BRAIN IMAGE SYNTHESIS USING INCOMPLETE MULTIMODAL DATA

Yanfu Zhang[1], *Guodong Liu*[5], *Runxue Bao*[2], *Liang Zhan*[3], *Paul Thompson*[4], *Heng Huang*[5]

[1]Department of Computer Science, William and Mary, Williamsburg, USA
[2]GE HealthCare, Chicago, USA
[3]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, USA
[4]Institute for Neuroimaging and Informatics, University of Southern California, Marina del Rey, USA
[5]Department of Computer Science, University of Maryland, College Park, USA

ABSTRACT

Multimodal medical image synthesis is an important task. Previous efforts mainly focus on the task domain of medical image synthesis using the complete source data and have achieved great success. However, data collection with completeness in real life might be prohibitive due to high expenses or other difficulties, particularly in brain imaging studies. In this paper, we address the challenging and important problem of medical image synthesis from incomplete multimodal data sources. We propose to learn the modal-wise representations and synthesize the targets accordingly. Particularly, a surrogate sampler is derived to generate the target representations from incomplete observations, based on which an interpretable attention-redistribution network is designed. The experimental results synthesizing PET images from MRI images demonstrate that the proposed method can solve different missing data scenarios and outperforms related baselines consistently.

***Index Terms*—** Multimodal, Neuroimaging, MRI, PET, Generative Network

1. INTRODUCTION

Brain imaging techniques have been widely used for the diagnosis of complex brain disorders, such as Alzheimer's disease (AD) [1, 2]. Different imaging techniques have been developed, including T1-weighted structural magnetic resonance imaging (T1), T2∗-weighted MRI (T2), T1-weighted-Fluid-Attenuated Inversion Recovery (Flair), perfusion-weighted imaging (Perfusion), and positron emission tomography (PET). Recently, the diagnosis of brain diseases (e.g. AD) benefits from integrating complementary information from multimodal brain imaging data [3–5]. However, in practice, the multimodal benefits usually cannot be fully exploited for brain disease diagnosis due to the missing data [6, 7]. The reason for data incompleteness is three-fold. First, the scanning time for each subject may be limited due to concerns about the total duration of radioactive exposure and the cost of scanning. Second, the data acquisition is dependent on the subjects' availability and the progress of the subjects' status. For example, during different visits, a subject may have different brain status, which makes the integration of multimodal data difficult even when they are available. And third, although there are many efforts to provide multimodal brain imaging data for research purposes, the data inconsistency is still inevitable due to various data collection sites and devices. For example, Alzheimer's Disease Neuroimaging Initiative (ADNI) [8], a multi-site longitudinal study toward AD, has been running since 2004, and PET data are not available for more than half of early participants. Different scanning standards also make the problem extremely challenging.

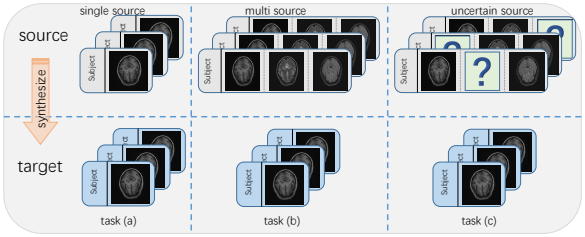

Fig. 1: Prior works (task (a) and (b)) synthesize target modality from known modalities. We consider the challenging heterogeneous image synthesis from incomplete multimodal data (task (c)), where each subject may have different missing data at various modalities, denoted by question marks.

To address the problem of data incompleteness, conventional methods either discard those subjects with missing data modality or focus on imputing hand-crafted features extracted from original data. Recently it has been shown that synthesizing missing modality is potentially beneficial for the diagnosis of AD [9, 10]. However, previous works typically presume source modalities are completely available, including generative adversarial networks [11–15] and variational auto-encoders [16–19]. In real applications, it is often not the case. Particularly, the modal incompleteness for different subjects is typically heterogeneous. Therefore, previous methods are not instantly applicable to this problem. Figure 1 illustrates the comparison of previous tasks and ours. In this paper, we address the challenging problem of synthesizing brain images

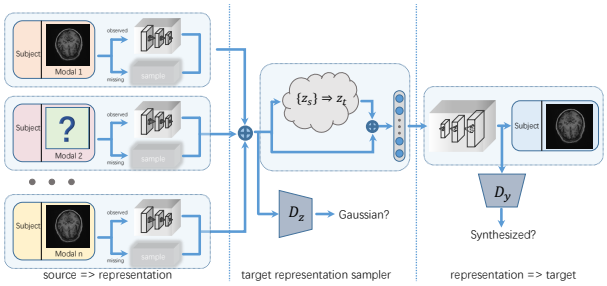

Fig. 2: The pipeline for synthesizing the target modality from incomplete multimodal source modality.

from data with randomly missing modalities. We propose a new unified model to generate target modality from source modalities with random missing patterns, and the experimental results demonstrate the proposed method outperforms related approaches.

2. METHODOLOGY

Notations. We use $\boldsymbol{x}_j^i$ and $\boldsymbol{y}_j$ to denote the i^{th} source domain and the target domain respectively, for the j^{th} subject. $\mathcal{X}_{s_j}$ denotes the set of all source domain, and $\mathcal{X}_{\mathring{s}_j} \subseteq \mathcal{X}_{s_j}$ denotes the available domains. s_j and $\mathring{s}_j$ are all modality and available modality for the j^{th} subject, respectively. $\boldsymbol{z}_j^i$ and $\boldsymbol{z}_t^i$ are the representation of $\boldsymbol{x}_j^i$ and $\boldsymbol{y}_j$. $\boldsymbol{\mathcal{Z}}_{s_j}$, $\boldsymbol{\mathcal{Z}}_{\mathring{s}_j}$ and $\boldsymbol{\mathcal{Z}}_t$ are the sets.

Problem Formulation. Provided with incomplete data $\mathcal{X}_{\mathring{s}}$, our task is to generate $\boldsymbol{y}_j$ with high fidelity. To this end, we need to learn the distribution $P(\boldsymbol{y}_j, \mathcal{X}_{\mathring{s}})$, with which the generation task is to compute $\operatorname{argmin}_{\boldsymbol{y}} P_\theta(\boldsymbol{y}|\mathcal{X}_{\mathring{s}})$. In our task, for each individual, some data modalities might be missing randomly. Although the data irregularity is frequently happening in real-world applications, standard methods including CycleGAN [12] typically fail to handle it since data completeness are explicitly required. To address this problem, we propose a new machine learning based multimodal medical image synthesis method with three key components: (1) the representations $\boldsymbol{\mathcal{Z}}_{\mathring{s}}$ are learned from the available source modalities individually; (2) we estimate the target modality representation $\boldsymbol{z}_t$ by integrating observed sources; (3) the synthesized results are generated from $\boldsymbol{z}_t$. The detailed methodology architecture is illustrated in Fig. 2.

Let $\boldsymbol{z_t}$ and $\boldsymbol{\mathcal{Z}}_{\mathring{\boldsymbol{s}}}$ are the representations of target and source domain respectively, the distribution $P(\boldsymbol{y}_j, \mathcal{X}_{\mathring{s}}) = \mathbb{E}_{\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}}}[P_\theta(\boldsymbol{y}_j, \mathcal{X}_{\mathring{s}}, \boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}})]$. $\mathbb{E}_{\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}}}$ is taking expectation over $\{\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{\boldsymbol{s}}}\}$. Let $\boldsymbol{\mathcal{Z}}_s \setminus \boldsymbol{\mathcal{Z}}_{\mathring{s}}$ be the missing modalities, we have $(\boldsymbol{y}_j, \mathcal{X}_{\mathring{s}}, \boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}}) = P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}}) P(\boldsymbol{y}|\boldsymbol{z}_t) \prod P(\boldsymbol{x}^i|\boldsymbol{z}_s^i) = \mathbb{E}_{\boldsymbol{\mathcal{Z}}_s \setminus \boldsymbol{\mathcal{Z}}_{\mathring{s}}}[P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s)] P(\boldsymbol{y}|\boldsymbol{z}_t) \prod P(\boldsymbol{x}^i|\boldsymbol{z}_s^i)$. $\prod P(\boldsymbol{x}^i|\boldsymbol{z}_s^i)$ is the term for the representation learning from source, namely step (1). $P(\boldsymbol{y}|\boldsymbol{z}_t)$ is the term describing the synthesizing, namely step (3). Many established methods [16, 17] can be exploited in the modelling of representation learning and synthesizing, and data can be represented with lower dimension and simpler structure. We are interested in step (2), the estimation of $\mathbb{E}_{\boldsymbol{\mathcal{Z}}_s \setminus \boldsymbol{\mathcal{Z}}_{\mathring{s}}}[P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s)]$.

The key in solving step (2) is to estimate a joint distribution of complete modality $P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s)$, based on which the incomplete case $P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}})$ can be computed as marginal distribution. With $P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_{\mathring{s}})$, $\boldsymbol{z}_t$ can be inferred straightforwardly according to maximum likelihood. We presume $P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s)$ is a zero-mean Gaussian distribution, with a block-wise co-variance matrix $\boldsymbol{\Lambda_z}$, where each block $\boldsymbol{\Lambda}_{\boldsymbol{z}_1, \boldsymbol{z}_2}$ represents the co-variance between modality $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. To learn $P(\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s)$, an intuitive option is to estimate its co-variance matrix. However, $\{\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s\}$ is typically of high dimension, making the estimation difficult and the computation expensive. Inspired by the reparameterization trick in the variational auto-encoder, we propose a surrogate sample method. For complete data $\{\boldsymbol{z}_t, \boldsymbol{\mathcal{Z}}_s\}$, we consider a sampler,

$$\boldsymbol{z}_t = \sum_{i \in \mathcal{S}} \boldsymbol{W}^i \left(\boldsymbol{z}_s^i + \boldsymbol{A}^i \boldsymbol{\sigma}^i\right) \sim P(\boldsymbol{z}_t | \boldsymbol{\mathcal{Z}}_s), \tag{1}$$

here $\boldsymbol{\sigma}^i$ is a modality-wise random perturbation in the representation, $\boldsymbol{W}^i$ and $\boldsymbol{A}^i$ are parameters related to the conditional distribution $P(\boldsymbol{z}_t|\boldsymbol{\mathcal{Z}}_s)$. For simplicity we assume the diagonals of $\boldsymbol{\Lambda_z}$ are identity matrices. It is easy to verify training $\boldsymbol{W}^i$ is equivalent to evaluating co-variance matrix $\boldsymbol{\Lambda}_{\boldsymbol{z}^i, \boldsymbol{z}^t}$, because mean of $\boldsymbol{z}_t|\boldsymbol{\mathcal{Z}}_s$ is $\boldsymbol{\mu}_{\boldsymbol{z}_t|\boldsymbol{\mathcal{Z}}_s} = \sum_{j \in \mathcal{S}} \boldsymbol{\Lambda}_{\boldsymbol{z}_{\mathring{s}}^i, \boldsymbol{z}_t} \boldsymbol{z}_{\mathring{s}}^i$. Since the sum of random variables needs to be adjusted to be consistent with the co-variance of the surrogate sampler (1), $\boldsymbol{A}^i$ is a scaling matrix $\boldsymbol{A}^i = \left(\boldsymbol{W}^i\right)^{-1} (\mathrm{I} - \boldsymbol{\Lambda}_{st})^{1/2} \boldsymbol{\Lambda}_{st}^{-1/2} \boldsymbol{\Lambda}^i$, where I is identity matrix, $\boldsymbol{\Lambda}^i$ is the co-variance of introduced perturbation, and $\boldsymbol{\Lambda}_{st} = \sum_{i \in \mathcal{S}} \boldsymbol{\Lambda}_{\boldsymbol{z}_{\mathring{s}}^i, \boldsymbol{z}^t} \boldsymbol{\Lambda}_{\boldsymbol{z}_{\mathring{s}}^i, \boldsymbol{z}^t}^{\mathsf{T}}$. One can verify the choice of $\boldsymbol{A}^i$ by computing the co-variance of $P(\boldsymbol{z}_t|\boldsymbol{\mathcal{Z}}_s)$, *i.e.*, $\boldsymbol{\Lambda}_{\boldsymbol{z}_t|\boldsymbol{\mathcal{Z}}_s} = \mathrm{I} - \boldsymbol{\Lambda}_{st} = \mathbb{E}\left(\left(\sum_{i \in \mathcal{S}} \boldsymbol{A}^i \boldsymbol{\sigma}^i\right)\left(\sum_{i \in \mathcal{S}} \boldsymbol{A}^i \boldsymbol{\sigma}^i\right)^{\mathsf{T}}\right)$. Regarding incomplete modalities, we can first compute marginal distribution $P_\theta\left(\boldsymbol{z}_t|\boldsymbol{\mathcal{Z}}_{\mathring{s}_j}\right)$, then derive a similar sampler,

$$\boldsymbol{z_t} = \sum_{i \in \mathring{\mathcal{S}}} \boldsymbol{W}^i \left(\boldsymbol{z}_s^i + \tilde{\boldsymbol{A}}^i \boldsymbol{\sigma}^i\right) + \sum_{j \in \mathring{\mathcal{S}} \setminus \mathcal{S}} \boldsymbol{W}^j \tilde{\boldsymbol{A}}^j \boldsymbol{\sigma}^j. \tag{2}$$

Note that $\boldsymbol{W}^i$ and $\boldsymbol{A}^i$ are determined by individual data. (2) potentially uses different $\tilde{\boldsymbol{A}}^i$ compared to (1), which can be interpreted as follow: the sample in (1) is evenly dependent on all modalities, but in (2), the sample is heavily based on observed modalities. In other words, the dependence on the missing modality is transferred to the present modality. In the next part, we describe a deep network to model this transfer, which is end-to-end trainable, and computationally efficient.

Integrating Incomplete Multimodal Data via Attention-Redistribution. Representation learning has been well studied, and we utilize state-of-the-art auto-encoder in the encoding of source modalities and reconstruction of target modality. In this part, we are mainly interested in designing a feasible sampling scheme *w.r.t* (2), a representation transformer from the incomplete source domain to the target domain.

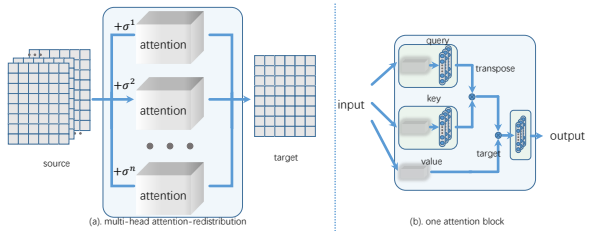

Fig. 3: The structure of attention-redistribution block.

According to to (2), $\boldsymbol{W}^i$ can be formulated using a one-layer network, and additionally, we need to estimate the mean and perturbation for each modality and re-scale the perturbation accordingly. The perturbation $\boldsymbol{\sigma}^i$ can be computed via a neural network, similarly in variational auto-encoder. Let all representations are m dimension vectors, $\tilde{\boldsymbol{A}}^i \in \mathbb{R}^{m\times m}$ is a large matrix which is prohibitive in computation [20]. To simply the computation, for each modality we compute a scalar score $\tilde{\boldsymbol{a}}^i$ instead of the full matrix, indicating the importance of the modality. Specifically, we first reparameterize the modality-wise representation as $\tilde{\boldsymbol{z}}_s^i = \boldsymbol{z}_s^i + \epsilon\boldsymbol{\sigma}^i$, here ϵ is a small random number. And for missing modalities, we let $\boldsymbol{z}_s^i = \boldsymbol{0}$ and $\boldsymbol{\sigma}^i = \boldsymbol{I}$. Then, we concatenate all source representations $\tilde{\boldsymbol{Z}}_k = [\tilde{\boldsymbol{z}}_s^1, \tilde{\boldsymbol{z}}_s^2, \dots]$. For source modality i, let the query vector be computed from the representations $\boldsymbol{q}_i = f_q(\boldsymbol{z}_s^i)$, and compute the key vectors $\boldsymbol{V}_i = f_k\tilde{\boldsymbol{Z}}_k$. Here, both f_q and f_k are one layer neural network, and the input and the ourput share the same size. The importance score $\tilde{\boldsymbol{a}}^i$ can be computed as $\max(\boldsymbol{q}_i\boldsymbol{V}_i^\intercal)$. At last, we have the transformed representation $\boldsymbol{z_t} = \sum_{i\in\mathcal{S}} \boldsymbol{W}^i\left(\tilde{\boldsymbol{a}}^i\tilde{\boldsymbol{z}}_s^i\right)$.

Our definition of keys and queries are similar to the attention mechanism [21–23]. Similar to the self-attention mechanism [24, 25], we can introduce multi-head attention into the importance score computation, and take the average as the final score. Specifically, the attention score here is interpreted as the correlation between the source domain and the target domain. And the correlation is estimated through computing the available domains. Ideally, if all modalities are available, the generation is dependent on all modalities. However, if there is a missing modality, the attention will be distributed to the available ones. That is the reason that we name the proposed structure *attention-redistribution*.

Optimizing Objective. We design three loss terms for the proposed method and optimize these objectives sequentially. *Reconstruction Loss*: to improve the representation ability, the auto-encoder units are trained using $\mathcal{L}_r = \sum \|f_m(\boldsymbol{x}_m) - \boldsymbol{x}_m\|_2$, here f_m is the auto-encoder for each modality, and m is the observed modality. *Sample Loss*: to keep the consistency of $\boldsymbol{z}_t$ and $\tilde{\boldsymbol{z}}_t$, we exploit the term $\mathcal{L}_s = \log P(\boldsymbol{z}|\mathcal{X}_{j_s}) \propto \|\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_t\|_2^2/\sigma_t^2$. *Discriminator Loss*: the representation for all modals are presumed to be Gaussian, thus we introduce a per-modality $\mathcal{L}_d = \mathbb{D}_{JS}(\boldsymbol{z}_m, \boldsymbol{z}), \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \lambda\boldsymbol{I})$, here $\mathbb{D}_{JS}$ is Jenson-Shannon divergence, $\boldsymbol{z}$ is a random variable, $\mathcal{N}(\boldsymbol{0}, \lambda\boldsymbol{I})$ is Gaussian distribution with $\boldsymbol{0}$ mean and $\lambda\boldsymbol{I}$ variance.

3. EXPERIMENTS

Data Description. Totally 264 subjects from ADNI were included in this study. The subjects are selected to ensure the available brain scanning is within 180 days for consistency across different modalities. T1, T2, Perfusion, and Flair-weighted MRI data were used to predict Pet data. Among them, 205 subjects have complete data, 264 T1, 235 T2, 244 Flair, 264 Perfusion, and 252 Pet. All imaging data were prepossessed using standard techniques [26] and registered to the ICBM common space of T1. 20% of complete data are chosen as the test set and the rest are used for training.

Experimental Setting. We use 2-D slices from each modality to train the model. We use ResNet as the encoder and a deconvolution network as the decoder with embedding size 512. For the attention-redistribution block, we use eight head and 1-D convolution for the query, key, and value computation. The model is trained for 200 epochs using Adam with learning rate 0.001 and batch size 128.

3.1. Comparison with Baselines

The main purpose of the comparison is to demonstrate that the proposed method attains competent performances on the difficult incomplete generation, compared to simple deterministic generation using previous methods. We implemented CycleGAN and VAEGAN as baselines. Both methods cannot solve the random missing problem. As such, for baselines, we simplify the task into fixed modal generation. The evaluation of the proposed is based on the more challenging task of an incomplete generation. We consider two scenarios for training. In the first, the proposed model is trained on subjects with complete source modalities. And in the second, the subjects in training have missing source modalities. Five testing scenarios are considered for testing to cover different missing severity, including using all source modalities, discarding fixed modalities, and randomly discarding modalities. The results are summarized in Table 1.

Regarding the proposed method, the performances of Sce.0.a and Sce.0.b are similar, indicating that the model performance is not affected by incomplete data, instead, it improves by using all available data. The results under different missing severity show the proposed method can benefit from more observed data, meanwhile yield competent results for high-level incompleteness. From baselines, it can be also observed that different source modalities have distinct prediction abilities for the target modality. For example, Pet synthesized from Perfusion is relatively worse than other sources particularly for VAEGAN, which is potentially caused by the fuzzy structure of perfusion scanning. Meanwhile, T1, T2, and Flair can yield decent predictions. In a multi-source situation, in general, the prediction is improved, for all methods. Nevertheless, the proposed method can achieve the highest performance compared to baselines. The model trained on all

Table 1: For the proposed modality, Sce.0.a: trained and tested both on complete data; Sce.0.b: trained on full data, tested on complete data; Sce.1.a: trained on full data, tested on discarded T1; Sce.1.b: trained on full data, tested on discarded T1 and perfusion; Sce.2.a: trained on full data, tested on random discard one modality; Sce.2.b: trained on full data, tested on random discard two modalities.

model	metrics	SSIM	PSNR	MSE
CycleGAN	T1	0.6960 ± 0.0789	21.5113 ± 1.0688	0.0091 ± 0.0036
	T2	0.6543 ± 0.1197	21.1860 ± 0.9028	0.0078 ± 0.0016
	Flair	0.6968 ± 0.0777	22.2292 ± 0.9390	0.0061 ± 0.0013
	Perfusion	0.3412 ± 0.2507	16.5147 ± 6.1676	0.2057 ± 0.2603
	T1 + Flair	0.7261 ± 0.0202	22.1737 ± 1.8330	0.0067 ± 0.0031
	T2 + Perfusion	0.6465 ± 0.1473	20.0400 ± 1.7807	0.0107 ± 0.0041
	All	0.7395 ± 0.0238	22.3542 ± 1.7640	0.0052 ± 0.0013
VAEGAN	T1	0.6123 ± 0.1136	21.7730 ± 1.9694	0.0073 ± 0.0032
	T2	0.6812 ± 0.0780	21.1647 ± 0.9858	0.0078 ± 0.0018
	Flair	0.6603 ± 0.1528	22.2167 ± 1.4206	0.0064 ± 0.0021
	Perfusion	0.4359 ± 0.1943	21.9777 ± 0.8825	0.0163 ± 0.0034
	T1 + Flair	0.7072 ± 0.0880	21.8273 ± 0.9465	0.0067 ± 0.0015
	T2 + Perfusion	0.5580 ± 0.1417	22.2876 ± 0.8006	0.0096 ± 0.0018
	All	0.6780 ± 0.1050	22.0836 ± 1.4341	0.0066 ± 0.0022
Proposed	Sce.0.a	0.8001 ± 0.0300	23.4038 ± 2.2272	0.0053 ± 0.0021
	Sce.0.b	$\mathbf{0.8004 \pm 0.0292}$	$\mathbf{24.0629 \pm 1.9586}$	0.0049 ± 0.0027
	Sce.1.a	0.7855 ± 0.0274	22.6806 ± 1.7987	0.0062 ± 0.0035
	Sce.1.b	0.7540 ± 0.0247	22.8832 ± 1.6017	0.0059 ± 0.0019
	Sce.2.a	0.7716 ± 0.0268	22.8814 ± 1.6906	$\mathbf{0.0048 \pm 0.0018}$
	Sce.2.b	0.7691 ± 0.0267	22.5484 ± 1.9522	0.0063 ± 0.0016

Table 2: Results of variants of the proposed structure. Time is in $second/epoch$. The variants include: n-*head*, refers to transformers with different number of headers; *shared*, refers to using one auto-encoder for all modalities; *joint*, refers to jointly training of transformer and auto-encoder.

Variant	SSIM	PSNR	MSE	Time
1-head	0.7794	22.8970	0.0074	284
4-head	0.7915	23.5318	0.0068	328
16-head	0.7965	23.9324	0.0049	511
shared	0.7880	22.2601	0.0078	156
joint	0.7961	22.6811	0.0054	397

data is also slightly better than the model on complete data. The inference from incomplete data also demonstrates decent results. For example, one can compare the results of Sce.2.a of the proposed method and T1+FLAIR in CycleGAN. To sum up, the proposed method outperforms the related baselines with similar observed data and is flexible to different scenarios by avoiding training additional models. Given that here are $\frac{s(s-1)}{2}$ combinations of missing patterns with s source modalities, the proposed method has better scalability, as the generation based on available source modalities is feasible within a single model.

Ablation Study We studied the number of heads in the attention-redistribution block. From Table 2, we found that increasing the number of heads can improve the performances, though may consume more computation and memory resources. We choose eight heads as a balance of computation burden and performances. Through the experiments, we also found that the model size can be reduced at the cost of moderate performance degeneration. At last, sequentially training transformer and auto-encoder can slightly improve the performance, compared to joint training. Empirically, the sequential training of auto-encoder, transformer, and discriminator improves the performances. Besides, if we jointly optimize the reconstruction loss and the transformer loss, additional hyper-parameters shall be introduced to balance the weighted summation. Instead, the hyper-parameter is not necessary for the sequential training, which simplifies the tuning of the model.

4. CONCLUSION

In this paper, we studied the challenging problem of medical image synthesis from incomplete multimodal data sources. We proposed a surrogate sampler method to infer the target representation from incomplete source representations and designed a multi-head attention-redistribution network for efficient computation. We conducted extensive experiments on synthesizing PET images from MRI images, and the results demonstrated that the proposed new method consistently outperforms the related approaches in various settings.

5. ACKNOWLEDGEMENTS

This study was partially supported by NIH (R01AG071243, R01MH125928, and U01AG068057) and NSF (IIS 2045848, IIS 2348159, and IIS 2319450).

References

[1] Haoteng Tang et al., "A comprehensive survey of complex brain network representation," *Meta-Radiology*, p. 100046, 2023.

[2] Haoteng Tang et al., "Signed graph representation learning for functional-to-structural brain network mapping," *Medical image analysis*, vol. 83, pp. 102674, 2023.

[3] Vince D Calhoun and Jing Sui, "Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness," *Biological psychiatry: cognitive neuroscience and neuroimaging*, vol. 1, no. 3, pp. 230–244, 2016.

[4] Kai Ye et al., "Bidirectional mapping with contrastive learning on multimodal neuroimaging data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 138–148.

[5] Haoteng Tang et al., "Contrastive brain network learning via hierarchical signed graph pooling model," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] Mingxia Liu et al., "View-aligned hypergraph learning for alzheimer's disease diagnosis with incomplete multi-modality data," *Medical image analysis*, vol. 36, pp. 123–134, 2017.

[7] Reza Shirkavand, Liang Zhan, Heng Huang, Li Shen, and Paul M Thompson, "Incomplete multimodal learning for complex brain disorders prediction," *arXiv preprint arXiv:2305.16222*, 2023.

[8] Clifford R Jack Jr et al., "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[9] Yongsheng Pan et al., "Disease-image specific generative adversarial network for brain disease diagnosis with incomplete multi-modal neuroimages," in *MICCAI*. Springer, 2019, pp. 137–145.

[10] Yan Jin et al., "Brain mri to pet synthesis and amyloid estimation in alzheimer's disease via 3d multimodal contrastive gan," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2023, pp. 94–103.

[11] Phillip Isola et al., "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[12] Jun-Yan Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[13] Dongwook Lee et al., "Collagan: Collaborative gan for missing image data imputation," in *CVPR*, 2019, pp. 2487–2496.

[14] Xi Chen et al., "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NeurIPS*, 2016, pp. 2172–2180.

[15] Hongwei Li et al., "Diamondgan: Unified multi-modal generative adversarial networks for mri sequences synthesis," in *MICCAI*. Springer, 2019, pp. 795–803.

[16] Diederik P Kingma et al., "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] Irina Higgins et al., "beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR*, vol. 2, no. 5, pp. 6, 2017.

[18] Alireza Makhzani et al., "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[19] Zhiting Hu et al., "On unifying deep generative models," *arXiv preprint arXiv:1706.00550*, 2017.

[20] Han Zhang et al., "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[21] Dzmitry Bahdanau et al., "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] Ashish Vaswani et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] Jianpeng Cheng et al., "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[24] Xiaolong Wang et al., "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[25] Tao Xu et al., "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018, pp. 1316–1324.

[26] Julian Maclaren et al., "Prospective motion correction in brain imaging: a review," *Magnetic resonance in medicine*, vol. 69, no. 3, pp. 621–636, 2013.