HyperMM : Robust Multimodal Learning with Missing Imaging Modalities

Anonymous

Anonymous Organization
@.**

Abstract. Combining multiple imaging modalities carrying complementary information through multimodal learning (MML) has shown considerable benefits for diagnosing multiple pathologies. However, the robustness of multimodal models to missing modalities is often overlooked. Most works assume modality completeness in the input data, while in clinical practice, it is common to have incomplete modalities. Existing solutions that address this issue rely on modality imputation strategies before using supervised learning models. These strategies, however, are complex, computationally costly and can strongly impact subsequent prediction models. Hence, they should be used with parsimony in sensitive applications such as healthcare. We propose an end-to-end framework designed for supervised MML with missing imaging modalities without using imputation before training. We introduce a novel strategy for training an *universal* feature extractor using a conditional hypernetwork, and propose a permutation invariant neural network that can handle inputs of varying dimensions to process the extracted features, in a two-phase *task-agnostic* framework. We experimentally demonstrate the advantages of our method on two tasks: Alzheimer's disease detection and breast cancer classification. We demonstrate that our strategy is robust to high rates of missing data and that its flexibility allows it to handle varying-sized datasets beyond the scenario of missing modalities. We make all our code and experiments available at `https://link/hidden/during/review/process`.

Keywords: Multimodal learning · Missing modalities · Classification

1 Introduction

Multimodal imaging techniques are widely used both in clinical practice and medical research. Simultaneous acquisition and analysis of multiple imaging modalities, such as Emission Tomography (PET), Computed Tomography (CT), or Magnetic Resonance Imaging (MRI), has shown to be beneficial in the diagnosis of Alzheimer's disease [20], or detection of cancers [21], among others. Accordingly, deep learning methods designed to learn from multimodal medical images have seen rapid growth [3, 14]. This development has been favored by the emergence of multimodal learning (MML), a field of machine learning combining modalities from various sources that depict a single subject from multiple views, thus providing both shared and complementary information. MML has shown considerable advantages in multiple domains [1, 25]. However, most current models assume completeness of the training and testing data, which is rare for real-world datasets. In particular, in routine clinical practice obtaining several imaging modalities for the same subject is not a standard. Having varying numbers of modalities per patient is common, which results in incomplete multimodal datasets where one or more modalities can be missing. This makes MML challenging as it prevents the straightforward use of the existing methods.

Related work. Existing solutions to address missing modalities mostly consist of first learning a generative model on a complete dataset, and using it to impute missing modalities before learning a discriminative model [7, 2, 8, 18, 27]. This approach has considerable limitations in practice. First, an unreasonable number of samples may be needed for training a good missing modality imputation model. Second, the difficulty of the prediction model strongly depends on the choice of imputation model. Both networks need to be adapted to one another [9, 11], which can be difficult to ensure in practice. Some studies [19, 23] address this limitation by focusing on learning jointly the modality imputation and prediction tasks, but these models rely on complex and computationally costly training strategies. Lastly, poorly imputed data can compromise the interpretability and feature importance of subsequent predictors [15], which is a crucial aspect to consider in applications such as healthcare where it can lead to incorrect conclusions about the impact of a feature on the outcome.

Contributions. In this work, we address the practical limitations of existing methods by proposing an end-to-end imputation-free strategy for multimodal supervised learning with missing imaging modalities. Building from conditional hypernetworks [4], we formulate a novel strategy for training an *universal* modality-agnostic feature extractor using a single large pre-trained network. We then reformulate the problem of predicting multimodal observations with missing modalities as one of predicting *sets* of observations of varying size, thus relaxing the requirement of fixed-dimensional data inputs of most machine learning models. We implement this approach through a permutation-invariant neural network [26], which eliminates the need to impute missing modalities as done by previous works [7, 2, 18, 27]. By combining these elements in a two-step training framework, we formulate HyperMM, a novel *task* and *model agnostic* strategy for MML from incomplete datasets, without the use of imputation.

2 Methodology

We consider a dataset $\mathcal{D}$ of $n \in \mathbb{N}$ independent input and output pairs such that $\mathcal{D} := \{(X_1, Y_2), \ldots, (X_n, Y_n)\}$, and for which the goal is to predict Y given X. Each $X := \{x_1, \ldots, x_d\}$ corresponds to a d-modal observation, where each x_i represents one of the available modalities. Let us now introduce the indicator vector $v \in \{0, 1\}^d$ to denote the positions of missing modalities in X, such that

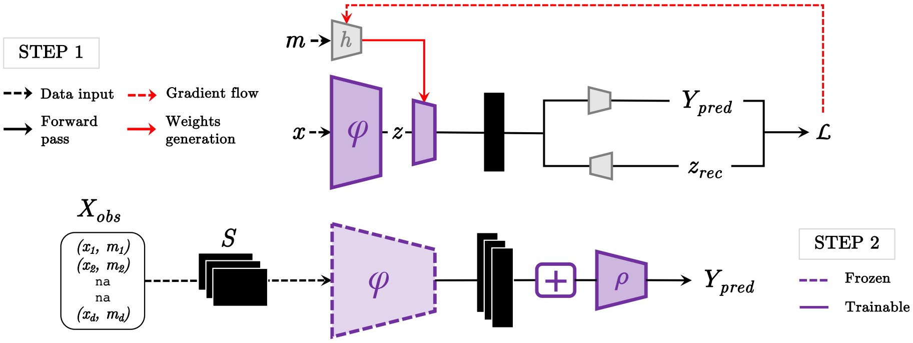

Fig. 1: Overview of our HyperMM framework.

$v_i = 1$ if x_i is missing, and 0 otherwise. The observed data of X can be expressed as $X_{obs} = (1 - v) \odot X + v \odot \texttt{na}$, where $\odot$ is the term-by-term product. In this setting, the learning goal becomes the prediction of Y given X_{obs}.

We intend on learning without the use of any form of imputation of missing modalities, and therefore, with entries of different dimensions. However, standard machine learning models, including MML models, are built to handle data inputs of a fixed size. In contrast, we aim to learn a sum-decomposable function f of the form $f = \rho(\sum \varphi(x_i))$, operating on *sets* and thus relaxing the requirement of fixed-dimensional data. We propose a two-step framework that we call HyperMM to implement our method. Figure 1 presents an overview of our strategy. In a first step, we learn an *unique* neural network φ that can extract features from any modality present in $\mathcal{D}$. Then in a second step, we freeze the learned φ, use it to encode each element of X_{obs}, and feed the combination of the encoded inputs to a classifier ρ through a permutation-invariant architecture.

2.1 Universal Feature Extractor

A single network φ that can encode all observed modalities in $\mathcal{D}$ is a requirement for learning a set function as described above. We propose to achieve this by first learning such universal feature extractor φ using a conditional hypernetwork [4]. In this pre-training step, we train a *main* network on all available images x in the dataset, without any modality pairing. As illustrated in Figure 1, we introduce an *auxiliary* network h that takes as input m, the modality identifier corresponding to x, and generates conditional weights for the last layer of the encoder φ of the main network. By doing so, the last step of the feature extraction is different for each modality, but still performed by the same network. Specifically, the modality-specific layers are generated through a common hypernetwork, which facilitates sharing of information across modality-specific layers.

To ensure that the features learned by φ are relevant, the network is trained to both predict y from the single modality images, and reconstruct z, the features outputed by the second-to-last layer of φ. This is achieved by optimizing a loss function of the form $\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CE}$, where $\mathcal{L}_{MSE}$ denotes the mean squared error between z and z_{rec}, and $\mathcal{L}_{CE}$ the cross-entropy loss between y and y_{pred}.

2.2 Permutation Invariant Architecture

Once we have learned φ, we use it to implement a permutation invariant network for supervised MML with missing modalities. To do so, we define S, the set representation of the $q = |S|$ observed elements of X_{obs}, such that $S := \{s_1, \dots, s_q\}$, with $q \leq d$. Each element s_j is represented as a tuple (x_i, m_i) consisting of an observed modality x_i, and the corresponding modality identifier m_i. This reformulation allows observations of varying dimensions. Thereby, it does not require nor expects all observations to have the same number of elements and it fully allows observations with missing modalities. A d-modal observation X_{obs} containing `na` values can simply be expressed as a set S of size $q \leq d$ where the `na` values are not represented anymore.

Using this definition, we leverage on the findings of [26], who proposed a learning framework that considers permutation invariant functions operating over sets. We reformulate our learning goal as one of learning a set function f of the form

$$f(X_{obs}) = \rho\left(\frac{1}{|S|}\sum_{s_k \in S} \varphi(s_k)\right), \tag{1}$$

where the function $\varphi : \mathbb{R} \times \{r \times r\}^d \to \mathbb{R}^{d_l}$ corresponds the encoder obtained from the pre-training phase, the function $\rho : \mathbb{R}^{d_l} \to \mathbb{R}$ is implemented as neural network, r is the size of each image and $d_l \in \mathbb{N}^+$ is the dimensionality of the latent space of φ.

As illustrated in Figure 1, a given observation X_{obs} with missing modalities is encoded as a set S. Each element $s_k \in S$ is then transformed into a representation $\varphi(s_k) := \varphi(x_i|m_i)$ through the pre-trained network φ conditioned by the modality identifier m. The representations $\varphi(s_k)$ are aggregated using a permutation invariant operation such as the sum, the mean or the maximum. The aggregation is processed through the network ρ, which allows to predict the target Y corresponding to the input X_{obs}. The proposed architecture interprets each observation S of a dataset as a set of unordered modalities, where all information available in X_{obs} is conserved and no new information, such as imputed images, is added. By transforming individual elements s_k of S at a time and then aggregating the transformations, our network encodes sets of arbitrary sizes into a fixed representation $\sum \varphi(s_k)$. This aspect is particularly relevant and further justifies handling our dataset with missing modalities as unordered sets.

Our permutation invariant model is learned by optimizing the loss function

$$\mathcal{L}(\theta) := \mathbb{E}_{(S,Y)\in\mathcal{D}}\Big[\ell\Big(Y, \rho_\theta\Big(\sum_{s_k \in S} \varphi(s_k)\Big)\Big)\Big], \tag{2}$$

where ρ is parametrized by θ, and ℓ is the cross-entropy loss. As φ is optimized in the pre-training step, its weights are not updated in this step.

3 Experiments

3.1 Alzheimer's Disease Detection

In a first application, we illustrate the performances of HyperMM and its robustness to missing modalities on the task of binary classification of Alzheimer's disease (AD) using multimodal images from the ADNI dataset [13]. We select a subset of 300 patients for which both T1-weighted MRIs and FDG-PET images are available, resulting in 165 cognitively normal (CN) and 135 AD observations. Before learning, all the samples are skull stripped using HD-BET [5], resampled through bicubic interpolation to set an uniform voxel size, standardized, and normalized using min-max scaling.

Baselines. We first evaluate the advantages of our strategy for MML with complete data. We compare the performances of HyperMM against: **Uni-CNN**, an unimodal CNN as implemented by [10]; **Multi-CNN**, a multimodal CNN as proposed by [22]; and **Multi-VAE**, a multimodal VAE [24] that we adapt for classification. Then, we compare our method against state-of-the-art techniques for MML with missing modalities in two scenarios: complete MRIs +50% of PETs available for training and testing, and complete PETs +50% of MRIs available. Specifically we compare to: **pix2pix**, a strategy where an image-to-image translation model [6] is trained on the subset of the training data containing only modality-complete samples, is then used to impute the missing modality of the incomplete data, and once imputed the data is classified using a Multi-CNN; and **cycleGAN**, the same strategy, only using a cycleGAN [28] for imputation.

Implementation details. Due to the high training time of GAN-based baselines, we perform the experiment on a single fold using a large subset for testing. The data is randomly split into train, validation and test sets with a 6:1:3 ratio on the patient-level For simplicity and fairness, we use the same feature extraction strategy (Figure 2) in all baselines. Specifically, 3D MRI and PET images are processed as batches of 2D slices that are each fed to a pre-trained frozen VGG11 [16] feature extractor. A 1D max pooling on the slice dimension of the input is then applied to the resulting blocks of feature maps, thus converting them into a single block. The resulting block is passed through a 1×1 convolution layer to adapt the pre-trained features into AD-specific ones. In our HyperMM

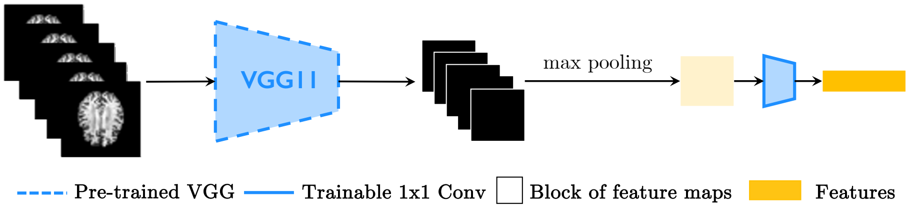

Fig. 2: Feature extraction strategy used in all the ADNI baselines.

Table 1: Performances on the ADNI dataset. **Bold** means best.

	Acc. (↑)	AUC (↑)	F1 (↑)	Prec. (↑)	Rec. (↑)	Time (↓)
Complete unimodal						
Uni-CNN PET	0.61	0.58	0.58	0.65	0.31	< 20 min
Uni-CNN MRI	0.71	0.69	0.58	**0.85**	0.43	< 20 min
Complete multimodal						
Multi-VAE classifier	0.66	0.64	0.53	0.73	0.41	< 30 min
Multi-CNN	0.70	0.70	0.67	0.67	0.68	< 30 min
HyperMM w/o pre-train (ours)	0.62	0.61	0.53	0.61	0.46	< 20 min
HyperMM w/ pre-train (ours)	**0.74**	**0.73**	**0.70**	0.70	**0.70**	< 1 h
100% MRI + 50% PET						
pix2pix	0.65	0.65	**0.62**	**0.62**	**0.61**	> 14+1 h
cycleGAN	0.63	0.62	0.57	0.61	0.53	> 30+1 h
HyperMM (ours)	**0.67**	**0.66**	0.60	0.60	0.60	< 1 h
100% PET + 50% MRI						
pix2pix	0.63	0.62	0.54	0.61	0.48	> 14+1 h
cycleGAN	0.61	0.59	0.47	0.61	0.39	> 30+1 h
HyperMM (ours)	**0.64**	**0.63**	**0.60**	**0.61**	**0.61**	< 1 h

implementation, this corresponds to the pre-training step, in which we simply make the last 1×1 convolution layer conditional. All models are implemented with PyTorch, and trained on an Nvidia TITAN Xp GPU for a maximum of 100 epochs using an early stopping strategy, a batch size of 1 and an Adam optimizer with an initial learning rate of $1e-4$.

Results. The performances of all models are reported in Table 1. Several observations can be drawn from these results. First, MML shows significant improvements over unimodal baselines. In particular, HyperMM achieves the best performances for binary classification of AD using complete multimodal data and considerably improves the F1-score, recall metric, and precision/recall balance. Second, MML with missing modalities still achieves better results than unimodal models. Notably, HyperMM trained on MRIs available even for only 50% of the patients performs better than an unimodal model trained on PETs only. Inversely, having access to PETs for 50% of the patients improves the F1-score and recall of learning from MRIs only. Third, HyperMM outperforms state-of-the-art strategies on MML with missing modalities. While GAN-based strategies can handle missing PETs in the input data, they are considerably less efficient in terms of precision/recall balance when the missing modality is MRI. In this scenario, the missing high-resolution MRIs need to be translated from the available low-resolution PETs before learning. In contrast, as HyperMM does not rely on any imputation, it performs well in both scenarios, and trains in significantly less time than competitors. Lastly, these results highlight the importance of the pre-training and conditioning step of the HyperMM framework.

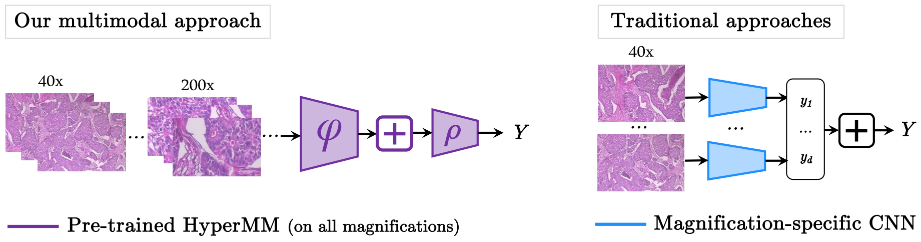
Our multimodal approach
40x
200x
φ
ρ
Y
Pre-trained HyperMM (on all magnifications)
Traditional approaches
40x
y1
...
yd
Y
Magnification-specific CNN

Fig. 3: Comparison of decision strategies for patient-level tumor classification.

3.2 Breast Cancer Classification

In a second application, we demonstrate the flexibility of HyperMM and its benefits for learning with varying-sized datasets, beyond the scenario of missing modalities. We perform binary classification of breast cancer using histopathological images from the BreaKHis dataset [17]. BreaKHis contains multiple images per sample (i.e. patient) of benign or malignant tumors observed through different microscopic magnifications: 40×, 110×, 200×, and 400×. We select a balanced subset of the data composed of samples of 24 benign and 29 malignant tumors, resulting in 5,575 images in total. We use the images as they are for learning, and do not perform any pre-processing or data augmentation.

Baselines. In clinical practice, pathologists combine the complimentary information present in images captured under different magnifications in order to make a patient-level decision. Nonetheless, most current learning approches consist of magnification-specific models, due to the difficulty of processing images of different natures with a single model. Moreover, because the number of available images can vary a lot from one patient to another, traditional algorithms cannot be applied at the patient-level. Existing methods rather predict from individual images, and later combine the predictions in order to form a global decision. Instead, we propose to approach this problem as one of MML with missing data, where each magnification level represents a modality. We classify tumors at patient-level by combining all available images during training directly. The differences between our approach and traditional ones are further illustrated in Figure 3. We evaluate the benefits of HyperMM for learning from histopathology data, and compare its performances with: **CNN**, a strategy in which a magnification-specific CNN is trained to classify tumor types from individual images, and patient-level prediction is obtained by averaging the classification scores of individual images, following [17]; and **Incremental-CNN**, in which a magnification-agnostic CNN is trained by incrementally updating its weights on successive batches of 40×, 100×, 200× then 400× magnifications, as proposed in [12]. The patient-level decision is obtained similarly to the previous baseline.

Implementation details. We randomly split the data into train and test sets with a 8:2 ratio on the patient-level, and repeat all experiments 5 times. We

Table 2: Performances (mean ± std) on the BreaKHis dataset. **Bold** means best.

	Acc. (↑)	AUC (↑)	F1 (↑)	Prec. (↑)	Rec. (↑)
Magnification-specific					
CNN 40×	0.83±0.07	0.81±0.07	0.83±0.06	0.85±0.08	0.83±0.08
CNN 100×	0.85±0.08	0.85±0.08	0.87±0.06	0.85±0.07	0.90±0.07
CNN 200×	0.84±0.07	0.84±0.09	0.84±0.05	0.80± 0.11	0.90± 0.09
CNN 400×	0.83±0.09	0.83±0.09	0.85±0.10	0.88±0.11	0.83±0.15
Magnification-agnostic					
Incremental-CNN	0.89±0.11	0.88±0.12	0.90±0.10	0.88±0.12	**0.93±0.09**
HyperMM (ours)	**0.92±0.06**	**0.91±0.07**	**0.90±0.08**	**0.94±0.09**	0.88±0.10

use a pre-trained VGG11 [16] feature extractor for all baselines, similarly to Sec. 3.1, only without 1D max pooling. All models are trained for a maximum of 50 epochs using an early stopping strategy, and an Adam optimizer with an initial learning rate of $1e-4$. We use a batch size of 16 for image-level baselines (i.e. CNN and Incremental-CNN) and 1 for HyperMM.

Results. All performances averaged over 5 repetitions are reported in Table 2. They underline the clear benefits of HyperMM for cancer classification from histopathological images. In particular, our method outperforms magnification-specific models, and is closely followed by Incremental-CNN, which highlights the benefits of combining the information carried by different magnifications. Moreover, while Incremental-CNN maximizes the recall score of the task, HyperMM maximizes precision, and overall improves upon Incremental-CNN. This shows that learning to predict an early latent combination of features (i.e. combining multiple images of a same patient during model training directly) yields better performances than combining predictions made on individual images.

4 Conclusion

We have demonstrated the advantages of HyperMM for robust MML with missing modalities: our method eliminates the need to use complex and computationally costly imputation strategies, thus significantly decreasing model training time; and unlike competitors, its performances are not dependant on which modality is missing in the data. In addition, we have shown that the flexibility of HyperMM alleviates the constraints usually met in applications with varying-sized datasets and opens up a whole new range of possible learning strategies. Our framework is *task-agnostic*, and can be easily used beyond the two applications we have presented. For instance, it could be extended to multivariate time series analysis, where incomplete data is common (e.g. damaged channels in EEG recordings). Moreover, while we used pre-trained feature extractors in all our experiments for simplicity, HyperMM is *model-agnostic* and adaptable to any neural network-based feature extractor or predictor.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[2] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166, 2018.

[3] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim. Multimodal spatial attention module for targeting multimodal pet-ct lung tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3507–3516, 2021.

[4] D. Ha, A. M. Dai, and Q. V. Le. Hypernetworks. *CoRR*, 2016.

[5] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[7] N. Jaques, S. Taylor, A. Sano, and R. Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 202–208, 2017.

[8] J.-C. Kim and K. Chung. Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE Access*, 8:104933–104943, 2020.

[9] M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What's a good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

[10] G. Liang, X. Xing, L. Liu, Y. Zhang, Q. Ying, A.-L. Lin, and N. Jacobs. Alzheimer's disease classification using 2d convolutional neural networks. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3008–3012, 2021.

[11] Z. Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] M. S. Mayouf and F. Dupin de Saint-Cyr. Curriculum incremental deep learning on breakhis dataset. In *Proceedings of the 2022 8th International Conference on Computer Technology Applications*, pages 35–41, 2022.

[13] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[14] M. Odusami, R. Maskeliūnas, R. Damaševičius, and S. Misra. Machine learning with multimodal neuroimaging data to classify stages of alzheimer's disease: a systematic review and meta-analysis. *Cognitive Neurodynamics*, pages 1–20, 2023.

[15] T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torné, E. Sala, P. Lió, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015.

[18] W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang. Semi-supervised multimodal image translation for missing modality imputation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4320–4324, 2021.

[19] Q. Suo, W. Zhong, F. Ma, Y. Yuan, J. Gao, and A. Zhang. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, volume 3534, page 3540, 2019.

[20] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann, et al. Multimodal imaging in alzheimer's disease: validity and usefulness for early detection. *The Lancet Neurology*, 14(10):1037–1053, 2015.

[21] C. M. Tempany, J. Jayender, T. Kapur, R. Bueno, A. Golby, N. Agar, and F. A. Jolesz. Multimodal imaging for improved diagnosis and treatment of cancers. *Cancer*, 121(6):817–827, 2015.

[22] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[23] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[24] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

[25] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[26] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[27] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou. Unified multi-modal image synthesis for missing modality imputation. *arXiv preprint arXiv:2304.05340*, 2023.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.