# Synthesize Missing Modality Based on Latent Space Model

Kai Zhang
*College of Intelligence and Computing*
*Tianjin University*
Tianjin, China
zhangkai_@tju.edu.cn

Ruonan Liu
*College of Intelligence and Computing*
*Tianjin University*
Tianjin, China
ruonan.liu@tju.edu.cn

Dongyue Chen
*College of Intelligence and Computing*
*Tianjin University*
Tianjin, China
dyue_chen@163.com

Wenlong Yu
*College of Intelligence and Computing*
*Tianjin University*
Tianjin, China
751909848@qq.com

Yusheng Pu
*College of Intelligence and Computing*
*Tianjin University*
Tianjin, China
pys@tju.edu.cn

Di Cao
*College of Intelligence and Computing*
*Tianjin University*
Tianjin, China
cd1147583042@126.com

2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT) | 978-1-6654-2630-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACAIT53529.2021.9731283

***Abstract*—Multi-modal learning is currently a research hotspot in the field of artificial intelligence. Multi-modal learning effectively improves learning performance since it comprehensively utilizes information from multiple modalities. However, in real-world application scenarios, it is often impossible to collect complete multi-modal data, which limits the wide application of multi-modal learning. High-quality missing modality synthesis is still a challenge. In this work, we propose a novel method to synthesize the missing modality. Specifically, we utilize the common latent representation space model to adaptively fuse the consistent and complementary information in existing modalities, and then the synthesis network with 1d-CNN layers and MLP is employed to synthesize the missing modality. In addition, "threshold-loss" is proposed to tackle the over-optimizing phenomenon during the testing stage. Experiments demonstrate the proposed method outperforms other existing methods.**

***Index Terms*—multi-modal, latent space, synthesis, missing modality, latent representation, multi-modal learning**

## I. Introduction

In the era of big data, the ways of data collection are simple and diverse, multi-modality data can be easily acquired. Multi-modality data refers to the data collected from different perspectives of the same datum point, these data can provide complementary information about the datum point [1], multi-modality data can be widely obtained in various fields. For example, data from web pages contains various modal data including image, text, audio, etc.; in the medical field, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) are commonly used multi-modality data in brain diseases diagnosis and treatment; remote sensing technology can obtain a variety of different modal data, including RGB images, Lidar, and multi-spectral data [2]. Based on the demand to make full use of multi-modality data, multi-modal learning has become a hotspot of research.

Using these data from multiple perspectives correctly and reasonably can significantly improve learning performance. Researchers have made efforts to fully tap the potential of multi-modality data in multiple fields. Literature [3] proposed a multi-modal medical image fusion method in the spatial domain. Firstly, a moving frame-based decomposition framework is adopted to decompose source images to texture components and approximation components, and then a novel weight map refined strategy based on image properties and guide filtering is implemented to produce the final map which is applied to fuse the approximation components; Liao et al. proposed Multi-Seq2Seq-Att for hotspot traffic speed prediction in [4], it's a multi-modal sequence learning model that deals with two sequences in different modalities, the main idea is learning to fuse the multi-modal sequence with content attention which helps to address the modality gap. Cui et al. [5] introduced the phase congruency model with illumination and contrast invariance for image matching, achieving a more robust effect on the task of automatic matching of multi-modal remote sensing images. In [6], a novel Multi-Modal Distance Metric Learning ($MD^2ML$) method is proposed, which learns a data-dependent similarity metric from multi-modal media data, aiming at assisting the brands to retrieve appropriate media data from social networks for potential customer discovery. Yang et al. [7] propose a method using multi-modal graph edge variational autoencoders to learn the latent distribution of multiple relations underlying each user link. It can be seen that due to the ability to fully explore the potential complementary relationships of different modalities and making better use of multi-modal data, multi-modal learning is widely used.

Nevertheless, in real application scenarios, the collected multi-modality data is often incomplete due to various factors. For example, in the clinic, patients often choose inspection methods with lower costs due to the economic factor, so it may not easy to collect complete multi-modal medical data. In remote sensing observations, some modalities may be missing in the data collected by the sensors due to the effect of internal and external interference. Unfortunately, the existing

multi-modal learning methods can hardly deal with these data with missing modalities, which limits the application scenarios of multi-modal learning. A naive solution is to remove the samples with missing modalities or abandon the incomplete modalities, but this method exacerbates the few-shots problem, and a large amount of valuable information is discarded. Therefore, simply removing samples with missing modalities is unreasonable.

In order to make better use of data with missing modalities, a variety of methods of modality imputation and synthesis have been developed. Some researchers proposed methods to complete missing modalities based on adversarial strategy [8–10], these methods have an outstanding performance in synthesis tasks. The method based on cascaded residual autoencoder also has been proposed in [11], the key idea is to continuously reduce the residual between synthesis data and real data by using residual autoencoder modules in cascade. To complete the missing modalities, the pivotal problem is how to effectively integrate information from existing modalities. To deal with this challenge, Zhou et al. [12] use autoencoders to extract features in levels from separate modalities and fuse them level by level into the final feature map which will be utilized to synthesis the missing modality by adversarial strategy. Peng et al. [13] proposed cross-media multiple deep network (CMDN) to hierarchically combine the inter-media and intra-media representations to further learn the rich cross-media correlation by a deeper two-level network strategy, and finally obtains a shared representation by a stacked network style. These methods showed notable results in various tasks. However, these methods mentioned above didn't fully consider the balance of consistency and complementarity between modalities. This will cause information loss of some modality or overly bias toward to some certain modality in the final synthesis modality data. As a result, a bad performance will appear in the downstream tasks. It's a challenging problem to trade off the consistency and complementarity in synthesis tasks.

To deal with the challenge mentioned above, this paper propose method based on a shared latent representation space model to synthesis missing modality. Specifically, in the training stage, the proposed method firstly obtained the shared latent representation and mapping network corresponding to existing modalities by optimizing the reconstruction loss of existing modalities, and the shared representation can naturally capture the information of existing modalities without biases. Then the synthesis network is trained to generate missing modality with the shared representation as input, and in which the convolutional neural network is implemented to extract deep level information to improve the synthesis performance. During the testing stage, shared representation among existing modalities is firstly acquired by minimizing reconstruction loss with mapping networks fixed, then the final synthesis modality is obtained by inputting the shared representation into the synthesis network. In addition, to tackle the over-optimizing problem during the testing stage, "threshold-loss" is proposed. The effectiveness of the proposed method has been verified by experiments.

The main contributions of this paper are as follows:

- We present a missing modality synthesis framework based on a shared latent representation space model which can adaptively tradeoff the consistency and complementarity among existing modalities.
- To deal with the problem of over-optimizing, "threshold-loss" is proposed.
- With extensive experiments, the effectiveness of our method has been verified.

The rest of this paper will proceed as follows. In Section II we discuss related works. Our method will be described in Section III, and be followed with experiments in Section IV. Then we conclude in Section V.

## II. Related Works

### A. Multi-modal Learning

Due to the wide scenarios of multi-modal learning in real-world applications, researchers have conducted extensive research on multi-modal learning. A classic way based on canonical correlation analysis (CCA) [14] is to project different modalities into a common subspace that can maximize the correlation between multiple modalities, and in this way, the consistent components among multiple modalities are extracted. The representative methods include kernelized canonical correlation analysis (KCCA) [15], deep canonical correlation analysis (DCCA) [16], deep variational canonical correlation analysis (DVCCA) [17], etc. Multiple kernel learning (MKL) [18] exploits kernels that naturally correspond to different views and combine kernels either linearly or non-linearly to improve learning performance. A promising approach is latent multi-view subspace clustering (LMSC) [19], which seeks the underlying latent representation and simultaneously performs data reconstruction based on the learned latent representation.

### B. Missing Modality Synthesis

As mentioned above, the problem of missing modalities is very common in real-world applications, to deal with this problem, some promising methods have been proposed. Adversarial strategy-based methods show potential in synthesis tasks. For example, in [9] a novel approach for view imputation via generative adversarial networks (VIGAN) is proposed, which combines GAN and DAE to enable the knowledge integration for domain mappings and view correspondences to effectively recover the missing view. Cai et al. proposed to take the existing modality as input and generates the missing modality by employing auxiliary adversarial loss to generate high-quality missing modality images in [8]. Different from these adversarial strategy-based methods, work in [11] proposed to reduce the residual between synthesis modality and missing modality with cascaded residual autoencoder module, which also shows considerable performance.

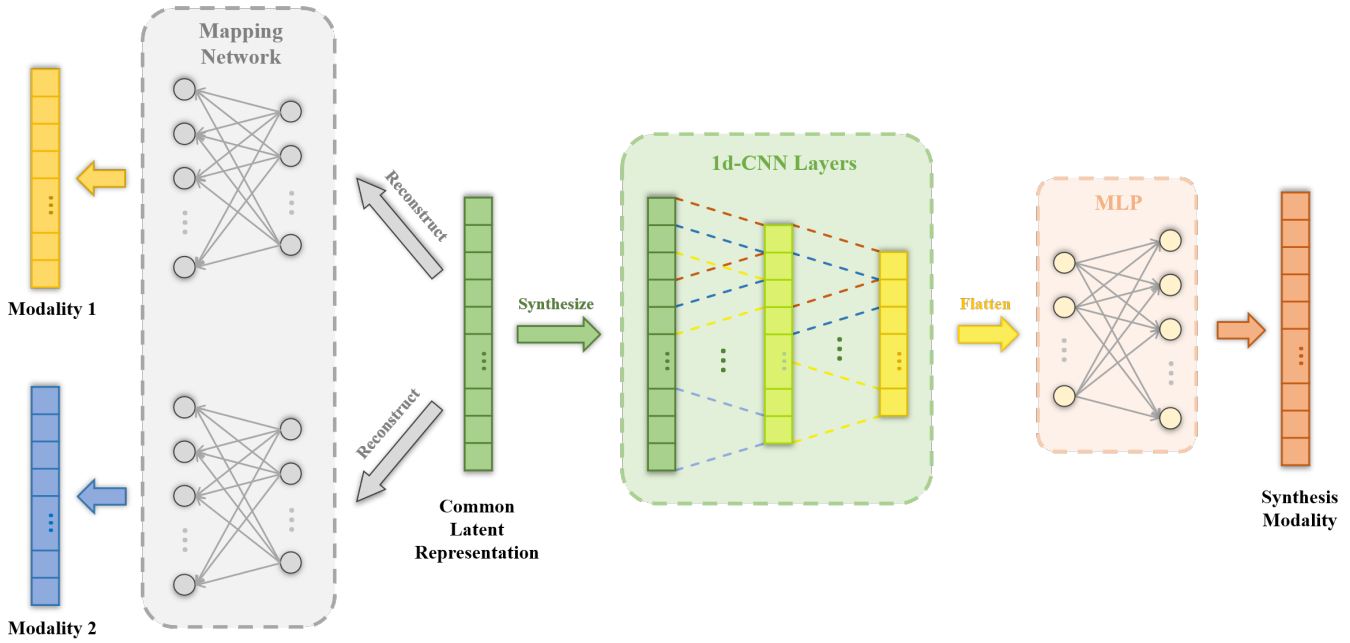


Fig. 1. The framework of the proposed method for missing modality synthesis. There are two main components in the proposed framework, including the mapping network and synthesis network. The mapping network fuses the existing modalities by reconstructing them with common latent representation as input, and 1d-CNN layers and MLP are employed in the synthesis network to synthesize the missing modality.

## III. Proposed Method

In this section, the details of the proposed method will be given. There are two main parts in the proposed model, namely mapping network and synthesis network, which will be described in detail in III-B and III-C, respectively.

### A. Multi-modal Latent Representation Model

Inspired by [19], this paper proposes to apply the multi-modal latent representation model into the missing modality synthesis task, which can naturally balance the consistency and complementary information among the existing modalities in the synthesis modality. The basic assumption is that existing modalities are mapped from a common latent representation in the common latent space, and each modality contains information from an individual perspective. Formally, given $N$ samples $\left\{\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(M)}\right\}_{i=1}^{N}$ with $M$ existing modalities, we need to find the missing modality $\mathbf{x}_i^{(M+1)}$ in each sample. Consider the assumption mentioned above, given $N$ corresponding latent representation vectors $\{\mathbf{h}_i\}_{i=1}^{N}$, and mapping network $f_{\theta_m}(\cdot)$ corresponding to modality existing $m$, then we have

$$\mathbf{x}_i^{(m)} = f_{\theta_m}(\mathbf{h}_i) + e_i^{(m)} \tag{1}$$

where $e_i^{(m)}$ denotes the reconstruction error corresponding to existing modality $m$ in sample $i$.

Our goal is to find parameter $\theta_m$ and latent representations $\{\mathbf{h}_i\}_{i=1}^{N}$ to minimize the reconstruction error $e_i^{(m)}$, which mean the information in existing modalities are adaptively fused into the latent representation. Accordingly, the objective function is

$$\begin{aligned} &\min_{\{\mathbf{h}_i\}_{i=1}^{N}, \{\theta_m\}_{m=1}^{M}} \mathcal{L}_r(\mathbf{X}, \hat{\mathbf{X}}), \\ with \quad &\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{(1)} & \cdots & \mathbf{x}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^{(M)} & \cdots & \mathbf{x}_N^{(M)} \end{bmatrix} \\ and \quad &\hat{\mathbf{X}} = \begin{bmatrix} f_{\theta_1}(\mathbf{h}_1) & \cdots & f_{\theta_1}(\mathbf{h}_N) \\ \vdots & \ddots & \vdots \\ f_{\theta_M}(\mathbf{h}_1) & \cdots & f_{\theta_M}(\mathbf{h}_N) \end{bmatrix}, \end{aligned} \tag{2}$$

where $\mathbf{h}_i$ is randomly initialized. By solving (2), the consistent and complementary information in existing modalities are naturally fused into latent representation. With the latent representation $\{\mathbf{h}_i\}_{i=1}^{N}$ obtained, we can model the synthesis process from the fused information to missing modality.

### B. Synthesis Network

In order to synthesize the missing modality, the synthesis network will be trained with latent representation as input, and missing modality as output. Due to the powerful ability of feature extraction of convolutional neural network (CNN), here we use a three-layer 1d-CNN network to extract high-level features from the latent representation vector. After that, multi-layer perception network (MLP) with the extracted feature as input is employed to synthesize the missing modality. The activation function here we use is the rectified linear unit (ReLU) function. The detailed structure is illustrated in Fig. 1. The objective function is

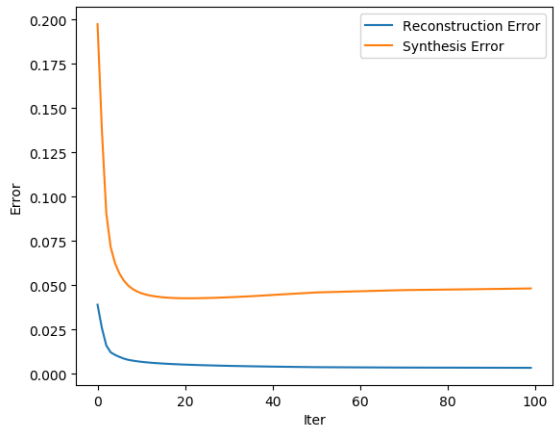


Fig. 2. The reconstruction error and synthesis error during the testing stage of the proposed method without threshold. It shows that as the reconstruction error decreasing, the synthesis error will first decrease to the lowest value, and then increase and stabilize at a certain value.

$$\begin{aligned} &\min_{\phi} \mathcal{L}_s(\mathbf{X}^{(M+1)}, \hat{\mathbf{X}}^{(M+1)}), \\ &with \quad \mathbf{X}^{(M+1)} = \begin{bmatrix} \mathbf{x}_1^{(M+1)} & \cdots & \mathbf{x}_N^{(M+1)} \end{bmatrix} \\ &and \quad \hat{\mathbf{X}}^{(M+1)} = \begin{bmatrix} g_\phi(\mathbf{h}_1) & \cdots & g_\phi(\mathbf{h}_N) \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathbf{x}_i^{(M+1)}$ is the ground-truth of the missing modality of sample $i$, $g_\phi(\cdot)$ denotes the synthesis network which contains 1d-CNN layers and MLP layers, and $\phi$ is the parameter of synthesis network. By minimizing (3), the synthesis network is obtained.

### C. Testing Stage

During the testing stage, we have $N_{test}$ testing samples $\left\{\mathbf{x}_j^{(1)}, \mathbf{x}_j^{(2)}, \ldots, \mathbf{x}_j^{(M)}\right\}_{j=1}^{N_{test}}$ with $M$ existing modalities, our goal is to obtain the common latent representations $\{\mathbf{h}_j\}_{j=1}^{N_{test}}$ of testing samples to fuse the information of the existing modalities. The basic idea is fixing the mapping network parameter and minimizing the reconstruction error by solving the objective function

$$\begin{aligned} &\min_{\{\mathbf{h}_j\}_{j=1}^{N_{test}}} \mathcal{L}_r(\mathbf{X}_{test}, \hat{\mathbf{X}}_{test}), \\ &with \quad \mathbf{X}_{test} = \begin{bmatrix} \mathbf{x}_1^{(1)} & \cdots & \mathbf{x}_{N_{test}}^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^{(M)} & \cdots & \mathbf{x}_{N_{test}}^{(M)} \end{bmatrix} \\ &and \quad \hat{\mathbf{X}}_{test} = \begin{bmatrix} f_{\theta_1}(\mathbf{h}_1) & \cdots & f_{\theta_1}(\mathbf{h}_{N_{test}}) \\ \vdots & \ddots & \vdots \\ f_{\theta_M}(\mathbf{h}_1) & \cdots & f_{\theta_M}(\mathbf{h}_{N_{test}}) \end{bmatrix}. \end{aligned} \tag{4}$$

Once the latent representations $\{\mathbf{h}_j\}_{j=1}^{N_{test}}$ obtained, the missing modality can be synthesized by inputting the latent representation to the synthesis network. However, notice that due to the difference between the objective function of the training stage and the testing stage, there is a gap between the latent representations of the two stages. Specifically, there is an over-optimizing phenomenon in the latent representations of the testing stage. In other words, as the reconstruction error decreasing, the synthesis error will first decrease to the lowest value, then increase and stabilize at a certain value, as Fig. 2 illustrated. That is, the best latent representations are not obtained. Unfortunately, the ground truth of missing modality doesn't exist in real-world applications, so we cannot simply save the best latent representations according to the lowest synthesis error. To tackle this problem, we propose a "threshold-loss" and the final objective function (4) becomes

$$\begin{aligned} &\min_{\{\mathbf{h}_j\}_{j=1}^{N_{test}}} \mathcal{L}_{threshold}(\mathbf{X}_{test}, \hat{\mathbf{X}}_{test}) \\ &\qquad = \left| \mathcal{L}_r(\mathbf{X}_{test}, \hat{\mathbf{X}}_{test}) - \alpha \right|, \\ &with \quad \mathbf{X}_{test} = \begin{bmatrix} \mathbf{x}_1^{(1)} & \cdots & \mathbf{x}_{N_{test}}^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^{(M)} & \cdots & \mathbf{x}_{N_{test}}^{(M)} \end{bmatrix} \\ &and \quad \hat{\mathbf{X}}_{test} = \begin{bmatrix} f_{\theta_1}(\mathbf{h}_1) & \cdots & f_{\theta_1}(\mathbf{h}_{N_{test}}) \\ \vdots & \ddots & \vdots \\ f_{\theta_M}(\mathbf{h}_1) & \cdots & f_{\theta_M}(\mathbf{h}_{N_{test}}) \end{bmatrix}, \end{aligned} \tag{5}$$

where $\alpha$ is a hyper-parameter. In this way, as the optimization progresses, the reconstruction error will stop decrease at a threshold, which means over-optimizing phenomenon is alleviated.

## IV. Experiments and Results

In this section, our experimental settings are described, including the detail of dataset, comparison methods, evaluation metrics, and implementation details. The results are given in IV-B.

### A. Experimental Settings

Our experiments are conducted on **Handwritten**[1] dataset. This dataset consists of features of 10 handwritten numerals ('0' to '9') extracted from a collection of Dutch utility maps. 200 samples per class (for a total of 2,000 samples) have been digitized in binary images. These digits are represented in terms of the following six feature sets (files):

- mfeat-fou: 76 Fourier coefficients of the character shapes.
- mfeat-fac: 216 profile correlations.
- mfeat-kar: 64 Karhunen-Love coefficients.
- mfeat-pix: 240 pixel averages in 2 × 3 windows.
- mfeat-zer: 47 Zernike moments.
- mfeat-mor: 6 morphological features.

In our experiments, we choose "mfeat-pix" as the missing modality to be synthesized since this modality can be visualized, and "mfeat-kar" and "mfeat-fou" as existing modality

[1]https://archive.ics.uci.edu/ml/datasets/Multiple+Features

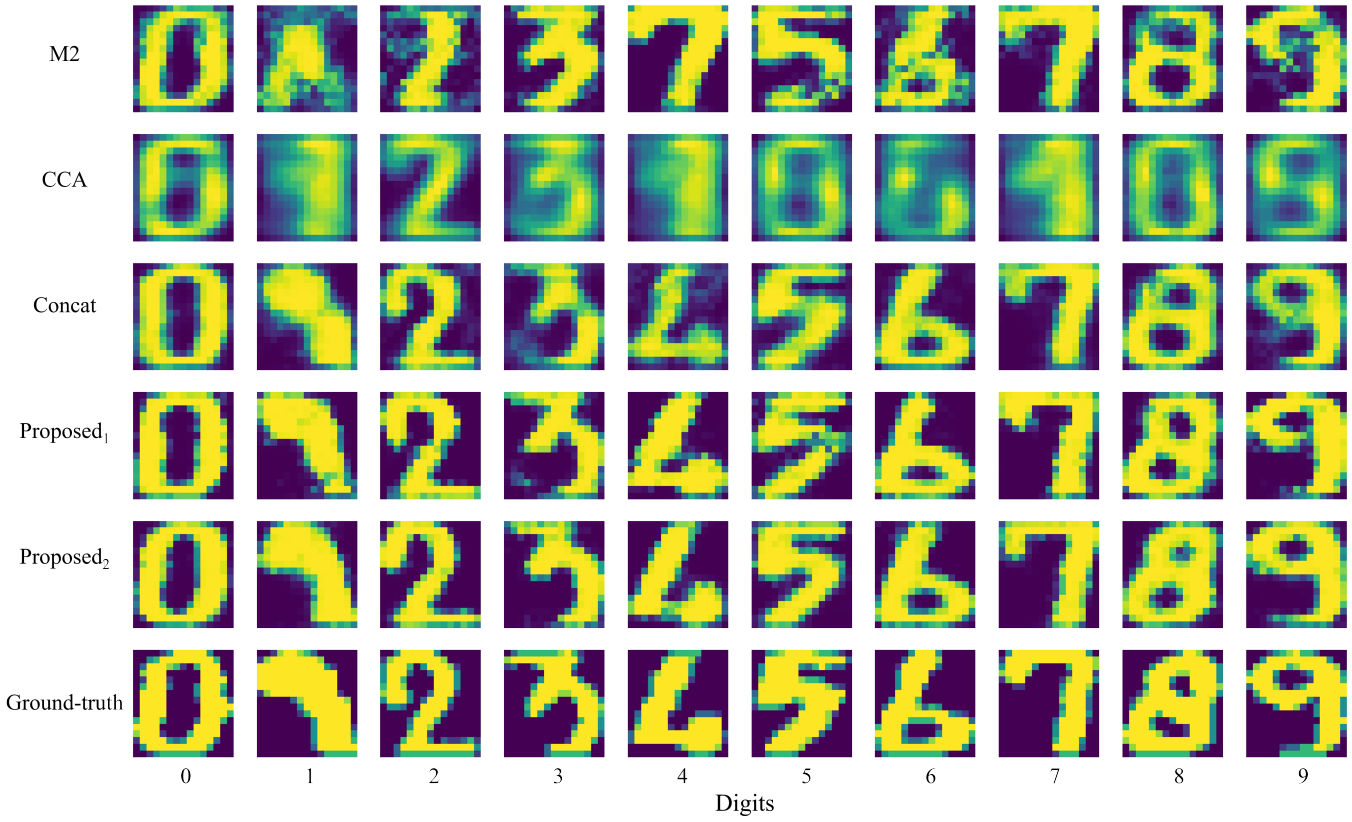


Fig. 3. All results of the proposed method and comparison methods. The "M2" means "Modality 2", and "Concat" means "Concatenate". The "Proposed$_1$" denotes the proposed method without threshold, and the "Proposed$_2$" is opposite. Intuitively, the performance of the proposed method with threshold is the best. All results are from the same testing sample.

TABLE I
THE EXPERIMENTS RESULTS OF ALL COMPARISON METHODS (MEAN ± STANDARD DEVIATION). RMSE AND PSNR ARE EMPLOYED TO EVALUATE THE PERFORMANCE OF PROPOSED METHOD, AND THE PROPOSED METHOD WITH THRESHOLD OUTPERFORM ALL OTHER METHODS.

| **Methods** | **M2** | **CCA** | **Concat** | **Proposed$_1$** | **Proposed$_2$** |
|---|---|---|---|---|---|
| **RMSE**↓ | 0.385±0.009 | 0.367±0.001 | 0.219±0.003 | 0.219±0.004 | **0.204±0.003** |
| **PSNR**↑ | 8.293±0.205 | 8.710±0.006 | 13.210±0.129 | 13.244±0.151 | **13.790±0.132** |

1 and modality 2, respectively. 1,600 samples are randomly selected for training, 400 samples for testing.

We compared proposed method with following methods:

- **Modality 2**: Only modality 2 is used to synthesize the missing modality.
- **CCA**[14]: Project existing modalities into common low-dimensional subspace, and concatenate the low-dimensional feature vectors to synthesize the missing modality.
- **Concatenate**: Just simply concatenate the existing modalities without any fusion, then synthesize the missing modality with the concatenated feature vector.

The details of implements of the proposed method are as follows: the dimension of latent representation vector is 256, and we simply model the mapping network as four-layers MLP with three hidden layers where 512, 256, and 128 hidden units are employed, respectively. In the synthesis networks, we employed three 1d-CNN layers with kernel size equals 3 and stride equals 1, and the channels number are 128, 128, and 128, respectively. ReLU activation function and max-pooling sub-sampling operation are employed after each 1d-CNN layer. After the 1d-CNN layer, a two-layers MLP with 4096 hidden units is employed to synthesize the missing modality.

The evaluation metrics used in our experiments are Root Mean Square Error (**RMSE**) and Peak Signal-to-Noise Ratio (**PSNR**). Given output value $\hat{x}_i$ and ground-truth $x_i$ of sample $i$, RMSE is calculated with

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{6}$$

where $N$ is the number of samples. And PSNR is defined as

$$PSNR = 20 \cdot \lg \frac{MAX}{RMSE} \tag{7}$$

where $MAX$ denotes the maximum value of output, and it equals 1 since we normalize the output of the synthesis network. Note that a lower RMSE, a higher PSNR mean a high quality of synthesis modality.

### B. Results

We evaluated the proposed method by comparing it with the methods mentioned above, the results are shown in Table I. The "M2" in column 2 means "Modality 2", and "Concat" in column 3 means "Concatenate". The "Proposed$_1$" denotes the proposed method without threshold, and the "Proposed$_2$" is opposite. There are two evaluation metrics in our experiments, and from Table I we can see that the proposed method outperforms all compared methods in all two metrics. All methods are run 10 times and the mean values and standard deviations are reported.

Firstly, the proposed method is far superior to the result of using only one existing modality for the synthesis task, which means that it is meaningful to fuse multi-modal information in the missing modality synthesis task, and the proposed method is effective on this. It can be seen from column 2 of Table I that the proposed method performs better than CCA, this suggests that it is not enough to only use the consistency information in the existing modalities, the useful information in the existing modalities is seriously lost in CCA since it reduces the dimensions of data to 1. Compared with simply concatenating the existing modalities, the proposed method also performs better since the concatenating operation cannot well balance the consistency and complementary information of the existing modalities, and the proposed method can handle this problem well. In addition, compared with the proposed method without the threshold, the method with the threshold performs better, which shows that the proposed threshold-loss can effectively alleviate the over-optimizing phenomenon during the test stage. In order to qualitatively compare the performance of the proposed method with other methods, all results of synthesis are shown in Fig. 3.

## V. Conclusion

In this work, we propose a novel method to synthesize the missing modality. Specifically, we utilize the common latent representation space model to adaptively fuse the consistent and complementary information in existing modalities, and then a synthesis network with 1d-CNN layers and MLP is employed to synthesize the missing modality. The experimental results suggest that our method outperforms other methods.

## References


[1] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls. Multimodal classification of remote sensing images: a review and future directions. Proceedings of the IEEE,103(9):1560–1584, 2015.

[2] Audebert N , Saux B L , Lefevre S . Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2017.

[3] Yan H , Li Z . A Multi-modal Medical Image Fusion Method in Spatial Domain[C]// 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2019.

[4] Liao B , Tang S , Yang S , et al. Multi-modal Sequence to Sequence Learning with Content Attention for Hotspot Traffic Speed Prediction[J]. Pacific Rim Conference on Multimedia, 2018.

[5] Song C , Zhong Y . Multi-Modal Remote Sensing Image Registration Based on Multi-Scale Phase Congruency[C]// 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS). 2018.

[6] Yang L , Gu Z , Ko T H , et al. Multi-Modal Media Retrieval via Distance Metric Learning for Potential Customer Discovery[C]// 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). ACM, 2018.

[7] Yang C , Zhang J , Wang H , et al. Relation Learning on Social Networks with Multi-Modal Graph Edge Variational Autoencoders[C]// Wsdm 20: the Thirteenth Acm International Conference on Web Search & Data Mining. ACM, 2020.

[8] Lei Cai,Zhengyang Wang,Hongyang Gao,Dinggang Shen,Shuiwang Ji. Deep Adversarial Learning for Multi-Modality Missing Data Completion[P]. Knowledge Discovery & Data Mining,2018.

[9] Shang C , Palmer A , Sun J , et al. VIGAN: Missing View Imputation with Generative Adversarial Networks[J]. 2017 IEEE International Conference on Big Data (Big Data), 2017.

[10] Dong, Nie, Roger, et al. Medical Image Synthesis with Deep Convolutional Adversarial Networks[J]. IEEE Transactions on Biomedical Engineering, 2018, 65(12):2720-2730.

[11] Tran L , Liu X , Zhou J , et al. Missing Modalities Imputation via Cascaded Residual Autoencoder[C]// Computer Vision & Pattern Recognition. IEEE, 2017:4971-4980.

[12] Zhou T , Fu H , Chen G , et al. Hi-Net: Hybrid-fusion Network for Multi-modal MR Image Synthesis[J]. IEEE Transactions on Medical Imaging, 2020.

[13] Peng Y , Xin H , Qi J . Cross-media shared representation by hierarchical learning with multiple deep networks. 2016.

[14] Hardoon D R , Szedmak S , Shawe-Taylor J . Canonical Correlation Analysis: An Overview with Application to Learning Methods[J]. Neural Computation, 2004, 16(12):2639-2664.

[15] Akaho S . A kernel method for canonical correlation analysis[J]. 2006.

[16] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in ICML, 2013, pp. 1247–1255.

[17] Wang W , Lee H , Livescu K . Deep Variational Canonical Correlation Analysis[J]. 2016.

[18] Lin Y Y , Liu T L , Fuh C S . Multiple Kernel Learning for Dimensionality Reduction[J]. IEEE Trans-

actions on Pattern Analysis & Machine Intelligence, 2011, 33(6):1147-1160.

[19] Zhang C , Hu Q , Fu H , et al. Latent Multi-view Subspace Clustering[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.