used for the cycle-consistency loss in the CycleGANs. Similar to experiment 4.1, Adam optimizer is used with settings β_1=0.5 and β_2=0.999. Initial learning rates 0.0002 for the generators and 0.0001 for the discriminators are used during the first 30,000 iterations followed by linearly decaying the learning rate values to 0 over the remaining 30,000 iterations. Model validation is performed every 1000 iterations. All models are trained with batch size 1.

Training was performed on an NVIDIA Tesla V100 SXM2 (32 GB) GPU provided by Data Science Research Infrastructure (DSRI) [3] of Maastricht University. The 3D training of the Pix2Pix, CycleGAN-naive, and CycleGAN-balanced models took 9.5 hours, 13 hours, and 15.25 hours, respectively.

Evaluation settings

The set of six image quality and similarity metrics described in 3.5.1 are used to assess the quality of HX4-PET-syn images, given the corresponding ground truth HX4-PET-reg images. The intensity scale is divided into 100 bins while computing NMI (for entropy calculation) and global histogram distance.

Additionally, a systematic visual inspection of the synthetic images is performed to identify failure modes that are typical to each GAN model. This would also enable validating the assessment of the image quality metrics. For each patient in the validation set, each model's predicted HX4-PET-syn images were visualized and inspected along all three axes – axial, sagittal and coronal – using the *3D Slicer* visualization software [34] under a window of [0, 1.3] SUV. This window is chosen so that the most of the variation in image intensities could be covered. The following set of image degradation criteria are defined to guide the visual inspection:

1. *Presence of background noise patterns:* It is well known that checkerboard artifacts can occur in GAN generated images usually due to the the nature of the transposed convolution operation used in upsampling layers [35]. This criterion aims at judging the images based on the presence of such periodic noise patterns in the region outside the body where the hypoxia signal is supposed to be zero. This criterion would be satisfied only when severe amount of noise is present such that it is readily visible to the observer in most slices along all three axes of the 3D image.

2. *Presence of noise patterns in the body region:* Checkerboard and other periodic noise patterns in the background may be absent in better performing models because all models were trained on body-masked images whose background intensities were set to zero signal value (0 SUV in PET and -1024 HU (air) in CT). Regardless, such noise can also possibly occur *within* the body region, i.e. the foreground of the image. This criterion

[3] DSRI documentation: `https://maastrichtu-ids.github.io/dsri-documentation/`

aims to account for this, and to satisfy it the noise must be severe enough to be easily detectable by the observer, must be wide spread in roughly the central parts of the body area (i.e. away from boundary regions) and recurring across multiple slices along all three image axes.

3. *Presence of hallucinated structures:* Generated scans can possibly contain structures which wouldn't be characteristic of human physiology, which might have been constructed by the model to increase the resemblance of the image to the target distribution samples. Many of such structures could be conspicuous enough to be noticeable to non-experts, however others would be too challenging even for trained professionals to discover. This is because unlike CT which shows anatomical structure in high detail, PET images, and especially HX4-PET, show only vague and diffused forms. The synthetic HX4-PET is bound to have at least some differences in these forms as compared to the ground truth, and therefore it is difficult to draw a line past which the differences can be deemed a hallucinated object. Hence, this criterion focuses on the former, more conspicuous, type of artificial structures. For example, isolated high-intensity objects with globular or ring-like shape, similar to common tumor hypoxia signatures, in regions where they shouldn't exist. Using the ground truth as reference, detection of even a single such structure in a predicted image would satisfy this criterion.

4. *Breaks in the existing structure:* Abrupt breaks in structure, for example, holes and gaps with sharp edges, inside the patient's body region in the generated images can occur when the model has not learned relevant anatomical features. Many such glitches can be easily detected by an untrained observer, and this criterion aims to account for their presence in the synthetic HX4-PET images. One or more severe gaps of such type in the image foreground would satisfy this criterion.

4.2.2 Result and Analysis 1: Evaluating Fully Trained Models

Performance of the three 3D models as measured by the six image metrics is reported in Table 4.2. Pix2Pix emerges to be the best performing model as evaluated by all six metrics, followed by CycleGAN-balanced whose performance values are very close to those of Pix2Pix. CycleGAN-naive performs the worst, as agreed by all the metrics. Borda counting based on the individual performance rankings of the metrics resulted in the same global ranking on the models – Pix2Pix as the best model, followed by CycleGAN-balanced and finally CycleGAN-naive.

Results of the visual inspection are reported in Table 4.3. Figure 4.4 compares the predictions of the models for different patients and highlights failure modes typical to each model.

None of Pix2Pix predictions contained severe background noise patterns, as opposed to the predicted images from CycleGAN-naive, all of which were

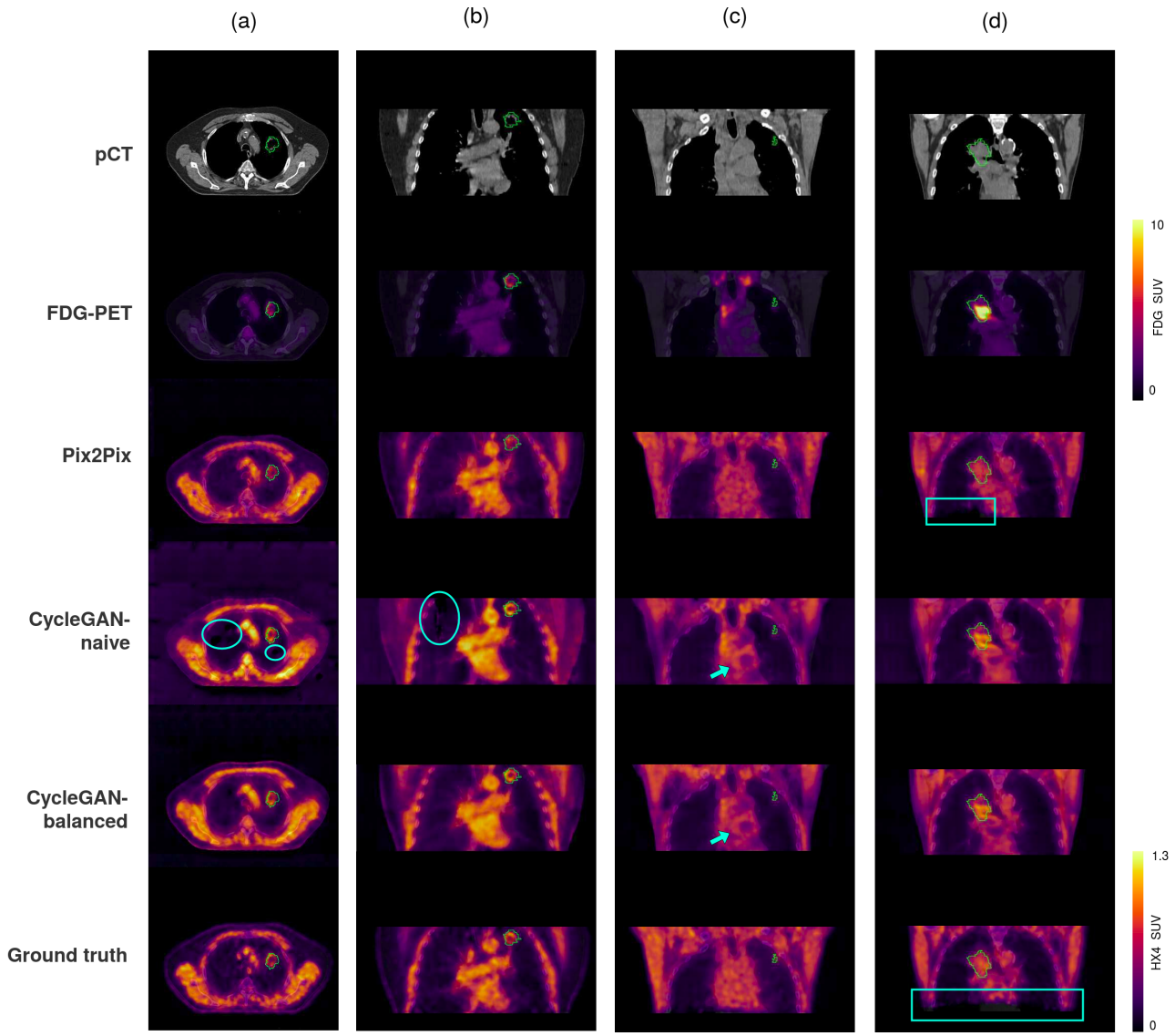

Figure 4.4: Corresponding slices of the inputs, the ground truth and the predictions. All PET images are overlaid on the corresponding pCT. Gross tumor volumes (GTV) are delineated with green contours. Severe background noise is visible in CycleGAN-naive outputs, which also penetrates into the foreground as seen in (a). Sharp edged holes in (a) and a rift in (b) (cyan ellipses), a common occurrence in CycleGAN-naive predictions. In (d), the Pix2Pix output shows structure break at the bottom likely caused due to model being trained on faulty ground truth (cyan bounding boxes). In (c), the outputs of CycleGAN-naive and CycleGAN-balanced have a hallucinated ring-shaped hypoxia signature near the heart (cyan arrows).

Method →	Pix2Pix	CycleGAN-naive	CycleGAN-balanced
MSE	**0.009** ± 0.006	0.043 ± 0.004	*0.016* ± 0.008
MAE	**0.049** ± 0.008	0.185 ± 0.006	*0.067* ± 0.010
PSNR	**27.467** ± 1.946	20.217 ± 2.032	*25.119* ± 1.812
SSIM	**0.582** ± 0.086	0.190 ± 0.044	*0.5* ± 0.07
NMI	**1.215** ± 0.016	1.099 ± 0.009	*1.181* ± 0.015
Histogram distance	**0.563** ± 0.160	1.289 ± 0.201	*0.662* ± 0.124
Performance ranking	**1**	3	*2*

Table 4.2: Performance on the Maastro Lung HX4-PET validation set. Best and second-to-best values are highlighted with bold and italics font, respectively. Performance of the unpaired CycleGAN-balanced model reaches close to that of the paired Pix2Pix model. The final performance ranking is calculated by combining each ranking produced by metric, with Borda count.

Method →	Pix2Pix	CycleGAN-naive	CycleGAN-balanced
Background noise	**0%** (0/19)	100% (19/19)	*42.1%* (8/19)
Foreground noise	**26.3%** (5/19)	100% (19/19)	*42.1%* (8/19)
Hallucinated structures	**10.5%** (2/19)	57.9% (11/19)	*31.6%* (6/19)
Broken structures	**57.9%** (11/19)	100% (19/19)	*89.5%* (17/19)

Table 4.3: Fraction of the total validation set images meeting the corresponding image degradation criteria. Best and second-to-best values are highlighted with bold and italics font, respectively.

severely corrupted by noise in both background and foreground. CycleGAN-balanced stands almost half-way between the other two models in terms of background and foreground noise, although none of its outputs contained noise as severe as in case of CycleGAN-naive. The background noise is easily visible in Figure 4.4a. The presence of greater noise in CycleGAN-naive is likely to be related to its conceptual problem related to non-invertibility. The $A \rightarrow B$ translation involves predicting hypoxia map from pCT and FDG-PET, which can have a unique solution. Whereas its reverse, the $B \rightarrow A$ translation, involves computing ideally exact pCT and FDG-PET given only the hypoxia map information, which is not possible unless the model encodes the extra information, as noise or artifacts, in the hypoxia image.

Structural breaks and holes were observed in a majority of all models' outputs. These were, however, more severe and numerous in case of CycleGAN-naive and CycleGAN-balanced, more in the former than in the latter. Pix2Pix had comparably lower number of these faults. A notable difference is that while many of such structural faults appeared in roughly the middle and upper regions of the body in predicted images in case of the two CycleGANs, most of them in Pix2Pix predictions were located in the bottom-most axial slices. These are shown in Figures 4.4a, 4.4b and 4.4d. A likely cause of this is the registration imperfections in the ground truth HX4-PET-reg images, many of which had missing signal in parts of the bottom axial slices. Pix2Pix training, which assumes voxel-to-voxel similarity across input and output images, might have been affected by this training noise resulting in a model that couldn't ac-

curately predict the hypoxia patterns in the bottom slices. On the other hand, the CycleGAN models which were trained on unpaired and unregistered HX4-PET-unreg images did not exhibit structural breaks of this kind, and in fact, during validation, were able to predict plausible hypoxia patterns in the bottom slices of their predictions whose corresponding ground truths themselves were faulty. This is visible in Figure 4.4d. The fact that the validation set's ground truth HX4-PET-reg images have this issue of missing signal may create another problem – if the reference image itself is partly incomplete in some region and if a model predicts a plausible hypoxia signal in that region of its predicted image, evaluating the image quality, especially using the voxel-wise difference metrics, can result in an incorrect assessment. However, this may not pose a serious issue here because first, only a small portion of the signal in entire ground truth image is missing, and second, the models being compared have significant differences in performance as the visual inspection revealed.

Finally, the type of image degradation related to the presence of hallucinated structures was observed more commonly in case of CycleGAN-naive and CycleGAN-balanced as compared to Pix2Pix. This was mainly characterized by the insertion of ring-shaped hypoxia signatures in regions near the heart and the upper abdomen. However, such signatures are a common feature only within the tumor where parts of its periphery are actually hypoxic. A likely reason of this pattern to occur in other regions in the predictions of the two CycleGANs is the failure to learn relevant anatomical features from the pCT and depending more heavily on features related to metabolic activity from FDG-PET in these regions. The Pix2Pix model trained directly in a supervised manner mostly managed to avoid this issue. This can be seen in Figure 4.4c.

In summary, the results of the visual inspection reflect the quantitative evaluation results. Pix2Pix produced synthetic HX4-PET images that are visually highly similar to the ground truth, mostly owing to its supervised training. However, because it was trained in the supervised manner on imperfect ground truth images, it also produced locally specific structure breaks in its predictions. Among the unpaired GAN systems, CycleGAN-balanced showed great improvement over CycleGAN-naive and was able to greatly reduce degradation caused by noise and hallucinated structures.

4.2.3 Result and Analysis 2: Analyzing Model Convergence during Training

Adversarial training of CycleGAN

Informally, model convergence refers to a case where the machine learning model's loss approaches a certain optimal value with a decreasing rate over the training iterations. In GAN training, two networks are simultaneously optimized with conflicting objectives. In theory, a Vanilla GAN model asymptotically converges to an equilibrium state where the generator perfectly models the data distribution and the discriminator's prediction over real and fake data is 0.5 on average, i.e. it can only do random guessing [14]. This stands, of

course, only under the condition that the discriminator is trained to optimality for each iteration of the generator's update. In practice, the discriminator is updated just once per generator update, but since the discriminator's task of binary classification is simpler than the generator's task of image synthesis, the discriminator is likely to learn faster and remain competitive with the generator in the corresponding iterations. In this part of the experiment, GAN convergence is analyzed empirically for the CycleGAN-naive and CycleGAN-balanced systems, by reasoning based on their recorded training losses.

The least-squares adversarial loss used in our CycleGAN models is shown in Equation 4.1.

$$\begin{aligned} L(G) &= E_{x\sim p_{data}(x)}[(1 - D(G(x)))^2] \\ L(D) &= E_{x\sim p_{data}(x)}[D(G(x))^2] + E_{y\sim p_{data}(y)}[(1 - D(y))^2] \end{aligned} \tag{4.1}$$

where G and D represent a generator (G_{AB} or G_{BA}) and the corresponding discriminator (D_B or D_A), respectively. x represents an image from the input domain of the generator G, and y represents an image from its target domain. This general representation is used here for simplicity.

During training, G and D are updated in an alternating manner. As the training progresses and if the training is stable, both G and D improve at a similar rate maintaining a competitive behavior. Depending on the extent of the stability, losses of both models either stay constant reaching an equilibrium or oscillate around the equilibrium values. Such an equilibrium state is reached when, after every update, G's output images are good enough to force D into random guessing and predicting a validity value of 0.5 on average for both real and fake images. Then according to Equation 4.1, the equilibrium state loss values $L(G^*)$ and $L(D^*)$ are 0.25 and 0.5, respectively.

Figure 4.5 shows the adversarial training loss plots for CycleGAN-naive and CycleGAN-balanced. The aforementioned general reasoning is now applied independently to the two generator-discriminator pairs of the CycleGAN system. Consider the $A{\rightarrow}B$ direction first. The adversarial loss $L(G_{AB})$ of CycleGAN-balanced was closer to the equilibrium value than the that of CycleGAN-naive was, whose G_{AB} was much more unstable and diverged away more significantly from around $25{,}000^{th}$ iteration. This effect relates to the corresponding discriminator D_B of CycleGAN-naive starting to overpower its G_{AB} at around the same period. Discriminator D_B of CycleGAN-balanced was maintained more stably around its equilibrium loss value of 0.5 over the full training period. Now looking at the $B{\rightarrow}A$ direction, a similar pattern is observed. Except here, the generator-discriminator pair G_{BA} and D_A of CycleGAN-balanced appeared to destabilize and deviate away between $40{,}000^{th}$ and $50{,}000^{th}$ iteration, although eventually regaining stability near the end of the training.

Additionally, cycle-consistency in CycleGAN-balanced was maintained to a greater extent almost throughout the training period as compared to CycleGAN-naive. The desirable training characteristics shown by the CycleGAN-balanced system can be attributed to its design modification.

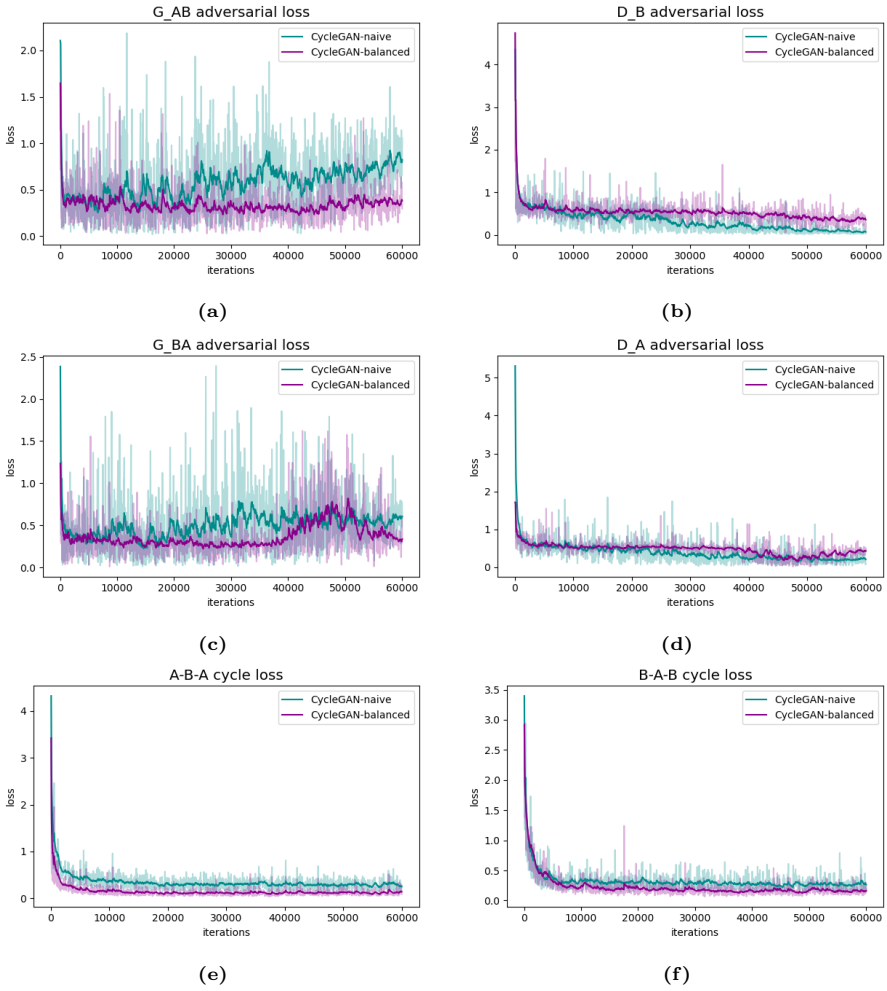

Figure 4.5: Adversarial and cycle consistency loss curves for the two CycleGAN models. Exponential moving average smoothing was applied to each to show the general trend.

Generalization performance

Next, we examine whether the same trend as was observed in the training convergence parameters is also reflected in the models' generalization performance on the validation set. Figure 4.6 plots the six validation metrics for the CycleGAN models over the training period. The voxel-wise difference metrics – MSE, MAE and PSNR – show a generally consistent improvement in CycleGAN-balanced reflecting its relatively stable training. However, the metrics based on image intensity statistics – SSIM, NMI and histogram χ^2 distance – show a brief, yet prominent, performance dip in the model halfway into the training

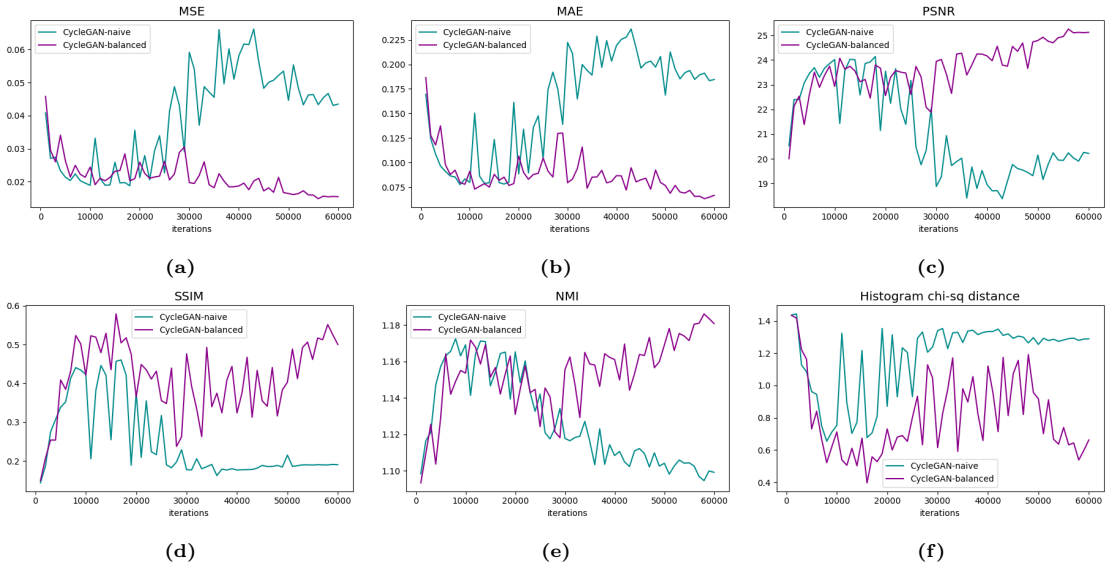

Figure 4.6: Validation metrics for the two CycleGAN models over the training period.

suggesting a temporary drop in its generalizability. CycleGAN-naive followed a generally similar trend as CycleGAN-balanced up until approximately 25,000th iteration – i.e. gradual improvement in the beginning followed by a performance dip. Though, contrary to CycleGAN-balanced, it was unable to recover its generalizability and continued to deviate. This correlates with the significant instability and divergence of the CycleGAN-naive model observed in the training convergence parameters at about the same period in the training process.

In summary, the image quality metrics applied on validation data were indeed informative about model convergence. Additionally, since some of them measured different properties of the generated images, they can be helpful in effectively tracking model generalizability.

4.3 Experiment 3: Application-specific Downstream Tasks

This experiment focuses on clinically focused evaluation of the synthetic hypoxia PET images by quantifying hypoxia inside the gross tumor volume (GTV).

4.3.1 Experiment Setup

HX4-PET-syn images are computed for each patient in the validation set using the fully trained models, their units converted to SUV scale, and stored as NRRD files. The images are then loaded and analyzed. Each HX4-PET-syn

image and its corresponding ground truth HX4-PET-reg are first cropped to a bounding box containing just the GTV. The GTV mask is then applied to the images, and intensities outside the mask are set to 0. Next, the GTV SUV values are converted to the tumor-to-background ratio (TBR) by dividing each GTV voxel's SUV with $SUV_{aorta-mean}$. From this point onward, the different hypoxia quantification metrics are computed separately as follows:

1. *MSE-GTV and SSIM-GTV:* MSE and SSIM are calculated only on the GTV voxels using array masking.

2. *Hypoxic tumor classification:* One of the ways of classifying the tumor as hypoxic or non-hypoxic is by calculating the total physical volume of the hypoxic region in the GTV, known as hypoxic volume (HV), and applying a threshold. We calculate the HV by first performing point-wise intensity thresholding using standard TBR threshold of 1.4 (Zegers et al. [11]) to obtain a binary image of hypoxic voxels followed by calculating the total physical volume (in mm^3) occupied by them. Then, an HV threshold of 1 cm^3 is applied beyond which the tumor is classified as hypoxic, similar to Even et al. [6]. Tumor classification is performed for each predicted image and its corresponding ground truth, and the mean accuracy is derived.

3. *Hypoxic region segmentation:* The 3D hypoxic region in the GTV is segmented by applying the standard TBR threshold of 1.4. This is performed for each predicted image and its corresponding ground truth image. Then, the Dice Similarity Coefficient (DSC) is used to measure the overlap between the hypoxic regions.

4.3.2 Result and Analysis

Table 4.4 reports the mean values of the hypoxia quantification measures over the validation set. The quantitative results provide no conclusive evidence on the model performances and are, in fact, slightly contradictory to previous results. MSE-GTV and SSIM-GTV were very similar for all the models, and Pix2Pix was less competent in accurately predicting the hypoxic regions. CycleGAN-balanced didn't show an improvement over CycleGAN-naive here. A visualization of the predicted tumor hypoxia patterns is shown in Figure 4.7. Two important observations can be made are. First, the Pix2Pix outputs show significant checkerboard-like noise patterns. Note that these are of different nature compared the image-wide foreground noise defined as a failure criterion earlier in experiment 4.2 in that these are of higher frequency and are localized to high-intensity regions, such as the tumor locality shown in the figure. Predictions from the CycleGANs show minimal noise. Second, Pix2Pix predictions do not contain sufficiently high intensity regions to be capable of being segmented and their spatial patterns do not match the hypoxic patterns in the ground truth. Since only a small number of voxels carry sufficiently high intensity values, there existed many false-negative voxels which collectively amounted to many hypoxic tumors being misclassified as non-hypoxic. In case of the CycleGANs,

their predicted hypoxic patterns are more similar to the metabolic patterns from FDG-PET, instead of the ground truth.

Method →	Pix2Pix	CycleGAN-naive	CycleGAN-balanced
MSE-GTV	**0.067** ± 0.040	0.079 ± 0.046	*0.068* ± 0.042
SSIM-GTV	0.870 ± 0.082	*0.884* ± 0.066	**0.886** ± 0.066
Tumor classif. accuracy	61.2% (12/19)	**78.9**% (15/19)	*73.7%* (14/19)
Hypoxic region seg. Dice	0.058 ± 0.092	**0.141** ± 0.184	*0.127* ± 0.174

Table 4.4: Results of tumor hypoxia quantification. Best and second-to-best values are highlighted with bold and italics font, respectively. For tumor classification, the accuracy is given in percentage values and fraction of patients with correctly classified tumors.

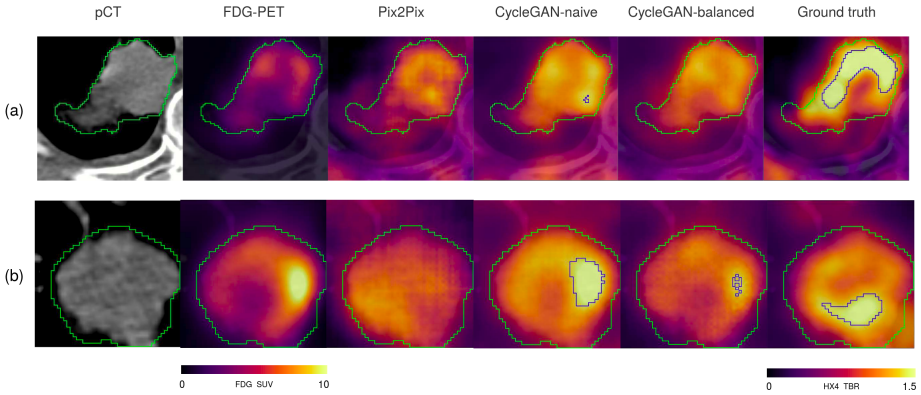

Figure 4.7: Tumor hypoxia patterns generated by the models. All PET images are overlaid on the corresponding pCT. GTV is delineated with green contour, and the hypoxic region in each of three models' prediction and the ground truth is marked with blue contours. In both examples, Pix2Pix predictions didn't contain sufficiently intense regions, hence no hypoxic areas are seen segmented in these slices.

One of the reasons Pix2Pix fails here appears to be related to its supervised training. The ground truth images contain registration errors that are more pronounced at the small scale in which this evaluation is performed. The fine hypoxia patterns in the ground truth, including those within the tumor, might have shifted post-registration causing misalignment with the CT and FDG-PET features, thereby inducing noise in the paired training data. Therefore, although the Pix2Pix model could produce synthetic hypoxia images of an overall higher quality, these images reveal large inaccuracies when examined at the scale of the tumor. In case of the two CycleGANs, the tumor hypoxia patterns matched the spatial pattern of FDG uptake suggesting that the models learned to excessively depend on FDG-PET features. Another reason for poor clinical performance of all three models might be the small size of the training dataset. Even when using a patch-based training approach, the number of sufficiently varied training samples was very low.

Chapter 5

Discussion

In experiments 4.1 and 4.2, we observe that the paired Pix2Pix approach can generate higher quality images as compared to unpaired CycleGAN, although at the cost of acquiring and preparing the ground truth images. Among the two CycleGAN systems, CycleGAN-balanced was able to circumvent the non-invertibility issue of CycleGAN-naive due to its design modification, and produced better results as indicated by both quantitative and qualitative analyses. Our quantitative evaluation of the synthetic HX4-PET images involved using a set of six metrics, of which three – MSE, MAE, and PSNR – measure the voxel-level accuracy of the synthetic images with respect to the ground truth, whereas the remaining three – SSIM, NMI, and histogram distance – account for the fidelity of image structure and statistics. Furthermore, each metric has its unique property and since the "quality" of an image is a multifaceted attribute, using a population of such metrics for image assessment can capture the image quality more effectively as compared to using a subset of them. We thereby address our first research question. In 4.2.3, we perform an analysis of CycleGAN training losses over the training period and observe that the CycleGAN-balanced model was more stable and displayed better convergence properties as opposed to the CycleGAN-naive model, which diverged away from its objectives. The validation metrics collectively reflected similar trends as the convergence parameters while also being informative about model generalizability, thereby answering our third research question. Finally, through experiment 4.3, we address our second research question by identifying and performing clinically relevant downstream tasks to determine the clinical value of our synthetic HX4-PET images. Tumor hypoxia measurement derived from synthetic images from Pix2Pix showed remarkably poor accuracy compared with the two CycleGANs. While a majority of the tumors in the synthetic images produced by the CycleGAN models were classified correctly, the classification rate in Pix2Pix predictions is lower. However, as indicated by a poor segmentation score, none of the models could predict accurately the spatial distribution of high hypoxia. The primary reason could be the lack of sufficient training data, and the second reason, specifically for Pix2Pix, could be the noise in its supervision signal caused due to imperfect

ground truth images. We believe that given sufficient training data, unpaired models like CycleGAN, especially the modified version of it, could be more suitable for the HX4-PET synthesis task. As supplementary material, we make available the *WandB* training reports for the depth-estimation task (experiment 4.1) [1] and the HX4-PET synthesis task (experiment 4.2) [2] which include training losses, validation metrics and intermediate outputs of the models.

There are several limitations of this thesis that can provide opportunities for future work. First, the small size of our Maastro Lung HX4-PET dataset was a serious limitation on model training and validation. A larger training dataset would allow training better translation models and a sufficiently large and diverse validation dataset would enable more effective evaluation of the models. Second, it was observed that both the CycleGAN models predicted tumor hypoxia patterns that closely resembled the FDG uptake signatures rather than the actual spatial distribution of hypoxia. The models might have learned to heavily rely on FDG-PET features while ignoring CT information. It would be interesting to verify this by performing an ablation study on these models. This could be conducted, for instance, by performing inference with the fully trained models with the CT signal set to zero value and observing the generator outputs. Third, the quality assessment of the synthetic HX4-PET images was performed using a set of six general-purpose image quality and similarity metrics that were logically chosen. However, this portfolio of evaluation metrics can be improved. It would be valuable to consult theoretical literature on image metrics and understand in a greater depth their mathematical properties. This would help in selecting more suitable and diverse metrics thereby improving the comprehensiveness of image assessment. Fourth, as an extension of the simple automated clinical evaluation performed in experiment 4.3, a version of the Turing test can be performed by presenting the synthetic HX4-PET images to a radiation oncologist, with a certain clinically relevant goal, for instance, determining the prognostic value of the synthetic images.

[1] Training report for the depth-estimation task: `https://bit.ly/3x452Qi`
[2] Training report for the HX4-PET synthesis task: `https://bit.ly/35Y8dx2`

Chapter 6

Conclusion

In this work, we formulated the problem of predicting hypoxia from FDG-PET and CT scans as image-to-image translation and investigated image translation GANs for synthesizing full HX4-PET images from the multimodal input. Using the paired Pix2Pix and the unpaired CycleGAN, we observed that the paired approach produces superior quality images both in our simulated translation task and in the HX4-PET synthesis task. We argued that the naive application of CycleGAN to the HX4-PET synthesis task has a conceptual flaw related to non-invertibility of the inter-domain mappings, and proposed an alternative strategy to circumvent this problem while still relying on unpaired training data. This modified CycleGAN system showed substantial performance improvement over the default CycleGAN in terms of image quality. To assess the quality of the synthetic HX4-PET images in a comprehensive manner, we constructed a set of six measures of image quality and image similarity that measure broadly two different aspects of the images – voxel-wise accuracy and (local and global) image statistics. A systematic visual inspection of the synthetic HX4-PET images validated the assessment of these metrics and also revealed common failure modes for each model. Images from Pix2Pix contained, on average, the least amount of degradation and artifacts, although they suffered from structural breaks in specific areas. These faults can be attributed to the supervised training of Pix2Pix which used ground truth images containing registration imperfections. Despite the clear image-quality-based performance differences across the three image translation models, the clinical evaluation of their synthetic images conducted via tumor hypoxia quantification tasks produced mixed and inconclusive results. These results were overall unsatisfactory indicating the insufficiency of the models in meeting the clinical requirements. The potential for clinical implementation of these image translation methods needs to be further investigated using larger and more diverse training and validation datasets. Between the paired and unpaired translation approaches, the latter could be more suitable due to their considerably more lenient data requirements and their lack of direct dependence on spatially aligned ground truth images that makes them immune to issues caused by misaligned training data.

Bibliography

[1] Michael Hockel and Peter Vaupel. Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects. *Journal of the National Cancer Institute*, 93(4):266–276, 2001.

[2] Barbara Muz, Pilar de la Puente, Feda Azab, and Abdel Kareem Azab. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia*, 3:83, 2015.

[3] Peter Vaupel and Arnulf Mayer. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer and Metastasis Reviews*, 26(2):225–239, 2007.

[4] Ian N Fleming, Roido Manavaki, Philip J Blower, Catharine West, Kaye J Williams, Adrian L Harris, Juozas Domarkas, Simon Lord, Claire Baldry, and Fiona J Gilbert. Imaging tumour hypoxia with positron emission tomography. *British journal of cancer*, 112(2):238–250, 2015.

[5] Sebastian Sanduleanu, Alexander Wiel, Relinde IY Lieverse, Damiënne Marcus, Abdalla Ibrahim, Sergey Primakov, Guangyao Wu, Jan Theys, Ala Yaromina, Ludwig J Dubois, et al. Hypoxia pet imaging with [18f]-hx4—a promising next-generation tracer. *Cancers*, 12(5):1322, 2020.

[6] Aniek JG Even, Bart Reymen, Matthew D La Fontaine, Marco Das, Arthur Jochems, Felix M Mottaghy, José SA Belderbos, Dirk De Ruysscher, Philippe Lambin, and Wouter van Elmpt. Predicting tumor hypoxia in non-small cell lung cancer by combining ct, fdg pet and dynamic contrast-enhanced ct. *Acta Oncologica*, 56(11):1591–1596, 2017.

[7] Sebastian Sanduleanu, Arthur Jochems, Taman Upadhaya, Aniek JG Even, Ralph TH Leijenaar, Frank JWM Dankers, Remy Klaassen, Henry C Woodruff, Mathieu Hatt, Hans JAM Kaanders, et al. Non-invasive imaging prediction of tumor hypoxia: A novel developed and externally validated ct and fdg-pet-based radiomic signatures. *Radiotherapy and Oncology*, 153:97–105, 2020.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[10] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[11] Catharina ML Zegers, Wouter van Elmpt, Roel Wierts, Bart Reymen, Hoda Sharifi, Michel C Öllers, Frank Hoebers, Esther GC Troost, Rinus Wanders, Angela van Baardwijk, et al. Hypoxia imaging with [18f] hx4 pet in nsclc patients: defining optimal imaging parameters. *Radiotherapy and Oncology*, 109(1):58–64, 2013.

[12] Catharina ML Zegers, Wouter Van Elmpt, Bart Reymen, Aniek JG Even, Esther GC Troost, Michel C Öllers, Frank JP Hoebers, Ruud MA Houben, Jonas Eriksson, Albert D Windhorst, et al. In vivo quantification of hypoxic and metabolic status of nsclc tumors using [18f] hx4 and [18f] fdg-pet/ct imaging. *Clinical Cancer Research*, 20(24):6389–6397, 2014.

[13] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):1–9, 2014.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[16] Anirudh Chandrashekar, Ashok Handa, Natesh Shivakumar, Pierfrancesco Lapolla, Vicente Grau, and Regent Lee. A deep learning approach to generate contrast-enhanced computerised tomography angiography without the use of intravenous contrast agents, 2020.

[17] Johannes Haubold, René Hosch, Lale Umutlu, Axel Wetter, Patrizia Haubold, Alexander Radbruch, Michael Forsting, Felix Nensa, and Sven Koitka. Contrast agent dose reduction in computed tomography with deep learning using a conditional generative adversarial network. *European Radiology*, pages 1–9, 2021.

[18] Avi Ben-Cohen, Eyal Klang, Stephen P. Raskin, Shelly Soffer, Simona Ben-Haim, Eli Konen, Michal Marianne Amitai, and Hayit Greenspan. Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection, 2018.

[19] Lei Bi, Jinman Kim, Ashnil Kumar, Dagan Feng, and Michael Fulham. Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans). In *molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment*, pages 43–51. Springer, 2017.

[20] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[21] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. Mri cross-modality neuroimage-to-neuroimage translation, 2018.

[22] Per Welander, Simon Karlsson, and Anders Eklund. Generative adversarial networks for image-to-image translation on multi-contrast mr images - a comparison of cyclegan and unit, 2018.

[23] Wouter van Elmpt, Dirk De Ruysscher, Anke van der Salm, Annemarie Lakeman, Judith van der Stoep, Daisy Emans, Eugène Damen, Michel Öllers, Jan-Jakob Sonke, and José Belderbos. The pet-boost randomised phase ii dose-escalation trial in non-small cell lung cancer. *Radiotherapy and Oncology*, 104(1):67–71, 2012.

[24] Darcy Mason. pydicom: An open source DICOM library. `https://github.com/pydicom/pydicom`, 2018. [Online; accessed July-06-2021].

[25] Bradley Christopher Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of simpleitk. *Frontiers in neuroinformatics*, 7:45, 2013.

[26] Vincent Andrearczyk, Valentin Oreiller, Mario Jreige, Martin Vallières, Joel Castelli, Hesham Elhalawani, Sarah Boughdad, John O Prior, and Adrien Depeursinge. Overview of the hecktor challenge at miccai 2020: automatic head and neck tumor segmentation in pet/ct. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pages 1–21. Springer, 2020.

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[28] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[29] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 174–182. Springer, 2018.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[31] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642. IEEE, 2020.

[32] Ibrahim Hadzic, Suraj Pai, Chinmay Rao, and Jonas Teuwen. ganslate-team/ganslate: Initial public release, September 2021.

[33] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[34] Ron Kikinis, Steve D Pieper, and Kirby G Vosburgh. 3d slicer: a platform for subject-specific image analysis, visualization, and clinical support. In *Intraoperative imaging and image-guided therapy*, pages 277–289. Springer, 2014.

[35] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.