This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2025.3586484

Adaptive Interaction and Correction Attention Network for Audio-Visual Matching

Jiaxiang Wang, Aihua Zheng*, Lei Liu, Chenglong Li, Ran He, Jin Tang

***Abstract*—Audio-visual matching techniques aim to recognize and match information across different identities by learning a similarity metric across modalities. However, modal differences arise from insufficient cross-modal correlations and noise interference, which substantially hinder the performance of traditional deep metric learning methods in audio-visual matching tasks. To address the modal differences issue, we propose a novel Adaptive Interactive and Correction Attention Network (AICANet). This network efficiently captures deep information connections, generating modality-consistent feature embeddings within a unified metric framework. The core of AICANet is its two-pronged approach to reducing modal differences. First, we propose the Adaptive Interactive Attention (AIA) module, which flexibly establishes associations among cross-modal local features using dynamically generated pseudo-labels. Second, we propose the Adaptive Correction Attention (ACA) mechanism, which employs an adaptive threshold to de-interference effectively and accurately adjust the representation of local feature associations. Notably, the ACA mechanism is suitable for both intra-modal and inter-modal refined attention correction. Additionally, we design a relative distance stretching metric loss ($\mathcal{L}_{RDSM}$), which reinforces the similarity invariance of feature embeddings in a uniform space and enhances matching accuracy. Extensive tests on the VoxCeleb and VoxCeleb2 datasets demonstrate that AICANet outperforms leading existing algorithms across several evaluation metrics, validating its superior performance. The codes can be found at https://github.com/w1018979952/AICANet.**

***Index Terms*—Audio-Visual Matching, Adaptive Interaction Attention, Adaptive Correction Attention, Modal Differences.**

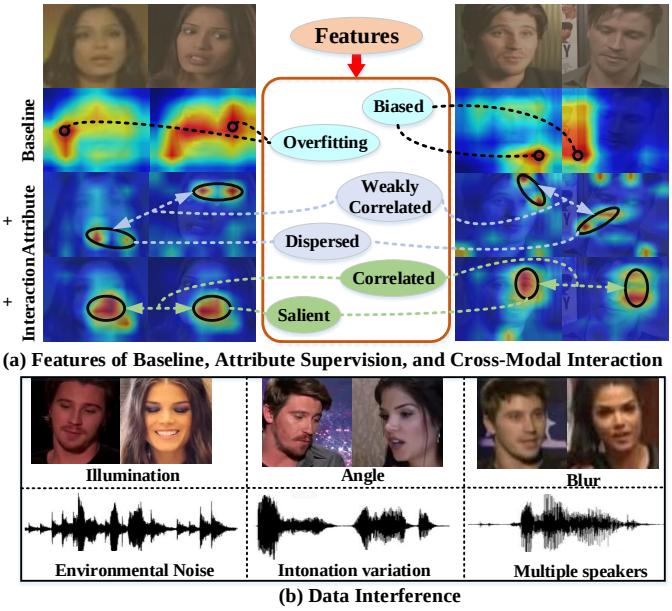


Fig. 1. (a) Comparison of feature distributions for baseline, attribute supervision, and cross-modal interaction. (d) Intermodal interference components (e.g., background noise or visual occlusion) significantly degrade interaction effectiveness.

I. Introduction

Audio-visual matching represents a significant research direction in machine learning and computer vision, concentrating on efficiently learning discriminative features to assess similarity and identify potential matching relationships. Metric learning reduces the distance between similar samples by mapping data from various modalities into a unified embedding space while preserving the separability of different classes. To this end, Deep Metric Learning (DML) employs deep neural networks to capture highly nonlinear feature-embedding semantic information. It has been extensively applied in downstream tasks such as cross-modal retrieval [1]–[4], face recognition [5]–[9], and person and vehicle re-identification [10]–[12]. Furthermore, DML has significantly advanced audio-visual matching techniques, contributing to applications such as audio-visual speech separation [13], [14], speaker recognition [15]–[18], and audio-visual localization [19], [20].

Recent audio-visual matching advancements powered by deep neural network technology have significantly outperformed human recognition capabilities. However, the fundamental differences between the sensory modalities of vision and hearing remain a barrier to further improving matching accuracy. To address this issue, Wang *et al.* [21] and Nawaz *et al.* [22] introduced a feature embedding strategy designed to reduce disparities between sensory features by mapping them into a shared space. Despite this effort, the embedding

This research is supported in part by the Scientific Research Foundation for High-level Talents of Anhui University of Science and Technology (2024yjrc95, 2024yjrc94), the National Natural Science Foundation of China (62372003), the University Synergy Innovation Program of Anhui Province (GXXT-2022-036), and the Natural Science Foundation of Anhui Province (2408085QF199, and 2308085Y40). The corresponding author is Aihua Zheng.

J. Wang and L. Lei are with the School of Artificial Intelligence, Anhui University of Science and Technology, Hefei 232001, China (e-mail: Netizen-wjx@foxmail.com; liulei970507@163.com).

A. Zheng and C. Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).

R. He is with the University of Chinese Academy of Sciences, Beijing 101408, China, and also with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100049, China (e-mail: rhe@nlpr.ia.ac.cn).

J. Tang is with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: tangjin@ahu.edu.cn).

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2025.3586484

process fails to eliminate inherent sensory differences due to insufficient supervision. Following this, Zheng *et al.* [23] and Cheng *et al.* [24] applied Generative Adversarial Networks (GANs) to achieve a Nash equilibrium, effectively transforming cross-modal features into modality-independent representations. These features were then processed using DML techniques to identify matching pairs. However, direct DML cannot process sensory information directly, limiting its capacity to establish complex associations between samples and enhance feature differentiation.

To address this limitation, Wang *et al.* [25] proposed an attribute-guided feature interaction and enhancement network, designed to improve cross-modal sample interactions and enable DML to learn more precise matching relationships. As illustrated in Fig. 1 (a), attribute supervision mitigates model overfitting and prediction bias compared to the baseline. However, it tends to become distracted when processing multiple attributes simultaneously. The lack of an effective cross-modal interaction mechanism can limit the correlation across attribute features, leading to performance bottlenecks. Thus, an adaptive cross-modal interaction architecture is necessary to extract correlation salient features for audio-visual matching. In real-world scenarios, audio-visual data often contains interference, which exacerbates modal differences, as shown in Fig. 1 (b). Existing attention-based cross-modal interaction methods are particularly vulnerable to such interference, impairing matching performance. To address this, Yu *et al.* [26] and Ning *et al.* [27] both proposed feature separation architectures to suppress interfering features and strengthen face-speech correlations. However, in practice, interfering and valid features are often entangled, making complete separation infeasible. Wen *et al.* [28] introduced a dynamic weighting strategy to reduce interference at the sample level. Still, this approach diminishes sample utilization and fails to fully leverage informative features in challenging samples. Moreover, sample diversity complicates matching relationship learning, as merely minimizing intra-class distances and maximizing inter-class distances proves insufficient. This underscores the need for dynamic DML methods. Current metric-based approaches [23]–[25], [28] primarily focus on alignment feature learning, often neglecting these requirements. Consequently, modal discrepancies in audio-visual data remain a significant challenge for matching models.

To mitigate cross-modal differences in DML, we propose three key innovations: (1) an Adaptive Interactive Attention (AIA) module, (2) an Adaptive Corrective Attention (ACA) mechanism, and (3) Relative Distance Stretching Metric loss ($\mathcal{L}_{RDSM}$). As illustrated in Fig. 2 (a), the AIA module leverages weak audio-visual correlations to guide cross-modal interactions. It employs skip connections to stabilize gradient propagation and incorporates dynamic feature fusion to suppress noise, enabling selective feature alignment. Addressing the noise interference shown in Fig. 2 (b), ACA employs modality-specific filters for adaptive corrective attention to selectively enhance robust feature representations, thereby effectively reducing cross-modal distribution divergence. While existing methods construct modal associations via shared feature space embedding (Fig. 2 (c)), they often lack sufficient intra-class

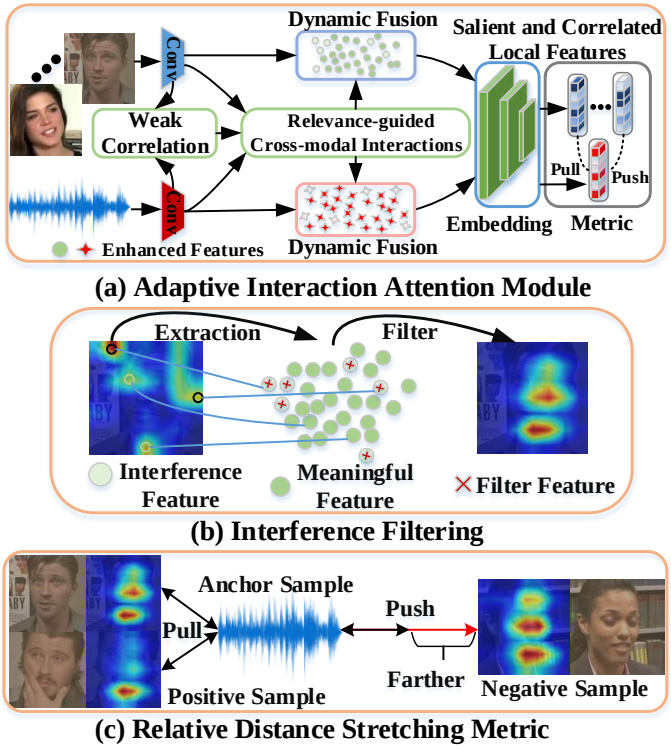


Fig. 2. (a) The raw signals exhibit both weak cross-modal correlations and noise interference; (b) The image and audio information have interference factors; (c) Intra-class features vary significantly, while inter-class features exhibit similarity.

compactness and inter-class separation. $\mathcal{L}_{RDSM}$ addresses this by dynamically optimizing intra-modal variance and inter-modal correlations through relative distance constraints while suppressing spurious feature associations.

In this paper, we propose a novel adaptive interaction and correction attention network that aims to enhance the audio-visual DML capability by manipulating intra- and inter-modal attention, mining cross-modal potential connections, and excluding irrelevant features. First, we propose the Adaptive Interactive Attention (AIA) module that utilizes dynamically generated relational pseudo-labels to guide cross-modal first-order sparse attentional interactions. Subsequently, the second-order intramodal attentional interactions are implemented to evoke visual and audio modality intrinsic semantic connections. Then, we propose the Adaptive Correction Attention (ACA) mechanism, which effectively filters interfering feature representations by the adaptive threshold to regulate the representation of associations between local features precisely. The mechanism is suitable for both intra-modal and inter-modal refined correction of attention. In addition, we design a relative distance stretching metric loss ($\mathcal{L}_{RDSM}$), which aims to enhance the DML embedding of audio-visual features' representations in a uniform space, thereby improving the matching robustness.

The core contributions of this paper can be summarized as follows.

- We propose the adaptive interactive attention (AIA) module, which explores intrinsic local connections between

cross-modal semantic features. Due to the modal differences, the AIA module employs second-order attention to capture potential associations between audio-visual features.

- We propose an adaptive correction attention (ACA) mechanism that regulates the representation of local feature associations by filtering out interfering features through an adaptive threshold. This mechanism can be corrected for both intra-modal and inter-modal attention.
- We design a relative distance stretching metric loss ($\mathcal{L}_{RDSM}$), which leverages the distance relationships between anchor, positive, and negative samples. $\mathcal{L}_{RDSM}$ aims to enhance the DML embedding representations of audio-visual features in a unified space, facilitating robust audio-visual matching.
- Comprehensive experiments conducted on the VoxCeleb [29] and VoxCeleb2 [30] datasets validate the model's effectiveness. These experiments also confirm the complementarity and effectiveness of each component, demonstrating superior performance compared to state-of-the-art (SOTA) algorithms.

II. Related Works

Audio-Visual Attention Mechanism. Attention mechanisms, with their ability to autonomously focus on relevant information, have been widely applied across various domains, particularly in enhancing cross-modal semantic associations. In cross-modal interactions, attention mechanisms help concentrate on the most relevant region-word pairs while filtering out weaker matches [31]. Zhang *et al.* [32] proposed an enhanced semantic similarity learning approach, extending this metric to dynamic, learnable frameworks to explore multidimensional correspondences between visual and textual features. However, unlike image-text cross-modal retrieval, audio-visual matching lacks explicit matching objects, making applying existing attention mechanisms less effective. To address this, Mercea *et al.* [33] proposed a cross-attention module that captures shared information between audio and visual representations to enhance cross-modal semantic feature learning. Similarly, Saeed *et al.* [34] suggested generating rich fusion embeddings using bimodal complementary cues for orthogonal identity feature clustering under constraints. However, in unknown identity-matching scenarios, these methods struggle to enhance feature-guided learning accurately. To overcome this limitation, Wang *et al.* [25] proposed an attribute-guided interaction enhancement module, which significantly improves audio-visual matching by enhancing the differentiation of local features in both modalities. Given the inherent inter-modal differences, directly learning cross-modal feature associations remains challenging. To address this, we propose the Adaptive Interactive Attention (AIA) module, which leverages a second-order attention mechanism to explore semantic connections between visual and audio modalities deeply.

Audio-Visual De-Interference. Audio-visual matching data, originating from real-world scenes, inevitably introduces noise, resulting in cross-modal feature discrepancies. To address this issue, researchers apply separate representation learning to enable models to identify and differentiate independent variables, thereby minimizing the impact of interfering features [35]. In audio-visual matching, the correlation between voices and faces often depends on implicitly extracted high-level attributes such as gender, age, and ethnicity. Wen *et al.* [36] developed a disjoint mapping network that leverages attribute recognition mechanisms to explore diverse attribute features and reduce modal differences. However, the collection of attribute labels requires extensive manual effort, making data acquisition challenging. Ning *et al.* [27] proposed β-control disentanglement of latent variables, which separates identity-related features, filters modality-specific features, and facilitates cross-modal associations. However, the non-adaptive nature of β makes accommodating diverse tasks and datasets difficult. Yu *et al.* [26] designed a framework for decoupled cross-modal latent representation, aimed at removing interfering features and strengthening face-speech connections. Despite these advances, accurately distinguishing between interfering and discriminative features remains a challenge, limiting the generalization capability of existing audio-visual matching models. To mitigate this issue, Wen *et al.* [28] proposed a dynamic weighting strategy that learns to match global face and audio features and evaluates the importance of samples, excluding negative factors that hinder generalization. However, while sample filtering can reduce interference, it may also eliminate too many samples, weakening the model's capacity to learn effective features. To address noise-induced modal differences, we propose an Adaptive Correction Attention (ACA) mechanism that employs an adaptive threshold to filter interfering features and adjust local feature association representations.

Audio-Visual Deep Metric Learning. Audio-visual matching originated in psychology, where dual-stream networks, empowered by deep learning, surpassed human recognition capabilities, opening new research avenues [37]. However, the inherent differences between vision and hearing limit the effectiveness of audio-visual matching. To address this, Nagrani *et al.* [38] introduced joint learning with contrastive loss to improve feature representation, while Wang *et al.* [21] developed an end-to-end joint embedding network that applies bidirectional sorting, identity, and centrality constraints on a small dataset to explore audio-visual features for deep metric learning. Yet, feature embedding alone struggles to overcome the challenge of modal differences. Inspired by adversarial mechanisms, Zheng *et al.* [23] proposed an adversarial metric to effectively reduce modal discrepancies, and Cheng *et al.* [24] introduced an adversarial embedding network that combines triplet loss with modal centrality loss, reinforcing audio-visual connections in deep matching. However, focusing solely on modal differences is insufficient, as a deeper exploration of audio-visual features is required. In response, Wang *et al.* [39] proposed a dual-enhanced Siamese-adversarial network that not only enhances audio-visual feature representation but also deepens metric learning by leveraging a Siamese-adversarial mechanism, thus improving audio-visual matching performance. Despite these advances, the distribution metric of audio-visual features in a unified feature space remains underexplored, which is crucial for audio-visual matching. Therefore, we propose a Relative Distance

This article has been accepted for publication in IEEE Transactions on Information Forensics and Security. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2025.3586484

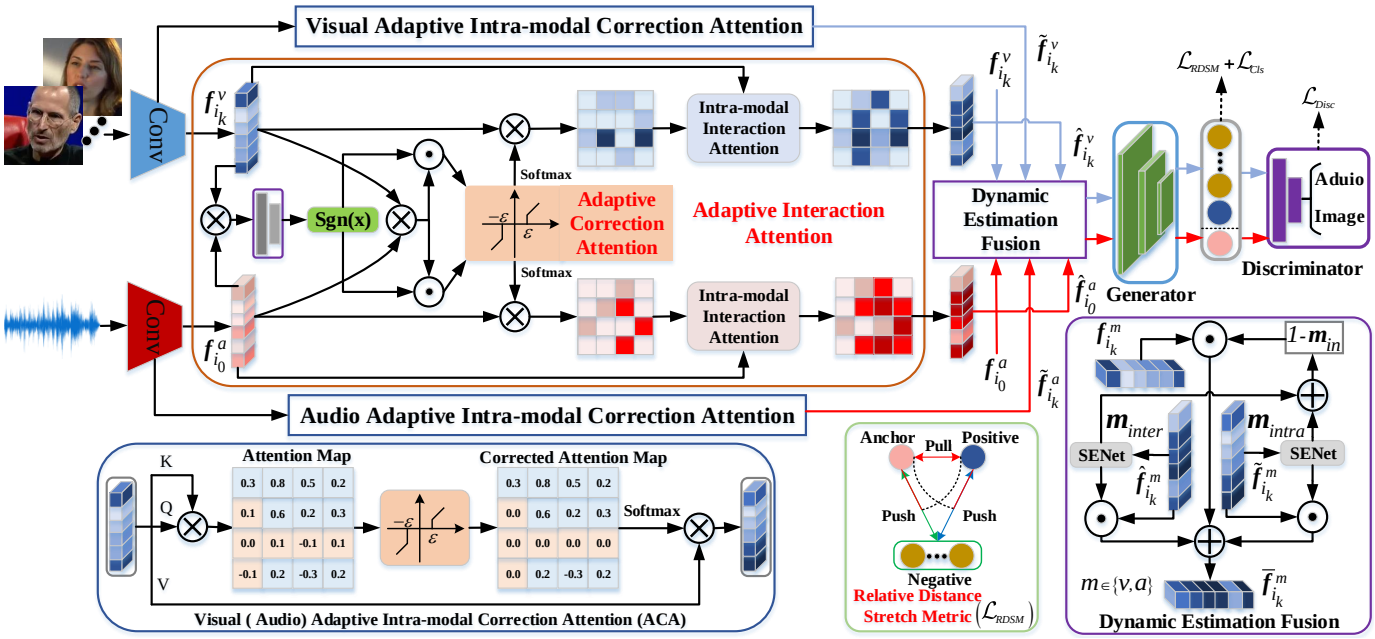


Fig. 3. The AICANet architecture integrates three core components: (1) an Adaptive Interactive Attention (AIA) module that models cross-modal correlations via second-order attention, (2) an Adaptive Correction Attention (ACA) mechanism that suppresses noise through adaptive thresholding, and (3) a Relative Distance Stretch Metric loss ($\mathcal{L}_{RDSM}$) that optimizes feature space geometry. AIA addresses cross-modal discrepancies by capturing local semantic relationships, while ACA mitigates noise amplification in conventional attention. To overcome pseudo-label dependency limitations, $\mathcal{L}_{RDSM}$ enhances alignment through relative distance optimization, imposing strict penalties for mismatches and improving matching robustness.

Stretching Metric loss ($\mathcal{L}_{RDSM}$), which aims to improve the discriminative embedding of audio-visual features in a unified space, enhancing robust audio-visual matching by the clever use of distances between anchor samples, positive samples, and negative samples.

III. Method

To address the problem of audio-visual modal differences due to insufficient cross-modal correlation and noise interference, we propose a novel Adaptive Interaction and Correction Attention Network (AICANet) framework, as shown in Fig. 3. First, we provide an overview of audio-visual matching and introduce the Adaptive Interaction Attention (AIA) module, which learns cross-modal feature associations. Next, we describe the Adaptive Correction Attention (ACA) mechanism, designed to filter out interference information. Finally, we explain matching cross-entropy loss for training the network and the Relative Distance Stretching Metric loss ($\mathcal{L}_{RDSM}$) for enhanced audio-visual matching accuracy.

A. Overview

The audio-visual matching task establishes identity correspondence through cross-modal recognition: given an audio clip (or face image), the system identifies matching identities from a gallery of candidate face images (or audio clips). This task operates in two distinct modes: (1) In the voice-to-face (V-F) scenario, audio serves as the anchor to matching identities among k candidate face images. (2) In the face-to-voice (F-V) scenario, face serves as the anchor to matching identities among k candidate audio clips. The V-F scenario operates as follows: an anchor audio clip a_{i_0} serves as the identity query, while k face images $\{v_{i_1}, v_{i_2}, \cdots, v_{i_k}\}$ constitute the matching gallery. Feature extraction employs ResNet18 [40] for face images and ResNetSE [40] for audio clips. For the i-th training tuple: the audio clip is denoted as $\boldsymbol{f}^a_{i_0}$, while the face image is denoted as $\boldsymbol{f}^v_i = \{\boldsymbol{f}^v_{i_1}, \cdots, \boldsymbol{f}^v_{i_k}\}$. The designation for different tasks is determined by the number of candidate libraries k values, where $k > 1$ denotes the matching task and $k = 1$ denotes the verification task. The batch size is set to N, where i indexes the i-th training sample tuple.

B. Adaptive Interaction Attention Module

Attention mechanisms autonomously focus on relevant information, enabling cross-modal interactions to strengthen audio-visual feature associations and reduce modality gaps caused by insufficient audio-visual connectivity. Audio-visual matching requires identifying unique matches from multiple candidate samples, necessitating explicit matching labels to guide interaction associations. However, these labels are frequently unavailable, particularly during testing. To address the modal differences due to insufficient audio-visual associations, we propose an adaptive interaction attention module that uses predicted relevance as pseudo-labels to guide cross-modal interaction, thereby activating intrinsic semantic links between visual and audio modalities.

Relevant Learning Network: To enable audio-visual interaction, we first design a relevant learning network that captures correlations between audio and visual modalities, using these relevances as pseudo-labels to guide cross-modal interaction. Specifically, the relevant learning network includes

a Compact Bilinear Pooling Network (CBPNet) [41] and a fully connected layer to compute audio-visual relevance, formulated as follows:

$$R_{i_{0k}} = FC(CBPNet(\boldsymbol{f}^a_{i_0}, \boldsymbol{f}^v_{i_k})), \tag{1}$$

where $R_{i_{0k}}$ denotes cross-modal relevance which is used as pseudo-labeling. To ensure the accuracy of the network estimation, we utilize the real labels in the training set for further supervision and perform the estimation directly in the testing phase. Its computational relevant loss $\mathcal{L}_{Rel}$ is shown below:

$$\mathcal{L}_{Rel} = \frac{1}{2k}(1+\cos(\frac{epoch}{N}\pi))\sum_{j=1}^{k}(\sigma(\tau R_{i_{0k}}) - l_{i_k})^2, \tag{2}$$

where $epoch$ and N represent the current and total number of iterations, respectively, while σ denotes the sigmoid activation function. The temperature control parameter τ is set to 5. During training, the model receives l_{i_k} as the ground-truth matching relation label. This label guides the network to estimate correlations aligning with the matching relationships, thereby making the adaptive interaction relationships explicit. However, since accurately predicting correspondence remains challenging and the model can only estimate correlations as regression values, we employ a sign function to rectify these values, preserving only their positive/negative directions while disregarding their magnitudes. To implement this approach effectively, we incorporate simulated annealing weights [42] to reduce correlation estimates when computing the loss function progressively.

Relevance-guided Cross-modal Interactions: We then compute the similarity between audio clips and face images to mine the semantics of cross-modal interaction features relevant to the matching task. The cross-modal attention is shown below:

$$E^m_{i_k} = \boldsymbol{f}^m_{i_k}(\boldsymbol{f}^m_{i_k})^T, \tag{3}$$

where k denotes the sample index, corresponding to audio samples when $k = 0$ and to image samples when $k > 0$. The term $m \in \{a, v\}$ represents either the audio or visual modality. The variable $E^m_{i_k}$ represents the correlation matrix between the audio clip and the face image. Therefore, we employ relevance-guided cross-modal attention can be denoted as follows:

$$\widehat{E}^m_{i_k} = sign(R_{i_{0k}}) \odot E^m_{i_k}, \tag{4}$$

where $\odot$ is the product of elements. $m = a$, then $k = 0$. while $m = v$, then $k \neq 1$. $\widehat{E}^m_{i_k}$ is the cross-modal attention behind the relevance-guided. $sign$ is the sign function, defined as follows:

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}. \tag{5}$$

Inspired by Jiang *et al.* [43], we observed that relying solely on cross-modal attention mechanisms to reduce inter-modal differences does not necessarily improve downstream task performance. Therefore, we utilize a sparse correlation matrix to explore the most potential local correlation feature regions between audio and visual modalities, which is formulated as follows:

$$\overrightarrow{f^m_{i_k}} = softmax(\delta \widehat{E}^m_{i_k})\boldsymbol{f}^m_{i_k}, \tag{6}$$

where the hyperparameter δ adjusts the similarity matrix to produce sparse attention, which is set to 10. $\overrightarrow{f^m_{i_k}}$ represents the audio-visual features enhanced by first-order cross-modal interactions, primarily focusing on the sparse association region. However, this approach does not fully capture cross-modal associations. Inspired by Fan *et al.* [44], we employ a self-support mechanism for second-order attentional interaction. This approach utilizes first-order interaction-enhanced features as key (K) elements while maintaining original modal features as both query (Q) and value (V) components, enabling comprehensive exploration of audio-visual semantic associations. The self-support mechanism operates as follows:

$$\hat{f}^m_{i_k} = softmax(\boldsymbol{f}^m_{i_k}(\overrightarrow{f^m_{i_k}})^T)\boldsymbol{f}^m_{i_k}, \tag{7}$$

where $\hat{f}^m_{i_k}$ denotes the feature representation of audio and face images after an adaptive interaction, which has a strong cross-modal correlation.

Dynamic Estimation Fusion: As exact modal alignment is often suboptimal for downstream prediction tasks, optimizing performance depends on capturing meaningful underlying modal structures rather than achieving perfect alignment [43]. To address this, we design an adaptive interaction attention module that produces features based on a dynamic estimation fusion of interaction and original modal features. This approach implicitly captures statistical modal information, enhancing feature semantics in downstream tasks. The dynamic estimation fusion module selectively incorporates intra- and inter-modal features through estimated mask values, combining them with original features to facilitate effective information transfer.

$$\begin{aligned}\overline{\boldsymbol{f}}^m_{i_k} = (1 - m_{in}) \odot \ & \boldsymbol{f}^m_{i_k} + m_{inter} \odot \ IN(\hat{f}^m_{i_k}) \\ & + m_{intra} \odot \ IN(\tilde{f}^m_{i_k}),\end{aligned} \tag{8}$$

where $\tilde{f}^m_{i_k}$ is the intra-modal feature representation after self-attention. m_{Inter} and m_{Intra} denote the mask values derived from inter-modal interaction features and intra-modal correction features, respectively, as calculated by SENet [45]. $m_{in} = m_{inter} + m_{intra}$. In the dynamic estimation fusion process, we apply instance normalization (IN) to reduce the variance of modal features.

C. Adaptive Correction Attention Mechanism

Audio-visual data in real-world scenarios inevitably contain noise, leading to discrepancies between modalities. While attentional mechanisms can effectively enhance task-relevant semantic features, they may also unintentionally amplify noise. To address modality differences caused by noise interference, we propose an adaptive Correction attention mechanism that reduces noise effects by applying the adaptive threshold to the correction attention correlation matrix.

Adaptive Threshold: The adaptive threshold has been a cornerstone of signal denoising for decades [46]. In traditional

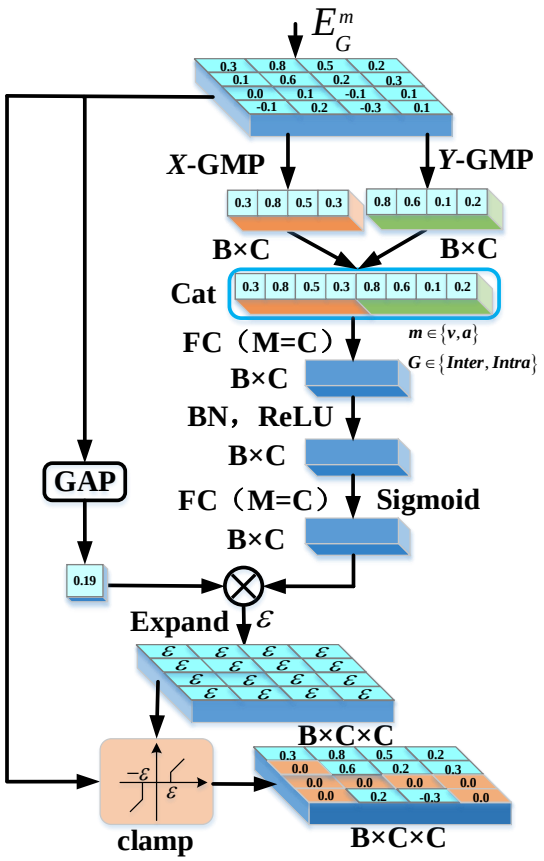


Fig. 4. Diagram of the architecture of the adaptive correction attention mechanism.

wavelet thresholding, for instance, denoising typically involves three stages: wavelet decomposition, soft thresholding, and wavelet reconstruction. The adaptive threshold converts valuable information into positive or negative features and filters out noise [47]. However, different signal types require specifically tailored filters, which demand considerable expertise.

Deep learning offers an alternative by learning signal correlations through convolutional methods, analogous to traditional filtering functions. Therefore, a similar denoising effect can be achieved by integrating a soft thresholding layer as a nonlinear transformation layer within the network architecture. As shown in Fig. 4, this layer performs Global Maximum Pooling (GMP) along the *X*- and *Y*-axes of the features and combines the results into a single vector. This vector is subsequently processed by a two-layer fully connected (FC) network to calculate the threshold parameter ε, constrained to the range (0,1), as shown below:

$$\varepsilon = \sigma(FC(cat(P_x(E_G^m), P_y(E_G^m)))) \odot GAP(E_G^m), \quad (9)$$

where σ denotes the sigmoid activation function and GAP denotes global average pooling. P_x and P_y represent the maximum pooling operations along different matrix directions. Here, E_G^m denotes the correlation matrix, which can either represent inter-modal correlations ($G = Inter$) or intra-modal autocorrelations ($G = Intra$). By adjusting the correlation matrix for different types of attention, interference can be minimized. As an example, the inter-modal correlation matrix is shown below, and the correction process in Eq. (4) is as follows:

$$\widehat{E}_{Inter}^m = sign(R_{i_{0k}}) \odot clamp(E_{Inter}^m, \pm\varepsilon), \quad (10)$$

where the *clamp* symbol is a truncation operator that adjusts according to an adaptive threshold. E_{Inter}^m corresponds to $E_{i_k}^m$. As a result, the intra-modal and inter-modal feature representations that are enhanced by correcting the attention are shown below:

$$f_G^m = softmax(\delta \widehat{E}_G^m) \boldsymbol{f}_{i_k}^m, \quad (11)$$

where $G \in \{Inter, Intra\}$, f_{Inter}^m corresponds to $\hat{f}_{i_k}^m$, while f_{Intra}^m corresponds to $\tilde{f}_{i_k}^m$.

D. Objective Function

Audio-visual modal interaction enables learning cross-modal associations, while adaptive attention correction minimizes the influence of distracting features, both of which help mitigate modal discrepancies. However, significant differences in deep audio-visual features persist. Traditional approaches address this by embedding features in a low-dimensional space through adversarial methods to minimize modal feature differences while retaining relevant features. Accordingly, we also apply an adversarial approach to generate audio features $\boldsymbol{f}_{i_0}^a$ and face image features $\{\boldsymbol{f}_{i_1}^v, \cdots, \boldsymbol{f}_{i_k}^v\}$ into modality-independent representations $\{\boldsymbol{h}_{i_0}, \cdots, \boldsymbol{h}_{i_k}\} \in \mathcal{H}$. In this adversarial process, we introduce discriminators to improve generation quality, with the discriminative loss defined as follows:

$$\mathcal{L}_{Disc} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=0}^{k} N_{i_j} \log D(\boldsymbol{h}_{i_j}), \quad (12)$$

where M represents the number of training data tuples, with N_{i_j} as the modality label of the jth sample in the ith data tuple, and $D(\boldsymbol{h}_{i_j})$ as the modality probability output by the discriminator D. The cross-entropy [48] calculates the loss for the discriminator's predictions. The identification of matching candidate objects is determined by estimating probabilities through a nonlinear discriminative network implemented using a multilayer perceptron.

$$\mathcal{L}_{Cls} = -\frac{1}{M} \sum_{i=1}^{M} (l_{i_j} \log C_m([exp(\boldsymbol{h}_{i_0} - \boldsymbol{h}_{i_1}), \cdots, exp(\boldsymbol{h}_{i_0} - \boldsymbol{h}_{i_k})]), \quad (13)$$

where C_m denotes the matching classification. The l_i is the matched identity label.

Relative Distance Stretching Metric Loss: Both Adaptive Interactive Attention and Adaptive Correction Attention implicitly reduce modal differences. However, the generated modal-independent features still require an explicit matching metric to further improve the model's recognition performance. To address this, we propose a matching metric that optimizes intra-class proximity while maximizing inter-class separations. Considering the diversity of samples within

classes, we design this metric as a relative distance stretching loss, which can utilize the distance relationship between samples to enhance the audio-visual matching feature embedding representations, thus promoting robust audio-visual matching.

$$\mathcal{L}_{RDSM} = \frac{1}{2M}\sum_{i=1}^{M} max(D_i, 0), \tag{14}$$

$$D_i = \mu_1 log(1 + \mu_2(\max_{j\in[1,k]} w_{l_i} e^{\theta - d_{i_0,i_j}} e^{d_{i_0,i_p} - d_{i_0,i_j}} + \max_{q\in[1,k]} w_{l_i} e^{\theta - d_{i_j,i_q}} e^{d_{i_0,i_p} - d_{i_j,i_q}})) + d_{i_0,i_p}, \tag{15}$$

where μ_1 and μ_2 are hyperparameters set to 10 and 4, respectively, while θ is set to 1.2. We compute Euclidean distances among anchor, positive, and negative samples. The distance between the anchor sample $\boldsymbol{h}_{i_0}$ and the positive sample $\boldsymbol{h}_{i_p}$ ($p \in [1, k]$) is denoted by d_{i_0,i_p}. Similarly, d_{i_0,i_j} denotes the distance between the anchor sample $\boldsymbol{h}_{i_0}$ and the negative sample $\boldsymbol{h}_{i_j}$, and d_{i_j,i_q} represents the distance between the positive sample $\boldsymbol{h}_{i_j}$ and the negative sample $\boldsymbol{h}_{i_p}$. In the V-F scenario, audio clips serve as anchor samples, with corresponding face images as positive samples and all others as negative samples. Conversely, in the F-V scenario, face images are anchor samples, matching audio clips represent positive samples, and non-matching ones constitute negative samples. We define w_{l_i} as the identity match label, where $w_{l_i} = 1$ if there is no match and 0 otherwise. This label is used to compute distances among negative samples within candidate targets. Among multiple candidate negative samples, the closest negative sample to the positive sample is identified, and its distance is used to enforce the relative stretching distance constraint.

The overall objective function loss is computed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{Disc} + \alpha\mathcal{L}_{RDSM} + \beta\mathcal{L}_{Cls} + \gamma\mathcal{L}_{Rel}, \tag{16}$$

where α, β, and γ are hyperparameters setting by hyperparametric analysis experiments.

IV. Experiments

A. Dataset Description.

VoxCeleb [29] is a substantial dataset extensively employed in speaker recognition research, sourced from publicly available YouTube videos. Through automated extraction of audio clips from these videos and subsequent categorization by speaker, VoxCeleb [29] ensures the diversity and authenticity of its data. This dataset comprises over 149354 audio clips spanning 1225 distinct speakers. Correspondingly, the associated VGGFace dataset comprises 137060 facial images extracted from the same videos. Formerly, a protocol was established [28], [37] where individuals whose names begin with "A" or "B" were allocated for verification purposes, while those with names starting with "C," "D," or "E" were designated for testing. The remaining individuals, with initials ranging from "F" to "Z," were designated for training. All subsequent experiments in this study adhere to this data partitioning protocol unless stated otherwise.

VoxCeleb2 [30] dataset, also derived from public YouTube videos, has five times the amount of data compared to VoxCeleb and contains all the data of VoxCeleb [29]. This augmentation enriches the dataset's diversity and underscores the potential of voiceprint recognition techniques. Hence, we conducted experiments utilizing the VoxCeleb [29] training model with VoxCeleb2 test data to assess its capacity for audio-visual matching generalization. The test data encompasses 118 identities across 4911 videos, adhering to the official protocol of VoxCeleb2 for division details.

TABLE I
THE MODEL LEARNING RATE ADJUSTMENT PROCESS.

Epochs	$1 \sim 25$	$26 \sim 40$	$41 \sim 50$
Feature extractor	5×10^{-2}	5×10^{-3}	5×10^{-4}
Generator	5×10^{-3}	5×10^{-4}	5×10^{-5}
Discriminator	5×10^{-3}	5×10^{-4}	5×10^{-5}
Classifier	5×10^{-2}	5×10^{-3}	5×10^{-4}
AIA	5×10^{-3}	5×10^{-4}	5×10^{-5}
ACA	5×10^{-3}	5×10^{-4}	5×10^{-5}

B. Implementation Details

1) Network architecture. The proposed AICANet method is implemented using the NVIDIA RTX 3090 graphics card and the PyTorch framework. For feature extraction, face images and audio clips are processed using pre-trained ResNet18 [40] and ResNetSE [40] models, respectively, with ResNet18 utilizing pre-trained weights from the ImageNet dataset [50]. The input data consists of $224 \times 224 \times 3$ face images and audio sequences of 160000 in length. The Adaptive Interactive Attention (AIA) module facilitates feature correlation across modalities, while the Adaptive Correction Attention (ACA) mechanism removes noisy features, both preserving the output feature dimensions at $512 \times 3 \times 3$ for audio-visual features. Additionally, an adversarial network addresses modal differences by transforming the features into a uniform feature space, making them modality-independent. The $\mathcal{L}_{RDSM}$ function is designed to improve the deep matching capability of auditory features within this space. Finally, the fused audio-visual features are compressed to 128 dimensions by a multilayer perceptron and input into a classifier to predict the matching probability.

2) Training parameters. The proposed AICANet method requires the independent optimization of the feature extractor, generator, discriminator, classifier, Adaptive Interactive Attention (AIA) module, and Adaptive Correction Attention (ACA) mechanism. The initial learning rates for each module are shown in Table I, with a decay factor of 0.1 applied at the 25th and 40th epochs over a total of 50 training epochs. The model is optimized using the Adaptive Moment Estimation (Adam) optimizer, with a batch size of 50, a momentum factor of 0.9, and a weight decay of 0.0005. In the validation experiments, the model addresses a specific matching task by classifying 256-dimensional features to determine whether candidate samples match. For the matching task, the input to the classification network consists of $K \times 128$-dimensional features, derived by subtracting each facial feature from the

TABLE II
THE QUALITATIVE RESULTS OF THE MATCHING TASK ARE EVALUATED BY VALIDATION ON VOXCELEB [29] FOR DIFFERENT SCENARIOS. WHEN K=1, IT DENOTES A VALIDATION; A BINARY OUTCOME INDICATES A 1:2 MATCH, WHILE A MULTI-WAY RESULT INDICATES A 2:K (K=10) MATCH.

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (ACC)	
		V-F	F-V	V-F	F-V	V-F	F-V
SVHF [37]	CVPR2018	81.0	79.5	34.5	×	-	-
DIMNet [36]	ICLR2019	81.3	81.9	38.4	36.2	81.0	81.2
Wang's [21]	ACM2020	83.4	84.2	39.7	36.4	82.6	82.9
Wen's [28]	CVPR2021	87.2	86.5	48.2	44.8	87.2	87.0
AML [23]	TMM2022	90.2	86.3	46.2	43.7	86.4	86.2
DCLR [26]	ICDM2022	86.79	87.45	-	-	86.76	86.89
DSANet [39]	TMM2023	92.5	88.4	49.1	46.8	87.4	91.5
P^2VANet [49]	TCSVT2024	93.1	90.4	**50.6**	48.1	88.5	88.7
ACIENet [25]	TIFS2024	96.0	92.3	49.5	47.1	90.1	91.9
Baseline	-	94.8	89.8	48.5	45.6	88.2	91.2
AICANet	**Ours**	**97.6**	**97.3**	49.8	**48.3**	**90.5**	**93.3**

corresponding audio feature. The network outputs K match probabilities, with the highest probability indicating the matching sample. Performance is evaluated using the accuracy rate (ACC) to assess the results of the audio-visual matching experiments.

C. Comparison to the State-of-the-Arts

To validate the effectiveness of the proposed method, we compare AICANet against nine state-of-the-art methods, including SVHF [37], DIMNet [36], Wang's [21], Wen's [28], AML [23], DCLR [26], DSANet [39], P^2VANet [49], and ACIENet [25]. Among these, AML employs both adversarial and metric learning strategies to minimize inter-modal differences, while DSANet enhances audio-visual matching performance by combining augmentation learning and Siamese adversarial networks. P^2VANet and ACIENet further improve model generalization by leveraging attribute supervision and adversarial learning to explore correlations between audio-visual deep features.

Table II shows the experimental results of executing the Wen-based data split method [28] on **VoxCeleb** [29] data. These results indicate that our method significantly outperforms AML and DSANet in validation, binary matching, and multi-way matching tasks. Our approach enhances audio-visual matching by leveraging cross-modal correlation and interference filtering, effectively reducing modal discrepancies. Compared to attribute-supervised and adversarial joint strategies, our method also demonstrates superior performance across diverse scenario tasks, which proves its superiority and effectiveness.

To compare more classical algorithms, we present the experimental results of the PINs-based data split method [38] on the **VoxCeleb** [29] dataset, as shown in Table III. The results demonstrate that our approach achieves state-of-the-art performance in validation, binary matching, and multi-way matching tasks. This confirms that our method effectively mitigates modal differences between audio-visual data, thereby enhancing audio-visual matching performance. To further showcase AICANet's capabilities, we conducted a $2:K$ multi-way audio-visual matching experiment for comparison. As illustrated in Fig. 5, the model's performance decreases as the number of matching candidates increases, highlighting the growing difficulty of identity differentiation. Our method is significantly better than other methods when there are fewer than five candidate samples, and is still highly competitive when there are more than five candidate samples. We propose

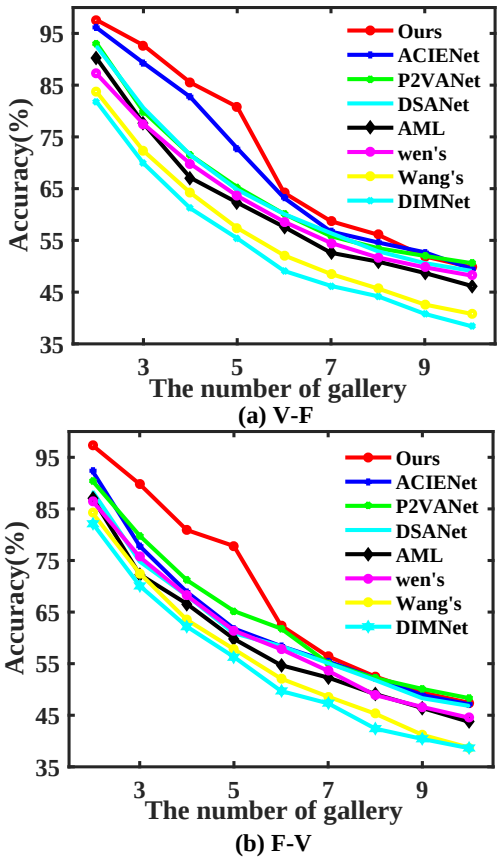


Fig. 5. The quantitative results of $2:K$ $(K=10)$ matching task in V-F and F-V scenarios on VoxCeleb [29].

TABLE III
COMPARISON RESULTS OF AUDIO-VISUAL MATCHING WITH THE SOTA METHOD IN DIFFERENT SCENARIOS ON VOXCELEB [29]. THE EXPERIMENTAL RESULTS IN THE TABLE ARE OBTAINED FOLLOWING THE DATA SETTINGS PROPOSED BY PINS.

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (ACC)
		V-F	F-V	V-F	F-V	
DIMNet [36]	ICLR2019	84.12	84.03	39.75	-	83.2
PINs [38]	ECCV2018	84.00	-	31.00	-	78.5
SSNet [22]	DIC2019	78.00	78.50	30.00	30.05	78.8
β-VAE [27]	TMM2021	84.15	84.22	41.30	40.02	84.64
AML [23]	TMM2022	92.72	93.3	43.45	39.35	80.6
CMPC [51]	IJCAI2022	82.2	81.7	-	-	84.6
FOP [34]	ICASSP2022	89.3	83.5	-	-	83.5
SBNet [52]	ICASSP2023	82.4	82.4	-	-	82.5
DSANet [39]	TMM2023	95.25	94.28	46.83	43.36	78.0
ACIENet [25]	TIFS2024	96.4	95.6	46.9	44.1	84.8
AICANet	**Ours**	**98.5**	**98.2**	**47.5**	**47.8**	**85.6**

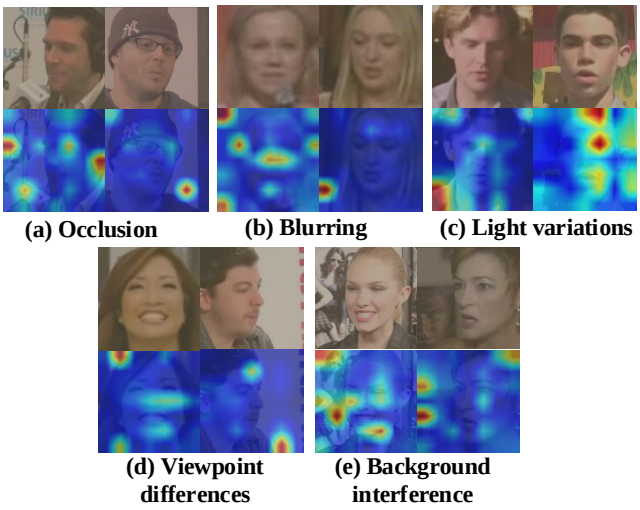


Fig. 6. Display of visualization features of some complex challenges on VoxCeleb [29]

AICANet to address cross-modal differences. While effective, AICANet still encounters challenges with complex conditions, including occlusion, blurring, lighting variations, changes in viewpoint, and background interference. As illustrated in Fig. 6, these factors occasionally prevent accurate foreground feature capture, contributing to audio-visual modal discrepancies. Thus, future work should enhance focus on foreground features and minimize interference, further reducing modal differences and achieving more robust matching.

VoxCeleb2 [30] provides a more comprehensive set of audio-visual samples, allowing for a better evaluation of our proposed method's generalizability. We conducted experiments on several classic methods with open-source code, as illustrated in Table IV. The experimental findings demonstrate state-of-the-art performance in both binary and multi-way matching tasks while achieving near-optimal results in verification tasks. These outcomes suggest that our approach effectively reduces modality discrepancies, thereby learning more generalized audio-visual matching capabilities. The multi-attribute supervised method, ACIENet, achieved optimal performance in verification tasks, which also highlighted the significance of exploring cross-modal latent feature correlations. We note that in previous methods, the audio-to-face image (V-F) matching task outperformed the face image-to-audio (F-V) matching task, which may stem from interference factors in the data. In contrast, our proposed method effectively reduces the differences between modalities and thus achieves comparable or better performance in multiple scenarios for both F-V and V-F tasks.

D. Ablation Study

1) Evaluation of Different Component Effectiveness: The AICANet method comprises three main components: the Adaptive Interactive Attention (AIA) module, the Adaptive Correction Attention (ACA) mechanism, and the Relative Distance Stretching Metric loss ($\mathcal{L}_{RDSM}$). To evaluate the effectiveness of each component, we conducted tests on two datasets, as shown in Table V. Note that Table V (a) presents the baseline approach, with the ACA mechanism applied as an insertion operation within the attention module. Comparative analysis of Tables V (a) and (b) reveals significantly improved matching performance on VoxCeleb, though with limited VoxCeleb2 generalization. We attribute this to overfitting from amplified inter-sample correlations. As Table V (e) demonstrates, effective audio-visual matching requires prior interference suppression for optimal feature space unification.

Similarly, the AIA module significantly enhances performance across most tasks, as demonstrated by comparisons between Table V (c) and (a), indicating that the AIA module learns latent connections between cross-modal features and mitigates modal differences. However, while the AIA module strengthens cross-modal associations, it may inadvertently introduce interfering feature effects. To address this, the ACA mechanism is proposed to filter these interfering features using the adaptive threshold, thereby reducing the modal differences caused by such interference. As shown in Table V (d) and (a), the combination of the ACA mechanism and the AIA module substantially enhances audio-visual matching performance. As

TABLE IV
THE QUALITATIVE RESULTS OF MATCHING TASKS ON VOXCELEB2 [30]. BINARY DENOTES THE 1:2 MATCHING WHILE MULTI-WAY DENOTES THE $2:K(K=10)$ MATCHING. '$\times$' INDICATES 'NOT CAPABLE' AND '-' INDICATES 'NO RESULTS'.

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (ACC)	
		V-F	F-V	V-F	F-V	V-F	F-V
SVHF-Net [37]	CVPR2018	68.7	67.9	$\times$	$\times$	–	–
DIMNet [36]	ICLR2019	68.5	69.0	–	–	–	–
AML [23]	TMM2021	80.2	81.4	41.2	40.7	80.6	78.4
DSANet [39]	TMM2023	82.9	83.6	42.3	41.2	78.8	77.5
P^2VANet [49]	TCSVT2024	87.3	85.2	<u>46.2</u>	<u>45.1</u>	84.9	82.1
ACIENet [25]	TIFS2024	<u>88.1</u>	<u>88.7</u>	45.6	44.3	**86.3**	**91.5**
AICANet	**Ours**	**90.4**	**91.6**	**46.2**	**46.5**	<u>85.4</u>	<u>90.3</u>

TABLE V
ABLATION EXPERIMENTS OF THE AICANET METHOD IN A VISUAL-AUDIO MATCHING TASK.

Component				VoxCeleb [29]						VoxCeleb2 [30]					
				Binary		Multi-way		Verification		Binary		Multi-way		Verification	
	AIA	ACA	$\mathcal{L}_{RDSM}$	V-F	F-V	V-F	F-V	V-F	F-V	V-F	F-V	V-F	F-V	V-F	F-V
(a)				94.8	89.8	48.5	45.6	88.2	91.2	87.7	87.1	42.2	41.5	83.8	84.0
(b)			✓	96.0	92.0	48.7	46.4	88.4	91.4	88.5	88.2	42.8	42.1	82.3	83.9
(c)	✓			96.6	96.9	48.8	47.6	88.7	92.4	89.1	90.5	42.5	45.4	83.1	89.6
(d)	✓	✓		97.2	97.2	49.3	48.1	90.1	93.1	90.0	91.4	45.6	46.1	85.2	89.9
(e)	✓	✓	✓	**97.6**	**97.3**	**49.8**	**48.3**	**90.5**	**93.3**	**90.4**	**91.6**	**46.2**	**46.5**	**85.4**	**90.3**

shown in Table V (d) vs. (c), the two components together perform better than the AIA module alone, which indicates that the removal of interfering features can also mitigate modal differences. Further, Table V (e) vs. (d) shows that the learned audio-visual feature representations enable more effective deep matching within a unified space, further improving matching accuracy. The ablation experiments confirm that the three components of AICANet effectively mitigate modal differences and enhance the robustness of audio-visual matching.

2) Evaluation of Adaptive Interactive Attention Module: The Adaptive Interaction Attention (AIA) module comprises three main components: Relevance-guided Cross-modal Interaction (RCI), Relevant loss ($\mathcal{L}_{Rel}$), and Dynamic Estimation Fusion (DEF). To assess the effectiveness of these components, we compare experimental results against baseline performance, as detailed in Table VI (a). The results indicate that RCI and DEF significantly improve model performance in matching and validation tasks, as shown in Table VI (b) vs. (a) and Table VI (d) vs. (b). This demonstrates that RCI effectively learns potential associations between audio-visual features, thereby mitigating audio-visual discrepancies. DEF integrates inter-modal correlation features with intra-modal augmentation, further enhancing cross-modal semantic matching. As indicated by the comparison between Table VI (c) and (b), the association loss causes a slight reduction in matching performance for the V-F task, which may be due to the use of face images as candidate matching samples, making it easier to distinguish positive from negative samples. Conversely, the F-V task uses audio as candidate matching samples, where relationships are less discernible, requiring perceptual loss to guide the association learning network. As accurate regression of correspondences is challenging, we introduced simulated annealing to gradually reduce the association loss impact, improving model robustness. Overall, a balanced integration of each component in the AIA module contributes positively to reducing audio-visual modality discrepancies.

TABLE VI
THE PROPOSED ADAPTIVE INTERACTIVE ATTENTION (AIA) MODULE PERFORMS ABLATION EXPERIMENTS EXECUTED IN BOTH VALIDATION AND BINARY MATCHING SCENARIOS ON VOXCELEB [29].

	AIA			Binary		Verification	
	RCI	$\mathcal{L}_{Rel}$	DEF	V-F	F-V	V-F	F-V
(a)				94.8	89.8	88.2	91.2
(b)	✓			96.3	91.6	88.5	91.8
(c)	✓	✓		95.5	93.8	86.6	92.1
(d)	✓	✓	✓	**96.6**	**96.9**	**88.7**	**92.4**

3) Evaluation of Self-support and Dynamic Estimation Fusion: To address modality differences, we propose the AIA module with an RCI mechanism. Single-stage cross-modal interaction can't fully capture feature associations due to audio-visual heterogeneity. Our two-stage approach first identifies sparse correlations and then refines them via an intra-modal self-supporting mechanism for comprehensive audio-visual feature exploration. Table VII demonstrates that the second-order self-support mechanism significantly improves audio-visual matching performance, validating the essential role of second-order interactions in RCI. Furthermore, inspired

TABLE VII
FURTHER ABLATION ON VOXCELEB [29] DATA FOR RCI AND DEF.

Method	AIA(RCI)		Binary		Verification	
	Without	With	V-F	F-V	V-F	F-V
Baseline			94.8	89.8	88.2	91.2
Self-support	✓		96.0	92.4	88.5	91.6
Self-support		✓	**96.6**	**96.9**	**88.7**	**92.4**
Method	AICANet		Binary		Verification	
	Without	With	V-F	F-V	V-F	F-V
DEF	✓		96.7	96.2	88.6	92.1
DEF		✓	**97.6**	**97.3**	**90.5**	**93.3**

TABLE VIII
THE ADAPTIVE THRESHOLD PARAMETER IN AICANET IMPACT OF THE SETTING OF ε ON AUDIO-VISUAL MATCHING IN DIFFERENT SCENARIOS ON VOXCELEB [29] IN THE V-F AND F-V TASKS.

Param(ε)	Binary		Multi-way		Verification	
	V-F	F-V	V-F	F-V	V-F	F-V
Baseline	94.8	89.8	48.5	45.6	88.2	91.2
0.01	96.4	95.4	49.4	48.2	88.3	**93.4**
0.05	97.3	97.1	49.7	47.0	88.0	**93.4**
0.1	96.1	**97.5**	49.5	46.9	87.5	93.2
0.2	96.3	97.0	**50.0**	47.3	89.2	93.1
Adaptive	**97.6**	97.3	49.8	**48.3**	**90.5**	93.3

by the skip connections, we designed the DEF that adaptively combines enhanced and original features via learned masks to reduce noise. Table VII demonstrates DEF's significant noise suppression advantage over conventional skip connections.

4) Evaluation of Adaptive Correction Attention Mechanism: The core of adaptive correction attention is an adaptive threshold that modulates the correlation matrix to avoid learning interference features. The adaptive threshold parameter (ε) in the adaptive correction attention mechanism is critical in controlling feature filtering and mitigating cross-modal variance caused by distracting information. We compare both empirical settings and model adaptations. As shown in Table VIII, the empirical setting of ε with values of [0.01, 0.05, 0.1, 0.2] has a significant effect on model performance, which suggests that an adaptive correction attention (ACA) mechanism can effectively mitigate interference. However, the performance of empirically set parameters varies significantly across tasks and scenarios. Therefore, we propose the adaptive correction attention mechanism to achieve optimal or near-optimal performance, further demonstrating the component's effectiveness.

While classical noise removal methods (shrinkage thresholding [47], principal component [53], frequency filtering [54], and disentanglement [27]) partially mitigate interference, their effectiveness remains limited due to: (1) challenges in optimal threshold/PCA ratio selection, and (2) inherent difficulty distinguishing subtle features from noise. Our proposed ACA mechanism overcomes these limitations through automated feature correlation filtering, eliminating manual parameter tuning. As Table IX demonstrates, ACA achieves superior and consistent noise suppression across diverse scenarios.

TABLE IX
COMPARISON OF DE-INTERFERENCE METHODS IN THE AUDIO-VISUAL MATCHING TASK ON VOXCELEB [29] DATA.

Methods	Binary		Verification	
	V-F	F-V	V-F	F-V
Shrinkage Thresholds [47]	95.9	96.3	89.6	92.8
Principal Component [53]	97.2	96.7	88.4	90.3
Frequency Filtering [54]	96.6	93.0	89.7	91.7
Disentanglement [27]	97.0	93.4	89.6	90.6
Adaptive Correction	**97.6**	**97.3**	**90.5**	**93.3**

TABLE X
COMPARISON OF DIFFERENT METRIC LOSSES IN THE AUDIO-VISUAL MATCHING TASK ON VOXCELEB [29] DATA.

Methods	Binary		Multi-way		Verification	
	V-F	F-V	V-F	F-V	V-F	F-V
$\mathcal{L}_{Trip}$ [55]	96.4	96.3	42.8	46.8	89.5	92.9
$\mathcal{L}_{LM}$ [56]	96.9	96.1	47.9	47.2	88.4	92.7
$\mathcal{L}_{SM}$ [23]	97.2	97.2	49.3	48.1	90.1	93.1
$\mathcal{L}_{RDSM}$	**97.6**	**97.3**	**49.8**	**48.3**	**90.5**	**93.3**

5) Evaluation of Different Metric Losses: Cross-modal feature metrics enhance identity matching and recognition. Traditional triplet loss [55] considers only inter-modal (anchor-positive) and intra-modal (anchor-positive/negative) distances, while lifted structure [56] and structure metric losses [23] incorporate broader relationships but neglect cross-sample distances. To overcome these limitations, we propose a relative distance stretching metric loss ($\mathcal{L}_{RDSM}$) that jointly optimizes inter-modal, intra-modal, and cross-sample distances, improving matching accuracy. As shown in Table X, single-sample constraints face challenges with multivariate variations (e.g., lighting, noise). However, they maintain robust cross-dataset generalization and multi-task adaptation, demonstrating effective modal alignment.

E. Hyper-parameters Analysis

Fig. 7 shows the hyperparameter weights for the various losses in Eq. (16), which are controlled by the variables α, β, and γ. These variables represent the weights for the matching metric loss, matching categorization loss, and relevance loss in the cross-modal matching task. The effects of these three hyperparameters on task performance fluctuate slightly between the V-F and F-V schemes, but both outperform current state-of-the-art methods. As λ increases, the performance of AICANet initially improves, reaching optimal levels at $\alpha = 2$, after which performance declines, indicating that an appropriate match metric enhances relationship learning. Performance peaks at $\alpha = 2$ for the V-F task and at $\alpha = 3$ for the F-V task. Regarding γ, setting it too low results in insufficient constraints on the Relevant Learning Network, while too high a value hampers accurate model estimation. Within the range [0.1, 1, 2, 3, 4], smaller values of γ benefit the V-F task, while larger values suit the F-V task.

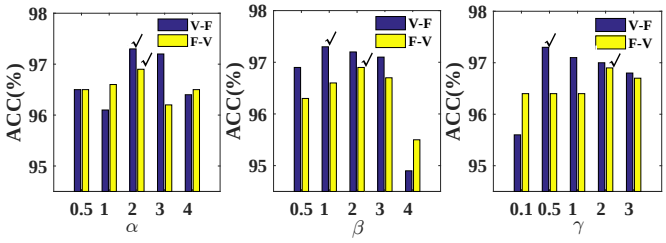


Fig. 7. The effects of hyperparameters of α, β, and γ on binary matching task on VoxCeleb [29].

V. Conclusion

This paper explores strategies to reduce audio-visual modal differences due to insufficient cross-modal correlation and noise interference and proposes a novel Adaptive Interactive and Correction Attention Network (AICANet). The main contributions of AICANet can be summarized in three points. To address modal differences due to insufficient correlation, we propose the Adaptive Interaction Attention (AIA) module, which employs relevance-guided cross-modal interactions and self-support mechanisms to work together to fully capture the potential correlations between audio-visual features. To mitigate modal differences caused by noise interference, we propose an adaptive Correction attention (ACA) mechanism that utilizes the adaptive threshold to adjust the correlation matrix, filtering out the influence of interference features. To enhance audio-visual deep matching, we design the Relative Distance Stretching Metric Loss ($\mathcal{L}_{RDSM}$), leveraging sample distance relationships to improve embedding representations of audio-visual features in a unified space. Experimental results indicate that AICANet achieves state-of-the-art performance across various scenarios. In future work, we will consider constructing a knowledge graph embedding model for audio-visual attributes to deepen the exploration of cross-modal feature associations.

References

[1] J. Wei, Y. Yang, X. Xu, X. Zhu, and H. T. Shen, "Universal weighting metric learning for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6534–6545, 2021.

[2] J. Yan, C. Deng, H. Huang, and W. Liu, "Causality-invariant interactive mining for cross-modal similarity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6216–6230, 2024.

[3] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin, "Cross-modal subspace learning via pairwise constraints," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5543–5556, 2015.

[4] Y. Yang, W. Hu, and H. Hu, "Unsupervised nir-vis face recognition via homogeneous-to-heterogeneous learning and residual-invariant enhancement," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2112–2126, 2024.

[5] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1578–1587, 2022.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

[7] W. Liu, Y. Wen, B. Raj, R. Singh, and A. Weller, "Sphereface revived: Unifying hyperspherical face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2458–2474, 2022.

[8] Z. Lei, S. Liao, R. He, M. Pietikainen, and S. Z. Li, "Gabor volume based local binary pattern for face representation and recognition," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1–6, 2008.

[9] J. Xin, Z. Wei, N. Wang, J. Li, and X. Gao, "Large pose face recognition via facial representation learning," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 934–946, 2024.

[10] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6398–6407, 2020.

[11] X. Zhou, Y. Zhong, Z. Cheng, F. Liang, and L. Ma, "Adaptive sparse pairwise loss for object re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19691–19701, 2023.

[12] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark," *Information Fusion*, vol. 100, p. 101901, 2023.

[13] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15495–15505, 2021.

[14] K. Yang, D. Marković, S. Krenn, V. Agrawal, and A. Richard, "Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8227–8237, 2022.

[15] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2020.

[16] S. Wang, Z. Zhang, G. Zhu, X. Zhang, Y. Zhou, and J. Huang, "Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 351–364, 2022.

[17] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023.

[18] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1696–1708, 2023.

[19] C. Xue, X. Zhong, M. Cai, H. Chen, and W. Wang, "Audio-visual event localization by learning spatial and semantic co-attention," *IEEE Transactions on Multimedia*, vol. 25, pp. 418–429, 2023.

[20] A. Greco, N. Petkov, A. Saggese, and M. Vento, "Aren: a deep learning approach for sound event recognition using a brain inspired representation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3610–3624, 2020.

[21] R. Wang, X. Liu, Y.-m. Cheung, K. Cheng, N. Wang, and W. Fan, "Learning discriminative joint embeddings for efficient face and voice association," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1881–1884, 2020.

[22] S. Nawaz, M. K. Janjua, I. Gallo, A. Mahmood, and A. Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *Proceedings of the Digital Image Computing: Techniques and Applications*, pp. 1–7, 2019.

[23] A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, and B. Luo, "Adversarial-metric learning for audio-visual cross-modal matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 338–351, 2021.

[24] K. Cheng, X. Liu, Y.-m. Cheung, R. Wang, X. Xu, and B. Zhong, "Hearing like seeing: Improving voice-face interactions and associations via adversarial deep semantic matching network," in *Proceedings of the ACM International Conference on Multimedia*, pp. 448–455, 2020.

[25] J. Wang, A. Zheng, Y. Yan, R. He, and J. Tang, "Attribute-guided cross-modal interaction and enhancement for audio-visual matching," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4986–4998, 2024. doi: 10.1109/TIFS.2024.3388949.

[26] Z. Yu, X. Liu, Y.-M. Cheung, M. Zhu, X. Xu, N. Wang, and T. Li, "Detach and enhance: Learning disentangled cross-modal latent representation for efficient face-voice association and matching," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 648–655, 2022.

[27] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 1763–1774, 2021.

[28] P. Wen, Q. Xu, Y. Jiang, Z. Yang, Y. He, and Q. Huang, "Seeking the shape of sound: An adaptive framework for learning voice-face association," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16347–16356, 2021.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Proceedings of the International Speech Communication Association*, pp. 2616–2620, 2017.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proceedings of the International Speech Communication Association*, pp. 1086–1090, 2018.

[31] Z. Pan, F. Wu, and B. Zhang, "Fine-grained image-text matching by cross-modal hard aligning network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19275–19284, 2023.

[32] K. Zhang, B. Hu, H. Zhang, Z. Li, and Z. Mao, "Enhanced semantic similarity learning framework for image-text matching," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[33] O.-B. Mercea, L. Riesch, A. Koepke, and Z. Akata, "Audio-visual generalised zero-shot learning with cross-modal attention and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10553–10563, 2022.

[34] M. S. Saeed, M. H. Khan, S. Nawaz, M. H. Yousaf, and A. Del Bue, "Fusion and orthogonal projection for improved face-voice association," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7057–7061, 2022.

[35] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7930–7938, 2020.

[36] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *Proceedings of the International Conference on Learning Representations*, 2019.

[37] 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*, pp. 8427–8436, 2018.

[38] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision*, pp. 71–88, 2018.

[39] J. Wang, C. Li, A. Zheng, J. Tang, and B. Luo, "Looking and hearing into details: Dual-enhanced siamese adversarial network for audio-visual matching," *IEEE Transactions on Multimedia*, vol. 25, pp. 7505–7516, 2023.

[40] 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*, pp. 770–778, 2016.

[41] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 317–326, 2016.

[42] A. Andonian, S. Chen, and R. Hamid, "Robust cross-modal representation learning with progressive self-distillation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16430–16441, 2022.

[43] Q. Jiang, C. Chen, H. Zhao, L. Chen, Q. Ping, S. D. Tran, Y. Xu, B. Zeng, and T. Chilimbi, "Understanding and constructing latent modality structures in multi-modal representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, 2023.

[44] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *Proceedings of the European Conference on Computer Vision*, pp. 701–719, 2022.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[46] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[47] K. Isogawa, T. Ida, T. Shiodera, and T. Takeguchi, "Deep shrinkage convolutional neural network for adaptive noise reduction," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 224–228, 2017.

[48] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 1, p. 6, 2018.

[49] A. Zheng, F. Yuan, H. Zhang, J. Wang, C. Tang, and C. Li, "Public-private attributes-based variational adversarial network for audio-visual cross-modal matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8698–8709, 2024.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[51] B. Zhu, K. Xu, C. Wang, Z. Qin, T. Sun, H. Wang, and Y. Peng, "Unsupervised voice-face representation learning by cross-modal prototype contrast," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3787–3794, 2022.

[52] M. S. Saeed, S. Nawaz, M. H. Khan, M. Z. Zaheer, K. Nandakumar, M. H. Yousaf, and A. Mahmood, "Single-branch network for multimodal training," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[53] J.-X. Wang, Z.-L. Sun, X. Chen, K.-M. Lam, and Z.-G. Zeng, "A csf-based cnr approach for small-size image sequences," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1808–1811, 2019.

[54] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5886–5895, 2023.

[55] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

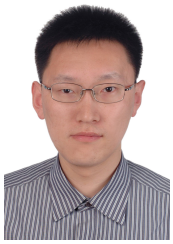
[56] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Jiaxiang Wang received the B.Eng. degree in automation from Anhui Institute of Information Technology, China, in 2016, and the M.Eng. degree in control engineering and the Ph.D. degree in computer science and technology from Anhui University, China, in 2020 and 2024, respectively. He is currently a lecturer in artificial intelligence at Anhui University of Science and Technology. His current research interests include vision-based artificial intelligence and pattern recognition. Especially on person/vehicle re-identification, audio-visual computing, and multimodal learning.

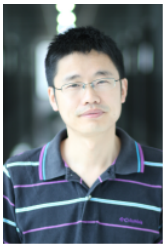
Aihua Zheng received B.Eng. degrees and finished Master-Docter combined program in Computer Science and Technology from Anhui University of China in 2006 and 2008, respectively. And received Ph.D. degree in computer science from University of Greenwich of UK in 2012. She visited University of Stirling and Texas State University during June to September in 2013 and during September 2019 to August 2020 respectively. She is currently a Professor and PhD supervisor at the School of Artificial Intelligence, Anhui University. Her main research interests include vision based artificial intelligence and pattern recognition. Especially on person/vehicle re-identification, audio-visual computing, and multimodal intelligence.

Lei Liu received the B.S. and Ph.D. degrees in the School of Computer Science and Technology from Anhui University, Hefei, China, in 2019 and 2024. He is currently working at the School of Artificial Intelligence, Anhui University of Science and Technology. His current research interests include computer vision and deep learning.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral Research Fellow with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. Dr. Li was the recipient of the ACM Hefei Doctoral Dissertation Award in 2016.



Ran He received the BE degree in computer science from the Dalian University of Technology, in 2001, the MS degree in computer science from the Dalian University of Technology, in 2004, and the PhD degree in pattern recognition and intelligent systems from CASIA, in 2009. Since September 2010, he joined NLPR, where he is currently a full professor. His research interests include information theoretic learning, pattern recognition, and computer vision. He serves as the editor board member of IEEE Transactions on Image Processing and Pattern Recognition, and serves on the program committee of several conferences. He is also a fellow of the IAPR.



Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in Computer Science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor and PhD supervisor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, pattern recognition, and machine learning.