Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Inf Process Med Imaging. Author manuscript; available in PMC 2022 February 15.

Published in final edited form as:
Inf Process Med Imaging. 2021 June ; 12729: 321–333. doi:10.1007/978-3-030-78191-0_25.

Representation Disentanglement for Multi-modal Brain MRI Analysis

Jiahong Ouyang[1], **Ehsan Adeli**[1], **Kilian M Pohl**[1,2], **Qingyu Zhao**[1], **Greg Zaharchuk**[1]

[1]Stanford University, Stanford, CA

[2]SRI International, Menlo Park, CA

Abstract

Multi-modal MRIs are widely used in neuroimaging applications since different MR sequences provide complementary information about brain structures. Recent works have suggested that multi-modal deep learning analysis can benefit from explicitly disentangling anatomical (shape) and modality (appearance) information into separate image presentations. In this work, we challenge mainstream strategies by showing that they do not naturally lead to representation disentanglement both in theory and in practice. To address this issue, we propose a margin loss that regularizes the similarity in relationships of the representations across subjects and modalities. To enable robust training, we further use a conditional convolution to design a single model for encoding images of all modalities. Lastly, we propose a fusion function to combine the disentangled anatomical representations as a set of modality-invariant features for downstream tasks. We evaluate the proposed method on three multi-modal neuroimaging datasets. Experiments show that our proposed method can achieve superior disentangled representations compared to existing disentanglement strategies. Results also indicate that the fused anatomical representation has potential in the downstream task of zero-dose PET reconstruction and brain tumor segmentation.

1 Introduction

Multi-modal MRIs using different pulse sequences (e.g., T1-weighted and T2 Fluid Attenuated Inversion Recovery) are widely used to probe complementary and mutually informative aspects of the brain structure, thereby playing a pivotal role in improving the understanding of neurodevelopment across the life span and diagnosis of neuropsychiatric disorders [15]. However, compared to uni-modal image analysis, models that operate on multi-modal data are more likely to encounter the issue of incomplete inputs (some cases have missing modalities) due to data corruption, when applied to larger MRI datasets [7].

To tackle these challenges, recent works [15,17] have suggested to explicitly disentangle anatomical and modality-specific information from multi-modal MRIs. Specifically, each image is encoded into two representations: an *anatomical representation* that encodes the morphological shape of brain anatomies and is mostly shared across all modalities

jiahongo@stanford.edu .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of the same subject, and a *modality representation* that encodes image appearance information specific to the modality. Such disentanglement is typically derived based on *cross-reconstruction* [8], i.e., by examining the quality of images synthesized from anatomical and modality representations from mixed sources. The resulting disentangled representations are shown to be useful for downstream tasks including cross-modality deformable registration [14], multi-modal segmentation [17], image harmonization [4], multi-domain image synthesis, and imputation of missing modalities [15].

All the above studies focused on evaluating the results of the downstream tasks. It remains *unclear* whether the learned representations are truly disentangled or not. In this work, we show that the cross-reconstruction strategies can easily lead to information leakage between representations, i.e., representations are still partly coupled after disentanglement. To address this issue, we propose a margin loss that regularizes the within-subject across-modality similarity between representations with respect to the across-subject within-modality similarity. Such regularization encourages the anatomical and modality information to fully disentangle in the representation space. Further, to obtain a robust training scheme, we use a modified conditional convolution to combine separate encoders associated with the modalities into a single coherent model. Lastly, we introduce a fusion function to combine the disentangled anatomical representations as a set of modality-invariant features, which can be used to solve various downstream tasks. We evaluate our method on three multi-modal neuroimaging datasets, including T1- and T2-weighted MRIs of 692 adolescents from the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) [19], T1-weighted and T2-FLAIR MRIs of 173 adults for zero-dose PET reconstruction, and multi-modal MRIs (T1, post-contrast T1, T2, and T2 Fluid Attenuated Inversion Recovery) from 369 subjects of the BraTS 2020 dataset [10]. Results indicate that our method achieves better disentanglement between anatomical and modality representations compared to several baseline methods. The fused modality-invariant representation shows potential in the downstream task of PET reconstruction and brain tumor segmentation (BraTS).

2 Related Works

Representation disentanglement is an active topic in image-to-image translation tasks lately [8]. The goal of these tasks is to disentangle the content (e.g., anatomical information) and style (e.g., modality, texture, appearance) information from an image so that images of the same content can be translated between different styles. The disentanglement is learned by optimizing a cross reconstruction loss on synthesized images, with content and style sampled from different training images [6,2]. A major issue is that these methods do not explicitly enforce the disentanglement, and hence the learned representations still suffer from information leakage.

Based on this observation, methods based on adversarial training [8,9,3,1] further regularize the content representations to be independent of the source style domain. For example, DRIT [8] couples adversarial training with a cross-cycle consistency loss to achieve between-domain translation based on unpaired data. MTAN [9] uses a multi-class adversarial loss for the style labels. DRNet [3] leverages the adversarial loss to disentangle the stationary and temporal components. Sagie et al. [1] proposed a zero loss to force the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

style encoder to capture information relevant to the specific domain. However, adversarial training can be unstable and easily stuck in a local optimum. In addition, only one of the two representations (usually the content representation) can be adversarially regularized, which can easily cause information leakage into the other representation. As an extreme scenario, the style representation can contain all the information about the input, while the content representation may carry little or random information, just enough to fool the discriminator.

Lastly, a common issue of the above methods is that they utilize a separate decoder for each domain. It means that regardless of being disentangled or not, the learned content representations can always produce satisfying results for domain translation or downstream tasks as long as they carry task-related information from the input. In conclusion, in the absence of visualization or evaluation of the disentangled representations, it is unclear if the representations learned by the above methods are still coupled or not.

3 Proposed Method

To address this ambiguity, we first introduce a robust model for disentangling the anatomical and modality representations for multi-modal MRI based on image-to-image translation in Section 3.1. Next, we introduce a strategy for fusing the disentangled anatomical representations from all available modalities of a subject into a modality-invariant representation, which can be used as the input for any downstream model.

3.1 Representation disentanglement by image-to-image translation

We assume each subject in the training set has MRIs of m modalities (sequences) and let $x_i \in \mathcal{X}_i$ denote the input image of the i-th modality. As shown in Fig. 1a, we aim to disentangle x_i into an anatomical representation s_i by an anatomical encoder $s_i = E_i^A(x_i)$ and a modality representation z_i by a modality encoder $z_i = E_i^M(x_i)$. We assume that the anatomical representation s_i encodes the morphological information of brain structures that is mostly impartial to the imaging modality, while z_i provides image appearance information specific to a modality. The decoder D then reconstructs x_i from a pair of anatomical and modality representations. Prior works [2] have suggested that such disentangled representations can be learned by optimizing the *self-reconstruction* and *cross-reconstruction* losses; Given a pair of s_i and z_j derived from images of any two modalities, D is supposed to synthesize an image that is similar to the input image $x_j \in \mathcal{X}_j$, whose synthesized domain corresponds to the j-th modality.

$$L_{self-recon} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{E}_{x_i \sim \mathcal{X}_i} [\| \widetilde{x_{ii}} - x_i \|_1],$$

$$L_{cross-recon} = \frac{\lambda_c}{m^2 - m} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} \mathbf{E}_{x_i \sim \mathcal{X}_i, x_j \sim \mathcal{X}_j} \left[\| \widetilde{x_{ij}} - x_j \|_1 \right],$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\widetilde{x_{ij}} = D\left(E_i^A(x_i), E_j^M(x_j)\right)$. In addition to these reconstruction losses, another loss function commonly used for training image-to-image translation [14] is the *latent consistency* loss, which encourages the representations derived from raw inputs to be similar to the ones from the synthesized images.

$$L_{latent} = \frac{\lambda_l}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{E}_{x_i \sim \mathcal{X}_i, x_j \sim \mathcal{X}_j} \left[\| \widetilde{z_{ji}} - z_i \|_1 \right], \tag{1}$$

where $\widetilde{z_{ji}} = E_i^M\left(D\left(E_j^A(x_j), E_i^M(x_i)\right)\right)$ is the modality representation derived from a synthesized image.

Enforcing Disentanglement by Similarity Regularization.—Although prior works have leveraged the above concept of disentanglement for several multi-modal downstream tasks, there is no theoretical guarantee that the cross reconstruction can encourage the encoder to disentangle anatomical and modality representations. In fact, information can freely leak between representations. As a naive example, both s_i and z_i can be an exact copy of x_i so that the decoder D can easily reconstruct the input.

We resolve this problem by exploring the similarity relationships between representations. As the brain's morphological shape is highly heterogeneous across subjects, we assume the anatomical representations s from the same subject but different modalities should be more similar than those from the same modality but different subjects. Note, s_i of the same subject are not necessary to be exactly the same, as multi-modal imaging is designed to capture distinct characteristics of brain anatomies. For instance, the brain tumor itself is more visible on T1-weighted MR with contrast (T1c) compared to T1 without contrast due to the injected contrast medium. On the other hand, modality representations z from the same modality but different subjects should be more similar than those from the same subject but different modalities. We propose to model such relationships using a similarity loss inspired by the margin-based hinge loss [5].

$$\begin{aligned} L_{sim} = & \frac{\lambda_s}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{E}\left[\max\left(0, \alpha_s - \cos\left(f(s_i^p), f\left(s_j^p\right)\right) + \cos\left(f(s_i^p), f(s_i^q)\right)\right)\right] + \\ & \frac{\lambda_z}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \mathbf{E}\left[\max\left(0, \alpha_z - \cos(z_i^p, z_i^q) + \cos\left(z_i^p, z_j^p\right)\right)\right] \end{aligned} \tag{2}$$

where p and q correspond to a pair of subjects randomly sampled in a minibatch, $\cos(\cdot,\cdot)$ denotes the cosine distance between two vectors, and f denotes a MaxPooling and flattening operation. Unlike the L_2-based similarity loss, Eq. (2) encourages the within-subject and across-subject distances to differ by the margins α_s and α_z and thereby avoids deriving identical representations.

Conditional convolution.—Another drawback of traditional multi-modal image translation methods is that each modality is associated with a pair of anatomical and modality encoders that are independently learned. However, these encoding tasks are highly dependent across modalities. Hence, each convolutional operation at a certain layer should

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

function similarly across networks. To enable robust multi-modal translation, we couple the set of independent anatomical encoders $\left\{E_i^A(x_i)\right\}$ into one coherent encoder model $E^A(x;i)$ and likewise couple all modality encoders $\left\{E_i^M(x_i)\right\}$ into $E^M(x;i)$ with i being an additional input to the model. We construct these two unified models using Conditional Convolution (CondConv) [16] as the fundamental building blocks. As inspired by [16], parameters of a convolutional kernel is conditioned on the input modality i using a mixture-of-experts model $CondConv(x;i) = \sigma\left(\left(\beta_1^i \cdot W_1 + \ldots + \beta_n^i \cdot W_n\right) \circledast x\right)$, where $\sigma(\cdot)$ is the sigmoid activation function, $\circledast$ denotes regular convolution, $\{W_1,\ldots, W_n\}$ are the learnable kernels associated with n experts, and $\left\{\beta_1^i, \ldots \beta_n^i\right\}$ are the modality-specific mixture weights. As such, the convolutional kernel exhibits correlated behavior across modalities as the n experts are jointly trained on data of all modalities.

3.2 Fusing disentangled representations for downstream tasks

As shown in Fig. 1b, after obtaining the disentangled representations, the anatomical representations from all available modalities of a subject are fused into one fixed-size encoding as the input for a downstream model T, which can be any state-of-the-art model for the downstream task. Note that the fusion function here should pool features of a various number of channels to a fixed number of channels. Let s be the concatenation of anatomical representations from the available modalities $Concat(s_i,\ldots,s_j)$. Then the fusion is the concatenation of several pooling functions: $Concat(MaxPool(s),MeanPool(s),MinPool(s))$. With this fusion operation, one can use two strategies to train the downstream model T. We can either solely train T based on the frozen s derived by the self-supervised learning of the encoders (Section 3.1), or fine-tune the encoders jointly with the downstream task model T. Though the joint training can potentially result in representations that better suit the downstream task, we confine our analysis to the first strategy (fixing encoders) in this work to emphasize the impact of representation disentanglement.

4 Experiments

We first describe the dataset and the experimental settings in Section 4.1 and 4.2. We then show in Section 4.3 that our proposed approach (Section 3.1) can effectively disentangle anatomical and modality representations on three neuroimaging datasets. We further show in Section 4.4 that the disentangled representations in combination with the fusion strategy (Section 3.2) can alleviate the missing modality problem in two downstream tasks.

4.1 Datasets

ZeroDose—The dataset comprised brain FDG-PET and two MR modalities (T1-weighted and T2 FLAIR) from 171 subjects with multiple diagnosis types including tumor, epilepsy, and dementia. The FLAIR and PET images were first registered to T1, and then all modalities were normalized to a standard template and resized to 192×160 in the axial plane. Intensities in the brain region of each image were converted to z-scores. The top and bottom 20 slices in each image were omitted from analysis. Each 3 adjacent axial slices were converted to a 3-channel image as the input to the encoder models. Random flipping of brain hemispheres was used as augmentation during training. Five-fold cross-validation was

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

conducted with 10% training cases used for validation. The downstream task was zero-dose PET reconstruction, i.e., to synthesize high quality FDG-PET from multi-modal MRIs. This is useful in practice as the injected radiotracer in current PET imaging protocols can lead to the risk of primary or secondary cancer in scanned subjects. Moreover, PET is more expensive than MRI and not offered in the majority of medical centers worldwide.

NCANDA—Based on the public data release[1], we used the T1 and T2 MRIs of 692 adolescents with no-to-low alcohol drinking from the NCANDA dataset [19]. The images were preprocessed by a pipeline [18] composed of denoising, bias field correction, skull stripping, aligning T1 and T2 to a template, and resizing to 192 × 160. Other settings were the same as the ZeroDose dataset. As this dataset only contained healthy individuals, we used it solely for evaluating the representation disentanglement based on the middle 40 slices in each image. **BraTS** Multimodal Brain Tumor Segmentation Challenge 2020 [10] provides multi-modal brain MRI of 369 subjects with four modalities: T1, post-contrast T1 (T1Gd), T2, and T2-FLAIR (FLAIR). Three categories were labeled for brain tumor segmentation, i.e., Gd-enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET). We used the 55 middle axial slices and cropped the image size to 192×160. Other preprocessing steps and settings were kept the same.

4.2 Experimental Settings

Implementation Details—The anatomical encoder E^A was a U-Net type model. Let C_k denote a Convolution-BatchNorm-ReLU block with k filters (4 × 4 spatial filters with stride 2), and CD_k an Upsample-Convolution-BatchNorm-ReLU block. The architecture was designed as C_{32}-C_{64}-C_{128}-C_{256}-C_{256}-CD_{256}-CD_{128}-CD_{64}-CD_{32}. A convolution then mapped the resulting representations to 4 channels with softmax activation as the anatomical representation. The modality encoder E^M consisted of 5 convolution layers of 3 × 3 filters and stride 2 with LeakyReLU of a 0.2 slope. Numbers of filters were 16-32-64-128-128. A fully connected layer mapped the resulting features to a 16-D representation. The decoder D was based on SPADE [13] with the architecture used in [2]. The networks were trained for 50 epochs by the Adam optimizer with learning rate of 2×10^{-4} and weight decay of 10^{-5}. The regularization rates were set to $\lambda_c = 2.0$, $\lambda_l = 0.1$, $\lambda_s = 10.0$, $\lambda_z = 2.0$. The margins in the similarity loss were set to $\alpha_s = \alpha_z = 0.1$. For the downstream model T in the zero-dose PET reconstruction, a U-Net based model with attention modules [12] was adopted. The downstream model for BraTS brain tumor segmentation was the BraTS 2018 challenge's winner NVNet [11].

Competing Methods—We first implemented the encoders using traditional convolution (training separate encoders), denoted as **Conv**. Based on this implementation, our disentanglement approach incorporating the similarity losses is denoted as **+Sim**. We then compared our approach with two types of methods. We term the first type [6,2] that regularized the disentanglement merely using cross-reconstruction and latent consistency loss as **+NA**. The other type [8,9,3,1,14,15,17] that utilized adversarial training on

[1]NCANDA_PUBLIC_4Y_STRUCTURAL_V01 (DOI: 10.7303/syn22216457); collection was supported by NIH grants AA021697, AA021695, AA021692, AA021696, AA021681, AA021690, and AA02169

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the anatomical representations are termed as **+Adv**. To make fair comparison on the disentanglement strategies, all comparison methods used the same network structure for the encoder and decoder. Finally, we replaced **Conv** with **CondConv** (training a single encoder) to show the advantage of using conditional convolution.

4.3 Evaluation on disentangled representation

We first evaluated the methods on representation disentanglement and image cross reconstruction based on 5-fold cross-validation. We derived the anatomical and modality representations of the test subjects learned by the approaches. Fig. 2 visualizes the 4-channel anatomical representations of one subject from NCANDA. We observe that s_1 and s_2 (extracted from T1 and T2) learned by the two baselines (Conv+NA and Conv+Adv) were *substantially different*, indicating that they might still contain modality-specific information. On the other hand, our approach (Conv+Sim) produced visually more similar anatomical representations than the baselines. This indicates that the proposed similarity regularization can decouple the modality-specific appearance features from the structural information shared between modalities.

This result is also supported by the visualization of the learned representation spaces. As shown in Fig. 3a–c, we randomly selected 200 modality representations in the test set of the BraTS dataset and projected them into a 2D space by t-SNE. Only our approach clearly separated the representations of different modalities into 4 distinct clusters (Fig. 3c), which was in line with the regularization on z in Eq. (2). The clustering with respect to modalities was not evident for the projections of the baseline approaches (Fig. 3a,b), indicating that complimentary information had leaked into the modalities representations. Moreover, the baseline approaches failed to disentangle T1 and T1Gd, two contrasts with high visual resemblance, as the red and blue dots were coupled in the representation space. Likewise, we visualized the space of anatomical representations in Fig. 3d–f. We randomly selected 4 subjects in the BraTS test set and projected the pooled anatomical representation $f(s_i)$ of 4 consecutive slices into a 2D space. Now the 4 distinct clusters of our approach were defined with respect to subjects as opposed to modalities, and there was no apparent bias of modality in each subject's cluster (Fig. 3f), indicating the representations solely encoded subject-specific yet modality-invariant information. The representation spaces learned by the two baselines contained both subject-specific anatomical and modality information (Fig. 3d,e); that is, although the projections could be separated by subjects (black circles), each subject-specific cluster could be further stratified by modalities (color of the markers).

The improved disentanglement also resulted in better cross-reconstruction. Fig. 4 shows the results of a test subject from the BraTS dataset. In each panel, $\widetilde{x_{ij}}$ is displayed on the jth row and i^{th} column; Diagonal images correspond to self-reconstruction and off-diagonal ones are cross-reconstruction. The proposed Conv+Sim achieved the best visual quality (accurate structural details), especially the FLAIR reconstruction highlighted in red boxes, where the tumor area was more precisely reconstructed. This improvement was quantitatively supported by the higher similarity between ground-truth and synthesized images in terms of peak-signal-noise ratio (PSNR) and structural similarity index (SSIM) (Table 1). For each image of a specific modality, we synthesized it based on its own modality representation

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and the anatomical representation from another modality (For each image in the BraTS dataset, we computed the average metrics over all three cross reconstruction results). According to Table 1, Conv+NA achieved the lowest reconstruction quality for both SSIM and PSNR on all three datasets. The quality improved when adding adversarial training on the anatomical representations (Conv+Adv) but was still lower than the two implementations with the proposed similarity loss (except for T1 reconstruction in NCANDA). Of the two models with +Sim, CondConv+Sim recorded better performance on the ZeroDose and NCANDA datasets. This indicates that CondConv enabled more stable reconstruction results by coupling highly dependent modality-specific encoders. However, on the BraTS dataset, Conv+Sim achieved better cross-reconstruction for T2 and FLAIR. The reason could be that CondConv shared anatomical and modality encoders across modalities at the expense of model capacity, especially when more than two modalities were involved. It is also worth mentioning that all methods achieved higher performance on NCANDA because it was the only dataset of healthy controls. Taken all together, only our approach resulted in true disentanglement between anatomical and modality representations, which was not guaranteed by the baselines.

4.4 Evaluation on downstream tasks

Deep learning models that rely on multi-modal input often suffer from the missing input problem. When a model is trained on data with complete inputs and tested on data with one modality missing, standard approaches either fill in all zero values (**Standard+Zero**) or use the average image of that modality over the entire cohort (**Standard+Avg**). Here, we demonstrate that an alternative solution is to train and test the model on the fusion of disentangled anatomical representations from all available modalities (**Ours**, CondConv+Sim). We show that this strategy can largely alleviate the impact of missing inputs.

In each run of the cross-validation, we first learned the disentangled representations and then trained the downstream models (based on the raw multi-modal images for the standard methods or fused anatomical representations for our proposed method) for zero-dose PET reconstruction. Then, the downstream model was tested on data with or without missing modalities. When all modalities were used for testing, the standard and proposed approaches achieved comparable reconstruction accuracy in terms of both PSNR and SSIM (N/A in Table 2), but we observe that our approach could generally result in more accurate tumor reconstruction (Fig. 5), which could not be reflected in the whole-brain similarity measure. The improvement of our approach became evident when one of the modalities was missing (Table 2). In particular, missing FLAIR induced larger performance drop for all approaches, as ZeroDose contained a large number of images with tumor, which was more pronounced in FLAIR than T1.

Next, we replicated this experiment for the downstream task of brain tumor segmentation on the BraTS dataset and measured the performance using the dice coefficient (DICE; Each cell in Table 2 records the average DICE across three categories: ET, ED, and NCR/NET). In line with the results of the ZeroDose experiment, the standard and proposed methods both obtained similar DICE scores, when complete inputs were used during testing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Standard+Zero recorded the lowest accuracy when missing any modality. Standard+Avg was less impacted when T1 or T2 was missing, but more impacted by T1Gd and FLAIR as the standard model relied mostly on those two modalities for localizing the tumor. The proposed method achieved the highest DICE score in all scenarios, among which the missing T2 recorded the largest drop on DICE. This might be because T2 had the most distinct appearance compared to other modalities, thus having the largest impact on the fused representation.

5 Conclusion

In this paper, we first proposed a novel margin loss to regularize the within-subject across-modality similarity between representations with respect to the across-subject within-modality similarity. It alleviates the information leakage problem in existing disentanglement methods. We further introduced a modified conditional convolution layer to enable training a single model for multiple modalities. Lastly, we proposed a fusion function to combine the disentangled anatomical representations from available modalities as a set of modality-invariant features for downstream tasks. Experiments on three brain MR datasets and two downstream tasks demonstrated that the proposed method achieved meaningful and robust disentangled representations compared with the existing methods. Though we only evaluated on brain images, the method is likely to generalize to other organs as long as the assumption on the within-subject across-modality and across-subject within-modality similarity holds.

Acknowledgement:

This work was supported by NIH funding AA021697 and by the Stanford HAI AWS Cloud Credit.

References

1. Benaim S, Khaitov M, Galanti T, Wolf L: Domain intersection and domaindifference. In: ICCV. pp. 3445–3453 (2019)
2. Chartsias A, Papanastasiou G, Wang C, Semple S, Newby D, Dharmakumar R, Tsaftaris SA: Disentangle, align and fuse for multimodal and zero-shot image segmentation. IEEE Transactions on Medical Imaging (2020)
3. Denton EL, et al. : Unsupervised learning of disentangled representations fromvideo. NeurIPS 30, 4414–4423 (2017)
4. Dewey BE, Zuo L, Carass A, He Y, Liu Y, Mowry EM, Newsome S,Oh J, Calabresi PA, Prince JL: A disentangled latent space for cross-site mri harmonization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 720–729. Springer (2020)
5. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T: Devise: A deep visual-semantic embedding model. In: NeurIPS. pp. 2121–2129 (2013)
6. Huang X, Liu MY, Belongie S, Kautz J: Multimodal unsupervised image-to-image translation. In: ECCV. pp. 172–189 (2018)
7. Lee D, Kim J, Moon WJ, Ye JC: Collagan: Collaborative gan for missing image data imputation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2487–2496 (2019)
8. Lee HY, Tseng HY, Huang JB, Singh M, Yang MH: Diverse image-to-image translation via disentangled representations. In: ECCV. pp. 35–51 (2018)
9. Liu Y, Wang Z, Jin H, Wassell I: Multi-task adversarial network for disentangled feature learning. In: CVPR. pp. 3743–3751 (2018)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

10. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J,Burren Y, Porz N, Slotboom J, Wiest R, et al. : The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34(10), 1993–2024 (2014) [PubMed: 25494501]
11. Myronenko A: 3d mri brain tumor segmentation using autoencoder regularization.In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018)
12. Ouyang J, Chen K, Zaharchuk G: Zero-dose pet reconstruction with missinginput by u-net with attention modules. Medical Imaging Meets NeurIPS (2020)
13. Park T, Liu MY, Wang TC, Zhu JY: Semantic image synthesis withspatially-adaptive normalization. In: CVPR. pp. 2337–2346 (2019)
14. Qin C, Shi B, Liao R, Mansi T, Rueckert D, Kamen A: Unsuperviseddeformable registration for multi-modal images via disentangled representations. In: IPMI. pp. 249–261. Springer (2019)
15. Shen L, Zhu W, Wang X, Xing L, Pauly JM, Turkbey B, Harmon SA,Sanford TH, Mehralivand S, Choyke P, et al. : Multi-domain image completion for random missing input data. arXiv preprint arXiv:2007.05534 (2020)
16. Yang B, Bender G, Le QV, Ngiam J: Condconv: Conditionally parameterizedconvolutions for efficient inference. In: NeurIPS. pp. 1307–1318 (2019)
17. Yang J, Li X, Pak D, Dvornek NC, Chapiro J, Lin M, Duncan JS: Cross-modality segmentation by self-supervised semantic alignment in disentangled content space. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, pp. 52–61. Springer (2020)
18. Zhao Q, Adeli E, Pfefferbaum A, Sullivan EV, Pohl KM: Confounder-awarevisualization of ConvNets. In: International Workshop on Machine Learning in Medical Imaging. LNCS, vol. 11861, pp. 328–336. Springer (2019)
19. Zhao Q, Sullivan EV, Honnorat N, Adeli E, Podhajsky S, De Bellis MD,Voyvodic J, Nooner KB, Baker FC, Colrain IM, et al. : Association of heavy drinking with deviant fiber tract development in frontal brain systems in adolescents. JAMA psychiatry (2020)

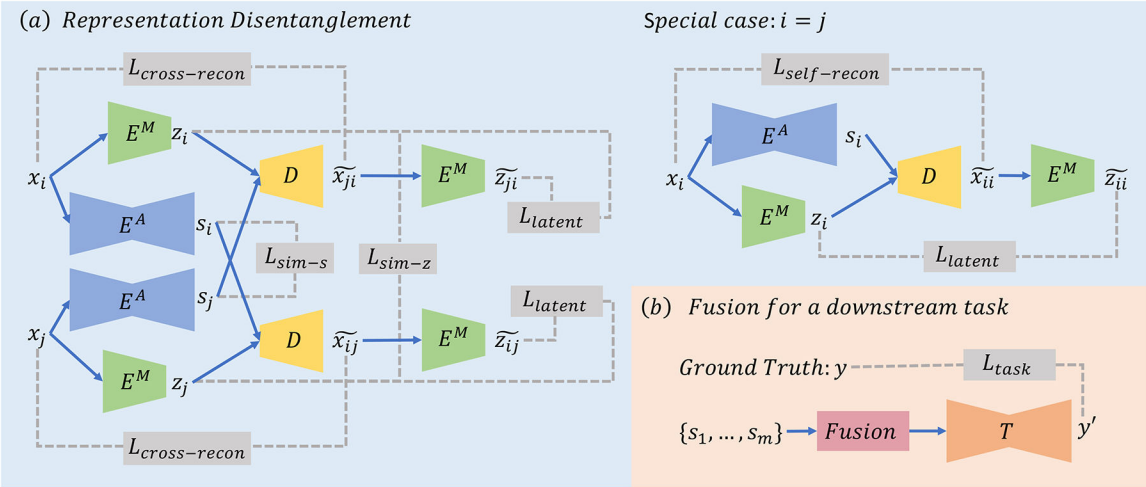

Fig.1:

Overview: (a) An image x_i is disentangled into an anatomical representation s_i and a modality representation z_i by E_A and E_M. The decoder D reconstructs the input from the two representations. These networks are trained by the reconstruction and latent consistency losses. We propose to add a similarity regularization L_{sim} that models the relationships between the representations from different images; (b) The disentangled anatomical representations of a subject are fused into one modality-invariant encoding that can be used as an input to a downstream model T.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

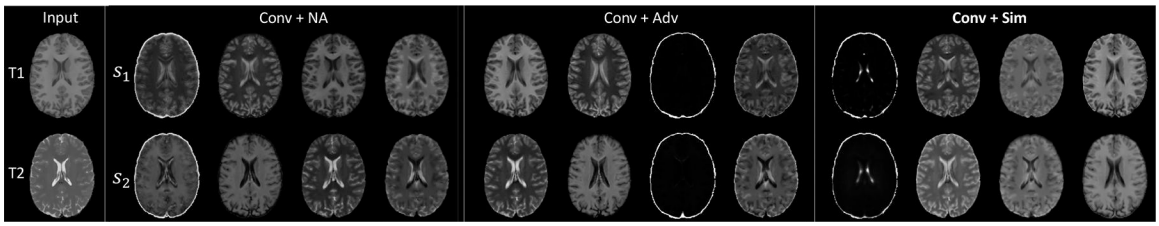

Fig.2:

Visualization of s_1 and s_2 of one NCANDA subject. Only our approach (+Sim) resulted in visually similar anatomical representations from T1 and T2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

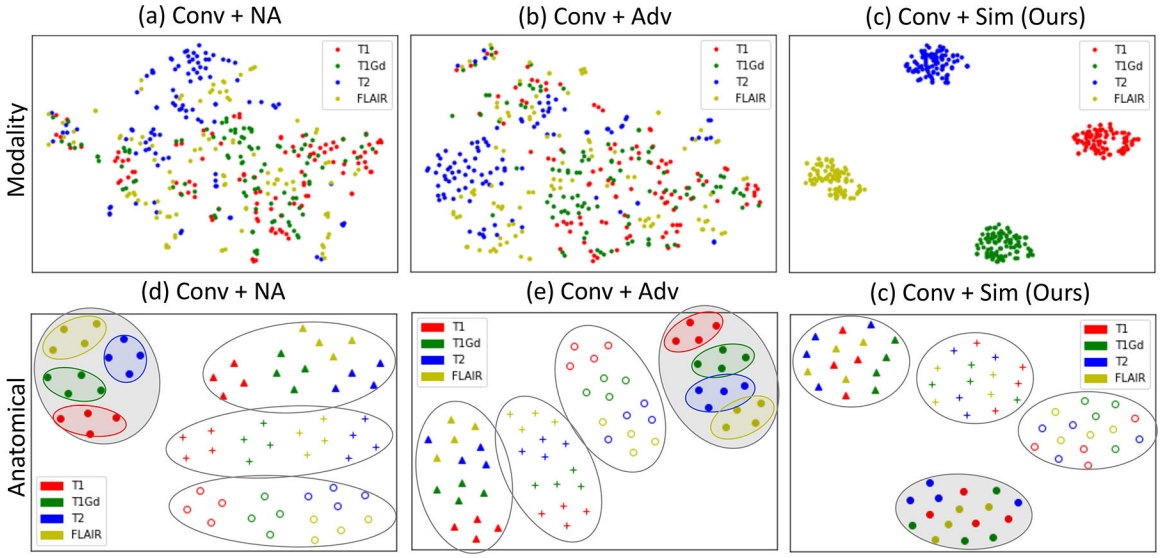

Fig. 3:

t-SNE visualization of the z space (a,b,c) and the s space (d,e,f) for BraTS dataset in 2D spaces. Fully disentangled z should cluster by modality (denoted by color); Fully disentangled s should cluster by subjects (denoted by marker style) with no modality bias (sub-clusters by modality).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

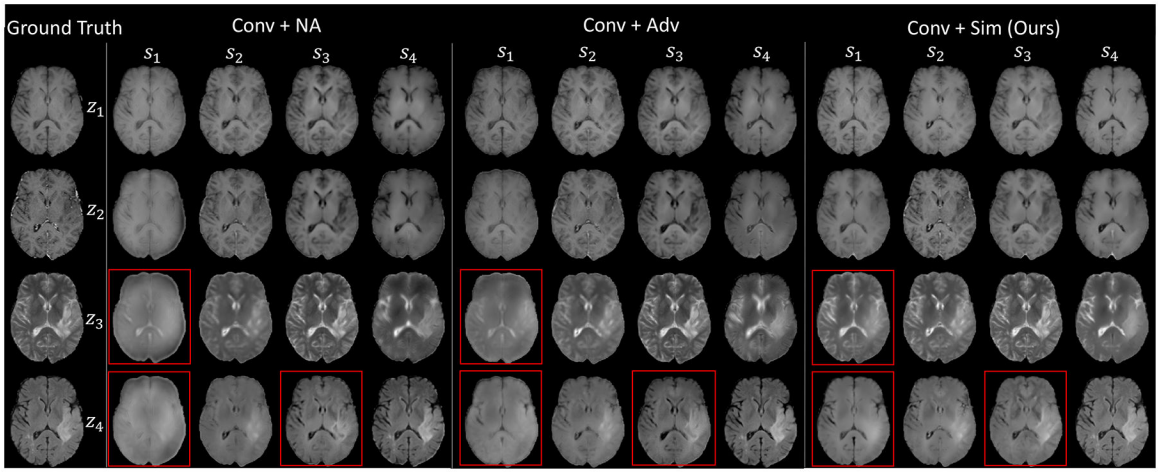

Fig.4:

Cross reconstruction results for one test subject from the BraTS dataset.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

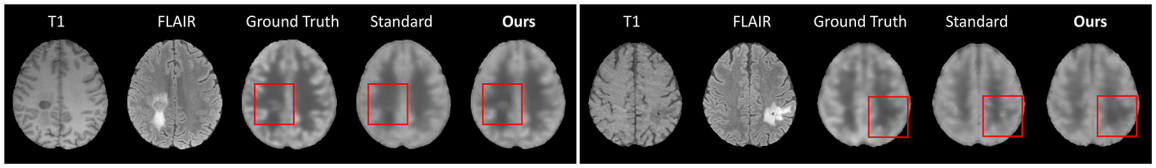

Fig.5:
ZeroDose PET reconstruction from T1 and FLAIR for two test subjects.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1:

5-fold cross-validation for quantitative cross-reconstruction evaluation.

(a) ZeroDose		
Methods	**T1**	**FLAIR**
	PSNR/SSIM	**PSNR/SSIM**
Conv+NA	25.603/0.682	24.435/0.612
Conv+Adv	27.131/**0.742**	25.846/0.674
Conv+Sim(Ours)	27.222/0.735	25.970/0.667
CondConv+Sim (Ours)	**27.231/0.742**	**25.978/0.681**

(b) NCANDA		
Methods	**T1**	**T2**
	PSNR/SSIM	**PSNR/SSIM**
Conv+NA	29.719/0.849	28.077/0.808
Conv+Adv	**30.421/0.866**	27.950/0.807
Conv+Sim(Ours)	30.266/0.863	28.367/0.825
CondConv+Sim (Ours)	30.331/0.865	**28.451/0.832**

(c) BraTS				
Methods	**T1**	**TIGd**	**T2**	**FLAIR**
	PSNR/SSIM	**PSNR/SSIM**	**PSNR/SSIM**	**PSNR/SSIM**
Conv+NA	27.304/0.717	24.897/0.638	25.148/0.621	25.166/0.617
Conv+Adv	27.485/0.717	25.385/0.656	25.951/0.658	26.135/0.642
Conv+Sim(Ours)	27.892/**0.756**	26.114/0.723	**26.479/0.744**	**26.588/0.692**
CondConv+Sim(Ours)	**27.916**/0.752	**26.221/0.731**	26.445/0.735	26.489/0.687

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Performance of two downstream tasks with incomplete input.

Methods	ZeroDose			BraTS				
	N/A	T1	FLAIR	N/A	T1	T1Gd	T2	FLAIR
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	DICE				
Standard+Zero	**25.475/0.739**	18.122/0.436	18.8863/0.435	**0.826**	0.364	0.240	0.616	0.298
Standard+Avg		24.425/0.676	23.137/0.631		0.724	0.279	0.733	0.452
Ours	25.386/0.729	**24.610/0.682**	**24.193/0.674**	0.821	**0.782**	**0.779**	**0.758**	**0.772**

Left: zero-dose PET reconstruction; Right: brain tumor segmentation.