Bimodal PET/MRI generative reconstruction based on VAE architectures

Valentin Gautier, Alexandre Bousse, Florent Sureau, Claude Comtat, Voichita Maxim, Bruno Sixou

Bimodal PET/MRI Generative Reconstruction Based on VAE Architectures

V. Gautier[1], A. Bousse[2], F. Sureau[3], C. Comtat[3], V. Maxim[1], B. Sixou[1]

[1] Université de Lyon, INSA-Lyon, UCBL 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France.
[2] Univ. Brest, LaTIM, Inserm UMR 1101, 29238 Brest, France.
[3] BioMaps, Université Paris-Saclay, CEA, CNRS, Inserm, SHFJ, 91401 Orsay, France.

E-mail: `valentin.gautier@creatis.insa-lyon.fr`

Abstract.

- **Objective:** In this study, we explore positron emission tomography (PET)/magnetic resonance imaging (MRI) joint reconstruction within a deep learning (DL) framework, introducing a novel synergistic method.
- **Approach:** We propose a new approach based on a variational autoencoder (VAE) constraint combined with the alternating direction method of multipliers (ADMM) optimization technique. We explore three VAE architectures, joint VAE (JVAE), product of experts (PoE)-VAE and multimodal JS divergence (MMJSD), to determine the optimal latent representation for the two modalities. We then trained and evaluated the architectures on a brain PET/MRI dataset.
- **Main results:** We showed that our approach takes advantage of each modality sharing information to each other, which results in improved peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as compared with traditional reconstruction, particularly for short acquisition times. We find that the one particular architecture, MMJSD, is the most effective for our methodology.
- **Significance:** The proposed method outperforms conventional approaches especially in noisy and undersampled conditions by making use of the two modalities together to compensate for the missing information.

1. Introduction

Medical imaging plays a pivotal role in modern healthcare, enabling non-invasive visualization and assessment of internal anatomical structures and physiological processes. Among the various imaging modalities, positron emission tomography (PET) and magnetic resonance imaging (MRI) are powerful tools that provide complementary information. On one hand, PET is a functional medical imaging method that uses radioactive tracers to observe and monitor various bodily processes, providing crucial information for cancer detection and benefiting fields like cardiology and neurology. On the other hand, MRI is a technique that creates detailed images of the body's anatomy

and physiological processes using powerful magnetic fields, magnetic field gradients, and radio waves. PET systems can be accompanied by an MRI system, providing anatomical information as well as gamma-rays attenuation correction (Catana 2020).

Image reconstruction is an ill-posed inverse problem. Analytical reconstruction methods such as filtered backprojection (Natterer 2001) can be used for PET but the resulting images often suffer from noise amplification, making them impractical for low-statistics acquisitions. Similarly, MRI reconstruction can be performed using an inverse fast Fourier Transform (IFFT); however, this method is ineffective for undersampled acquisitions. To remedy this, model-based iterative reconstruction techniques have been deployed, namely the maximum-likelihood expectation-maximization (MLEM) algorithm (Shepp and Vardi 1982) and its regularized versions (De Pierro 1995; Ahn and Fessler 2003) in PET as well as compressed sensing approaches for undersampled MRI (Fessler 2020).

Furthermore, PET and MRI provide different perspectives of the same object and therefore share some mutual information. This has led to an entire field of research focused on multimodal synergistic reconstruction. The paradigm of synergistic reconstruction is that leveraging mutual information can yield improved reconstruction results with less data, which in turn means lower patient dose from the PET acquisition and faster MRI acquisition. Variational methods have shown that using a combination of modalities can improve the quality of reconstructions (Ehrhardt et al. 2015; Mehranian et al. 2018; Arridge et al. 2021). These methods aim at representing the prior knowledge of the target images with regularization terms that promote structural similarity between the two modalities. However, they often create "artificial similarities" between the PET and the MRI, which is undesirable as these imaging systems have significantly different intrinsic resolutions and specific information.

The use of machine learning (ML) can tackle this issue by training a model that can learn dependencies between several images. Examples from the literature include dictionary learning for PET/MRI (Sudarshan et al. 2020) as well as convolutional dictionary learning in dual-energy computed tomography (Perelli et al. 2022). Dictionary learning techniques are limited to sparse representation using linear mappings, while deep learning (DL) architectures with nonlinear mappings offer more flexibility for more complex manifolds. The obtained compact representation acts as a regularization in inverse problems that outperforms traditional ML methods. For instance convolutional neural network-based magnetic resonance (MR)-guided PET reconstruction (Xie et al. 2021; Schramm et al. 2021) and post-processing (Chen, Gong, et al. 2019; Chen, Toueg, et al. 2021; Da Costa-Luis and Reader 2020; Bousse et al. 2024) deliver promising results.

Among DL methods, variational autoencoders (VAEs) have gained significant attention as powerful generative models capable of learning data encodings and capturing complex relationships within the data. VAEs have been successfully applied in various imaging tasks, facilitating the generation of high-fidelity images and enabling feature disentanglement. In multimodal imaging, accurately representing features that are both common and unique to each modality in a low-dimensional latent space is an

important challenge, necessitating exploration of novel methodologies. Multimodal VAEs achieve this through their loss function which builds a latent space that encodes both the mutual and specific information of each modality (see Suzuki and Matsuo (2022) for a comprehensive review).

We propose in this paper a novel approach to PET/MRI synergistic reconstruction by leveraging the potential of VAEs as a constraint within the alternating direction method of multipliers (ADMM) framework. ADMM is an optimization technique that iteratively decomposes complex problems into simpler subproblems, enabling the incorporation of prior knowledge and constraints. By integrating the VAE as a regularization term, we exploit its ability to learn meaningful latent representations and encourage the reconstruction process to adhere to the underlying data distribution, resulting in more accurate and physiologically plausible reconstructions. Our method can be interpreted as a synthesis method that constrains the solution to be in the range of the decoder of a VAE.

The main contributions of this work are as follows:

- We developed a standard VAE-based model that combines PET and MRI to exploit inter-modality information.
- We propose an ADMM-based synergistic reconstruction algorithm that incorporates our VAE-based model.
- We compare the results obtained using several multimodal VAE architectures and loss functions.
- We evaluate the proposed framework using extensive experiments on data simulated from real clinical brain PET/MRI acquisitions.

The remainder of this paper is organized as follows: Section 2 provides a presentation of the classical VAE and of the multimodal VAE architectures we investigated. Section 3 presents the new reconstruction methodology, detailing the VAE-based ADMM framework, namely deep latent reconstruction (DLR), and its implementation. Section 4 gives the experimental setup and section 5 presents the results, followed by a thorough discussion in Section 6. Finally, Section 7 concludes the paper, summarizing our contributions and outlining future research directions.

2. Variational Autoencoders

We first present the monomodal VAEs and then their multimodal extension. In the following, for some $k \in \mathbb{N}$, 0_k denotes the k-dimensional null vector and $I_k \in \mathbb{R}^{k\times k}$ is the identity matrix in $\mathbb{R}^k$.

2.1. Mono-modal VAE

The VAE is a probabilistic deep latent-variable model (DLVM) that tries to learn the distribution of the data under the assumption that it is generated from an unobserved

latent variable while regularizing the latent space to give it a smooth shape (Diederik P. Kingma and Welling 2019). This enables controlled generation of images from the latent space.

2.1.1. Decoder and Encoder In this paper, $x \in \mathbb{R}^m$ plays the role of a 2-dimensional (2-D) (or 3-dimensional) image with m pixels (or voxels). The goal is to estimate the real probability distribution function (PDF) $p^*\colon \mathbb{R}^m \to \mathbb{R}_+$ of x, based on observations from a training dataset with PDF p_{data}. The conventional approach consists in approximating p^* with a parametrized PDF p_θ, θ being a trainable finite-dimensional parameter. DLVMs change the paradigm by incorporating a generative model $p_\theta(\cdot|z)$ conditioned by a latent variable $z \in \mathbb{R}^d$, $d \ll m$, such that $p_\theta(x)$ is obtained by marginalizing out z:

$$p_\theta(x) = \int_{\mathbb{R}^d} p_\theta(x, z)\, \mathrm{d}z \tag{1}$$

$$p_\theta(x, z) \triangleq p_\theta(x \mid z) \cdot p_0(z) \tag{2}$$

where $p_0(z)$ is a known prior distribution on the latent space chosen to be a standard Gaussian distribution in our case, i.e.,

$$\begin{aligned} z &\sim p_0 \\ &\sim \mathcal{N}(0_d, I_d)\,, \end{aligned}$$

and where the conditional PDF $p_\theta(\cdot|z)$ is assumed to be isotropic Gaussian with known standard deviation η and mean $G_\theta(z) \in \mathbb{R}^m$, i.e.,

$$x \mid z \sim p_\theta(\cdot \mid z) \tag{3}$$

$$\sim \mathcal{N}(G_\theta(z), \eta^2 I_m)\,. \tag{4}$$

The mapping G_θ is often referred to as the *generator*, or *decoder*. Later in the paper (Section 3.3) we will assume that the model is deterministic, i.e., $x = G_\theta(z)$.

The integral in Equation (1) does not have a closed-form and cannot be differentiated with respect to θ. The maximum likelihood estimation is thus intractable (Diederik P Kingma and Welling 2013). Furthermore, the intractability of $p_\theta(x)$ is related to the intractability of the posterior PDF $p_\theta(z|x)$ through the Bayes' rule:

$$p_\theta(z \mid x) = \frac{p_\theta(x, z)}{p_\theta(x)}\,. \tag{5}$$

Thus, a parametric inference model $q_\phi(z|x)$ is introduced to approximate $p_\theta(z|x)$, ϕ being a new trainable parameter, which makes the whole problem tractable. This distribution is chosen to be Gaussian, i.e.,

$$z \mid x \sim q_\phi(\cdot \mid x) \tag{6}$$

$$\sim \mathcal{N}\left(\mu_\phi(x), \sigma_\phi^2(x) I_d\right) \tag{7}$$

where $\mu_\phi(x) \in \mathbb{R}^d$ and $\sigma_\phi(x) \in \mathbb{R}_+^*$. We introduce the *encoder* E_ϕ defined as

$$E_\phi(x) = (\mu_\phi(x), \sigma_\phi(x))\,, \tag{8}$$

where μ_ϕ and σ_ϕ are respectively referred to as the *encoder mean* and the *encoder standard deviation (STD)*.

Finally, G_θ and E_ϕ are represented by two neural networks (NNs) with trainable weights θ and ϕ.

2.1.2. Training The training should be performed by maximum likelihood, i.e., by maximizing $\log p_\theta(x)$ over a training dataset. Denoting by KL the Kullback-Leibler (KL) divergence, it can be shown using Jensen's inequality that the function

$$\mathcal{L}_{\theta,\phi}(x) = \underbrace{\mathop{\mathbb{E}}_{z\sim q_\phi(\cdot|x)}\left[\log\ p_\theta(x \mid z)\right]}_{(i)} - \underbrace{\mathrm{KL}\left(q_\phi(\cdot \mid x) \,\|\, p_0\right)}_{(ii)} \tag{9}$$

is a lower bound for the log-likelihood of the data $\log p_\theta(x)$. Therefore, maximizing the evidence lower bound (ELBO) approximately maximizes $\log p_\theta(x)$ (Diederik P. Kingma and Welling 2019). This function is refereed to as the ELBO, and its negative is used as a loss function. We can distinguish two terms in the ELBO (9):

(i) A data fidelity term that aims at generating data according to the distribution $p_\theta(x)$; the expectation is computed by sampling z following $q_\phi(\cdot|x)$.

(ii) A regularization term that acts on the latent space encouraging the encoder to generate latent variables according to the prior distribution p_0.

Finally, the best parameters $\hat{\theta}$ and $\hat{\phi}$ are learned by maximizing $\mathcal{L}_{\theta,\phi}$ for x drawn from an empirical PDF p_{data} that corresponds to the training dataset:

$$(\hat{\theta}, \hat{\phi}) \in \arg\max_{\theta,\phi} \mathop{\mathbb{E}}_{x\sim p_{\text{data}}}\left[\mathcal{L}_{\theta,\phi}(x)\right]. \tag{10}$$

Solving (10) is generally achieved with a stochastic gradient ascent where the gradients are obtained by back-propagation through the NNs G_θ and E_ϕ. The resulting PDF $p_{\hat{\theta}}$ is then used as an approximation for the true PDF p^*.

VAEs have the following properties:

- Continuity: if two points are close within the latent space, their decoded representations should also be close to each other. More specifically, for any sequence (z_p) such that $z_p \to z$, we have $G_\theta(z_p) \to G_\theta(z)$.
- Completeness: selecting a random point from the latent space distribution should yield a result that aligns with the distribution p_{data} of the data.

The KL regularization term in the ELBO (9) helps to smooth and compact the VAE's latent space. Figure 1 illustrates this aspect on a 2-D latent space example. The images (triangle, square and circle) are represented in the latent space by Gaussian PDFs given by the encoder mean and the encoder STD. Without regularization, the PDFs are isolated and decoding an image from a random z will produce an image that is not useful (i.e., the purple spiral). The regularization term brings these PDFs together. As a result, moving z around the centers of these Gaussian PDFs generates small variations of the originally encoded figures (cf. the gray shape that is a "mixture" of the three images).

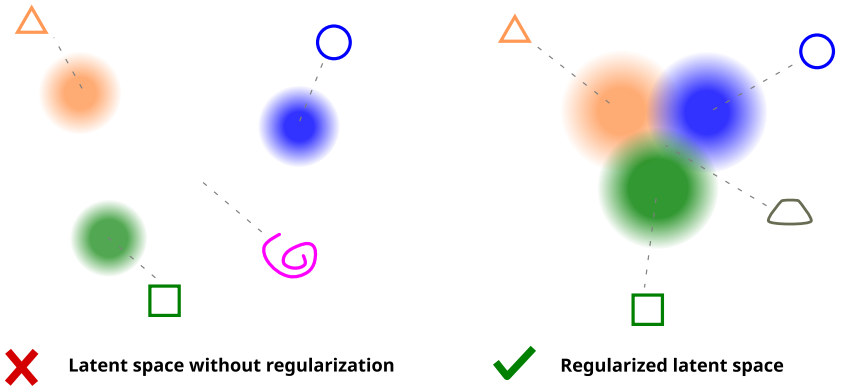


Figure 1: Illustration of the VAE's regularized latent space.

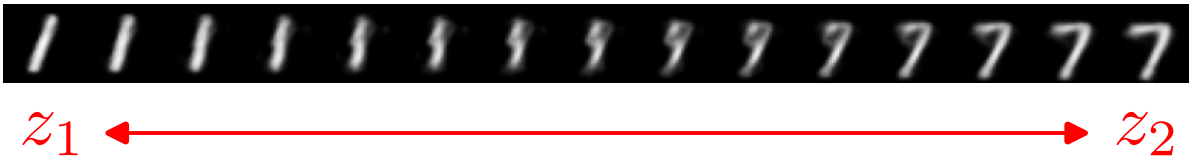


Figure 2: Illustration of the continuity and completeness on an MNIST example. The images were generated from latent variables on the segment $\{(1-t)\cdot z_1 + t\cdot z_2\,; t\in[0,1]\}$ where z_1 and z_2 were obtained by encoding x_1 and x_2 respectively. Shifting t from 0 to 1 progressively transitions from x_1 to x_2.

We further illustrate these properties in Figure 2 with images from the Modified National Institute of Standards and Technology (MNIST) dataset (Deng 2012). Two MNIST images x_1 and x_2 were encoded with the encoder mean μ_ϕ as $z_1 = \mu_\phi(x_1)$ and $z_2 = \mu_\phi(x_2)$, and we displayed the decoded images $G_\theta((1-t)\cdot z_1 + t\cdot z_2)$, $t\in[0,1]$. We observe that by progressively shifting t from 0 to 1 the resulting decoded image progressively transitions from x_1 to x_2, which illustrates the smoothness of the decoder.

In practice however, the β-VAE (Higgins et al. 2017) is preferred over the classical VAE to control the regularization in latent space. With this model, the ELBO is rewritten as :

$$\mathcal{L}^{\beta}_{\theta,\phi}(x) = \mathop{\mathbb{E}}_{z\sim q_\phi(\cdot|x)}\left[\log\ p_\theta(x\mid z)\right] - \beta\mathrm{KL}\left(q_\phi(\cdot\mid x)\parallel p_0\right). \tag{11}$$

The introduction of the hyperparameter β allows to control the regularization the latent space and to favor the disentanglement of the latent variable (Higgins et al. 2017). If β is too large, the model will collapse into one general distribution while z will be decoded into a mean blurry image. Conversely if β is too small, we fall back to the classical autoencoder with no regularization over the latent space. We tuned this parameter manually so that the KL divergence and the data fidelity term are about the same magnitude at the end of the training.

2.2. Multimodal VAE

VAEs are well-suited to handling multimodal data as they are able to compress multiple modalities into one shared latent variable that can extract the mutual information in a lower-dimensional common space. This latent space is often utilized for feature extraction or classification (Cheng et al. 2021). The framework used for monomodal imaging can be extended to N-modal imaging considering a collection $X = (x_1, \dots, x_N)$ of N random vectors in $\mathbb{R}^m$. We assume that X follows a true distribution $p^*: \mathbb{R}^{m \times N} \to \mathbb{R}_+$. The goal is to train a mutimodal generative model $G_\theta^{\text{mult}} = (G_\theta^1, \dots, G_\theta^N): \mathbb{R}^d \to \mathbb{R}^{m \times N}$ with a single latent variable $z \in \mathbb{R}^d$ that represents the common information from the N modalities while still being able to represent each modality x_n individually. The model assumes that the x_ks are independent conditionally to z:

$$p_\theta(X \mid z) = \prod_{k=1}^{N} p_\theta(x_k \mid z) \tag{12}$$

with

$$x_k \mid z \sim \mathcal{N}(G_\theta^k(z), \eta^2 I_m)\,. \tag{13}$$

Again, the conditional PDF $p_\theta(z|x_k)$ is approximated with a variational distribution $q_\phi(z \mid x_k)$ with ϕ being a trainable parameter.

In the following we introduce several multimodal VAEs. These models encode the multimodal image X into a single latent variable z from which a multimodal image is generated. Two encoding strategies can be considered: (i) using a single encoder to process all modalities together (Section 2.2.1) or (ii) using one encoder per modality (Section 2.2.2 and Section 2.2.3), thus providing more flexibility at the expense of an increased number of parameters.

2.2.1. Joint Variational Autoencoder (JVAE) The most straightforward approach is the joint VAE (JVAE) which consists of training the parameters with the following ELBO:

$$\mathcal{L}_{\theta,\phi}(X) = \mathop{\mathbb{E}}_{z \sim q_\phi(\cdot|X)} \left[\sum_{k=1}^{N} \log p_\theta(x_k \mid z) \right] - \mathrm{KL}\left(q_\phi(\cdot \mid X) \parallel p_0\right) \tag{14}$$

where $q_\phi(z|X)$ is a multichannel encoder distribution, which can be implemented using a multimodal encoder $E_\phi^{\text{mult}}(X) = (\mu_\phi(X), \sigma_\phi(X))$ such that

$$\begin{aligned} z \mid X &\sim q_\phi(\cdot \mid X) \\ &\sim \mathcal{N}\left(\mu_\phi(X), \sigma_\phi^2(X) I_d\right)\,. \end{aligned}$$

In practice, this can be implemented using one encoder that will encode the different modalities together. This approach, which was utilized by Pinton et al. (2024), is memory efficient since it needs only one encoder but may result in a loss of modality-specific features.

2.2.2. Product of Experts (PoE) and Mixture of Experts (MoE) Another approach proposed by Wu and Goodman (2018) consists in defining one encoder distribution $p_\theta(z|x_k)$ per modality k and to define the multichannel encoder distribution using the product of experts (PoE):

$$q_\phi\left(z \mid X\right) \propto p_0(z) \cdot \prod_{k=1}^{N} \tilde{q}_\phi\left(z \mid x_k\right), \tag{15}$$

where the functions $\tilde{q}_\phi(z \mid x_k)$ are the unimodal inference distributions, referred to as the *experts*. With this formulation, each encoder predicts a modality specific distribution. The PoE is then computed from the experts' distributions, usually defined as Gaussian PDFs, and the joint latent variable z is sampled from (15). This formulation is numerically convenient as the product of Gaussian PDFs is still a Gaussian (up to a normalization constant) and the mean $\mu_\phi(X)$ and STD $\sigma_\phi(X)$ are computed from the mean $\mu_\phi(x_k)$ and STD $\sigma_\phi(x_k)$ of $\tilde{q}_\phi\left(\cdot \mid x_k\right)$, $k = 1, \ldots, K$. One issue is that, due to the multiplicative nature of this model, a single expert can overshadow all others. This can be an issue since the goal of synergistic reconstruction is to obtain balanced influence of the various modalities. The training is also achieved by maximizing the ELBO (14).

Another way to make use of those experts is by using a mixture of experts (MoE) instead of a product (Shi et al. 2019):

$$q_\phi\left(z \mid X\right) = \sum_{k=1}^{N} \pi_k \tilde{q}_\phi(z \mid x_k) \tag{16}$$

where $\pi_k \in [0, 1]$, $\sum \pi_k = 1$. Sampling z is achieved by drawing an integer in $\{1, \ldots, N\}$ with probabilities π_k, $k = 1, \ldots, N$, then by sampling a Gaussian of parameter $(\mu_\phi(x_k), \sigma_\phi(x_k))$. The resulting VAE allows to better learn each expert individually, leading to better performances in case of missing modalities compared to the PoE.

2.2.3. Mixture of Expert Multimodal Jensen-Shannon Divergence (MMJSD) To get the best of both PoE and MoE, Sutter et al. (2020) proposed to replace the KL divergence in the ELBO (14) with a Jensen-Shannon (JS) divergence. The ELBO then becomes:

$$\mathcal{L}_{\theta,\phi}(X) = \underset{z \sim q_\phi(\cdot|X)}{\mathbb{E}} \left[\log\ p_\theta(X \mid z)\right] - \mathrm{JS}_{\Pi}^{N+1}\left(\left\{\{\tilde{q}_\phi(\cdot \mid x_k)\}_{k=1}^{N}, p_0\right\}\right) \tag{17}$$

where:

- $\Pi = \{\pi_k\}_{k=1}^{N+1} \in \mathbb{R}_+^{N+1}$, $\sum_k \pi_k = 1$, is a $(N+1)$-tuple of weights;
- $\mathrm{JS}_{\Pi}^{N+1}\left(\{q_k\}_{k=1}^{N+1}\right) = \sum_{k=1}^{N+1} \pi_k \mathrm{KL}\left(q_k \,\middle\|\, f_{\mathcal{M}}\left(\{q_\nu\}_{\nu=1}^{N+1}\right)\right)$ (in our case $q_\nu = \tilde{q}_\phi(\cdot \mid x_\nu)$ for $\nu = 1, \ldots, N$ and $q_{N+1} = p_0$);
- $f_{\mathcal{M}}$ is an abstract mean function that maps a collection of $N+1$ PDFs to a single PDF;
- $q_\phi\left(z \mid X\right) \propto \prod_{k=1}^{N} \tilde{q}_\phi\left(z \mid x_k\right)$.

Using the PoE as the abstract mean function $f_{\mathcal{M}}$ corresponds to computing KL divergences between Gaussian distributions. In the following we will also add a weighting factor β on the JS divergence to balance it with the data fidelity term in the same way we do with the β-VAE. This architecture is referred to as multimodal JS divergence (MMJSD).

3. Deep Latent Reconstruction

In this subsection we detail the bimodal inverse problem formulation for PET/MRI. The reconstruction problem is formulated as an optimization problem under constraints implemented by the decoder. A DLR method is proposed for the numerical resolution. We denote $X = (x_{\text{pet}}, x_{\text{mr}}) \in \mathbb{R}^{m\times 2}$ the bimodal image to reconstruct, where we assumed both images have m pixels (or voxels).

3.1. PET Model

The PET system matrix $P \in \mathbb{R}^{n\times m}$, n denoting the number of detector pairs, is defined such that each entry $[P]_{i,j}$ is the probability that a positron emission at the jth pixel (or voxel) is detected at the ith detector pair, taking into account physical factors such as the attenuation. The expected counts vector is

$$\bar{y}_{\text{pet}} = \alpha \cdot (Px_{\text{pet}} + r + s), \tag{18}$$

where $r \in \mathbb{R}^n$ and $s \in \mathbb{R}^n$ are the expected number of randoms and scatters respectively. $[x_{\text{pet}}]_j$ is the mean number of pairs produced at voxel j and $\alpha > 0$ is a scaling factor that encompasses the acquisition time and/or detector sensitivity. The observed PET counts are stored in a vector $y_{\text{pet}} \in \mathbb{R}^n$.

Assuming the counts are independent and follow a Poisson distribution, denoted $p_{\text{pet}}(y_{\text{pet}}|x_{\text{pet}})$, the PET data fidelity term is related to the negative log-likelihood by

$$\mathscr{D}_{\text{pet}}(x_{\text{pet}}) = -\log p_{\text{pet}}(y_{\text{pet}} \mid x_{\text{pet}}) + C \tag{19}$$

$$= \sum_{i=1}^{n} ([\bar{y}_{\text{pet}}]_i - [y_{\text{pet}}]_i \log([\bar{y}_{\text{pet}}]_i)) \tag{20}$$

(with the convention $0 \cdot \log 0 = 0$) where C is a constant independent of x_{pet}. PET reconstruction is then achieved by minimizing $\mathscr{D}_{\text{pet}}(x_{\text{pet}})$, for example by MLEM (Shepp and Vardi 1982) or modified versions to incorporate a convex penalty (De Pierro 1995).

3.2. MRI Model

The MRI model is

$$\bar{y}_{\text{mr}} = Ex_{\text{mr}}, \tag{21}$$

where $\bar{y}_{\text{mr}} \in \mathbb{R}^p$ are the expected K-space data and $E \in \mathbb{C}^{p\times m}$ is the MR forward operator defined as

$$E = UF, \tag{22}$$

where $F \in \mathbb{C}^{m\times m}$ is the discrete Fourier transform matrix and $U \in \mathbb{R}^{p\times m}$ performs a radial subsampling of factor $R = m/p$.

Assuming the MR measurement $y_{\text{mr}} \in \mathbb{R}^p$ follows an isotropic Gaussian distribution $p_{\text{mr}}(y_{\text{mr}}|x_{\text{mr}})$ with variance σ^2_{mr} and centered on $\bar{y}_{\text{mr}}$, the data fidelity loss is

$$\begin{aligned}\mathscr{D}_{\text{mr}}(x_{\text{mr}}) = & -\log p_{\text{mr}}(y_{\text{mr}} \mid x_{\text{mr}}) + C' && (23)\\ = & \frac{1}{2\sigma^2_{\text{mr}}}\|Ex_{\text{mr}} - y_{\text{mr}}\|_2^2\,, && (24)\end{aligned}$$

where C' is a constant independent of x_{mr}. Similarly to PET, $\mathscr{D}_{\text{mr}}(x_{\text{mr}})$ can be minimized with iterative techniques (Fessler 2020).

3.3. Proposed Joint PET/MRI Reconstruction Framework

To solve the joint PET/MRI reconstruction inverse problem, we proceed using a maximum *a posteriori* (MAP) framework with a learned prior approximating the distribution p^* of $X = (x_{\text{pet}}, x_{\text{mr}})$. The multimodal generator produces $N = 2$ images (PET and MR); we introduce the notations $G^{\text{pet}}_\theta = G^1_\theta$, $G^{\text{mr}}_\theta = G^2_\theta$ and $G^{\text{mult}}_\theta \triangleq (G^{\text{pet}}_\theta, G^{\text{mr}}_\theta)\colon \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^m$. Instead of directly maximizing the joint posterior $p(X|Y)$ with $X = (x_{\text{pet}}, x_{\text{mr}})$ and $Y = (y_{\text{pet}}, y_{\text{mr}})$, the maximization will be performed through the latent space of a VAE, as proposed previously by Bora et al. (2017). We thus maximize the posterior $p(X, z \mid Y)$. By the Bayes' law we have:

$$p(X, z \mid Y) = p(Y \mid X)p_\theta(X \mid z)p_0(z)/p(Y) \tag{25}$$

(when X is known, z do not provide any additional information for Y so that $p(Y|X) = p(Y|X, z)$). Moreover, the noise realizations on the PET and MR acquisitions being independent and we obtain:

$$\begin{aligned}-\log\, p(X, z \mid Y) = & \mathscr{D}_{\text{pet}}(x_{\text{pet}}) + \mathscr{D}_{\text{mr}}(x_{\text{mr}}) - \log p_\theta(X \mid z) && (26)\\ & + \frac{1}{2}\|z\|_2^2 + \log p(Y)\,. && (27)\end{aligned}$$

Now, letting $\eta \to 0$ in the generative model (13), $p_\theta(X \mid z)$ becomes a Dirac distribution centered on $G^{\text{mult}}_\theta(z)$. Let $\mathscr{C}$ be the set defined as

$$\mathscr{C} = \{(x_{\text{pet}}, x_{\text{mr}}, z) \mid (x_{\text{pet}}, x_{\text{mr}}) = G^{\text{mult}}_\theta(z)\}. \tag{28}$$

The MAP optimization problem can be now rewritten as a constrained optimization problem on $\mathscr{C}$,

$$(\hat{x}_{\text{pet}}, \hat{x}_{\text{mr}}, \hat{z}) = \underset{(x_{\text{pet}}, x_{\text{mr}}, z)\in\mathscr{C}}{\arg\min}\ \mathscr{D}_{\text{pet}}(x_{\text{pet}}) + \mathscr{D}_{\text{mr}}(x_{\text{mr}}) + \frac{1}{2}\|z\|_2^2. \tag{29}$$

In our case, the regularization term $\frac{1}{2}\|z\|_2^2$ has no real impact on the results. For the sake of simplicity, it was not considered in the following. Solving (29) can be achieved using the ADMM framework (Boyd et al. 2010). Denoting by $\mu = (\mu_{\text{pet}}, \mu_{\text{mr}})$ the Lagrange multiplier, the bimodal DLR algorithm for iteration q is:

$$x^{(q+1)}_{\text{pet}} = \underset{x_{\text{pet}}}{\arg\min}\left(\mathscr{D}_{\text{pet}}(x_{\text{pet}}) + \frac{\rho_{\text{pet}}}{2}\left\|x_{\text{pet}} - G^{\text{pet}}_\theta\left(z^{(q)}\right) + \mu^{(q)}_{\text{pet}}\right\|_2^2\right) \tag{30}$$

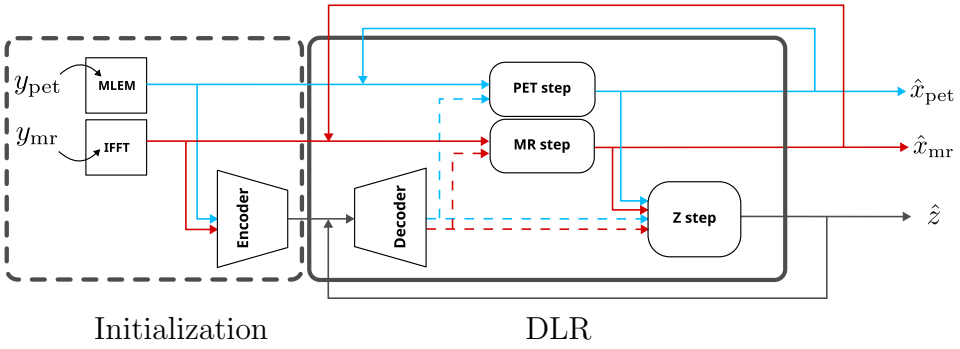


Figure 3: Representation of the DLR algorithm. Solid blue arrows represent the currently reconstructed PET image and solid red arrows represent the currently reconstructed MRI. Dashed arrows represent the current prediction of the decoder for the corresponding modality. The grey arrows represent the current latent variable.

$$x_{\text{mr}}^{(q+1)} = \arg\min_{x_{\text{mr}}} \left(\mathscr{D}_{\text{mr}}(x_{\text{mr}}) + \frac{\rho_{\text{mr}}}{2} \left\| x_{\text{mr}} - G_\theta^{\text{mr}}\left(z^{(q)}\right) + \mu_{\text{mr}}^{(q)} \right\|_2^2 \right) \tag{31}$$

$$z^{(q+1)} = \arg\min_{z} \left\| G_\theta^{\text{mult}}(z) - \left(X^{(q+1)} + \mu^{(q)}\right) \right\|_2^2 \tag{32}$$

$$\mu^{(q+1)} = \mu^{(q)} + X^{(q+1)} - G_\theta^{\text{mult}}\left(z^{(q+1)}\right) \tag{33}$$

where we denoted $X^{(q)} = \left(x_{\text{pet}}^{(q)}, x_{\text{mr}}^{(q)}\right)$, $\mu^{(q)} = \left(\mu_{\text{pet}}^{(q)}, \mu_{\text{mr}}^{(q)}\right)$. We used a surrogate function for optimization transfer to solve (30) (De Pierro 1995), while we used a conjugate gradient algorithm to solve the quadratic problem (31) (details in Appendix A). The inner iteration number in (31) was chosen to achieve a residual error below 10^{-5}. The latent variable update (32) is performed with gradient descent. We use the Adam algorithm (Diederik P. Kingma and Ba 2017) that includes tuning of the step size. The initialization is performed with first-guess images obtained with 10 iterations of MLEM for $x_{\text{pet}}^{(0)}$ and IFFT for $x_{\text{mr}}^{(0)}$ while $z^{(0)}$ is obtained by applying the encoder to $x_{\text{pet}}^{(0)}$ and $x_{\text{mr}}^{(0)}$. The workflow of the DLR algorithm is represented in Figure 3.

A practical issue that needs to be addressed is the scaling of the decoder G_θ^{mult}. It is common practice to train the decoder on standardized data (with a mean of 0 and STD of 1 across the pixel values). We standardized the images for training as follows:

$$x_{\text{mod}}^{\text{standard}} = \frac{1}{\text{std}(x_{\text{mod}})}(x_{\text{mod}} - \text{mean}(x_{\text{mod}}))\,, \quad \text{mod} \in \{\text{pet}, \text{mr}\}\,, \tag{34}$$

where the mean and the STD are computed separately for each image x_{mod}. However, the reconstructed images are not necessarily standardized. Thus, we have to rescale independently the decoded images during the iterations to match with the current reconstructions. To do so, we unstandardize the output of G_θ^{mult} to match the current reconstruction of PET and MR images at each iteration q as follows::

$$G_\theta^{\text{mod}}\left(z^{(q)}\right) = \widetilde{G}_\theta^{\text{mod}}\left(z^{(q)}\right) \cdot \text{std}\left(x_{\text{mod}}^{(q)}\right) + \text{mean}\left(x_{\text{mod}}^{(q)}\right)\,, \tag{35}$$

$$\text{with} \quad \text{mod} \in \{\text{pet}, \text{mr}\} \tag{36}$$

where $\widetilde{G}_\theta^{\text{mod}}\left(z^{(q)}\right)$ is the actual output of the decoder and $G_\theta^{\text{mod}}\left(z^{(q)}\right)$ is the rescaled output Other methods to tackle this include histogram equalization (Gonzalez et al. 2009).

The ADMM penalty parameters ρ_{pet} and ρ_{mr} have a strong influence on the optimization and are often empirically tuned based on validation data. In this work, we implemented the adaptive update scheme proposed in (Wohlberg 2017) for linear constraints. This scheme consists in balancing the relative primal and dual residuals while computing the residuals separately then updating ρ_{pet} and ρ_{mr}. We have also implemented the stopping criterion proposed in the same work with $\epsilon = 0.02$.

4. Numerical Experiments

This section describes our implementation of the VAEs, DLR as well as the experiments carried out to evaluate them.

4.1. Dataset and Training

A collection of 48 brain [^{18}F]FDG PET/ T1-weighted MRI volumes acquired on a SIGNA PET/MRI system (GE Healthcare) at the Service Hospitalier Frédéric Joliot, Orsay, France, was used in this study. PET and MR volumes were rigidly co-registered using the DICOM header parameters. Each volume was resized into a stack of 20 256×256 slices, for a total of 960 PET/MRI slice pairs. Each pair $x_{\text{pet}}, x_{\text{mr}}$ correspond to the same slice. We used 900 pairs to train the models and 60 for testing. This data partition does not separate the patients, i.e., some slices from the testing dataset may correspond to a different slice from a patient already seen in the training dataset. This is a deliberate choice as our training dataset is not large enough to allow the trained network to fully generalize. This aspect will be further discussed in Section 6 and in Appendix B.

The models we considered are the following three multimodal VAEs: JVAE, PoE-VAE and MMJSD. We examined the effect of varying the dimensionality d of the latent space. To accomplish this, we have trained two versions of each VAE: one with $d = 64$ and another with $d = 32$. Additionally, monomodal VAEs, i.e., using either PET or MR images, were considered in order to assess the benefit of reconstructing two modalities simultaneously.

The structure of the VAEs is shown in Figure 4 for the JVAE and Figure 5 for PoE-VAE and MMJSD. The NNs are built as stacks of ConvBlocks and DeconvBlocks. A ConvBlock consist of a convolution layer, batch normalization layer and activation layer with the ReLU activation function. The DeconvBlocks are similar and only replace the convolution layer with a transposed convolution layer. Each convolution and transposed convolution is described with its format (number of channels, filter size, stride) and the dense layers are described by their number of neurons. The sampling layer is used to perform the re-parametrization trick (Diederik P Kingma and Welling 2013) and takes a mean vector and a standard deviation vector to sample a latent variable z.

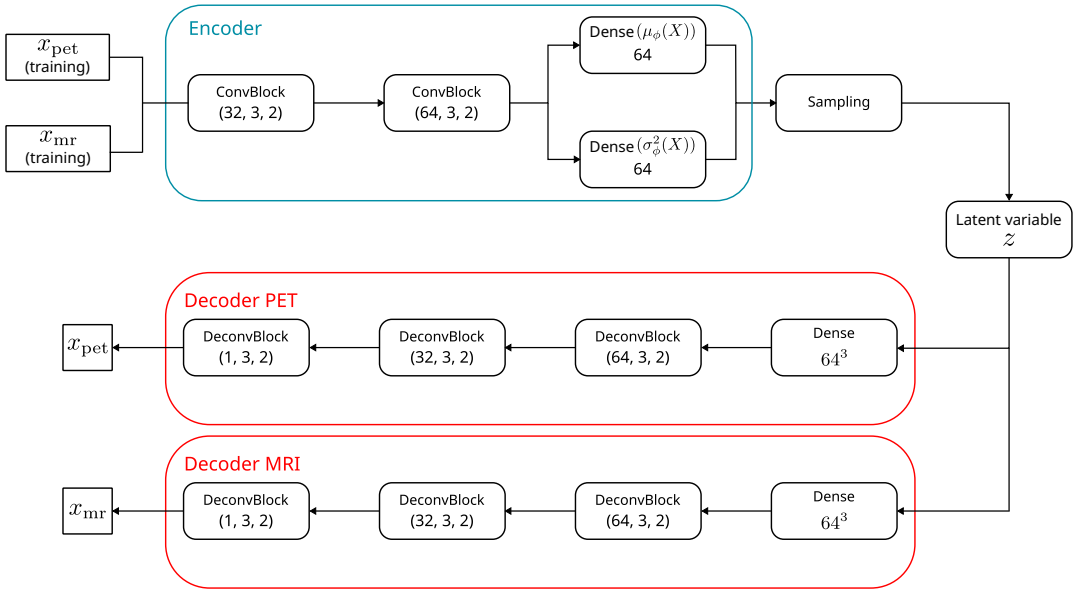


Figure 4: Architecture of the joint VAE.

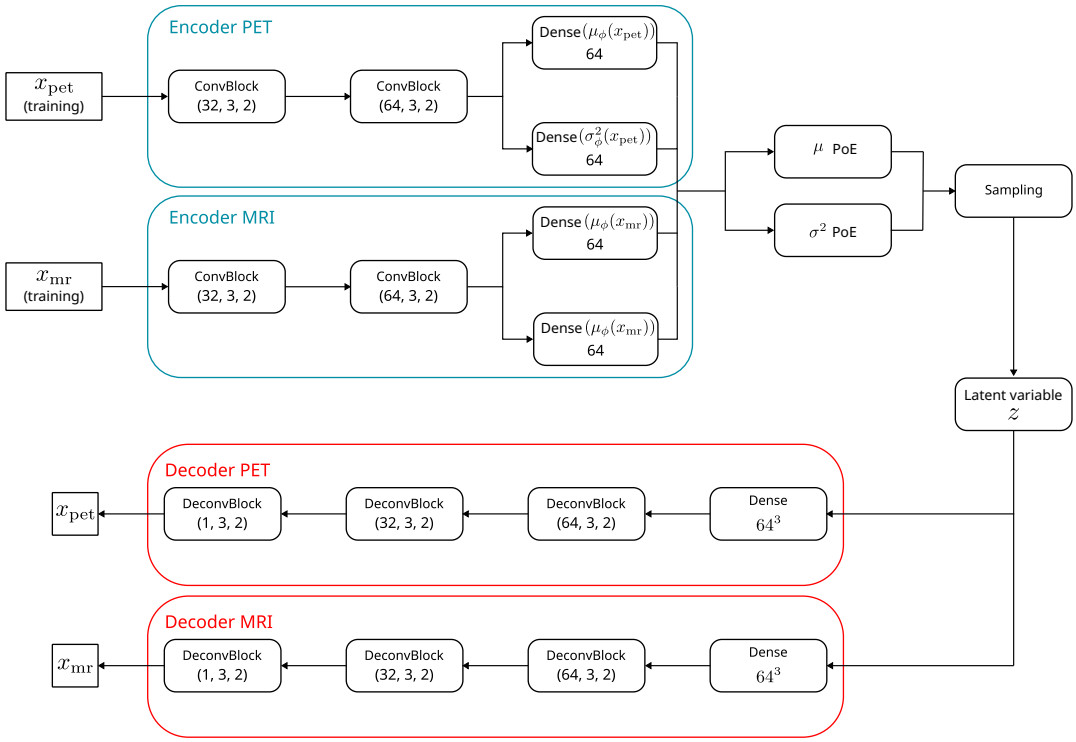


Figure 5: Architecture of the PoE-VAE and MMJSD.

The training was implemented using the open-source library Keras 2.2.5 with Tensorflow backbone and performed with an NVIDIA RTX A2000 mobile. The network was trained for 500 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 64.

4.2. Experiment 1: Image Generation

An important aspect to evaluate is the ability of the trained generative model $G_\theta^{\text{mult}} = (G_\theta^{\text{pet}}, G_\theta^{\text{mr}})$ to generate PET/MR images that consistently represent the same brain. Given a PET/MR target image pair $(x_{\text{pet}}^\star, x_{\text{mr}}^\star)$ from the testing dataset, a latent variable z_{pet} is computed by maximizing the posterior conditionally to $x_{\text{pet}}^\star$, and another latent variable z_{mr} latent variable is computed by maximizing the posterior conditionally to $x_{\text{mr}}^\star$, i.e

$$\begin{aligned}\hat{z}_{\text{pet}} \in & \arg\max_z p_\theta(z \mid x_{\text{pet}}^\star) \\ = & \arg\min_z \frac{1}{2\eta^2} \left\| x_{\text{pet}}^\star - G_\theta^{\text{pet}}(z) \right\|_2^2 + \frac{1}{2}\|z\|_2^2 \end{aligned} \tag{37}$$

and

$$\begin{aligned}\hat{z}_{\text{mr}} \in & \arg\max_z p_\theta(z \mid x_{\text{mr}}) \\ = & \arg\min_z \frac{1}{2\eta^2} \|x_{\text{mr}}^\star - G_\theta^{\text{mr}}(z)\|_2^2 + \frac{1}{2}\|z\|_2^2 . \end{aligned} \tag{38}$$

For a deterministic generator, as the one considered in this paper, η tends to zero and instead we solve

$$\hat{z}_{\text{pet}} \in \arg\min_z \left\| x_{\text{pet}}^\star - G_\theta^{\text{pet}}(z) \right\|_2^2 \tag{39}$$

and

$$\hat{z}_{\text{mr}} \in \arg\min_z \|x_{\text{mr}}^\star - G_\theta^{\text{mr}}(z)\|_2^2 . \tag{40}$$

Thus, the latent variable $\hat{z}_{\text{pet}}$ is estimated from the PET image only and $\hat{z}_{\text{mr}}$ is estimated from the MR image only, and images $G_\theta^{\text{pet}}(\hat{z}_{\text{pet}})$ and $G_\theta^{\text{mr}}(\hat{z}_{\text{mr}})$ are respectively the *PET-fitted* image and *MR-fitted* image, that is to say, they represent the best separate predictions of $x_{\text{pet}}^\star$ and $x_{\text{mr}}^\star$ by the model $G_\theta^{\text{mult}} = (G_\theta^{\text{pet}}, G_\theta^{\text{mr}})$. If G_θ^{mult} is properly trained to represent consistent image pairs, we expect to have

$$G_\theta^{\text{mr}}(\hat{z}_{\text{mr}}) \approx x_{\text{mr}}^\star \quad \text{and} \quad G_\theta^{\text{pet}}(\hat{z}_{\text{mr}}) \approx x_{\text{pet}}^\star , \tag{41}$$

where $G_\theta^{\text{mr}}(\hat{z}_{\text{pet}})$ is the *predicted MR image* from the PET image, and conversely $G_\theta^{\text{pet}}(\hat{z}_{\text{mr}})$ is the *predicted PET image* from the MR image. We performed this test on the three VAEs with $d = 32$ and $d = 64$.

4.3. Experiment 2: Image Reconstruction

We used a collection of PET/MR $(x_{\text{pet}}^\star, x_{\text{mr}}^\star)$ as ground truths (GTs) to generate raw data for reconstruction. First, the [^{18}F]FDG distributions $x_{\text{pet}}^\star$ are normalized, i.e, $\|x_{\text{pet}}^\star\|_1 = 1$. Then the PET raw data y_{pet} are generated as

$$y_{\text{pet}} \sim \text{Poisson}\left(\alpha(Px_{\text{pet}}^\star + r + s)\right) , \tag{42}$$

for different values of the dose-related parameter $\alpha > 0$, thus allowing to change the statistics. The projector P and its transpose were implemented using the ASTRA

Toolbox (Aarle et al. 2016). Attenuation factors were ignored. While randoms can be corrected for by real-time subtraction of a delayed coincidence channel (Knoll 2010), scatter can be estimated by simulations (Watson et al. 1996) or learned with a deep architecture (Laurent et al. 2023). In this work, we assumed $r+s$ to be a constant vector that was adjusted to account for approximately 30% of the counts.

The MR raw data y_{mr} were generated as

$$y_{\text{mr}} = Ex^{\star}_{\text{mr}} + \epsilon\,, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{\text{mr}})\,, \tag{43}$$

where σ_{mr} is adjusted to add 5% of the measured data as noise and for different values of the subsampling factor $R = m/p$ (see 22).

PET and MR images were jointly reconstructed using the bimodal and monomodal DLR methods described in Section 3.3 with the different settings described in Section 4.1 (JVAE, PoE-VAE and MMJSD, $d = 64$). The monomodal versions of DLR were implemented using monomodal VAEs. Additionally, they were individually reconstructed with (isotropic) total variation (TV) regularization as:

$$x_{\text{mod}} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^m} \mathscr{D}_{\text{mod}}(x) + \gamma \| \, |\nabla x| \, \|_1\,, \quad \text{mod} \in \{\text{pet}, \text{mr}\} \tag{44}$$

where $\nabla\colon \mathbb{R}^m \to \mathbb{R}^{2\times m}$ is the 2-D discrete gradient and $|\cdot|\colon \mathbb{R}^{2\times m} \to \mathbb{R}^m$ is the ℓ^2-norm applied on each pixel. We used the MLEM-TV algorithm from (Sawatzky et al. 2008) for PET and the Chambolle-Pock algorithm for MRI (Chambolle and Pock 2011; Sidky et al. 2012). These two reconstructions are referred to as maximum-likelihood with TV regularization (ML-TV). Automatic choice of the regularization parameter is a challenging task, in particular for low-count Poisson distributed data. We performed a broad sweep of the regularization parameter γ values to minimize the mean squared error with the GT while avoiding overfitting.

In addition, we trained a collection of U-Nets for image post-processing (Ronneberger et al. 2015; Pain et al. 2022). For a given PET/MR image pair $(x^{\star}_{\text{pet}}, x^{\star}_{\text{mr}})$ from the training dataset, let us denote by $x^{\text{mlem}}_{\text{pet}}$ (i.e., with $\gamma = 0$ in (44)) the MLEM-reconstructed image from data simulated by (42) (with $x^{\star}_{\text{pet}}$ as GT and with dose parameter α) and by $x^{\text{ifft}}_{\text{mr}}$ the IFFT-reconstructed image from data simulated by (43) (with $x^{\star}_{\text{mr}}$ as GT and under-sampling parameter R). We trained a MR-guided PET denoiser U-Net $F_{\varphi}\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ with parameter φ as

$$\min_{\varphi} \mathbb{E}\left[\left\| F_{\varphi}\left(x^{\text{mlem}}_{\text{pet}}, x^{\star}_{\text{mr}}\right) - x^{\star}_{\text{pet}} \right\|_2^2 \right], \tag{45}$$

and an MR post-processing U-Net $H_{\psi}\colon \mathbb{R}^m \to \mathbb{R}^m$ with parameter ψ as

$$\min_{\psi} \mathbb{E}\left[\left\| H_{\psi}\left(x^{\text{ifft}}_{\text{mr}}\right) - x^{\star}_{\text{mr}} \right\|_2^2 \right], \tag{46}$$

where both expectations are taken over the training dataset. The trainings (45) and (46) are performed for each considered values of α and R. We will show the following U-Net outputs: (i) the post-processed MR $H_{\psi}\left(x^{\text{ifft}}_{\text{mr}}\right)$, (ii) the MR-guided denoised PET $F_{\varphi}\left(x^{\text{mlem}}_{\text{pet}}, x^{\text{ifft}}_{\text{mr}}\right)$ and (iii) the MR-guided denoised PET with post-processed MR $F_{\varphi}\left(x^{\text{mlem}}_{\text{pet}}, H_{\psi}\left(x^{\text{ifft}}_{\text{mr}}\right)\right)$ (denoted U-net+PP).

The results will be shown for different values of α and or R.

4.4. Evaluation metrics

The comparison between the methods was performed on standardized images (see Equation 34), using peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). The PSNR between two images x and x_{ref} is given by:

$$\text{PSNR}(x, x_{\text{ref}}) = 20 \log \left(\frac{\text{range}(x_{\text{ref}})}{\sqrt{\text{MSE}(x, x_{\text{ref}})}} \right), \tag{47}$$

where $\text{range}(x_{\text{ref}}) = \max(x_{\text{ref}}) - \min(x_{\text{ref}})$ is the range of intensities of the image x_{ref}. The SSIM between the two images is given by

$$\text{SSIM}(x, x_{\text{ref}}) = \frac{(2\mu_x\mu_{\text{ref}} + c_1)(2\sigma_x\sigma_{\text{ref}} + c_2)(\text{cov}(x, x_{\text{ref}}) + c_3)}{(\mu_x^2 + \mu_{\text{ref}}^2 + c_1)(\sigma_x^2 + \sigma_{\text{ref}}^2 + c_2)(\sigma_x\sigma_{\text{ref}} + c_3)}, \tag{48}$$

where

- μ_x and μ_{ref} and are the mean of x and x_{ref} respectively,
- σ_x and σ_{ref} are the STD of x and x_{ref} respectively,
- $\text{cov}(x, x_{\text{ref}})$ is the covariance of x and x_{ref},
- $c_1 = (k_1\text{range}(x_{\text{ref}}))^2$, $c_2 = (k_2\text{range}(x_{\text{ref}}))^2$ and $c_3 = \frac{c_2}{2}$,
- $k_1 = 0.01$ and $k_2 = 0.03$

5. Results

5.1. Experiment 1: Image Generation

The objective of this experiment is to demonstrate that the bimodal generative models have been adequately trained to generate images pair-wise. Assuming that the latent variable z represents "the patient", the generated images $G_\theta^{\text{pet}}(\hat{z})$ and $G_\theta^{\text{mr}}(\hat{z})$ must be consistent with each other. We consider two target images $x_{\text{pet}}^\star$ and $x_{\text{mr}}^\star$ and we fit both models by solving (39) and (40).

Figure 6 shows the PET-fitted images $G_\theta^{\text{pet}}(\hat{z}_{\text{pet}})$ (first row), i.e., the PET images generated from a latent variable $\hat{z}_{\text{pet}}$ of dimension $d = 64$ that was estimated by fitting the model to the target PET image $x_{\text{pet}}^\star$ only, as well as the predicted MR images $G_\theta^{\text{mr}}(\hat{z}_{\text{pet}})$, for the three considered generative models JVAE, PoE-VAE and MMJSD; the SSIM and PSNR values were computed with respect to the target PET $x_{\text{pet}}^\star$ and MR $x_{\text{mr}}^\star$ which are shown on the left column. First we observe that the PET-fitted images are similar to the target PET, which shows that $x_{\text{pet}}^\star$ is close to the range of the generator. We also observe that the predicted MR images (second row) are also similar—to a lesser extend—to the target MR images. This illustrates the "transfer of information" from the PET image to the MR image of our bimodal generative models.

Figure 7 show the results of the reverse experiment, that is to say the MR-fitted images $G_\theta^{\text{mr}}(\hat{z}_{\text{mr}})$ (second row) and the predicted PET images $G_\theta^{\text{pet}}(\hat{z}_{\text{mr}})$ (first row). The MR-fitted images are also similar to the target $x_{\text{mr}}^\star$ and the predicted PET images $G_\theta^{\text{pet}}(\hat{z}_{\text{mr}})$ are somehow similar to the target PET image $x_{\text{pet}}^\star$.

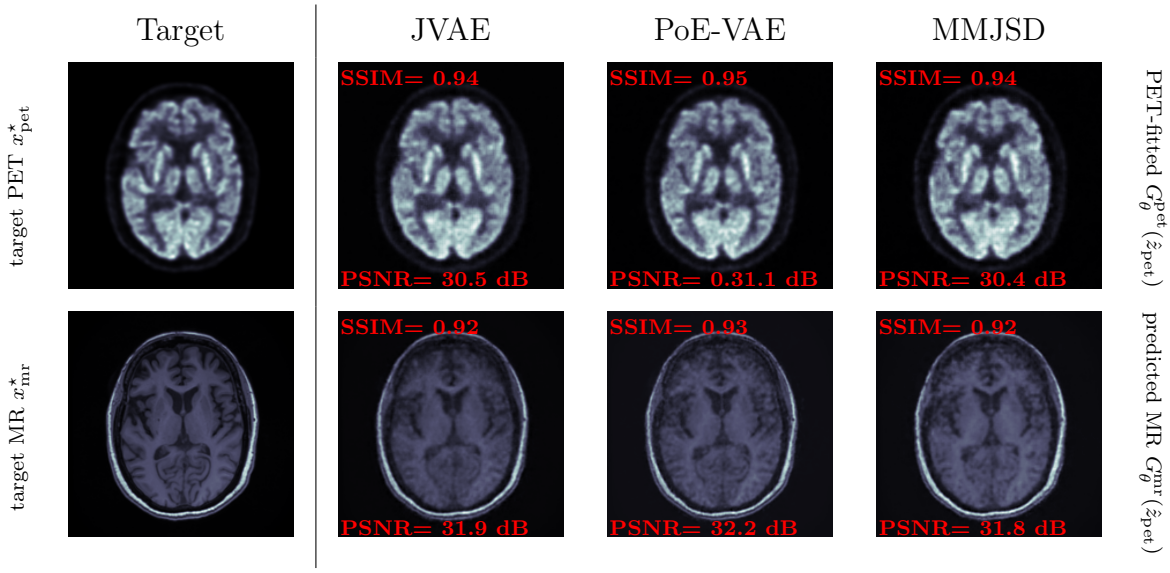


Figure 6: Experiment 1 ($d = 64$)—PET-fitted images $G^{\text{pet}}_\theta(\hat{z}_{\text{pet}})$ and predicted MR images $G^{\text{mr}}_\theta(\hat{z}_{\text{pet}})$ where $\hat{z}_{\text{pet}}$ is given by (39) using the top-left image as the target $x^\star_{\text{pet}}$.

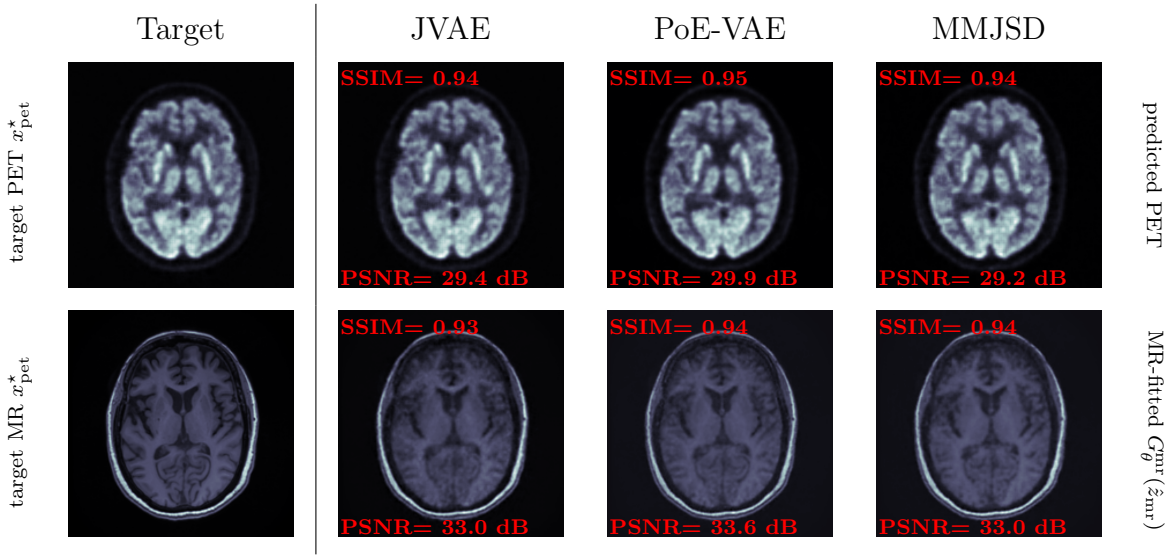


Figure 7: Experiment 1 ($d = 64$)—Predicted PET images $G^{\text{pet}}_\theta(\hat{z}_{\text{mr}})$ and MR-fitted images $G^{\text{mr}}_\theta(\hat{z}_{\text{mr}})$ where $\hat{z}_{\text{mr}}$ is given by (40) using the bottom-left image as the target $x^\star_{\text{mr}}$.

This experiment shows that our bimodal models are capable of conveying information between modalities. More precisely, extracting information from one modality (by model fitting) can provide information on the other modality in both ways.

Figure 8 and Figure 9 show the results of the same experiments, but this time with $d = 32$. We observe in both experiments that the results are slightly behind the ones with $d = 64$ for both modalities. This highlights the importance of the latent space dimension to represent the images accurately.

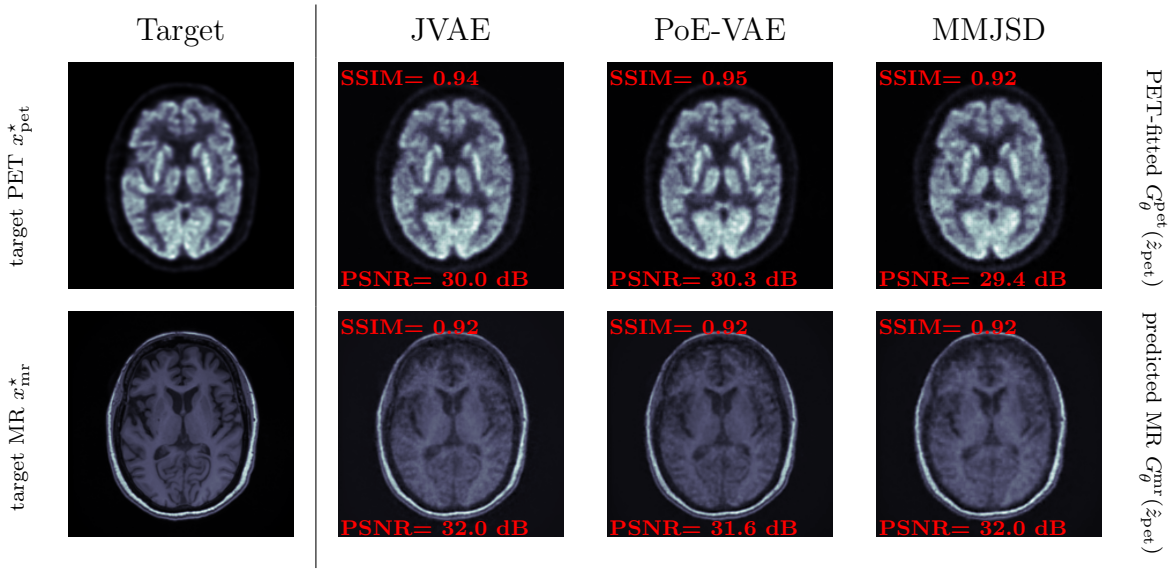


Figure 8: Experiment 1 ($d = 32$)—PET-fitted images $G^{\text{pet}}_\theta(\hat{z}_{\text{pet}})$ and predicted MR images $G^{\text{mr}}_\theta(\hat{z}_{\text{pet}})$ where $\hat{z}_{\text{pet}}$ is given by (37) using the top-left image as the target $x^\star_{\text{pet}}$.

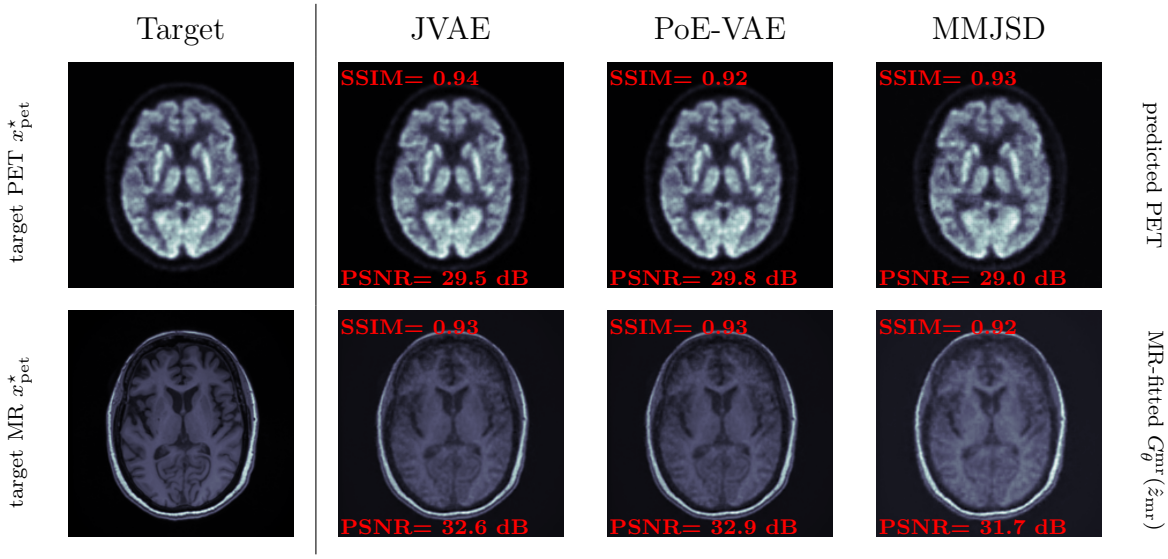
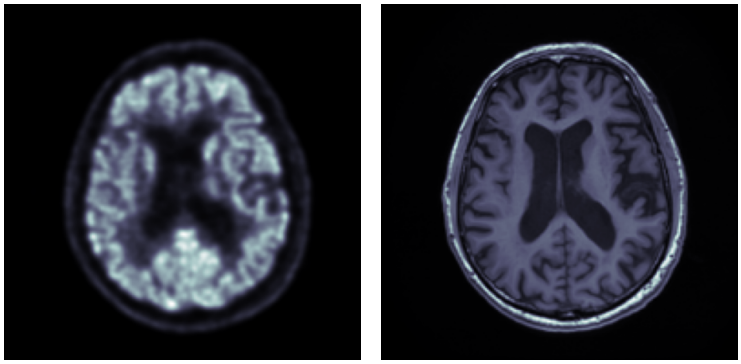


Figure 9: Experiment 1 ($d = 32$)—Predicted PET images $G^{\text{pet}}_\theta(\hat{z}_{\text{mr}})$ and MR-fitted images $G^{\text{mr}}_\theta(\hat{z}_{\text{mr}})$ where $\hat{z}_{\text{mr}}$ is given by (38) using the bottom-left image as the target $x^\star_{\text{mr}}$.

5.2. Experiment 2: Image Reconstruction

In this section we show the reconstruction results from data simulated with different settings. We proceeded with showing reconstructing images from data acquired first with $\alpha = 10^5$, $R = 20$ then with $\alpha = 10^6$, $R = 40$, using the methods described in Section 4.3, with models trained with a latent space dimension $d = 64$. These values correspond to highly-altered projections, as can be seen from the non-regularized reconstructions.



(a) GT PET (b) GT MR

Figure 10: Experiment 2—GT images (from testing dataset) used to generate the raw data used for reconstruction (Figure 11 and Figure 12).

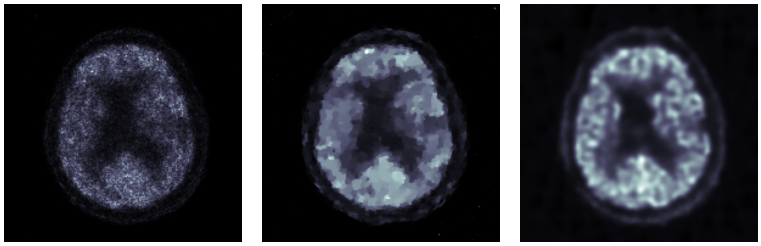
First, in terms of computational cost of DLR, most of it comes from the gradient descent for the z optimization step. For the initialization of the PET image we only perform a few iterations of MLEM (15 in our case) which is quite fast compared to the main body of the algorithm. The same goes for the MR image with IFFT. Finally, z is initialized from these two first approximations but using the encoder after learning is very fast and is also negligible in time compared to the main algorithm. As a result, reconstructing one slice takes between 20 and 25 seconds, 3 of which are for the initialization.

Figure 10 shows one of the GT PET/MR image pairs from the testing dataset that was used to generate raw data for reconstruction following (42) and (43).

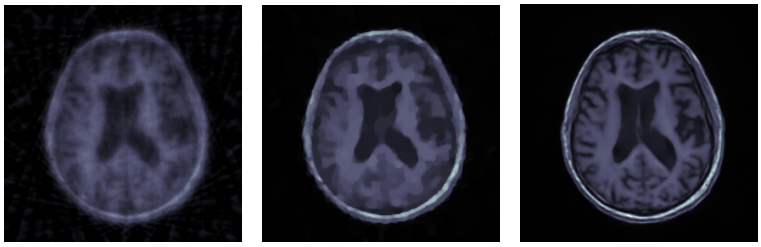
Figure 11 shows the results of the reconstructions from data acquired with $\alpha = 10^5$, $R = 20$ and $d = 64$. The U-Net PET output in Figure 11c was obtained from the MLEM PET (Figure 11a) and IFFT MR (Figure 11d). The MLEM-reconstructed PET image suffer from noise while the inverse fast Fourier Transform-reconstructed image suffer from heavy streak artifacts due to undersampling. TV regularization reduces noise in PET images and streak artifacts in MR images, however this results in typical patch artifacts. The DLR method strikes a balance between suppressing noise and artifacts while still reproducing the structures from the images. The U-Net output MR image appears similar to the GT, while the U-Net output PET image is slightly blurry.

Figure 12 shows the same reconstructed images for $\alpha = 10^6$ and $R = 40$. Compared to the previous experiment, PET images are expected to improve, while anatomical structures in the MR images are expected to be more affected by artifacts. This is indeed visible in the non-regularized and in the TV-regularized images. The DLR images are very similar to the previous case although it can be noticed that the MR images are slightly blurrier while the PET images are slightly sharper. Once again, the U-Net output MR image is similar to the GT while U-Net output PET image is slightly blurry and noisy.

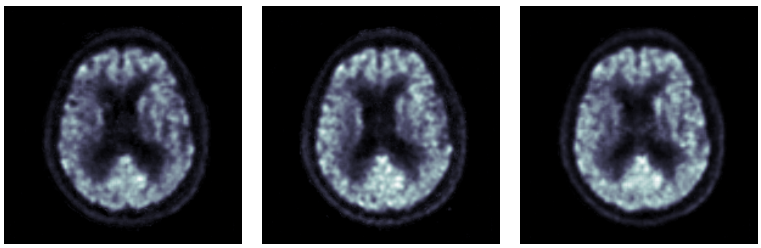
Figures 13 and 14 display the PSNR and SSIM metrics for the PET and MR



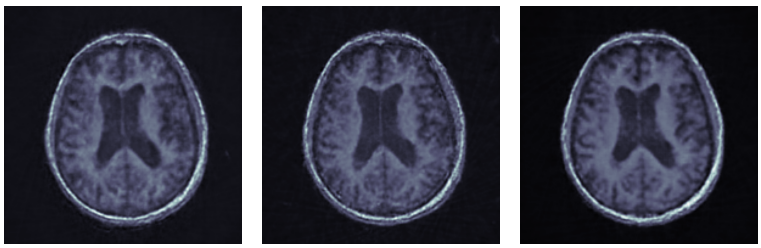
(a) MLEM PET (b) ML-TV PET (c) U-Net PET



(d) MR IFFT (e) MR ML-TV (f) U-Net MR

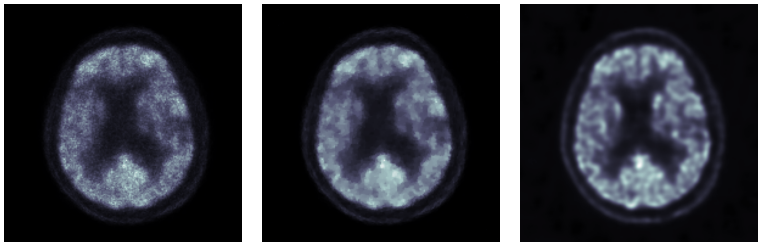


(g) JVAE PET (h) PoE-VAE PET (i) MMJSD PET

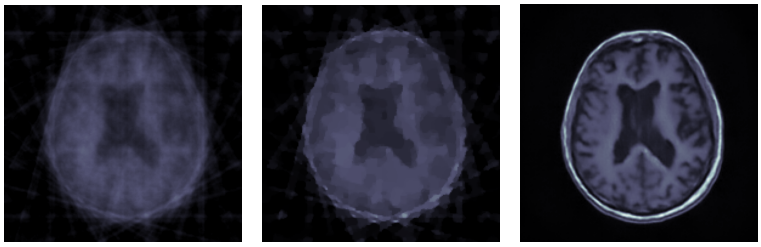


(j) JVAE MR (k) PoE-VAE MR (l) MR MMJSD

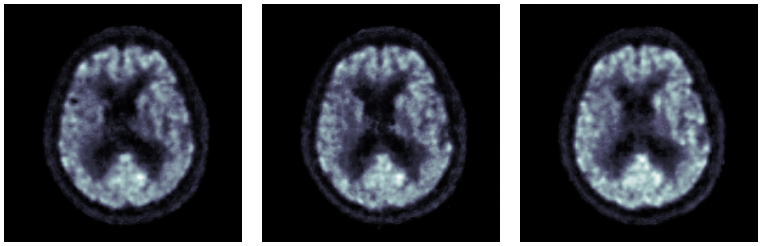
Figure 11: Experiment 2—Reconstructed images from data simulated with $\alpha = 10^5$ and $R = 20$ using the GT images in Figure 10. The VAEs used by the DLR methods use a latent space of dimension $d = 64$, while the U-Net PET output (c) was obtained from the MLEM PET (a) and IFFT MR (d).



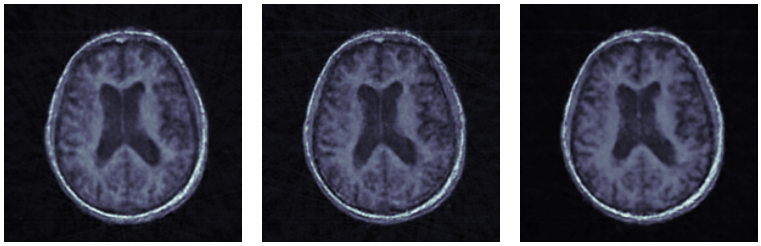
(a) MLEM PET (b) ML-TV PET (c) PET U-Net



(d) IFFT MR (e) ML-TV MR (f) U-Net MR



(g) JVAE PET (h) PoE-VAE PET (i) MMJSD PET



(j) JVAE MR (k) PoE-VAE MR (l) MMJSD MR

Figure 12: Experiment 2—Reconstructed images from data simulated with $\alpha = 10^6$ and $R = 40$ using the GT images in Figure 10. The VAEs used by the DLR methods use a latent space of dimension $d = 64$, while the U-Net PET output (c) was obtained from the MLEM PET (a) and IFFT MR (d)

reconstructions averaged across the testing dataset. In Figure 13, four values of α varying from $0.1 \cdot 10^6$ to 10^6 are considered while R is maintained constant at $R = 40$. This corresponds to a high subsampling rate for the MR acquisition coupled with high to very high levels of noise in the PET acquisition. In Figure 14, four values of the sub-sampling factor R between 10 and 40 are considered with $\alpha = 10^5$. This corresponds to high Poisson noise level in PET data, coupled to high to very high sub-sampling rates for the MR acquisition.

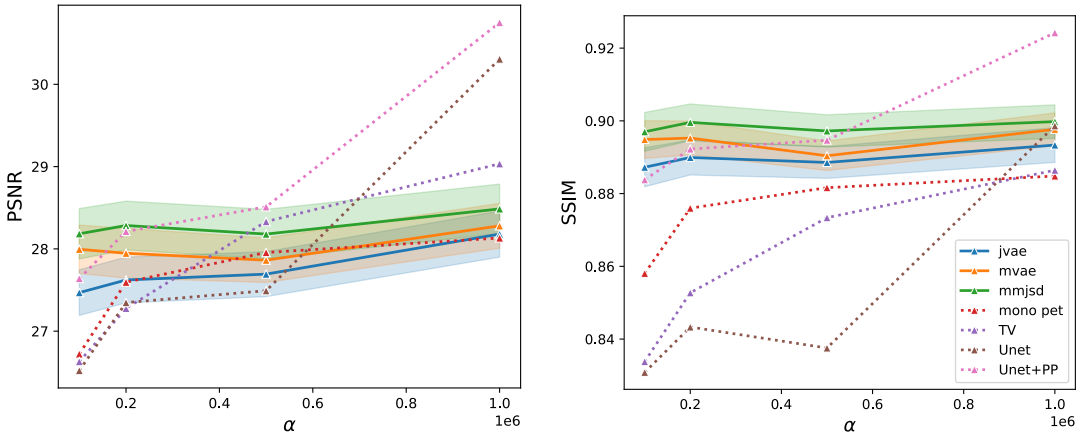
Seven reconstruction methods for PET and six for MRI are compared, the monomodality VAE reconstructions being added as well as the PET U-Net output with post-processed MR guidance, while the non-regularized MLEM and IFFT reconstructions are not shown.

For both scenarios the TV reconstructions outperform DLRs in high-sampling and high-count settings; an exception is observed for the SSIM metric in the case $R = 40$ and $\alpha = 10^6$ and for the PET image. However, DLR achieves better results in both SSIM and PSNR in low-count settings.

For PET reconstruction, monomodal DLR and ML-TV experience significant drops in PSNR and SSIM when α decreases while these metrics remain fairly similar for the multimodal DLRs. Unimodal DLR is on par with bimodal DLR with high counts, whereas bimodal DLRs outperforms it with lower counts. In terms of PSNR, the unimodal DLR is outperformed by at least 5% by bimodal DLRs with low counts and is on on par for highest counts. In terms of SSIM, the bimodal DLR is consistently above the unimodal one with a large gap (up to 3%). Both U-Nets outperform DLRs in terms of PSNR in the high-count setting, but their performances drop by 7% when α decreases for the one without processed MR image and by 5% for the one with the denoised MR image. In particular, the U-Net approach with non-processed MR is significantly outperformed by the DLRs in terms of PSNR.

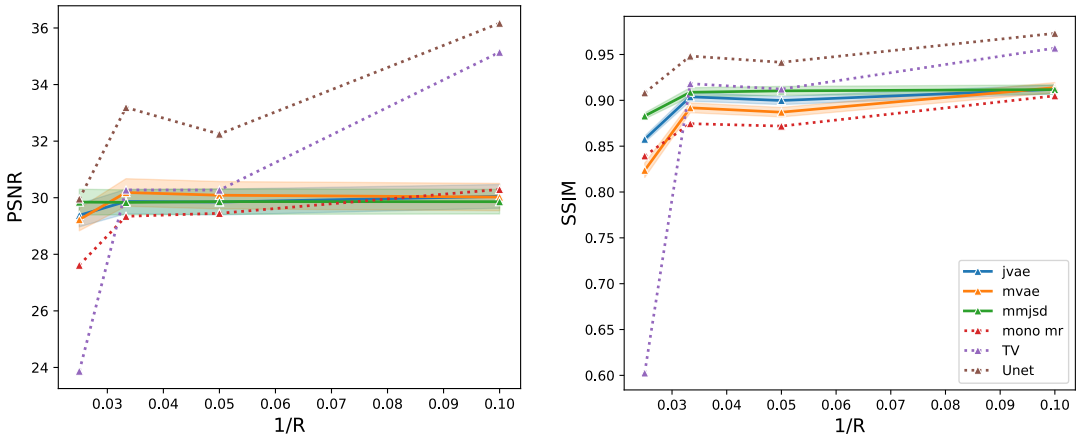
For the MR reconstruction, we observe that the results for the unimodal DLR degrade compared to multimodal DLRs as the subsampling factor R increases . As opposed to the PET reconstruction, we do not observe a significant increase of the performance gap between unimodal and bimodal DLRs. The U-Net approach outperforms all methods for all values of R for both metrics.

The different DLR approaches can be ranked based on their performance for PET reconstruction. For both PSNR and SSIM, MMJSD performs the best, followed by multimodal VAE and then JVAE. The difference in SSIM results between the different models is not significant however, with SSIM varying 1% at the most across all values of α. For MR, the PSNR results are similar . However, they perform differently in terms of PSNR with the same ranking as with PET. The performance gap increases significantly with the subsampling factor, with a difference of approximately 6% in SSIM between the best and the worst performances.



(a) PSNR vs α (PET) (b) SSIM vs α (PET)

Figure 13: Experiment 2—PSNR and SSIM values of the reconstructed PET images with α varying from 10^5 to 10^6, $R = 40$, $d = 64$.



(a) PSNR vs R (MR) (b) SSIM vs R (MR)

Figure 14: Experiment 2—PSNR and SSIM values of the reconstructed MR images with R varying from 10 to 40, $\alpha = 10^5$, $d = 64$.

6. Discussion

The use of a generative model trained on reference images within an iterative reconstruction framework compensates for high noise in PET or undersampling in MRI. The reconstructed images are free of noise and artifacts, exhibiting enhanced detail and clarity.

We observe that PET reconstruction using bimodal DLR is less affected by data deterioration compared to the monomodal version. This indicates that the method's performance is not solely due to the use of a VAE for image generation. The PET reconstruction benefits from the inclusion of MRI data, which helps to control the noise.

However, the results for MRI are less conclusive. Both monomodal and bimodal versions of DLR show similar performance trends with varying subsampling factors. Despite this, the bimodal performance appears to be slightly superior to the monomodal one, suggesting that the presence of the PET image helps in finding a latent variable that better represents the target image.

U-Net post-processing approaches deliver good results and outperforms DLRs at several instances, especially for MRI where it outperforms DLRs for low noise levels and undersampling rates. This suggests that using VAEs for MR image representation may be suboptimal. This outcome was somehow expected as VAE are known to produce blurry images that may not be well-suited for MRI. However, the training of the VAEs used in DLR are unsupervised and can be performed on any PET/MRI dataset, while the training of the U-Nets are supervised and depend on α and R, which is not practical as it requires to train multiple NNs.

Comparing the different VAE architectures, we find that the results of each model are similar for $d = 64$. However, MMJSD consistently performs slightly better on both PET and MRI, especially for the shortest acquisition times. The dimension of the latent space is a crucial hyperparameter. Experiment 1 demonstrates that reducing the latent space dimension coincides with a loss in terms of generated image quality. There is, however, a tradeoff to be reached here. Increasing the latent space dimension improves details retention but also escalates memory requirements and computing time. Moreover, a larger latent space increases the number of network parameters, necessitating more training data. However, the availability of medical data is limited, posing a significant challenge.

Other hyperparameters can be fine-tuned to improve the results. In particular, the influence of the number of MLEM iterations for the initial image has been tested and the best results were obtained with 10 iterations.

The dataset we used for training was relatively small, leading to overfitting of the model. It is important to note that while the test and training sets of slices are disjoint, the test and training volumes are not separated. When the model is tested on a separate volume, the results are worse than those presented in the paper, showcasing the difficulty to generalize to new patients (see Appendix B). To investigate further, we tested our method on the simpler problem of Gaussian denoising using bimodal MNIST data (Deng 2012). The MNIST dataset is larger than our medical image dataset and consists of smaller images. The results we obtained (not shown in the paper) were significantly better. We also experimented with various data augmentation techniques, which improved the results somewhat, but not to a satisfactory extent.

Upon close inspection, it is evident that some details in the images are removed, and features not present in the ground truth image may appear. These 'hallucinations' are a result of the constraints. The algorithm's output closely resembles the VAE's output and therefore is impacted by VAE's performance. This also affects the image resolution, as VAEs are known to produce blurry images (Dumoulin et al. 2017). To improve the algorithm, the utilization of more advanced architectures like DiffuseVAE (Pandey et al.

2022) could be beneficial. However, these architectures typically require even more data. An alternative approach could be to reformulate the problem with a 'softer' constraint, where the output is a combination of the VAE's output and classical reconstruction

Another limitation of the proposed scheme is its reliance on the selection of several hyperparameters, particularly the parameter which balances the fidelity terms between PET and MR data during VAE training.

This work opens interesting directions for future research. Recently, several new VAE models with more complex generative processes have been proposed. For instance, segregating the latent representation space into a combination of private and shared latent spaces, as proposed in Sutter et al. (2021), could enhance both the generative power of the VAE and the latent space quality.

7. Conclusion

In this work, we introduce DLR, a novel method for synergistic VAE/MRI reconstruction utilizing VAE constraints. Our approach leverages both data and physical models to identify optimal latent variables within the VAE's latent space. These variables, when decoded, generate results comparable to traditional acquisitions with standard acquisition times. Although lacking formal theoretical guarantees, empirical experiments demonstrate that DLR outperforms conventional methods in retrieving missing data, especially under time-constrained conditions. By integrating information from complementary modalities, DLR effectively compensates for missing data, thereby enhancing reconstruction accuracy.

Additionally, we evaluate various VAE architectures, each trained with distinct loss functions. While achieving comparable results overall, the MMJSD architecture shows superior performance. However, potential overfitting suggests that performance rankings may vary with new datasets. Future research should focus on validating these findings with larger, more diverse datasets to elucidate differences among these models.

Furthermore, improving VAE performance remains crucial. Issues such as "hallucinations," where VAEs generate unrealistic outputs unrelated to the physical model, persist. Moreover, our current VAE models, trained without anomalies like lesions or tumors, cannot effectively produce such features. Addressing these challenges requires exploring methods to enhance the generative process of VAEs by incorporating domain-specific knowledge into model training.

Appendix A. Resolution of the DLR update

We solve the minimization problem (30) by using an optimization transfer approach with a convex surrogate (Guobao Wang and Jinyi Qi 2012). We define a surrogate function Q for $\mathscr{D}_{\text{pet}}$ as

$$Q\left(x \mid x^{(k)}\right) = \sum_j Q_j\left(x_j \mid x^{(k)}\right) \tag{A.1}$$

with

$$Q_j\left(x_j \mid x^{(k)}\right) = p_j\left(x_j - [x^{(k)}_{\text{pet,em}}]_j \log x_j\right) \tag{A.2}$$

$$+ \frac{\rho_{\text{pet}}}{2}\left(x_j - \left[G^{\text{pet}}_\theta(z^{(k)})\right]_j + \mu_j\right)^2 \tag{A.3}$$

where $p_j = \alpha \sum_i P_{i,j}$ and $x^{n+1}_{\text{pet,em}}$ is obtained by doing one MLEM step:

$$\left[x^{(k)}_{\text{pet,em}}\right]_j = \frac{\left[x^{(k)}_{\text{pet}}\right]_j}{p_j} \sum_i P_{i,j} \frac{[y_{\text{pet}}]_i}{\left[P x^{(k)}_{\text{pet}}\right]_i + r_i + s_i} \tag{A.4}$$

The minimization problem in Equation (30) is then transferred to the surrogate function for every pixel j:

$$\left[x^{(k+1)}_{\text{pet}}\right]_j = \underset{[x_{\text{pet}}]_j}{\arg\min}\, Q_j\ \left([x_{\text{pet}}]_j \ \mid\ x^{(k)}_{\text{pet}}\right) \tag{A.5}$$

By using first-order optimality condition, we end up solving a quadratic equation with positivity constraint which gives the following update $x^{(k+1)}_{\text{pet}}$ at each pixel j:

$$\left[x^{(k+1)}_{\text{pet}}\right]_j = \frac{1}{2}\Bigg(\left[G^{\text{pet}}_\theta(z^{(k)})\right]_j - \left[\mu^{(k)}_{\text{pet}}\right]_j - \frac{p_j}{\rho_{\text{pet}}} + \tag{A.6}$$

$$\sqrt{\left(\left[G^{\text{pet}}_\theta(z^{(k)})\right]_j - \left[\mu^{(k)}_{\text{pet}}\right]_j - \frac{p_j}{\rho_{\text{pet}}}\right)^2 + \frac{4p_j\left[x^{(k+1)}_{\text{pet,em}}\right]_j}{\rho_{\text{pet}}}}\ \Bigg)$$

Equation (31) is a penalized least square problem. This problem can be solved using first order optimality condition, i.e., by solving:

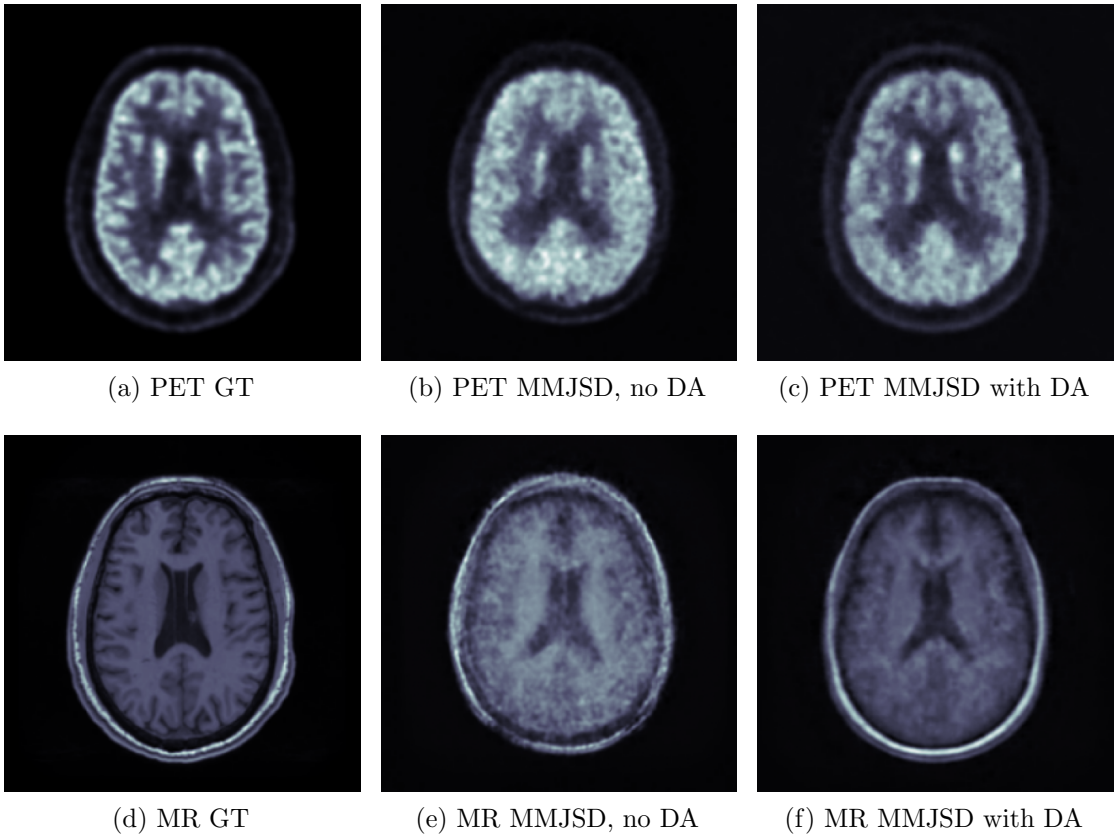
$$(\rho_{\text{mr}} I + E^H E) x^{(k+1)}_{\text{mr}} = E^H y_{\text{mr}} + \rho_{\text{mr}}(\text{Decoder}(z^{(k)})_{\text{mr}} - \mu^{(k)}_{\text{mr}})\,, \tag{A.7}$$

where E^H is the Hermitian adjoint of E. The solution of this equation can be obtained with a few iterations of the conjugate gradient algorithm (Pruessmann et al. 2001).

Appendix B. Test on a separate patient dataset

We qualitatively assessed the generalization power of MMJSD with an evaluation based on data simulated from a patient that does not belong to the training dataset. We tested two models: (i) one trained on the standard dataset and (ii) one trained on the same dataset completed with data augmentation (DA), including random rotations and dilations.

Figure B1 shows the results, alongside the GT images used to generate the raw data following (42) and (43). We observe that the MMJSD-reconstructed images without DA are somehow noisy and suffer from several artifacts, while the MMJSD-reconstructed images with DA are significantly closer to the GT images. This experiments shows that it is possible to improve the performances of DLR by increasing the size of the training dataset.



(a) PET GT (b) PET MMJSD, no DA (c) PET MMJSD with DA

(d) MR GT (e) MR MMJSD, no DA (f) MR MMJSD with DA

Figure B1: GT images and MMJSD-reconstructed images using models trained with and without DA.

Acknowledgment

We would like to thank the Service Hospitalier Frédéric Joliot for providing us with the data used in this study as well as Thomas M. Sutter for helping us with the implementation of the mmJSD model.

All authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this paper.

This work was supported by the French National Research Agency (ANR) under grant No `ANR-20-CE45-0020` and by France Life Imaging under grant No ANR-11-INBS-0006.

Ethical statement

This study was performed in line with the principles of Helsinki. Approval was granted by the executive board of the CEA/SHFJ department (2022/11/10).

References

Aarle, Wim van, Willem Jan Palenstijn, Jeroen Cant, Eline Janssens, Folkert Bleichrodt, Andrei Dabravolski, Jan De Beenhouwer, K. Joost Batenburg, and Jan Sijbers (Oct. 2016). "Fast and flexible X-ray tomography using the ASTRA toolbox". In: *Opt. Express* 24.22, pp. 25129–25147. DOI: 10.1364/OE.24.025129. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-24-22-25129.

Ahn, Sangtae and Jeffrey A Fessler (2003). "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms". In: *IEEE transactions on medical imaging* 22.5, pp. 613–626.

Arridge, Simon R., Matthias J. Ehrhardt, and Kris Thielemans (June 2021). "(An overview of) Synergistic reconstruction for multimodality/multichannel imaging methods". en. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2200, p. 20200205. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2020.0205.

Bora, Ashish, Ajil Jalal, Eric Price, and Alexandros G Dimakis (2017). "Compressed sensing using generative models". In: *International conference on machine learning*. PMLR, pp. 537–546.

Bousse, Alexandre, Venkata Sai Sundar Kandarpa, Kuangyu Shi, Kuang Gong, Jae Sung Lee, Chi Liu, and Dimitris Visvikis (2024). "A Review on Low-Dose Emission Tomography Post-Reconstruction Denoising With Neural Network Approaches". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 8.4, pp. 333–347. DOI: 10.1109/TRPMS.2023.3349194. URL: https://arxiv.org/abs/2401.00232.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2010). "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundations & Trends in Machine Learning* 3.1, pp. 1–122.

Catana, Ciprian (2020). "Attenuation correction for human PET/MRI studies". In: *Physics in Medicine & Biology* 65.23, 23TR02.

Chambolle, Antonin and Thomas Pock (2011). "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of mathematical imaging and vision* 40, pp. 120–145.

Chen, Kevin T, Enhao Gong, Fabiola Bezerra de Carvalho Macruz, Junshen Xu, Athanasia Boumis, Mehdi Khalighi, Kathleen L Poston, Sharon J Sha, Michael D Greicius, Elizabeth Mormino, et al. (2019). "Ultra–low-dose 18F-florbetaben amyloid PET imaging using deep learning with multi-contrast MRI inputs". In: *Radiology* 290.3, pp. 649–656.

Chen, Kevin T, Tyler N Toueg, Mary Ellen Irene Koran, Guido Davidzon, Michael Zeineh, Dawn Holley, Harsh Gandhi, Kim Halbert, Athanasia Boumis, Gabriel Kennedy, et al. (2021). "True ultra-low-dose amyloid PET/MRI enhanced with deep learning for clinical interpretation". In: *European journal of nuclear medicine and molecular imaging* 48, pp. 2416–2425.

Cheng, Jianhong, Min Gao, Jin Liu, Hailin Yue, Hulin Kuang, Jun Liu, and Jianxin Wang (2021). "Multimodal disentangled variational autoencoder with game theoretic interpretability for glioma grading". In: *IEEE Journal of Biomedical and Health Informatics* 26.2, pp. 673–684.

Da Costa-Luis, Casper O and Andrew J Reader (2020). "Micro-networks for robust MR-guided low count PET imaging". In: *IEEE transactions on radiation and plasma medical sciences* 5.2, pp. 202–212.

De Pierro, Alvaro R (1995). "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography". In: *IEEE transactions on medical imaging* 14.1, pp. 132–137.

Deng, Li (2012). "The MNIST database of handwritten digit images for machine learning research [best of the web]". In: *IEEE signal processing magazine* 29.6, pp. 141–142.

Dumoulin, Vincent, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville (Feb. 2017). *Adversarially Learned Inference*. en. arXiv:1606.00704 [cs, stat]. URL: http://arxiv.org/abs/1606.00704 (visited on 01/30/2024).

Ehrhardt, Matthias J, Kris Thielemans, Luis Pizarro, David Atkinson, Sébastien Ourselin, Brian F Hutton, and Simon R Arridge (Jan. 2015). "Joint reconstruction of PET-MRI by exploiting

structural similarity". en. In: *Inverse Problems* 31.1, p. 015001. ISSN: 0266-5611, 1361-6420. DOI: 10.1088/0266-5611/31/1/015001.

Fessler, Jeffrey A (2020). "Optimization methods for magnetic resonance image reconstruction: Key models and optimization algorithms". In: *IEEE signal processing magazine* 37.1, pp. 33–40.

Gonzalez, Rafael C., Richard E. Woods, and Barry R. Masters (2009). "Digital Image Processing, Third Edition". en. In: *Journal of Biomedical Optics* 14.2, p. 029901. ISSN: 10833668. DOI: 10.1117/1.3115362. URL: http://biomedicaloptics.spiedigitallibrary.org/article.aspx?doi=10.1117/1.3115362 (visited on 02/23/2024).

Guobao Wang and Jinyi Qi (Dec. 2012). "Penalized Likelihood PET Image Reconstruction Using Patch-Based Edge-Preserving Regularization". en. In: *IEEE Transactions on Medical Imaging* 31.12, pp. 2194–2204. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2012.2211378. URL: http://ieeexplore.ieee.org/document/6257498/ (visited on 12/18/2023).

Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner (2017). "β-VAE: learning basic visual concepts with a constrained variational framework". en. In.

Kingma, Diederik P and Max Welling (2013). "Auto-Encoding Variational Bayes". In: DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Kingma, Diederik P. and Max Welling (2019). "An Introduction to Variational Autoencoders". en. In: *Foundations and Trends® in Machine Learning* 12.4. arXiv:1906.02691 [cs, stat], pp. 307–392. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000056.

Knoll, Glenn F (2010). *Radiation detection and measurement*. John Wiley & Sons.

Laurent, Baptiste, Alexandre Bousse, Thibaut Merlin, Stephan Nekolla, and Dimitris Visvikis (2023). "PET scatter estimation using deep learning U-Net architecture". In: *Physics in Medicine & Biology* 68.6, p. 065004.

Mehranian, Abolfazl, Martin A. Belzunce, Claudia Prieto, Alexander Hammers, and Andrew J. Reader (Jan. 2018). "Synergistic PET and SENSE MR Image Reconstruction Using Joint Sparsity Regularization". en. In: *IEEE Transactions on Medical Imaging* 37.1, pp. 20–34. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2017.2691044. URL: http://ieeexplore.ieee.org/document/7903631/ (visited on 11/15/2021).

Natterer, Frank (2001). *The mathematics of computerized tomography*. SIAM.

Pain, Cameron Dennis, Gary F. Egan, and Zhaolin Chen (July 2022). "Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement". en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 49.9, pp. 3098–3118. ISSN: 1619-7089. DOI: 10.1007/s00259-022-05746-4. URL: https://doi.org/10.1007/s00259-022-05746-4 (visited on 08/26/2024).

Pandey, Kushagra, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar (2022). *DiffuseVAE: Efficient, Controllable and High-Fidelity Generation from Low-Dimensional Latents*. DOI: 10.48550/ARXIV.2201.00308. URL: https://arxiv.org/abs/2201.00308.

Perelli, Alessandro, Suxer Alfonso Garcia, Alexandre Bousse, Jean-Pierre Tasu, Nikolaos Efthimiadis, and Dimitris Visvikis (2022). "Multi-channel convolutional analysis operator learning for dual-energy CT reconstruction". In: *Physics in Medicine & Biology* 67.6, p. 065001.

Pinton, Noel Jeffrey, Alexandre Bousse, Catherine Cheze-Le-Rest, and Dimitris Visvikis (2024). "Multi-Branch Generative Models for Multichannel Imaging with an Application to PET/CT Joint Reconstruction". In: *arXiv preprint arXiv:2404.08748*. URL: https://arxiv.org/abs/2404.08748.

Pruessmann, Klaas P., Markus Weiger, Peter Börnert, and Peter Boesiger (Oct. 2001). "Advances in sensitivity encoding with arbitrary k -space trajectories: SENSE With Arbitrary k -Space Trajectories". en. In: *Magnetic Resonance in Medicine* 46.4, pp. 638–651. ISSN: 07403194. DOI:

10.1002/mrm.1241. URL: https://onlinelibrary.wiley.com/doi/10.1002/mrm.1241 (visited on 11/24/2021).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". en. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

Sawatzky, Alex, Christoph Brune, Frank Wubbeling, Thomas Kosters, Klaus Schafers, and Martin Burger (Oct. 2008). "Accurate EM-TV algorithm in PET with low SNR". en. In: *2008 IEEE Nuclear Science Symposium Conference Record*. Dresden, Germany: IEEE, pp. 5133–5137. ISBN: 978-1-4244-2714-7. DOI: 10.1109/NSSMIC.2008.4774392. URL: http://ieeexplore.ieee.org/document/4774392/ (visited on 02/09/2024).

Schramm, Georg, David Rigie, Thomas Vahle, Ahmadreza Rezaei, Koen Van Laere, Timothy Shepherd, Johan Nuyts, and Fernando Boada (2021). "Approximating anatomically-guided PET reconstruction in image space using a convolutional neural network". In: *Neuroimage* 224, p. 117399.

Shepp, L. A. and Y. Vardi (1982). "Maximum Likelihood Reconstruction for Emission Tomography". In: *IEEE Transactions on Medical Imaging* 1.2, pp. 113–122. DOI: 10.1109/TMI.1982.4307558.

Shi, Yuge, N. Siddharth, Brooks Paige, and Philip H. S. Torr (Nov. 2019). *Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models*. en. arXiv:1911.03393 [cs, stat]. URL: http://arxiv.org/abs/1911.03393 (visited on 11/02/2023).

Sidky, Emil Y., Jakob H. Jørgensen, and Xiaochuan Pan (May 2012). "Convex optimization problem prototyping for image reconstruction in computed tomography with the Chambolle-Pock algorithm". en. In: *Physics in Medicine and Biology* 57.10. arXiv:1111.5632 [physics], pp. 3065–3091. ISSN: 0031-9155, 1361-6560. DOI: 10.1088/0031-9155/57/10/3065. URL: http://arxiv.org/abs/1111.5632 (visited on 02/12/2024).

Sudarshan, Viswanath P, Gary F Egan, Zhaolin Chen, and Suyash P Awate (2020). "Joint PET-MRI image reconstruction using a patch-based joint-dictionary prior". In: *Medical image analysis* 62, p. 101669.

Sutter, Thomas M., Imant Daunhawer, and Julia E. Vogt (Nov. 2020). *Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence*. en. arXiv:2006.08242 [cs, stat]. URL: http://arxiv.org/abs/2006.08242 (visited on 12/05/2023).

— (June 2021). *Generalized Multimodal ELBO*. en. arXiv:2105.02470 [cs, stat]. URL: http://arxiv.org/abs/2105.02470 (visited on 10/31/2023).

Suzuki, Masahiro and Yutaka Matsuo (Mar. 2022). "A survey of multimodal deep generative models". en. In: *Advanced Robotics* 36.5-6. arXiv:2207.02127 [cs, stat], pp. 261–278. ISSN: 0169-1864, 1568-5535. DOI: 10.1080/01691864.2022.2035253. URL: http://arxiv.org/abs/2207.02127 (visited on 08/30/2023).

Watson, Charles C, DMEC Newport, and Mike E Casey (1996). "A single scatter simulation technique for scatter correction in 3D PET". In: *Three-dimensional image reconstruction in radiology and nuclear medicine*. Springer, pp. 255–268.

Wohlberg, Brendt (2017). "ADMM Penalty Parameter Selection by Residual Balancing". In: DOI: 10.48550/ARXIV.1704.06209. URL: https://arxiv.org/abs/1704.06209.

Wu, Mike and Noah Goodman (Nov. 2018). *Multimodal Generative Models for Scalable Weakly-Supervised Learning*. en. arXiv:1802.05335 [cs, stat]. URL: http://arxiv.org/abs/1802.05335 (visited on 11/02/2023).

Xie, Zhaoheng, Tiantian Li, Xuezhu Zhang, Wenyuan Qi, Evren Asma, and Jinyi Qi (2021). "Anatomically aided PET image reconstruction using deep neural networks". In: *Medical Physics* 48.9, pp. 5244–5258. DOI: https://doi.org/10.1002/mp.15051. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.15051. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.15051.