Mitigating analytical variability in fMRI results with style transfer

Elodie Germani, Camille Maumet, Elisa Fromont

Graphical Abstract

Mitigating analytical variability in fMRI with style transfer

Elodie Germani, Camille Maumet, Elisa Fromont

Highlights

Mitigating analytical variability in fMRI with style transfer

Elodie Germani, Camille Maumet, Elisa Fromont

- We explore the ability to convert fMRI statistic maps across different pipelines using generative models (Generative Adversarial Networks (GAN) and Diffusion models (DM) frameworks, in supervised and unsupervised settings).
- To enhance conversion performance, we explore several modifications of traditional DM frameworks by conditioning on the average latent representation of multiple target images.
- Our results show that images can be converted successfully using DM, but with lower similarity with the ground truth compared to GAN, in particular in supervised settings.

Mitigating analytical variability in fMRI with style transfer

Elodie Germani[a], Camille Maumet[a,1,*], Elisa Fromont[b,1]

[a] *Univ Rennes, Inria, CNRS, Inserm, Rennes, France*
[b] *Univ Rennes, IUF, Inria, CNRS, Rennes, France*

Abstract

We propose a novel approach to improve the reproducibility of neuroimaging results by converting statistic maps across different functional MRI pipelines. We make the assumption that pipelines used to compute fMRI statistic maps can be considered as a style component and we propose to use different generative models, among which, Generative Adversarial Networks (GAN) and Diffusion Models (DM) to convert statistic maps across different pipelines. We explore the performance of multiple GAN frameworks, and design a new DM framework for unsupervised multi-domain style transfer. We constrain the generation of 3D fMRI statistic maps using the latent space of an auxiliary classifier that distinguishes statistic maps from different pipelines and extend traditional sampling techniques used in DM to improve the transition performance. Our experiments demonstrate that our proposed methods are successful: pipelines can indeed be transferred as a style component, providing an important source of data augmentation for future medical studies.

Keywords: style transfer, generative models, analytical variability, functional MRI, data re-use

1. Introduction

Over the last decades, the question of understanding brain functions took an important place in many research fields ranging from medicine and psychology to artificial intelligence and philosophy. With the development

*Corresponding author: camille.maumet@inria.fr
[1]Joint senior authorship

of brain imaging techniques such as task-based functional Magnetic Resonance Imaging (task-fMRI), researchers can now explore brain activity of individuals while they perform predefined tasks, and get a better understanding of the neural correlates of different cognitive processes. The number of published studies making use of this modality exploded in the last ten years: in 2018, more than one thousand studies registered in the website `clinicaltrial.gov` were using fMRI as an outcome measure (Sadraee et al., 2021).

However, the "reproducibility crisis" that affected scientific research raised concerns regarding the reliability of published findings, see for example in neuroimaging (Button et al., 2013; Poldrack et al., 2017; Botvinik-Nezer and Wager, 2023). In particular, the low statistical power of task-fMRI studies has been criticized, as it led to lower probabilities of identifying true effects but also to higher probabilities of reporting false positive findings in the literature (Ioannidis, 2005). Since then, efforts have been made to increase sample sizes and thus, statistical power, for instance by acquiring raw data from a larger number of participants for a few number of cognitive tasks (*e.g.* UK Biobank (Sudlow et al., 2015) or Human Connectome Project (Van Essen et al., 2013b)) or for a small number of participants on a larger number of cognitive tasks (e.g. Individual Brain Charting (Pinho et al., 2018)). But the number of research questions that can be explored is always limited by the characteristics of each dataset. To tackle these challenges, a promising solution would be to combine together data from different studies. Moreover, with the increased adoption of data sharing (Poline et al., 2012), more and more neuroimaging data are made available on dedicated platforms (*e.g.* OpenNeuro (Markiewicz et al., 2021) or NeuroVault (Gorgolewski et al., 2015)). Re-using shared data in combined studies would allow researcher to explore new research questions, with larger and more diverse datasets, while bypassing the difficulties associated with acquiring new data.

Yet, neuroimaging datasets are often very heterogeneous. In the case of task-fMRI, raw data are 4-dimensional matrices and correspond to a 3-dimensional brain volume that is acquired at different time points. To study the activity of the brain during the task, these raw data are first preprocessed to correct for spatial and temporal noise. Preprocessed data are then submitted to a first-level statistical analysis (*i.e.* at the run or subject-level), and potentially to a second-level statistical analysis, also known as group-level analysis. This chain of processing and analysis steps applied to raw data is called a "pipeline". In the end, these pipelines result in statistic maps that

represent the activation of the brain during the task. Derived data, such as statistic maps at the first- or second-level, could be combined instead of raw data through meta- and mega-analyses (Costafreda, 2009). This process is easier due to reduced privacy requirements, but also because it avoids having to perform costly re-computations. However, pipelines in neuroimaging, and in particular in task-fMRI, are highly flexible (Carp, 2012), and these shared derived data were often processed differently. Indeed, at each step to build their pipeline, researchers have to make choices between different computing environments, different software packages and different algorithms.

Over the last decade, multiple studies explored the impact of analytical choices on the results of neuroimaging studies and found that a slight change in a pipeline could lead to variations not only in the statistic maps, but also in the final findings (Botvinik-Nezer et al., 2020; Bhagwat et al., 2021; Glatard et al., 2015). This phenomenon, also known as "analytical variability", takes part in the "reproducibility crisis" (Botvinik-Nezer and Wager, 2023), but also put into question the ability to re-use derived data computed with different pipelines. In a recent study, Rolland et al. (2022) showed that combining derived data computed with different pipelines in mega-analyses could lead to a higher risk of false positives. In such context, it would be useful to find an approach to mitigate the effect of analytical variability to benefit from the large amount of shared derived data.

For similar purposes, *i.e.* to mitigate the effect of different sources of variability, researchers usually perform data harmonisation. In particular, recent advances in computer vision gave rise to techniques such as style transfer (Gatys et al., 2016) that allows to learn mappings between different domains to convert and harmonize datasets. Style transfer frameworks make use of generative models, such as Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) or Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020). In supervised settings, frameworks such as Pix2Pix (Isola et al., 2017) or Palette (Saharia et al., 2022) can be trained to learn a mapping between pairs of data from different domains and apply this mapping to new unseen data. Unsupervised frameworks (Zhu et al., 2017; Sasaki et al., 2021; Liu et al., 2017) do not necessitate pairs of data in different domains for their training as they use constraints like cycle consistency in CycleGAN (Zhu et al., 2017) or shared latent space assumption (Liu et al., 2017; Sasaki et al., 2021). They provide a good opportunity to benefit from large unlabeled databases to learn complex mapping without any ground-truth target data. By conditioning on domain-specific features (*e.g.*

class vector, latent space of auxiliary classifiers, etc.) instead of full target images, unsupervised frameworks also extend to multi-domain transitions (Choi et al., 2018, 2021; Ho and Salimans, 2021) to learn transfer between multiple domains in a single model.

In medical imaging, style transfer frameworks have already been used for multiple tasks (Kaji and Kida, 2019), including modality transition (Armanious et al., 2020; Denck et al., 2021; Jin et al., 2019; Kong et al., 2021; Lyu and Wang, 2022; Nie et al., 2018; Ozbey et al., 2023; Qin et al., 2022; Wolterink et al., 2017a; Yang et al., 2020), image denoising (Yang et al., 2018; Wolterink et al., 2017b; Armanious et al., 2020), or data harmonisation across different sites or scanner (Bashyam et al., 2022; Liu et al., 2021). These frameworks allow researchers to more easily extract information from heterogeneous datasets, in particular when building multi-modal or multi-centric datasets. This can be done by learning to generate data from missing or noisy modality or by learning to mitigate the effect of variability in data coming from different sites. For instance, using a conditional GAN coupled with a perceptual loss and a style transfer loss, MedGAN (Armanious et al., 2020) showed its performance in PET to CT-scan translation as well as PET denoising and correction of MRI artifacts. In supervised settings, Nie et al. (2018) used a variant of Pix2Pix (Isola et al., 2017) with a gradient-based loss function for MRI to CT translation. Yang et al. (2018) also used a conditional GAN for low-dose to high-dose CT translation, with pixelwise loss associated with a minimization of the Wasserstein distance and a perceptual similarity loss. For the same application, Wolterink et al. (2017b) proposed to get rid of paired datasets and showed the potential of CycleGAN. This model also showed its potential for stain normalization in histological images (Shaban et al., 2019).

Since their emergence, more and more DDPM (Ho et al., 2020) style transfer frameworks were developed (Lyu and Wang, 2022; Ozbey et al., 2023; Pan et al., 2023; Dorjsembe et al., 2024; Jiang et al., 2023). Lyu and Wang (2022) showed the superiority of diffusion models compared to GAN for the conversion between MRI and CT using a supervised framework. In unsupervised settings, Pan et al. (2023) developed a cycle-guided framework composed of two DDPM that condition each other to generate synthetic images from two different MRI pulse sequences. Similarly, Ozbey et al. (2023) proposed SynDiff with a source-conditional adversarial projector that denoises the target image sample with guidance from the source image.

In this work, we explore the ability of style transfer frameworks to convert

task-fMRI derived data, *i.e.* statistic maps, between pipelines. Our goal is to propose a solution to mitigate the effect of analytical variability in fMRI statistic maps to build more valid mega-analyses and benefit from the large amount of derived data shared on public databases. To be useful in real practice, the proposed method should rely on unpaired data (*i.e.* could be trained without access to the ground-truth target images) and perform multi-domain transitions (*i.e.* learn multiple transfers using a single model). However, to the best of our knowledge, this application of style transfer to conversion of data between different analysis pipelines is new and off the shelf methods do not directly apply as these were not designed on the same type of data.

Moreover, DDPM are challenging to control when the objective is to generate images that maintain the intrinsic properties of the source images while transferring the extrinsic properties to the target domain. Indeed, these are iterative generative models, *i.e.* they learn to model the transition from a Gaussian distribution to a target data distribution. Thus, data generated by the DDPM depend on the initial samples drawn from the Gaussian distribution, usually at random.

To tackle these challenges, we made the following contributions:

- We are the first to make the assumption that pipelines can be considered as a style property of statistic maps which can be transferred between maps.
- We re-implement three state-of-the-art style transfer frameworks based on GAN, namely Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) and StarGAN (Choi et al., 2018), and adapt them to our 3-dimensional statistic maps.
- We extend a state-of-the-art conditional DDPM (Ho and Salimans, 2021) and adapt it for style transfer by conditioning the sampling on the source image.
- We explore different types of conditioning for our DDPM framework: using class vectors, and using the latent space of multiple target images in a classifier trained to distinguish statistic maps between pipelines, a task previously unexplored.
- We compare the performance of these different frameworks to convert statistic maps between pipelines with different degrees of distance.

2. Materials and Methods

2.1. Dataset

We use group-level statistic maps from the *HCP multi-pipeline dataset*. More details about this dataset can be found in Germani et al. (2023). Briefly, this dataset is composed of subject-level (1,080 participants) and group-level (1,000 groups) statistic and contrast maps derived from raw data of the Human Connectome Project Young Adult S1200 release (Van Essen et al., 2013b). In this dataset, raw fMRI data for the motor task were analyzed with 24 different pipelines for the 5 contrasts: *right-hand*, *right-foot*, *left-hand*, *left-foot* and *tongue*. The pipelines used in this dataset vary in terms of software package, smoothing kernel Full-Width at Half-Maximum (FWHM), number of motion regressors and derivatives of the Haemodynamic Response Function (HRF) included in the first-level analysis.

We explore in particular the statistic maps obtained with four different pipelines that differ in terms of software package (SPM (Penny et al., 2011) or FSL (Jenkinson et al., 2012)) and presence or absence of the derivatives of the HRF for the first-level analysis. We use all the available group-level statistic maps ($N = 1,000$) for each pipeline for the contrast *right-hand*. In the following, these pipelines will be labelled with "<software>-<derivatives>", for instance "fsl-1" means use of FSL software package and HRF derivatives.

The selected group-level statistic maps are resampled to a size of 48 x 56 x 48 and masked using the intersection mask of all groups. The voxel values are normalized between -1 and 1 for each statistic maps using a min-max operation. The 1,000 groups are split into train and test with a 80/20 ratio and all models are trained and evaluated on the same sets. Further investigation about possible data leakage across groups is provided in Supplementary Figure 1 (Germani et al., 2024b).

2.2. GAN frameworks

First, we assess the potential of GAN frameworks to convert statistic maps between pipelines. In particular, we evaluate the performance of Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) and StarGAN (Choi et al., 2018). A detailed description of each framework is available in the corresponding papers and we provide a quick description of the main properties of these models in Table 1.

Framework	Learning	Transition	Loss
Pix2Pix (Isola et al., 2017)	Supervised	One-to-one	Adversarial Reconstruction
CycleGAN (Zhu et al., 2017)	Unsupervised	One-to-one	Adversarial Cyclic
StarGAN (Choi et al., 2018)	Unsupervised	Multi-domain	Adversarial Cyclic Classification

Table 1: Description of GAN frameworks

Architecture and training. We use the default architecture of these models, as described in their respective papers, and we only modify the 2-dimensional convolutions and batch normalization layers to cope with our 3-dimensional statistic maps. These were implemented using PyTorch (Paszke et al., 2019) and each framework was trained for 200 epochs on 1 GPU NVIDIA Tesla V100.

2.3. DDPM frameworks

Due to the promising performance of DDPM on natural images and medical imaging (Dhariwal and Nichol, 2021), we also assess the potential of DDPM frameworks. However, there is only few DDPM frameworks developed for style transfer applications, and to our knowledge, all of them rely on paired datasets (Saharia et al., 2022) or learn only one-to-one transitions (Pan et al., 2023). Thus, to perform multi-domain transitions, we adapt an existing conditional DDPM to style transfer tasks. In particular, we use the framework from Ho and Salimans (2021), which generates images conditioned using a one-hot encoding of the class (*i.e.* class vector). We also extend this model to a conditioning based on the latent space of the classifier, inspired from Preechakul et al. (2022). Both are unsupervised frameworks, learning multi-domains transitions. A more detailed description of the original framework is available in Ho and Salimans (2021).

In Figure 1, we illustrate the design of our DDPM framework, with the main modifications applied to the basis of Ho and Salimans (2021). Figure 1 (A), (C) and (D) represent the conditional diffusion used in Ho and Salimans (2021), that we enhanced using source content preservation and classifier conditioning (Figure 1 (B)). In the following, we describe in more details these modifications.

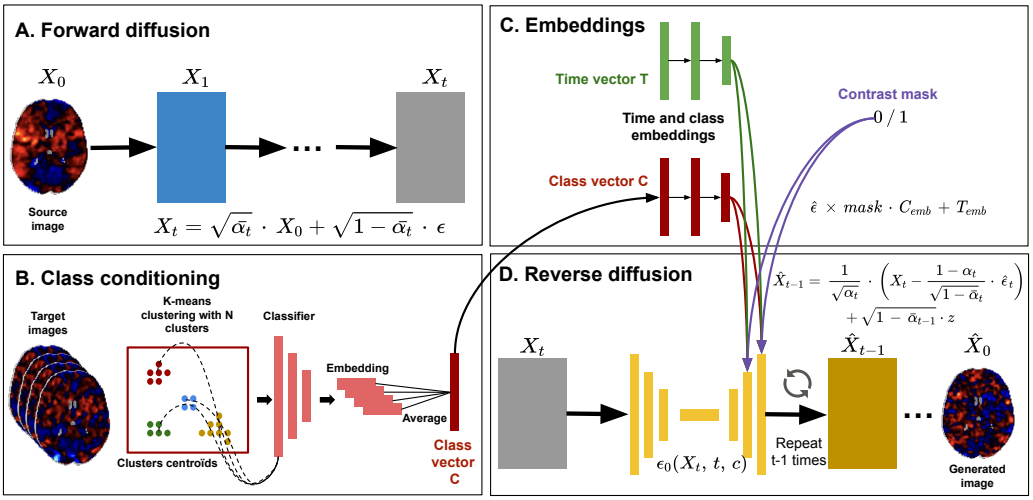

Figure 1: Diagram of the workflow. During the forward diffusion (A), original maps X_0 are turned into X_t after t steps of noise addition ϵ. (B) Class conditioning uses latent vectors extracted from a classifier. These are averaged across $N = 10$ images, which are the centroids of $N = 10$ clusters identified using a K-Means algorithm. (C) Time and class are embedded using two Multi-Layers Perceptrons (MLP). A mask is applied to the class conditioning vector to jointly train an unconditional model with a pre-defined probability. (D) During the reverse diffusion, the neural network $\epsilon_\theta(X_t, t, c)$ learns to predict the noise added to the image and reconstructs X_{t-1} iteratively until $t = 0$.

Source content preservation. To adapt our DDPM framework to style transfer, our main objective is to find a solution to generate images that still contain the intrinsic properties of the source image. In Saharia et al. (2022), authors concatenated the source image along with random Gaussian noise to initialize the diffusion. Here, we propose to fix the initial state of the DDPM by directly using the forward diffusion process of traditional DDPM to generate a noisy version of the source image X_t. Then, the noisy source image is iteratively denoised using the predicted noise and the reverse diffusion process with an additional conditioning on the target domain.

Classifier conditioning. We also develop an extension of the model from Ho and Salimans (2021) to condition the generation based on the latent space of a classifier (see Figure 1 (B)). Indeed, in Ho and Salimans (2021), the diffusion is conditioned using a one-hot encoding of the domain, which decreases the diversity of samples. In Preechakul et al. (2022), a semantic encoder is used to guide sampling. Thus, we extend this idea by conditioning the model

using a latent feature vector extracted from a pre-trained CNN. This CNN was pre-trained to predict the pipeline used to obtain the statistic maps (*i.e.* the component we are trying to transfer between statistic maps). The features are extracted just before the fully connected layer, to get a good representation, useful to distinguish images across pipelines.

Multi-target images. To condition on the latent space of this classifier during sampling, some target images must be selected. In Choi et al. (2021), authors showed that conditioning on multiple images generates images that share coarse or fine features with the target ones depending on the number of selected images. Selecting multiple target images to convert images between domains can help to generate images that represent the diversity of the target domain. In practice, the whole set of images available in the target domain could be used. This is impractical for large datasets and might lead the model to focus on specific patterns of the target domain if these are over-represented in the dataset. Here, we chose to condition the sampling on the voxel-to-voxel mean of the selected target images. We implemented several variations to explore the impact of the choice of these target images.

- **Number of target images**: N=5, 10 or 20.

- **Target images selection**: random (∞), using a K-means algorithm, or using a K-Nearest Neighbors algorithm.

For the target image selection, we proposed several algorithms. We used the K-Means algorithm (MacQueen, 1967) to identify N clusters of images in the target domain (see Figure 1 (B)). Then, we extract the centroid of these clusters and average their latent vector for conditioning. We also compared the selection process with a random sampling of target images and with a sampling based on the identification of images that are close to the source image using a K-Nearest Neighbors algorithm (Mucherino et al., 2009).

Architecture and training. The neural network used in the DDPM to predict the noise follows a simple U-Net architecture (Ronneberger et al., 2015) with two downsampling and upsampling blocks with 3D convolutions layers and skip connections. The hyperparameters of the DDPM are the following: $t = 500$ diffusion steps; linear noise schedule with variances in the range of $\beta_1 = 10^4$ and $\beta_t = 0.02$; batch size of 8 and learning rate of 1e-4. The weight w used to control the conditional guidance is optimized on the validation

set by comparing $w = 0, w = 0.5$ and $w = 2$ and a value of 0.5 was found to give the best results in terms of Pearson's correlation coefficient between the target ground-truth and the generated image on this set. The model is implemented using PyTorch (Paszke et al., 2019) and trained for 200 epochs on 1 GPU NVIDIA Tesla V100.

The CNN used to extract class conditional features is composed of five 3-dimensional convolution layers with 3-dimensional batch normalization and leaky rectified linear units (ReLU) activation functions, followed by a fully connected layer. The latent space corresponds to a $4,096$ flatten vector which is injected as conditioning to the U-Net. It is trained for 150 epochs using a learning rate of 1e-4 and a batch size of 64 on 1 GPU NVIDIA Tesla V100.

2.4. Evaluation of performance

We evaluated the performance of the GAN and DDPM frameworks using different metrics. In the following equations, we use X_A, X_B and X_{AB} to respectively define the source image, target image and translated image.

We used two types of metrics: Pearson's correlation coefficient and Mean Squared Error (MSE) to study the adequacy of generated images to the ground truth target, and Inception Score (Salimans et al., 2016) (IS) to explore the quality and diversity of the generated images. IS combines the confidence of the class predictions (*i.e.* each image's label distribution $p(Y|X)$) with the variety in the output of the model (*i.e.* the marginal label distribution for the whole set of images $P(Y)$).

- Pearson's correlation (Corr.) in percent

$$r = \frac{\sum_{i=1}^{n}(X_{AB_i} - \overline{X_{AB}})(X_{B_i} - \overline{X_B})}{\sqrt{\sum_{i=1}^{n}(X_{AB_i} - \overline{X_{AB}})^2}\sqrt{\sum_{i=1}^{n}(X_{B_i} - \overline{X_B})^2}} \tag{1}$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n} * \sum_{i=1}^{n}(X_{AB_i} - X_{B_i})^2 \tag{2}$$

- Inception Score (IS) (Salimans et al., 2016) computed by passing the generated images of each model through the pipeline classifier to obtain probability distributions of labels that are used to compute the

score. Note that, we did not compute this score for frameworks learning only one-to-one transitions because we would then have obtained a score computed with images generated by different generators. In the following equation, X refers to any generated image, and Y refers to the corresponding target label.

$$IS(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(Y|X) \parallel p(Y))) \quad (3)$$

As an additional evaluation criterion, we used the pipeline classifier developed for our DDPM framework to classify the generated images and verify if these images were correctly classified in the target pipeline class.

3. Results

3.1. GAN frameworks

		fsl-1 → spm-0	spm-0 → fsl-1	fsl-1 → spm-1	fsl-1 → fsl-0
	IS	Mean correlations (Std. errors)			
Initial	*3.69*	*78.2 (0.5)*	*78.2 (0.5)*	*82.8 (0.3)*	*92.3 (0.5)*
Pix2Pix	-	**91.4 (0.1)**	**89.1 (0.2)**	**90.1 (0.2)**	**97.4 (0.1)**
CycleGAN	-	85.5 (0.3)	67.1 (0.4)	70.0 (0.5)	71.2 (0.4)
StarGAN	3.63	90.5 (0.4)	86.8 (0.5)	87.6 (0.5)	91.5 (0.3)

	fsl-1 → spm-0	spm-0 → fsl-1	fsl-1 → spm-1	fsl-1 → fsl-0
	Mean MSE (Std. errors)			
Initial	*0.0076 (0.0003)*	*0.0076 (0.0003)*	*0.0041 (0.0002)*	*0.0022 (0.0001)*
Pix2Pix	**0.0027 (0.0001)**	**0.0014 (0.0001)**	**0.0025 (0.0001)**	**0.0005 (0.0)**
CycleGAN	0.0049 (0.0002)	0.0049 (0.0002)	0.0072 (0.0002)	0.0048 (0.0001)
StarGAN	0.0035 (0.0002)	0.002 (0.0001)	0.0035 (0.0001)	0.0017 (0.0001)

Table 2: Performance in pipeline-to-pipeline transfer for the GAN frameworks. IS is the "Inception Score" across all transfers. Pearson's correlation (%) (upper table) and Mean Squared Error (MSE) (lower table) computed between generated and ground truth image and averaged across 20 images per transfer. *Initial* represents the metrics between the source image (before transfer) and the ground-truth target image. **Boldface marks the top model** for each of the 4 transfers studied. Note: as explained before, Inception score was not computed for the one-to-one transition models such as Pix2Pix and CycleGAN.

In Table 2, we show the performance of the GAN frameworks for four transfers, between pipelines with: a different software and a different HRF (columns 1-4), a different software and the same HRF (columns 4-6) and, the

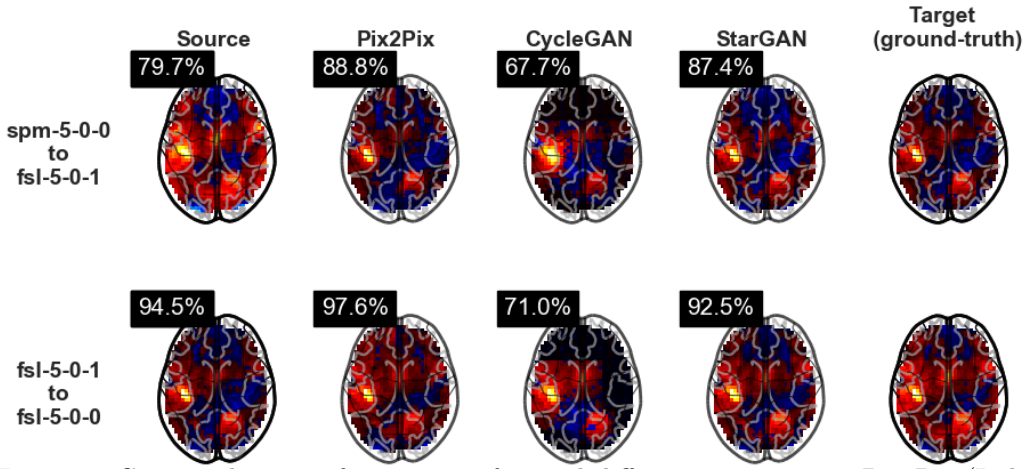

Figure 2: Generated images for two transfers and different competitors: Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) and starGAN (Choi et al., 2018). Correlation with the target are indicated above each image (in percent).

same software and a different HRF (columns 6-8). Overall, using Pix2Pix (Isola et al., 2017) and StarGAN (Choi et al., 2018), the conversion of statistic maps between pipelines is successful, with increased correlations between target and generated maps compared to correlations between source and target (similar observations are made with MSE), *e.g.* 91.4% for target-generated compared to 76.2% for source-target with Pix2Pix for conversion "fsl-1" to "spm-0".

We can point out the large superiority of the supervised Pix2Pix framework compared to the 3 alternatives, which are all unsupervised. By benefiting from paired data, Pix2Pix succeeds in generating images closer to the target image than to the source image for all transfers. Correlations between target and generated images are close to 0.9, which is nearly perfect. On the other hand, the CycleGAN (Zhu et al., 2017) framework gives surprising results, relatively low compared to the other GAN frameworks. While it makes use of a cyclic loss in unsupervised settings, similarly to StarGAN (Choi et al., 2018), this framework only learns transfers between two domains. We can suppose that StarGAN leverages the data from the multiple source domains and benefits from the additional classification loss, leading to higher performance in similar settings.

In Figure 2, we illustrate two transfers: (first row) between pipelines with different software packages and different HRF (spm-0 to fsl-1) and (second

row) between pipelines with the same software package and different HRF (fsl-1 to fsl-0). For these two transfers, a random statistic maps of the source pipeline was chosen and we generated the corresponding converted map in the target pipeline. We also display the ground-truth statistic maps in the target pipeline (*i.e.* the same raw data as the source but a different pipeline). Maps generated using Pix2Pix are closer to the target ground-truth, with more similar patterns, as seen with the similarity metrics.

3.2. DDPM frameworks

		fsl-1 → spm-0	spm-0 → fsl-1	fsl-1 → spm-1	fsl-1 → fsl-0
	IS	Mean correlations (Std. errors)			
Initial	*3.69*	*78.2 (0.5)*	*78.2 (0.5)*	*82.8 (0.3)*	*92.3 (0.5)*
One-hot	3.66	83.9 (0.7)	75.0 (0.9)	78.8 (0.8)	81.1 (0.6)
N=1	3.70	85.4 (0.6)	77.4 (0.8)	80.1 (0.8)	82.8 (0.8)
N=10, ∞	**3.86**	**86.1 (0.4)**	**78.9 (0.6)**	**81.5 (0.4)**	**84.1 (0.6)**

	fsl-1 → spm-0	spm-0 → fsl-1	fsl-1 → spm-1	fsl-1 → fsl-0
	Mean MSE (Std. errors)			
Initial	*0.0076 (0.0003)*	*0.0076 (0.0003)*	*0.0041 (0.0002)*	*0.0022 (0.0001)*
One-hot	0.0097 (0.0014)	0.0048 (0.0007)	0.0088 (0.0014)	0.0043 (0.0003)
N=1	0.0053 (0.0003)	0.0037 (0.0003)	0.0073 (0.0009)	0.0037 (0.0003)
N=10, ∞	**0.0043 (0.0003)**	**0.0028 (0.0002)**	**0.0049 (0.0003)**	**0.0029 (0.0002)**

Table 3: Performance in pipeline-to-pipeline transfer for DDPM frameworks. IS is the "Inception Score" across all transfers. Pearson's correlation (%) (upper table) and Mean Squared Error (MSE) (lower table) computed between generated and ground truth image and averaged across 20 images per transfer. *Initial* represents the metrics between the source image (before transfer) and the target image. **Boldface marks the top model** for each of the 4 transfers studied.

In Table 3, we show the performance of the DDPM frameworks for the same four transfers as those studied in GANs (see Table 2). Different frameworks are compared: one-hot encoding conditioning from Ho and Salimans (2021), classifier-conditioning with $N = 1$ target image selected randomly, inspired from Preechakul et al. (2022), and classifier-conditioning with $N = 10$ target images selected randomly (named $N = 10, \infty$ in the Table).

Using such frameworks, the conversion between pipelines seems more difficult than with the GAN-based methods. While all models succeed in changing the class identified by a pipeline classifier to the target domain, the

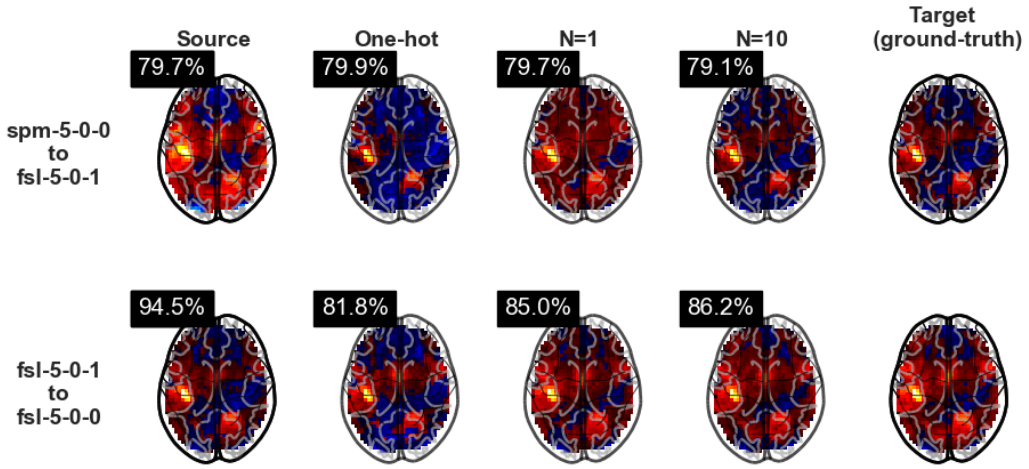

Figure 3: Generated images for two transfer and different competitors: conditioning with one-hot encoding (Ho and Salimans, 2021), with a classifier-conditioning N=1 and N=10 target images with random selection. Correlation with target ground-truth are indicated above generated and source images (in percent).

success of the conversion in terms of similarity to the target ground-truth image is variable across transfers. For instance, all DDPM frameworks succeed in converting statistic maps for the transfer "fsl-1" to "spm-0", while none is successful for the transfer "fsl-1" to "fsl-0". These low performance could be explained by the difficulty of the models to learn differences between close pipelines (*i.e.* whose results display very similar activation patterns). In Table 4, we show the similarity of features extracted from the pipeline classifier for each pair of pipelines explored in this study. In particular, we observe that features learned at Layer 4 (*i.e.* the features used for conditioning) are similar (*i.e.* higher correlation) for pipelines sharing the same software, even with different use of HRF. This proximity of features used for conditioning might explain the difficulty to perform the transfer.

The use of a DDPM with classifier-conditioning and multiple target images ($N = 10, \infty$) improves the performance compared to the alternative DDPM frameworks. Both quality and diversity of images is increased ($IS = 3.86$), and in terms of similarity to the ground-truth target image, this frameworks outperforms the other DDPM models by up to 4% in correlations between target ground-truth and generated image compared to Ho and Salimans (2021) for the transfer "spm-0" to "fsl-1" and up to 3% for "fsl-1" to

Pipelines	Layer 1	Layer 2	Layer 3	Layer 4
Same software, different parameters				
fsl-5-0-0 / fsl-5-0-1	86.5	91.4	95.4	99.2
spm-5-0-0 / spm-5-0-1	86.5	90.9	94.2	98.4
Same parameters, different software				
fsl-5-0-0 / spm-5-0-0	88.8	88.2	93.6	98.2
fsl-5-0-1 / spm-5-0-1	84.8	85.8	92.4	98.0
Different software, different parameters				
fsl-5-0-0 / spm-5-0-1	74.5	81.0	88.7	97.1
fsl-5-0-1 / spm-5-0-0	74.8	77.7	88.2	97.3

Table 4: Mean correlations between features maps learned at each layers for each pair of pipelines

"spm-0".

The first row of Figure 3 illustrates a transfer between pipelines with different software packages and different HRF ("spm-0" to "fsl-1"). The second row shows a transfer between pipelines with the same software package and different HRF ("fsl-1" to "fsl-0"). The DDPM with multiple target images generates statistic maps close to the ground truth for both transfer, representing the intrinsic properties of the map while modifying its extrinsic properties to the target domain. Using the one-hot encoding conditioning, the generated statistic maps seem far from the target image, failing to represent the whole characteristics of the target domain. When using only one target image, statistic maps are more similar to the target in terms of activation area.

The performance of the DDPM frameworks remain notably inferior to the ones obtained with Pix2Pix (Isola et al., 2017) or StarGAN (Choi et al., 2018). This superiority can be explained by the differences between frameworks: GAN methods use adversarial training and StarGAN improves this by using a classifier loss and a cyclic-reconstruction loss. Moreover, the sampling process of GAN relies on the source image directly and do not require to set an initial state, which might facilitate the source content preservation. However, we can note that Inception Scores (IS) obtained with DDPM frameworks are better than the one obtained with StarGAN, which indicates that images generated by DDPM frameworks are more diverse. This observation is consistent with the literature (Dhariwal and Nichol, 2021) and the particular sampling process of DDPM frameworks which includes some

randomness.

3.3. Impact of multi-target images

		fsl-1 → spm-0		spm-0 → fsl-1		fsl-1 → spm-1		fsl-1 → fsl-0	
	IS	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
N=5, ∞	3.89	86.5	0.005	79.1	0.003	82.0	0.005	84.2	0.003
N=10, ∞	3.86	86.5	0.005	79.0	0.003	81.8	0.005	84.3	0.003
N=20, ∞	3.85	86.7	0.005	79.3	0.003	81.5	0.005	84.4	0.003
N=5, Kmeans	3.86	86.4	0.005	78.7	0.003	81.2	0.005	84.5	0.003
N=10, Kmeans	3.86	86.1	0.005	79.0	0.003	81.2	0.005	84.1	0.003
N=20, Kmeans	3.87	86.1	0.005	79.2	0.003	81.3	0.005	83.9	0.003
N=10, KNN	3.75	84.9	0.005	78.7	0.003	81.6	0.005	83.6	0.003

Table 5: Performance associated with four transfers with DDPM frameworks with different implementation. IS means "Inception Score" across all transfers. Pearson's correlation (%) and Mean Squared Error (MSE) computed between generated and ground-truth target image for 20 images per transfer. *Initial* represents the metrics between the source image (before transfer) and the ground-truth target image. ∞ means random sampling.

In Table 5, we show the influence of the number of target images and of the selection methods. The number of images does not seem to impact the performance, correlations are very similar between $N = 5, N = 10 \quad \text{and} \quad N = 20$. Performing the selection using the K-Means algorithm does not seem to improve the performance compared to a random selection, for any N values, probably due to the low diversity in our dataset (*i.e.* participants in groups are sampled from the same study (HCP Young Adults) and there is a small overlap between participants in groups (see Supplementary Figure 1 (Germani et al., 2024b))). However, selection using a K-Nearest Neighbors (KNN) algorithm decreases the performance from 1.6%, meaning that the diversity of the target images is beneficial for a good transfer.

4. Discussion

In this work, we made the assumption that statistic maps could be converted between pipelines to facilitate the re-use of derived data in mega-analyses (Costafreda, 2009). We explored different frameworks based on GAN and DDPM with the aim to develop an unsupervised multi-domain framework that researchers could re-train and use to convert the derived data available in public databases such as NeuroVault (Gorgolewski et al., 2015). Our results are promising, with satisfying performance in converting

statistic maps between pipelines, in particular for pipelines with the most dissimilar results (*e.g.* from different software packages). In these cases, generated statistic maps are much closer to the target image than the original ones, and generated statistic maps are all classified in the target domain by the pipeline classifier. In a follow-up work of Rolland et al. (2022), available as preprint (Germani et al., 2024d), we saw that combining data from different pipelines in mega-analyses leads to invalid results with low to high false positive rates depending on the combination of pipelines. In particular, studies combining data from different software packages are the ones that were leading to the highest false positive rates, and thus, the largest invalidity. The ability to transfer statistic maps between software packages using style transfer frameworks frameworks may therefore greatly help the future of data re-use.

We compared several frameworks and found that, in our case, GAN frameworks always outperformed DDPM in terms of adequacy to the target image. While the largest performance of DDPM was demonstrated in many papers (Dhariwal and Nichol, 2021; Müller-Franzes et al., 2023), we believe that our results are related to the specific properties of style transfer and of fMRI data. The two studies above demonstrated the superiority of DDPM compared to GAN for the task of image synthesis, in both natural and medical images, but not for style transfer. The traditional sampling strategy of DDPM is not suited for such task, as it relies on random noise, which makes it difficult to maintain intrinsic properties of the source images while changing the style. On the contrary, GAN sampling relies on the source images directly and do not require to set an initial state, which might facilitate the source content preservation. In addition, DDPM are trained to minimize an MSE loss between the predicted noise and the actual noise added to the image, without any component related to style transfer, whereas in the GAN frameworks, and in particular StarGAN (Choi et al., 2018), the classifier loss seems to greatly improve performance. Another issue related to DDPM is the high dimensionality of images, here 3-dimensional images with hundreds of thousands of values, which, associated with our small sample sizes, the large number of trainable parameters of the model and the complexity of the learning process, makes it difficult to train efficient models. Recently, the potential of latent diffusion models was shown, these frameworks act in the latent space of a Variational AutoEncoder to reduce the size of data and facilitate training (Rombach et al., 2022). In future work, we would like to experiment such frameworks and compare the results to see if they can

compete with the ones obtained with GAN.

Across GAN frameworks, unsurprisingly, we obtained better performance with the supervised framework (Pix2Pix (Isola et al., 2017)) compared to the unsupervised ones, in particular for conversion between pipelines giving already close results (*e.g.* same software package, different parameters). However, gathering paired data is impractical and far from real life practice. In large databases, for instance NeuroVault (Gorgolewski et al., 2015), we have no information about the pipeline used to obtain statistic maps and potentially no access to raw data to build paired datasets. Our goal is to build a model that could be applied on two or more datasets with different statistic maps of the same task, but obtained with different pipelines. In such unsupervised settings, performance of StarGAN (Choi et al., 2018) is statisfying, the framework succeeds in generating data that are close to the target image for all transfers. For transfer with pipelines giving very distant results, the performance of this framework almost reach the performance of Pix2Pix. We believe that this model could be a good candidate for further development in real-life practice.

However, the practical usability of the proposed frameworks remain questionable. Further work would be needed to assess the potential of these newly transferred statistic maps for statistical studies. In Germani et al. (2024d), we computed false positive rates of mega-analyses combining subject-level data obtained from different pipelines. This method could be applied with between-group analyses composed of 1) a group with data from the target pipeline and 2) a group with data originally obtained with another pipeline and that have been converted to the target pipeline. However, for now, due to the standardization of the data used as input to deep learning models and the architecture of the models, voxel values in generated maps are constrained between -1 and 1. First attempts have been made to de-normalize data using a scale factor derived from the source map. Further work would be needed to deal with the case of maps coming from different software packages. For instance for a transfer from FSL to SPM, differences in percent BOLD change (*i.e.* unit of fMRI contrast maps would have to be taken into account) (Nichols, 2012).

Ideally, to combine their data, researchers should be able to re-use a pretrained style transfer framework to convert statistic maps. However, this pretrained framework may have been trained using statistic maps from specific tasks or from participants with specific characteristics. In preliminary works (see Supplementary Materials (Germani et al., 2024b)), we observed

that frameworks pretrained on statistic maps from a particular task suffer from a large performance drop when applied on statistic maps from another task. In another work (Germani et al., 2024a), we explored the stability of the relationships between pipelines results to evaluate the potential robustness to dataset shifts (*e.g.* different tasks, groups of participants) of our solution. In future work, we would like to explore the potential of techniques such as transfer learning to build more generalizable style transfer frameworks. Transductive transfer learning (Arnold et al., 2007) aims to improve the learning of a target task in a target domain using knowledge from a similar task in a source domain. In particular, in unsupervised transductive transfer learning, there is no labeled data from the target domain. Further experiments would be needed to investigate whether transfer learning frameworks could be helpful to obtain more robust frameworks in our context.

5. Conclusion

In this study, we explored the potential of style transfer frameworks on the task of converting fMRI statistic maps between different pipelines. We showed that the StarGAN framework, trained on unsupervised and multi-domain data, could be easily trained and applied to generate statistic maps that maintain the intrinsic properties of brain activity while changing the style of the image. These could be used to build valid mega-analyses on heterogeneous datasets and hence increase sample sizes in fMRI data analysis.

6. Data statement

This study was performed using derived data from the HCP Young Adult (Van Essen et al., 2013b), publicly available at ConnectomeDB. Data usage requires registration and agreement to the HCP Young Adult Open Access Data Use Terms available at: Van Essen et al. (2013a).

The HCP multi-pipeline dataset (Germani et al., 2023) is in the process of being made publicly available on Public-nEUro (at Rigshopitalet, 2023) (data are uploaded and we are waiting for a link to access data on the platform).

7. Ethics

This study was performed using derived data from the HCP Young Adult (Van Essen et al., 2013b). No experimental activity involving the human participants was made by the authors. Only publicly released data were used.

Written informed consent was obtained from participants and the original study was approved by the Washington University Institutional Review Board.

We agreed to the HCP Young Adult Open Access Data Use Terms available at: Van Essen et al. (2013a).

8. Code and data availability

All the scripts used to perform the study (models training, testing and performance evaluation) are available on Software Heritage: swh:1:snp:b0b52aa88bef8f4411bdd7e00a2d71715d7830bb (Germani et al., 2024c).

Derived data such as pretrained models and computed metrics are available on Zenodo (Germani et al., 2024b).

9. Acknowledgements

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

10. Author contributions

Elodie Germani: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft. **Camille Maumet**: Supervision, Writing - Review & Editing. **Elisa Fromont**: Supervision, Writing - Review & Editing.

11. Funding sources

This work was funded by Region Bretagne (ARED MAPIS) and ANR project ANR-20-THIA-0018). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: Medical image translation using GANs. Computerized Medical Imaging and Graphics 79, 101684. doi:10.1016/j.compmedimag.2019.101684.

Arnold, A., Nallapati, R., Cohen, W.W., 2007. A Comparative Study of Methods for Transductive Transfer Learning, in: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 77–82. doi:10.1109/ICDMW.2007.109.

Bashyam, V.M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C.L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Satterthwaite, T.D., Wolf, D.H., Gur, R.E., Gur, R.C., Morris, J.C., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, N.R., Wittfeld, K., Bülow, R., Wolk, D.A., Shou, H., Nasrallah, I.M., Davatzikos, C., The iSTAGING and PHENOM consortia, 2022. Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. Journal of Magnetic Resonance Imaging 55, 908–916. doi:10.1002/jmri.27908.

Bhagwat, N., Barry, A., Dickie, E.W., Brown, S.T., Devenyi, G.A., Hatano, K., DuPre, E., Dagher, A., Chakravarty, M., Greenwood, C.M.T., Misic, B., Kennedy, D.N., Poline, J.B., 2021. Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. GigaScience 10, giaa155. doi:10.1093/gigascience/giaa155.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., Adcock, R.A., Avesani, P., Baczkowski, B.M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R.G., Berkers, R.M.W.J., Bhanji, J.P., Biswal, B.B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K.L., Bowring, A., Braem, S., Brooks, H.R., Brudner, E.G., Calderon, C.B., Camilleri, J.A., Castrellon, J.J., Cecchetti, L., Cieslik, E.C., Cole, Z.J., Collignon, O., Cox, R.W., Cunningham, W.A., Czoschke, S., Dadi, K., Davis, C.P., Luca, A.D., Delgado, M.R., Demetriou, L., Dennison, J.B., Di, X., Dickie, E.W., Dobryakova, E., Donnat, C.L., Dukart, J., Duncan, N.W., Durnez, J., Eed, A., Eickhoff, S.B., Erhart,

A., Fontanesi, L., Fricke, G.M., Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J.J., Golowin, S.A.E., González-García, C., Gorgolewski, K.J., Grady, C.L., Green, M.A., Guassi Moreira, J.F., Guest, O., Hakimi, S., Hamilton, J.P., Hancock, R., Handjaras, G., Harry, B.B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.P., Huettel, S.A., Hughes, M.E., Iacovella, V., Iordan, A.D., Isager, P.M., Isik, A.I., Jahn, A., Johnson, M.R., Johnstone, T., Joseph, M.J.E., Juliano, A.C., Kable, J.W., Kassinopoulos, M., Koba, C., Kong, X.Z., Koscik, T.R., Kucukboyaci, N.E., Kuhl, B.A., Kupek, S., Laird, A.R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M.Y.C., Lim, P.C., Lintz, E.N., Liphardt, S.W., Losecaat Vermeer, A.B., Love, B.C., Mack, M.L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J.T., Melero, H., Méndez Leal, A.S., Meyer, B., Meyer, K.N., Mihai, G., Mitsis, G.D., Moll, J., Nielson, D.M., Nilsonne, G., Notter, M.P., Olivetti, E., Onicas, A.I., Papale, P., Patil, K.R., Peelle, J.E., Pérez, A., Pischedda, D., Poline, J.B., Prystauka, Y., Ray, S., Reuter-Lorenz, P.A., Reynolds, R.C., Ricciardi, E., Rieck, J.R., Rodriguez-Thompson, A.M., Romyn, A., Salo, T., Samanez-Larkin, G.R., Sanz-Morales, E., Schlichting, M.L., Schultz, D.H., Shen, Q., Sheridan, M.A., Silvers, J.A., Skagerlund, K., Smith, A., Smith, D.V., Sokol-Hessner, P., Steinkamp, S.R., Tashjian, S.M., Thirion, B., Thorp, J.N., Tinghög, G., Tisdall, L., Tompson, S.H., Toro-Serey, C., Torre Tresols, J.J., Tozzi, L., Truong, V., Turella, L., van 't Veer, A.E., Verguts, T., Vettel, J.M., Vijayarajah, S., Vo, K., Wall, M.B., Weeda, W.D., Weis, S., White, D.J., Wisniewski, D., Xifra-Porxas, A., Yearling, E.A., Yoon, S., Yuan, R., Yuen, K.S.L., Zhang, L., Zhang, X., Zosky, J.E., Nichols, T.E., Poldrack, R.A., Schonberg, T., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88. doi:10.1038/s41586-020-2314-9.

Botvinik-Nezer, R., Wager, T.D., 2023. Reproducibility in Neuroimaging Analysis: Challenges and Solutions. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 8, 780–788. doi:10.1016/j.bpsc.2022.12.006.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size

undermines the reliability of neuroscience. Nature Reviews Neuroscience 14, 365–376. doi:10.1038/nrn3475.

Carp, J., 2012. On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. Frontiers in Neuroscience 6. doi:10.3389/fnins.2012.00149.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S., 2021. ILVR: conditioning method for denoising diffusion probabilistic models, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE. pp. 14347–14356.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 8789–8797.

Costafreda, S.G., 2009. Pooling fMRI Data: Meta-Analysis, Mega-Analysis and Multi-Center Studies. Frontiers in Neuroinformatics 3, 33. doi:10.3389/neuro.11.033.2009.

Denck, J., Guehring, J., Maier, A., Rothgang, E., 2021. MR-contrast-aware image-to-image translations with generative adversarial networks. International Journal of Computer Assisted Radiology and Surgery 16, 2069–2078.

Dhariwal, P., Nichol, A.Q., 2021. Diffusion models beat GANs on image synthesis, in: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems.

Dorjsembe, Z., Pao, H.K., Odonchimed, S., Xiao, F., 2024. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. IEEE Journal of Biomedical and Health Informatics , 1–10doi:10.1109/JBHI.2024.3385504.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image Style Transfer Using Convolutional Neural Networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423. doi:10.1109/CVPR.2016.265.

Germani, E., Fromont, E., Maumet, C., 2024a. Uncovering communities of pipelines in the task-fMRI analytical space. URL: https://hal.science/

hal-04331232. accepted at the 2024 IEEE International Conference on Image Processing.

Germani, E., Fromont, E., Maurel, P., Maumet, C., 2023. The hcp multi-pipeline dataset: an opportunity to investigate analytical variability in fmri data analysis. URL: https://arxiv.org/abs/2312.14493, arXiv:2312.14493.

Germani, E., Maumet, C., Fromont, E., 2024b. Mitigating analytical variability in fMRI with style transfer - Supplementary Materials. URL: https://doi.org/10.5281/zenodo.13748563, doi:10.5281/zenodo.13748563.

Germani, E., Maumet, C., Fromont, E., 2024c. Software heritage archive for the github repository "style_transfer_diffusion". URL: https://archive.softwareheritage.org/swh:1:snp:b0b52aa88bef8f4411bdd7e00a2d71715d7830bb;origin=https://github.com/elodiegermani/style-transfer_diffusion.

Germani, E., Rolland, X., Maurel, P., Maumet, C., 2024d. On the validity of fmri studies with subject-level data processed through different pipelines. arXiv:2402.12900.

Glatard, T., Lewis, L.B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.E., Sherif, T., Deelman, E., Khalili-Mahani, N., Evans, A.C., 2015. Reproducibility of neuroimaging analyses across operating systems. Frontiers in Neuroinformatics 9. doi:10.3389/fninf.2015.00012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S.S., Maumet, C., Sochat, V.V., Nichols, T.E., Poldrack, R.A., Poline, J.B., Yarkoni, T., Margulies, D.S., 2015. NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front. Neuroinform. .

Ho, J., Jain, A., Abbeel, P., 2020. Denoising Diffusion Probabilistic Models, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 6840–6851.

Ho, J., Salimans, T., 2021. Classifier-Free Diffusion Guidance, in: "Deep Generative Models and Downstream Applications" Workshop@NeurIPS'21.

Ioannidis, J.P.A., 2005. Why Most Published Research Findings Are False. PLoS Medicine 2, e124. doi:10.1371/journal.pmed.0020124.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62, 782–790.

Jiang, L., Mao, Y., Wang, X., Chen, X., Li, C., 2023. CoLa-Diff: Conditional Latent Diffusion Model for Multi-modal MRI Synthesis, in: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland. pp. 398–408. doi:10.1007/978-3-031-43999-5_38.

Jin, C.B., Kim, H., Liu, M., Jung, W., Joo, S., Park, E., Ahn, Y.S., Han, I.H., Lee, J.I., Cui, X., 2019. Deep CT to MR Synthesis Using Paired and Unpaired Data. Sensors 19. doi:10.3390/s19102361.

Kaji, S., Kida, S., 2019. Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. Radiological Physics and Technology 12, 235–248. doi:10.1007/s12194-019-00520-y.

Kong, L., Lian, C., Huang, D., Li, Z., Hu, Y., Zhou, Q., 2021. Breaking the dilemma of medical image-to-image translation, in: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems.

Liu, M., Breuel, T.M., Kautz, J., 2017. Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems 30 (NIPS), pp. 700–708.

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N., 2021. Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 313–322. doi:10.1007/978-3-030-87199-4_30.

Lyu, Q., Wang, G., 2022. Conversion between ct and mri images using diffusion and score-matching models. arXiv:2209.12104.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. volume 5.1, pp. 281–298.

Markiewicz, C.J., Gorgolewski, K.J., Feingold, F., Blair, R., Halchenko, Y.O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncavles, M., Jwa, A., Poldrack, R., 2021. The OpenNeuro resource for sharing of neuroscience data. eLife 10, e71774.

Mucherino, A., Papajorgji, P.J., Pardalos, P.M., 2009. k-nearest neighbor classification, in: Data Mining in Agriculture. Springer, pp. 83–106.

Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J.N., Truhn, D., 2023. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Scientific Reports 13, 12098. doi:10.1038/s41598-023-39278-0.

Nichols, T., 2012. SPM plot units. URL: https://web.archive.org/web/20230606094719/https://blog.nisox.org/2012/07/31/spm-plot-units.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical Image Synthesis with Deep Convolutional Ad-

versarial Networks. IEEE Transactions on Biomedical Engineering 65, 2720–2730. doi:10.1109/TBME.2018.2814538.

Ozbey, M., Dalmaz, O., Dar, S.U.H., Bedel, H.A., Ozturk, S., Gungor, A., Cukur, T., 2023. Unsupervised Medical Image Translation With Adversarial Diffusion Models. IEEE transactions on medical imaging 42, 3524–3539. doi:10.1109/tmi.2023.3290149.

Pan, S., Chang, C.W., Peng, J., Zhang, J., Qiu, R.L.J., Wang, T., Roper, J., Liu, T., Mao, H., Yang, X., 2023. Cycle-guided Denoising Diffusion Probability Model for 3D Cross-modality MRI Synthesis. arXiv .

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32 , 8024–8035doi:10.48550/arXiv.1912.01703.

Penny, W., Friston, K., Ashburner, J., Kiebel, S., Nichols, T.E., 2011. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Elsevier ed.

Pinho, A.L., Amadon, A., Ruest, T., Fabre, M., Dohmatob, E., Denghien, I., Ginisty, C., Becuwe-Desmidt, S., Roger, S., Laurier, L., Joly-Testault, V., Médiouni-Cloarec, G., Doublé, C., Martins, B., Pinel, P., Eger, E., Varoquaux, G., Pallier, C., Dehaene, S., Hertz-Pannier, L., Thirion, B., 2018. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. Scientific Data 5, 180105. doi:10.1038/sdata.2018.105.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nature Reviews Neuroscience doi:10.1038/nrn.2016.167.

Poline, J.B., Breeze, J.L., Ghosh, S.S., Gorgolewski, K., Halchenko, Y.O., Hanke, M., Helmer, K.G., Marcus, D.S., Poldrack, R.A., Schwartz, Y., Ashburner, J., Kennedy, D.N., 2012. Data sharing in neuroimaging research. Frontiers in Neuroinformatics 6. doi:10.3389/fninf.2012.00009.

Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S., 2022. Diffusion autoencoders: Toward a meaningful and decodable representation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE. pp. 10609–10619.

Qin, Z., Liu, Z., Zhu, P., Ling, W., 2022. Style transfer in conditional GANs for cross-modality synthesis of brain magnetic resonance images. Computers in Biology and Medicine 148, 105928. doi:10.1016/j.compbiomed.2022.105928.

at Rigshopitalet, N.R.U., 2023. Public nEUro. URL: https://public-neuro.github.io/index.html.

Rolland, X., Maurel, P., Maumet, C., 2022. Towards efficient fmri data reuse: can we run between-group analyses with datasets processed differently with spm ?, in: ISBI 2022 - IEEE International Symposium on Biomedical Imaging, pp. 1–4.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 10674–10685.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Sadraee, A., Paulus, M., Ekhtiari, H., 2021. fMRI as an outcome measure in clinical trials: A systematic review in clinicaltrials.gov. Brain and Behavior 11, e02089. doi:10.1002/brb3.2089.

Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M., 2022. Palette: Image-to-image diffusion models, in: SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, ACM. pp. 15:1–15:10.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X., 2016. Improved Techniques for Training GANs, in: Lee, D.,

Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.

Sasaki, H., Willcocks, C.G., Breckon, T.P., 2021. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. arXiv .

Shaban, M.T., Baur, C., Navab, N., Albarqouni, S., 2019. Staingan: Stain Style Transfer for Digital Histological Images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 953–956. doi:10.1109/ISBI.2019.8759152.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine 12.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., 2013a. Human connectome project: Data usage agreement. https://www.humanconnectome.org/study/hcp-young-adult/document/wu-minn-hcp-consortium-open-access-data-use-terms.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., 2013b. The WU-Minn Human Connectome Project: An overview. NeuroImage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I., 2017a. Deep MR to CT Synthesis Using Unpaired Data, in: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham. pp. 14–23. doi:10.1007/978-3-319-68127-6_2.

Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2017b. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. IEEE Transactions on Medical Imaging 36, 2536–2545. doi:10.1109/TMI.2017.2708987.

Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E.I.C., Xu, Y., 2020. MRI Cross-Modality Image-to-Image Translation. Scientific Reports 10, 3753. doi:10.1038/s41598-020-60520-6.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G., 2018. Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. IEEE Transactions on Medical Imaging 37, 1348–1357. doi:10.1109/TMI.2018.2827462.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, ICCV, IEEE Computer Society. pp. 2242–2251.