Restoring Speaking Lips from Occlusion for Audio-Visual Speech Recognition

Jiadong Wang[1,3], **Zexu Pan**[1], **Malu Zhang**[2*], **Robby T. Tan** [1], **Haizhou Li** [3,1]

[1]National University of Singapore
[2]University of Electronic Science and Technology of China
[3]Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
jiadong.wang@u.nus.edu, maluzhang@uestc.edu.cn

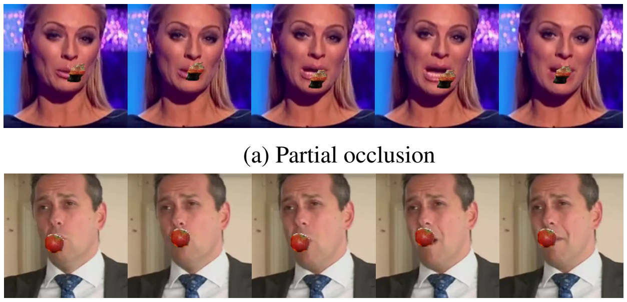
Abstract

Prior studies on audio-visual speech recognition typically assume the visibility of speaking lips, ignoring the fact that visual occlusion occurs in real-world videos, thus adversely affecting recognition performance. To address this issue, we propose a framework that restores occluded lips in a video by utilizing both the video itself and the corresponding noisy audio. Specifically, the framework aims to achieve these three tasks: detecting occluded frames, masking occluded areas, and reconstruction of masked regions. We tackle the first two issues by utilizing the Class Activation Mapping (CAM) obtained from occluded frame detection to facilitate the masking of occluded areas. Additionally, we introduce a novel synthesis-matching strategy for the reconstruction to ensure the compatibility of audio features with different levels of occlusion. Our framework is evaluated in terms of Word Error Rate (WER) on the original videos, the videos corrupted by concealed lips, and the videos restored using the framework with several existing state-of-the-art audio-visual speech recognition methods. Experimental results substantiate that our framework significantly mitigates performance degradation resulting from lip occlusion. Under -5dB noise conditions, AV-Hubert's WER increases from 10.62% to 13.87% due to lip occlusion, but rebounds to 11.87% in conjunction with the proposed framework. Furthermore, the framework also demonstrates its capacity to produce natural synthesized images in qualitative assessments.

Introduction

Audio-visual speech recognition aims to transcribe spoken words by utilizing both lip movements and speech signals. It has recently gained substantial attention because of the lip-reading module's capability to maintain robustness in the presence of acoustic noise, which used to significantly degrade speech recognition performance. Additionally, this lip-reading module finds widespread application in other speech-related tasks, such as speech extraction (Pan et al. 2021; Pan, Ge, and Li 2022) and speaker verification (Shi et al. 2022).

Many current methods in lip-reading or audio-visual speech recognition assume that the lips are clearly visible in the video, without any occlusion. However, this assumption is overly optimistic in the real world, as a microphone or a hand can easily obstruct the view of a speaking lip (Hong et al. 2023). Unfortunately, state-of-the-art audio-visual speech recognition methods (Hong et al. 2022; Ma, Petridis, and Pantic 2021) are known to be susceptible to lip occlusion, resulting in an increase in error rate of up to 65% (Hong et al. 2023). To address this significant degradation in audio-visual speech recognition, a study introduced a module for dynamically assessing modality reliability. This module is integrated into the audio-visual speech recognition network to circumvent the utilization of corrupted visual frames in videos affected by lip occlusion (Hong et al. 2023). However, this approach does not involve the restoration of occluded lips in corrupted videos, and as a result, it cannot directly leverage the capabilities of existing audio-visual speech recognition methods that assume no lip occlusion.



(a) Partial occlusion

(b) Full occlusion

Figure 1: Examples of lip occlusion.

Inspired by speech enhancement techniques (Zhang et al. 2020, 2022) that restore clear speech from noisy recordings, we aim to mitigate the adverse effects of lip occlusion on audio-visual speech recognition by restoring occluded lips. Considering the partially occluded lip shown in Fig. 1a as an example, our goal is to remove the object occluding the lip and restore the occluded region. When the lip is completely occluded, as illustrated in Fig. 1b, restoring the lip using solely the input image becomes a challenging task. In such instances, we depend on synchronized audio to supplement the visual information.

[*]Corresponding author

To achieve lip restoration, our objectives encompass three tasks: (1) detecting occluded frames, (2) masking occluded areas, and (3) reconstruction of masked regions.

Based on the concept of Class Activation Map (CAM) (Zhou et al. 2016; Jin, Sharma, and Tan 2021), which demonstrates how a classification network can roughly localize a subject class within an image, we employ the CAM from occluded frame detection to aid in masking occluded areas. Here, optimizing the detection of occluded frames involves a loss function that classifies an image into occluded or occlusion-free categories.

The diverse range of occlusion levels presents a significant challenge in the reconstruction of masked regions. While an audio feature correlates with the shape of a fully synchronized lip, completing the lip with varying degrees of occlusion using the same audio feature poses a complex task. To tackle this challenge, we introduce a novel synthesis-matching strategy for the reconstruction that employs audio features to envision a complete lip during the synthesis stage and then enhances the complete lip using unmasked regions during the matching stage.

Specifically, during the synthesis step, we follow the approach of talking face generation (Prajwal et al. 2020; Wang et al. 2023; Park et al. 2022) by masking the lower half of the occluded face. This supplementary mask confines the audio feature's role to imagining only the lower half of the face containing the complete lips. Additionally, we utilize a detected occlusion-free frame as a visual reference for the mouth's appearance, aiding the audio feature in envisioning the complete lip. Nonetheless, the synthesized image from the synthesis step could be inaccurate due to the presence of noisy audio features. Thus, in the matching step, we input a combination of the synthesized image and the occlusion-masked face into an auto-encoder-decoder network, producing a high-fidelity image. In summary, this paper makes the following contributions:

- Given the corresponding noisy audio and the video, we address the adverse effect of lip occlusion on audio-visual speech recognition by restoring occluded lips in a video. We propose a framework of audio-visual lip restoration (AVLR) to achieve this target.
- We introduce a novel synthesis-matching strategy to make audio features compatible with various degrees of occlusion while obtaining high-fidelity images.
- Experiments confirm that our method can substantially alleviate the deterioration of audio-visual speech recognition resulting from lip occlusion. Under -5dB noise conditions, AV-Hubert's WER increases from 10.62 to 13.87 due to lip occlusion, but rebounds to 11.87 in conjunction with the proposed framework.

Related Work

Audio-visual Speech Recognition Audio-visual speech recognition leverage lip reading to complement speech recognition, since speech recognition is vulnerable under noise but more informative than lip reading without noise (Wang, Qian, and Li 2022).

Audio-visual fusion is a unique module in audio-visual speech recognition, distinguishing it from speech recognition or lip reading. Concatenation or multiplication of features from feature extractors or encoders are common strategies (Afouras et al. 2018; Petridis et al. 2018; Yu et al. 2020). Besides, cross-modal attention is adopted to enhance interaction between audio and visual modalities (Sterpu, Saam, and Harte 2020; Paraskevopoulos et al. 2020). Different from feature fusion, there are some works that conduct a fusion of prediction from different modalities via some strategies, such as addition with fixed weights (Luettin, Potamianos, and Neti 2001) or dynamic weights (Stewart et al. 2013; Abdelaziz, Zeiler, and Kolossa 2015). However, the above fusion methods only consider audio corruption and do not pay attention to the corruption of videos.

In (Hong et al. 2023), the authors claim to be the first to address video corruption in audio-visual speech recognition. They tackle this issue by enhancing fusion methods, specifically by calculating reliability scores for modalities and merging features using attention and scores. In (Afouras, Chung, and Zisserman 2019), for audio-visual enhancement, the authors employ a two-step strategy to obtain speaker embedding in the first round when the visual modality is corrupt. However, unlike our AVLR, both two methods do not attempt to restore occluded lips from corrupted videos.

In (Yu et al. 2021), the authors evaluate their audio-visual speech recognition method in various scenarios, including occluded videos and occluded videos treated with an inpainting method. However, this inpainting method does not utilize audio information, posing a challenge when dealing with fully occluded lips.

Talking Face Generation Talking face generation is to synthesize videos of lip movements consistent with given audio and identity images (Prajwal et al. 2020; Wang et al. 2023) and optional pose cues (Zhou et al. 2021) and emotional cues (Liang et al. 2022). Methods of talking face generation can be categorized into reconstruction and intermediate methods (Park et al. 2022). Intermediate methods follow a media (e.g. landmarks (Chen et al. 2019; Zhou et al. 2020) or 3D meshes (Song et al. 2022)) to bridge inputs and outputs. Reconstruction methods mostly employ an auto-encoder-decoder architecture with extra discriminators to penalise degraded visual quality (Zhou et al. 2019), inaccurate synchronization (Prajwal et al. 2020), and reading intelligibility (Wang et al. 2023).

Talking face generation is different from audio-visual lip restoration as follows: 1). Lip restoration can only use noisy audio while talking face generation usually adopts clean speech. 2). Talking face generation constructs entire lips depending on audio features but lip restoration might need to restore lips under different degrees of occlusion given the same audio clip.

Proposed Method

The use of our audio-visual lip restoration (AVLR) framework for audio-visual speech recognition under adverse conditions of acoustic noise and lip occlusion is depicted in Fig. 2. Given a pair of a corrupted video affected by lip occlusion

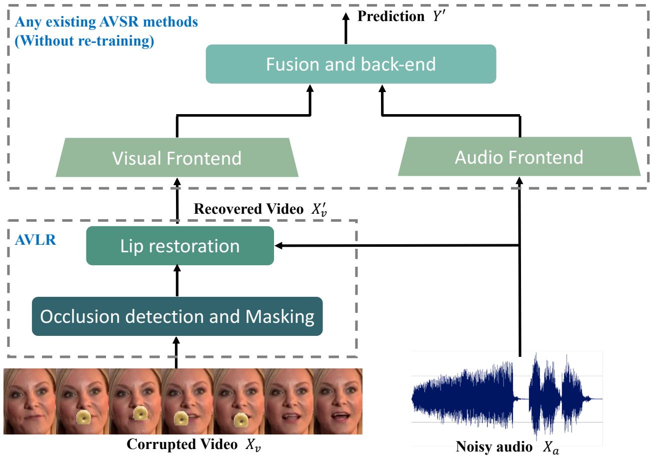


Figure 2: Audio-visual speech recognition with audio-visual lip restoration (AVLR).

$X_v \in \mathbb{R}^{T\times 3\times H\times W}$ and a noisy audio feature $F_a \in \mathbb{R}^{T\times d_a}$ extracted from noisy audio X_a, our AVLR restores occluded lips and outputs the recovered video X_v', where T, d_a, d_v are numbers of frames, visual feature and audio feature dimensions, H and W are height and width. Subsequently, any existing trained audio-visual speech recognition methods mentioned in the related work section can be employed to process X_v' and X_a to predict spoken words $Y' \in \mathbb{R}^{L\times C}$ without re-training, where L and C are the length of spoken words and the number of potential word classes, respectively.

AVLR, illustrated in Fig. 3, restores occluded lips to mitigate the degradation of audio-visual speech recognition caused by lip occlusion. Specifically, we describe the AVLR using two subsections below: (1) detecting occluded frames and generating masks of occlusion to mask the occlusion; and (2) reconstructing masked regions given an occlusion-masked image I_{orm}, an image providing a cue of mouth appearance I_s, and a synchronized noisy audio clip X_a.

Occluded-Frame Detection and Mask Generation

In this phase, our goal is to determine the occluded frames within a video and compute the respective occlusion masks for lip restoration. Furthermore, an additional objective for each occluded frame is to choose the temporally nearest unobstructed frame as a reference for mouth appearance.

As shown in Fig. 3, the detection network contains a downsampling module, a Resnet (He et al. 2016), a global average pooling layer and a linear layer. The detection network is trained with a binary cross-entropy loss:

$$\mathcal{L}_{\text{det}} = y_d \cdot log(\mathcal{F}_d(x_v^i)) + (1 - y_d) \cdot log(1 - \mathcal{F}_d(x_v^i)), \quad (1)$$

where $y_d \in [0, 1]$ is the binary ground truth to indicate whether an image contains a concealed lip, x_v^i denotes the i-th image in a video.

As the detection network progressively learns to differentiate whether the lip in an image is occluded, CAM can provide a rough localization of the occlusion (Zhou et al. 2016), which is expressed as:

$$m_{cam} = \sum^{k} w^k \cdot f_g^k, \quad (2)$$

where f_g is the feature before the global average pooling layer, w is the weight of the linear layer after the global average pooling layer, and k is the channel index of f_g.

To refine CAM, we adopt another Resnet and feed it with the concatenation of the CAM and f_{down}. The refined mask m_{fine} is optimized also by a binary cross-entropy loss with mask labels:

$$\mathcal{L}_{\text{mask}} = \frac{1}{HW}\sum_{H}\sum_{W}(m_{gt}^{h,w}\log(m_{fine}^{h,w}) + (1 - m_{gt}^{h,w})\log(1 - m_{fine}^{h,w})), \quad (3)$$

where $m_{gt}^{h,w} \in [0, 1]$ represents the mask ground truth, indicating whether the pixel at position ”h, w” within an image belongs to an occlusion region.

Subsequently, we apply the mask m_{fine} on the corresponding occluded image as Eq. 4. This leads to the creation of an occlusion-masked image denoted as I_{orm}.

$$I_{orm} = m_{fine} \cdot I_{occ}. \quad (4)$$

Reconstruction of Missing Lip Regions

In this phase, our goal is to complete the masked regions of a lip given an occlusion-masked image I_{orm}, an image I_s providing mouth appearance, and a noisy audio clip X_a. Mouth appearance herein refers to a human-dependent aspect instead of human-independent visemes which are analogous to phonemes in the context of lip reading (Bear and Harvey 2017).

To reduce the disparity in pose between I_{orm} and I_s, we opt for an occlusion-free image that is temporally closest to I_{orm} as the appearance cue I_s. During training, we randomly select I_s to prevent the chance of the nearest occlusion-free frame having a similar mouth shape (viseme) to that of I_{orm}. As mentioned in the introduction, we propose a synthesis-matching strategy to make audio features compatible with different extents of occlusion. In the synthesis module, audio clips are designed to generate the full lip. To this end, we mask the lower-half face of $Iorm$ and obtain I_{2m}, which removes the influence of the residual visible mouth on the synthesis module and ensures that the synthesis module generates a mouth that depends solely on F_a and the appearance cue I_s. This pre-processing is widely used in talking face generation (Prajwal et al. 2020; Wang et al. 2023; Park et al. 2022).

To synthesize an image with a lip shape consistent with the audio clip, the synthesis module adopts an auto-encoder-decoder architecture. The encoder of the synthesis module down-samples the concatenation of the I_{2m} and the I_s to a smaller size (1×1). The decoder gradually up-samples the audio feature, with skip connections between the encoder and the decoder, and generates a face image I_{syn} at the end.

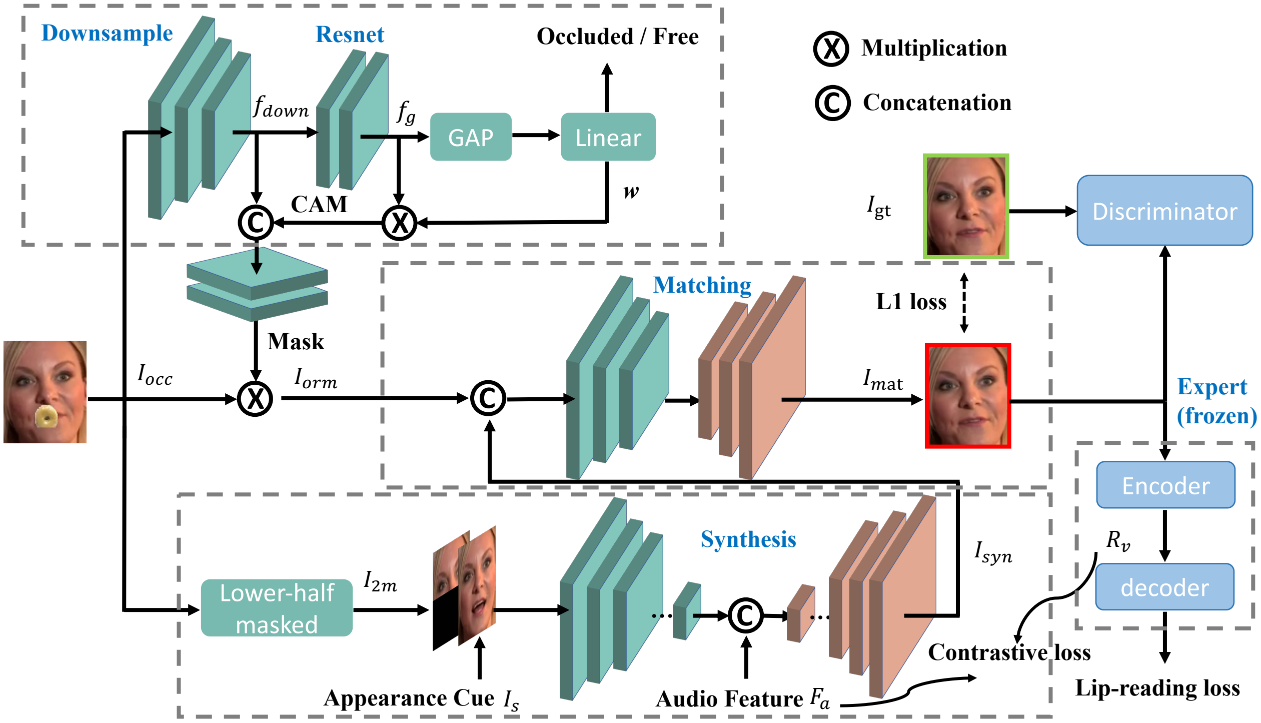


Figure 3: The framework of audio-visual lip restoration.

However, the I_{syn} may be inaccurate due to noisy audio features. Additionally, even though the lower-half face in the I_{syn} could seem reasonable, it might lack sharpness and not necessarily be properly aligned with the corresponding region in the I_{orm}.

To make the I_{syn} more consistent with the I_{orm}, a matching module is employed. We concatenate the I_{syn} with the I_{orm} and pass them through another auto-encoder-decoder with skip connections. Finally, an I_{mat} is obtained. It should be noted that the matching module only conducts shallow downsampling and does not reduce the concatenation to the size of 1 × 1. We use an $L1$ loss to optimize the synthesis module and the matching module, as shown in Eq. 5:

$$\mathcal{L}_{\text{rec}} = |I_{mat} - I_{gt}| + |I_{syn} - I_{gt}|. \tag{5}$$

where I_{gt} is the ground-truth image.

To ensure the visual realism of I_{mat}, we apply a GAN loss (Park et al. 2022; Liang et al. 2022; Goodfellow et al. 2020), which is widely employed in talking face generation (Prajwal et al. 2020; Wang et al. 2023; Park et al. 2022):

$$\mathcal{L}_{\text{gen}} = \mathbb{E}[\log(1 - D(I_{mat}))], \tag{6}$$

$$\mathcal{L}_{\text{disc}} = \mathbb{E}[\log(1 - D(I_{gt}))] + \mathbb{E}[\log(D(I_{mat}))], \tag{7}$$

where D is a discriminator. $\mathcal{L}_{\text{gen}}$ reduces implausible content in images by learning to fool the discriminator.

Lip Reading Loss

A frozen lip reading network is proven to be effective to penalize the synthesized lip that looks reasonable but inconsistent with ground-truth spoken word (Wang et al. 2023). We employ this loss to suppress inaccuracy brought by noisy audio. The frozen lip-reading network, which acts as a lip-reading expert, takes recovered video X'_v as input and generates predicted spoken words Y'. Following (Wang et al. 2023), we choose the lip-reading network of AV-Hubert (Shi et al. 2022) as the lip-reading expert, which is formed by 1) a front-end (3D Convolution Neural Network (CNN) and ResNet-18 (He et al. 2016)) to extract local temporal features 2) a transformer encoder to extract global context feature $R_v \in \mathbb{R}^{T \times f}$ where f is the feature dimension 3) a transformer decoder to predict spoken words Y'. Finally, the cross-entropy loss is applied between Y'_{text} and ground-truth text Y:

$$\mathcal{L}_{\text{lip}} = -Y \log P(Y'|X'_v). \tag{8}$$

Even though the lip-reading expert is frozen, the gradient from the $\mathcal{L}_{\text{lip}}$ still back-propagates to X'_v. As there are some synthesized images present in X'_v, the gradients will also further back-propagate to AVLR and ensure that the synthesized images have correct lip shapes.

Contrastive Loss

Since the audio is noisy, we improve the audio features F_a by contrastive loss, which is proven to be effective to enhance audio-visual synchronization and visual intelligibility on talking face generation based on clean speeches (Wang et al. 2023). Contrastive loss aims to reduce the distance between a frame of F_a and its temporally-aligned frame of visual features R_v, and increase the distance with other frames of F_a. We select infoNCE (Oord, Li, and Vinyals 2018) to achieve this loss, expressed as:

$$\theta(x, x') = \exp(\mathcal{F}_c(x) \cdot \mathcal{F}_c(x')/\tau), \tag{9}$$

$$\mathcal{L}_{\text{cont}} = -\sum_{i \in \Upsilon} \log \frac{\theta(F_a^i, R_v^i)}{\theta(F_a^i, R_v^i) + \sum_{i \neq j}^{j \in \Upsilon} \theta(F_a^i, F_a^j)} \tag{10}$$

where Υ indicates a set covering all frames detected as occluded, $\mathcal{F}_c$ is a linear layer and τ is a pre-defined constant. x and x' denote two feature vectors.

Summarizing all losses, our proposed AVLR is optimized as follows, we omit weights of losses for concision:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{lip}} + \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{disc}}. \tag{11}$$

Experiments

Quantitative Evaluation Metrics

Our main objective is to mitigate the negative impact of lip occlusion on audio-visual speech recognition. To evaluate this, we employ several state-of-the-art (SOTA) audio-visual speech recognition methods. We conduct experiments on each method to compare the WER on original clean videos without lip occlusion, videos with lip occlusion X_v, and videos that have undergone lip occlusion recovery through AVLR X'_v. It's important to note that these methods are solely trained on original clean videos and are not re-trained using the recovered videos.

Methods for Evaluation

To prove our AVLR is compatible with different audio-visual speech recognition methods without re-training, we adopt some methods with different fusion and decoding models: TM-CTC (Afouras et al. 2018), Conformer (Ma, Petridis, and Pantic 2021), P&U net (Wang, Qian, and Li 2022), AV-Hubert (Shi et al. 2022). TM-CTC (Afouras et al. 2018) is a simple but effective method with Connectionist Temporal Classification (CTC) loss which assumes that frames are temporally independent, and is widely employed by other works as a baseline. Conformer (Ma, Petridis, and Pantic 2021) employ both attention loss and CTC loss to get rid of CTC's assumption and repress non-monotonic alignments of attention loss. Inspired by human speech perception studies that lip movements come in advance and cue listeners when and on which frequency of speech should be focused, P&U net (Wang, Qian, and Li 2022) firsts decode spoken words from lip movements and provides it as a prior to speech via a factorized-excitation feed-forward network. AV-Hubert (Shi et al. 2022) is a self-supervised method which employs clustering indices of Mel-Frequency Cepstral Coefficients (MFCC) as pseudo labels to make use of massive audio-visual data without text annotation. In order to better show the influence of video modality, all WER are calculated between ground truth and hypothesis without a language model involved.

Dataset

Our AVLR is trained on LRS2 (Afouras et al. 2018) which encompasses 224 hours of audio-visual data along with text annotation. There are about 144482 utterances that have a duration of less than 6 seconds.

Method	Occlusion	10dB	5dB	0dB	-5dB
TM-CTC	✗	13.8	18.8	29.5	46.8
	✓	16.4	25.5	42.7	66.7
AVLR	✓	13.9	19.4	31.7	52.5
AV-Hubert	✗	3.48	3.80	5.18	10.62
	✓	3.48	3.95	6.58	13.87
AVLR	✓	3.45	3.87	5.69	11.87
P&U net	✗	5.7	7.4	11.4	21.2
	✓	6.7	9.2	16.3	32.6
AVLR	✓	6.1	8.1	13.2	25.0
Conformer	✗	8.2	11.1	19.0	33.4
	✓	8.9	13.1	25.8	48.7
AVLR	✓	8.4	11.6	20.5	37.2

Table 1: Audio-visual speech recognition on the LRS2 dataset. For each method, we present a triple performance assessment (original videos, corrupted videos with occlusion, and recovered videos using our AVLR) across varying Signal-to-Noise Ratio (SNR) (from 10dB to -5dB). Values in the table are WERs.

In the evaluation phase, we assess all chosen audio-visual speech recognition methods based on their WER on LRS2 and LRS3 datasets (Afouras et al. 2018). The evaluation is conducted under three conditions: original clean videos, corrupted videos X_v, and recovered videos X'_v. The LRS3 dataset is a large audio-visual speech recognition dataset derived from TED talks, offering a distinct environment compared to LRS2, which is assembled from BBC programs. Both datasets share a common audio sampling rate of 16 kHz and a video frame rate of 25 frames per second.

Implementation Details

Before applying the AVLR procedure, we follow the preprocessing steps outlined in (Prajwal et al. 2020; Wang et al. 2023). This includes utilizing a face detection module to identify faces, cropping them using bounding boxes, and resizing them to dimensions of 96 × 96. We simulate lip occlusion following (Hong et al. 2023) which employs objects in the Naturalistic Occlusion Generation dataset (Voo, Jiang, and Loy 2022). In detail, we add an object to about 30% frames by aligning its centre with one of the mouth landmarks. The downsample block in the occluded-frame detection and the matching module downsamples reduces the size of input images to 1/4. The discriminator only distinguishes lower-half faces between I_{mat} and I_{gt}. The selection of the lip-reading expert follows (Wang, Qian, and Li 2022) which chooses AV-Hubert base which is pre-trained on the Voxceleb2 (Chung, Nagrani, and Zisserman 2018) and LRS3 (Afouras et al. 2018), and finetuned on the LRS2 (Afouras et al. 2018). Note that AV-Hubert pre-processes images with an affine transformation to eliminate scalar and rotation influences. In this work, we fine-tune the lip-reading expert without employing the affine transformation, allowing it to be adaptable to a broader range of scenarios.

Method	Occlusion	10dB	5dB	0dB	-5dB
AV-Hubert	✗	1.8	2.3	4.5	11.7
	✓	1.8	2.5	5.2	15.2
AVLR	✓	1.9	2.4	4.8	13.3
P&U net	✗	4.0	5.8	11.2	27.9
	✓	5.7	9.1	18.3	42.1
AVLR	✓	4.4	6.3	12.1	31.1
Conformer	✗	6.1	9.7	21.3	42.8
	✓	6.7	10.9	26.7	56.8
AVLR	✓	6.2	9.8	22.5	46.5

Table 2: Audio-visual speech recognition on the LRS3 dataset. For each method, we present a triple performance assessment (original videos, corrupted videos with occlusion, and recovered videos using our AVLR) across varying SNR (from 10dB to -5dB). Values in the table are WERs.

Quantitative Results

In Tab. 1, we assess the efficacy of our AVLR across various SOTA audio-visual speech recognition models. It is evident that the WERs for all methods increase in the presence of lip occlusion. However, upon utilizing our AVLR model to restore occluded facial features, the WERs decrease substantially, approaching levels comparable to those observed without occlusion, particularly at noise levels exceeding 0 dB. When noise is more dominant such as a noise level of -5 dB, although the gap between WERs on original videos and recovered videos slightly widens, the AVLR still manages to alleviate a significant degree of degradation. Specifically, the AVLR reduces degradation by 71.4% (14.2/19.9), 61.5%(2/3.25), 66.7%(7.6/11.4), and 75%(11.5/15.3) for TM-CTC, AV-Hubert, P&U net, and conformer, respectively.

We further confirm the effectiveness of our AVLR by validating it on the LRS3 dataset and show results in Tab. 2. Notably, our AVLR successfully mitigates the degradation resulting from lip occlusion across all methods on the LRS3 dataset, especially for P&U net (Wang, Qian, and Li 2022) and Conformer (Ma, Petridis, and Pantic 2021). For these two methods, AVLR almost relieves degradation when SNR exceeds 0. Even under the challenging noise condition of -5 SNR, AVLR also achieves significant improvement. Thanks to self-supervision on a larger dataset, AV-Hubert is much more robust to lip occlusion. However, AVLR still relieves degradation under noise lower than 0 dB. These results also demonstrate that our AVLR has good generalization across datasets since we do not train our AVLR on LRS3. It's worth noting that the AV-Hubert (large) model utilized in Table 1 is finetuned without the affine transformation preprocessing, thus exhibiting slight performance variations compared to the model reported in (Shi et al. 2022).

Qualitative Results

In this section, we illustrate the effectiveness of our proposed AVLR by evaluating the quality of the recovered images. In Figure 4, we showcase 4 instances of restored lip images. For each sample, I_s is selected from the nearest frame to I_{occ} that is devoid of any occlusion, ensuring that the chosen I_s possesses a similar pose to I_{occ}. The AVLR model is trained to restore I_{mat} to capture the precise viseme from the ground truth I_{gt}. Here, a viseme corresponds to a group of phonemes that share identical lip shapes (Bear and Harvey 2017). We can see the I_{mat} approaches the I_{gt} well.

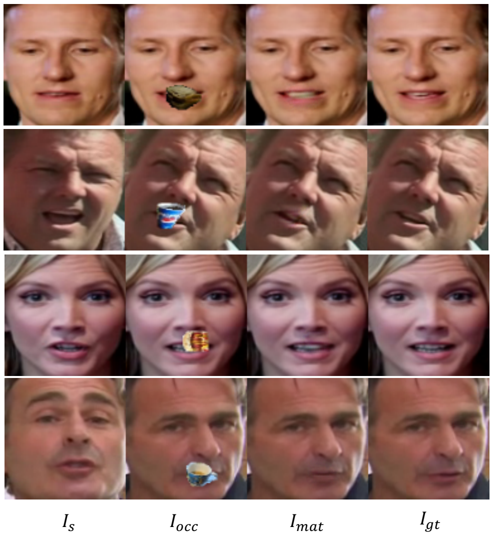


Figure 4: The visualization of restoration of occluded lips. I_s, I_{occ}, I_{mat} and I_{gt} refer to the appearance cue, the occluded input, the output of the AVLR, and the ground-truth image, respectively.

Results of Detection and Masking

We use two simple metrics "recall" and "precision" to measure the accuracy of the detection of occluded frames and the masking of occluded areas. Specifically, the recall and the precision are True Positives / (True Positives + False Negatives) and True Positives / (True Positives + False Positives), respectively.

For the detection, the recall is the number of the correct prediction of occluded frames divided by that of all occluded frames, which is 99.74%. The precision in the detection is 99.91%. For the masking, the recall is the number of the correct prediction of occluded pixels divided by that of all occluded pixels, which is 96.04%. The precision is 98.40%.

Comparison of Synthesis-Matching Strategy and Talking Face Generation

Intuitively, methods of talking face generation also have the potential to complete occluded-masked images. Here, we compare the synthesis-matching strategy with two talking face generation methods, i.e. Wav2lip (Prajwal et al. 2020) which works on enhancing synchronization and Talklip (Wang et al. 2023) which is proven to generate videos with good intelligibility. The input to Wav2lip and Talklip is a combination of an occlusion-masked image, an image of the appearance cue and an audio clip.

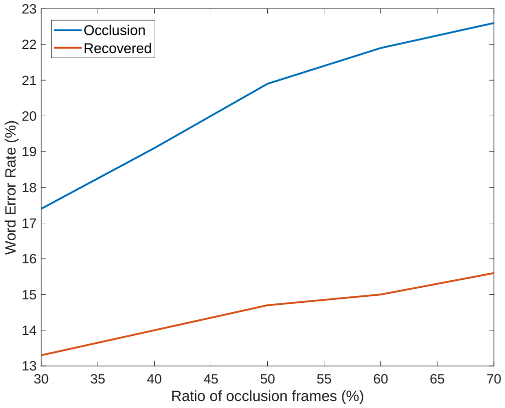


Figure 5: The relationship between performance and the ratio of occluded frames using the LRS2 dataset. Results are computed utilizing the P&U net and a noise level of 0 dB.

Method	SSIM	PSNR	LSE-C	LSE-D	WER%
Wav2lip	93.30	**34.14**	7.58	6.70	35.26
TalkLip	92.50	32.99	7.49	6.69	32.39
Ours	**93.66**	33.26	**7.77**	**6.48**	**29.87**
- $\mathcal{L}_{\text{cont}}$	93.83	33.24	7.67	6.53	31.70
- $\mathcal{L}_{\text{cont,lip}}$	93.22	32.42	7.57	6.63	33.77

Table 3: Comparison with talking face generation methods and ablation study of contrastive loss and lip-reading loss. "- $\mathcal{L}_{\text{cont}}$" means contrastive loss is not utilized. WER is the lip-reading performance based on AV-Hubert. **Ours** means the synthesis-matching strategy.

We adopt PSNR (Vougioukas, Petridis, and Pantic 2020; Park et al. 2022; Jin, Yang, and Tan 2022) and SSIM (Wang et al. 2004; Liang et al. 2022; Jin et al. 2022) metrics to measure visual quality and LSE-D, LSE-C (Prajwal et al. 2020) to measure audio-visual synchronization, and WER to measure lip reading performance (Wang et al. 2023). The superiority of our proposed synthesis-matching approach over two SOTA talking face generation methods is readily apparent in terms of lip-speech synchronization and visual reading intelligibility. Although the improvement in visual quality is not distinct from quantitative results. We find that the synthesized images by Wav2Lip and TalkLip are not natural, remaining a subtle occlusion mask or artifacts. We will provide a qualitative comparison in the supplementary.

Ablation Study of Contrastive Loss and Lip-Reading Loss

Despite the effectiveness demonstrated in enhancing reading intelligibility and synchronization in (Wang et al. 2023), we undertake an ablation study of these two losses. This is because the context of the lip restoration task is not the same as that of talking face generation. The results of this study are presented in Tab. 3. It is observed that the contrastive loss and lip-reading loss are capable of enhancing synchronization and reading intelligibility, which is consistent with that in (Wang et al. 2023).

Type	SI-SDRi↑	SDRi↑	PESQi↑	STOIi↑
Original	11.4	11.8	1.02	0.20
Corrupted	9.1	9.8	0.83	0.15
Recovered	11.3	11.7	1.00	0.20

Table 4: Performance of our AVLR on the audio-visual speech extraction task.

AVLR on Other Audio-Visual Tasks

To validate that the recovered videos are task-independent, we make a comparison of original videos, corrupted videos, and recovered videos on the audio-visual speech extraction task using the USEV network(Pan, Ge, and Li 2022). The SI-SDRi and SDRi represent the improvements in the signal quality of the extracted speech qualities. The PESQi and STOIi represent the improvements in the perceptual quality and intelligibility of the extracted speech. The USEV network is pre-trained and fixed, without re-training on occlusion data. We listed the results in Tab. 4, and it is seen that the performance of all metrics drops significantly when the video is corrupted compared to the original video. When our proposed AVLR is used to restore the occluded lips, the performance of USEV is on par with the original video, showing the effectiveness of our AVLR.

Ablation of the Ratio of Occluded Frames

Since the ratio of occluded frames can vary in real-world scenarios, we illustrate the correlation between audio-visual speech recognition performance and this ratio, as shown in Fig. 5. It is evident that degradation intensifies as the ratio increases, regardless of whether AVLR is applied. Nevertheless, we can discern that the discrepancy between the performance on corrupted videos and recovered videos widens, signifying that the AVLR mitigates the detrimental impact of lip occlusion on audio-visual speech recognition.

Conclusion

This paper addresses a relatively less explored issue: the detrimental impact of lip occlusion on audio-visual speech recognition, by proposing a framework (AVLR) to restore occluded lips. To achieve the final target, the AVLR framework encompasses three main tasks: detecting occluded frames, masking occluded areas, and reconstruction of masked regions. To ensure audio features align well with varying degrees of occlusion in the reconstruction of masked regions, we introduce a novel synthesis-matching inpainting strategy. Empirical findings validate its superiority to one-step talking face generation methods. Besides the reconstruction, AVLR can accurately detect occluded frames in a video and estimate the mask of occlusion. Overall, our AVLR can markedly mitigate the degradation in audio-visual speech recognition attributed to lip occlusion, both within individual datasets and across them.

Acknowledgements

This work is supported by Human Robot Collaborative AI under its AME Programmatic Funding Scheme (Project No. A18A2b0046), the National Natural Science Foundation of China (Grant No. 62271432, 62106038), Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002), and the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant No. T00120220002.

References

Abdelaziz, A. H.; Zeiler, S.; and Kolossa, D. 2015. Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(5): 863–876.

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Afouras, T.; Chung, J. S.; and Zisserman, A. 2019. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*.

Bear, H. L.; and Harvey, R. 2017. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95: 40–67.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7832–7841.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, J.; Kim, M.; Choi, J.; and Ro, Y. M. 2023. Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18783–18794.

Hong, J.; Kim, M.; Yoo, D.; and Ro, Y. M. 2022. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. *arXiv preprint arXiv:2207.06020*.

Jin, Y.; Sharma, A.; and Tan, R. T. 2021. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5027–5036.

Jin, Y.; Yan, W.; Yang, W.; and Tan, R. T. 2022. Structure Representation Network and Uncertainty Feedback Learning for Dense Non-Uniform Fog Removal. In *Proceedings of the Asian Conference on Computer Vision*, 2041–2058.

Jin, Y.; Yang, W.; and Tan, R. T. 2022. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European Conference on Computer Vision*, 404–421. Springer.

Liang, B.; Pan, Y.; Guo, Z.; Zhou, H.; Hong, Z.; Han, X.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3387–3396.

Luettin, J.; Potamianos, G.; and Neti, C. 2001. Asynchronous stream modeling for large vocabulary audio-visual speech recognition. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, 169–172. IEEE.

Ma, P.; Petridis, S.; and Pantic, M. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7613–7617. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pan, Z.; Ge, M.; and Li, H. 2022. USEV: Universal speaker extraction with visual cue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 3032–3045.

Pan, Z.; Tao, R.; Xu, C.; and Li, H. 2021. MuSE: Multi-Modal Target Speaker Extraction with Visual Cues. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 6678–6682.

Paraskevopoulos, G.; Parthasarathy, S.; Khare, A.; and Sundaram, S. 2020. Multiresolution and multimodal speech recognition with transformers. *arXiv preprint arXiv:2004.14840*.

Park, S. J.; Kim, M.; Hong, J.; Choi, J.; and R, Y. M. 2022. SyncTalkFace: Talking Face Generation with Precise Lip-syncing via Audio-Lip Memory. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 234–778.

Petridis, S.; Stafylakis, T.; Ma, P.; Tzimiropoulos, G.; and Pantic, M. 2018. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 513–520. IEEE.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492.

Shi, B.; Hsu, W.-N.; Lakhotia, K.; and Mohamed, A. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations*.

Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2022. Everybody's talkin': Let me talk as you want. *IEEE Trans. on Information Forensics and Security*, 17: 585–598.

Sterpu, G.; Saam, C.; and Harte, N. 2020. How to teach DNNs to pay attention to the visual modality in speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1052–1064.

Stewart, D.; Seymour, R.; Pass, A.; and Ming, J. 2013. Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE transactions on cybernetics*, 44(2): 175–184.

Voo, K. T.; Jiang, L.; and Loy, C. C. 2022. Delving into high-quality synthetic face occlusion segmentation datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4711–4720.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5): 1398–1413.

Wang, J.; Qian, X.; and Li, H. 2022. Predict-and-Update Network: Audio-Visual Speech Recognition Inspired by Human Speech Perception. *arXiv preprint arXiv:2209.01768*.

Wang, J.; Qian, X.; Zhang, M.; Tan, R. T.; and Li, H. 2023. Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14653–14662.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Yu, J.; Zhang, S.-X.; Wu, B.; Liu, S.; Hu, S.; Geng, M.; Liu, X.; Meng, H.; and Yu, D. 2021. Audio-visual multichannel integration and recognition of overlapped speech. *IEEE/ACM TASLP*, 29: 2067–2082.

Yu, J.; Zhang, S.-X.; Wu, J.; Ghorbani, S.; Wu, B.; Kang, S.; Liu, S.; Liu, X.; Meng, H.; and Yu, D. 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)*, 6984–6988. IEEE.

Zhang, Q.; Nicolson, A.; Wang, M.; Paliwal, K. K.; and Wang, C. 2020. DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1404–1415.

Zhang, Q.; Qian, X.; Ni, Z.; Nicolson, A.; Ambikairajah, E.; and Li, H. 2022. A Time-Frequency Attention Module for Neural Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 462–475.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, 9299–9306.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15.