POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

# GMM-DRIT: Learning Disentangled Latent Spaces for MRI Harmonization via Gaussian Mixture Attribute Modeling

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

**Author:** Eleonora Cometa

**Advisor:** Dr. Lara Cavinato

**Co-advisor:** Luca Caldera

**Academic year:** 2024-2025

## 1. Introduction

Collecting brain images via magnetic resonance imaging (MRI) across multiple batches, such as sites or scanners, is essential for advancing clinical studies of the human brain [3]. However, multi-site neuroimaging data suffer from non-biological variability introduced by differences in acquisition settings [3]. MRI harmonization addresses this issue by reducing scanner-dependent variability while preserving biological information. More generally, image harmonization maps images from different domains into a common domain, enabling their analysis as if they were acquired under the same setting [3].

Existing harmonization methods can be divided into feature-based and image-based approaches. While feature-based methods are limited by their dependence on feature extraction, image-based approaches overcome this limitation by operating directly on images. Among image-based methods, disentanglement-based image-to-image (I2I) translation is a promising approach, since it learns to separate scanner-invariant content (anatomy) from scanner-specific attributes (appearance) and generates harmonized images by recombining content with a target style. DRIT++ [4] is a representative framework in this setting, as its cross-cycle mechanism helps preserve anatomical structure, although its unimodal attribute prior may be too restrictive to capture heterogeneous scanner variability. Conversely, GMM-UNIT [5] uses multimodal attribute modeling to better capture scanner-dependent appearance, but lacks the explicit cross-cycle consistency mechanism of DRIT++.

In this work, we build on these complementary perspectives and propose GMM-DRIT, a novel harmonization framework that combines the cross-cycle consistency of DRIT++ with multimodal attribute modeling inspired by GMM-UNIT. Specifically, we extend DRIT++ with a Gaussian Mixture Model (GMM) structure on the attribute space through a Gaussian Mixture Variational Autoencoder (GM-VAE) [2].

The contributions of this thesis are the introduction of a novel multi-domain unpaired MRI harmonization framework based on Gaussian-mixture attribute modeling; the formulation of a unified approach to both domain-to-domain and scanner-free harmonization; and an evaluation showing improved anatomical preservation and scanner-dependent appearance alignment compared to DRIT++.

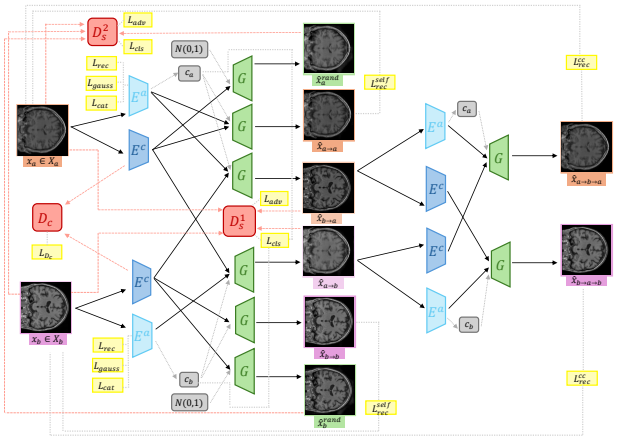

Figure 1: Training iteration in GMM-DRIT.

## 2. Problem Definition

Let $\mathcal{D} = \{1, \ldots, K\}$ be the set of acquisition domains (scanners), and let the dataset be

$$\mathcal{X} = \bigcup_{d \in \mathcal{D}} \mathcal{X}_d, \qquad x \in \mathcal{X}_d \subset \mathbb{R}^{1 \times H \times W}, \qquad (1)$$

where $x$ is a grayscale MRI slice.

During training, the model learns to represent each image by a disentangled pair $(z^c, z^a)$, with $z^c \in \mathcal{C}$ encoding scanner-invariant content (anatomy) and $z^a \in \mathcal{A}$ encoding scanner-dependent attribute (appearance).

At test time, the model supports two harmonization settings. In the first, an input image $x \in \mathcal{X}_a$ can be translated toward a specific target scanner domain $b \in \mathcal{D}$, enabling *domain-to-domain harmonization*, while preserving the anatomical content $z_c$ of the source image. In the second, images from any source domain can be mapped to a common scanner-invariant domain, enabling *scanner-free harmonization* without bias toward a particular scanner.

## 3. Network Architecture

Figure 1 summarizes the main architectural components of GMM-DRIT and the interactions among them. The model consists of a content encoder $E^c$, an attribute encoder $E^a$, a generator $G$, two scanner discriminators $D_s^1, D_s^2$, and a content discriminator $D_c$.

**Content encoder.** The content encoder $E^c$ is a convolutional encoder that maps an input image $x$ to a scanner-invariant content code $z^c = E^c(x)$. This representation is intended to retain anatomical structure while discarding scanner-specific appearance factors.

**Attribute encoder (GM-VAE).** The attribute encoder $E^a$ maps an input image $x$ to a domain-specific attribute code $z^a$ and an inferred mixture assignment $\tilde{y}$ ($(z_a, \tilde{y}) = E^a(x)$). The attribute space is regularized by a $K$-component GMM, and concretely implemented as the inference network of a GM-VAE [2].

**Generator.** The generator $G$ recombines content and attribute representations with the target domain code to synthesize harmonized images in both domain-to-domain and scanner-free settings. In the former, the attribute code $z^a$ is taken from the latent attribute space of the target domain, either extracted from a target-domain image during training or sampled from the corresponding distribution at test time. In the latter, it is sampled from a standard Gaussian, i.e., $z^a_{\text{rand}} \sim \mathcal{N}(0, I)$, and combined with the scanner-free domain code $c_{\text{sf}} = \mathbf{0}$.

**Discriminators.** GMM-DRIT uses adversarial discriminators to encourage realism and scanner consistency in the generated images. The scanner discriminator $D_s$ distinguishes real from generated images and predicts the scanner label, encouraging translations to match the target acquisition domain. The content discriminator $D_c$ predicts the scanner label from the content representation $z^c$, while the content encoder adversarially prevents scanner information from being encoded in $z^c$.

## 4. Training Process

Training is performed in two stages: first, the GM-VAE is pretrained; then, GMM-DRIT is trained with the attribute encoder initialized

from the pretrained model.

**Stage 1 (GM-VAE pretraining).** The GM-VAE is pretrained to learn a structured multimodal attribute space aligned with scanner variability. Its overall objective is

$$\mathcal{L}_{\text{GMVAE}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{gauss}}\mathcal{L}_{\text{gauss}} + \lambda_{\text{cat}}\mathcal{L}_{\text{cat}},$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss, implemented as a binary cross-entropy on the decoder output, $\mathcal{L}_{\text{gauss}}$ regularizes the inferred continuous posterior toward the component-conditioned Gaussian prior, and $\mathcal{L}_{\text{cat}}$ regularizes the inferred component assignments, implemented as a Kullback–Leibler divergence, with respect to a uniform categorical prior.

After convergence, the pretrained GM-VAE initializes the attribute encoder $E^a$ in GMM-DRIT and remains frozen during Stage 2.

**Stage 2 (GMM-DRIT training).** The full harmonization model is trained on unpaired image pairs $\{(x_a, c_a), (x_b, c_b)\}$ from two scanners $a, b \in \mathcal{D}$, where $c_a, c_b \in \{0,1\}^K$ are one-hot domain codes inferred by the GM-VAE. As shown in Figure 1, each training iteration consists of four steps.

**(1) Encode content and attributes.**

$$z_a^c = E^c(x_a), \qquad (z_a^a, c_a) = E^a(x_a),$$
$$z_b^c = E^c(x_b), \qquad (z_b^a, c_b) = E^a(x_b).$$

**(2) Generate translations.** Cross-domain translations are obtained by swapping attribute codes and conditioning on the target domain code:

$$\hat{x}_{a\to b} = G(z_a^c, z_b^a, c_b), \qquad \hat{x}_{b\to a} = G(z_b^c, z_a^a, c_a).$$

Scanner-free translations use $z^a_{\text{rand}} \sim \mathcal{N}(0, I)$ and the scanner-free code $c_{\text{sf}} = \mathbf{0} \in \{0,1\}^K$:

$$\hat{x}_a^{\text{rand}} = G(z_a^c, z^a_{\text{rand}}, c_{\text{sf}}), \quad \hat{x}_b^{\text{rand}} = G(z_b^c, z^a_{\text{rand}}, c_{\text{sf}}).$$

**(3) Re-encoding and reconstruction.** Translated images are re-encoded to compute cross-cycle reconstructions, which enforce consistency of anatomy across translations:

$$\hat{x}_{a\to b\to a} = G(\tilde{z}_b^c, \tilde{z}_a^a, c_a), \quad \hat{x}_{b\to a\to b} = G(\tilde{z}_a^c, \tilde{z}_b^a, c_b),$$

where $(\tilde{z}_b^c, \tilde{z}_b^a)$ and $(\tilde{z}_a^c, \tilde{z}_a^a)$ are obtained by re-encoding $\hat{x}_{a\to b}$ and $\hat{x}_{b\to a}$, respectively. We also compute self-reconstructions:

$$\hat{x}_{a\to a} = G(z_a^c, z_a^a, c_a), \qquad \hat{x}_{b\to b} = G(z_b^c, z_b^a, c_b),$$

to ensure that the latent representations retain sufficient information to reproduce the input when no domain change is required.

**(4) Parameter update.** Using the latent codes and generated samples from steps 1–3, Stage 2 updates all components, except $E^a$, whose parameters are kept frozen.

For image-level supervision, two scanner discriminators are used: $D_s^1$ supervises *encoded* cross-domain translations and $D_s^2$ supervises *scanner-free* translations. Let $i \in \{1, 2\}$ denote the branch ($i = 1$ encoded, $i = 2$ scanner-free). For each branch, the scanner discriminator objective is denoted by $\mathcal{L}_{D_s^i}$ and combines an adversarial real/fake term, $\mathcal{L}^{\text{adv}}_{D_s^i}$, with a scanner-classification term on real images, $\mathcal{L}^{\text{cls}}_{D_s^i}$, implemented as binary cross-entropy (BCE) on the inferred one-hot scanner label.

The generator objective for branch $i$ is denoted by $\mathcal{L}_{G_i}$ and combines an adversarial term, $\mathcal{L}^{\text{adv}}_{G_i}$, which encourages generated images to fool $D_s^i$, and a classification term, $\mathcal{L}^{\text{cls}}_{G_i}$, which encourages the generated output to be classified as belonging to the intended scanner domain.

To preserve anatomy and stabilize unpaired training, we add a self-reconstruction loss $\mathcal{L}^{\text{self}}_{\text{rec}}$ and a cross-cycle reconstruction loss $\mathcal{L}^{\text{cc}}_{\text{rec}}$, both implemented as $L_1$ reconstruction terms.

Finally, the content discriminator objective $\mathcal{L}_{D_c}$ trains $D_c$ to predict scanner labels from content codes $z^c$, implemented as a BCE on the inferred scanner label.

## 5. Dataset and Preprocessing

We evaluated GMM-DRIT using healthy-subject T1-weighted scans from four acquisition domains, for a total of 931 3D volumes. The data come from three public datasets: PPMI (41 volumes) [6], IXI (396 volumes) [1], and SALD (494 volumes) [7]. We define four domains (scanners/acquisition settings): *Triotim* (PPMI; 41), *SALD* (494), *Gyroscan Intera* (IXI; 322), and *Unspecified* (IXI; 74). Although PPMI

and SALD were both acquired on Triotim scanners, we treated them as distinct domains because they come from different acquisition centers, and site-specific protocols may introduce non-biological appearance variability.
For computational efficiency, each 3D volume was converted into grayscale 2D coronal slices of size $182 \times 182$. The resulting dataset comprises 424 Unspecified slices, 963 Gyroscan Intera slices, 888 SALD slices, and 375 Triotim slices.

## 6. Experimental Setup

The experimental setup is designed to evaluate GMM-DRIT along three complementary aspects: the quality of the learned attribute space, the quality of image-level harmonization, and the alignment between generated and real image distributions. To this end, we define the following evaluation protocol.
**(E1) Learned attribute-space quality.** We assessed whether the pretrained GM-VAE captures meaningful scanner-related structure in the attribute space $\mathcal{A}$ through quantitative and qualitative analyses. Quantitatively, we report clustering metrics, namely Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Qualitatively, we present a 2D PCA projection of the learned mixture components to visualize scanner-dependent modes in the latent space.
**(E2) Structural SSIM for scanner-free harmonization.** We evaluated anatomical preservation in the scanner-free setting using the structural component of the Structural Similarity Index Measure (SSIM), computed between each input image and its scanner-free harmonized output.
**(E3) Luminance SSIM for domain-to-domain harmonization.** We evaluated target-domain appearance consistency using the luminance component of SSIM. For each target scanner, the metric was computed on $M = 5000$ randomly sampled pairs, where each pair consists of a real image from the target domain and an image translated into that domain. As a reference, we also report the within-scanner real-vs-real luminance SSIM baseline.
**(E4) Cross-scanner Luminance SSIM Before and After Harmonization** To quantify the reduction of inter-scanner intensity differences, we computed luminance SSIM between randomly sampled image pairs from different domains before and after harmonization.
**(E5) Distribution similarity in feature space.** We report Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) in a learned feature space to quantify how closely translated images match the distribution of real target-domain images. Since these metrics require sufficiently large sample sets, this evaluation is restricted to the two largest target domains (SALD and Gyroscan Intera).
**(E6) Qualitative assessment.** We qualitatively inspect scanner-free harmonization and domain-to-domain translations across all scanners to assess whether anatomy is preserved while scanner-dependent appearance is adapted to the desired target setting.
**(E7) Intensity distribution analysis.** We compared voxel-intensity distributions across scanners before and after scanner-free harmonization to evaluate the reduction of scanner-dependent intensity variability.

### 6.1. Baseline Comparison

We compared GMM-DRIT with DRIT++ [4], a multi-domain image-to-image translation model with a unimodal attribute prior, evaluated, whenever possible, under the same settings.

## 7. Results

This section reports the main results of the proposed evaluation protocol (E1–E7), including comparisons with DRIT++.

### 7.1. (E1) Learned Attribute Space

Figure 2 shows a 2D PCA projection of the GM-VAE mixture components in the attribute space $A$, obtained with batch size = 64, Gaussian size = 64, $\lambda_{\text{gauss}} = 2.0$, $\lambda_{\text{categ}} = 1.0$, and $\lambda_{\text{rec}} = 0.7$. Clustering performance is ACC = 83.251%, NMI = 0.6219, and ARI = 0.6028.
The attribute space shows a clear multi-modal structure, with distinct regions associated with scanner-dependent appearance modes. Partial overlap was expected, as scanner effects coexist with shared anatomical variability in MRI data. Overall, the PCA projection suggests that the GM-VAE mainly captures acquisition-related variation rather than purely anatomical differences; in particular, the second principal

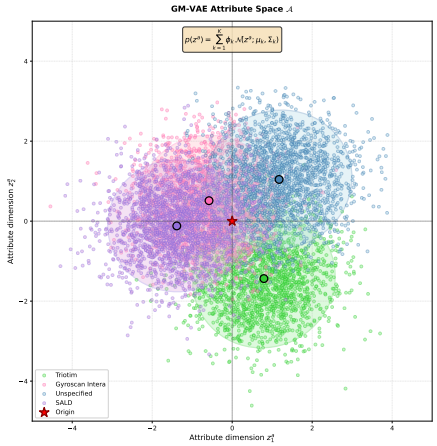


Figure 2: PCA projection of the learned GM-VAE attribute space.

Table 1: Structural SSIM ($\mu \pm \sigma$).

| Scanner | GMM-DRIT | DRIT++ | Δ (95% CI) |
|---|---|---|---|
| SALD | 0.96 ± 0.01 | 0.28 ± 0.08 | [0.67, 0.69] |
| Gyr. Intera | 0.97 ± 0.01 | 0.30 ± 0.08 | [0.66, 0.68] |
| Unspecified | 0.96 ± 0.01 | 0.21 ± 0.11 | [0.73, 0.77] |
| Triotim | 0.96 ± 0.01 | 0.32 ± 0.07 | [0.63, 0.66] |

component may reflect scanner-dependent appearance factors, such as global intensity and contrast, which help separate Unspecified from Triotim in the projected attribute space.

### 7.2. (E2) Structural SSIM for Scanner-free Harmonization

Table 1 shows that GMM-DRIT preserves anatomical structure across all scanners, with consistently high structural SSIM, indicating stable anatomical preservation after harmonization. DRIT++ instead yields markedly lower scores and higher variability. The 95% bootstrap confidence intervals are strictly positive for all scanners, confirming the advantage of GMM-DRIT in anatomical preservation.

### 7.3. (E3) Luminance SSIM for Domain-to-domain Translation

Table 2 shows that GMM-DRIT achieves high luminance similarity across all target scanners, closely matching the within-scanner real-vs-real baseline. DRIT++ also improves luminance alignment but remains consistently below GMM-DRIT, with strictly positive 95% bootstrap confidence intervals of the mean difference for all targets. The larger variability observed

Table 2: Luminance SSIM ($\mu \pm \sigma$).

| Target | GMM-DRIT | DRIT++ | Real | Δ (95% CI) |
|---|---|---|---|---|
| SALD | 0.99 ± 0.01 | 0.97 ± 0.03 | 0.99 ± 0.01 | [0.02, 0.03] |
| Gyr. Int. | 0.99 ± 0.01 | 0.98 ± 0.02 | 0.99 ± 0.01 | [0.01, 0.01] |
| Unspec. | 0.95 ± 0.04 | 0.92 ± 0.05 | 0.95 ± 0.06 | [0.01, 0.03] |
| Triotim | 0.99 ± 0.01 | 0.98 ± 0.02 | 0.99 ± 0.0030 | [0.02, 0.03] |

Table 3: Cross-scanner luminance SSIM before/after scanner-free harmonization.

| Method | Pre | Post | Δ (95% CI) |
|---|---|---|---|
| DRIT++ | 0.97±0.04 | 0.99±0.01 | [0.02,0.02] |
| GMM-DRIT | 0.97±0.04 | 0.99±0.004 | [0.02,0.02] |

for the Unspecified domain is consistent with its lower within-scanner baseline and likely reflects higher intrinsic heterogeneity rather than harmonization errors.

### 7.4. (E4) Cross-scanner Luminance SSIM Before and After Harmonization

Table 3 shows that both methods improve cross-scanner luminance similarity after harmonization. However, GMM-DRIT achieves a higher post-harmonization score, lower variability, indicating a stronger and more consistent scanner-free harmonization effect across domains.

### 7.5. (E5) Distribution Similarity in Feature Space (FID and KID)

GMM-DRIT achieves substantially lower FID and KID than DRIT++ (FID = 0.7611, KID = 0.000578 vs. 11.7100 and 0.013435), indicating that the feature distribution of translated images matches more closely the distribution of real images from the target scanner.

### 7.6. (E6) Qualitative Assessment

Figure 3 shows representative examples of scanner-free and domain-to-domain harmonization across the four domains. For each source image, we report the input slice, the scanner-free output, and the translations to the target domains. Qualitatively, GMM-DRIT reduces scanner-dependent intensity/contrast differences while preserving anatomical structures in the scanner-free setting. In domain-to-domain harmonization, image appearance adapts to the selected target scanner without altering anatomical structure.

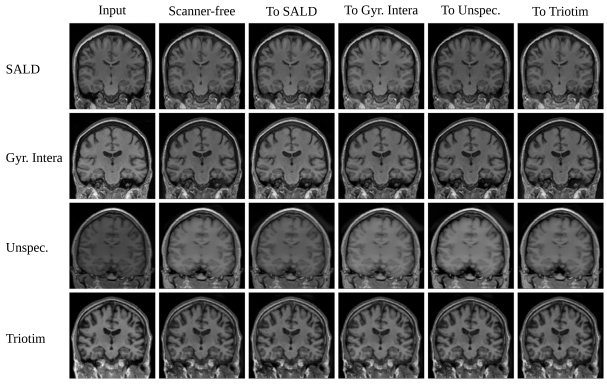


Figure 3: Qualitative results of GMM-DRIT.

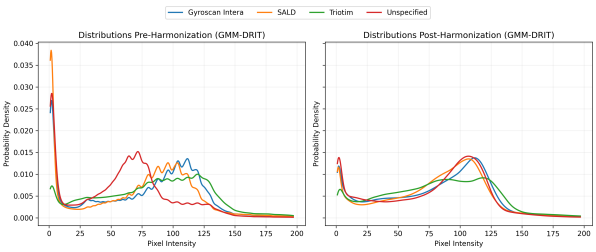


Figure 4: Voxel-intensity distributions before and after scanner-free harmonization.

### 7.7. (E7) Intensity Distribution Analysis

Figure 4 compares voxel-intensity distributions across scanners before and after scanner-free harmonization. Before harmonization, the distributions show scanner-dependent differences in peak position and spread. After scanner-free harmonization, the curves become more consistent across scanners, indicating reduced scanner-dependent intensity variability.

## 8. Conclusions

This thesis addressed unpaired multi-domain MRI harmonization in realistic multi-site settings, where paired scans across scanners are rarely available and anatomical fidelity must be preserved. To this end, we proposed GMM-DRIT, a disentanglement-based image-to-image translation framework that extends DRIT++ with a Gaussian-mixture attribute space learned through a GM-VAE. The main contributions of this thesis are the introduction of this framework, the definition of a unified approach to both domain-to-domain and scanner-free harmonization, and its empirical evaluation against DRIT++.

The results show that Gaussian-mixture attribute modeling provides a more expressive representation of scanner-dependent variability and improves harmonization performance. Compared with DRIT++, GMM-DRIT achieves stronger anatomical preservation, better scanner-dependent appearance alignment, and improved feature-space distribution matching, with qualitative findings consistent with the quantitative analysis. Overall, these results indicate that representing the attribute space with a Gaussian Mixture Model is an effective design choice for capturing the complexity of scanner-dependent appearance variability while preserving anatomical content.

This study is limited by the use of 2D slices, the restricted size and imbalance of the available datasets, and the limited exploration of hyperparameters due to computational constraints. Future work could therefore extend the framework to 3D volumes, evaluate it on larger and more balanced datasets, and assess its impact on downstream clinical tasks such as segmentation or disease prediction.

## References

[1] Brain Development Project. IXI – Information eXtraction from Images Dataset, 2019.

[2] Dilokthanakul et al. Deep unsupervised clustering with gaussian mixture variational autoencoders. 2016.

[3] Hu et al. Image Harmonization: A Review of Statistical and Deep Learning Methods for Removing Batch Effects and Evaluation Metrics for Effective Harmonization. 2023.

[4] Lee et al. DRIT++: Diverse Image-to-Image Translation via Disentangled Representations. 128:2402 – 2417, 2019.

[5] Liu et al. GMM-UNIT: Unsupervised Multi-Domain and Multi-Modal Image-to-Image Translation via Attribute Gaussian Mixture Modeling, 2020.

[6] Marek et al. The Parkinson progression marker initiative (PPMI). 2011.

[7] Wei et al. Structural and functional MRI from a cross-sectional Southwest University Adult lifespan Dataset (SALD). 2017.