Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities

Jinming Zhao[*]
School of Information
Renmin University of China
zhaojinming@ruc.edu.cn

Ruichen Li[*]
School of Information
Renmin University of China
ruichen@ruc.edu.cn

Qin Jin[†]
School of Information
Renmin University of China
qjin@ruc.edu.cn

Abstract

Multimodal fusion has been proved to improve emotion recognition performance in previous works. However, in real-world applications, we often encounter the problem of missing modality, and which modalities will be missing is uncertain. It makes the fixed multimodal fusion fail in such cases. In this work, we propose a unified model, Missing Modality Imagination Network (MMIN), to deal with the uncertain missing modality problem. MMIN learns robust joint multimodal representations, which can predict the representation of any missing modality given available modalities under different missing modality conditions. Comprehensive experiments on two benchmark datasets demonstrate that the unified MMIN model significantly improves emotion recognition performance under both uncertain missing-modality testing conditions and full-modality ideal testing condition. The code will be available at https://github.com/AIM3-RUC/MMIN.

1 Introduction

Automatic multimodal emotion recognition is very important to natural human-computer interactions (Fragopanagos and Taylor, 2002). It aims to understand and interpret human emotions expressed through multiple modalities such as speech content, voice tones and facial expression. Previous works have shown that these different modalities are complimentary for emotion expression, and proposed many effective multimodal fusion methods to improve the emotion recognition performance (Baltrušaitis et al., 2018; Tsai et al., 2019; Zhao et al., 2018). However, in real applications, many common causes can lead to the missing modality problem. For example, the camera is turned off or blocked due to privacy issues; the speech content is unavailable due to automatic speech recognition errors; the voice and text are missing due to the silence of the user; or the faces cannot be detected due to lighting or occlusion issues as shown in Figure 1. Existing multimodal fusion models trained on full-modality samples usually fail when partial modalities are missing (Aguilar et al., 2019; Pham et al., 2019; Cai et al., 2018; Parthasarathy and Sundaram, 2020).

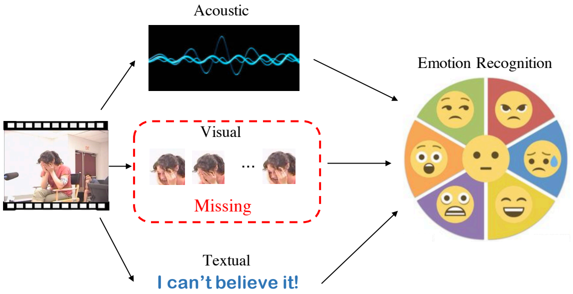


Figure 1: Illustration of a missing modality scenario for multimodal emotion recognition systems. As shown in this video segment, we encounter the missing visual modality problem due to the person's face was obscured by her hands.

The missing modality problem has attracted more research attention in the past years, and the existing solutions for this problem are mainly based on learning joint multimodal representation so that all modality information can be encoded. Han et al. (Han et al., 2019) propose a joint training approach that implicitly fuses multimodal information from auxiliary modalities, which improves the monomodal emotion recognition performance. The recent cross-modality sequential translation-based methods proposed in (Pham et al., 2019; Wang et al., 2020) learn the joint multimodal representations via translating a source modality to multiple target modalities, which improves the performance

[*] Equal Contribution
[†] Corresponding Author

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 2608–2618
August 1–6, 2021. ©2021 Association for Computational Linguistics

of the source modality as input at the test time. However, these methods can only deal with the scenario where the source modality is input to the trained model. Different models need to be built for different missing modality cases[1]. Additionally, the sequential translation-based models require translation and generation of videos, audios, and text, which are difficult to train especially with limited training samples (Li et al., 2018; Pham et al., 2019).

In this work, we propose a novel unified model, Missing Modality Imagination Network (MMIN), to address the above issues. Specifically, the proposed MMIN learns the robust joint multimodal representations through cross-modality imagination with Cascade Residual Autoencoder (CRA) (Tran et al., 2017) and Cycle Consistency Learning (Zhu et al., 2017) based on sentence-level modality-specific representations, as the sentence-level representation is more reasonable for modeling the cross-modality emotion correlation. The imagination module aims to predict the sentence-level emotional representation of the missing modality from the other available modalities. To the best of our knowledge, this is the first work that investigates a unified model for multimodal emotion recognition with uncertain missing-modality.

Extensive experiments are carried out on two benchmark datasets, IEMOCAP and MSP-IMPROV, under both uncertain missing-modality and full-modality conditions. The proposed MMIN model as a unified multimodal emotion recognition model can learn robust joint multimodal representations and outperforms the standard multimodal fusion models on both benchmark datasets under both the uncertain missing-modality and the full-modality conditions. Furthermore, to evaluate the imagination ability of our MMIN model, we visualize the distributions of the imagined representations of the missing modalities and its ground-truth representations and find they are very similar, which demonstrates that MMIN can imagine the representations of the missing modalities based on the representations of the available modalities.

In summary, the main contributions of this work are: 1) We propose a unified model, Missing Modality Imagination Network (MMIN), to improve the robustness of emotion recognition systems under uncertain missing-modality testing conditions. 2) We design cross-modality imagination based on paired multimodal data and adopt Cascade Residual Autoencoder (CRA) and Cycle Consistency Learning to learn the robust joint multimodal representations. 3) Extensive experiments on two benchmark datasets demonstrate the effectiveness of the proposed model which improves the emotion recognition performance under both the uncertain missing-modality and the full-modality conditions.

2 Related Work

Multimodal Emotion Recognition Many previous works have focused on fusing multimodal information to improve emotion recognition performance. Temporal attention-based methods are proposed to use the attention mechanism to selectively fuse different modalities based on the frame-level or word-level temporal sequence, such as Gated Multimodal Unit (GMU) (Aguilar et al., 2019), Multimodal Alignment Model (MMAN) (Xu et al., 2019) and Multi-modal Attention mechanism (cLSTM-MMA) (Pan et al., 2020). These methods use different uni-modal sub-networks to model the contextual representations for each modality and then use the multimodal attention mechanism to selectively fuse the representations of different modalities. Liang et al. (Liang et al., 2020) propose a semi-supervised multimodal (SSMM) emotion recognition model which uses cross-modality emotional distribution matching to leverages unlabeled data to learn the robust representations and achieves state-of-the-art performance.

Missing Modality Problem Existing methods for missing modality problem can mainly be divided into three groups. The first group features the data augmentation approach, which randomly ablates the inputs to mimic missing modality cases (Parthasarathy and Sundaram, 2020). The second group is based on generative methods to directly predict the missing modalities given the available modalities (Li et al., 2018; Cai et al., 2018; Suo et al., 2019; Du et al., 2018). The third group aims to learn the joint multimodal representations that can contain related information from these modalities (Aguilar et al., 2019; Pham et al., 2019; Han et al., 2019; Wang et al., 2020).

Data augmentation methods: Parthasarathy et al. (Parthasarathy and Sundaram, 2020) propose a strategy to randomly ablate visual inputs during

[1] If there are audio(a),visual(v) and textual(t) three modalities, then the system needs 6 models trained under 6 missing modality conditions {a}, {v}, {t}, {a,v}, {a,t} and {v,t}, plus one model trained under the full-modality data.

training at the clip or frame level to mimic real-world missing modality scenarios for audio-visual multimodal emotion recognition, which improves the recognition performance under missing modality conditions.

Generative methods: Tran et al. (Tran et al., 2017) propose Cascaded Residual Autoencoder (CRA) to utilize the residual mechanism over the autoencoder structure, which can take the corrupted data and estimate a function to well restore the incomplete data. Cai et al. (Cai et al., 2018) propose an encoder-decoder deep neural network to generate the missing modality (Positron Emission Tomography, PET) given the available modality (Magnetic Resonance Imaging, MRI), and the generated PET can provide complementary information to improve the detection and tracking of Alzheimers disease.

Learning joint multimodal representations: Han et al. (Han et al., 2019) propose a joint training model that consists of two modality-specific encoders and one shared classifier, which implicitly fuse the audio and visual information as joint representations and improve the performance of the mono-modality emotion recognition. Pham et al. (Pham et al., 2019) propose a sequential translation-based model to learn the joint representation between the source modality and multiple target modalities. The hidden vectors of the source modality encoder work as the joint representations, which improve the emotion recognition performance of the source modality. Wang et al. (Wang et al., 2020) follow this translation-based method and propose a more efficient transformer-based translation model with parallel translation including textual features to acoustic features and textual features to visual features. Moreover, the above two translation-based models adopt the forward translation and backward translation training strategy to ensure that joint representations can retain maximal information from all modalities.

3 Method

Given a set of video segments S, we use $x = (x^a, x^v, x^t)$ to represent the raw multimodal features for a video segment $s \in S$, where x^a, x^v and x^t represent the raw features of acoustic, visual and textual modalities respectively. $|S|$ represents the number of video segments in set S. We denote the target set $Y = \{y_i\}_{i=1}^{|S|}, y_i \in \{0, 1, \ldots, C\}$, where y_i is the target emotion category of the video segment s_i and $|C|$ is the number of emotion categories. Our proposed method aims to recognize the emotion category y_i for every video segment s_i with full modalities, or with only partial modalities available, for the example shown in Figure 1, there exist only acoustic and textual modalities when visual modality is missing.

	(available,missing)	unified triplet format pairs
1	$((x^a), (x^v, x^t))$	$((x^a, x^v_{miss}, x^t_{miss}), (x^a_{miss}, x^v, x^t))$
2	$((x^v), (x^a, x^t))$	$((x^a_{miss}, x^v, x^t_{miss}), (x^a, x^v_{miss}, x^t))$
3	$((x^t), (x^a, x^v))$	$((x^a_{miss}, x^v_{miss}, x^t), (x^a, x^v, x^t_{miss}))$
4	$((x^a, x^v), (x^t))$	$((x^a, x^v, x^t_{miss}), (x^a_{miss}, x^v_{miss}, x^t))$
5	$((x^a, x^t), (x^v))$	$((x^a, x^v_{miss}, x^t), (x^a_{miss}, x^v, x^t_{miss}))$
6	$((x^v, x^t), (x^a))$	$(x^a_{miss}, x^v, x^t), (x^a, x^v_{miss}, x^t_{miss}))$

Table 1: The six possible missing-modality conditions and their unified format cross-modality pairs.

3.1 Missing Modality Imagination Network

In order to learn robust joint multimodal representations, we propose a unified model, Missing Modality Imagination Network (MMIN), which can deal with different uncertain missing-modality conditions in real application scenarios. Figure 2 illustrates the framework of our proposed MMIN model which contains three main modules: 1) Modality Encoder Network for extracting modality-specific embeddings; 2) Imagination Module based on the Cascade Residual Autoencoder (CRA) and Cycle Consistency Learning for imagining the representations of missing modalities given the representations of the corresponding available modalities. The latent vectors of the autoencoders in CRA are collected to form the joint multimodal representations; 3) Emotion classifier for predicting the emotion category based on the joint multimodal representations. We introduce each module in details in the following subsections.

3.1.1 Modality Encoder Network

The Modality Encoder Network is used to extract the modality-specific utterance-level embeddings based on the raw modality features x. As shown in Figure 2(b), we first pretrain the Modality Encoder Network in a multimodal emotion recognition model and it is further trained within MMIN model. We define the modality-specific embeddings of each modality as $h^a = \mathrm{EncA}(x^a)$, $h^v = \mathrm{EncV}(x^v)$, $h^t = \mathrm{EncT}(x^t)$, where EncA, EncV and EncT represent the acoustic, visual and textual encoders respectively, and h^a, h^v and h^t represent the modality-specific embeddings generated by the corresponding encoders respectively.

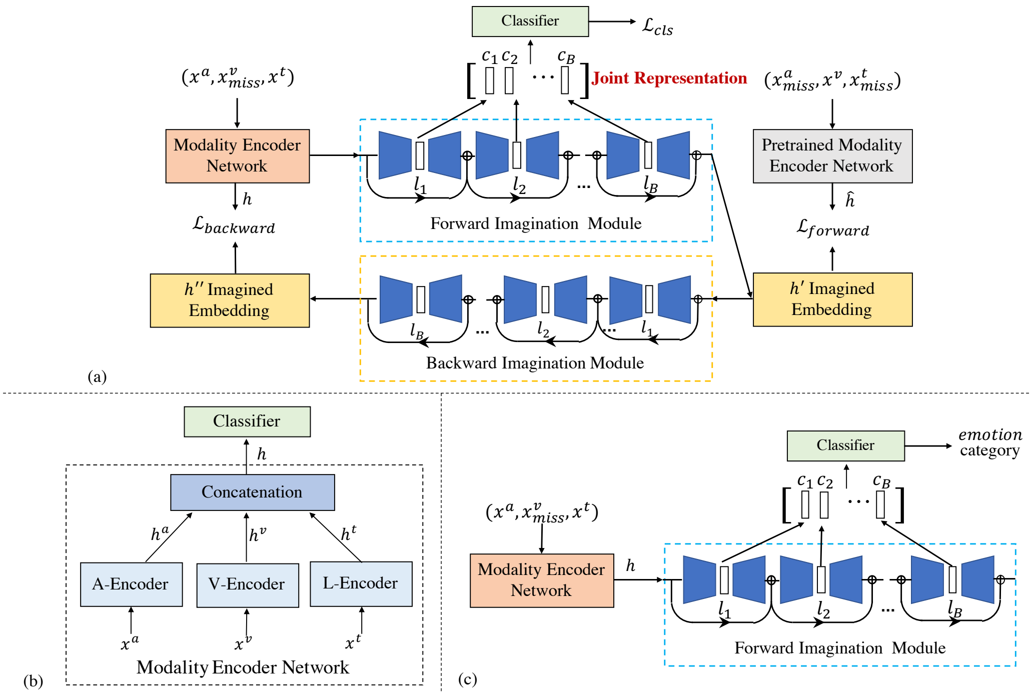


Figure 2: Illustration of the Missing Modality Imagination Network (MMIN) framework. (a) MMIN at the training stage (taking the visual modality missing condition as example). MMIN is trained with all six possible missing modality conditions (Table 1). (b) Modality Encoder Network. The modality encoder network is pretrained in the multimodal emotion recognition task on the full-modality data and then it is updated during the MMIN training as shown in the orange colored block in MMIN. The pretrained modality encoder network (gray colored block in MMIN) is similar to the modality encoder network, and the only difference is that it is fixed during training. (c) Missing Modality Imagination Network (MMIN) at the inference stage (taking the visual modality missing condition as an example). MMIN can inference under different missing modality conditions.

3.1.2 Missing Modality Condition Creation

Given a training sample with all three modalities (x^a, x^v, x^t), there are 6 different possible missing-modality conditions as shown in Table 1. We can build a cross-modality pair $(available, missing)$ under each missing-modality condition, where the $available$ and $missing$ mean the available modalities and the corresponding missing modalities respectively. In order to ensure a unified model that can handle various missing-modality conditions, we enforce a unified triplet input format for the modality encoder network as (x^a, x^v, x^t). Under the missing-modality conditions, the raw features of the corresponding missing modalities are replaced by zero vectors. For example, the unified format input of the available modalities under the visual modality missing condition (case 1 in Table 1) is formatted as (x^a, x^v_{miss}, x^t), where x^v_{miss} refers to zero vectors.

Under the missing-modality training conditions, the input includes the cross-modality pairs referring to available modalities and missing modalities in the unified triplet format (as shown in Table 1). The multimodal embeddings of these cross-modality pairs can be represented as (taking the visual modality missing condition as example):

$$
\begin{aligned}
h &= concat(h^a, h^v_{miss}, h^t) \\
\hat{h} &= concat(h^a_{miss}, h^v, h^t_{miss})
\end{aligned}
\tag{1}
$$

where h^a_{miss}, h^v_{miss} and h^t_{miss} represent the modality-specific embedding when the corresponding modality is missing, which is produced by the corresponding modality encoder with input zero vectors.

3.1.3 Imagination Module

We propose an autoencoder-based Imagination Module to predict the multimodal embeddings of the missing modalities given the multimodal embeddings of the available modalities. The Imagination Module is expected to learn the robust joint multimodal representations through the cross-modality imagination. As illustrated in Figure 2(a), we employ the Cascade Residual Autoencoder (CRA) (Tran et al., 2017) structure, which has sufficient learning capacity and more stable convergence than the standard autoencoder. The

CRA structure is constructed by connecting a series of Residual Autoencoders (RAs). We further employ cycle consistency learning (Zhu et al., 2017; Wang et al., 2020) with a coupled net architecture with two independent networks to perform imagination in two directions, including the Forward ($available \rightarrow missing$) and Backward ($missing \rightarrow available$) imagination directions.

To be specific, we use a CRA model with B RAs and each RA is represented by ϕ_k, $k = 1, 2, \ldots, B$, and the calculation of each RA can be defined as:

$$\begin{cases} \Delta z_k = \phi_k(h), \quad k = 1 \\ \Delta z_k = \phi_k(h + \sum_{j=1}^{k-1} \Delta z_j), \quad k > 1 \end{cases} \tag{2}$$

where h is the extracted multimodal embedding based on the available modalities in a unified cross-modality pair format (Eq.(1)) and Δz_k represents the output of the k^{th} RA. Taking the visual modality missing condition as example (as shown in Figure 2(a)), the forward imagination aims to predict the multimodal embedding of the missing visual modality based on the available acoustic and textual modalities. The forward imagined multimodal embedding is expressed as:

$$h^{'} = imagine_{forward}(h) = h + \sum_{k=1}^{B} \Delta z_k \tag{3}$$

where $imagine(\cdot)$ represents the function of the Imagination Module. The backward imagination aims to predict the multimodal embedding of the available modalities based on the forward imagined multimodal embedding $h^{'}$ (Eq.(3)). The backward imagined multimodal embedding is expressed as:

$$h^{''} = imagine_{backward}(h^{'}) \tag{4}$$

3.1.4 Classifier

We collect the latent vectors of each auto-encoder in the forward imagination module and concatenate them together to form the joint multimodal representation: $R = concat(c_1, c_2, \ldots, c_B)$, where c_k is the latent vector of the autoencoder in the k^{th} RA. Based on the joint multimodal representation R, we calculate the probability distribution q as:

$$q = softmax(f_{cls}(R)) \tag{5}$$

where $f_{cls}(\cdot)$ denotes the emotion classifier that consists of several fully-connected layers.

3.2 Joint Optimization

The loss function for MMIN training includes three parts: the emotion recognition loss $\mathcal{L}_{cls}$, forward imagination loss $\mathcal{L}_{forward}$, and backward imagination loss $\mathcal{L}_{backward}$:

$$\begin{aligned} \mathcal{L}_{cls} &= -\frac{1}{|S|} \sum_{i=1}^{|S|} H(p, q) \\ \mathcal{L}_{forward} &= \frac{1}{|S|} \sum_{i=1}^{|S|} \left\| \hat{h}_i - h_i^{'} \right\|_2^2 \\ \mathcal{L}_{backward} &= \frac{1}{|S|} \sum_{i=1}^{|S|} \left\| h_i - h_i^{''} \right\|_2^2 \end{aligned} \tag{6}$$

where p is the true distribution of one-hot label and q is the prediction distribution calculated in Eq.(5). $H(p, q)$ is the cross-entropy between distributions p and q. h_i and $\hat{h_i}$ are the ground-truth representations extracted by the modality encoder network as shown in Eq.(1). We combine all the three losses into the joint objective function as below to jointly optimize the model parameters:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{forward} + \lambda_2 \mathcal{L}_{backward} \tag{7}$$

where λ_1 and λ_2 are weighting hyper parameters for $\mathcal{L}_{forward}$ and $\mathcal{L}_{backward}$ respectively.

4 Experiments

4.1 Dataset

We evaluate our proposed model on two benchmark multimodal emotion recognition datasets, Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008) and MSP-IMPROV (Busso et al., 2016). The statistics of the two datasets are shown in Table 2.

IEMOCAP contains recorded videos in 5 dyadic conversation sessions. In each session, there are multiple scripted plays and spontaneous dialogues between a male and a female speaker and 10 speakers in total in the database. We follow the emotional label processing in (Xu et al., 2019; Liang et al., 2020) to form the four-class emotion recognition setup.

MSP-IMPROV contains recorded segments videos in dyadic conversation scenarios with 12 actors. We first remove videos that are shorter than 1 second. Then we select the videos in the "Other-improvised" group which are recorded during the improvisation scenarios with *happy*, *anger*, *sadness*, or *neutral* labels to form the four-class emotion recognition setup.

dataset	Happy	Anger	Sadness	Neutral	Total
IECMOAP	1636	1103	1084	1708	5531
MSP-IMPROV	999	460	627	1733	3819

Table 2: Data Statistics of datasets

4.1.1 Missing-Modality Training Set

We first define the original training set which contains all the three modalities as the full-modality training set. Based on the full-modality training set, we construct another training set that contains cross-modality pairs to simulate the possible missing-modality conditions and we define it as the missing-modality training set, which we use to train the proposed MMIN. Six different cross-modality pairs (Table 1) for each training sample are generated. Therefore, the number of the generated cross-modality pairs is six times as large as the number of the full-modality training samples.

4.1.2 Missing-Modality Testing Set

We first define the original testing set which contains all the three modalities as the full-modality testing set. To evaluate the performance of the proposed MMIN under the uncertain missing-modality conditions, we construct six different missing modality testing subsets corresponding to the six possible missing modality conditions respectively. For example, in the inference stage, under the missing visual modality condition as shown in Figure 2(c), the raw feature of a missing-modality testing sample in the unified format is (x^a, x^v_{miss}, x^t). We combine all the six missing-modality testing subsets together and denote it as the missing-modality testing set.

4.2 Raw Feature Extraction

We follow feature extraction methods described in (Liang et al., 2020; Pan et al., 2020) and extract the frame-level raw features of each modality [2].

Acoustic features: OpenSMILE toolkit (Eyben et al., 2010) with the configuration of "IS13_ComParE" is used to extract frame-level features, which have similar performance with the IS10 utterance-level acoustic features used in (Liang et al., 2020). We denote the features as "ComParE" and the feature vectors are in 130 dimensions.

Visual features: We extract the facial expression features using a pretrained DenseNet (Huang et al., 2017) which is trained based on the Facial Expression Recognition Plus (FER+) corpus (Barsoum et al., 2016). We denote the facial expression features as "Denseface". The "Denseface" are frame-level sequential features based on the detected faces from the video frames, and the feature vectors are in 342 dimensions.

Textual features: We extract contextual word embeddings using a pretrained BERT-large model (Devlin et al., 2019) which is one of the state-of-the-art language representations. We denote the word embeddings as "Bert" and the features are in 1024 dimensions.

4.3 Higher-level Feature Encoder

To generate more efficient sentence-level modality-specific representations for the Imagination Module, we design different modality encoders for different modalities.

Acoustic Modality Encoder (EncA)**:** We apply a Long Short-term Memory (LSTM) network (Sak et al., 2014) to capture the temporal information based on the sequential frame-level raw acoustic features x^a. Then we use max-pooling to get utterance-level acoustic embedding h^a based on the LSTM hidden states.

Visual Modality Encoder (EncV)**:** We adopt a similar method with $EncA$ on the sequential frame-level facial expression features x^v and get utterance-level visual embedding h^v.

Textual Modality Encoder (EncT)**:** We apply a TextCNN (Kim, 2014) to get the utterance-level textual embedding as h^t based on the sequential word-level features x^t.

4.4 Recognition Baselines

Our baseline model takes the structure as shown in Figure 2(b), which is trained based on the full-modality training set and we use it as our **full-modality baseline**. To improve the system robustness against the missing modality problem, one intuitive solution is to add samples under the missing-modality conditions into the training set. We, therefore, pool the missing-modality training set and full-modality training set together to train the baseline model and use it as our **augmented baseline**.

4.5 Implementation Details

Table 3 presents our implementation details. We use the 10-fold and 12-fold speaker-independent cross-validation to evaluate the models on IEMOCAP and MSP-IMPROV respectively. For the experiments on IEMOCAP, we take four sessions for

[2] To facilitate fair comparison with the sequential translation-based missing modality method MCTN, we adopt frame-level features which can be directly used in the MCTN method

Acoustic Encoder	single layer LSTM with hidden size of 128
Visual Encoder	single layer LSTM with hidden size of 128
Textual Encoder	3 Conv blocks in TextCNN with kernel size {3,4,5} and output layer with 128 channels
Emotion Classifier	3 FC layers of size {128,64,4}
CRA	5 residual-RAs with RA-layers in size 384-256-128-64-128-256-384 (latent-vector size: 64)
parameters λ_1, λ_2	both set as 0.1
Learning rate	Adam optimizer with learning rate of 0.001, ReLU activation

Table 3: Implementation Details

	train	test	WA	UA
Our full-modality baseline	$\{a, v, t\}$	$\{a, v, t\}$	**0.7651**	**0.7779**
cLSTM-MMA(Pan et al., 2020)			0.7394	–
SSMM(Liang et al., 2020)			0.7560	0.7450

Table 4: Multimodal Emotion Recognition Results on IEMOCAP under full-modality condition.

training, and the remaining session is split by speakers into the validation and testing sets. For MSP-IMPROV, we take the utterances of 10 speakers for training, the remaining 2 speakers are divided into validation set and testing set by speakers. We train the model with at most 100 epochs for each experiment. We select the best model on the validation set and report its performance on the testing set. To demonstrate the robustness of our models, we run each model three times to alleviate the influences of random initialization of parameters and apply a significance test for model comparison. All models are implemented with Pytorch deep learning toolkit and run on a single Nvidia GTX 1080Ti graphic card.

For the experiments on IEMOCAP, we use two evaluation metrics: weighted accuracy (WA) and unweighted accuracy (UA). Due to the imbalance of emotion categories on MSP-IMPROV, we use the f-score as the evaluation metric.

4.6 Full-modality Baseline Results

We first compare our **full-modality baseline** with several state-of-the-art multimodal recognition models under the full-modality condition. Results in Table 4 show that our full-modality baseline outperforms other state-of-the-art models, which proves that our modality encoder network can extract effective representations for multimodal emotion recognition.

4.7 Uncertain Missing-Modality Results

Table 5 presents the experimental results of our proposed MMIN model under different missing-modality testing conditions and full-modality testing condition. On IEMOCAP, comparing to the "full-modality baseline" results in Table 4, we see a significant performance drop under uncertain missing-modality testing conditions, which indicates that the model trained under the full-modality condition is very sensitive to the missing modality problem. The intuitive solution "Augmented baseline", which combines the missing-modality training set with the full-modality training set to train the baseline model, does significantly improves over the full-modality baseline under missing-modality testing conditions, which indicates that data augmentation can help alleviate the problem of data mismatch between training and testing. More notably, our proposed MMIN significantly outperforms both the full-modality baseline and the augmented baseline under every possible missing-modality testing condition. It also outperforms the two baselines under the full-modality testing condition, even though the MMIN model does not use the full-modality training data. These results indicate that our proposed MMIN model can learn robust joint multimodal representation so that it can achieve consistently better performance under both the different missing-modality and the full-modality testing conditions. This is because our proposed MMIN method not only has the data augmentation capability, but also can learn better joint representation, which can preserve information of other modalities.

We further analyze the performance under different missing modality conditions. Our MMIN model achieves significant improvement under one modality available conditions ($\{a\}$, $\{v\}$, or $\{t\}$) compared with the augmented baseline, especially for the weak modalities $\{a\}$ and $\{v\}$. It brings some improvements as well over the augmented baseline even for the strong modality combinations, such as $\{a, t\}$. These experimental results indicate that the learned joint representation via MMIN did learn complementary info from the other modalities to compensate for the weak modalities.

The bottom block in Table 5 shows the performance comparison on the MSP-IMPROV dataset. Our proposed MMIN model again significantly outperforms the two baselines under different missing-modality and full-modality testing conditions, which demonstrates the good generalization ability of MMIN across different datasets.

We also compare to the MCTN (Pham et al., 2019) model which is the state-of-the-art model for the missing modality problem. As MCTN can-

Dateset	Model	Metric	Testing Condition							
			$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
IEMOCAP	Full-modality baseline	WA(↑)	0.4190	0.4574	0.5646	0.5488	0.7018	0.6217	0.5522	0.7651
		UA(↑)	0.4719	0.3966	0.5549	0.5762	0.7257	0.5971	0.5537	0.7779
	Augmented baseline	WA(↑)	0.5303	0.4864	0.6564	0.6395	0.7251	0.7082	0.6243$*$	0.7617
		UA(↑)	0.5440	0.4598	0.6691	0.6434	0.7435	0.7162	0.6293$*$	0.7767
	proposed MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410$*$▲**	0.7650
		UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524$*$▲**	**0.7812$*$▲**
	MCTN (Pham et al., 2019)	WA(↑)	0.4975	0.4892	0.6242	0.5634	0.6834	0.6784	0.5894$*$	–
		UA(↑)	0.5162	0.4573	0.6378	0.5584	0.6946	0.6834	0.5913$*$	–
MSP-IMPROV	Full-modality baseline	F1(↑)	0.2824	0.3295	0.4576	0.4721	0.5655	0.5368	0.4543	0.6523
	Augmented baseline	F1(↑)	0.4278	0.4185	0.5544	0.5396	0.6038	0.6295	**0.5455$*$**	**0.6663$*$**
	proposed MMIN	F1(↑)	**0.4647**	**0.4471**	**0.5573**	**0.5740**	**0.6188**	**0.6411**	**0.5649$*$▲**	**0.6855$*$▲**
	MCTN (Pham et al., 2019)	F1(↑)	0.3285	0.3810	0.5050	0.4683	0.5611	0.5886	0.4721$*$	–

Table 5: Performance comparison under six possible missing-modality testing conditions and the full-modality testing condition (i.e. testing condition "$\{a\}$" means that only the acoustic modality is available and both visual and textual modalities are missing. "$\{a, v, t\}$" refers to the full-modality testing condition where all acoustic, visual and textual modalities are available) "Average" refers to the average performance over all six missing-modality conditions. T-test is conducted on Average and {a,v,t} column. $*$ indicates that *p-value* < 0.05 (compared with Full-modality baseline). ▲ indicates that *p-value* < 0.05 (compared with Augmented baseline).

not handle different missing-modality conditions in one unified model, so we have to train a particular model under each missing-modality condition[3]. The comparison results demonstrate that our proposed MMIN model not only can handle both the different missing-modality and the full-modality testing condition with a unified model, but also can consistently outperform the MCTN models under all missing-modality conditions.

4.8 Ablation Study

We conduct experiments to ablate the contributions of different components in MMIN, including the structure of the imagination module and the cyclic consistency learning.

Structure of the imagination module. We first investigate the impact of different network structures on the performance in the imagination module. Specifically, we compare the Autoencoder and the CRA structure in MMIN, and we adopt the same parameter scale to ensure the fairness of the comparison. As shown in Table 6, the performance of the imagination module with Autoencoder structure "MMIN-AE" is worse than that with the CRA structure under both different missing-modality and full-modality testing conditions. The performance comparison indicates that the CRA has a stronger imagination ability than the Autoencoder model.

Cycle Consistency Learning. To evaluate the impact of the cyclic consistency learning in MMIN, we conduct experiments using MMIN with or without cycle consistency learning. As shown in Table 6, the model trained without cycle consistency learning results in performance loss under all conditions, which indicates that the cycle consistency learning can enhance the imagination ability and learn more robust joint multimodal representations.

4.9 Analysis of MMIN Core Competence

We conduct detailed experiments on IEMOCAP to demonstrate the joint representation learning ability and the imagination ability of our MMIN model.

Joint representation learning ability: Since the joint representation is expected to retain information of multiple modalities, we conduct experiments to evaluate the joint representation learning ability of MMIN. We compare MMIN to the baseline model under the matched-modality condition in which the training data and the test data contain the same modalities. As shown in Table 7, comparing to the baseline model, MMIN achieves on par with or even better performance, which demonstrates that MMIN has the ability to learn effective joint multimodal representations. We also notice that the data-augmented model cannot beat the corresponding matching partial-modality baseline model, which indicates the data-augmented model cannot learn the joint representation.

Imagination ability: Figure 3 visualizes the distribution of the ground-truth multimodal embeddings ($\hat{h}$ in Figure 2) and MMIN imagined multimodal embeddings ($h^{'}$ in Figure 2) for a male speaker and female speaker using t-SNE (Maaten and Hinton, 2008). We observe that the distribution of

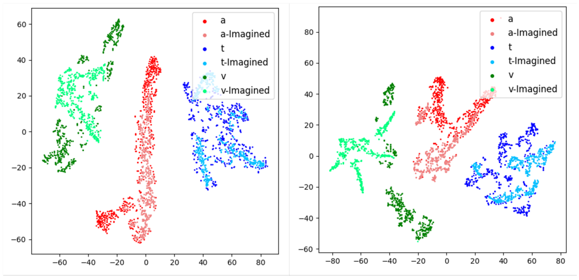
[3] We use features described in Sec. 4.3 and follow the training setting in (Pham et al., 2019) to conduct the MCTN experiments. The MCTN model cannot be evaluated under the full-modality testing condition because the target modalities cannot be None.

Model	Metric	Testing Condition							
		$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
MMIN-AE	WA(↑)	0.5404	0.5025	0.6588	0.6115	0.7203	0.7125	0.6244	0.7619
	UA(↑)	0.5625	0.4836	0.6689	0.6246	0.7374	0.7187	0.6368	0.7677
MMIN-NoCycle	WA(↑)	0.5503	0.5116	0.6577	0.6239	0.7185	0.7202	0.6304	0.7498
	UA(↑)	0.5821	0.5006	0.6705	0.6454	0.7438	0.7301	0.6454	0.7709
MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410**	**0.7650**
	UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524**	**0.7812**

Table 6: Experimental results for component contribution evaluation on IEMOCAP. "MMIN-AE" denotes replacing the CRA structure with the Autoencoder structure in the imagination module. "MMIN-NoCycle" denotes removing the cycle consistency learning in MMIN.

	train	test	Baseline	Augmented	MMIN
ComparE	a	a	0.5760	0.5440	**0.5900**
Denseface	v	v	0.5064	0.4598	**0.5160**
Bert	t	t	0.6873	0.6691	0.6802
ComparE+Denseface	a,v	a,v	0.6380	0.6434	**0.6543**
ComparE+Bert	a,t	a,t	0.7533	0.7435	0.7514
Bert+Denseface	v,t	v,t	0.7177	0.7162	**0.7361**
ComparE+Bert+Denseface	a,v,t	a,v,t	0.7779	0.7767	**0.7812**

Table 7: Evaluation (UA) of the joint representation learning ability on IEMOCAP. "Baseline" denotes the results individually train with cross-entropy loss on partial modalities samples. "Augmented" and "MMIN" denote the evaluation results of our unified data-augmented baseline model and MMIN model under different test conditions, which are the same as in Table 5.



(a) A Female Speaker (b) A Male Speaker

Figure 3: Visualization of the ground-truth and imagined multimodal embeddings. For example, a denotes the ground-truth multimodal embeddings of the acoustic modality. $a\_imagined$ denotes the MMIN imagined multimodal embeddings of the acoustic modality based on visual and textual modalities.

the ground-truth embeddings and imagined embeddings are very similar, although the distribution of visual modality embeddings deviates a little, it is mainly because the quality of the visual modality is poor in this dataset. It demonstrates that MMIN can imagine the representations of the missing modalities based on the available modalities.

5 Conclusion

In this paper, we propose a novel unified multimodal emotion recognition model, Missing Modality Imagination Network (MMIN), to improve the emotion recognition performance under uncertain missing-modality conditions in real application scenarios. The proposed MMIN can learn the robust joint multimodal representations through cross-modality imagination via the Cascade Residual Autoencoder and Cycle Consistency Learning. Extensive experiments on two public benchmark datasets demonstrate the effectiveness and robustness of our proposed model, which significantly outperforms other baselines under both uncertain missing-modality and full-modality conditions.

In the future work, we will explore ways to further improve the robust joint multimodal representation.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, National Natural Science Foundation of China (No. 62072462), National Natural Science Foundation of China (No. 61772535), Beijing Natural Science Foundation (No. 4192028).

References

Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and multi-view models for emotion recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 991–1002.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, ICMI '16, page 279283, New York, NY, USA. Association for Computing Machinery.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 108–116.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462.

N Fragopanagos and J. G. Taylor. 2002. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.

Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller. 2019. Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5861–5865. IEEE.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jingjun Liang, Ruichen Li, and Qin Jin. 2020. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2852–2861.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Zexu Pan, Zhaojie Luo, Jichen Yang, and Haizhou Li. 2020. Multi-modal attention for speech emotion recognition. *Proc. Interspeech 2020*, pages 364–368.

Srinivas Parthasarathy and Shiva Sundaram. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. 2019. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, pages 3534–3540.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1405–1414.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, pages 2514–2520.

Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. 2019. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*.

Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. 2018. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 65–72.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.