

MIR-GAN: Refining Frame-Level Modality-Invariant Representations with Adversarial Network for Audio-Visual Speech Recognition

Yuchen Hu¹, Chen Chen¹, Ruizhe Li², Heqing Zou¹, Eng Siong Chng¹
¹Nanyang Technological University, Singapore ²University of Aberdeen, UK
 {yuchen005@e., chen1436@e., heqing001@e., aseschn@}ntu.edu.sg,
 ruizhe.li@abdn.ac.uk

Abstract

Audio-visual speech recognition (AVSR) attracts a surge of research interest recently by leveraging multimodal signals to understand human speech. Mainstream approaches addressing this task have developed sophisticated architectures and techniques for multi-modality fusion and representation learning. However, the natural heterogeneity of different modalities causes distribution gap between their representations, making it challenging to fuse them. In this paper, we aim to learn the shared representations across modalities to bridge their gap. Different from existing similar methods on other multimodal tasks like sentiment analysis, we focus on the temporal contextual dependencies considering the sequence-to-sequence task setting of AVSR. In particular, we propose an adversarial network to refine frame-level modality-invariant representations (MIR-GAN), which captures the commonality across modalities to ease the subsequent multimodal fusion process. Extensive experiments on public benchmarks LRS3 and LRS2 show that our approach outperforms the state-of-the-arts¹.

1 Introduction

Human perception of the world intrinsically comprises multiple modalities, including vision, audio, text, etc. (McGurk and MacDonald, 1976; Baltrušaitis et al., 2018). Audio-visual speech recognition (AVSR) leverages both audio and visual modalities to understand human speech, improving the noise-robustness of audio-only speech recognition (Chen et al., 2022a,b, 2023a,b; Hu et al., 2022b,a, 2023a,c,b; Zhu et al., 2023a,b) with noise-invariant lip movement information (Sumbly and Pollack, 1954). Thanks to recent advances of deep learning techniques, AVSR research has gained a remarkable progress (Afouras et al., 2018a; Ma et al., 2021; Shi et al., 2022b).

¹Code is available at <https://github.com/YUCHE N005/MIR-GAN>.

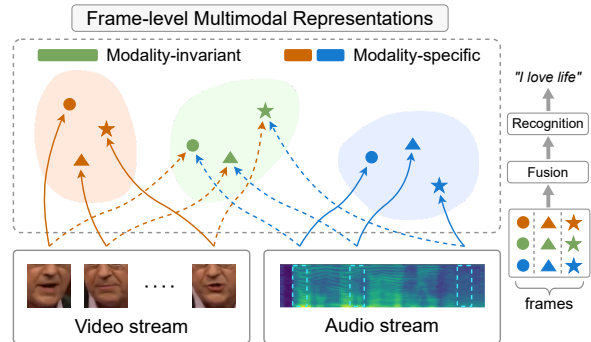


Figure 1: Multimodal learning of frame-level modality-invariant and -specific representations.

Currently, the mainstream AVSR approaches are centered around developing sophisticated architectures and techniques for multi-modality fusion, including simple feature concatenation (Makino et al., 2019; Ma et al., 2021; Pan et al., 2022; Chen et al., 2022c; Zhu et al., 2023c), recurrent neural network (Petridis et al., 2018; Xu et al., 2020) and cross-modal attention (Afouras et al., 2018a; Lee et al., 2020; Hu et al., 2023d). Despite the advances, these approaches are often challenged by the representation gap persisting between naturally heterogeneous modalities (Hazarika et al., 2020).

Recently in some other multimodal tasks like sentiment analysis (Hazarika et al., 2020; Yu et al., 2021; Yao and Mihalea, 2022) and cross-modal retrieval (Xiong et al., 2020), there have been research works proposing to learn two distinct representations to benefit multimodal learning. The first representation is *modality-invariant*, where multiple modalities of a same utterance are mapped to a shared space, indicating the homogeneous semantic meaning from the speaker. In addition, they also learn *modality-specific* representations that are private to each modality. Given an utterance, each modality contains some unique features with respect to speaker-sensitive information (Tsiros, 2013). Combing these two representations provides a holistic view of multimodal data for downstream tasks (Yang et al., 2022). However, these meth-

ods focus on utterance-level representations that could be easily mapped to either shared or individual modality space using similarity cost functions, which does not apply to AVSR task that requires sequence-to-sequence mapping with temporal contextual dependencies (Petridis et al., 2018).

Motivated by above observations, we propose an adversarial network to refine frame-level modality-invariant representations (MIR-GAN) for capturing the commonality across modalities, which bridges their heterogeneous gap to ease the subsequent multimodal fusion. In particular, we first design a MIR generator to learn modality-invariant representations over the shared audio-visual modality space. Meanwhile, a modality discriminator is proposed to strengthen its modality agnosticism via adversarial learning. Moreover, to further enrich its contextual semantic information, we propose a mutual information maximization strategy to align the refined representations to both audio and visual modality sequences. Finally, both modality-invariant and -specific representations are fused for downstream speech recognition. Empirical results demonstrate the effectiveness of our approach. In summary, our main contributions are:

- We present MIR-GAN, an AVSR approach to refine frame-level modality-invariant representations, which captures the commonality across modalities and thus bridges their heterogeneous gap to ease multimodal fusion.
- We first learn modality-invariant representations with a MIR generator, followed by another modality discriminator to strengthen its modality agnosticism via adversarial learning. Furthermore, we propose a mutual information maximization strategy to enrich its contextual semantic information. Finally, both modality-invariant and -specific representations are fused for downstream recognition.
- Our proposed MIR-GAN outperforms the state-of-the-arts on LRS3 and LRS2 benchmarks. Extensive experiments also show its superiority on ASR and VSR tasks.

2 Related Work

Audio-Visual Speech Recognition. Current mainstream AVSR methods focus on sophisticated architectures and techniques for audio-visual modality fusion. Prior methods like RNN-T (Makino

et al., 2019), Hyb-Conformer (Ma et al., 2021) and MoCo+wav2vec (Pan et al., 2022) employ simple feature concatenation for multimodal fusion, other works including Hyb-RNN (Petridis et al., 2018) and EG-seq2seq (Xu et al., 2020) leverage recurrent neural network for audio-visual fusion. In addition, cross-modal attention has also become popular recently for multimodal interaction and fusion in AVSR tasks, such as TM-seq2seq (Afouras et al., 2018a), DCM (Lee et al., 2020) and MMST (Song et al., 2022). Despite the effectiveness, these fusion techniques are often challenged by the representation gap between naturally heterogeneous modalities. Recently, multimodal self-supervised learning has been popular for capturing unified cross-modal representations, like AV-HuBERT (Shi et al., 2022a) and u-HuBERT (Hsu and Shi, 2022), which achieve the state-of-the-art but require abundant unlabeled data and computing resources. In this work, we propose a supervised learning scheme to efficiently refine modality-invariant representations for bridging the heterogeneous modality gap.

Modality-Invariant and -Specific Representations. Recent studies in many multimodal tasks suggest that the model benefits from both shared and individual modality representations, including multimodal sentiment analysis (Hazarika et al., 2020; Yu et al., 2021; Yang et al., 2022), person re-identification (Wei et al., 2021; Huang et al., 2022), cross-modal retrieval (Zeng et al., 2022) and image-sentence matching (Liu et al., 2019), etc. MISA (Hazarika et al., 2020) maps the multimodal features into two spaces as modality-invariant and -specific representations, and then fuses them for downstream classification. MCLNet (Hao et al., 2021) learns modality-invariant representations by minimizing inter-modal discrepancy and maximizing cross-modal similarity. VI-REID (Feng et al., 2019) builds an individual network for each modality, with a shared identity loss to learn modality-invariant representations. However, these methods map utterance-level representations to modality-invariant or -specific spaces via similarity cost functions, while AVSR is sequence-to-sequence task that requires contextual semantic information. To this end, we propose an adversarial network with mutual information maximization to refine frame-level modality-invariant representations that subjects to temporal contextual dependencies.

Adversarial Network. The concept of adversarial network starts from GAN (Goodfellow et al., 2014),

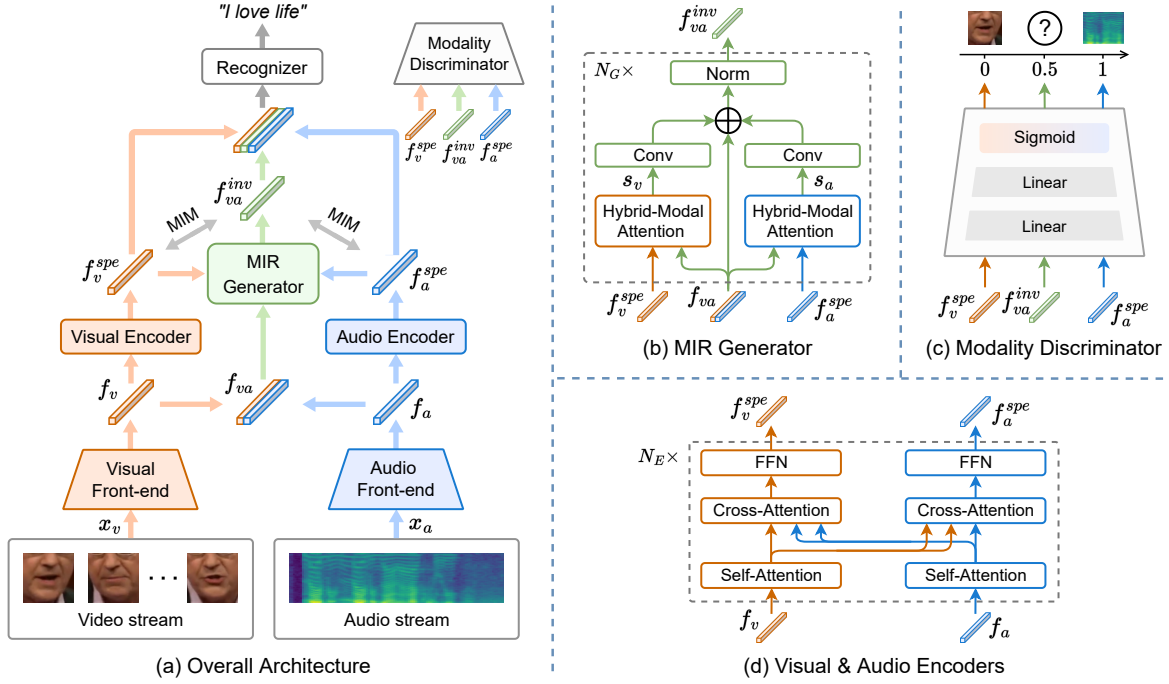


Figure 2: Illustration of our MIR-GAN. (a) Overall architecture. (b) MIR generator that learns modality-invariant representation f_{va}^{inv} . (c) Modality discriminator that strengthens the modality agnosticism of f_{va}^{inv} . (d) Visual and audio encoders that learn modality-specific representations f_v^{spe}, f_a^{spe} . “MIM” is mutual information maximization.

which has attracted a surge of research interests due to its strong ability of generating high-quality novel samples according to existing data. The best-known applications include image-to-image translation (Isola et al., 2017) and image synthesis (Denton et al., 2015; Radford et al., 2015). Recently, GAN is further applied to multimodal tasks such as text-to-image synthesis (Reed et al., 2016; Tan et al., 2020), video captioning (Yang et al., 2018; Bai et al., 2021) and cross-modal retrieval (Qian et al., 2021). In this work, we leverage the strong distinguishing ability of adversarial network to strengthen the modality agnosticism of the learned modality-invariant representations.

3 Methodology

3.1 Overview

The overall architecture of our proposed MIR-GAN is illustrated in Fig. 2. First, we have two front-end modules² to process the input streams, which generate two modality sequences, *i.e.*, $f_v, f_a \in \mathbb{R}^{T \times D}$, where T is number of frames and D is embedding size. These two sequences are then fed by visual and audio encoders respectively to generate modality-specific representations, *i.e.*, $f_v^{spe}, f_a^{spe} \in \mathbb{R}^{T \times D}$. Based on that, we propose a MIR generator

to learn modality-invariant representations by extracting the shared information of two modalities, *i.e.*, $f_{va}^{inv} \in \mathbb{R}^{T \times D}$. Meanwhile, we design a modality discriminator to strengthen its modality agnosticism via adversarial learning. In addition, to further enrich its contextual semantic information, we propose a mutual information maximization (MIM) strategy to align the refined representations to both audio and visual modality sequences. Finally, both modality-invariant and -specific representations are fused for downstream speech recognition.

3.2 Visual & Audio Encoders

As illustrated in Fig. 2 (d), we introduce a pair of visual and audio encoders to learn modality-specific representations. Following Transformer (Vaswani et al., 2017) architecture, they first employ self-attention modules to capture the contextual dependencies within each modality, followed by cross-attention modules for interaction between two modalities, which can initially narrow their gap to benefit the subsequent modality-invariant representation learning. Finally, there are feed-forward networks to generate the modality-specific outputs.

3.3 MIR-GAN

With learned modality-specific representations, we propose MIR-GAN to refine frame-level modality-invariant representations. First, we design a MIR

²Details are presented in Appendix A.3.

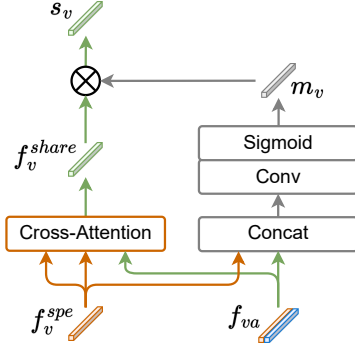


Figure 3: Illustration of the Hybrid-Modal Attention. Here we take the visual modality for example ($m = v$).

generator to extract the shared information of two modalities, which generates a modality-invariant representation $f_{va}^{inv} \in \mathbb{R}^{T \times D}$. Meanwhile, we design a modality discriminator to strengthen its modality agnosticism via adversarial learning.

3.3.1 MIR Generator

Fig. 2 (b) details the architecture of proposed MIR generator G , where we design a hybrid-modal attention (HMA) module to extract out the part of information in each modality-specific representation that is related to both modalities:

$$s_m = HMA(f_m^{spe}, f_{va}), \quad m \in \{v, a\}, \quad (1)$$

where the subscript m denotes modality. The resulted features are then added to input sequence f_{va} to form the final modality-invariant representation:

$$f_{va}^{inv} = \text{Norm}(f_{va} + \sum_{m \in \{v, a\}} \text{Conv}(s_m)), \quad (2)$$

where the ‘‘Norm’’ denotes layer normalization (Ba et al., 2016), ‘‘Conv’’ denotes 1×1 convolution followed by PReLU activation (He et al., 2015).

Hybrid-Modal Attention (HMA) first involves a cross-attention sub-module to extract the information in each modality-specific representation that is related to both modalities, with the query input f_{va} comprising both visual and audio sequence information, as shown in Fig. 3:

$$f_m^{share} = \text{Cross-Attention}(f_{va}, f_m^{spe}, f_m^{spe}), \quad (3)$$

To further make the extracted feature invariant to modalities, we design a parallel convolutional network to learn a mask for filtering out the modality-specific information:

$$s_m = f_m^{share} \otimes \sigma(\text{Conv}(f_m^{spe} \parallel f_{va})), \quad (4)$$

where ‘‘Conv’’ denotes 1×1 convolutional layer, \parallel denotes feature concatenation, σ denotes Sigmoid activation, \otimes denotes element-wise multiplication.

As a result, the output representation s_m from HMA involves information regarding both visual and audio modalities, making the final output f_{va}^{inv} (in Eq. 2) invariant to modalities.

3.3.2 Modality Discriminator

With the generated modality-invariant representation, we further design a modality discriminator D to strengthen its modality agnosticism via adversarial learning. As shown in Fig. 2 (c), the discriminator consists of two linear layers followed by Sigmoid activation to predict a scalar between 0 and 1 for each frame, indicating which modality it belongs to (*i.e.*, 0 for visual and 1 for audio):

$$D(f) \in \mathbb{R}^{T \times 1}, \quad f \in \{f_v^{spe}, f_a^{spe}, f_{va}^{inv}\}, \quad (5)$$

Therefore, for frames in modality-specific representations f_v^{spe} and f_a^{spe} , we hope the discriminator can correctly classify the modality type, *i.e.*, 0 or 1. In contrast, in order to strengthen the modality agnosticism of refined representation f_{va}^{inv} , we hope it can confuse the discriminator with the output around 0.5, *i.e.*, a medium between two modalities.

With above designs of generator and discriminator, the adversarial training objective of MIR-GAN can be mathematically formulated as:

$$\begin{aligned} \mathcal{L}_{GAN} &= \mathcal{L}_D + \mathcal{L}_G \\ &= \mathbb{E}_f[\log D(f_a^{spe}) + \log(1 - D(f_v^{spe}))] \\ &\quad + \mathbb{E}_f[-\log D(f_{va}^{inv}) - \log(1 - D(f_{va}^{inv}))], \end{aligned} \quad (6)$$

where $f_{va}^{inv} = G(f_v^{spe}, f_a^{spe}, f_{va})$, \mathbb{E} denotes the expectation over all the temporal frames in current data batch. Details of the corresponding optimization strategy are illustrated in Alg. 1.

3.4 Mutual Information Maximization

The MIR-GAN successfully refines the modality-invariant representation by focusing on the modality commonality and agnosticism, while the original semantic information may not be preserved. To this end, we further design a mutual information maximization (MIM) strategy via contrastive learning to enrich the contextual semantic information in refined modality-invariant representation.

In particular, we formulate a contrastive loss function to maximize the mutual information between modality-invariant representation f_{va}^{inv} and the modality-specific representations f_v^{spe}, f_a^{spe} :

$$\mathcal{L}_{MIM} = - \sum_{i=1}^T \log \frac{\exp(\langle f_{va_i}^{inv}, f_{v_i}^{spe} \rangle / \tau)}{\sum_{j=1}^T \exp(\langle f_{va_i}^{inv}, f_{v_j}^{spe} \rangle / \tau)} - \sum_{i=1}^T \log \frac{\exp(\langle f_{va_i}^{inv}, f_{a_i}^{spe} \rangle / \tau)}{\sum_{j=1}^T \exp(\langle f_{va_i}^{inv}, f_{a_j}^{spe} \rangle / \tau)}, \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, τ is temperature parameter. The subscripts i and j denote frame index, where $f_{va}^{inv} / f_v^{spe} / f_a^{spe} \in \mathbb{R}^{T \times D}$.

The constructed positive and negative samples are distinguished by frame index. As same frame of different representations express similar semantic meanings, we assign them as positive samples to strengthen consistency, while the mismatched frames are pulled apart from each other. As a result, the MIM strategy can enrich the semantic information in final modality-invariant representation.

3.5 Optimization

The optimization strategy of MIR-GAN is detailed in Alg. 1. After the forward-propagation process, we calculate \mathcal{L}_{GAN} and \mathcal{L}_{MIM} according to Eq. 6 and Eq. 7. Meanwhile, the downstream speech recognition loss \mathcal{L}_{rec} is calculated as the cross-entropy between recognized text and the ground-truth transcription. The final training objective of MIR-GAN can therefore be written as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{GAN} \cdot \mathcal{L}_{GAN} + \lambda_{MIM} \cdot \mathcal{L}_{MIM}, \quad (8)$$

where $\lambda_{GAN}, \lambda_{MIM}$ are weighting parameters to balance different training objectives.

Inspired by GAN training strategy (Goodfellow et al., 2014), we split the back-propagation process into two steps. First, we *maximize* \mathcal{L}_{GAN} to update the discriminator, where the generator is detached from optimization. According to Eq. 6, maximizing the first term of \mathcal{L}_{GAN} (*i.e.*, \mathcal{L}_D) trains the discriminator to correctly classify the two modalities, while increasing the second term amounts to informing discriminator that f_{va}^{inv} is modality-specific and can either be visual or audio³ (this is opposite to what we desire as modality-invariant). Second, we freeze discriminator and update the rest network, where *minimizing* \mathcal{L}_G pushes the discrimination output of f_{va}^{inv} to 0.5,³ which is a medium between visual and audio modalities, *i.e.*, modality-agnostic.

³Function $\log(x) + \log(1-x)$ reaches maximum at $x = 0.5$, and the minimum is obtained around $x = 0$ and $x = 1$.

Algorithm 1 MIR-GAN Optimization.

Require: Training data D that contains visual-audio pairs (x_v, x_a) and the text transcription y . The MIR-GAN network θ that consists of front-ends θ_{vf} and θ_{af} , encoders θ_{vae} , MIR generator θ_G , modality discriminator θ_D and downstream speech recognition model θ_{rec} . Hyper-parameter weights $\lambda_{GAN}, \lambda_{MIM}$.

- 1: Randomly initialize the entire system θ .
 - 2: **if** select *self-supervised setting* **then**
 - 3: Load the pre-trained AV-HuBERT for speech recognition model θ_{rec} and front-ends θ_{vf}, θ_{af}
 - 4: **end if**
 - 5: **while** not converged **do**
 - 6: **for** $(x_v, x_a) \in D$ **do**
 - 7: FORWARD-PROPAGATION:
 - 8: $f_v = \theta_{vf}(x_v), f_a = \theta_{af}(x_a)$ ▷ front-ends
 - 9: $f_v^{spe}, f_a^{spe} = \theta_{vae}(f_v, f_a)$ ▷ encoders
 - 10: $f_{va} = f_v \parallel f_a$
 - 11: $f_{va}^{inv} = \theta_G(f_v^{spe}, f_a^{spe}, f_{va})$ ▷ Generator
 - 12: $\hat{y} = \theta_{rec}(f_v^{spe} \parallel f_a^{spe} \parallel f_{va}^{inv})$ ▷ recognition
 - 13: TRAINING OBJECTIVES:
 - 14: \mathcal{L}_{GAN} (\mathcal{L}_D and \mathcal{L}_G) in Eq. 6 ▷ Discriminator
 - 15: \mathcal{L}_{MIM} in Eq. 7 ▷ MI maximization
 - 16: $\mathcal{L}_{rec} = \text{CrossEntropy}(\hat{y}, y)$
 - 17: BACK-PROPAGATION: ▷ adversarial training
 - 18: UPDATE DISCRIMINATOR: ▷ unfreeze θ_D
 - 19: $\arg \max_{\theta_D} \mathcal{L}_{GAN}$
 - 20: UPDATE THE REST NETWORK: ▷ freeze θ_D
 - 21: $\arg \min_{\theta \setminus \theta_D} \mathcal{L}_{rec} + \lambda_{GAN} \cdot \mathcal{L}_G + \lambda_{MIM} \cdot \mathcal{L}_{MIM}$
 - 22: **end for**
 - 23: **end while**
-

In addition, \mathcal{L}_{rec} optimizes the downstream speech recognition model and \mathcal{L}_{MIM} implements the MIM strategy. The entire system is trained in an end-to-end manner with well-tuned weighting parameters.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two large-scale public benchmarks, LRS3 (Afouras et al., 2018b) and LRS2 (Chung et al., 2017). LRS3 dataset collects 433 hours of transcribed English videos in TED and TEDx talks from over 5000 speakers, which is the largest publicly available labeled audio-visual speech recognition dataset. LRS2 dataset contains 224 hours of video speech, with a total of 144K clips from BBC programs.

Model Configurations. We first build a base model with only front-ends and downstream speech recognition module, which follows Transformer architecture with 24 encoder layers and 9 decoder layers. Based on that, we build the MIR-GAN with $N_E = 3$ visual & audio encoder layers and $N_G = 3$ MIR generator layers. To maintain similar model size, we only use 12 encoder layers and 9 decoder layers in the recognition model. The number

Method	Backbone	Criterion	Unlabeled data (hrs)	Labeled data (hrs)	DataAug	LM	WER(%)		
							Clean	Noisy	
<i>Supervised</i>									
TM-seq2seq (2018a)	Transformer	S2S	-	1,519	✓	✓	7.2	-	
EG-seq2seq (2020)	RNN	S2S	-	590	✓	-	6.8	-	
RNN-T (2019)	RNN	RNN-T	-	31,000	-	-	4.5	-	
Hyb-Conformer (2021)	Conformer	S2S + CTC	-	590	✓	✓	2.3	-	
<i>Self-Supervised</i>									
AV-HuBERT (2022b)	Transformer	S2S	1,759	433	✓	-	1.4	5.8	
u-HuBERT (2022)	Transformer	S2S	2,211	433	✓	-	1.2	-	
<i>Proposed (Supervised)</i>									
Ours	Base model	Transformer	S2S	-	433	✓	-	3.5	14.8
	MIR-GAN							2.8	11.7
	Base model	Conformer	S2S	-	433	✓	-	2.5	10.9
	MIR-GAN							2.1	8.5
<i>Proposed (Self-Supervised)</i>									
Ours	Base model	Transformer	S2S	1,759	433	✓	-	1.4	5.8
	MIR-GAN							1.2	5.6

Table 1: WER (%) of our MIR-GAN and prior works on LRS3 benchmark. ‘‘S2S’’ denotes sequence-to-sequence loss (Watanabe et al., 2017), ‘‘CTC’’ denotes CTC loss (Graves et al., 2006), ‘‘DataAug’’ denotes noise augmentation, ‘‘LM’’ denotes language model rescoring. The noisy test set is synthesized using MUSAN noise (Snyder et al., 2015).

of parameters in our base model and MIR-GAN are 476M and 469M respectively. We also use Conformer (Gulati et al., 2020) as our backbone. In addition, we implement a self-supervised setting by loading pre-trained AV-HuBERT⁴. Following prior work (Shi et al., 2022b), we employ data augmentation and noisy test set based on MUSAN noise (Snyder et al., 2015). More detailed settings are presented in Appendix A.3 - A.5.

Baselines. To evaluate our proposed MIR-GAN, we select some popular AVSR methods for comparison, which can be roughly divided into two groups. The first is supervised learning method, including TM-seq2seq/CTC (Afouras et al., 2018a), RNN-T (Makino et al., 2019), EG-seq2seq (Xu et al., 2020) and Hyb-Conformer (Ma et al., 2021). Another one is the recently popular self-supervised learning method such as MoCo+wav2vec (Pan et al., 2022), AV-HuBERT (Shi et al., 2022b) and u-HuBERT (Hsu and Shi, 2022).

4.2 Main Results

We conduct experiments on two public datasets under *supervised* and *self-supervised* settings, depending on whether use the AV-HuBERT pre-trained model. Results show that our proposed MIR-GAN achieves the state-of-the-art under both settings.

LRS3 Benchmark. Table 1 presents the AVSR

⁴https://github.com/facebookresearch/av_hubert

Method	Backbone	WER(%)		
		Clean	Noisy	
<i>Supervised</i>				
TM-seq2seq (2018a)	Transformer	8.5	-	
TM-CTC (2018a)	Transformer	8.2	-	
Hyb-RNN (2018)	RNN	7.0	-	
LF-MMI TDNN (2020)	TDNN	5.9	-	
Hyb-Conformer (2021)	Conformer	3.7	-	
<i>Self-Supervised</i>				
MoCo+wav2vec (2022)	Transformer	2.6	-	
<i>Proposed (Supervised)</i>				
Ours	Base model	Transformer	5.4	21.2
	MIR-GAN		4.5	16.7
	Base model	Conformer	3.9	15.8
	MIR-GAN		3.2	11.9
<i>Proposed (Self-Supervised)</i>				
Ours	Base model	Transformer	2.3	7.3
	MIR-GAN		2.2	7.0

Table 2: WER (%) of our MIR-GAN and prior works on the LRS2 benchmark. Detailed configurations are further presented in Table 6.

performance of our proposed MIR-GAN and prior methods on LRS3 benchmark. Under supervised setting, our MIR-GAN achieves significant improvement over the base model in both clean and noisy testing conditions, and the best performance achieves new state-of-the-art (2.1% vs. 2.3%) while without using the language model rescoring. In addition, the Conformer backbone consistently outperforms Transformer (2.1% vs. 2.8%, 8.5% vs. 11.7%). Under self-supervised setting, MIR-

Model	TF-Sup-3		CF-Sup-3		TF-SelfSup-3		TF-Sup-2		CF-Sup-2		TF-SelfSup-2	
	Clean	Noisy	Clean	Noisy	Clean	Noisy	Clean	Noisy	Clean	Noisy	Clean	Noisy
MIR-GAN (Full)	2.8	11.7	2.1	8.5	1.2	5.6	4.5	16.7	3.2	11.9	2.2	7.0
<i>Importance of Representations</i>												
w/o Modality-Invariant	3.3	13.7	2.4	10.1	1.3	5.8	5.3	19.9	3.7	14.9	2.3	7.2
w/o Modality-Specific	3.2	13.2	2.3	9.8	1.4	5.7	5.1	19.5	3.7	14.6	2.2	7.1
<i>Importance of Modules</i>												
w/o Visual & Audio Encoders	3.0	12.1	2.1	8.9	1.2	5.6	4.8	18.1	3.4	13.1	2.2	7.0
w/o MIR Generator	3.1	12.8	2.2	9.2	1.3	5.7	4.9	18.7	3.6	13.8	2.2	7.1
w/o Modality Discriminator	3.2	13.3	2.3	9.7	1.4	5.8	5.2	19.4	3.7	14.5	2.3	7.2
<i>Importance of Strategies</i>												
w/o Adversarial Training	3.1	13.0	2.3	9.5	1.3	5.7	5.1	19.2	3.6	14.1	2.2	7.2
w/o MIM Strategy	2.9	12.0	2.1	9.0	1.2	5.6	4.7	17.8	3.5	12.6	2.2	7.1

Table 3: Ablation study on LRS3 and LRS2 benchmarks. Results are reported on six configurations in the format “[Backbone]-[Setting]-[Test set]”, where “TF”/“CF” denote Transformer/Conformer backbone, “Sup”/“SelfSup” denote supervised/self-supervised setting, “3”/“2” denote LRS3/LRS2 test set.

GAN also improves the performance of base model, which surpasses or matches previous state-of-the-art (1.2% vs. 1.2%, 5.6% vs. 5.8%) while using less unlabeled data for pre-training.

LRS2 Benchmark. Table 2 compares the AVSR results of MIR-GAN and baselines on LRS2 benchmark. We can observe that the proposed MIR-GAN outperforms previous state-of-the-art by a large margin under both supervised and self-supervised settings (3.2% vs. 3.7%, 2.2% vs. 2.6%). In addition, we also observe promising gains of performance in noisy testing conditions.

As a result, our proposed MIR-GAN achieves new state-of-the-art under both supervised and self-supervised settings on two public benchmarks, which demonstrates its superiority on AVSR task.

4.3 Ablation Study

Table 3 presents the ablation study of each component in MIR-GAN. There are three parts of ablation that are independent with each other, *i.e.*, each study is conducted where other two components are kept same as the full MIR-GAN.

Importance of Representations. We first investigate the importance of modality-invariant and -specific representations by discarding each of them. When removing the refined modality-invariant representations from multi-modality fusion, the downstream speech recognition performance degrades a lot under all configurations, which verifies its significance of bridging the modality gap. Similarly, we observe that the modality-specific representations also plays an important role in AVSR.

Importance of Modules. In this part, we study the role of each module in the proposed MIR-GAN.

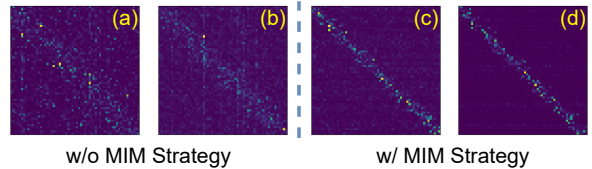


Figure 4: Alignment (attention map) between modality-invariant and -specific representations with and without MIM strategy: (a)(c) $f_v^{inv} \leftrightarrow f_v^{spe}$, (b)(d) $f_v^{inv} \leftrightarrow f_a^{spe}$.

The visual and audio encoders are designed to extract deep modality-specific representations, which contributes to performance gains of MIR-GAN. Then we replace the core module - MIR generator with simple feature concatenation in refining modality-invariant representations, which results in significant performance degradation. Another key module - modality discriminator also contributes a lot in MIR-GAN by strengthening the modality agnosticism of refined representations from MIR generator. In this sense, we conclude that all the modules in proposed MIR-GAN contribute positively to the multimodal representation learning.

Importance of Strategies. With the adversarial training strategy illustrated in Alg. 1, the proposed modality discriminator effectively strengthens the modality agnosticism of the refined representations from generator. To verify its effectiveness, we remove the adversarial training strategy from MIR-GAN, which results in similar performance degradation to the previous case without modality discriminator. Therefore, it demonstrates the key role of this strategy in learning modality-invariant representations, where further visualization is shown in Fig. 5. Meanwhile, we design a MIM strategy to enrich the contextual semantic information in the

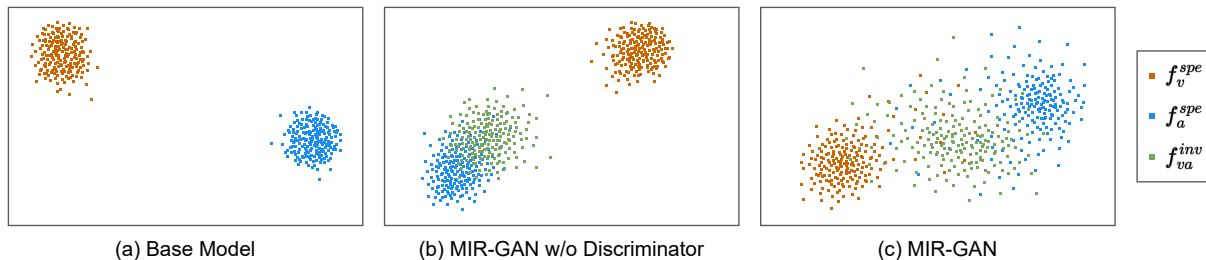


Figure 5: The t-SNE visualization of modality-invariant and -specific representations from (a) base model, (b) MIR-GAN without modality discriminator and (c) MIR-GAN. The orange and blue points denote visual and audio modality-specific representations respectively, and green points denote modality-invariant representations. This study is conducted on frame-level representations using a portion of LRS3 test set.

refined modality-invariant representations, and similar performance drops can be observed in absence of such strategy. Furthermore, we visualize the attention maps in Fig. 4 to show its effectiveness. The clear diagonals in (c) and (d) indicate the strong ability of MIM strategy to align modality-invariant and -specific representations, which enriches the contextual semantic information in the former.

Visualizations of Modality-Invariant and -Specific Representations. Fig. 5 presents the t-SNE visualization of modality-invariant and -specific representations to illustrate the principle of MIR-GAN. First, we observe from (a) base model that the two modality-specific representations are distantly separated, indicating the heterogeneous gap between different modalities (Hazarika et al., 2020). With the proposed MIR-GAN (no modality discriminator), the two modalities are pushed closer by the interaction between encoders, and the refined modality-invariant representations serve as a medium between them. However, these refined representations are still entangled with audio modality-specific representations⁵, making them less modality-invariant. Finally, the proposed discriminator effectively strengthens their modality agnosticism via adversarial learning, which are dispersed between two modalities to capture their commonality and thus bridge the heterogeneous modality gap. As a result, the subsequent multi-modality fusion process would be eased and generate better features for downstream recognition.

Comparison with Utterance-Level Approaches

As illustrated in §2, prior works have investigated utterance-level modality-invariant and -specific representations with similarity cost functions, including MISA (Hazarika et al., 2020), MCLNet (Hao et al., 2021) and VI-REID (Feng et al., 2019). We

Method	WER(%)	
	Clean	Noisy
Base Model	3.5	14.8
+ MCLNet (Hao et al., 2021)	3.4	14.5
+ VI-REID (Feng et al., 2019)	3.3	14.0
+ MISA (Hazarika et al., 2020)	3.3	13.7
MIR-GAN (ours)	2.8	11.7

Table 4: Comparison between MIR-GAN and utterance-level multimodal approaches on LRS3 benchmark.

implement them in our framework as comparison to our proposed MIR-GAN, where we employ their designed similarity cost functions on frame-level representations. As illustrated in Table 4, these utterance-level approaches can also improve AVSR results but still underperforms our proposed approach by a large margin.

Performance on Single-Modality Inputs. Furthermore, Table 5 presents the performance of our MIR-GAN on single-modality inputs. First, we observe that in all models using both modalities performs better than single modality, and the audio-only case achieves much better results than visual-only case, which shows the dominance of audio modality in AVSR task. Under two single-modality cases, our proposed MIR-GAN both achieves significant improvement over the base model, and the best performance outperforms or matches previous state-of-the-arts in both supervised and self-supervised settings (2.3% vs. 2.3%, 34.2% vs. 33.6%; 1.3% vs. 1.4%, 26.6% vs. 26.9%). Therefore, even with missing modality, our MIR-GAN can still refine effective modality-invariant representations to benefit the downstream speech recognition, which further verifies the generality of our approach.

5 Conclusion

In this paper, we propose MIR-GAN, an adversarial network to refine frame-level modality-invariant

⁵Audio modality plays the dominant role in AVSR task.

Method	Backbone	WER(%)		
		AV	A	V
<i>Supervised</i>				
TM-seq2seq (2018a)	Transformer	7.2	8.3	58.9
EG-seq2seq (2020)	RNN	6.8	7.2	57.8
RNN-T (2019)	RNN	4.5	4.8	33.6
Hyb-Conformer (2021)	Conformer	2.3	2.3	43.3
<i>Self-Supervised</i>				
Distill-Pretrain (2022)	Conformer	-	-	31.5
AV-HuBERT (2022b)	Transformer	1.4	1.5	26.9
u-HuBERT (2022)	Transformer	1.2	1.4	27.2
<i>Proposed (Supervised)</i>				
Ours	Base model	3.5	4.7	63.5
	MIR-GAN	2.8	3.5	48.6
	Base model	2.5	3.0	40.2
	MIR-GAN	2.1	2.3	34.2
<i>Proposed (Self-Supervised)</i>				
Ours	Base model	1.4	1.6	28.6
	MIR-GAN	1.2	1.3	26.6

Table 5: Performance on single-modality inputs with LRS3 benchmark. “AV”, “A” and “V” indicate the input modality during both finetuning and inference stages. The missing modality is replaced by zero embeddings.

representations for AVSR, which captures the commonality across modalities to ease the multimodal fusion process. MIR-GAN first learns modality-invariant representation with MIR generator, followed by a modality discriminator to strengthen its modality agnosticism via adversarial learning. Furthermore, we propose a mutual information maximization strategy to enrich its contextual semantic information. Finally, both modality-invariant and -specific representations are fused to provide a holistic view of multimodal data for downstream task. Experiments on public benchmarks show that our MIR-GAN achieves the state-of-the-art.

Limitations

The main novelty of our proposed MIR-GAN is refining frame-level modality-invariant representations via adversarial learning. It is promising to combine this approach with the popular self-supervised pre-training to learn unified multimodal representations. In this work, we only load pre-trained AV-HuBERT for the front-ends and speech recognition model, while the proposed modules (*i.e.*, encoders, generator, discriminator) are still trained from scratch. In future, we may include the entire MIR-GAN into self-supervised learning scheme, together with the adversarial learning to refine better multimodal representations.

Ethics Statement

All the data used in this paper are publicly available and are used under the following five licenses: the Creative Commons BY-NC-ND 4.0 License and Creative Commons Attribution 4.0 International License, the TED Terms of Use, the YouTube’s Terms of Service, and the BBC’s Terms of Use. The data is collected from TED and BBC and contain thousands of speakers from a wide range of races. To protect the anonymity, only the mouth area of a speaker is visualized wherever used in the paper.

Acknowledgements

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (<https://www.nsccl.org.sg>).

References

- Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Senior. 2018a. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.
- Triantafyllos Afouras, Joon Son Chung, and Andrew Senior. 2018b. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.
- Yang Bai, Junyan Wang, Yang Long, Bingzhang Hu, Yang Song, Maurice Pagnucco, and Yu Guan. 2021. Discriminative latent semantic graph for video captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3556–3564.
- 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.
- Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng. 2022a. Noise-robust speech recognition with 10 minutes unparalleled in-domain data. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4298–4302. IEEE.

- Chen Chen, Yuchen Hu, Nana Hou, Xiaofeng Qi, Heqing Zou, and Eng Siong Chng. 2022b. Self-critical sequence training for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3688–3692. IEEE.
- Chen Chen, Yuchen Hu, Weiwei Weng, and Eng Siong Chng. 2023a. Metric-oriented speech enhancement using diffusion probabilistic model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Chen Chen, Yuchen Hu, Qiang Zhang, Heqing Zou, Beier Zhu, and Eng Siong Chng. 2022c. Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning. *arXiv preprint arXiv:2212.05301*.
- Chen Chen, Yuchen Hu, Heqing Zou, Linhui Sun, and Eng Siong Chng. 2023b. Unsupervised noise adaptation using data simulation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Senior. 2017. Lip reading sentences in the wild. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3444–3453. IEEE.
- Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, 28.
- Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*.
- Zhanxiang Feng, Jianhuang Lai, and Xiaohua Xie. 2019. Learning modality-specific representations for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 29:579–590.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.
- Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.
- Anmol Gulati, James Qin, Chiu Chung-Cheng, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, pages 5036–5040.
- Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16403–16412.
- Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Wei-Ning Hsu and Bowen Shi. 2022. u-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In *Advances in Neural Information Processing Systems*.
- Yuchen Hu, Chen Chen, Ruizhe Li, Qiushi Zhu, and Eng Siong Chng. 2023a. Gradient remedy for multi-task learning in end-to-end noise-robust speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yuchen Hu, Chen Chen, Qiushi Zhu, and Eng Siong Chng. 2023b. Wav2code: Restore clean speech representations via codebook lookup for noise-robust asr. *arXiv preprint arXiv:2304.04974*.
- Yuchen Hu, Chen Chen, Heqing Zou, Xionghu Zhong, and Eng Siong Chng. 2023c. Unifying speech enhancement and separation with gradient modulation for end-to-end noise-robust speech separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yuchen Hu, Nana Hou, Chen Chen, and Eng Siong Chng. 2022a. Dual-path style learning for end-to-end noise-robust speech recognition. *arXiv preprint arXiv:2203.14838*.
- Yuchen Hu, Nana Hou, Chen Chen, and Eng Siong Chng. 2022b. Interactive feature fusion for end-to-end noise-robust speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6292–6296. IEEE.
- Yuchen Hu, Ruizhe Li, Chen Chen, Heqing Zou, Qiushi Zhu, and Eng Siong Chng. 2023d. Cross-modal global interaction and local alignment for audio-visual speech recognition. *arXiv preprint arXiv:2305.09212*.

- Nianchang Huang, Jianan Liu, Yongjiang Luo, Qiang Zhang, and Jungong Han. 2022. Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification. *Pattern Recognition*, page 109145.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.
- Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.
- Yong-Hyeok Lee, Dong-Won Jang, Jae-Bin Kim, Rae-Hong Park, and Hyung-Min Park. 2020. Audio-visual speech recognition based on dual cross-modality attentions with the transformer model. *Applied Sciences*, 10(20):7263.
- Ruoyu Liu, Yao Zhao, Shikui Wei, Liang Zheng, and Yi Yang. 2019. Modality-invariant image-text embedding for image-sentence matching. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–19.
- Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE.
- Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *arXiv preprint arXiv:2202.13084*.
- Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE.
- Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
- Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. 2022. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4491–4503, Dublin, Ireland. Association for Computational Linguistics.
- Stavros Petridis, Themis Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE.
- Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2440–2448.
- Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.
- Bowen Shi, Wei-Ning Hsu, Kushal Lakhota, and Abdelrahman Mohamed. 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*.
- Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022b. Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763*.
- David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.
- Qiya Song, Bin Sun, and Shutao Li. 2022. Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems*.
- William H Sumby and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.
- Hongchen Tan, Xiuping Liu, Meng Liu, Baocai Yin, and Xin Li. 2020. Kt-gan: knowledge-transfer generative adversarial network for text-to-image synthesis. *IEEE Transactions on Image Processing*, 30:1275–1290.
- Augoustinos Tsiros. 2013. The dimensions and complexities of audio-visual association. *Electronic Visualisation and the Arts (EVA 2013)*, pages 149–156.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

- Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.
- Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. 2021. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234.
- Haixia Xiong, Weihua Ou, Zengxian Yan, Jianping Gou, Quan Zhou, and Anzhi Wang. 2020. Modality-specific matrix factorization hashing for cross-modal retrieval. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–15.
- Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. 2020. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442.
- Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. 2022. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1708–1717.
- Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen, and Yanli Ji. 2018. Video captioning by adversarial lstm. *IEEE Transactions on Image Processing*, 27(11):5600–5611.
- Yiqun Yao and Rada Mihalcea. 2022. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834.
- Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. 2020. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE.
- Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.
- Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. 2022. Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*.
- Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng. 2019. Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks. *IEEE Transactions on Image Processing*, 29:1061–1073.
- Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, and Li-Rong Dai. 2023a. A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. 2023b. Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2023c. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*.

A Experimental Details

A.1 Datasets

LRS3⁶ (Afouras et al., 2018b) is currently the largest public sentence-level lip reading dataset, which contains over 400 hours of English video extracted from TED and TEDx talks on YouTube. The training data is divided into two parts: pretrain (403 hours) and trainval (30 hours), and both of them are transcribed at sentence level. The pretrain part differs from trainval in that the duration of its video clips are at a much wider range. Since there is no official development set provided, we randomly select 1,200 samples from trainval as validation set (~ 1 hour) for early stopping and hyper-parameter tuning. In addition, it provides a standard test set (0.9 hours) for evaluation.

LRS2⁷ (Chung et al., 2017) is a large-scale publicly available labeled audio-visual (A-V) datasets, which consists of 224 hours of video clips from BBC programs. The training data is divided into three parts: pretrain (195 hours), train (28 hours) and val (0.6 hours), which are all transcribed at sentence level. An official test set (0.5 hours) is provided for evaluation use.

A.2 Data Preprocessing

The data preprocessing for above two datasets follows the LRS3 preprocessing steps in prior work (Shi et al., 2022a). For the audio stream, we extract the 26-dimensional log filter-bank feature at a stride of 10 ms from input raw waveform.

⁶https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html

⁷https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html

Method	Backbone	Criterion	Unlabeled data (hrs)	Labeled data (hrs)	DataAug	LM	WER(%)		
							Clean	Noisy	
<i>Supervised</i>									
TM-seq2seq (2018a)	Transformer	S2S	-	1,519	✓	✓	8.5	-	
TM-CTC (2018a)	Transformer	CTC	-	1,519	✓	✓	8.2	-	
Hyb-RNN (2018)	RNN	S2S + CTC	-	397	✓	✓	7.0	-	
LF-MMI TDNN (2020)	TDNN	LF-MMI	-	224	-	✓	5.9	-	
Hyb-Conformer (2021)	Conformer	S2S + CTC	-	381	✓	✓	3.7	-	
<i>Self-Supervised</i>									
MoCo+wav2vec (2022)	Transformer	S2S + CTC	60,000	381	✓	-	2.6	-	
<i>Proposed (Supervised)</i>									
Ours	Base model	Transformer	S2S	-	224	✓	-	5.4	21.2
	MIR-GAN							4.5	16.7
	Base model	Conformer	S2S	-	224	✓	-	3.9	15.8
	MIR-GAN							3.2	11.9
<i>Proposed (Self-Supervised)</i>									
Ours	Base model	Transformer	S2S	1,759	224	✓	-	2.3	7.3
	MIR-GAN							2.2	7.0

Table 6: WER (%) of our MIR-GAN and prior works on LRS2 benchmark. ‘‘S2S’’ denotes sequence-to-sequence loss (Watanabe et al., 2017), ‘‘CTC’’ denotes CTC loss (Graves et al., 2006), ‘‘DataAug’’ denotes noise augmentation, ‘‘LM’’ denotes language model rescoring. The noisy test set is synthesized using MUSAN noise (Snyder et al., 2015).

For the video clips, we detect the 68 facial keypoints using dlib toolkit (King, 2009) and align the image frame to a reference face frame via affine transformation. Then, we convert the image frame to gray-scale and crop a 96×96 region-of-interest (ROI) centered on the detected mouth. During training, we randomly crop a 88×88 region from the whole ROI and flip it horizontally with a probability of 0.5. At inference time, the 88×88 ROI is center cropped without horizontal flipping. To synchronize these two modalities, we stack each 4 neighboring acoustic frames to match the image frames that are sampled at 25Hz.

A.3 Model Settings

Front-ends. We introduce the modified ResNet-18 from prior work (Shi et al., 2022a) as visual front-end, where the first convolutional layer is replaced by a 3D convolutional layer with kernel size of $5 \times 7 \times 7$. The visual feature is flattened into an 1D vector by spatial average pooling in the end. For audio front-end, we use one linear projection layer followed by layer normalization (Ba et al., 2016). **MIR-GAN.** We build the MIR-GAN framework based on Transformer, where the embedding dimension/feed-forward dimension/attention heads in each Transformer layer are set to 1024/4096/16 respectively. In addition, we also employ Conformer as backbone, where the depth-wise convolution kernel size is set to 31. We use a dropout of $p = 0.1$ after the self-attention block within each

Transformer layer, and each Transformer layer is dropped (Fan et al., 2019) at a rate of 0.1.

A.4 Data Augmentation

Following prior work (Shi et al., 2022b), we use many noise categories for data augmentation. We select the noise categories of ‘‘babble’’, ‘‘music’’ and ‘‘natural’’ from MUSAN noise dataset (Snyder et al., 2015), and extract some ‘‘speech’’ noise samples from LRS3 dataset. All categories are divided into training, validation and test partitions.

During training process, we randomly select one noise category and sample a noise clip from its training partition. Then, we randomly mix the sampled noise with input clean audio, at signal-to-noise ratio (SNR) of 0dB with a probability of 0.25.

At inference time, we evaluate our model on clean and noisy test sets respectively. Specifically, the system performance on each noise type is evaluated separately, where the testing noise clips are added at five different SNR levels: $\{-10, -5, 0, 5, 10\}dB$. At last, the testing results on different noise types and SNR levels will be averaged to obtain the final noisy WER result.

A.5 Training Details

Training. We follow the sequence-to-sequence finetuning configurations of AV-HuBERT (Shi et al., 2022b) to train our systems. We use Transformer decoder to decode the encoded features into unigram-based subword units (Kudo, 2018), where

the vocabulary size is set to 1000. The temperature τ in Eq. 7 is set to 0.1, and the weighting parameters $\lambda_{GAN}/\lambda_{MIM}$ in Eq. 8 are set to 0.01/0.005 respectively. The entire system is trained for 60K steps using Adam optimizer (Kingma and Ba, 2014), where the learning rate is warmed up to a peak of 0.001 for the first 20K updates and then linearly decayed. The finetuning process takes \sim 1.4 days on 4 NVIDIA-V100-32GB GPUs.

Inference. No language model is used during inference. We employ beam search for decoding, where the beam width and length penalty are set to 50 and 1 respectively. All the hyper-parameters in our systems are tuned on validation set. Since our experimental results are quite stable, a single run is performed for each reported result.

A.6 Baselines

In this section, we describe the baselines for comparison.

- **TM-seq2seq** (Afouras et al., 2018a): TM-seq2seq proposes a Transformer-based AVSR system to model the A-V features separately and then attentively fuse them for decoding, and uses sequence-to-sequence loss (Watanabe et al., 2017) as training criterion.
- **TM-CTC** (Afouras et al., 2018a): TM-CTC shares the same architecture with TM-seq2seq, but uses CTC loss (Graves et al., 2006) as training criterion.
- **Hyb-RNN** (Petridis et al., 2018): Hyb-RNN proposes a RNN-based AVSR model with hybrid seq2seq/CTC loss (Watanabe et al., 2017), where the A-V features are encoded separately and then concatenated for decoding.
- **RNN-T** (Makino et al., 2019): RNN-T adopts the popular recurrent neural network transducer (Graves, 2012) for AVSR task, where the audio and visual features are concatenated before fed into the encoder.
- **EG-seq2seq** (Xu et al., 2020): EG-seq2seq builds a joint audio enhancement and multi-modal speech recognition system based on the element-wise attention gated recurrent unit (Zhang et al., 2019), where the A-V features are concatenated before decoding.
- **LF-MMI TDNN** (Yu et al., 2020): LF-MMI TDNN proposes a joint audio-visual speech separation and recognition system based on time-delay neural network (TDNN), where the A-V features are concatenated before fed into the recognition network.
- **Hyb-Conformer** (Ma et al., 2021): Hyb-Conformer proposes a Conformer-based (Gulati et al., 2020) AVSR system with hybrid seq2seq/CTC loss, where the A-V input streams are first encoded separately and then concatenated for decoding.
- **MoCo+wav2vec** (Pan et al., 2022): MoCo+wav2vec employs self-supervised pre-trained audio and visual front-ends, *i.e.*, wav2vec 2.0 (Baevski et al., 2020) and MoCo v2 (Chen et al., 2020), to generate better audio-visual features for fusion and decoding.
- **AV-HuBERT** (Shi et al., 2022a,b): AV-HuBERT employs self-supervised learning to capture deep A-V contextual information, where the A-V features are masked and concatenated before fed into Transformer encoder to calculate masked-prediction loss for pre-training, and cross-entropy based sequence-to-sequence loss is used for finetuning.
- **u-HuBERT** (Hsu and Shi, 2022): u-HuBERT extends AV-HuBERT to a unified framework of audio-visual and audio-only pre-training.
- **Distill-Pretrain** (Ma et al., 2022): Distill-Pretrain proposes a Conformer-based VSR framework with additional distillation from pre-trained ASR and VSR models.