Enhanced Self-Supervised Multi-View Representations with Modality-Missing Robustness for Audio-Visual Speech Recognition

Anonymous ICME submission

***Abstract*—Audio-Visual Speech Recognition (AVSR) leverages visual information to enhance speech understanding. However, current models often assume stable, frontal viewpoints, suffering significant performance drops with non-frontal angles or when video signals are partially missing. Our approach first employs a multi-view data generation strategy using 3D head avatar reconstruction, synthesizing viewpoint-diverse data to handle varying head poses. Then, we introduce a self-supervised multi-view representation learning model (MVL) , ensuring viewpoint-invariant and domain-agnostic embeddings. Moreover, a Unified Modality Adapter (UMA) module enables the model to gracefully revert to near audio-only performance levels when visual inputs are unavailable. Experimental results demonstrate that, compared with the baseline AV-HuBERT model, our approach achieves a 3% improvement in lip-reading accuracy under large pose deviations and yields a 16% overall gain in AVSR performance. Moreover, the system significantly outperforms the baseline on real data, confirming the generalizability of our approach under challenging multi-view and modality-missing scenarios. *Code and Data:* https://yakumostudio.github.io/yakumo.github.io/mvss/mvss.html.**

***Index Terms*—Self-Supervised, Multi-view Representations, Modality-Missing, Audio-visual Speech Recognition**

I. Introduction

While early studies highlighted the potential of visual cues (e.g., lip movements) to enhance audio-based speech recognition [1], [2], most existing AVSR approaches assume that the speaker is front-facing and that visual inputs are stable and clear [3], [4]. However, real-world scenarios frequently involve speakers captured from varying angles, partial occlusions, or even intermittent video streams [5], [6]. Such conditions severely degrade performance, sometimes resulting in worse accuracy than simpler audio-only models. Recent advancements in self-supervised and large-scale pre-training strategies, such as AV-HuBERT [7], have improved AVSR performance on standard benchmarks. Nonetheless, these models remain vulnerable to non-frontal viewing angles and missing video modalities [8], [9]. The lack of comprehensive training data reflecting diverse head poses and the absence of robust adaptation techniques for modality-missing scenarios limit their real-world applicability [10], [11]. Hence, addressing the dual challenge of multi-view variability and missing visual inputs is essential for deploying AVSR systems in unconstrained environments.

To this end, we propose a unified framework that enriches training data coverage across a wide range of head poses via multi-view synthetic data generation using 3D head avatar reconstruction [12], [13]. This enables the model to learn viewpoint-invariant lip-reading capabilities. Furthermore, we introduce a self-supervised multi-view representation learning paradigm that integrates a Multi-View Consistency Loss and a Representation Domain Alignment Loss, ensuring that learned embeddings are robust to viewpoint shifts and domain variations. We develop a Unified Modality Adapter (UMA) model to seamlessly revert to audio-only performance levels when visual inputs are unavailable, maintaining the benefits of audio-visual modeling under modality-missing conditions [14].Experimental results on challenging datasets demonstrate significant improvements in recognition accuracy under diverse viewing conditions including cases where visual inputs are partially or fully missing.

II. Related Work

Lipreading and Multi-View Representation Learning. Lipreading, which aims to decode speech solely from visual cues of the speaker's lips, has witnessed substantial progress. Early methods relied on constrained settings and often focused on isolated words or simple phrases [16]. With the introduction of large-scale resources such as LRS2 [17] and LRS3 [3], along with increasingly sophisticated deep learning models [18]–[20], lipreading has advanced towards continuous, sentence-level transcription. Furthermore, some multi-modal fusion methods [21]–[23] have been proposed to enhance audio-visual speech recognition by more effectively combining complementary cues from both modalities. However, these approaches primarily focus on modality integration rather than explicitly addressing viewpoint variations, limiting their effectiveness for large pose deviations or non-frontal angles. Our approach adopts self-supervised pretraining strategy to learn multi-view representations from unlabeled data [24]–[26], enabling the capture of robust cross-modal correlations and visual patterns even under challenging multi-view conditions.

3D Facial Avatar Reconstruction. Early work often avoided explicit 3D structures and relied on image-based warping or implicit representations [27], [28]. However, these approaches typically struggled with significant pose shifts due to limited geometric consistency. 3D Morphable Models, such as FLAME [29], [30], introduced low-dimensional shape and expression spaces, serving as foundations for mesh-based methods that ensure geometric accuracy and temporal stability [31]. Recent advances in 3D head reconstruction have been driven

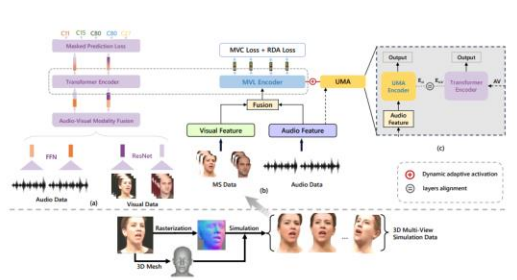

Fig. 1. Overview of the proposed framework. (a) AV-HuBERT-based Audio-Visual Encoder, computes the masked multi-modal prediction loss [15] on real data. FFN denotes a feed-forward network for audio extraction, and ResNet is a modified ResNet for video feature extraction. (b) The self-supervised Multi-View Representation Learning (MVL) model, which leverages multi-view simulated data (MS Data) generated by the lower pipeline. Parameters in the MVL Encoder are shared with the Transformer Encoder to ensure consistent representation learning. (c) The Unified Modality Adapter (UMA) model, where "AV" represents fused audio-visual features. "$\mathbf{E}_a$" denotes UMA encoder audio embeddings and "$\mathbf{E}_{av}$" denotes the audio-visual encoder embeddings for the same input. When UMA is activated, audio features skip the modality fusion step and pass through UMA, providing robustness to missing visual inputs.

by approaches that unify neural rendering and deep learning-based geometry estimation. Grassal et al. [12] proposed a method to generate neural head avatars from monocular RGB videos, leveraging deep neural networks to produce dense 3D geometry and personalized appearance. Neural rendering, combined with parametric priors, now enables photo-realistic and controllable avatars that better preserve subtle lip movements critical for AVSR under multi-view conditions [12], [32], [33].

Robustness to Missing Video Modality. Real-world audio-visual speech recognition must cope with scenarios where visual inputs are degraded, partially visible, or entirely absent [8], [34]. Early strategies attempted random masking of video frames, encouraging models to rely more on audio cues when the visual modality faltered [5], [35]. Beyond simple dropout, current methods explore architectures and training regimes that ensure the presence of visual features never diminishes performance [11]. Novel frameworks incorporate multi-view data and modality-adaptive modules to seamlessly revert to audio-only quality levels under adverse conditions [21], [36], thereby offering robust performance across a wide range of viewing angles and video availability scenarios.

III. Methodology

Our proposed framework is illustrated in Fig. 1, the methodology involves synthesizing multi-view data via 3D reconstruction, learning viewpoint-invariant and domain-aligned representations through self-supervised training.

A. Multi-View Data Generation Strategy

1) Data Preparation: As shown in Fig. 2, We first perform head pose estimation on each video segment by solving a Perspective-n-Point (PnP) problem via 2D facial landmarks, thereby extracting Euler angles [37], [38]. Subsequently, we select video segments exhibiting substantial rotational changes as the primary source material for multi-view simulation.

2) Multi-View Data Simulation: We proceed to generate additional viewing angles using a 3D head avatar reconstruction approach. Instead of relying on multiple cameras or extensive capture setups, we employ a learned neural avatar model that reconstructs a detailed, animatable 3D representation of the speaker from RGB video sequence [12], [27], [28]. We adopt a 3D morphable model (e.g., FLAME [29]) as a geometric backbone. Given shape parameters $\boldsymbol{\alpha}$, expression parameters $\boldsymbol{\epsilon}$, and pose parameters $\boldsymbol{\theta}$, the base model provides a coarse mesh:

$$\mathbf{M}_{\text{base}}(\boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\theta}) \in \mathbb{R}^{N_v \times 3}, \tag{1}$$

where N_v is the number of vertices. We use geometry refinement function $\mathbf{G}$ to capture fine details not represented by the

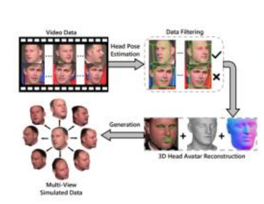

Fig. 2. Multi-view data generation strategy.

template:

$$\mathbf{M}(\boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\theta}) = \mathbf{M}_{\text{base}}(\boldsymbol{\alpha}, \boldsymbol{\epsilon}, \boldsymbol{\theta}) + \mathbf{G}(\boldsymbol{\theta}). \tag{2}$$

The coarse geometric alignment and fine-grained texture refinement ensure that lip contours align closely with the original frames. [21], [39], [40].

B. Self-Supervised Multi-View Representation Learning

Leveraging self-supervised training to capture intrinsic audio-visual correlations [23], [25], we utilize a feature extractor and encoder structure inspired by AV-HuBERT [24] to learn viewpoint-invariant, domain-aligned audio-visual embeddings.

1) Multi-View Consistency Loss: Consider a batch of N paired samples, where $\{\mathbf{x}_{r,i}\}_{i=1}^N$ are real audio-visual sequences and $\{\mathbf{x}_{s,i}\}_{i=1}^N$ are their corresponding synthesized multi-view counterparts. We seek an embedding function $f(\cdot)$ that maps these inputs into a latent space invariant to viewpoint changes. Let $f(\mathbf{x}_{r,i})$ and $f(\mathbf{x}_{s,i})$ be embeddings with dimensions $T \times D$, where T is the combined spatio-temporal dimension and D is the feature channel dimension. We first consider an element-wise alignment cost to ensure that real and synthetic embeddings match at the feature level. Defining a mean squared error term:

$$L_{\text{mse}} = \frac{1}{N} \sum_{i=1}^{N} \| f(\mathbf{x}_{r,i}) - f(\mathbf{x}_{s,i}) \|_2^2. \tag{3}$$

However, achieving semantic alignment requires more than just element-wise agreement. To address this, we impose a correlation alignment term. Consider normalized and standardized embeddings $\mathbf{Z}_r, \mathbf{Z}_s \in \mathbb{R}^{T \times D}$. Each embedding $\mathbf{Z}_m$ is first normalized appropriately (mean subtraction, standard deviation division, and row-wise ℓ_2 normalization). We measure the difference between these correlation matrices using the Frobenius norm:

$$L_{\text{corr}} = \frac{1}{N} \sum_{i=1}^{N} \| \text{Corr}(\mathbf{Z}_{r,i}) - \text{Corr}(\mathbf{Z}_{s,i}) \|_F^2. \tag{4}$$

Minimizing L_{corr} ensures that real and synthetic embeddings share not only similar feature magnitudes but also analogous internal semantic structures. Combining these two terms:

$$L_{\text{MVC}} = \alpha L_{\text{corr}} + (1 - \alpha) L_{\text{mse}}. \tag{5}$$

2) Representation Domain Alignment Loss: To bridge the gap between real and synthetic distributions, a contrastive objective is employed to facilitate the learning of domain-invariant features. Specifically, we anchor on near-frontal (or minimally rotated) synthetic samples of the same speaker as positive samples, pairing them with corresponding real samples. Negative samples are constructed using embeddings from other data within the same batch. This setup encourages the model to focus on critical speech-related features, such as lip motion, while disregarding irrelevant factors like background, texture, and lighting. Formally, let $f(\cdot)$ be the embedding function, and $(\mathbf{x}_{\text{real},i}, \mathbf{x}_{\text{sim},i})$ be a real-synthetic pair of the same speaker with minimal head rotation differences. Positive similarity is defined as:

$$\text{pos\_sim} = \frac{\langle f(\mathbf{x}_{\text{real},i}), f(\mathbf{x}_{\text{sim},i}) \rangle}{\| f(\mathbf{x}_{\text{real},i}) \| \cdot \| f(\mathbf{x}_{\text{sim},i}) \|}. \tag{6}$$

For negatives, we use samples $(\mathbf{x}_{\text{neg},j})$ from other data, treated as dissimilar:

$$\text{neg\_sim} = \frac{\langle f(\mathbf{x}_{\text{real},i}), f(\mathbf{x}_{\text{neg},j}) \rangle}{\| f(\mathbf{x}_{\text{real},i}) \| \cdot \| f(\mathbf{x}_{\text{neg},j}) \|}. \tag{7}$$

L_{RDA} becomes:

$$L_{\text{RDA}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{pos\_sim}/\tau)}{\sum_{j=1}^{M} \exp(\text{neg\_sim}/\tau)}, \tag{8}$$

where τ is a temperature parameter and M is the number of negative samples. To align with downstream phoneme or viseme categories for effective lip-reading, we integrate a Masked Multi-Modal Prediction Loss (L_{MMP}). The MVL model training objective combines the three losses:

$$L_{\text{MVL}} = \lambda_{\text{MVC}} L_{\text{MVC}} + \lambda_{\text{RDA}} L_{\text{RDA}} + \lambda_{\text{MMP}} L_{\text{MMP}}, \tag{9}$$

where λ_{MMP}, λ_{MVC}, and λ_{RDA} are weighting factors.

C. Unified Modality Adapter

The UMA model consists of transformer layers, each receiving two inputs: the output of the corresponding layer of the MVL encoder (only audio feature) and the output of the previous UMA layer. This architectural setup ensures a progressive refinement of the audio-only features towards audio-visual-like representations. Let $\mathbf{E}_a$ denote the UMA-enhanced audio embeddings and $\mathbf{E}_{av}$ the corresponding audio-visual encoder embeddings for the same input content. We define a combined loss:

$$L_{\text{UMA}} = \gamma L_{\text{MMP}} + (1 - \gamma) L_{\text{FS}}, \tag{10}$$

where L_{MMP} is the masked multi-modal prediction loss and L_{FS} is the feature similarity loss. We designate a set of early UMA layers as P and deeper layers as Q. For layers in P, we align the internal correlation structures between $\mathbf{E}_{av}$ and $\mathbf{E}_a$ to ensure semantic consistency, while for layers in Q, we directly minimize an element-wise feature discrepancy:

$$L_{\text{FS}} = \sum_{p \in P} \frac{\lambda_p}{T \times T} \| \text{Corr}(\mathbf{E}_{av}^{(p)}) - \text{Corr}(\mathbf{E}_a^{(p)}) \|_F^2 + \sum_{q \in Q} \frac{1 - \lambda_p}{T \times D} \| \mathbf{E}_{av}^{(q)} - \mathbf{E}_a^{(q)} \|_F^2. \tag{11}$$

where λ_p is a weighting factor controlling the importance of correlation alignment versus direct feature-level alignment.

IV. Experiments

A. Experimental Settings

We evaluate our approach on LRS3 [3] and OuluVS2 [41]. Two subsets of LRS3 are considered: a 30-hour subset (30h) and the full 433-hour set (433h). During 3D reconstruction, hyperparameters for geometry and texture refinement are tuned to preserve crucial lip motion features. Specifically, we use 150 epochs for geometry offset optimization, 80 for texture refinement, and 50 for joint geometry-texture optimization. Angles range from -25° to 25°, sampled every 5°, plus random angles in [-10°,10°] for subtle variations. After synthesis, we form multi-view extended sets, ms30h: 30h set combining 30% synthetic and 70% real data. ms433h: 433h set combining 40% synthetic and 60% real data. OuluVS2, which comprises more than 20k video recordings captured simultaneously from five different views (0°, 30°, 45°, 60°, 90°), is utilized to evaluate generalization under more realistic and challenging conditions. For modality-missing evaluation, we randomly mask the video modality at rates from 10% to 100%.

B. Training Details

Pretraining: We initialize from the 5th iteration AV-HuBERT base model, which provides cluster pseudo-labels for the 6th iteration. This baseline is trained and fine-tuned following the official protocols [15]. Our MVL model's encoder consists of 12 transformer blocks, and adopt a ResNet18-based front-end for visual feature extraction. Lip videos are resized to 88×88 with a single channel. For the 30h setup, during the first half of the updates, we set λ_{MMP}=0.8, λ_{MVC}=0.1, λ_{RDA}=0.1, τ=0.07. In the latter half, as the model gradually learns multi-view representations, λ_{MVC} and λ_{RDA} are gradually reduced, stopping once the negative sample similarity in λ_{RDA} falls below 0.002. For the 433h data, the same schedule applies, but λ_{MVC} and λ_{RDA} weights are reduced over the first third of the updates. When training the UMA model, the encoder weights remain frozen. Audio-only inputs pass through both the encoder and the UMA adapter. We set $\gamma = 0.8$ and $\lambda_p = 0.2$.

Fine-tuning: For fine-tuning the MVL model with visual modality only, we use an attention-based sequence-to-sequence cross-entropy loss. For UMA model fine-tuning, we include both modalities, apply video masking at rates of {0.0, 0.1, 0.2, …, 1.0}, expanding 30 hours of original data into a total of 330 hours. All other hyperparameters remain consistent with the AV-HuBERT 5th iteration baseline to ensure a fair comparison.

TABLE I. *Comparison of our approach and AV-HuBERT. "0" corresponds to the standard LRS3 test subset. For VSR, we incorporate multi-view synthetic data with angles of ±5° or ±10°. For AVSR, we apply 10% and 30% masking rate on video frames. † indicates that AV-HuBERT results are reproduced using the official code from [15], [24].*

Model	Labeled data (hrs)	Unlabeled data (hrs)	Task	WER (%) 0	5°	10°	0.1	0.3
AV-HuBERT†	30	30	VSR	81.3	89.6	93.2	-	-
	30	433	VSR	54.3	69.0	77.1	-	-
	433	433	VSR	43.7	63.9	73.3	-	-
	30	30	AVSR	6.7	-	-	6.8	6.7
	30	433	AVSR	5.8	-	-	6.1	6.1
Ours (MVL)	ms30	ms30	VSR	74.5	86.4	90.9	-	-
	ms30	ms433	VSR	55.0	68.9	74.8	-	-
	ms433	ms433	VSR	**43.2**	**62.9**	**71.9**	-	-
Ours (MVL+UMA)	ms30+mask	ms30	AVSR	5.6	-	-	5.5	5.7
	ms30+mask	ms433	AVSR	**5.0**	-	-	**5.2**	**5.2**

C. Experimental Results and Analysis

1) Overall Performance: Table I presents a comprehensive comparison of our models with the baseline AV-HuBERT. Without additional rotations or modality perturbations, the baseline AV-HuBERT achieves a WER of 5.8% when using 30 labeled hours and 433 unlabeled hours for AVSR tasks. By integrating the proposed MVL model, which leverages multi-view data and self-supervised losses (L_{MVC}, L_{RDA}, L_{MMP}), the WER further improves. Specifically, for the same data setting, our MVL model attains a lower WER of 43.2% for lipreading. Further incorporating the UMA module bolsters robustness against missing visual cues, maintaining a 5.0% WER in full audio-visual scenarios and showing stable performance even when video input is partly unavailable. In general, the final performance of the AVSR improves by about 16% relative to the baseline, highlighting the effectiveness of multi-view training and modality-invariant adaptation.

2) Evaluation on Multi-View and Modality-Missing data: We further evaluate the model under various multi-view conditions, including horizontal yaw angles up to 20°, combined yaw-pitch deviations, and partial modality-missing scenarios (show in Fig. 3). In the 0°–20° range, our approach achieves an overall 6% improvement in lip-reading performance, demonstrating superior robustness compared to the baseline. Even under more severe pose variations in both yaw and pitch, the MVL models maintain stable accuracy levels, consistently outperforming the baseline. When UMA is introduced to handle missing visual inputs, the model gains significant resilience to missing video inputs: at 20% video masking rate, it surpasses the baseline by 16.7%. This gain stems not only from the robustness of UMA to incomplete modalities but also from the ability of MVL to extract richer lip-motion cues from diverse viewpoints, thereby enhancing overall audio-visual speech recognition capability.

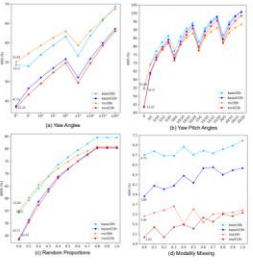

Fig. 3. Evaluation on multi-view and modality-missing data. *base30h* and *base433h* denote the AV-HuBERT base model finetuned on 30h and 433h datasets. *mv30h* and *mv433h* represent the MVL model under ms30h and ms433h. *ma30h* and *ma433h* indicate the UMA model finetuned on 30h and 433h data. (a) Yaw Angles: gradually introducing different horizontal angles into the test set; (b) Yaw Pitch Angles: incorporating both horizontal and vertical pose variations; (c) Random Proportions: evaluating varying proportions of multi-view synthetic data; (d) Modality Missing: applying distinct masking ratios to the test data.

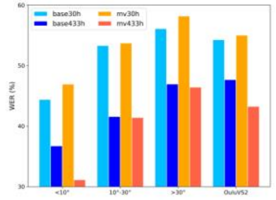

Fig. 4. Generalization performance evaluation. For the LRS3 test set, we split samples by face yaw into three subsets: $<10°$, 10°-30°, and $>30°$. For OuluVS2, we follow the data evaluation protocol of [42].

3) Generalization Performance Evaluation: We conduct evaluations on four categories of genuine test data: three LRS3 subsets grouped by increasing face yaw angles ($<10°$, 10°-30°, $>30°$) and the OuluVS2 dataset, which presents substantial viewpoint variations. As illustrated in Figure 4, when yaw angles stay within 10°, the MVL model trained on *ms433h* data achieves a WER of 31.1%, significantly outperforming the baseline. Even as yaw angles widen, our models continue to exhibit robust recognition performance on authentic multi-view samples. Notably, on OuluVS2, which features many large or even profile-angle views, the MVL approach leverages its strong multi-view lip-motion modeling to attain a WER of 43.25%, thereby showcasing its effectiveness under extreme pose conditions.

4) Ablation Studies: To further dissect the contributions of each component, we analyze performance under different configurations. Table II presents that without MVL, the baseline model struggles to align multi-view data, causing inconsistent gains or even degradations. Adding masking alone under a 30h fine-tuning setup provides limited improvements but does not fully capture multi-view representations. Notably, using only L_{MVC} cannot effectively address background, illumination, or other extraneous factors, leading to suboptimal representation learning. Introducing L_{RDA} substantially enhances the model's multi-view capability (WER from 65.8% to 55.0%). Finally, equipping the model with UMA effectively mitigates the impact of missing or partially corrupted video signals. When multi-view data and the associated losses are collectively employed, the resulting model not only gains robustness to visual modality absence but also utilizes lip-motion information more effectively, yielding further improvements in audio-visual speech recognition.

TABLE II. *Comparison among AV-HuBert BASE models and our proposed models under diverse data and loss settings on the test set. AV-HuBert is pre-trained using 433h data and fine-tuned on 30h data. "Multi-View" indicates models pre-trained on* `ms433` *data and fine-tuned on* `ms30` *data, while "Mask" signifies additional fine-tuning with masked samples.*

Method	Data: Multi-View	Data: Mask	Loss: L_{MVC}	Loss: L_{RDA}	Loss: L_{FS}	WER (%)
AV-HuBERT	-	-	-	-	-	**54.3**
	✓	-	-	-	-	65.0
	-	✓	-	-	-	6.8
	✓	✓	-	-	-	7.1
+ MVL	✓	-	✓	-	-	65.8
	✓	-	-	✓	-	56.3
	✓	-	✓	✓	-	55.0
+ UMA	-	✓	-	-	✓	6.4
	✓	✓	✓	✓	✓	**5.4**

V. Conclusion

In this paper, we presented an enhanced self-supervised AVSR model that incorporates multi-view representation learning and modality-missing robustness. By leveraging 3D facial avatar reconstruction, we generated multi-view speaker face videos, enabling the model to learn robust lip-reading capabilities across different head poses. The adapter module allows the model to handle scenarios where the visual modality is unreliable or missing by adjusting its reliance on visual features. Our experiments demonstrate that the proposed model consistently surpasses the AV-HuBERT model in both multi-view lip-reading and audio-visual speech recognition under missing-modality conditions. Furthermore, evaluations on real-world multi-view datasets validate the effectiveness of our approach in practical scenarios.

References

[1] Eric David Petajan, *Automatic lipreading to enhance speech recognition (speech reading)*, University of Illinois at Urbana-Champaign, 1984.
[2] Gerasimos Potamianos, Hans Peter Graf, and Eric Cosatto, "An image transform approach for hmm based automatic lipreading," in *Proc. ICIP*. IEEE, 1998, pp. 173–177.
[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
[4] Kenji Mase and Alex Pentland, "Automatic lipreading by optical-flow analysis," *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67–76, 1991.
[5] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[6] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proc. AAAI*, 2020, vol. 34, pp. 279–286.
[7] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *Proc. ICLR*, 2022.
[8] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia, "Modality attention for end-to-end audio-visual speech recognition," in *ICASSP*. IEEE, 2019, pp. 6565–6569.
[9] Koji Iwano, Tomoaki Yoshinaga, Satoshi Tamura, and Sadaoki Furui, "Audio-visual speech recognition using lip information extracted from side-face images," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. 1–9, 2007.
[10] Shiliang Zhang, Ming Lei, Bin Ma, and Lei Xie, "Robust audio-visual speech recognition using bimodal dfsmn with multi-condition training and dropout regularization," in *Proc. ICASSP*. IEEE, 2019, pp. 6570–6574.
[11] Srinivas Parthasarathy and Shiva Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Proc. ICMI*, 2020, pp. 400–404.
[12] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies, "Neural head avatars from monocular rgb videos," in *Proc. CVPR*, 2022, pp. 18653–18664.
[13] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," *afouras2018lrs3SIGGRAPH*, pp. 187–194, 1999.
[14] Xinmeng Chen, Xuchen Gong, Ming Cheng, Qi Deng, and Ming Li, "Cross-modal assisted training for abnormal event recognition in elevators," *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.
[15] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.
[16] Christoph Bregler and Yochai Konig, ""eigenlips" for robust speech recognition," in *Proc. ICASSP*. IEEE, 1994, pp. II–669.
[17] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, 2017, pp. 6447–6456.
[18] Themos Stafylakis and Georgios Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proc. Interspeech*, 2017, pp. 3652–3656.
[19] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.
[20] Yiting Li, Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *Proc. ICIS*. IEEE, 2016, pp. 1–6.
[21] Liqiang Nie, Mengzhao Jia, Xuemeng Song, Ganglu Wu, Harry Cheng, and Jian Gu, "Multimodal activation: Awakening dialog robots without wake words," in *Proc. SIGIR*, 2021, pp. 491–500.
[22] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. ICASSP*, 2021, pp. 7613–7617.
[23] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "ASR is all you need: Cross-modal distillation for lip reading," in *Proc. ICASSP*, 2020, pp. 2143–2147.
[24] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," in *Proc. Interspeech*, 2022, pp. 2118–2122.
[25] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, et al., "Large-Scale Visual Speech Recognition," in *Proc. Interspeech*, 2019, pp. 4135–4139.
[26] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," in *Proc. CVPR*, 2018.
[27] Aliaksandr Siarohin, Stephane Lathuiliere, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "First order motion model for image animation," in *Proc. NeurIPS*, 2019.
[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, et al., "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proc. ECCV*, 2020, pp. 405–421.
[29] Zhenhua Li, Tao Yu, Hao Li, Guo-Jun Zhang, and Qionghai Dai, "Learning a model of facial shape and expression from 4d scans," in *Proc. SIGGRAPH Asia*, 2017.
[30] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces.," in *Proc. SIGGRAPH*, 1999, vol. 99, pp. 187–194.
[31] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. CVPR*, 2016, pp. 2387–2395.
[32] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, et al., "Deep video portraits," *arXiv preprint arXiv:1805.11714*, 2018.
[33] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. CVPR*, 2021, pp. 13325–13333.
[34] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," *CVPR*, pp. 3444–3453, 2017.
[35] Takaki Makino et al., "Recurrent neural network transducer for audio-visual speech recognition," in *Proc. ASRU*. IEEE, 2019, pp. 905–912.
[36] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic, "LiRA: Learning Visual Speech Representations from Audio Through Self-Supervision," in *Proc. Interspeech*, 2021, pp. 3011–3015.
[37] Xiangxin Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. CVPR*, 2015, pp. 787–796.
[38] Luan Tran and Xiaoming Liu, "On learning 3d face morphable model from in-the-wild images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 157–171, 2021.
[39] Xiaoning Song, Zhen-Hua Feng, Guosheng Hu, Josef Kittler, and Xiao-Jun Wu, "Dictionary integration using 3d morphable face models for pose-invariant collaborative-representation-based classification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2734–2745, 2018.
[40] Kejun Wang, Guofeng Zou, Guixia Fu, and Lei Yuan, "An approach to fast eye location and face plane rotation correction," *Journal of Computer-Aided Design & Computer Graphics*, vol. 25, no. 6, pp. 865–872, 2013.
[41] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proc. FG*. IEEE, 2015, vol. 1, pp. 1–5.
[42] Houjeung Han, Sunghun Kang, and Yoo, "Multi-view visual speech recognition based on multi task learning," in *Proc. ICIP*. 2017, pp. 3983–3987, IEEE Computer Society.