Noise- and Occlusion-Robust Mask-Guided Adaptive Fusion for Audio-Visual Person Verification

Juntao Wang[1], Qiuming Zhao[2], Xiaolong Wu[3], Mingxing Xu[2], Askar Hamdulla[1*], Thomas Fang Zheng[2*]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Center for Speech and Language Technology, Tsinghua University, Beijing, China
[3]School of Basic Sciences for Aviation, Naval Aviation University, Yantai, China

{wangjuntao,wuxl}@stu.xju.edu.cn, zqm23@mails.tsinghua.edu.cn,
askar@xju.edu.cn, {xumx,fzheng}@tsinghua.edu.cn

***Abstract*—Audio-visual multimodal person verification has attracted increasing attention for improving robustness in challenging conditions by leveraging the complementarity between speech and face cues. Despite recent progress, existing feature-level fusion methods—e.g., soft-attention and gating—typically infer modality reliability implicitly from identity embeddings, without dedicated degradation-aware cues for acoustic noise or facial occlusion. Consequently, the learned modality weights can be unreliable when degradations are severe or asymmetric, yielding suboptimal fusion representations. In this work, we propose a mask-guided adaptive fusion framework (MGAF) that explicitly estimates modality reliability from degradation-aware masks and uses this signal to guide fusion. Specifically, MGAF predicts a time–frequency mask to characterize noise-corrupted speech and learns a feature-level visual mask on convolutional representations to enhance reliable responses; the masks are aggregated into explicit reliability cues, based on which MGAF performs mask-driven re-weighting of unimodal embeddings and constructs the fused representation via weighted concatenation. This design allows the model to emphasize the more reliable modality under strong corruption while retaining complementary information when both modalities are informative. We train MGAF on VoxCeleb2 and evaluate it on VoxCeleb1 under systematically simulated noise and occlusion settings, including asymmetric degradations. On the standard VoxCeleb1 test sets, MGAF achieves EERs of 0.207%, 0.212%, and 0.433% on VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H, respectively, and achieves consistent gains over representative fusion baselines under degraded scenarios—especially under asymmetric degradations—without over-suppressing either modality when both are reliable.**

***Index Terms*—person verification, multi-modal fusion, audio-visual, occluded face, noise-robust**

I. Introduction

Biometric authentication has been widely used in forensic investigation, commercial identity verification, and access control. Among various modalities, speech and face are particularly attractive because they are easy to acquire and contain rich identity information. Recent advances in network architectures and discriminative loss functions have greatly improved speaker verification (SV) and face verification (FV) on standard benchmarks [1]–[4]. However, real-world deployments often involve asymmetric modality degradation: SV is vulnerable to environmental noise and reverberation [5], while FV is easily affected by occlusions such as masks, sunglasses, and hand blockage [6], which weakens the reliability of unimodal systems.

To improve robustness, prior unimodal studies have explored degradation-specific strategies. In SV, enhancement-based methods can reduce noise interference, but they may also suppress speaker-discriminative cues, motivating approaches that jointly optimize denoising and speaker representation learning [7]–[9]. In FV, occlusion robustness is commonly addressed by recovering occluded content or suppressing corrupted responses in the feature space [10], [11]. Although effective within each modality, these methods cannot exploit the complementarity between speech and face under severe degradation.

This limitation has driven increasing interest in audio-visual multimodal person verification [12], [13]. Early studies mainly adopted score-level fusion [14]–[16], which is simple but lacks flexibility under asymmetric degradation. More recent methods have shifted to feature-level fusion [17]–[19], using soft attention [20], [21], gating [22], and cross-attention [23]. However, most of these methods estimate modality importance implicitly from identity embeddings rather than from explicit degradation cues. As a result, the fusion weights may become unreliable when speech and face are degraded to different degrees.

To address this problem, we propose a mask-guided adaptive fusion framework (MGAF). MGAF predicts a time–frequency mask for speech and a feature-level mask for visual representations, and uses them as explicit reliability cues for fusion. Guided by these cues, the model adaptively re-weights unimodal embeddings and performs weighted concatenation, allowing it to emphasize the more reliable modality while preserving complementary information from the other. Trained on VoxCeleb2 [24] and evaluated on VoxCeleb1 [25] under simulated noise and occlusion conditions, MGAF achieves consistent improvements in degraded scenarios while remain-

* Corresponding authors.

ing competitive on clean data.

II. RELATED WORK

Our work is related to three research directions: occlusion-robust face recognition, noise-robust speaker verification, and audio-visual multimodal person verification.

A. Occlusion-Robust Face Recognition

Facial occlusion remains a major challenge for face recognition. Existing methods for occluded face recognition mainly follow two directions: reconstructing occluded regions to recover missing information, or suppressing occlusion-corrupted features to reduce interference.

Reconstruction-based methods use generative models to inpaint occluded regions and recover complete face images for recognition [26], [27]. However, under severe occlusion, reconstruction may introduce artifacts or identity-inconsistent textures, which can reduce recognition robustness.

Song et al. [28] propose a mask-learning strategy based on annotated occlusion locations, where occlusion detection and feature processing are performed in two stages. To avoid such stage-wise design, Qiu et al. [10] develop FROM, which dynamically predicts feature masks in an end-to-end manner. Huang et al. [11] further introduce mask-supervised occlusion prediction and joint spatial–channel suppression, and combine original and purified feature maps to obtain the final embedding.

Inspired by these studies on occlusion-aware mask prediction and feature modulation, we enhance a ResNet backbone with an explicit occlusion-aware suppression module. The predicted occlusion mask not only improves robustness under occlusion, but also serves as an intermediate cue for subsequent fusion.

B. Noise-Robust Speaker Verification

Speaker verification (SV) performance degrades substantially in noisy environments. A common solution is to apply speech enhancement before SV [29]. However, enhancement models often prioritize perceptual quality and may suppress speaker-discriminative information. To address this issue, VoiceID loss [7] introduces speaker-related supervision into enhancement training, encouraging the model to suppress components that are harmful to speaker discrimination.

To further reduce the mismatch between enhancement and recognition, joint optimization of speech enhancement and speaker embedding networks has been widely studied [9]. Wu et al. [8] further propose asynchronous sub-region optimization to alleviate gradient conflicts between enhancement and speaker classification objectives during joint training.

Similar to these joint training strategies, we jointly optimize a speech enhancement front-end and a speaker embedding network. This design improves robustness under additive noise and produces a time–frequency mask that can be used as a reliability cue during inference.

C. Audio-Visual Person Verification

With the development of multimodal fusion [30], audio-visual person verification has achieved substantial progress. Early studies mainly adopted score-level fusion [14]–[16], which can improve overall performance but often cannot adapt modality contributions under asymmetric degradation. More recent work has shifted toward feature-level fusion by combining voice and face embeddings, denoted as $\mathbf{e}_v$ and $\mathbf{e}_f$.

Two representative weighting-based strategies are simple soft attention (SSA) and gated fusion. In both cases, the unimodal embeddings are first projected into a shared space:

$$\tilde{\mathbf{e}}_v = f_{\text{trans},v}(\mathbf{e}_v), \quad \tilde{\mathbf{e}}_f = f_{\text{trans},f}(\mathbf{e}_f). \tag{1}$$

SSA predicts two scalar attention weights and performs weighted summation:

$$\hat{\mathbf{a}} = f_{\text{att}}([\mathbf{e}_v, \mathbf{e}_f]), \quad \alpha_i = \frac{\exp(\hat{a}_i)}{\sum_{k\in\{v,f\}} \exp(\hat{a}_k)}, \quad \mathbf{e}_p = \sum_{i\in\{v,f\}} \alpha_i \tilde{\mathbf{e}}_i, \tag{2}$$

where α_i denotes the modality weight [20], [21]. In contrast, gated fusion extends scalar weighting to element-wise modulation:

$$\mathbf{z} = \sigma\big(f_{\text{att}}([\mathbf{e}_v, \mathbf{e}_f])\big), \quad \mathbf{e}_p = \mathbf{z}\odot\tanh(\tilde{\mathbf{e}}_v) + (1-\mathbf{z})\odot\tanh(\tilde{\mathbf{e}}_f), \tag{3}$$

which allows dimension-wise control of the two modalities [13].

Beyond weighting-based fusion, some studies explore earlier feature interaction [12], cross-attention-based sequence modeling [23], or improved objectives and training strategies [31]–[34]. However, existing methods usually infer modality importance implicitly from identity embeddings, without explicit degradation-aware evidence tied to acoustic noise or facial occlusion. Under severe or asymmetric degradation, such implicit weighting may become unreliable. To address this issue, we explicitly predict unimodal degradation masks and derive modality reliability from them for adaptive fusion.

III. METHODOLOGY

We propose the Mask-Guided Adaptive Fusion (MGAF) framework for robust person verification under modality-specific corruptions. As shown in Fig. 1, MGAF consists of a speech branch, a face branch, and a mask-guided adaptive fusion module.

The speech branch combines a learnable time-frequency masking front-end with an ECAPA-TDNN [4] encoder. The face branch is built on a ResNet18 backbone [35] and includes a mask prediction module and an occlusion mode predictor for occlusion-aware modeling.

The fusion module uses the predicted masks to estimate modality reliability and adaptively balance the two modalities under asymmetric degradations. Details are given in the following subsections.

A. Voice Embedding Extraction

We formulate speech enhancement as a time–frequency masking problem in the STFT magnitude domain. While the

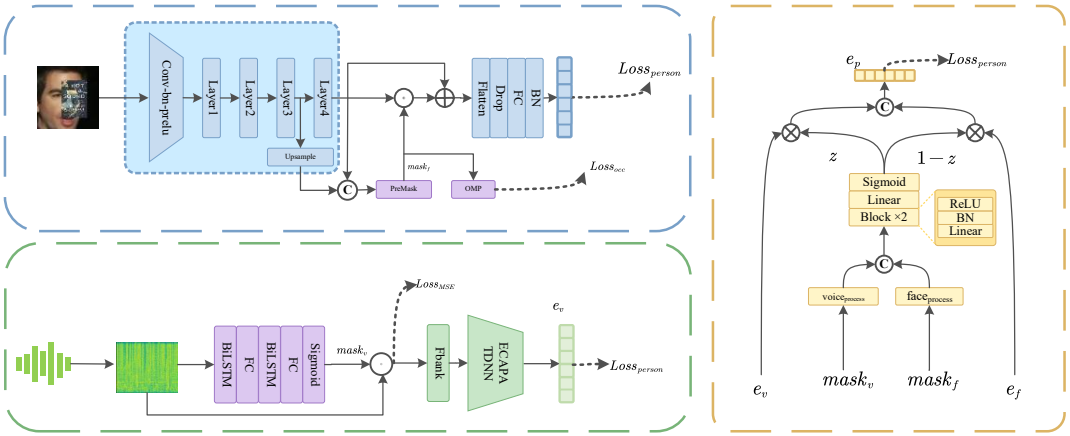

Fig. 1. Overall architecture of the proposed MGAF framework. The face branch predicts an occlusion mask $mask_f$ to modulate facial feature maps and extract a face embedding $\mathbf{e}_f$, while the speech branch estimates a time–frequency mask $mask_v$ to enhance the speech spectrum and extract a speaker embedding $\mathbf{e}_v$. The fusion module further encodes $(mask_v, mask_f)$ as reliability cues to infer a weight z, and adaptively fuses $\mathbf{e}_v$ and $\mathbf{e}_f$ into a person embedding $\mathbf{e}_p$.

additive model strictly holds in the time/complex-STFT domain, we adopt the common magnitude-domain approximation:

$$\mathbf{Y} \approx \mathbf{X} + \mathbf{N}, \tag{4}$$

where $\mathbf{X}$, $\mathbf{N}$, and $\mathbf{Y}$ denote the STFT magnitude spectrograms of the clean speech, noise, and their mixture, respectively. Given the mixture magnitude $\mathbf{Y}$, we feed its log-magnitude spectrum into a mask estimation network to predict a time–frequency mask:

$$\mathbf{mask}_v = f_{\text{mask}}(\log \mathbf{Y}). \tag{5}$$

The enhanced spectrum is obtained by element-wise masking:

$$\hat{\mathbf{X}} = \mathbf{Y} \odot \mathbf{mask}_v. \tag{6}$$

To supervise the enhancement front-end, we adopt a mean squared error objective in a power-law compressed spectral domain (pow-loss) [36], implemented as MSE on the square-root spectra:

$$\mathcal{L}_{\text{spec}} = \left\| \hat{\mathbf{X}}^{\frac{1}{2}} - \mathbf{X}^{\frac{1}{2}} \right\|_2^2. \tag{7}$$

In parallel, to learn discriminative speaker representations, we convert the enhanced spectrum into a power spectrum and extract Fbank features, which are fed into ECAPA-TDNN to produce a 192-dimensional speaker embedding:

$$\mathbf{F} = \text{Fbank}(\hat{\mathbf{X}}^2), \qquad \mathbf{e}_v = f_{\text{spk}}(\mathbf{F}), \qquad \mathbf{e}_v \in \mathbb{R}^{192}. \tag{8}$$

The speaker classification objective is implemented with Additive Angular Margin Softmax (AAM-Softmax) [3]:

$$\mathcal{L}_{\text{aam}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s \cdot \cos(\theta_{y_i} + m)}}{e^{s \cdot \cos(\theta_{y_i} + m)} + \sum_{j=1, j \neq y_i}^{C} e^{s \cdot \cos(\theta_j)}} \tag{9}$$

where N and C denote the batch size and the number of classes, respectively; s is the scale factor and m is the additive angular margin, and θ_j is the angle between the embedding and the j-th class weight.

Finally, the speech branch is optimized end-to-end with a weighted sum of the speaker classification loss and the spectral enhancement loss:

$$\mathcal{L}_{\text{voice}} = \mathcal{L}_{\text{aam}} + \lambda \mathcal{L}_{\text{spec}}, \tag{10}$$

Fig. 2. When $K = 2$, the pattern set contains 10 classes: four 1×1 blocks, two 1×2 rectangles, two 2×1 rectangles, one 2×2 full-coverage pattern, and one non-occluded pattern.

where $\lambda = 5$. This joint training encourages the enhancement module to suppress noise while preserving identity-relevant information, yielding more robust speech embeddings for subsequent fusion.

B. Face Embedding Extraction

We adopt a ResNet18 backbone (removing global average pooling and the original FC head) to extract hierarchical facial features. Let the last-stage feature map be $\mathbf{F}_4 \in \mathbb{R}^{512 \times 7 \times 7}$ and an intermediate feature map be $\mathbf{F}_3 \in \mathbb{R}^{256 \times 14 \times 14}$. To incorporate multi-scale cues, we upsample $\mathbf{F}_3$ to match the spatial resolution of $\mathbf{F}_4$ and concatenate them along the channel dimension to form a multi-scale representation:

$$\mathbf{F}_{\text{ms}} = \text{Cat}(\mathbf{F}_4,\ \text{Up}(\mathbf{F}_3)) \in \mathbb{R}^{768 \times 7 \times 7}. \tag{11}$$

Based on $\mathbf{F}_{\text{ms}}$, the mask prediction module generates an occlusion mask with the same shape as $\mathbf{F}_4$:

$$\mathbf{mask}_f = f_{\text{PreMask}}(\mathbf{F}_{\text{ms}}), \quad \mathbf{mask}_f \in \mathbb{R}^{512 \times 7 \times 7}. \tag{12}$$

$f_{\text{PreMask}}(\cdot)$ is a lightweight two-layer 3×3 convolutional head with PReLU and BatchNorm, followed by a Sigmoid to produce a soft occlusion mask in $[0, 1]$.

The predicted mask modulates the backbone feature in a residual manner to obtain an occlusion-aware feature map:

$$\hat{\mathbf{F}} = \mathbf{F}_4 + \mathbf{F}_4 \odot \mathbf{mask}_f. \tag{13}$$

In addition, we predict occlusion-pattern logits from the predicted mask for auxiliary supervision:

$$\mathbf{k} = f_{\text{OMP}}(\mathbf{mask}_f), \quad \mathbf{k} \in \mathbb{R}^{N_{\text{patterns}}}. \tag{14}$$

$f_{\text{OMP}}(\cdot)$ is a lightweight head with Conv-BN-PReLU, global average pooling, and a linear projection to output occlusion-mode logits.

Finally, we flatten $\hat{\mathbf{F}}$ and apply a projection head to obtain a 512-dimensional face embedding:

$$\mathbf{e}_f = f_{\text{proj}}(\text{Flatten}(\hat{\mathbf{F}})) \in \mathbb{R}^{512}. \tag{15}$$

To obtain a more accurate visual mask $\mathbf{mask}_f$, we introduce occlusion-pattern prediction as an auxiliary supervision signal. We approximate each occlusion as an axis-aligned rectangle and partition the face region into a $K \times K$ grid. Based on all possible combinations of contiguous grid intervals along the horizontal and vertical directions, we construct a set of rectangular occlusion patterns $\{\mathcal{P}_i\}_{i=1}^{N_{\text{patterns}}}$, where the total number of patterns is

$$N_{\text{patterns}} = \left(\frac{K(K+1)}{2}\right)^2 + 1, \tag{16}$$

and the extra "+1" corresponds to the non-occluded case [10]. As illustrated in Fig. 2, when $K = 2$, this yields 10 patterns. In this paper, we set $K = 5$.

Given the ground-truth occlusion region O (obtained from synthesis), we assign the occlusion label by maximum IoU matching:

$$l_{\text{occ}} = \arg\max_i \frac{|\mathcal{P}_i \cap O|}{|\mathcal{P}_i \cup O|}, \tag{17}$$

where $\mathcal{P}_i$ denotes the region covered by the i-th occlusion pattern. Let $\mathbf{k} = [k_1, \ldots, k_{N_{\text{patterns}}}]$ denote the occlusion-mode logits predicted by OMP. The occlusion classification loss is then defined as the cross-entropy:

$$\mathcal{L}_{\text{occ}} = -\log \frac{\exp(k_{l_{\text{occ}}})}{\sum_{j=1}^{N_{\text{patterns}}} \exp(k_j)}. \tag{18}$$

The face branch is trained with a multi-task objective that combines occlusion classification and identity supervision:

$$\mathcal{L}_{\text{face}} = \mathcal{L}_{\text{occ}} + \mathcal{L}_{\text{aam}}, \tag{19}$$

which encourages occlusion-aware mask learning while preserving identity-discriminative face embeddings.

C. Fusion Module

As shown in Fig. 1, we design a *mask-weighted fusion* module to adaptively balance audio and visual embeddings according to the predicted reliability of each modality. Given the voice embedding $\mathbf{e}_v \in \mathbb{R}^{192}$, the face embedding $\mathbf{e}_f \in \mathbb{R}^{512}$, and their corresponding masks $\mathbf{mask}_v \in \mathbb{R}^{F\times T}$ and $\mathbf{mask}_f \in \mathbb{R}^{C\times H\times W}$, the fusion module outputs a fused embedding $\mathbf{e}_p \in \mathbb{R}^{704}$.

a) Mask feature extraction: We convert each predicted mask into a fixed-dimensional representation. For the voice mask, a voice mask processor (process_v) maps $\mathbf{mask}_v$ to a 512-D vector:

$$\mathbf{f}_v = \text{process}_v(\mathbf{mask}_v) \in \mathbb{R}^{512}. \tag{20}$$

Concretely, process_v first compresses the frequency axis using 1×1 Conv-BN-ReLU, then applies a lightweight temporal encoder with two Conv-BN-ReLU blocks, and finally performs global statistical pooling followed by a linear projection to 512 dimensions.

For the visual mask, a face mask processor (process_f) maps $\mathbf{mask}_f$ to a 512-D vector:

$$\mathbf{f}_f = \text{process}_f(\mathbf{mask}_f) \in \mathbb{R}^{512}. \tag{21}$$

Specifically, process_f uses global average pooling and a linear layer with BN-ReLU-Dropout to obtain the mask representation.

b) Mask-driven Gated weighting: We concatenate the mask representations from the audio and visual branches and predict a scalar gate $z \in (0, 1)$ using a lightweight MLP with a sigmoid activation:

$$z = g\big([\mathbf{f}_v; \mathbf{f}_f]\big), \tag{22}$$

where $[\cdot;\cdot]$ denotes concatenation. The gate z modulates the contribution of the voice embedding, while $1 - z$ modulates the face embedding.

c) Weighted fusion: Given unimodal embeddings $\mathbf{e}_v$ and $\mathbf{e}_f$, we perform mask-guided re-weighting and concatenate them to form the fused representation:

$$\mathbf{e}_p = \text{LN}\Big([\, z\,\mathbf{e}_v;\ (1-z)\,\mathbf{e}_f\,]\Big) \in \mathbb{R}^{704}. \tag{23}$$

where LN denotes layer normalization.

The fusion module is trained with the same identity supervision as the face branch, i.e., $\mathcal{L}_{\text{person}} = \mathcal{L}_{\text{aam}}$, implemented using the Additive Angular Margin Softmax (AAM-Softmax) loss.

IV. EXPERIMENTS

A. Datasets

VoxCeleb1 [25] and VoxCeleb2 [24] are used in our experiments. VoxCeleb is an audio-visual dataset containing over 2,000 hours of speech clips collected from YouTube interview videos. We use the DEV split of VoxCeleb2 for training, which contains 5,994 speakers and 1,092,009 utterances, and evaluate on VoxCeleb1 using the three official trial lists: Vox1-O, Vox1-E, and Vox1-H.

To assess robustness under different degradation levels, we further construct degraded test conditions based on Vox1-O by perturbing both modalities. For speech, we add MUSAN noise [37] at five SNR levels (0, 5, 10, 15, and 20 dB). For faces, as shown in Fig. 3, we apply synthetic occluders from [10] with three scaling factors (×0.75, ×1.0, and ×1.25).

TABLE I
SPEAKER VERIFICATION EER (%) UNDER DIFFERENT SNR CONDITIONS.

SNR (dB)	ECAPA-TDNN	ECAPA-TDNN-Mask
Clean	1.064	**1.016**
0	3.739	**3.064**
5	2.186	**1.995**
10	1.659	**1.580**
15	1.382	**1.362**
20	1.202	**1.154**
Average	1.872	**1.695**

TABLE II
FACE VERIFICATION EER (%) UNDER DIFFERENT OBSTACLE SIZES.

Obstacle size	ResNet18	ResNet18-Mask
Clean	1.597	**1.372**
×0.75	3.016	**2.516**
×1.0	4.382	**3.739**
×1.25	6.797	**6.319**
Average	3.948	**3.486**

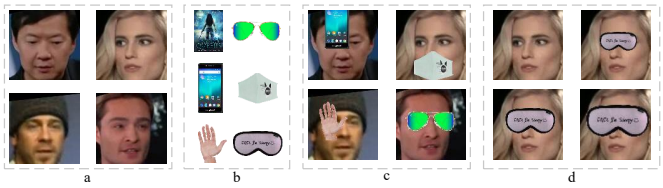

Fig. 3. Facial occlusion synthesis components and results. (a) Original face images. (b) Occlusion objects including eye masks, sunglasses, hands, phones, and face masks. (c) Synthesized occluded faces with diverse occlusion types. (d) Different occlusion severity levels(×0.75, ×1.0, ×1.25).

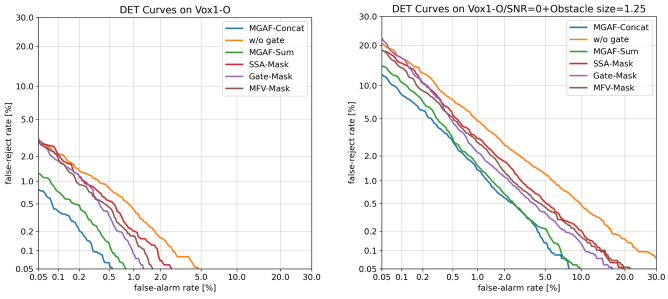

Fig. 4. DET curves for six audio-visual fusion methods.

B. Implementation Details

Speech features are extracted from STFT-based power spectra (512-point FFT, 25 ms window, 10 ms hop), followed by mask-guided corruption and conversion to 80-bin log-Mel filterbank (Fbank) features. Face frames are sampled at 1 fps, resized to 224×224, and aligned to 112×112 using InsightFace.

Training is conducted in three stages. First, we train the voice branch, which consists of a VoxCeleb2-pretrained ECAPA-TDNN backbone and a mask prediction module, using $\mathcal{L}_{\text{voice}}$. The learning rates are 1×10^{-3} for the speech mask estimation network and 1×10^{-4} for ECAPA-TDNN, with a decay factor of 0.97.

Second, we train the face branch with $\mathcal{L}_{\text{face}}$, initializing the ResNet backbone with VoxCeleb2-pretrained weights. The learning rates are 1×10^{-4} for the backbone and 1×10^{-3} for the remaining modules, with a decay factor of 0.90.

Finally, after freezing both unimodal branches, we optimize the fusion module using $\mathcal{L}_{\text{person}}$, with a learning rate of 5×10^{-4} and a decay factor of 0.97.

For $\mathcal{L}_{\text{aam}}$, we set $(s, m) = (30, 0.2)$ for the voice branch, $(64, 0.3)$ for the face branch, and $(20, 0.2)$ for the fusion module.

For evaluation, we use cosine scoring with adaptive s-norm [38], and report Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $P_{\text{target}} = 0.01$, $C_{\text{FA}} = 1$, and $C_{\text{Miss}} = 1$ [39].

V. RESULTS AND ANALYSIS

A. Unimodal Results

To evaluate unimodal robustness on Vox1-O, we construct degraded test conditions by adding noise to speech and synthetic occlusions to faces. For the speech branch, Table I compares ECAPA-TDNN with its mask-enhanced variant under different SNRs. The mask module consistently improves verification performance, with larger gains under lower SNR conditions. For the face branch, Table II compares ResNet18 and its mask-augmented version under different occlusion scales. Consistent improvements across all settings show that the proposed masking mechanism effectively suppresses degradation-induced interference and enhances identity-discriminative representations.

B. Audio-Visual Results

We compare our method with representative audio-visual fusion approaches, as summarized in Table III. For soft-attention- and gating-based baselines, we re-implement the fusion modules under a unified evaluation protocol: the unimodal branches are kept the same as ours, and only the fusion module is replaced for fair comparison. The suffix "-Mask" denotes variants built on our mask-enhanced unimodal branches. For methods such as P.Sun [31] and Selvakumar [33], which mainly improve training strategies while still adopting attention- or gating-based fusion, we directly report the results from their papers. We also include MFV [19], [40], a representative recent fusion model, and a self-supervised approach [34] for comparison. As shown in Table III, our method achieves competitive or superior performance on VoxCeleb1-O/E/H, demonstrating the effectiveness of the proposed fusion strategy.

C. Robustness under Cross-Modal Degradations

We construct four cross-modal degradation scenarios on Vox1-O by combining speech corruption (SNR) and face occlusion (obstacle size), and compare representative audio-visual fusion methods under a unified protocol (Table IV). MGAF achieves the best performance across all settings, including the clean condition, and maintains clear advantages under severe degradations. In addition, the mask-enhanced variants

TABLE III
PERFORMANCE COMPARISON ON VOXCELEB1 TEST SETS. V AND F DENOTE THE VOICE AND FACE MODALITIES.
NOTE: METHODS MARKED WITH * ARE REPORTED DIRECTLY FROM THE CORRESPONDING PAPERS.

Method	Modality	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER↓ (%)	minDCF↓	EER↓ (%)	minDCF↓	EER↓ (%)	minDCF↓
ECAPA-TDNN-Mask	V	1.016	0.080	1.285	0.087	2.424	0.148
ResNet18-Mask	F	1.372	0.101	1.318	0.089	1.922	0.120
Hörmann* [21]	VF	0.660	0.071	0.540	0.064	0.910	0.110
SSA-Mask [20]	VF	0.537	0.038	0.463	0.029	0.863	0.053
Gate-Mask [13]	VF	0.420	0.035	0.350	0.021	0.679	0.039
MFV-Mask [19]	VF	0.473	0.037	0.366	0.024	0.699	0.040
R. Tao* [34]	VF	0.330	–	0.430	–	0.820	–
Selvakumar* [33]	VF	0.244	–	0.252	–	0.441	–
P. Sun* [31]	VF	**0.180**	0.017	0.260	0.035	0.490	0.057
w/o gate	VF	0.654	0.038	0.527	0.030	0.824	0.049
MGAF-Sum	VF	0.261	0.020	0.293	0.018	0.559	0.031
MGAF-Concat	VF	0.207	**0.016**	**0.212**	**0.013**	**0.433**	**0.029**

TABLE IV
EER (%) UNDER COMBINED VOICE SNR AND FACE OBSTACLE-SIZE DEGRADATIONS.

SNR, Obstacle size	SSA	SSA-Mask	Gate	Gate-Mask	MFV	MFV-Mask	MGAF-Concat	MGAF-Sum	w/o gate
Clean, Clean	0.665	0.537	0.553	0.420	0.543	0.473	**0.207**	0.261	0.654
Clean, ×1.25	1.138	1.048	0.994	0.920	1.180	1.058	**0.468**	0.548	1.431
0dB, Clean	1.005	0.723	0.755	0.580	0.813	0.638	**0.356**	0.457	0.808
0dB, ×1.25	2.202	1.840	1.814	1.500	2.117	1.675	**1.165**	1.234	2.340
Average	1.253	1.037	1.029	0.855	1.163	0.961	**0.549**	0.625	1.308

(SSA-Mask and Gate-Mask) generally outperform their original versions, indicating that the masking mechanism is beneficial not only for unimodal modeling but also for robust fusion under cross-modal degradation.

To further compare performance across operating points, Fig. 4 shows the DET curves on clean Vox1-O and the most challenging degraded condition (SNR= 0 dB, obstacle size ×1.25). MGAF achieves lower false-reject rates than competing baselines over a wide range of false-alarm rates, with a clearer advantage under severe degradation.

D. Ablation on Gating and Fusion Operator

To evaluate the effectiveness of (i) mask-driven adaptive gating and (ii) the fusion operator, we conduct two ablation studies, as summarized in Tables III and IV.

a) MGAF-Concat (default): Our default design first estimates a scalar gate $z \in (0, 1)$ from mask representations and then re-weights the unimodal embeddings before concatenation:

$$\mathbf{e}_p = \mathrm{LN}\Big(\big[\, z\,\mathbf{e}_v;\ (1-z)\,\mathbf{e}_f \,\big]\Big). \tag{24}$$

b) w/o gate (equal-weight concatenation): To isolate the contribution of adaptive gating, we disable the gate by fixing $z = 0.5$ and keep the concatenation operator unchanged:

$$\mathbf{e}_p = \mathrm{LN}\Big(\big[\, 0.5\,\mathbf{e}_v;\ 0.5\,\mathbf{e}_f \,\big]\Big). \tag{25}$$

We compare this variant with MGAF-Concat to isolate the effect of mask-driven weighting in the fusion stage.

c) MGAF-Sum (Gated summation): To study the impact of the fusion operator, we replace concatenation with a Gate-weighted summation. Since the two unimodal embeddings have different dimensions, we first align them with linear projections and then apply Gated summation:

$$\mathbf{e}_p = \mathrm{LN}\big(z\,\phi_v(\mathbf{e}_v) + (1-z)\,\phi_f(\mathbf{e}_f)\big), \tag{26}$$

where $\phi_v(\cdot)$ and $\phi_f(\cdot)$ denote linear projection layers for dimension alignment. This variant is compared with MGAF-Concat to examine whether concatenation or summation is more suitable for mask-driven fusion.

d) Discussion: As shown in Tables III and IV, MGAF-Concat consistently outperforms w/o Gate, demonstrating that mask-driven adaptive gating is critical for suppressing the degraded modality and improving robustness under corruption. Moreover, MGAF-Concat also yields better performance than MGAF-Sum in all settings, suggesting that concatenation better preserves complementary cues from the two modalities, whereas Gate-weighted summation may mix the information and thus limit the discriminability of the fused embedding.

VI. CONCLUSIONS

In this paper, we proposed Mask-Guided Adaptive Fusion (MGAF) for audio-visual person verification under acoustic noise and facial occlusion. By predicting modality-specific degradation masks and converting them into reliability cues, MGAF performs adaptive re-weighting and weighted concatenation of speech and face embeddings. Experiments on VoxCeleb show that MGAF improves robustness under severe and

asymmetric degradations while remaining competitive on clean trials. Future work will explore more realistic degradations, sequence-level fusion, and lighter-weight designs for efficient deployment.

VII. Acknowledgment

This research was supported by the Tianshan Talents Cultivation Program - Leadings Talents for Scientific and Technological Innovation (No. 2024TSYCLJ0002).

References

[1] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[2] S. Wang, Z. Chen, K. A. Lee, Y. Qian, and H. Li, "Overview of speaker modeling and its applications: From the lens of deep speaker representation learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[5] D. Cai, W. Cai, and M. Li, "Within-sample variability-invariant loss for robust speaker recognition under noisy environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6469–6473.

[6] D. Zeng, R. Veldhuis, and L. Spreeuwers, "A survey of face recognition techniques under occlusion," *IET biometrics*, vol. 10, no. 6, pp. 581–606, 2021.

[7] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.

[8] Y. Wu, L. Wang, K. A. Lee, M. Liu, and J. Dang, "Joint feature enhancement and speaker recognition with multi-objective task-oriented network." in *Interspeech*, 2021, pp. 1089–1093.

[9] J.-h. Kim, J. Heo, H.-j. Shim, and H.-J. Yu, "Extended u-net for speaker verification in noisy environments," *arXiv preprint arXiv:2206.13044*, 2022.

[10] H. Qiu, D. Gong, Z. Li, W. Liu, and D. Tao, "End2end occluded face recognition by masking corrupted features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6939–6952, 2021.

[11] B. Huang, Z. Wang, K. Jiang, Q. Zou, X. Tian, T. Lu, and Z. Han, "Joint segmentation and identification feature learning for occlusion face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 875–10 888, 2022.

[12] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6194–6198.

[13] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1079–1092, 2021.

[14] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-visual person recognition in multimedia data from the iarpa janus program," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3031–3035.

[15] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, J. Hernandez-Cordero *et al.*, "The 2019 nist audio-visual speaker recognition evaluation." in *Odyssey*, 2020, pp. 259–265.

[16] J. Alam, G. Boulianne, L. Burget, M. Dahmane, M. D. Sánchez, A. Lozano-Diez, O. Glembek, P.-L. St-Charles, M. Lalonde, P. Matejka *et al.*, "Analysis of abc submission to nist sre 2019 cmn and vast challenge." in *Odyssey*, 2020, pp. 289–295.

[17] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[18] C.-Y. Wu, C.-C. Hsu, and U. Neumann, "Cross-modal perceptionist: Can face geometry be gleaned from voices?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 452–10 461.

[19] R. Tao, Z. Shi, Y. Jiang, D.-T. Truong, E.-S. Chng, M. Alioto, and H. Li, "Multi-stage face-voice association learning with keynote speaker diarization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 342–11 347.

[20] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3995–3999.

[21] S. Hörmann, A. Moiz, M. Knoche, and G. Rigoll, "Attention fusion for audio-visual person verification using multi-scale features," in *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*. IEEE, 2020, pp. 281–285.

[22] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb." in *Interspeech*, 2020, pp. 2252–2256.

[23] R. G. Praveen and J. Alam, "Dynamic cross attention for audio-visual person verification," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024, pp. 1–5.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[26] S. Ge, W. Guo, C. Li, J. Zhang, Y. Li, and D. Zeng, "Masked face recognition with generative-to-discriminative representations," *arXiv preprint arXiv:2405.16761*, 2024.

[27] Q. Fan, Z. Li, J. Li, and C. Cao, "Mode: Mixture of diffusion experts for any occluded face recognition," *arXiv preprint arXiv:2505.04306*, 2025.

[28] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 773–782.

[29] Y. Shi, Q. Huang, and T. Hain, "Robust speaker recognition using speech enhancement and attention model," *arXiv preprint arXiv:2001.05031*, 2020.

[30] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM computing surveys*, vol. 56, no. 9, pp. 1–36, 2024.

[31] P. Sun, S. Zhang, Z. Liu, Y. Yuan, T. Zhang, H. Zhang, and P. Hu, "Learning audio-visual embedding for person verification in the wild," *arXiv preprint arXiv:2209.04093*, 2022.

[32] ——, "A method of audio-visual person verification by mining connections between time series," in *Proc. Interspeech*, vol. 2023, 2023, pp. 3227–3231.

[33] A. Selvakumar and H. Fashandi, "Getting more for less: Using weak labels and av-mixup for robust audio-visual speaker verification," in *Proc. Interspeech 2024*, 2024, pp. 4728–4732.

[34] R. Tao, K. A. Lee, R. K. Das, V. Hautamäki, and H. Li, "Self-supervised training of speaker encoder with multi-modal diverse positive pairs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1706–1719, 2023.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2021, pp. 72–76.

[37] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[38] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition." in *Interspeech*, 2017, pp. 1567–1571.

[39] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[40] M. S. Saeed, S. Nawaz, M. S. Tahir, R. K. Das, M. Z. Zaheer, M. Moscati, M. Schedl, M. H. Khan, K. Nandakumar, and M. H. Yousaf, "Face-voice association in multilingual environments (fame) challenge 2024 evaluation plan," *arXiv preprint arXiv:2404.09342*, 2024.