MELBA Journal Machine Learning for Biomedical Imaging

Learning disentangled representations for unpaired synthesis of high-resolution dynamic MRI

Claire **Scavinner-Dorval** [1], Rodolphe **Bailly** [2], Bhushan **Borotikar** [3], Sylvain **Brochard** [4], Douraied **Ben Salem** [4], François **Rousseau** [1]

1 IMT Atlantique, LaTIM UMR 1101 INSERM, Brest, France
2 Fondation Ildys, LaTIM UMR 1101 INSERM, Brest, France
3 Symbiosis Centre for Medical Image Analysis, Pune, India
4 Univ Brest, CHU Brest, UMR 1101, F 29200 Brest, France

Abstract

Image-to-image (I2I) synthesis aims at learning the mapping between two visual domains. Due to the scarcity of paired datasets, learning such transformation may be challenging. With a growing number of publications each year, unpaired I2I synthesis has drawn attention from the research community and has found a great application field in medical imaging. Disentangled representation learning constitutes a significant portion of these methods. By relying on the factorization of an image into independent variation latent codes, this approach offers greater control over the result of the synthesis than GAN-based ones. However, disentangling latent representations is not a trivial task and may be influenced by particular inductive bias. In this work, we propose to apply disentangled representation learning to the unpaired synthesis of high-resolution dynamic MRI. We study the impact of both the entanglement module and the addition of a segmentation auxiliary task on the result of the synthesis and the disentanglement of the representations. Our results demonstrate that the choice of the entanglement module greatly influences the learning of a good representation, and that the addition of a segmentation auxiliary task leads to better synthesis performances. Our code is available at `https://github.com/cScavinner/Unpaired_image_synthesis`.

Keywords

Dynamic MRI, Unpaired image synthesis, Disentangled representations learning, Cerebral palsy

Article informations

https://doi.org/https://doi.org/10.59275/j.melba.2025-AAAA

Received: 2024-04-03, Published 2025-02-28
Corresponding author: claire.scavinner-dorval@imt-atlantique.fr

1. Introduction

Image synthesis is defined as the process of generating an image according to specific characteristics. It can be either conditional or unconditional. Conditional image synthesis generates images based on provided input data, which can include text descriptions (prompts), existing images, or even sounds. It has found extensive applications within computer vision tasks such as style transfer (Jiang et al., 2020; Gatys et al., 2016), super-resolution (Ledig et al., 2017; Hu et al., 2021) and semantic image synthesis (Park et al., 2019; Tang et al., 2020), etc.). Furthermore, image synthesis is making significant contributions in the fields of medical imaging in reconstruction (de Haan et al., 2020), denoising (Chen et al., 2022), registration (Fu et al., 2020) or super-resolution (Sanchez and Vilaplana, 2018). Image-to-image (I2I) translation in practice faces significant challenges due to the inherent ambiguity of the task and the scarcity of suitable training data. The core issue lies in the lack of readily available "paired datasets" for many applications. Paired datasets consist of corresponding input-ground truth pairs for a given task. While datasets like Horse2Zebra allow pairing for classification tasks (horse vs. zebra labels), they become unpaired for I2I translation (horse image to zebra image). Constructing such paired datasets necessitates manual annotation by experts, which can be time-consuming (e.g., segmentation), expensive (multiple imaging acquisitions), or prone

to errors (imperfect registration across modalities). These difficulties are particularly amplified in medical imaging, making the development of robust I2I translation models a complex endeavor.

The scarcity of paired data for image synthesis tasks has driven the research community towards utilizing unpaired datasets. Since their emergence in 2014, Generative Adversarial Networks (Goodfellow et al., 2020) (GANs) have formed the foundation for numerous unpaired image synthesis methods (Zhu et al., 2017; Park et al., 2020; Taigman et al., 2022). However, GAN-based approaches often lack control over the synthesis outcome, potentially generating realistic but anatomically inaccurate images in medical applications. As an alternative, disentangled representations have emerged as a promising approach. This class of methods assumes that an image can be decomposed into a set of independent, interpretable features, each representing a specific aspect of variation within the image (Liu et al., 2022; Chen et al., 2023). Compared to GANs, disentangled representations provide a more interpretable generation process, greater control over the synthesis, and improved transferability across tasks.

Dynamic MRI is an imaging technique employed *in vivo* to study physiological dynamics, including cardiac cycles, blood flow, and joint biomechanics (Garetier et al., 2020; Pettigrew, 1989). Our study centers on pediatric dynamic MRI data acquired for investigating equinus (Makki et al., 2019). Affecting roughly 75% of children with cerebral palsy, equinus is the most prevalent musculoskeletal deformity in this population (Banks and Green, 1958). Characterized by a fixed plantarflexed foot position with a neutral hindfoot and extended knee, equinus significantly impacts mobility. The use of dynamic MRI to understand the effect of equinus on joint mechanics and bone deformities may lead to a better comprehension of the post-surgery recurrence rate and improve medical management of equinus deformity. However, capturing high-resolution dynamic sequences requires a long acquisition time and repeated motion patterns at a constant speed, making them difficult to handle for patients with musculoskeletal disorders. In order to minimize patient discomfort, the choice of sequence type should be balanced between reducing acquisition time and ensuring sufficient resolution and contrast between anatomical structures. To enhance comprehension of the underlying pathology, a combination of one high-resolution static image and three lower-resolution dynamic images was employed, with the aim of studying the *in vivo* biomechanics of the pediatric ankle joint (detailed in Figure 1). Estimating high-resolution dynamic sequences from static images exemplifies an unpaired image synthesis problem. This work delves into the application of a disentangled representation learning framework to address this challenge.

In this work, we present the following contributions:

1. We study the impact of the entanglement module on disentangle representations learning for unpaired image synthesis.
2. We investigate the impact of incorporating a segmentation auxiliary task on the synthesis results.
3. We study the uncertainty through test time data augmentation of dynamic MR image synthesis.

In a previous study (Scavinner-Dorval et al., 2024), a model called mDRIT++, based on the DRIT++ framework, was introduced. In terms of neural network architectures, the mDRIT++ framework is equivalent, within the context of this study, to the DRL method with the AdaIN entanglement module and without the auxiliary segmentation task. The present work investigates the influence of several architectural hyperparameters and the addition of a segmentation module on the result of the synthesis in a different experimental setup.

2. Related Works

2.1 Unpaired Image-to-Image Synthesis

The field of unpaired image synthesis is primarily driven by two dominant frameworks: Generative Adversarial Networks (GANs) and Disentangled Representation Learning (DRL) (Chen et al., 2023). More recently, diffusion models have gained significant interest from the research community due to their ability to generate highly realistic images. Diffusion models represent a class of generative models wherein images are progressively corrupted with an increasing level of noise, which the models then learn to reverse. This generative process is defined as the reverse of a Markovian process, whereby white noise is progressively denoised to produce an image. These models have already been applied in unpaired image synthesis in computer vision (Sasaki et al., 2021; Sun et al., 2023) and medical imaging (Luo et al., 2024; Özbey et al., 2023; Fan et al., 2024). Diffusion-based methods are capable of achieving diverse and highly realistic image synthesis, and have been shown to outperform GAN-based methods for certain tasks. However, the sampling process is performed in a non-disentangled data domain, such as the image domain. These methods are computationally expensive during training and inference steps. GANs are a family of deep learning models capable of generating new data based on existing samples. GAN training involves two competing networks: the Generator and the Discriminator. The Generator strives to synthesize new data that is indistinguishable from a real data set Y, while the Discriminator aims to differentiate real data from the synthesized ones (Goodfellow et al., 2020). In unconditional image synthesis, the

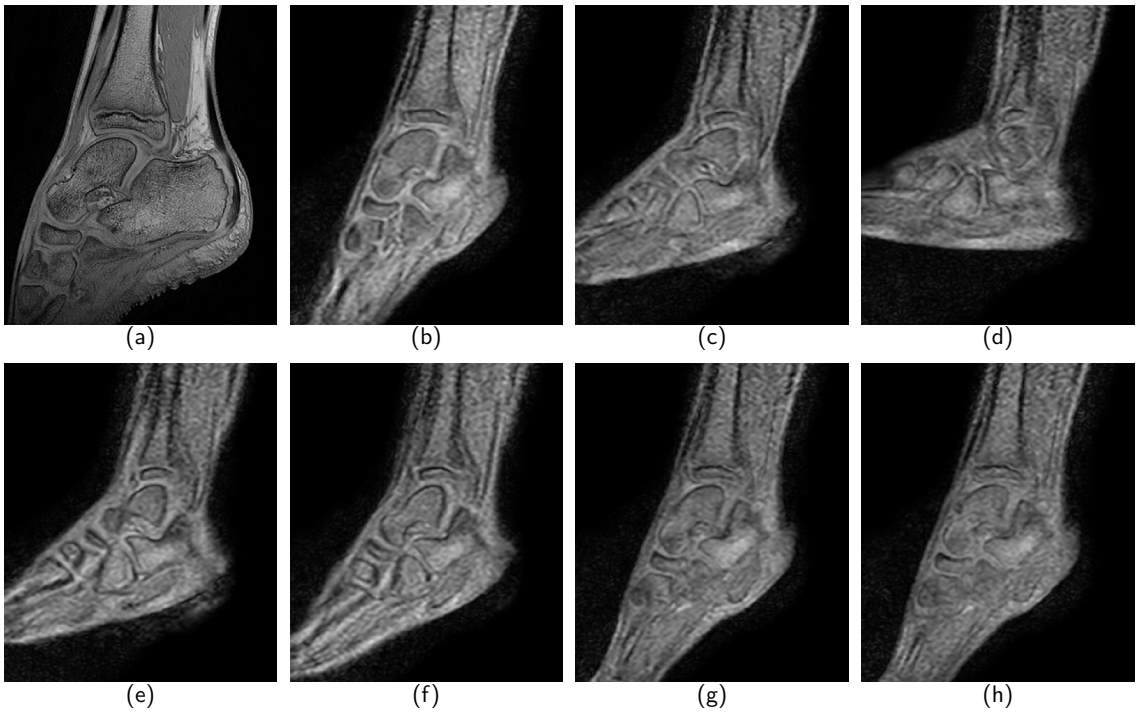

Figure 1: MRI source data. (a) Static 3DT1 image (b)-(h) Seven different time frames extracted from a dynamic MRI sequence. During each dynamic MRI sequences, the subject performs (actively or passively) a single cycle of dorsi-plantar flexion with a rotation speed between 4°/s and 5°/s. Each sequence consists in 15 3D time frames.

Generator takes a random vector z as input and outputs a synthetic data. In this scenario, the output is solely constrained by the set of real data Y that it attempts to mimic. Conversely, conditional image synthesis employs additional input such as an image (image-to-image synthesis) or text (text-to-image synthesis), to guide the generation process. In the context of image-to-image translation, cyclic GANs have emerged as a popular approach (Zhu et al., 2017). These models leverage two GANs in a closed loop, with each GAN focusing on one direction of the translation cycle.

Although GAN-based approaches are effective in image-to-image translation, they often suffer from limited control over the synthesis outcome, resulting in the generation of realistic but structurally inconsistent images that lack human interpretability. Disentangled representation learning (DRL) provides an alternative approach based on the assumption that an input image can be decomposed into a set of independent, meaningful features. Each of these features represents a specific mode of variation within the image and is encoded into a distinct dimension (Liu et al., 2022; Chen et al., 2023; Higgins et al., 2018). Encouraging a representation to be disentangled offers several advantages, including a more interpretable generation process, greater control over the synthesis, and improved transferability across tasks.

2.2 Disentangled representation learning

In deep learning, representation learning refers to the process of learning a representation of the data that facilitates information extraction for subsequent tasks (Bengio et al., 2013). Therefore, finding a "good" representation for a given task is a critical issue in deep learning. Disentangled representation learning aims to find the particular representation that captures the underlying factors of variation in the data distribution. Ideally, each of these factors represents an independent, semantically meaningful aspect of the data, human-understandable. Such a representation then allows to generate an image with control over each semantic aspect. According to Thomas et al. (2017), a representation is considered "disentangled" if changing a single factor only affects the corresponding mode of variation within the data. As there is no clear consensus on the definition of disentanglement, several approaches have been proposed in the last decade (Higgins et al., 2018; Do and Tran, 2019; Eastwood and Williams, 2018; Achille and Soatto, 2018; Bengio et al., 2013). However, there is a general agreement within the community regarding two key properties: the representation needs to be distributed, and each factor should be encoded in a distinct dimension within the latent space.

Disentangled representations, due to their inherent flexibility, find applicability across various modalities and tasks (Wang et al., 2023). They are commonly used in com-

puter vision for tasks such as generation (Karras et al., 2019; Chen et al., 2016; Kim and Mnih, 2018) and domain adaptation (Lee et al., 2020; Gonzalez-Garcia et al., 2018), and to a lesser extent in classification (Ferreira dos Santos and Mileo, 2023), semantic segmentation (Chu et al., 2021), and colorization (Lai et al., 2020). Furthermore, beyond these practical applications, disentangled representation learning frameworks hold great promise due to their tractability, generalizability, and controllability (Wang et al., 2023). Assuming a representation to be disentangled, each factor of variation is modeled by a corresponding latent representation that retrieves a particular semantic aspect of a set of images. This enables the generation of new images to be controllable and interpretable. Moreover, if the learned representation is truly disentangled, each factor is statistically independent from the others, leading to improved generalization abilities compared to classical algorithms.

Disentangling style and content is a non-trivial process, particularly in unpaired I2I, as there is no ground truth for the latent variable to be extracted. To ensure the correct information is extracted, a set of constraints can be applied. A good disentanglement may be favored by specific inductive biases, such as architectures, latent space manipulations, or learning schemes. Several techniques can be used to improve disentanglement, including entanglement modules, special learning schemes, and, in the case of unpaired I2I, the use of an equivariant task as a regularization. Section 2.3 details commonly used entanglement modules, while Section 2.4 reviews the main methods for promoting disentanglement in the latent space.

2.3 Entanglement module

The entanglement module plays a crucial role within a content-style disentanglement learning framework. Its function is to merge the different latent codes (disentangled factors) within a generator to produce a new image. The effectiveness of this merging process, often referred to as entanglement, is paramount in disentangled representation learning.

Many of the most popular techniques in the recent literature for information merging in neural networks are based on conditional normalization (CN). CN is a concept derived from normalization layers, such as Batch Normalization (Ioffe and Szegedy, 2015), where the affine parameters of the transformation are learned from conditioning information. These methods aim to modulate the network's intermediate activations in a domain-specific manner. The most popular CN methods include Conditional Instance Normalization (CIN) (Dumoulin et al., 2017), Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) and Adaptative Instance Normalization (AdaIN) (Huang and Belongie, 2017). The FiLM layer is proposed as a conditioning layer for visual reasoning tasks and as a generalization of CIN. It applies a channel-wise affine transformation to the intermediate activations of a neural network. The transformation parameters are learned from arbitrary input conditions by two multi-layer perceptrons (MLPs). This module has applications in medical image segmentation (Chartsias et al., 2019), soundscape editing (Jiang et al., 2024), and vision language control for robotics (Saxena et al., 2023). On the other hand, the AdaIN module was originally designed for arbitrary style transfer. However, AdaIN is widely used in literature for various conditional generation tasks, including text-to-image generation (Tewel et al., 2024), motion synthesis (Cao et al., 2022), colorization from audio (Zhao et al., 2024), text-to-speech synthesis (Li et al., 2024) or face generation (Karras et al., 2019). Like the FiLM layer, AdaIN acts as an affine transformation in the feature spaces. However, unlike FiLM, the affine transformation parameters are not learned but extracted from the input condition using first and second-order statistics. The transformation is achieved by adjusting the channel-wise mean and standard deviation of the network's intermediate activations based on those from the condition across spatial locations.

FiLM and AdaIN are frequently chosen as entanglement modules in content-style disentanglement frameworks. In this context, the content serves as the generator's direct input, while the style is provided as a conditioning input. Although early work such as Esser et al. (2018) used a simple concatenation of style and content, this approach may hinder proper disentanglement between the two factors (Liu et al., 2022). In the field of content-style disentanglement, some methods propose custom architectures for style-content fusion, as seen in Lee et al. (2020). However, to the best of our knowledge, there is currently a lack of research investigating the influence of the entanglement module on the quality of synthesized images in the field of content-style disentanglement.

2.4 Promote and measure the disentanglement

Learning disentangled representations remains a significant challenge, especially in unpaired settings. A successful framework for disentangled representation learning hinges on the independence between the latent representations. However, in most common applications of these frameworks, the ground truth for the latent codes does not exist. A significant challenge in these methods is to promote disentanglement and independence between the latent representations of the factors of variation, without knowledge of the exact realizations of these factors. Additional strategies are introduced to enforce disentanglement of the latent representations without explicit supervision of the latent variables.

Most of the time, unpaired I2I translation methods using disentangled representation rely on a cycle consistency loss to guide the translation process. This objective, introduced by Zhu et al. (2017), constitutes the state-of-the-art in unpaired I2I. This objective has proved its efficiency in guiding inter-domain translation, providing strong regularization, and stabilizing training (Li et al., 2017). However, cycle-consistency alone cannot guarantee a correct disentanglement of the latent representations. Recent approaches have introduced constraints into the latent spaces to further constrain the latent representations themselves and their disentanglement, instead of working only in the image spaces.

These techniques are mainly divided into two categories: those that guide the structure of the latent space and those that guide the content. The methods working on the structure of the latent representation are usually designed to condition the shape and distribution of the data point in the latent space (using a VAE (Kingma and Welling, 2014) with a latent space constrained to fit a normal distribution with zero mean and unit variance, or contrastive learning (Yu et al., 2021; Hjelm et al., 2018; Wu et al., 2018; Chen et al., 2020) which favors the proximity of similar data in the latent space and pushes away the data that are different). These methods can promote disentanglement by encouraging a human-interpretable structure and favoring certain properties of the latent representations. The second category of methods focuses on constraining the content of the latent representations. How to favour some semantic properties in the latent representations without knowing any realization of the latent codes? As previously mentioned, content representation in content-style disentanglement is particularly suitable for use in equivariant tasks. In imaging tasks, it mostly pertains to semantic segmentation. Requiring the content to convey the necessary anatomical properties for segmenting anatomical structures has been shown to improve the learned representation, as demonstrated by Chartsias et al. (2019) and reported by Liu et al. (2022). On the other hand, Lee et al. (2020) introduce a content discriminator to encourage the content representations to be indistinguishable between the two modalities. This type of prior enforces domain-invariance of the content and constrains the style to convey domain-specific information.

Although there is a growing number of approaches that focus on learning disentangled representations, the metrics for quantifying this disentanglement are still relatively unknown. Most of the proposed metrics are applicable in the following cases: the ground truths corresponding to the variation factors are known, or the disentanglement of the factors takes place within a single latent vector variable (such as GAN or VAE) (Carbonneau et al., 2022; Chen et al., 2018; Kim and Mnih, 2018; Higgins et al., 2022; Duan et al., 2019). When it comes to disentangling content and style, the ground truths for the variation factors are often inaccessible. In addition, the different representations are encoded as multiple latent variables of different forms (style is often represented as a vector, while content is often represented as a spatial tensor). In practice, few disentanglement metrics are designed to overcome the need for ground truth.

3. Methods

In this work, given a set of low-resolution dynamic MRI data X and a set of high-resolution static MRI data Y, we seek to estimate the transformation that synthesized a high-resolution dynamic MRI data from a low-resolution dynamic MRI data, meaning to estimate the transformation mapping an element $x \in X$ into an element of Y. This work is inspired by DRIT++ (Lee et al., 2020), an unpaired image synthesis framework built upon disentangled representation learning. We leverage this framework as a foundation to investigate the influence of the entanglement module on the synthesis outcome. Furthermore, we incorporate segmentation masks to bolster disentanglement and enhance the quality of the synthesized images through the introduction of a segmentation auxiliary task. This section is structured as follows: Section 3.1 details the original DRIT++ framework, Section 3.2 describes the different entanglement modules and Section 3.3 provides further details on the incorporation of the segmentation as an auxiliary task.

3.1 Unpaired Image Synthesis

Consider a pair of magnetic resonance images (MRI) acquired from the same subject, denoted as $x \in X$ and $y \in Y$. This work aims to disentangle the content, referred to as "anatomy" in medical imaging terminology, from its representation, which aligns with the concept of "style" in broader literature. The modality refers to the MRI acquisition parameters specific to each image and embodies modality-specific attributes. Typically represented as a vector, it does not encode spatial information but rather the rendering of the anatomical structures. Conversely, the anatomy represents the modality-invariant, spatial information within the image. Encoded as a tensor, this representation preserves the spatial correlations of the original images, making it suitable for tasks requiring equivariance.

For each image, dedicated modality-specific, fully-convolutional encoders extract two latent codes representing the disentangled factors. We denote by E^a_X and E^a_Y the anatomy encoders and E^m_X and E^m_Y the modality encoders. The anatomy encoders aims at mapping images into a modality-invariant, shared latent space while the modality encoders

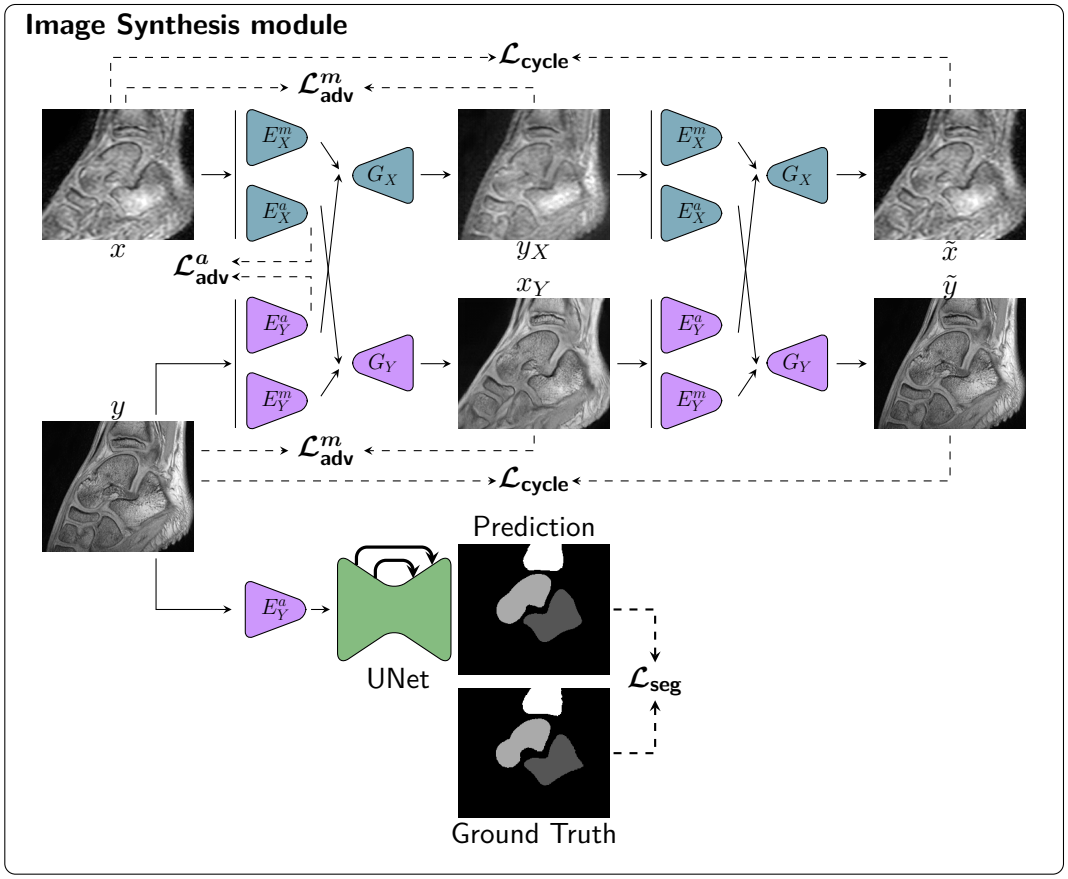

Figure 2: Overview of the synthesis process. Each image $x \in X$ and $y \in Y$ is factorized into anatomy and modality latent codes by E^a and E^m. Generators G recombine the latent codes and generate a new image according to the input latent codes. Once the cross-modality synthesis is done, the inverse operation is performed to recover the original images and ensure the cycle consistency. The segmentation network (UNet) takes as input the anatomy latent code of the image y and outputs the corresponding bones segmentation. The overall model is trained in an end-to-end manner. It should be noted that networks with the same notation are identical in terms of both structure and weights.

aims at mapping the images into modality specific, independent latent spaces. The extracted latent codes are denoted z^m for the modality latent code and z^a for the anatomy latent code. The disentanglement process is described by Equation (1).

$$\begin{cases} x = (z^a_x, z^m_x) = (E^a_X(x), E^m_X(x)) \\ y = (z^a_y, z^m_y) = (E^a_Y(y), E^m_Y(y)) \end{cases} \tag{1}$$

Since each pair (x, y) consists of images acquired from the same subject, they are expected to share the same anatomical properties despite differing modalities. Disentanglement is enforced through a discriminator working in the anatomy latent space and shared weights in the networks. This discriminator is denoted by D and aims to identify the source modality of a given anatomy latent code. This setup encourages the anatomy representation to be modality invariant. The corresponding adversarial loss is expressed by Equation (2).

$$\begin{aligned} \mathcal{L}^a_{\text{adv}}(E^a_X, E^a_Y, D) = \; & \mathbb{E}_x[\frac{1}{2}\log D(E^a_X(x)) \\ & + \frac{1}{2}\log(1 - D(E^a_X(x)))] \\ & + \mathbb{E}_y[\frac{1}{2}\log D(E^a_Y(y)) \\ & + \frac{1}{2}\log(1 - D(E^a_Y(y)))] \end{aligned} \tag{2}$$

Two generators, one for each modality, are conditioned on both anatomy and modality latent codes to generate a new image. G_X stands for the generator in the X domain while G_Y stands for those in the Y domain. A cross-modality synthesis from a source modality to a target modality is achieved by providing to a generator the anatomy latent code of the source image and the modality latent code of the image from the target modality. We denote x_Y and y_X the synthesized cross-modality images.

The generation process is described in Equation (3).

$$\begin{cases} x_Y = G_Y(z_x^a, z_y^m) \\ y_X = G_X(z_y^a, z_x^m) \end{cases} \tag{3}$$

To address the unpaired setting, the model is conditioned on both image and latent spaces. Similar to the CycleGAN framework, we first introduce an adversarial loss to constrain the generated images' data distribution to approximate that of the target modality's real images. Two discriminators, one for each modality and denoted by D_X and D_Y are introduced to this end. The corresponding adversarial loss, including both discriminators is expressed by Equation (4). Given x' and y' two "real" data from X and Y respectively:

$$\begin{aligned} \mathcal{L}_{\text{adv}}^m(G_X, D_X, G_Y, D_Y) = \\ \mathbb{E}_{x' \in X}[\log D_X(x')] \\ + \mathbb{E}_{x \in X, y \in Y}[\log(1 - D_X(G_X(z_y^a, z_x^m))] \\ + \mathbb{E}_{y' \in Y}[\log D_Y(y')] \\ + \mathbb{E}_{x \in X, y \in Y}[\log(1 - D_Y(G_Y(z_x^a, z_y^m))] \end{aligned} \tag{4}$$

Second, the framework leverages a cycle-consistency loss to stabilize the training, introduce a regularization, and enforce content preservation through the domain translation. $\tilde{x}$ and $\tilde{y}$ (expressed by Equation (5)) denote the images obtained after two inter-modality translations.

$$\begin{cases} \tilde{x} = G_X(z_{x_Y}^a, z_{y_X}^m) \\ \tilde{y} = G_Y(z_{y_X}^a, z_{x_Y}^m) \end{cases} \tag{5}$$

The cycle-consistency loss is then expressed by the following equation:

$$\mathcal{L}_{\text{cycle}} = ||x - \tilde{x}||_1 + ||y - \tilde{y}||_1 \tag{6}$$

In addition to the previously introduced objective functions, several others are used in order to ease the training. A self-reconstruction loss compare the original image with its reconstruction obtained by using latent codes from the original image (see Equation (7)).

$$\mathcal{L}_{\text{self}} = ||x - G_X(z_x^a, z_x^m)||_1 + ||y - G_Y(z_y^a, z_y^m)||_1 \tag{7}$$

A latent regression loss enforces an invertible mapping between image and modality latent space, forcing images to contain the information encoded in the modality representation (Huang et al., 2018). Given z_{rdn}^m a random modality latent code sampled from $\mathcal{N}(0, 1)$, the latent regression loss is expressed by Equation (8):

$$\begin{aligned} \mathcal{L}_{\text{latent}} = \frac{1}{n}\sum_{i=1}^{n} |z_{\text{rdn}}^m - E_X^m(G_X(z_y^a, z_{\text{rdn}}^m))| \\ + \frac{1}{n}\sum_{i=1}^{n} |z_{\text{rdn}}^m - E_Y^m(G_Y(z_x^a, z_{\text{rdn}}^m))| \end{aligned} \tag{8}$$

The global cost function is written:

$$\begin{aligned} \mathcal{L} = \lambda_{\text{adv}}^a \mathcal{L}_{\text{adv}}^a + \lambda_{\text{adv}}^m \mathcal{L}_{\text{adv}}^m + \lambda_{\text{latent}} \mathcal{L}_{\text{latent}} \\ + \lambda_{\text{self}} \mathcal{L}_{\text{self}} + \lambda_{\text{cycle}} \mathcal{L}_{\text{cycle}} \end{aligned} \tag{9}$$

where λ_{adv}^a, λ_{adv}^m, λ_{latent}, λ_{self}, and λ_{cycle} are the dedicated weighting for each particular objective function. An overview of the method is proposed on Figure 2.

3.2 Entanglement module

Let us consider the anatomy latent code, denoted by $z^a \in \mathbb{R}^{N \times C \times H \times W}$, and the modality latent code, denoted by $z^m \in \mathbb{R}^{N \times h}$. The variable N stands for the batch size, (C, H, W) refers to the spatial dimensions of the anatomy latent code, and h refers to the dimension of the modality latent code.

FiLM The FiLM module (Perez et al., 2018) learns the mapping function from z^m to the affine parameters γ and β. Both mapping functions are modeled by a MLP. γ and β are computed channel-wise and across spatial locations and aim at modulating intermediate activations of a neural network. The FiLM operator can be expressed by Equation (10) where $(.)$ stands for the element-wise multiplication and $(+)$ for the element-wise addition:

$$\text{FiLM}(z_{n,c}^a, z_{n,c}^m) = \gamma_{n,c}(z^m).z_{n,c}^a + \beta_{n,c}(z^m) \tag{10}$$

where $n \in [1, N]$, $c \in [1, C]$ and $\gamma, \beta \in \mathbb{R}^{N,C}$.

AdaIN AdaIN (Huang and Belongie, 2017) is initially designed for arbitrary style transfer and also acts as an affine transformation in the feature space. Unlike FiLM, the AdaIN module has no learnable parameters and the parameters of the affine transformation are obtained from z^m statistics. While both FiLM and AdaIN modules are used for information merging, their approaches differ. The FiLM module directly applies an affine transformation to the intermediate activations derived from z^a. In contrast, the AdaIN module focuses on aligning the channel-wise mean and standard deviation of each intermediate activation sample with the corresponding values from z^m:

$$\text{AdaIN}(z_{n,c}^a, z_{n,c}^m) = \sigma_{n,c}(z^m)\Big(\frac{z_{n,c}^a - \mu_{n,c}(z^a)}{\sigma_{n,c}(z^a)}\Big) + \mu_{n,c}(z^m) \tag{11}$$

Conv The method used in the DRIT++ framework (Lee et al., 2020) relies on successive concatenations between z^a and z^m, followed by convolution operations to merge both data. K_i denotes the $i \times i$ convolution kernel and $(\oplus)$ the concatenation operator:

$$\text{Conv}(z_{n,c}^a, z_{n,c}^m) = F(\text{IN}(F(z_{n,c}^a, z_{n,c}^m) * K_3), z_{n,c}^m) \tag{12}$$

where $F(f, z) = (a((f \oplus z) * K_1) * K_1)$, a is an activation function (e.g. ReLU) and IN is the Instance Normalization function.

3.3 Supervised Segmentation

This section details the introduction of a segmentation module into the framework detailed in Section 3.1. An overview of the proposed method is illustrated in Figure 2. As discussed in Section 2.4, auxiliary task such as segmentation may provide a way to enforce disentanglement of the latent representations by encouraging particular properties in the content representation (Chartsias et al., 2019; Liu et al., 2022). The content representation in image-to-image translation is particularly suitable for equivariant tasks such as segmentation (Chartsias et al., 2019). To this end, we leverage the ankle joint bones segmentation on the static MR images to introduce a supervised segmentation task from the anatomy representation.

This segmentation net takes as input the anatomy representation of the static MR image and outputs the corresponding segmentation of the ankle joint bones. Given segmentation maps of static data, this task is performed in a supervised manner. The goal is to constrain the anatomical information to be encoded into the anatomy representation in order to perform the segmentation while the modality-specific information remains in charge of the modality representation. The objective function, denoted by $\mathcal{L}_{\text{seg}}$, is a cross-entropy loss between the estimation provided by the segmentation net and the ground truth. This paired setting for the segmentation offer an additional constraint to guide the extraction of the latent representation.

The global cost function, including the segmentation module, is expressed as follows:

$$\begin{aligned}\mathcal{L} = \lambda^a_{\text{adv}}\mathcal{L}^a_{\text{adv}} + \lambda^m_{\text{adv}}\mathcal{L}^m_{\text{adv}} + \lambda_{\text{latent}}\mathcal{L}_{\text{latent}} \\ + \lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} \quad (13)\end{aligned}$$

4. Experimental Setup

This section describes the experimental setup, including the dataset used in the experiments in Section 4.1, the metrics considered for evaluation (see Section 4.2), and the implementation details in Section 4.3.

4.1 Dataset

This dataset consists of pediatric MRI data (Makki et al., 2019) of the ankle joint acquired in order to study equinus. Equinus is the most common deformity in children with cerebral palsy (Cobeljic et al., 2009). To enhance comprehension of pediatric ankle joint biomechanics, one high-resolution static image and several low-resolution dynamic sequences have been acquired. The study, which was approved by the regional ethics committee, included eleven typically developing children and nine children with equinus, aged between 6 and 14 years old. 3D MRI data were collected during a single visit after the parents signed informed consent forms. The data was acquired using a 3T MR scanner (Achieva dStream, Philips Medical Systems, Best, Netherlands). The acquisition protocol includes, per child, one high-resolution static 3D MR scan of the ankle joint with a resolution of $0.26 \times 0.26 \times 0.8$mm (T1-weighted gradient-echo, flip angle 10°, matrix 576576, FOV 150mm×150mm, TR/TE 7.81/2.75 ms, mean acquisition duration: 424.32 s) and three low-resolution dynamic sequences of the ankle joint while performing a single cycle of dorsi-plantar flexion. The dynamic sequences includes two passive sequences for repeatability measures and one active sequence, all acquired with knee angle kept in full extension (approximately between 0° and 10°). Each sequence consists of 15 3D time frames with a spatial resolution of $0.57 \times 0.57 \times 8$mm (flip angle 15°, matrix 352352, FOV 200mm×200mm, TR/TE 20.61/1.8 ms, acquisition duration: 18.98 s).

Each static MRI has been manually segmented by human experts. The segmentations include the three bones of interest of the ankle joint: the talus, the tibia, and the calcaneus. All the segmentations have been performed using ITK-Snap (Yushkevich et al., 2006). The age range of the subjects is from 6 to 14, resulting in variations in the growth plates among them. The segmentation of bones includes both bone and eventual growth plates, but articular cartilage is not included.

4.2 Metrics

Reconstruction metrics A large part of the image synthesis literature uses full-reference image quality metrics such as PSNR, MSE, or SSIM. These metrics rely on a strong physical background and have proven their efficiency and relevance. However, they are not suitable for unpaired datasets since they compare a synthesized image with its corresponding ground truth, which is not available in unpaired datasets.

No-reference image quality assessment metrics are designed to avoid the need for ground truth. These metrics are often built on comparisons between intermediate activations of secondary neural networks, rather than comparing pixels. Two commonly used metrics for evaluating the quality of generated images are Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018). The FID corresponds to an improvement of the Inception Score (Salimans et al., 2016). It compares features of both real and synthetic images using Inception v3's features from a net trained on

ImageNet, as described in equation (14).

$$\begin{aligned} \text{FID} &= d^2\Big((\mu, C), (\mu_w, C_w)\Big) \\ &= \|\mu - \mu_w\|_2^2 + \text{Tr}\Big(C + C_w - 2(CC_w)^{1/2}\Big) \end{aligned} \tag{14}$$

where d(.,.) is the Wasserstein-2 distance (Vaserstein, 1969), (μ, C) the multivariate Gaussian obtained from the intermediate activation of the Inception network fed by generated data and (μ_w, C_w) the Gaussian obtained from the intermediate activation of the Inception network fed by real data. μ stands for the mean and C for the covariance matrix. Although FID was initially developed for natural image analysis, there is evidence that it can also be relevant in medical imaging (Woodland et al., 2022).

The KID is a variation of FID, also based on Inception v3's activations. It reflects the shared visual similarities between real and synthetic images. It is defined as the squared Maximum Mean Discrepancy (MMD) between the Inception intermediate activation of real and generated images, using a polynomial kernel:

$$\text{KID} = MMD(f_{real}, f_{fake})^2 \tag{15}$$

f_{real} defines the intermediate features from real images and f_{fake} those from generated images. The kernel for MMD computation is defined as $k(x, y) = \Big(1/dx^T y + 1\Big)^3$.

However, these metrics suffer from a lack of reproducibility as their values can only be roughly compared between different setups and papers (Parmar et al., 2022).

Disentanglement metrics Liu et al. Liu et al. (2021) focus on quantifying the disentanglement in the case of style-content disentanglement. They propose using distance correlation Székely et al. (2007) and a metric called 'Information over Bias' to measure correlation and informativeness. Distance correlation (DC) measures the degree of dependence between two random variables of arbitrary dimensions, and unlike Pearson correlation, DC is bounded in the interval [0,1]. A DC=0 indicates that the two random variables are independent. In the case of disentangling style and content, let X and Y be two matrices with N rows corresponding to N examples and an arbitrary number of columns. X and Y can refer to style, content, or linked images. As variables corresponding to content or images are spatial tensors, they are reformatted as vectors, resulting in a 2D matrix. The distance correlation is written as follows:

$$\text{DC}(X, Y) = \frac{\text{dCov}(X, Y)}{\text{dVar}(X)\text{dVar}(Y)} \tag{16}$$

with $\text{dCov}(X, Y) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{A_{i,j}B_{i,j}}{N^2}}$ the distance covariance between X and Y and $\text{dVar}(X) = \text{dCov}(X, X)$ the distance variance. A and B are the respective distance matrix of X and Y. The DC can be compute between the anatomy and the modality latent codes. DC values closer to 0 indicate a higher disentanglement. The DC can also be compute between those latent codes and the resulting image, indicating the level of dependence between the image and its generating factor. However, such metric alone cannot retrieve the disentanglement ability of a system. As indicated in Liu et al. (2021), $\text{DC}(X, Y) = 0$ may indicate either X and Y encode unrelated information, or one of those encode all the information and the other encode noise, indicating full entanglement and posterior collapse. To this end, the authors introduced the Information over Bias (IoB) metric, which aim to measure the informativeness of generating factors relative to the corresponding image. IoB aim at comparing a reconstruction accuracy of particular images from uninformative representation and the estimated representation done by the disentangle representation learning model. Given z an estimated representation, $\mathbb{1}$ an uniformative constant tensor, and G_{θ_n} a neural network defined by its parameters θ_n aiming to reconstruct images I given a representation, IoB is defined as the expectation over the test images of the ratio between MSE obtained after the training of G_{θ_n} on the uninformative representation and the informative one:

$$\text{IoB}(I, z) = \mathbb{E}_i\left[\frac{\text{MSE}(I_i, G_{\theta_1}(\mathbb{1}))}{\text{MSE}(I_i, G_{\theta_2}(z_i))}\right] \tag{17}$$

When learning from the uninformative representation $\mathbb{1}$, G_{θ_n} only learn the dataset biais which can be modeled by θ_n, so, higher values of IoB can be associated with an higher amount of information encoded into the representation z. IoB=1 meaning that no information about iages I are encoded into the representation z.

4.3 Implementation details

The training set includes eleven subjects, five with equinus and six typically developing. The validation and test sets each include one subject, respectively typically developing and with equinus. All MR images were resampled to an intermediate resolution of 0.41×0.41×8mm. The resolution in the sagittal plane was chosen to be halfway between the static and dynamic resolutions to limit information loss in static images and interpolation artifacts in dynamic images. Finally, to limit the proportion of interpolated images in the training set and given the significant differences in resolution between static and dynamic images, the resolution in the frontal axis is kept the same as that of the dynamic images. The static MR images are preprocessed to be roughly registered to the dynamic images. This registration enables the use of image patches in the disentangled representation learning framework in an unpaired setting without using too dissimilar image patches for each modality. For exam-

ple, a patch sampled from the foot in the dynamic image and a patch sampled from the background in the static image should not be used in parallel. Throughout this work, we used 2D patches from MRI sequences for training. The use of 3D patches increases computation times and slows down the training process. Furthermore, the strong anisotropy in dynamic MR images between the resolution in the sagittal plane and along the frontal axis limits the accuracy and robustness of volumetric image processing. To increase the quantity of training data, data augmentation was performed using the TorchIO library (Pérez-García et al., 2021). TorchIO is an open-source library designed for medical imaging. It provides tools for loading, preprocessing, and augmenting data. It enables the generation of MRI-specific artifacts, such as magnetic field inhomogeneity and motion artifacts, as well as typical computer vision augmentations. All selected transformations can be applied with a specific probability and can be composed with others. Random transformations were applied to the data, including flips along the lateral axis, a bias field with a maximal magnitude of polynomial coefficient equal to 0.5, Gaussian noise with a standard deviation of 0.1, and affine transformations including scaling (with an amplitude of 0.2) and rotations (along the sagittal plane, with a 40° amplitude). The number of extracted patches was set to 95,400.

The model has been implemented using PyTorch. The framework is inspired by that of Lee et al. (2020) but using lighter network architectures. Each encoder is fully-convolutional. The anatomy encoders, denoted E^a, consist of three convolutional layers followed by two residual blocks. The last convolutional layers of E^a_X and E^a_Y are shared to enhance disentanglement (Lee et al., 2020). The modality encoders, denoted E^m, consist of five convolutional layers followed by an average pooling layer and a final convolutional layer. We set $z^m \in \mathbb{R}^8$ and $z^a \in \mathbb{R}^{256\times32\times32}$. Similar to Karras et al. (2019), each generator shares a common structure of a fully-connected mapping network that takes modality latent code z^m as input, followed by two residual blocks and three fractionally-strided convolutions. Each residual block shares the same fully-convolutional structure, except for the entanglement module. The content discriminator D is fully-convolutional and comprises three convolutional layers. Both modality-specific discriminators, D_X and D_Y, are PatchGAN discriminators introduced in CycleGAN (Zhu et al., 2017), each set with three layers and Instance normalization. The segmentation network is a UNet (Ronneberger et al., 2015), followed by fractionally-strided convolutions, with approximately 26K parameters.

The full model contains approximately 23 million parameters, compared to the original model's 87 million.

For training, we use Adam optimizer, with a learning rate of 10^{-5}. The batch size is set to 32, exponential decay rates to $(\beta_1, \beta_2) = (0.5, 0.999)$ and weight decay to 0.0001. For all experiments, we set $\lambda_{\text{cycle}} = 10$, $\lambda_{\text{latent}} = 10$, $\lambda_{\text{reg}} = 0.01$, $\lambda_{\text{self}} = 10$. For the segmentation task, we set $\lambda_{\text{seg}} = 10$. If not specified, all remaining weights are set to 0. Each DRL method is trained in an end-to-end fashion. The source code for this project is available at `https://github.com/cScavinner/Unpaired_image_synthesis`. As a baseline, we use the CycleGAN model (Zhu et al., 2017).

5. Results

We conduct the experiments on the dataset presented below. The disentangled representation learning framework, with and without the segmentation task, was compared to the CycleGAN model (Zhu et al., 2017). The results presented in this section are discussed in Section 6. Section 5.1 presents the results of high-resolution dynamic sequence synthesis. Section 5.2 describes the segmentation results of the disentangled representation learning framework, and Section 5.3 presents the uncertainty evaluation using test-time augmentation. An evaluation of the disentanglement is presented in Section 5.4.

5.1 Reconstruction

The disentangled representation learning framework and its variations are compared to CycleGAN. Figures 3 and 4 show the results of using the three entanglement modules with and without the segmentation module for both equinus and typically developing subjects. Table 1 and 2 provide a quantitative evaluation of the synthesis in terms of KID and FID. The Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD) are included for the methods that incorporate the segmentation module.

DRL methods that use segmentation as a task prior demonstrates superior performance in comparison to those without segmentation, in terms of both KID and FID, for both equinus and typically developing subjects. The use of segmentation as a task prior globally improved the synthesis performance for each entanglement module. The best results were obtained using the AdaIN entanglement module in combination with the segmentation task. The FiLM entanglement module demonstrated the poorer synthesis performance. Figure 3 and 4 provide a visual assessment of the synthesis quality. The source dynamic and static images are provided as a reference. It can be observed that the combination of the AdaIN module with the segmentation network provides the best trade-off between image quality and anatomical accuracy. Although the DRIT++ entanglement module seems to provide clean images with and without segmentation, it appears to be more prone to artifacts and produces less regular edges. The AdaIN

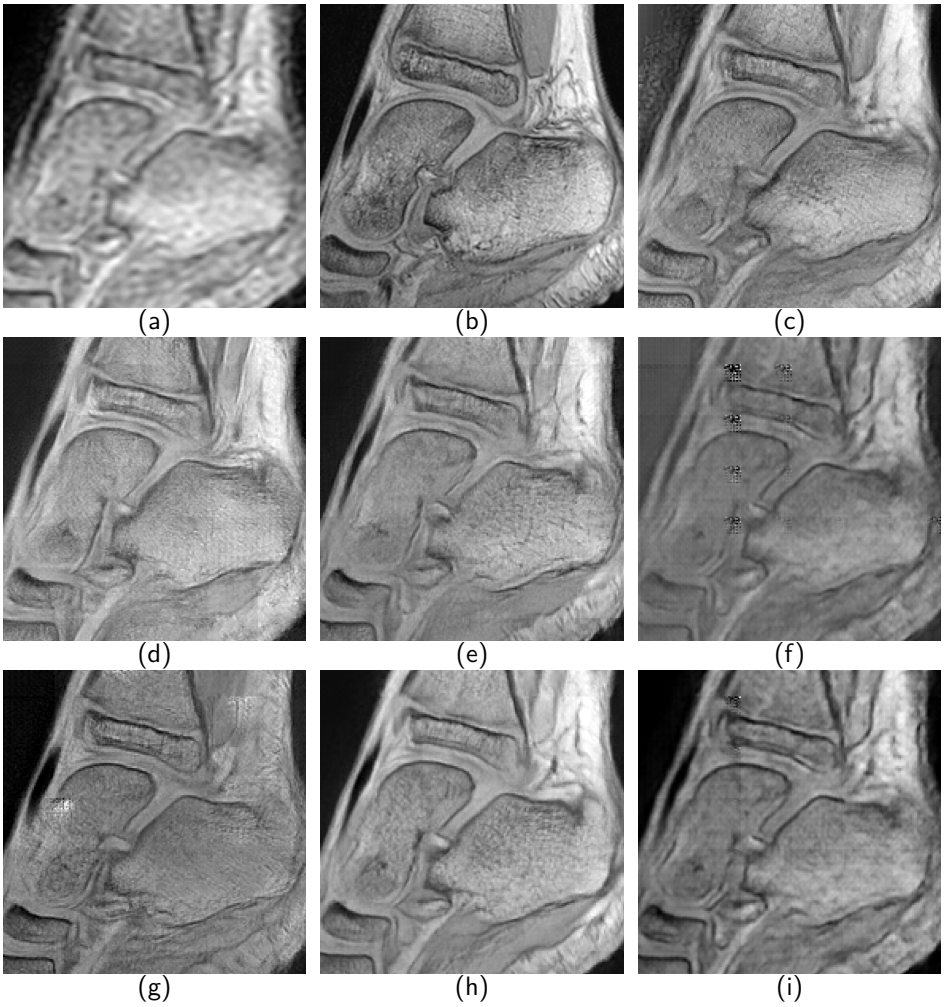

Figure 3: High-resolution dynamic sequence synthesis results using CycleGAN and DRL model for an **equinus subject**. (a) Dynamic image (b) Static image (c) Estimation with CycleGAN (d) Estimation with DRL - Conv (e) Estimation with DRL - AdaIN (f) Estimation with DRL - FiLM (g) Estimation with DRL - Conv with segmentation as an auxiliary task (h) Estimation with DRL - AdaIN with segmentation (i) Estimation with DRL - FiLM with segmentation.

module without segmentation produces a realistic synthesis of textures but generates inconsistent edges and lacks realistic bone and cartilage shapes.

A comparison between the quantitative evaluations of typically developing and equinus subjects reveal uneven performances of all the methods on each subject. If similar trends are observed between the different methods for both subjects, we observe a degradation in reconstruction and segmentation quality on the typically developing subject in comparison to the subject with equinus. With regard to reconstruction, the observed differences in metrics can be attributed to the sensitivity of these metrics to modifications in setup, particularly the number of samples employed for evaluation, which can influence metric reliability (Parmar et al., 2022). As the number of test images used for the typically developing subject is lower than for the equinus subject, a comparison of the two reconstruction performances may be compromised.

5.2 Segmentation

This section presents the segmentation results obtained with each entanglement module. The segmentation accuracy is evaluated using the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). The quantitative results for the segmentation task are presented in Table 1 for the equinus subject and Table 2 for the typically developing subject. Despite the disparity in synthesis quality, all methods exhibit comparable segmentation performance with regard to the DSC. Regarding the HD, the DRL methods using Conv and AdaIN modules exhibit comparable performance, whereas the AdaIN module exhibiting a higher variability. Despite the lower synthesis quality, the DRL method using the FiLM module consistently achieves the lowest HD among the different methods. The segmentation results are presented in Figure 5 for an equinus subject and 6 for a typically developing subject. The estimated segmentations are displayed as overlays on the source static

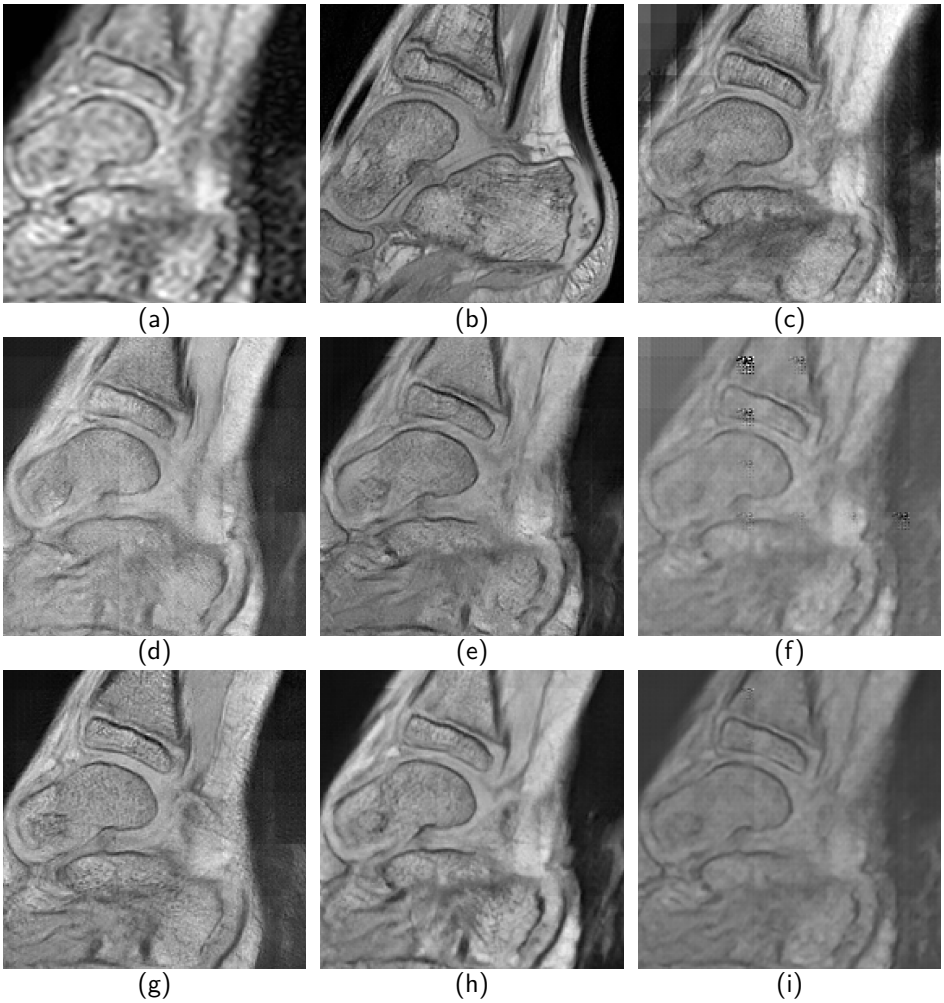

Figure 4: High-resolution dynamic sequence synthesis results using CycleGAN and DRL model for a **typically developing subject**. (a) Dynamic image (b) Static image (c) Estimation with CycleGAN (d) Estimation with DRL - Conv (e) Estimation with DRL - AdaIN (f) Estimation with DRL - FiLM (g) Estimation with DRL - Conv with segmentation as an auxiliary task (h) Estimation with DRL - AdaIN with segmentation (i) Estimation with DRL - FiLM with segmentation.

MRI in the first row. The ground truth is indicated as a reference on the left. The second row shows the absolute difference between the ground truth and the estimated segmentations. It could be observed that the segmentation performances are globally similar across the different entanglement modules. The absolute error is found to be evenly distributed along bone contours, regardless of whether the bone edges are smooth or sharp. Additionally, for each bone, the corresponding segmentation is free of holes and exhibits a single connected component. As illustrated by both DSC and HD, the segmentation accuracy is inferior for the typically developing subject, regardless of the DRL method employed.

5.3 Uncertainty

Data augmentation is commonly used during the training process, though it can also be used during the test time. The purpose is to generate multiple variations of a single image and then combine the predictions made by the model after reversing the transformation. Typically used as a way to improve the prediction of neural networks during training, especially in segmentation tasks, it also provides a measure of the uncertainty of the model with respect to particular transformations. In this case, we use it to evaluate the uncertainty of the high-resolution dynamic image synthesis process.

The transformations are randomly sampled from a set of rotations and scaling transformations. The amplitude of scaling is set to 0.2, and the rotation amplitude is set to 40° on the sagittal plane. A total of ten images are used, including the original images and nine randomly transformed images. High-resolution synthesis is performed for all ten images, and the inverse transformation is then applied to restore the original configuration. Statistics are then computed based on the ten resulting images, with a focus on the mean and standard deviation images.

Table 1: Evaluation of the synthesis performance for a **subject with equinus**. The evaluation is performed using two unpaired image metrics, KID and FID. The segmentation is evaluated using DSC and HD, computed on synthesised image segmentations. The best performance is indicated in bold.

	$\mathcal{L}_{seg}$	Entanglement	FID ↓	KID ↓	DSC ↑	HD (mm) ↓
DRL	✗	Conv	164.13	0.12	/	/
		AdaIN	195.5	0.17	/	/
		FiLM	367.5	0.44	/	/
DRL	✓	Conv	162.54	0.12	0.96 ± 0.003	2.8 ± 2.55
		AdaIN	**134.8**	**0.076**	0.96 ± 0.004	2.77 ± 3.0
		FiLM	201.4	0.15	0.96 ± 0.004	**1.66 ± 0.52**
CycleGAN			154.81	0.10	/	/

Table 2: Evaluation of the synthesis performance for a **typically developing subject**. The evaluation is performed using two unpaired image metrics, KID and FID. The segmentation is evaluated using DSC and HD, computed on synthesised image segmentations. The best performance is indicated in bold.

	$\mathcal{L}_{seg}$	Entanglement	FID ↓	KID ↓	DSC ↑	HD (mm) ↓
DRL	✗	Conv	227.4	0.19	/	/
		AdaIN	209.99	0.16	/	/
		FiLM	395.59	0.48	/	/
DRL	✓	Conv	183.32	**0.11**	0.93 ± 0.02	6.29 ± 8.02
		AdaIN	**181.45**	0.12	0.93 ± 0.009	8.06 ± 9.36
		FiLM	259.81	0.22	0.93 ± 0.007	**4.43 ± 4.71**
CycleGAN			277.63	0.28	/	/

Figure 7 illustrates the results for each method. The first and second rows of this figure correspond to the mean image and standard deviation image, respectively. Each mean image and standard deviation image shares the same contrast dynamic. A blurred mean image indicates greater variability in the image synthesis, whereas a sharper one indicates greater stability of the reconstructions. The standard deviation image reflects the amount of variability present in a given pixel. First, we observe that all the mean images from the DRL model methods seem to exhibit greater blurring than the CycleGAN one. However, while the mean image from the CycleGAN appears sharper than the others, the standard deviation within that area is considerably greater than in the other methods. The standard deviation image of the CycleGAN methods demonstrates a global variability in the image, while the other methods show high standard deviation values concentrated around the bones' edges. Secondly, it can be observed that the texture of the bone is rendered more accurately by the mean images generated by the AdaIN methods (with and without segmentation) than by those produced by the other methods. This is particularly visible on the calcaneus. The enhanced detail observed in the mean images produced by the AdaIN methods suggests that these methods yield more robust results than the other methods.

Additionally, we observe that the mean image of each method using the segmentation as an auxiliary task demonstrates significantly richer textures in terms of details, and a reduced variability, according to the standard deviation images. To support our assertions, three statistical measures—maximum, minimum, and mean—were computed for each standard deviation image. While the minimum values are observed to be comparable across all methods, the maximum and mean values exhibit notable discrepancies contingent on the method employed. The incorporation of the segmentation task results in a notable reduction in the maximum and mean values for both the FiLM and AdaIN modules. For the FiLM module, the maximum value declines from 43.75 to 24.02, while the mean value drops from 8.64 to 3.83. Similarly, for the AdaIN module, the maximum value decreases from 7.87 to 5.55, while the mean value decreases from 1 to 0.53. In the case of the Conv module, both the maximum and mean values are higher with the introduction of the segmentation task (without segmentation: max = 10.01 and mean = 0.75; with segmentation: max = 13.77 and mean = 1.63). This divergence from the other methods can be attributed to the sensitivity of Conv-based methods with regard to artifacts, clearly discernible in the results of the method employing both Conv and the segmentation task. These artifacts may manifest as high-intensity gray-level spikes that exert a considerable influence on the computation of the mean value.

These observations suggest that the addition of the

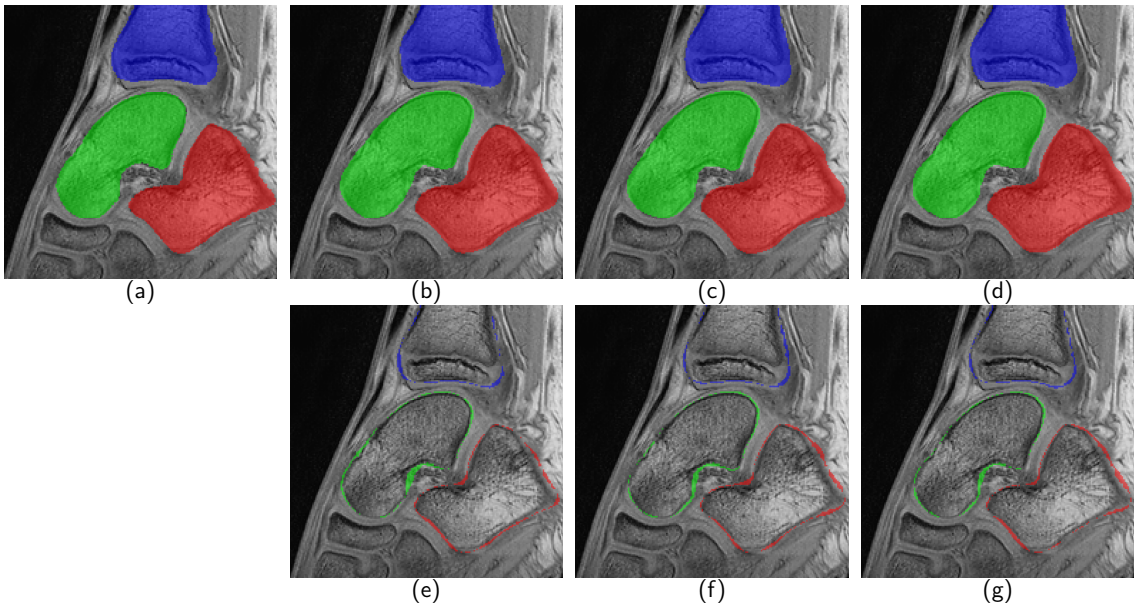

Figure 5: Segmentation results using DRL model on an **equinus subject**. Each segmentation is displayed as an overlay on the corresponding static image. (a) Ground truth segmentation (b) Estimation with DRL - Conv (c) Estimation with DRL - AdaIN (d) Estimation with DRL - FiLM (e) (f) and (g) corresponds to the difference between the ground truth and the above result.

segmentation reduces the variability among the synthesized images, producing a more robust model. The standard deviation images demonstrate the presence of artifacts, which are predominantly observed in the images generated by the methods that employ the Conv and FiLM modules.

5.4 Evaluation of the disentanglement

This section provides an approach to evaluate the disentanglement of the anatomy and modality latent representations for each of the disentangled representation learning scenarios previously introduced. The evaluation uses the DC and IoB metrics introduced in Section 4.2 and the quantitative results are provided in tables 3 and 4. The DC and IoB metrics are designed to assess distinct aspects of the disentanglement: the independence between each representation and the representativeness of these representations. The degree of dependence between the two latent representations is quantified by the DC (see first row in Tables 3 and 4). A lower value indicates a higher degree of independence between the representations. Conversely, both the DC and the IoB metrics computed between each representation and the source image, are designed to reflect the representativeness of these representations with respect to the source image. A higher DC between a given representation and the source image indicates a greater degree of correlation between the two. Additionally, the IoB metric is designed to assess the informativeness of the learned representations with respect to the source image. The results for a dynamic image are presented in Table 3, and for a static image in Table 4.

The method using the AdaIN entanglement module consistently exhibits a higher correlation between the source image and the extracted anatomy representation for both static and dynamic images, with and without the segmentation task. The DC between the image and the modality representation is globally even across all variations of the DRL framework and for both dynamic and static images. The primary impact of the segmentation task lies in the decorrelation between the two latent representations. The introduction of the segmentation task appears to increase the decorrelation rate between the two latent representations for both DRL-AdaIN and DRL-Conv, and both static and dynamic images, suggesting a better independence between the two latent representations, and thus a better disentanglement. In contrast, the introduction of the segmentation module does not yield the same improvement in decorrelation for the DRL-FiLM method as for the other two methods. However, this method produced particularly unrealistic results compared to the other two, suggesting that this entanglement module is less effective for learning disentangled representations than the other two. This would explain the differences in the observed disentanglement trends.

In the case of the IoB, no significant differences were observed in the values obtained with and without the introduction of the segmentation task. The primary insight derived from IoB is that the representations encoded from the static image appear to convey a higher degree of infor-

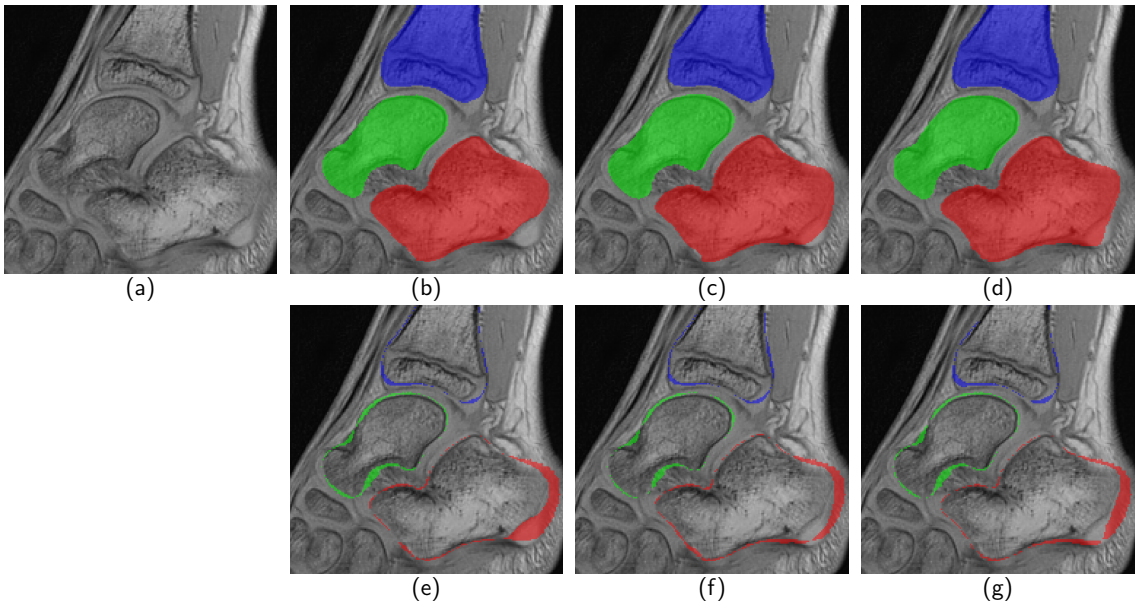

Figure 6: Segmentation results using DRL model on a **typically developing subject**. Each segmentation is displayed as an overlay on the corresponding static image. (a) Ground truth segmentation (b) Estimation with DRL - Conv (c) Estimation with DRL - AdaIN (d) Estimation with DRL - FiLM (e) (f) and (g) corresponds to the difference between the ground truth and the above result.

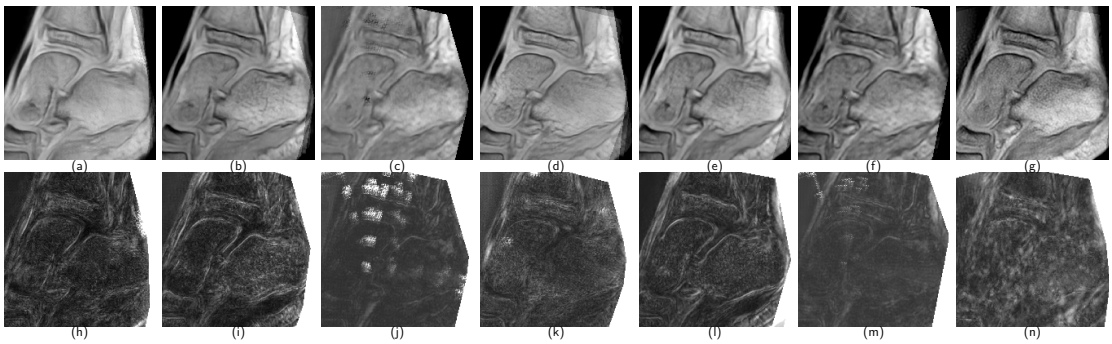

Figure 7: Results of the test-time augmentation for n=10. The first row corresponds to the mean images and the second corresponds to the standard deviation image. (a)(h) DRL - Conv (b)(i) DRL - AdaIN (c)(j) DRL - FiLM (d)(k) DRL - Conv with segmentation (e)(l) DRL - AdaIN with seg (f)(m) DRL - FiLM with seg (g)(n) CycleGAN.

mation about the source image than those encoded from the dynamic images. This observation is consistent with expectations, as static images typically contain a greater quantity of information and details than dynamic images, which are commonly affected by noise and blurring. The DC between the two representations is also lower in static images, and with both representations being more informative, this indicates a greater amount of information carried by the latent representations of the static images and a higher independence between them, and so a better disentanglement on the static images.

6. Discussion and Conclusion

This paper describes a disentangled representation learning framework for unpaired synthesis of high-resolution dynamic MRI. Three entanglement modules are compared, and the impact of an auxiliary segmentation module is investigated. The experiments demonstrate that the addition of an auxiliary segmentation task significantly improves synthesis quality, improves the robustness of the model to spatial transformations. While both the Conv and the AdaIN-based DRL framework produce visually realistic images, the method using the AdaIN module appears to be less prone to artifacts and produces more regular edges. The results demonstrate that, while the DRL methods using the AdaIN module consistently demonstrate the higher correlation between the anatomy latent representation and

Table 3: Evaluation of the disentanglement on a dynamic image for each method. We use Distance Correlation (DC) and Information over Bias introduced in Liu et al. (2021). The best performance is indicated in bold.

	without segmentation			segmentation		
	Conv	AdaIN	FiLM	Conv	AdaIN	FiLM
$\mathrm{DC}(z_x^a, z_x^m) \downarrow$	**0.56 ± 0.03**	0.57 ± 0.03	**0.56 ± 0.02**	**0.52 ± 0.03**	0.56 ± 0.03	0.57 ± 0.02
$\mathrm{DC}(z_x^a, x) \uparrow$	0.77 ± 0.02	**0.85 ± 0.03**	0.81 ± 0.03	0.76 ± 0.02	**0.83 ± 0.02**	0.8 ± 0.03
$\mathrm{DC}(z_x^m, x) \uparrow$	**0.90 ± 0.01**	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.02
$\mathrm{IoB}(z_x^a, x) \uparrow$	1.33 ± 0.03	1.33 ± 0.02	1.33 ± 0.02	1.32 ± 0.03	1.33 ± 0.03	**1.34 ± 0.03**
$\mathrm{IoB}(z_x^m, x) \uparrow$	**1.08 ± 0.06**	1.06 ± 0.02	1.03 ± 0.05	1.05 ± 0.02	**1.06 ± 0.03**	1.1 ± 0.02

Table 4: Evaluation of the disentanglement on a static image for each method. We use Distance Correlation (DC) and Information over Bias introduced in Liu et al. (2021). The best performance is indicated in bold.

	without segmentation			segmentation		
	Conv	AdaIN	FiLM	Conv	AdaIN	FiLM
$\mathrm{DC}(z_y^a, z_y^m) \downarrow$	**0.48 ± 0.03**	0.58 ± 0.04	0.54 ± 0.01	**0.47 ± 0.03**	0.54 ± 0.04	0.56 ± 0.04
$\mathrm{DC}(z_y^a, y) \uparrow$	0.71 ± 0.04	**0.83 ± 0.05**	0.78 ± 0.04	0.71 ± 0.04	**0.8 ± 0.05**	0.78 ± 0.05
$\mathrm{DC}(z_y^m, y) \uparrow$	**0.91 ± 0.03**	0.9 ± 0.03	0.77 ± 0.06	0.89 ± 0.03	0.9 ± 0.03	**0.92 ± 0.03**
$\mathrm{IoB}(z_y^a, y) \uparrow$	1.36 ± 0.04	**1.37 ± 0.07**	1.35 ± 0.05	**1.35 ± 0.04**	1.34 ± 0.06	**1.35 ± 0.05**
$\mathrm{IoB}(z_y^m, y) \uparrow$	1.12 ± 0.05	**1.13 ± 0.05**	**1.13 ± 0.04**	**1.12 ± 0.04**	1.08 ± 0.03	1.09 ± 0.04

the source image, the Conv-based DRL methods consistently exhibit higher disentanglement between the two latent representations. The CycleGAN framework generates a visually realistic texture of the bones, but appears to be more susceptible to irregular anatomical structures and artifacts. Furthermore, it has been demonstrated to be less robust than the DRL approaches.

Acknowledgments

The research leading to these results has received funding from ANR (AI4CHILD ANR-19-CHIA-0015-01), Région Bretagne (DynMRI project), Philips, Fondation de l'Avenir, Paris, France, and Fondation Motrice, Paris, France. This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD010314332 made by GENCI.

Ethical Standards

The studies involving human participants were reviewed and approved by Comité d'Ethique du CHU de Brest. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Conflicts of Interest

We declare we do not have conflicts of interest.

Data availability

Raw data from the Equinus dataset (ID.RCB: 2015-A01409-40) are not publicly available due to privacy and ethical concerns.

References

Alessandro Achille and Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9, San Diego, CA, February 2018. IEEE. ISBN 978-1-72810-124-8. . URL `https://ieeexplore.ieee.org/document/8503149/`.

Henry H. Banks and William T. Green. The Correction of Equinus Deformity in Cerebral Palsy. *JBJS*, 40(6):1359, December 1958. ISSN 0021-9355. URL `https://journals.lww.com/jbjsjournal/Abstract/1958/40060/The_Correction_of_Equinus_Deformity_in_Cerebral.13.aspx`.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. ISSN 1939-3539. . Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Inter-*

national Conference on Learning Representations, February 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Yi-Heng Cao, Vincent Jaouen, Vincent Bourbonne, François Lucia, Nicolas Boussion, Ulrike Schick, Julien Bert, and Dimitris Visvikis. Patient-specific 4DCT respiratory motion synthesis using tumor-aware GANs. Milan, Italy, November 2022. URL https://hal.science/hal-03811270.

Marc-André Carbonneau, Julian Zaïdi, Jonathan Boilard, and Ghyslain Gagnon. Measuring Disentanglement: A Review of Metrics. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. ISSN 2162-2388. . URL https://ieeexplore.ieee.org/abstract/document/9947342. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, December 2019. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841519300684.

Junhua Chen, Leonard Wee, Andre Dekker, and Inigo Bermejo. Improving reproducibility and performance of radiomics in low-dose CT using cycle GANs. *Journal of Applied Clinical Medical Physics*, 23(10):e13739, 2022. ISSN 1526-9914. . URL https://onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13739.

Junhua Chen, Shenlun Chen, Leonard Wee, Andre Dekker, and Inigo Bermejo. Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review. *Physics in Medicine & Biology*, 68(5):05TR01, February 2023. ISSN 0031-9155. . URL https://dx.doi.org/10.1088/1361-6560/acba74. Publisher: IOP Publishing.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, November 2020. URL https://proceedings.mlr.press/v119/chen20j.html. ISSN: 2640-3498.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html.

Sanghyeok Chu, Dongwan Kim, and Bohyung Han. Learning Debiased and Disentangled Representations for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 8355–8366. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/465636eb4a7ff4b267f3b765d07a02da-Abstract.html.

Goran Cobeljic, Marko Bumbasirevic, Aleksandar Lesic, and Zoran Bajin. The management of spastic equinus in cerebral palsy. *Orthopaedics and Trauma*, 23(3):201–209, June 2009. ISSN 1877-1327. . URL https://www.sciencedirect.com/science/article/pii/S1877132709000761.

Kevin de Haan, Yair Rivenson, Yichen Wu, and Aydogan Ozcan. Deep-Learning-Based Image Reconstruction and Enhancement in Optical Microscopy. *Proceedings of the IEEE*, 108(1):30–50, January 2020. ISSN 1558-2256. .

Kien Do and Truyen Tran. Theory and Evaluation Metrics for Learning Disentangled Representations. In *International Conference on Learning Representations*, September 2019. URL https://openreview.net/forum?id=HJgK0h4Ywr.

Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Chris Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised Model Selection for Variational Disentangled Representation Learning. September 2019. URL https://openreview.net/forum?id=SyxL2TNtvr.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A Learned Representation For Artistic Style. In *International Conference on Learning Representations (ICLR)*, page 9, April 2017.

Cian Eastwood and Christopher K. I. Williams. A Framework for the Quantitative Evaluation of Disentangled

Representations. In *International Conference on Learning Representations*, February 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A Variational U-Net for Conditional Appearance and Shape Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Esser_A_Variational_U-Net_CVPR_2018_paper.html.

Yuheng Fan, Hanxi Liao, Shiqi Huang, Yimin Luo, Huazhu Fu, and Haikun Qi. A survey of emerging applications of diffusion probabilistic models in MRI. *Meta-Radiology*, 2(2):100082, June 2024. ISSN 2950-1628. . URL https://www.sciencedirect.com/science/article/pii/S2950162824000353.

Eric Ferreira dos Santos and Alessandra Mileo. From Disentangled Representation to Concept Ranking: Interpreting Deep Representations in Image Classification Tasks. In Irena Koprinska, Paolo Mignone, Riccardo Guidotti, Szymon Jaroszewicz, Holger Fröning, Francesco Gullo, Pedro M. Ferreira, Damian Roqueiro, Gaia Ceddia, Slawomir Nowaczyk, João Gama, Rita Ribeiro, Ricard Gavaldà, Elio Masciari, Zbigniew Ras, Ettore Ritacco, Francesca Naretto, Andreas Theissler, Przemyslaw Biecek, Wouter Verbeke, Gregor Schiele, Franz Pernkopf, Michaela Blott, Ilaria Bordino, Ivan Luciano Danesi, Giovanni Ponti, Lorenzo Severini, Annalisa Appice, Giuseppina Andresini, Ibéria Medeiros, Guilherme Graça, Lee Cooper, Naghmeh Ghazaleh, Jonas Richiardi, Diego Saldana, Konstantinos Sechidis, Arif Canakoglu, Sara Pido, Pietro Pinoli, Albert Bifet, and Sepideh Pashami, editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Communications in Computer and Information Science, pages 322–335, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-23618-1. .

Yabo Fu, Yang Lei, Tonghe Wang, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine &amp$\mathsemicolon$ Biology*, 65(20):20TR01, October 2020. ISSN 0031-9155. . URL https://doi.org/10.1088/1361-6560/ab843e. Publisher: IOP Publishing.

Marc Garetier, Bhushan Borotikar, Karim Makki, Sylvain Brochard, François Rousseau, and Douraïed Ben Salem. Dynamic MRI for articulating joint evaluation on 1.5T and 3.0T scanners: setup, protocols, and real-time sequences. *Insights into Imaging*, 11(1):66, May 2020. ISSN 1869-4101. . URL https://doi.org/10.1186/s13244-020-00868-5.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. . URL http://ieeexplore.ieee.org/document/7780634/.

Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/dc6a70712a252123c40d2adba6a11d84-Abstract.html.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, October 2020. ISSN 0001-0782. . URL https://doi.org/10.1145/3422622.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations, December 2018. URL http://arxiv.org/abs/1812.02230. arXiv:1812.02230 [cs, stat].

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. July 2022. URL https://openreview.net/forum?id=Sy2fzU9gl.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, Septem-

ber 2018. URL https://openreview.net/forum?id=Bklr3j0cKX.

Yanting Hu, Xinbo Gao, Jie Li, Yuanfei Huang, and Hanzi Wang. Single image super-resolution with multi-scale information cross-fusion network. *Signal Processing*, 179:107831, February 2021. ISSN 0165-1684. . URL https://www.sciencedirect.com/science/article/pii/S0165168420303753.

Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-image Translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Xun_Huang_Multimodal_Unsupervised_Image-to-image_ECCV_2018_paper.html.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456. PMLR, June 2015. URL https://proceedings.mlr.press/v37/ioffe15.html. ISSN: 1938-7228.

Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A Simple and Versatile Framework for Image-to-Image Translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 206–222, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58580-8. .

Xilin Jiang, Cong Han, Yinghao Aaron Li, and Nima Mesgarani. Listen, Chat, and Edit: Text-Guided Soundscape Modification for Enhanced Auditory Experience, February 2024. URL http://arxiv.org/abs/2402.03710. arXiv:2402.03710 [cs, eess].

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2649–2658. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kim18b.html. ISSN: 2640-3498.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2014 International Conference on Learning Representations (ICLR)*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Chung-Sheng Lai, Zunzhi You, Ching-Chun Huang, Yi-Hsuan Tsai, and Wei-Chen Chiu. Colorization of Depth Map via Disentanglement. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 450–466, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58571-6. .

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Ledig_Photo-Realistic_Single_Image_CVPR_2017_paper.html.

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: Diverse Image-to-Image Translation via Disentangled Representations. *International Journal of Computer Vision*, 128(10):2402–2417, November 2020. ISSN 1573-1405. . URL https://doi.org/10.1007/s11263-019-01284-z.

Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/ade55409d1224074754035a5a937d2e0-Abstract.html.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2024. URL https://proceedings.

neurips.cc/paper_files/paper/2023/hash/3eaad2a0b62b5ed7a2e66c2188bb1449-Abstract-Conference.html.

Xiao Liu, Spyridon Thermos, Gabriele Valvano, Agisilaos Chartsias, Alison O'Neil, and Sotirios A. Tsaftaris. Measuring the Biases and Effectiveness of Content-Style Disentanglement. In *British Machine Vision Conference (BMVC)*, 2021.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, August 2022. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841522001633.

Yimin Luo, Qinyu Yang, Ziyi Liu, Zenglin Shi, Weimin Huang, Guoyan Zheng, and Jun Cheng. Target-Guided Diffusion Models for Unpaired Cross-Modality Medical Image Translation. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4062–4071, July 2024. ISSN 2168-2208. . URL https://ieeexplore.ieee.org/document/10508481/?arnumber=10508481. Conference Name: IEEE Journal of Biomedical and Health Informatics.

Karim Makki, Bhushan Borotikar, Marc Garetier, Sylvain Brochard, Douraied Ben Salem, and François Rousseau. In vivo ankle joint kinematics from dynamic magnetic resonance imaging using a registration-based framework. *Journal of Biomechanics*, 86:193–203, March 2019. ISSN 1873-2380. .

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.html.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive Learning for Unpaired Image-to-Image Translation. Lecture Notes in Computer Science, pages 319–345, Cham, 2020. Springer International Publishing. ISBN 9783030585457. .

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Parmar_On_Aliased_Resizing_and_Surprising_Subtleties_in_GAN_Evaluation_CVPR_2022_paper.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. . URL https://ojs.aaai.org/index.php/AAAI/article/view/11671. Number: 1.

Roderic I. Pettigrew. Dynamic Cardiac MR Imaging Techniques and Applications. *Radiologic Clinics of North America*, 27(6):1183–1203, November 1989. ISSN 0033-8389. . URL https://www.sciencedirect.com/science/article/pii/S0033838922012052.

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, September 2021. ISSN 0169-2607. . URL https://www.sciencedirect.com/science/article/pii/S0169260721003102.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. .

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Irina Sanchez and Veronica Vilaplana. Brain MRI super-resolution using 3D generative adversarial networks. In *Medical Imaging with Deep Learning (MIDL)*, 2018. URL https://openreview.net/forum?id=rJevSbniM.

Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models, April 2021. URL http://arxiv.org/abs/2104.05358. arXiv:2104.05358 [cs, eess].

Saumya Saxena, Mohit Sharma, and Oliver Kroemer. Multi-Resolution Sensing for Real-Time Control with Vision-Language Models. In *7th Annual Conference on Robot Learning*, August 2023. URL https://openreview.net/forum?id=WuBv9-IGDUA.

Claire Scavinner-Dorval, Rodolphe Bailly, Bhushan Borotikar, Sylvain Brochard, Douraied Ben Salem, and François Rousseau. Analysis of disentangled representation learning for high-resolution dynamic MRI synthesis. In *Medical Imaging 2024: Image Processing*, volume 12926, pages 694–699. SPIE, April 2024. . URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12926/129262U/Analysis-of-disentangled-representation-learning-for-high-resolution-dynamic-MRI/10.1117/12.3006829.full.

Shikun Sun, Longhui Wei, Junliang Xing, Jia Jia, and Qi Tian. SDDM: Score-Decomposed Diffusion Models on Manifolds for Unpaired Image-to-Image Translation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 33115–33134. PMLR, July 2023. URL https://proceedings.mlr.press/v202/sun23n.html. ISSN: 2640-3498.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, December 2007. ISSN 0090-5364, 2168-8966. . URL https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-6/Measuring-and-testing-dependence-by-correlation-of-distances/10.1214/009053607000000505.full. Publisher: Institute of Mathematical Statistics.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised Cross-Domain Image Generation. In *International Conference on Learning Representations*, July 2022. URL https://openreview.net/forum?id=Sk2Im59ex.

Hao Tang, Dan Xu, Yan Yan, Philip H. S. Torr, and Nicu Sebe. Local Class-Specific and Global Image-Level Generative Adversarial Networks for Semantic-Guided Scene Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Tang_Local_Class-Specific_and_Global_Image-Level_Generative_Adversarial_Networks_for_Semantic-Guided_CVPR_2020_paper.html.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-Free Consistent Text-to-Image Generation, February 2024. URL http://arxiv.org/abs/2402.03286. arXiv:2402.03286 [cs].

Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently Controllable Factors, August 2017. URL http://arxiv.org/abs/1708.01289. arXiv:1708.01289 [cs, stat].

L. N. Vaserstein. Markovian processes on countable space product describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969. ISSN 0555-2923.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled Representation Learning, August 2023. URL http://arxiv.org/abs/2211.11695. arXiv:2211.11695 [cs].

McKell Woodland, John Wood, Brian M. Anderson, Suprateek Kundu, Ethan Lin, Eugene Koay, Bruno Odisio, Caroline Chung, Hyunseon Christine Kang, Aradhana M. Venkatesan, Sireesha Yedururi, Brian De, Yuan-Mao Lin, Ankit B. Patel, and Kristy K. Brock. Evaluating the Performance of StyleGAN2-ADA on Medical Images. In Can Zhao, David Svoboda, Jelmer M. Wolterink, and Maria Escobar, editors, *Simulation and Synthesis in Medical Imaging*, Lecture Notes in Computer Science, pages 142–153, Cham, 2022. Springer International Publishing. ISBN 978-3-031-16980-9. .

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html.

Ziqi Yu, Yuting Zhai, Xiaoyang Han, Tingying Peng, and Xiao-Yong Zhang. MouseGAN: GAN-Based Multiple MRI Modalities Synthesis and Segmentation for Mouse Brain Structures. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12901, pages 442–450. Springer International Publishing, Cham, 2021. ISBN 978-3-030-87192-5 978-3-030-87193-2. . URL https://link.springer.com/10.1007/978-3-030-87193-2_42. Series Title: Lecture Notes in Computer Science.

Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C.

Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128, July 2006. ISSN 1053-8119. . URL https://www.sciencedirect.com/science/article/pii/S1053811906000632.

Pengcheng Zhao, Yanxiang Chen, Yang Zhao, Wei Jia, Zhao Zhang, Ronggang Wang, and Richang Hong. Audio-Infused Automatic Image Colorization by Exploiting Audio Scene Semantics, January 2024. URL http://arxiv.org/abs/2401.13270. arXiv:2401.13270 [cs].

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. URL https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html.

Muzaffer Özbey, Onat Dalmaz, Salman U. H. Dar, Hasan A. Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised Medical Image Translation With Adversarial Diffusion Models. *IEEE Transactions on Medical Imaging*, 42 (12):3524–3539, December 2023. ISSN 1558-254X. . URL https://ieeexplore.ieee.org/document/10167641/?arnumber=10167641. Conference Name: IEEE Transactions on Medical Imaging.