Synthetic Medical Dataset from Multimodal Masked Autoencoder

1st Yunhao Chen
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
1191200221@stu.jiangnan.edu.cn

***Abstract*—Lack of data has been one of the main obstacles to developing deep learning, especially with the rise of attention-based models, which can easily overfit datasets with over 1 million samples. Moreover, we can only access a minority of medical data for general public research due to legal concerns about patient privacy. What is more, multimodal has already been proven experimentally and theoretically effective. Consequently, a new multimodal masked autoencoder for cross-modal dataset generation was proposed. Firstly, a new training strategy called modal decoupling was proposed to improve the diversity of generated datasets. Then, a new masked autoencoder was proposed for cross-modal dataset generation with new fusion strategies. The whole pipeline is tested on CoronaHack -Chest X-Ray-Dataset and COUGHVID Dataset.**

***Index Terms*—Multimodal, MAE, Data Generation, Attention Mechanism**

I. Introduction

The modern deep learning technology applied in medical diagnosing is in great demand for publicly inaccessible labelled data, which is inaccessible due to proprietary and privacy reasons. It is difficult for medical professionals to make most medical images public without patient consent [5]. Besides, the usability of publicly accessible data is inferior due to the fact of lacking information o size and annotations. Specifically, most deep learning algorithms can fit nicely on the data samples they are trained on, yet have difficulty generalizing other samples, whose number is limited. At the same time, deep learning can easily overfit the dataset with millions of samples [3], with the rise of several attention-based models (transformer and similar architectures). In short, the potential and growth of deep learning is limited in medical diagnosis.

In natural language processing (NLP) and computer vision (CV), the self-supervised pretraining model effectively addresses the demand for data. Conceptually, the autoregressive language model in Generative Pre-Training [6], the masked autoencoder model in BERT [7], and MAE [1], remove a portion of the input, and then, train a model to forecast the masked information. These techniques obtain more than 100 billion parameters, when training models in NLP and CV.

The limits of the individual modalities can be overcome by combining other modalities. T2-weighted MRI, for instance, is effective at lowering the number of false-positive results when used in breast cancer screening, despite contrast-enhanced MRI's higher sensitivity in spotting breast tumours [8]. Moreover, human perceptual learning is greatly facilitated by simultaneous multimodal senses [9]. Intuitively, multimodal learning can aggregate information from multiple data sources, which helps to obtain the complete representation. In the case of video classification, for example, a multimodal model using textual information such as captions, audio information, and visual information is significantly better than a unimodal model with only one type of information [9], [10].

Synthetic dataset methods are proposed by [5], [19], [21]. However, these methods are all based on Generative Adversarial Network(GAN), which is well known for being delicate and unstable when training [30]. The main issues with GAN are likewise nonconvergence and diminished gradient. It is challenging to balance the generator and discriminator in a single training since reducing the loss function necessitates many training steps. The cost function is raised as a result [31]. Moreover, these methods can only generate single-modal dataset samples, which fails to take advantage of the potential provided by multi-modal data samples. In addition, as we have illustrated above, modern deep learning techiques can easily overfit. Consequently, GAN requires large amounts of data samples to be well trained which can be overcame by the masked autoencoder.

In this paper, the self-supervised strategy with cross-modal modification is introduced to improve accuracy through multi-modal learning and generate reliable datasets free of legal and privacy concerns.

In this paper, contributions are described as follows.

(1) A cross-modal masked autoencoder with new fusion strategies is proposed for data generation.

(2) A new training strategy is utilized for cross-modal classification.

(3) A effective loss function for the multimodal masked autoencoder is proposed.

II. Related Work

A. Masked Image Modelling

Masked image modelling (MIM) [1] is an efficient pretraining method for the Vision Transformer (ViT) [4], which is powerful but challenging to train due to a lack of inductive bias. Masking and reconstructing, namely masking a series of picture patches before sending them to the transformer and reconstructing the patches at the output, are the fundamental concepts of MIM. MIM encourages the network to gather data from the context and deduce the identity of the hidden target. Both vision and medical imaging tasks require the capacity to aggregate contextual information. Masked Autoencoder (MAE) [1] is one of the most straightforward and efficient MIM frameworks. Also, recent works on the MIM task for Medical Image Modelling [11] and Audio Classification [12] have demonstrated the practical application of MAE in medical images and audio. Consequently, it is viable for us to investigate the effectiveness of cross-modal MAE.

B. Multimodal Learning Applications and Theories

The fusion of signals from different modalities has become much easier owing to the rapid development of deep learning techniques. For instance, [8], [9], [13], [14] combine RGB and thermal Images to improve object detection or fuse video and audio to increase classification accuracy or explore the enhancement of segmentation effects with RGB and depth images integrated altogether. What is more, unlike deep learning itself, the effectiveness of multimodal can not only be verified by multiple experiments but also can be explained by solid mathematical derivations and proofs. For example, Total Correlation Gain Maximization (TCGM) [15] proved that TCGM could find the ground-truth Bayesian classifier provided that each modality is given. In addition, [10] theoretically justified that multimodal learning is genuinely better than single-modal learning when the dataset is the same by proving latent representation quality is connected to the final result.

C. Synthetic Data with GAN

Using synthetic data, researchers from various disciplines have made private data accessible to the general public. For instance, the US Census gathers personally identifiable information (PII) concerning the United States population, including occupation, education, income, and geography. Even if sources are de-identified and obscured, there is a sizable danger of deanonymization due to the inherent distinctiveness of the data [17]. Due to privacy concerns, access to this crucial data, which contains numerous potentially helpful statistical connections, is restricted. Reiter, a researcher at Duke University, developed artificial business census data to resolve this privacy issue [18]. As a result, the Synthetic Longitudinal Business Database [19], the first publicly accessible record-level database on business establishments, was made available in 2011.

Moreover, with the rise of deep learning techniques, the generative adversarial network has been implemented to generate medical datasets [5], [20], [21]. However, the GAN is not only very unstable when training but also can not generate labelled datasets. Moreover, these methods only rely on a single modal to generate another, resulting in instability compared with multimodal.

III. Method

A. Overlapping Latent Space Assumption

The manifold assumption posits that high-dimensional data can be represented in a lower-dimensional latent space [22], which can be more easily analyzed and has lower degrees of freedom. This latent space is learned using techniques such as auto-encoders and variational autoencoders [23]–[25] and helps to capture the underlying structure of the data within a single domain. However, this approach does not consider dependencies between different domains when no supervision is provided. The shared latent space assumption can alleviate this problem [26]

The shared latent space assumption derives from the manifold assumption and is commonly used in domain adaptation. This assumption posits that different domain data views can be mapped to a common latent space, or code, representing a shared structure between the domains. This shared latent space can be learned using techniques such as multi-task learning or transfer learning, which involve training a model on multiple tasks or domains simultaneously to improve performance on a target task or domain. By learning this shared latent space, the model can better utilize information from multiple domains and make more accurate predictions.

However, different from the assumption that every-to-one mapping can be seen as an over-simplification in the [26], we suppose that the cross-modal signals have the same features overlapped in one latent space (e.g. consider the audio and x-ray image of Covid-19 patients, the most essential latent space the signals have in common is whether they are the representations of infection.). The decoders, like the transformers, will be more than able to acquire the distinct features of different modalities.

Unlike the shared latent space assumption, the overlapping latent space assumption refers to the idea that a latent space or representation not only underlies the same structures of multi-modal data but also forms the basis of different features of separate modalities.

Consequently, the overlapping latent space can represent the same structure and patterns in complex cross-modal datasets. Moreover, this latent space has the information to construct dissimilar modalities separately. Specifically, we assume that a given data $\mathbf{x} := \left(x^{(1)}, \cdots, x^{(K)}\right)$ consists of K modalities, where $x^{(k)} \in \mathcal{X}^{(k)}$ which means the data x belongs to the modal k. Consequently, the input data space can be formalized as follows.

$$\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(K)} \tag{1}$$

We use $\mathcal{Y}$ to denote the target domain. The function $g^{\star} : \mathcal{X} \mapsto \mathcal{Z}$ represents the correct mapping from the input space (including all K modalities) to the latent space. The function $h^{\star}$ represents the correct mapping at the task layer, $h^{\star} : \mathcal{Z} \mapsto \mathcal{Y}$. Data $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_k}) \in \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(K)}$ is sampled from a distribution $\mathcal{D}$.

The assumption can be formalized in the following form. The $\mathcal{Z}_{overlapping}$ means the overlapping latent space.

$$\mathcal{Z}_{overlapping} = \bigcup_{i=1}^{K} \mathcal{Z}_i \tag{2}$$

Or formalized in probabilistic form as follows:

$$\mathbb{P}_{(\mathcal{X}^{(1)}\times\cdots\times\mathcal{X}^{(K)})}\left(\mathcal{Z}_{overlapping}\right) = \mathbb{P}_D\left(\bigcup_{i=1}^{K} g^\star(\mathbf{x}_i)\right) \tag{3}$$

Where $\mathbb{P}_i$ means the probabilties under the circumstance i. The overlapping latent space is shown in fig1.

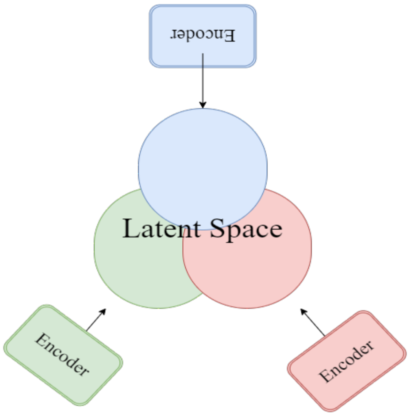

Fig. 1. Overlapping Latent Space Assumption.

B. Modal Decoupling Training Strategy

According to the overlapping latent space assumption(Equation 2,3), the shared latent space is what the modalities have in common (e.g. the cough and x-ray of patients infected with covid-19 have the shared latent space representing their disease type or health status). The complement of the shared latent space is the different features between modalities which we will refer to as dissimilar latent space and can be formalized as follows:

$$\mathbb{P}_{(\mathcal{X}^{(1)}\times\cdots\times\mathcal{X}^{(K)})}\left(\mathcal{Z}_{dissimilar}\right) = \mathbb{P}_D\left(\bigcap_{i=1}^{K} g^\star(\mathbf{x}_i)\right) \tag{4}$$

Where the $\mathcal{Z}_{dissimilar}$ means the dissimilar latent space.

Or in a simple form:

$$\mathcal{Z}_{dissimilar} = \bigcap_{i=1}^{K} \mathcal{Z}_i \tag{5}$$

Also, according to the [10], the better the latent space the model acquires, the better the results. Moreover, the generalization of classification and generation of datasets depends on (3) and (4). However, the limited paired data samples of medical datasets can not help the algorithm acquire a good representation of latent space. As a result, we decouple the modalities of paired data samples into separate ones and combine each other individually (an Image-Audio cross-modal example is illustrated in fig 2). This is possible because data with the same label in different modes have the same shared latent space. And their different features are represented in the dissimilar latent space, which does not interfere with the shared latent space. In addition, this is beneficial to the generalization of the model because the specific information of certain paired data is dissolved in this training strategy. Because the network can acquire, e.g. the overall structure of a patient's lung instead of the lung's lesion to identify whether the patient is infected due to the limited dataset samples and great generalization capabilities of neural network.

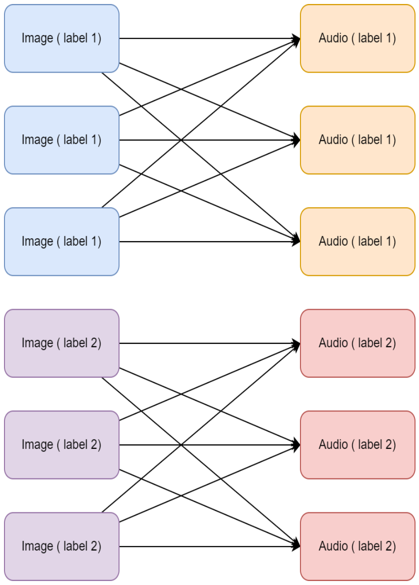

Fig. 2. Modal Decoupling Training Strategy.

C. Gradual Cross Attention Fusion Strategy

Based on the unique property that the learning algorithm is composed of different depths' layers of deep learning and the latent space assumption, the shallower the layer, the more modalities differ. For example, the input representations for audio and vision tasks differ significantly. While many state-of-the-art audio classification methods use short-term Fourier analysis to produce log-Mel spectrograms and feed them into CNNs designed for image data, the time-frequency representations of these spectrograms have different distributions compared to images. In particular, multiple acoustic objects can have energy at the same frequency, so the translation invariance property of CNNs may not be as valuable. In contrast, the visual stream in a video is three-dimensional, with two spatial dimensions and one temporal dimension. While different spatial regions of an image may correspond to different objects, there is a unique challenge of high redundancy across multiple frames. As a result, input representations, neural network architectures, and benchmarks tend to vary widely depending on the modality used [9]. As a result, the late fusion [9] is more reasonable. However, we still can not determine when to fuse each modality. Consequently, we use learnable parameters whose initialised values are greater as the layer goes deeper to determine the best way to fuse each modality. The fusion process is demonstrated in fig3. The fusion process of each layer can be formalized as follows:

$$\mathcal{Z}_i^{\mathcal{A}} = {}^{\mathcal{L}}g_i^{\mathcal{I}}\left({}^{\mathcal{C}}g_i^{\mathcal{A}}\left(\mathcal{Z}_{i-1}^{\mathcal{I}}, \mathcal{Z}_{i-1}^{\mathcal{A}}, \mathcal{Z}_{i-1}^{\mathcal{A}}\right)\xi_i^{\mathcal{A}} + \mathcal{Z}_{i-1}^{\mathcal{A}}\right) \tag{6}$$

$$\mathcal{Z}_i^{\mathcal{I}} = {}^{\mathcal{L}}g_i^{\mathcal{I}}\left({}^{\mathcal{C}}g_i^{\mathcal{I}}\left(\mathcal{Z}_{i-1}^{\mathcal{A}}, \mathcal{Z}_{i-1}^{\mathcal{I}}, \mathcal{Z}_{i-1}^{\mathcal{I}}\right)\xi_i^{\mathcal{I}} + \mathcal{Z}_{i-1}^{\mathcal{I}}\right) \tag{7}$$

where $\mathcal{Z}_i^{\mathcal{I}}$ means the i-th latent space of Image, $\mathcal{Z}_i^{\mathcal{A}}$ means the i-th latent space of Audio, $\xi_i^{\mathcal{I}}$ means the i-th learnable parameter of Image, $\xi_i^{\mathcal{A}}$ means the i-th learnable parameter of Audio, ${}^{L}g_i^{\mathcal{I}}$ means the i-th layer's transformer of Image, ${}^{\mathcal{L}}g_i^{\mathcal{A}}$ means the i-th layer's transformer of Audio, ${}^{\mathcal{C}}g_i^{\mathcal{I}}$ means the i-th layer's cross attention of Image, ${}^{\mathcal{C}}g_i^{\mathcal{A}}$ means the i-th layer's cross attention of Audio.

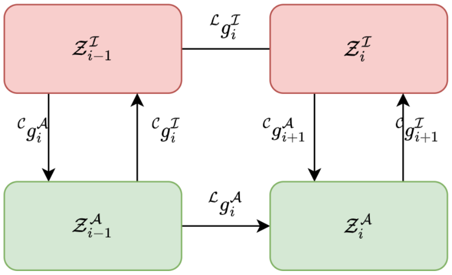
$\mathcal{Z}_{i-1}^{\mathcal{I}}$
${}^{\mathcal{L}}g_i^{\mathcal{I}}$
$\mathcal{Z}_i^{\mathcal{I}}$
${}^{\mathcal{C}}g_i^{\mathcal{A}}$
${}^{\mathcal{C}}g_i^{\mathcal{I}}$
${}^{\mathcal{C}}g_{i+1}^{\mathcal{A}}$
${}^{\mathcal{C}}g_{i+1}^{\mathcal{I}}$
$\mathcal{Z}_{i-1}^{\mathcal{A}}$
${}^{\mathcal{L}}g_i^{\mathcal{A}}$
$\mathcal{Z}_i^{\mathcal{A}}$

Fig. 3. Gradual Cross Attention Fusion Strategy.

D. Loss Backpropagation Decoupling Strategy for MAE

To better train the models, we have to normalize the Image data and Audio data before training. However, due to the intrinsic differences between the Image modal and the Audio modal, they have different scales after normalization. Conseqeuntly, if we add the losses from the Image and Audio together without further consideration, the backpropagation will be inaccurate due to the different scales between Image and Audio data. However, it is also inappropriate to normalize the audio's magnitude to zero and one. Unlike the image magnitude, which can be easily scaled to zero and one, recovering the audio's magnitude can be more difficult. Consequently, we decouple the loss backpropagation strategy. Namely, we backpropagate loss separately.

E. MultiModal Masked Autoencoder

We can now ensemble the Cross-Modal Masked Autoencoder according to the abovementioned methods. The whole structure is illustrated in fig 4.

IV. Experiments

A. Experiments Setup

We evaluate our approach on the CoronaHack -Chest X-Ray- Dataset [2] and COUGHVID Datase [3] for synthetic Medical Dataset. For the input tensor, Mel spectrogram is extracted with the size of (100,400) with a window length of 0.025s and the hop length of 0.01s. Also, what needs to emphasize is that each sample in the COUGHVID Database dataset is cut into 4 segments to fit the input size of our model. Consequently, audio samples have a size of (100,100). Also, the training batch size is 128. As a result, the input size is (128,100,100). The data sample of CoromaHack is resized to (224,224) for training. We used the AdamW optimizer in the experiments with a learning rate of 0.001, a weight-decay of 0.05 and other parameters are default. The loss we used in the experiments is the Cross-entropy-loss with label-smoothing of 0.1. The training epochs are 350 for classification and 20 for multimodal masked autoencoder. To prove the efficiency of our network, we did not use techniques like data augmentation, model transfer, EMA, pretraining, or others. We used 10-Fold for performance evaluation for classification and MSE for multimodal masked autoencoder.

B. Evaluation Metrics

Our pipeline produced a synthetic audio data sample from corresponding real audio and randomly chosen x-ray images of the same label. We used this data to train a ResNet network. We evaluated the ResNet network on test images from the original dataset. We also calculated the variance between the synthetic and real datasets through a Kullback–Leibler (KL) divergence score as the VAE [29], to measure the difference between two probability distributions.

We must analyze the adversarial divergence when considering generative models to calculate the statistical correlation between the generated and original data. The KL divergence score has been the standard to measure this for generative models, calculated by: Given two probability distributions, P and Q, the KL divergence from Q to P is defined as:

$$KL\left(P, Q\right) = \sum_i P_i \left(\ln \frac{P_i}{Q_i}\right) \tag{8}$$

The KL divergence measures how much information is lost when approximating P with Q.

We use simple accuracy to measure the classification outcome. This score can display the effectiveness of classification, which we use to compare the the classifier produced from our synthetical dataset and original dataset.

C. Quantitative Results

We received an accuracy rating of 0.62 for our synthetically trained ResNet and an accuracy of 0.614 for our DRIVE-trained ResNet on the audio dataset, the COUGHID dataset. The minor difference between the two scores illustrates the high calibre of the training data we generate.

To assess the variance of our datasets, we calculated the KL divergence, which indicates the difference between the distributions of the two datasets. The KL divergence score for the synthetic data compared to the real data was 6.84, while the KL divergence score for two random subsets of the real data was 5×10^{-4}. This small score for the subsets of real data is expected, as these subsets are drawn from the same dataset. The higher KL divergence score for the synthetic data compared to the real data demonstrates that the synthetic data does not simply replicate the original distribution.

Also, as the pixel-intensity distribution of three pairs of natural and synthetic images (fig 6) and multi-modal synthetic

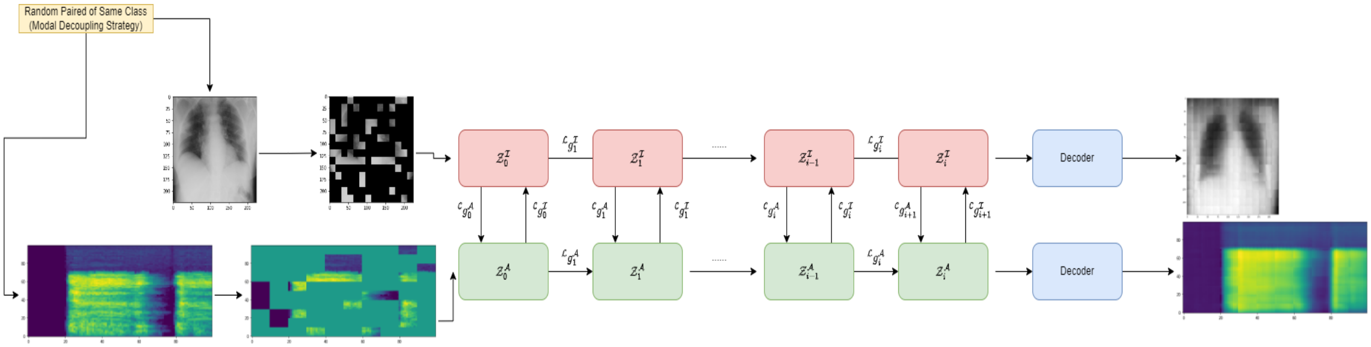

Fig. 4. MultiModal Masked Autoencoder.

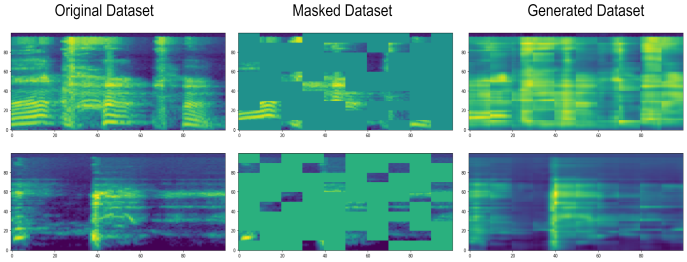

Fig. 5. Single Modal Synthetic Data samples.

data samples fig 7 demonstrated, the pipeline could generate data samples which are different yet acceptable new ones.

D. Abalation Study

To demonstrate the effectiveness of the cross-modal masked autoencoder, we conduct an ablation study between single-modal masked autoencoder(fig 5) and cross-modal masked autoencoder with different masking ratios.

As shown in fig 8, the variance of a single-modal masked autoencoder is significantly lower than the cross-modal masked autoencoder. Moreover, the cross-modal masked autoencoder's change in variance is more stable than the single-modal's. Consequently, we can come to the conclusion that the cross-modal MAE is better than the single-modal's.

V. Discussion

Two transformers cannot provide clear Images and audio due to the enormous variance in medical imaging and audio data (different illuminations, noise, patterns, intervals, etc.). The Masked Autoencoder cannot identify complicated structures, as seen by the audio spectrogram and poorly defined X-ray images. Only essential qualities like colour, shape, and brightness may be recognized by it.

Moreover, the masking procedure will lose details in the generated Image and audio. Through masking, we can improve the intrinsic variance of the generated dataset. However, if critical parts are masked, the decoders will more likely generate unusable datasets.

This lack of detail is unacceptable for medical Image and audio generation, as medical images and audio have many intricacies that must be accurately represented for the data to be usable.

Also, the variance of the dataset generated from the masked autoencoder is dependent on the existing dataset. Though the change of masking ratio and the implementation of cross-modal masked autoencoder can increase the variance of the generated dataset, the dependency on the existing dataset is a significant shortcoming that can not be easily overlooked.

VI. Conclusion

In this thesis, we propose a new latent space assumption, a new training strategy, a new cross-modal way and a cross-modal masked autoencoder for medical dataset generation. During the training session, we decouple the modal and pair the audio and Image of the label one by one and feed them into the masked autoencoder. Then we make use of the gradual cross-attention for modal synthesis. In the end, we use the decoder to reconstruct the Image and Audio.

Also, we test the whole pipeline with KL divergence and ResNet. The KL divergence proves that the pipeline can produce datasets with more diverse datasets with similar dataset qualities.

To conclude, the proposed method can generate preferable audio data samples with high variance than the single modal masked autoencoder.

References

[1] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.Image1). Masked Autoencoders Are Scalable Vision Learners. arXiv. https://doi.org/10.48550/arXiv.2111.06377

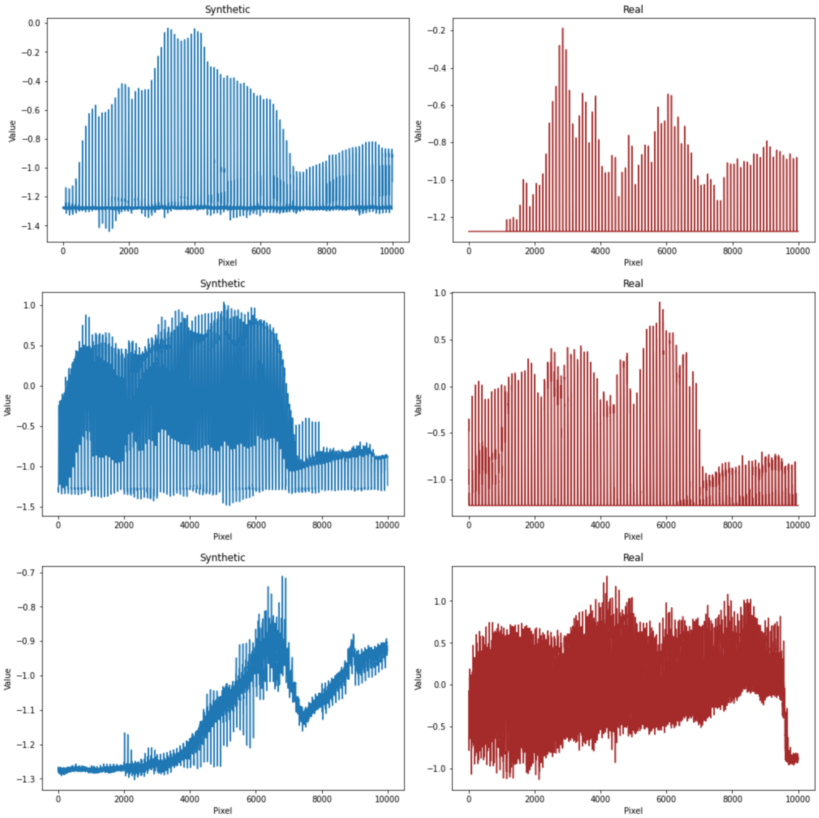

Fig. 6. Pixel-intensity distribution of real and synthetic audio.

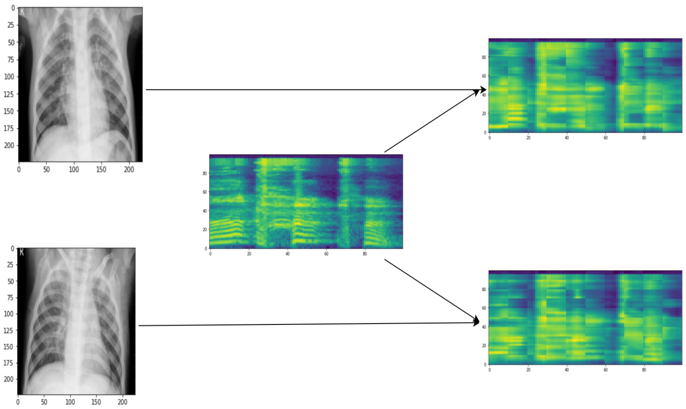

Fig. 7. Cross Modal Synthetic Data Samples. We randomly chose one Image and one piece of Audio with the same label and generated new data samples to improve the variance.

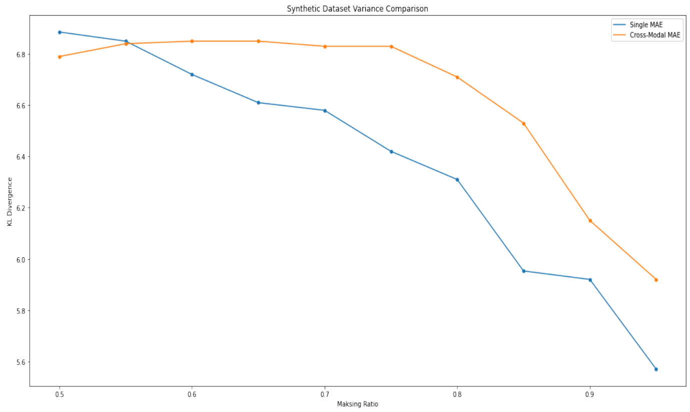

Fig. 8. Ablation Study. We compare the variance between single-modal masked autoencoder and cross-modal masked autoencoder with different masking ratios.

[2] COVID-19 Image Data Collection: Prospective Predictions Are the Future Joseph Paul Cohen and Paul Morrison and Lan Dao and Karsten Roth and Tim Q Duong and Marzyeh Ghassemi arXiv:2006.11988, https://github.com/ieee8023/covid-chestxray-dataset, 2020

[3] Orlandic, L., Teijeiro, T. Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci Data* 8, 156 (2021). https://doi.org/10.1038/s41597-021-00937-4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[5] Guibas, J. T., Virdi, T. S., Li, P. S. (2017). Synthetic Medical Images from Dual Generative Adversarial Networks. arXiv. https://doi.org/10.48550/arXiv.1709.01872G.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language

models are few-shot learners. In NeurIPS, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.

[8] Li, C., Sun, H., Liu, Z., Wang, M., Zheng, H., Wang, S. (2019). Learning Cross-Modal Deep Representations for Multi-Modal MR Image Segmentation. arXiv. https://doi.org/10.48550/arXiv.1908.01997

[9] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C. (2021). Attention Bottlenecks for Multi-modal Fusion. arXiv. https://doi.org/10.48550/arXiv.2107.00135

[10] Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., Huang, L. (2021). What Makes Multi-modal Learning Better than Single (Provably). arXiv. https://doi.org/10.48550/arXiv.2106.04538

[11] Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P. (2022). Self Pre-training with Masked Autoencoders for Medical Image Analysis. arXiv. https://doi.org/10.48550/arXiv.2203.05573

[12] Huang, P., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C. (2022). Masked Autoencoders that Listen. arXiv. https://doi.org/10.48550/arXiv.2207.06405

[13] Chen, Y., Shi, J., Ye, Z., Mertz, C., Ramanan, D., Kong, S. (2021). Multi-modal Object Detection via Probabilistic Ensembling. arXiv. https://doi.org/10.48550/arXiv.2104.02904

[14] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE international conference on computer vision, pages 4980–4989, 2017.

[15] Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In European Conference on Computer Vision, pages 171–188. Springer, 2020.

[16] Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. 2019.

[17] Aditi Ramachandran, Lisa Singh, Edward Porter, Frank Nagle. Exploring Re-identification Risks in Public Domains, Georgetown University, Harvard University, https://www.census.gov/srd/CDAR/rrs2012-13 Exploring Re-ident Risks.pdf

[18] Jarmin, R. and Louis, T. (2014). [ebook] Washington: U.S. Census Bureau, Center for Economic Studies,https://www2.census.gov/ces/wp/2014/CES-WP-14-10.pdf

[19] Satkartar K. Kinney, Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M.Abowd. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, 79(3):362–384, December 2011.

[20] Zein, H., Chantaf, S., Fournier, R. (2022). Generative Adversarial Networks for anonymous Acneic face dataset generation. arXiv. https://doi.org/10.48550/arXiv.2211.04214

[21] Tariq, U., Qureshi, R., Zafar, A., Aftab, D., Wu, J., Alam, T., Shah, Z., Ali, H. (2022). Brain Tumor Synthetic Data Generation with Adaptive StyleGANs. arXiv. https://doi.org/10.48550/arXiv.2212.01772

[22] Chapelle, O., Scholkopf, B., Zien, Eds., A.: Semi-supervised learning (chapelle, o.et al., eds.; 2006) [book reviews]. IEEE Transactions on Neural Networks 20(3),

[23] . Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)

[24] . Hinton, G., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets.

[25] Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2014)
Neural computation 18, 1527–54 (08 2006)
542–542 (2009)

[26] Mayet, T., Bernard, S., Chatelain, C., Herault, R. (2022). Domain Translation via Latent Space Mapping. arXiv. https://doi.org/10.48550/arXiv.2212.03361

[27] Lin, H., Cheng, X., Wu, X., Yang, F., Shen, D., Wang, Z., Song, Q., Yuan, W. (2021). CAT: Cross Attention in Vision Transformer. arXiv. https://doi.org/10.48550/arXiv.2106.05786

[28] Dietterich, T.G. Ensemble methods in machine learning. in International workshop on multiple classifier systems. 2000. Springer.

[29] Kingma, D. P., Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv. https://doi.org/10.48550/arXiv.1312.6114

[30] Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein GAN. arXiv. https://doi.org/10.48550/arXiv.1701.07875

[31] Asimopoulos, D.C., Nitsiou, M., Lazaridis, L., Fragulis, G.F. (2022). Generative Adversarial Networks: a systematic review and applications. SHS Web of Conferences.