Pattern Recognition 179 (2026) 113840

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Bidirectional intervention attention network for audio–visual matching

Jiaxiang Wang [a,b], Aihua Zheng [a,*], Dequan Li [b], Chenglong Li [a], Wenjuan Cheng [c], Ran He [d,e,1]

[a] School of Artificial Intelligence, Anhui University, Hefei 230601, China
[b] School of Artificial Intelligence, Anhui University of Science and Technology, Huainan 232001, China
[c] School of Computer and Information, Hefei University of Technology, Hefei 230009, China
[d] Center for Research on Intelligent Perception and Computing (CRIPAC), Beijing, 100089, China
[e] Institute of Automation, Chinese Academy of Sciences, Beijing, 100089, China

ARTICLE INFO

Keywords:
Audio–visual matching
Bidirectional Intervention Attention
Counterfactual inference
Curriculum Learning Strategy

ABSTRACT

The main challenge in audio–visual matching is accurately modeling the human ability to discern correlations between multimodal biosignals. Current audio–visual matching tasks are frequently affected by interfering data that can misdirect the model's attention, which leads to prediction bias. To address this challenge, we propose a novel Bidirectional Intervention Attention Network (BIANet), which performs bidirectional intervention by disentangling the correlation matrix to reduce bias. This bidirectional intervention is divided into a Positive Intervention Attention (PIA) module and a Negative Intervention Attention (NIA) module, establishing a counterfactual inference framework to improve the robustness of audio–visual matching. Specifically, the PIA module uses strong intra-modal correlations to aggregate identity-related features and reduce the impact of bias. However, interfering features within the data may also be aggregated, creating a bottleneck in model performance. Therefore, the NIA module enhances the characterization of these interfering features, enabling reverse bias elimination. Furthermore, we apply a Curriculum Learning Strategy (CLS) to incrementally reduce bias by progressively introducing reverse prediction results during training. Experiments on publicly available audio–visual datasets demonstrate that BIANet outperforms existing state-of-the-art algorithms. The code is released at https://github.com/w1018979952/BIANet.

1. Introduction

Audio–visual matching seeks to equip machines with cognitive "hearing" and "visual" capabilities, enabling the interpretation of multimodal biological signals and mimicking the human ability to perceive inter-modal associations. This technique holds particular value in fields such as criminal investigation, intelligent surveillance, and automated voice analysis. Guided by maximum likelihood principles, previous research has aimed to improve performance through cross-modal attention interactions [1,2], disentangled representation learning [3,4], and base adversarial metric [5–7]. Despite significant progress, the confounding interference factors in the data have often resulted in prediction bias.

Videos captured in real-world environments are inevitably influenced by surrounding factors, often resulting in prediction bias. For instance, as illustrated in Fig. 1, facial features in an interview may be affected by lighting conditions, viewing angles, hairstyles, and image clarity. Similarly, audio quality can be compromised by noise interference, variations in speech content, or overlapping voices from multiple speakers. Research indicates that these distractors can substantially reduce the accuracy of model prediction [8,9]. Unlike other multimodal collaborative prediction tasks [10,11], audio–visual matching models treat each sample as a unique recognition unit. Consequently, these distractors induce attention bias, negatively impacting overall prediction outcomes.

To address the problem of prediction bias, previous studies have focused mainly on two approaches. The first approach employs feature disentanglement [3,4], where features are divided into task-relevant and irrelevant categories using disentanglement architectures. This method aims to eliminate irrelevant interference features and reduce prediction bias. However, in practice, task-relevant and irrelevant features are often intertwined, making them difficult to separate and limiting the model's ability to mitigate bias effectively. Meanwhile, Wen et al. [8] propose a dynamic weighting strategy that filters face

* Corresponding author.
E-mail address: ahzheng214@foxmail.com (A. Zheng).
[1] *Fellow, IEEE.*

https://doi.org/10.1016/j.patcog.2026.113840
Received 24 March 2025; Received in revised form 14 January 2026; Accepted 22 April 2026
Available online 25 April 2026

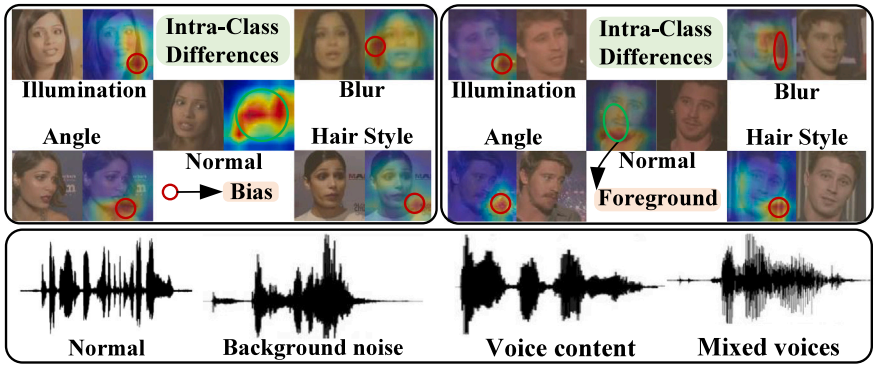

Fig. 1. Audio–visual matching aims to establish correlations between facial and audio biometric features. However, the presence of interference factors in the data often leads to overfitting and feature bias in the baseline, thereby limiting their overall performance. In Fig. 1, facial images of the same identity exhibit substantial intra-class differences caused by multiple interference factors, while the corresponding audio clips are similarly affected by disturbances in complex environments.

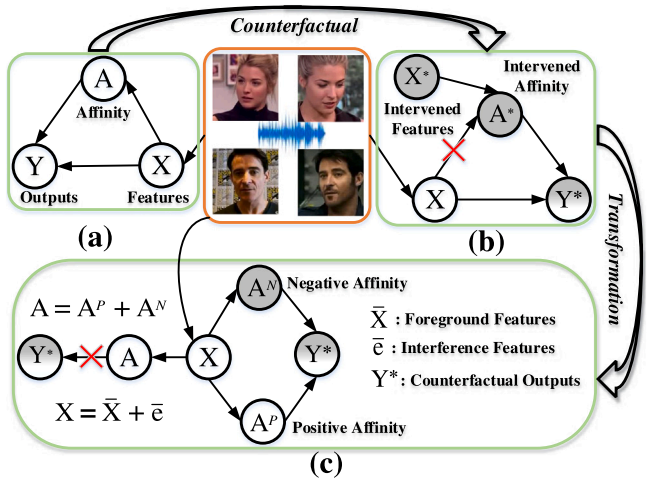

Fig. 2. (a) Example of a causal graph. (b) Example of counterfactual inference. (c) Bidirectional intervention-based counterfactual inference in this paper.

and speech samples to minimize sample-level interference. Nevertheless, this approach may insufficiently explore challenging yet informative features, thereby restricting cross-modal feature association. The second approach is based on counterfactual intervention [12,13], which enhances the model's robustness to disturbances and mitigates prediction bias by introducing controlled perturbations. However, due to the diverse nature of real-world disturbances, the learned virtual intervention features may not accurately eliminate bias effects. Therefore, removing bias remains a challenging problem.

Current audio–visual matching research on bias removal methods generally involves the following steps: (1) extracting intrinsic features from audio and facial image data; (2) exploring different feature branches to implement a disentanglement technique to learn relevant representations for cross-modal matching; and (3) developing effective inference strategies to integrate cross-modal features and predict matching relationships during inference. While existing methods rely on complex disentanglement model architectures [3,4] and advanced inference strategies [14], achieving significant improvements, extracted features still produce bias, creating bottlenecks in prediction performance. Disentanglement methods primarily address bias at the feature level, whereas inference strategies tackle bias effects directly during inference. Although both approaches have distinct advantages, few studies have integrated them to create a comprehensive adaptive paradigm for bias elimination.

To address interference factors in audio–visual data, we propose a bidirectional intervention attention network that combines disentanglement operations with inference strategies to eliminate bias effectively. Unlike traditional feature-detangling methods, our approach applies disentanglement to intramodal affine attention (denoted A), as illustrated in Fig. 2(a). The affine operation learns inherent feature relationships, allowing feature aggregation to focus on discriminative representations of identity [15]. Additionally, as shown in Fig. 2(b), the vanilla counterfactual intervention feature transformation method uses maximum likelihood estimation combined with total indirect effects (TIE) for causal inference. This approach enhances the reliability of model inference by calculating the difference in predicted outcomes between the original feature (X) and the intervention feature (X^*) transformed with intervened affinity (A^*) [13]. We extend this approach, as depicted in Fig. 2(c), by performing a disentanglement operation on the affine transformation. This process decomposes the information into two pathways, establishing a counterfactual inference framework. The core mechanism involves categorizing sample features into identity-relevant and irrelevant features to support counterfactual inference, thereby mitigating the influence of bias on predictions. Given the coupling between the interference components and the identity representations, positive correlations are introduced after the decomposition of the affine matrix to enhance the discrimination of identity characteristics and mitigate interference effects. However, due to this coupling, certain interference features inevitably persist, thereby weakening audio–visual matching performance. To address this issue, the negative intervention operation is introduced for counter-directional prediction, further reducing bias induced by interference. This innovative framework transcends traditional methods focused solely on statistical or spurious bias elimination, advancing the development of genuine causal effect analysis.

As a comprehensive adaptive bias removal framework, the proposed paradigm transforms the traditional global debiasing task into identifying and addressing bias in specific samples, driving audio–visual matching models toward unbiased predictions. To achieve this, we establish a linkage between feature disentanglement and counterfactual inference to enable mutual facilitation in learning, thereby accomplishing precise on-sample debiasing. Specifically, we introduce a novel Bidirectional Intervention Attention Network (BIANet), comprising a Positive Intervention Attention (PIA) module and a Negative Intervention Attention (NIA) module. This dual-pathway framework constructs a counterfactual inference mechanism that enhances the debiased prediction capability of audio–visual matching models. The PIA module aggregates identity-relevant features by leveraging strong intra-modal correlations. This enhances the discriminative power of the identity features, which mitigates model bias. In contrast, weak intra-modal correlations stem from interference features associated with

uncertain identities. The feature representations aggregated from these weak correlations are subsequently used for reverse prediction. This process synergizes with the forward intervention mechanism, collectively suppressing interference and effectively reducing prediction bias. During training, identity-relevant and irrelevant features often become entangled. Direct intervention risks impairing the learning of robust identity features. Because neural networks typically learn effective features before overfitting on noise, we employ a Curriculum Learning Strategy (CLS) to progressively integrate reverse prediction. This ensures the effective acquisition of robust features early in training, which is vital for achieving superior audio–visual matching performance.

The main contributions of this paper can be summarized as follows:

- We propose the Positive Intervention Attention (PIA) module, which mitigates model prediction bias by aggregating identity-related features.
- We propose the Negative Intervention Attention (NIA) module, which leverages identity-irrelevant features for back-prediction to address residual prediction bias.
- We design a Curriculum Learning Strategy (CLS), which incrementally integrates back-prediction results to maximize the debiasing effect.
- Extensive experiments and ablation tests conducted on the VoxCeleb1 [16] and VoxCeleb2 [17] datasets across various scenarios demonstrate that the complementarity and effectiveness of the BIANet components significantly improve model performance.

The paper is structured as follows. Section 2 reviews related work. Section 3 introduces our bidirectional intervention framework for counterfactual reasoning. Section 4 details the audio–visual matching task, the proposed network architecture, and its loss function. Section 5 presents experimental setups, compares our method with SOTA models, and provides ablation studies and visualizations. Finally, Section 6 summarizes the paper's contributions.

2. Related works

Audio–visual Matching. Audio–visual matching is a concept originating from psychological research. Deep learning-based dual-stream networks are capable of achieving performance beyond human capabilities in probabilistic recognition. This breakthrough is opening up new avenues for research in this field [18]. Subsequently, a series of metric-based methodologies [19] are emerging, which are trying to explore the interaction between audio and visual signals to advance the field. However, the modal heterogeneity within audio–visual data remains a significant obstacle to research development in this field. To address this problem, Wang et al. [20] and Nawaz et al. [21] respectively use shared feature embedding modules to mitigate the differences in feature distributions caused by modality heterogeneity.

Afterward, researchers further optimize feature distributions through distance metrics to learn global feature correlations across modalities. However, this metric is reliant on pre-labeled labels for supervised learning, which adjust the distribution based on the correlation between data pairs. Consequently, it fails to address the overall distribution problem across modalities directly. To address this limitation, the generative adversarial network (GAN)-based model [22,23] can eliminate modal heterogeneity at the global distribution level, which effectively reduces the disparity between face and audio features. Therefore, Zheng et al. [5] propose an audio–visual matching model that integrates adversarial learning with metric learning. The method utilizes generative adversarial networks to achieve a Nash equilibrium for obtaining modality-independent feature representations, facilitating effective audio–visual matching through metric learning. Additionally, Cheng et al. [6] propose a model achieving modal equilibrium through generative adversarial training. The method integrates triple loss and modal center loss to bolster the robustness of the audio–visual matching network. Wang et al. [7] propose the dual-enhanced siamese adversarial network model that emphasizes mining local details of audio and face features using fine-grained information, thereby significantly enhancing audio–visual matching performance. Wang et al. [2] and Zheng et al. [24] propose multi-attribute approaches to explore potential semantic features for robust audio–visual matches in adversarial learning, respectively.

Disentangled Representation Learning. This study utilizes audio–visual data sourced from real-world interview scenarios, which inherently have interference characteristics. Such interference can lead the model to prioritize irrelevant background features, thereby introducing bias. However, disentangled representation learning, a machine learning technique, enables models to identify and isolate biases in data, thereby improving their capacity to capture complex features. To achieve this, Higgins et al. [25] introduce the hyperparameter β in variational autoencoders (VAEs) to control the discretization of latent representations, allowing customization of these representations based on specific needs. Building on this, Ning et al. [3] propose the disentangled latent variable controlled by β to separate identity-related features and filter out modality-specific features for cross-modal association. However, β lacks adaptability, making it challenging to apply across diverse tasks and datasets. To address this limitation, Kim et al. [26] introduce forced disentanglement, which enforces independent and discrete distributions of latent variables by imposing constraints on the latent space. By expanding on this approach, Yu et al. [4] construct the cross-modal latent representation framework for disentanglement to examine the relationship between face and speech by removing independent features. Recognizing the inherent attribute correlations in audio–visual data, Wen et al. [19] propose the disjoint mapping network for audio–visual matching, employing an attribute-oriented mechanism to prevent feature overfitting. Despite the progress in disentanglement methods, separating bias caused by interference remains a significant challenge. To address this, Wen et al. [8] introduce a dynamic weighting strategy to mitigate bias effects at the sample level, enhancing model evaluation robustness. However, this approach can hinder the model's learning of challenging samples, which leads to inferior model generalization performance. We propose the BIANet method to mitigate bias in audio–visual matching. The network disentangles the correlation matrix to make forward predictions for identity features and reverse predictions for non-identity features. These predictions are integrated within a counterfactual inference framework, effectively mitigating the influence of bias in audio–visual matching tasks.

Causal Inference. Causal inference has become a pivotal tool in computer vision, enhancing the robustness of deep learning models by effectively addressing biases. Its applications span various tasks, including visual question answering [27], Re-identification [13], and multimodal sentiment analysis [12]. Mainstream causal inference research comprises intervention-based methods [28] and counterfactual approaches [29]. Intervention modifies the original feature distribution to uncover causal effects, while counterfactuals represent the outcomes of manipulated variables under different conditions [30].

As opposed to tackling spurious associations, we generate the counterfactual input through intervention to mitigate bias introduced by confounding factors in the data. While recent studies have addressed problems at the dataset level, they neglect bias at the sample level, which is crucial for accurately interpreting information within a single modality. Our framework integrates disentanglement operations with counterfactual inference to jointly eliminate sample-level biases, representing a significant step toward achieving unbiased predictions.

3. Bidirectional intervention-based counterfactual inference

In this section, we explore the application of causal reasoning to audio–visual matching tasks. First, we will explain in detail the core concepts of causal reasoning in the literature [31]. Then, we

analyze the reasons for the poor generalization performance of current methods.

Preliminary: Causal graphs reflect the causal relationships between variables, which are constructed via X, Y, and A as shown in Fig. 2(a). When variable X directly influences variable Y, Y is termed a subterm of X, indicated as $X \rightarrow Y$. If X indirectly affects Y through variable A, then A acts as a mediator between X and Y denoted as $X \rightarrow A \rightarrow Y$.

Counterfactual notation formula representation: when X is set to be x and A is set to be a, the resulting value of Y is denoted as:

$$Y_{x,a} = Y(X = x, A = a), \tag{1}$$

where $Y_{x,a}$ represents the result of reasoning obtained with X set to x, which A is a specific value when $X = x^{\star}$, and the result of counterfactual reasoning is denoted as $Y_{x,A_{x^{\star}}}$. Fig. 2(b) illustrates the counterfactual notation and its application process.

In causality research, the causal effect represents the change in an outcome variable resulting from the reference variable assuming a specific value. This effect, commonly referred to as the total effect (TE), can be exemplified by the impact of an experimental variable $X = x$ on the outcome Y. The total effect can be expressed mathematically as follows:

$$TE = Y_{x,A_x} - Y_{x^{\star},A_{x^{\star}}}. \tag{2}$$

The TE equals the natural direct effect (NDE) plus the total indirect effect (TIE). The NDE shows the direct effect of X on Y if mediator A is blocked, with A set to the value obtained at $X = x^{\star}$, i.e., the response of A to the treatment $X = x$ is disabled. Therefore, the NDE is denoted by the increase of Y as X changes from $x^{\star}$ to x. The NDE is defined as follows:

$$NDE = Y_{x,A_{x^{\star}}} - Y_{x^{\star},A_{x^{\star}}}. \tag{3}$$

The TIE can be computed from Eqs. (2) and (3) as:

$$TIE = TE - NDE = Y_{x,A_x} - Y_{x,A_{x^{\star}}}. \tag{4}$$

Bidirectional Intervention-based Counterfactual Inference: In practice, accurately determining the value of the intervention feature $x^{\star}$ is a significant challenge. The CIFT [13] method generates intervention feature samples through Gaussian sampling, but this approach may inadvertently introduce foreground feature interventions, thereby compromising the effectiveness of the intervention. In contrast, the CLUE [12] method learns shared virtual intervention features across samples by training the network. However, this limits its ability to intervene precisely on specific samples. Typically, noise and background information, treated as interference terms in prediction, can serve as intervention features to enhance the model's generalization performance. Therefore, we design a disentanglement threshold layer for feature correlation matrix decomposition to obtain an accurate intervention feature representation. The specific calculation of the intervention features is outlined as follows:

$$x^{\star} = \overline{e} A^{N}, \tag{5}$$

$$A^{N} = \overline{e}\overline{e}^{T}, \tag{6}$$

where $\overline{e}$ denotes interference features, but it is not easy to disentangle them from the X features. At the same time, the foreground feature ($\overline{X}$) and the interference feature ($\overline{e}$) should remain independent (i.e., ($\mathbf{0} = \overline{X}\overline{e}^{t} = \overline{e}\overline{X}^{t}$). A^{N} is the negative affinity, which enhances the interference features through the self-attention correlation matrix. The affinity transformation A of the X-feature can then be denoted as:

$$\begin{aligned} A &= XX^{T} = (\overline{X} + \overline{e})(\overline{X} + \overline{e})^{T} \\ &= \overline{X}\overline{X}^{T} + \overline{e}\overline{e}^{T} = A^{P} + A^{N}, \end{aligned} \tag{7}$$

where A^{P} represents the positive affinity transformation of foreground features, which enhances the key area feature representations. The network is prone to recognizing foreground features with high certainty and is shown to be gradual from shallow to deep in feature learning [32]. For this reason, the strong correlations tend to correspond to foreground features, while the weak correlations are more related to interference features with high uncertainty. Therefore, the disentanglement threshold is applied to the correlation matrix to disentangle diverse feature representations more efficiently. The total indirect effect (TIE) can be calculated as:

$$TIE = Y_{x,A^{P}_{\overline{x}}} - Y_{x,A^{N}_{x^{\star}}}. \tag{8}$$

We propose a bidirectional intervention framework integrating positive and negative intervention attention modules to enhance the model's debiasing capability. This framework reduces the interdependence between foreground features ($\overline{X}$) and intervention features ($\overline{e}$), thereby improving the accuracy of audio–visual matching.

4. Methodology

Interference in real-world data can introduce bias into cognitive models for matching audio and facial images across biometric information. To address this issue, we propose the BIANet method (illustrated in Fig. 3), which leverages correlation matrix disentanglement and counterfactual inference to eliminate prediction bias. In this model, features extracted from audio and facial images are processed into query features (Q), key features (K), and value features (V) through convolutional operations. Using a learned disentanglement threshold, the network decomposes the correlation matrix, defined as the covariance matrix between query features (Q) and key features (K). This matrix is multiplied by value features (V) via a softmax operation to produce disentangled features. These features are further transformed for counterfactual inference, and the resulting outputs are used to confirm the matching relationships.

4.1. Overview

The objective of the audio–visual matching task is to utilize audio clips to recognize a gallery of face images depicting multiple candidates and conduct identity matching. This scenario is denoted as V-F, while the reverse case is termed F-V. For identification, a baseline audio clip a_{i_0} is chosen, and a gallery of matched images, consisting of k face images $\{v_{i_1}, v_{i_2}, \ldots, v_{i_k}\}$, indexed by i, is utilized. ResNet [33] is employed for feature extraction. Audio clips are denoted as X_i^{a}, and visual face images are represented as $X_i^{v} = \{X_{i_1}^{v}, \ldots, X_{i_k}^{v}\}$. The variable k defines various matching scenarios, where $k > 1$ signifies a general matching case, and $k = 1$ denotes a special matching scenario, also considered as a validation scenario.

4.2. Positive intervention attention module

The attention mechanism is crucial in aggregating identity-related features in the audio–visual matching task. However, real-world audio–visual data often contains complex backgrounds and significant noise. These elements can amplify interference when the attention mechanism focuses on identity features, thereby weakening the model's generalization performance. Referring to Eq. (7), we compute the intramodal correlation matrix as an affine transformation, represented as follows:

$$A_{i_k}^{m} = X_{i_k}^{m} X_{i_k}^{m\,T} = (\overline{X}_{i_k}^{m} + \overline{e}_{i_k}^{m})(\overline{X}_{i_k}^{m} + \overline{e}_{i_k}^{m})^{T}, \tag{9}$$

where k denotes the index of the sample, with $k = 0$ corresponding to audio samples and $k > 0$ corresponding to image samples. The $m \in \{a, v\}$ refers to the audio or visual modality. $\overline{X}_{i_k}^{m}$ represents the undisturbed face image (audio) features, while $\overline{e}_{i_k}^{m}$ corresponds to the disturbed face image (audio) features. The matrices $A_{i_k}^{m}$ are the feature correlation matrices for face images (audio) and are designed to aggregate identity-related features while simultaneously enhancing interfering features. Due to the complex mixture of interference features, directly extracting interference-free identity features is challenging.

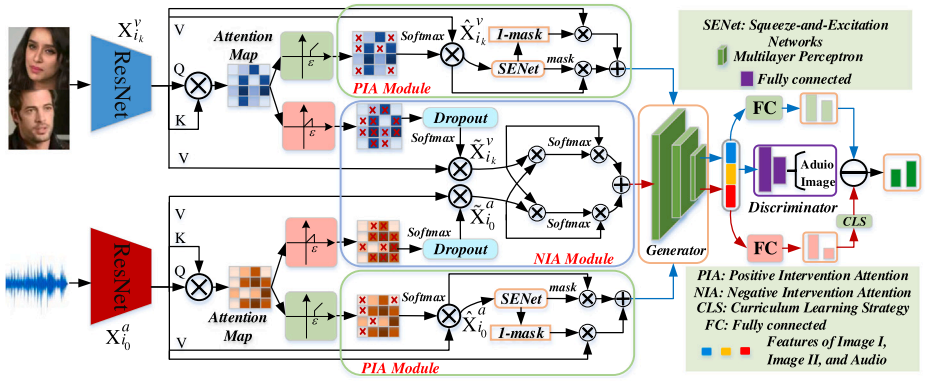

Fig. 3. Overview of the Bidirectional Intervention Attention Network (BIANet). The BIANet comprises Positive Intervention Attention (PIA) and Negative Intervention Attention (NIA). These modules generate outputs that facilitate counterfactual inference operations to eliminate bias. The PIA module mitigates bias by leveraging strong intra-modal correlations to aggregate identity-related features. Conversely, the NIA module efficiently transfers identity-irrelevant feature representations for reverse prediction, reducing prediction bias. Additionally, we introduce a Curriculum Learning Strategy (CLS) to incorporate reverse prediction incrementally results incrementally, achieving unbiased prediction audio–visual matching.

The model tends to prioritize deterministic foreground features over uncertain interfering features, resulting in a higher feature correlation to focus on the foreground features. Therefore, we can estimate the disentanglement threshold ε from the network to focus on different hierarchical features. The disentangled correlation matrix can be expressed as:

$$\hat{A}^m_{i_k} = clamp(A^m_{i_k}, \varepsilon), \tag{10}$$

$$\tilde{A}^m_{i_k} = Relu(-(Relu(A^m_{i_k}) - \varepsilon)), \tag{11}$$

where $Relu$ denotes the ReLU activation function, and $clamp$ signifies the truncation operation. The matrix $\hat{A}^m_{i_k}$ represents the attention matrix capturing strong correlations among image face (audio) foreground features, while $\tilde{A}^m_{i_k}$ denotes the attention matrix for spurious correlations among image face (audio) interference features. By applying $\hat{A}^m_{i_k}$, we can effectively aggregate identity-related features in both audio and face images. Conversely, using $\tilde{A}^m_{i_k}$ amplifies the interference features in both audio and face images.

$\hat{A}^m_{i_k}$ is determined by a disentanglement threshold that focuses on the correlation of foreground features. This approach helps to aggregate these features, thereby enhancing the performance of audio–visual matching. The precise computation of the aggregated features is as follows:

$$\hat{X}^m_{i_k} = softmax(\tau^m_1 \hat{A}^m_{i_k}) X^m_{i_k}, \tag{12}$$

$$\hat{A}^m_{i_k} = \overline{X}^m_{i_k}(\overline{X}^m_{i_k})^T + \overline{e}^m_{i_k}(\overline{X}^m_{i_k})^T = (A^m_{i_k})^P + \overline{e}^m_{i_k}(\overline{X}^m_{i_k})^T, \tag{13}$$

where τ^m_1 is set to 10, based on experimental results, to act as a temperature control mechanism that enhances correlation differences and emphasizes salient features. Ideally, according to Eq. (7), $\hat{A}i_k{}^m$ should convey valuable foreground information solely through the affine transformation $(A^m_{i_k})^P$. However, in practice, potential spurious correlations (denoted as $\overline{e}^m_{i_k}(\overline{X}^m_{i_k})^T$) may arise between the foreground and interfering information. These correlations propagate into the aggregated identity features, $\hat{X}i_k{}^m$, ultimately reducing the accuracy of the model's predictions.

To achieve stable optimization, we incorporate a skip connection design with learnable mask weights, which improves the model's ability to deliver features consistently. These weights dynamically regulate the fusion of input features ($X^m_{i_k}$) with identity features ($\hat{X}^m_{i_k}$), enabling effective audio–visual feature delivery. This mechanism facilitates direct input-to-output mapping, mitigating challenges such as gradient vanishing and explosion during backpropagation [33]. Additionally, the mask weights function as filters, reducing the transmission of irrelevant or interfering features. The fusion feature representation is formulated as follows:

$$\overline{X}^m_{i_k} = (1 - mask^m_{i_k}) X^m_{i_k} + mask^m_{i_k} \odot \ IN(\hat{X}^m_{i_k}), \tag{14}$$

where $mask^m_{i_k}$ denotes the feature weight within unimodal, calculated via SENet [34] and acquired through instance normalization (IN) to diminish the variance of modal features.

4.3. Negative intervention attention module

Interference features can adjust the original feature distribution to reveal causal effects, enabling counterfactual inference and generating results for different operational variables. Consequently, generating relevant intervening features is a critical step for achieving accurate counterfactual inference. Currently, aggregating identity features using attention mechanisms helps mitigate interferences, thereby reducing the model's positive prediction bias. However, as previously mentioned, such aggregation may inadvertently introduce spurious associations. To address this issue, weakly correlated $\tilde{A}^m_{i_k}$ features can be filtered out by applying an untangling threshold, which emphasizes interference feature representations as intervention features for causal inference. This inference process is constructed through the complementary integration of forward and backward predictions, further reducing prediction bias via enhanced counterfactual inference. Among these, the augmented interference feature $\tilde{X}^m_{i_k}$ is calculated as follows:

$$\tilde{X}^m_{i_k} = softmax(\tilde{A}^m_{i_k} + d^m_{i_k}) X^m_{i_k}, \tag{15}$$

$$\tilde{A}^m_{i_k} = \overline{e}^m_{i_k}(\overline{e}^m_{i_k})^T + \overline{X}^m_{i_k}(\overline{e}^m)^T = (A^m_{i_k})^N + \overline{X}^m_{i_k}(\overline{e}^m_{i_k})^T, \tag{16}$$

the affine transformation of the interference features is denoted as $(A^m_{i_k})^N$, according to Eq. (6). However, in practice, the affine transformation $\tilde{A}^m_{i_k}$ incorporates not only $(A^m_{i_k})^N$ but also spurious correlations between foreground and interference features. This leads to simultaneous enhancement of the interference features and discriminative foreground features, resulting in feature entanglement. To mitigate this issue, we introduce a feature dropout operation. Specifically, $d^m_{i_k}$ represents randomized Bernoulli distribution mask matrices for audio and image data, respectively. These matrices assign specific elements of the similarity matrix a value of $-\infty$, thereby enhancing the regularization of the attention weights. This approach reduces the model's tendency to overfit foreground features, improving the representation of interference features. The mask $d^m_{i_k}$ is computed as follows:

$$d^m_{i_k} = \begin{cases} 0 & 1 - rand(p) = 1 \\ -\infty & rand(p) = 0 \end{cases}, \tag{17}$$

where *rand* represents a random process, while p signifies the probability magnitude of the random percentage.

Interference features can be employed for back-prediction in counterfactual inference to achieve interference-free audio–visual associations. However, directly using these enhanced interference features does not effectively strengthen correlations between audio–visual features. This limitation arises from uncertain spurious correlations within cross-modal features, which complicates the accurate application of back-prediction to enhance the model's debiasing capabilities. To overcome this challenge, it is crucial to identify false correlations within the interference features and amplify the relevant components for inverse prediction. Therefore, we treat the interference feature as an intervention feature in causal inference, as described below:

$$\tilde{A}^m_{i_k} = \tilde{X}^m_{i_k}(\tilde{X}^m_{i_k})^T, \tag{18}$$

$$\overline{X}^m_{i_k} = softmax(\tau^m_2 \tilde{A}^m_{i_k})\tilde{X}^m_{i_k} + softmax(-\tau^m_2 \tilde{A}^m_{i_k})\tilde{X}^m_{i_k}, \tag{19}$$

where τ^m_2 serves as the temperature control parameter, defined as in the previous section. $\tilde{A}^m_{i_{0k}}$ represents the attention weights for audio or facial image interference features. Due to the uncertain matching relationships among candidate objects, these features can guide both positive and negative attention toward cross-modal features, enabling a bidirectional attention mechanism. $\overline{X}^m_{i_k}$ denotes spuriously associated interference features from face images (or audio), which assist the model in generating backward predictions. These predictions facilitate counterfactual inference, effectively reducing model bias.

To address the complexity of environmental interference, we propose a positive intervention attention module that aggregates identity features using a strongly correlated attention mechanism. Additionally, we propose a negative intervention attention module that emphasizes interference features through weakly correlated attention mechanisms for reverse prediction matching. However, these attention mechanisms may inadvertently associate foreground features with interference features, resulting in model prediction bias. To address this, we incorporate the features from both intervention modules into a counterfactual inference architecture to eliminate prediction bias. This integration ensures the independence of foreground features from interference features, thereby enhancing the generalization capability of the audio–visual matching model.

4.4. Counterfactual inference in audio–visual matching

Following the prior discourse, we employ an adversarial training method to mitigate discrepancies among various modal features [2]. Hence, we adhere to this approach, merging audio features $\overline{X}^a_{i_0}$ with multiple face image features $\{\overline{X}^v_{i_1}, \ldots, \overline{X}^v_{i_k}\}$ to produce shared features across modalities $\{\hat{h}_{i_0}, \ldots, \hat{h}_{i_k}\} \in \mathcal{H}$. This adversarial process has followed previous studies [2] to keep consistency. Subsequently, the generated cross-modality independent features require a specific matching metric to define the matching relationships clearly. This metric is based on the general pair weighting approach. Finally, we leverage the nonlinear properties of the multilayer perceptron (MLP) to assess the probability of a correct match, depicted as follows:

$$\mathcal{P}_{cp} = softmax(C_m([exp(\hat{h}_{i_0} - \hat{h}_{i_1}), \cdots, exp(\hat{h}_{i_0} - \hat{h}_{i_k})])), \tag{20}$$

where C_m and $\mathcal{P}_{cp}$ denote match classification and positive match probability, respectively. exp represents the natural logarithm with base e.

Using the generator's embedding structure, we embed the interference features ($\overline{X}^m_{i_k}$) as shared modal features, denoted as $\{\tilde{h}_{i_0}, \ldots, \tilde{h}_{i_k}\}$. These features are then used to compute the probability of reverse prediction matching, as follows:

$$\mathcal{P}_{cn} = softmax(C_m([exp(\tilde{h}_{i_0} - \tilde{h}_{i_1}), \cdots, exp(\tilde{h}_{i_0} - \tilde{h}_{i_k})])), \tag{21}$$

where $\mathcal{P}_{cn}$ denotes reverse match probability.

Curriculum Learning Strategy: We employ a classification network to predict the matching probability based on the features provided by the bidirectional intervention attention mechanism. Referring to Eq. (8), the positive and negative prediction results are used for counterfactual inference to achieve debiased predictions. The cross-entropy function [35] is used to compute the loss for counterfactual inference, as follows:

$$\mathcal{L}_{Cls} = -\frac{1}{M}\sum_{i=1}^{M}(Y_i \log(\mathcal{P}_{cp} - (1 - \cos(\frac{epoch}{N}\pi))\mathcal{P}_{cn})), \tag{22}$$

where $epoch$ denotes the current iteration and N represents the total number of iterations. Y_{i_j} ($Y_{i_j} \in [1, k]$) is the identity label for the match. We employ a curriculum learning strategy that incrementally introduces counterfactual inference to prevent the model from overemphasizing counterfactual inference in the early stages, thereby preserving valid features. This strategy is implemented using a cosine function weighting scheme [36], which progressively integrates backward prediction results, enabling the model to perform counterfactual inference and achieve unbiased predictions. Notably, our model utilizes the same counterfactual inference architecture for matching prediction during both the training and testing phases.

5. Experiments

5.1. Implementation details

(1) Network architecture. All experiments are performed on NVIDIA GeForce RTX 3090 graphics cards. The ResNet-18 [33] architecture is employed for facial image feature extraction, while the SE-ResNet-34 [33] architecture is utilized for audio clips processing. The selected architectures are consistent with established previous research [2], ensuring fair algorithm performance evaluation and comparable results.

The image feature extractor utilized ImageNet [37] pre-trained parameters, while the audio feature extractor is not pre-trained. Facial images are compressed into $3 \times 224 \times 224$ matrices for input. The processing flow for the input audio clips is as follows: the *librosa* library is first employed to read the audio data, which is then converted to decibel (dB) values using the *amplitude_to_db* function to align with human auditory perception. The data is subsequently normalized to the range (0, 1) to ensure computational stability and enable effective comparison. The final feature sequences consisted of 160000 dimensions. Shorter audio samples are loop-padded to ensure all sequences have the same length. The feature dimensions extracted by the network are uniformly adjusted to $512 \times 3 \times 3$ to minimize inter-modal differences in the GAN [5]. The multilayer perceptron served as the generator, reducing the audio–visual features to 128 dimensions, which are treated as modality-independent features. A binary classification network functioned as the discriminator to train the generator. We then employed a strong intra-modal correlation matrix to aggregate identity features and a weak intra-modal correlation matrix to highlight interference features. These features are transferred to a fully connected network for bidirectional prediction. The bidirectional prediction results are processed through counterfactual inference to derive the final matching probability. In summary, we propose a bidirectional intervention attention network that leverages correlation matrix disentanglement and counterfactual reasoning to achieve debiased predictions.

(2) Training parameters. During training, we use the Adaptive Moment Estimation (Adam) optimizer with a batch size of 50, momentum set to 0.9, and a weight decay rate of 0.0005. The initial learning rate for each feature extractor module is set to 5×10^{-2}, while the initial learning rates for the generator and discriminator are set to 5×10^{-3}. The matching classifier has an initial learning rate of 5×10^{-2}, and the learning rate for both the Positive Intervention Attention (PIA) module and the Negative Intervention Attention (NIA) module is set to 5×10^{-3}. Each module is trained for 20 and 40 epochs, respectively, with a delay

Table 1
A validation of different scenarios performed on the VoxCeleb1 [16] dataset to evaluate the performance of the matching task. Where k=1 denotes the validation task, 1:2 denotes the binary matching task, while 2:k (K=10) denotes the multi-way matching task.

Methods	Venue	Backbone	Binary		Multi-way		Verification	
			V-F	F-V	V-F	F-V	V-F	F-V
SVHF [18]	CVPR2018		81.0	79.5	34.5	×	–	–
DIMNet [19]	ICLR2019		81.3	81.9	38.4	36.2	81.0	81.2
Wang's [20]	ACM2020		83.4	84.2	39.7	36.4	82.6	82.9
Wen's [8]	CVPR2021		87.2	86.5	48.2	44.8	87.2	87.0
AML [5]	TMM2022		90.2	86.3	46.2	43.7	86.4	86.2
DCLR [4]	ICDM2022	CNN	86.79	87.45	–	–	86.76	86.89
DSANet [7]	TMM2023		92.5	88.4	49.1	46.8	87.4	91.5
P^2VANet [24]	TCSVT2024		93.1	90.4	**50.6**	**48.1**	88.5	88.7
ACIENet [2]	TIFS2024		96.0	92.3	49.5	47.1	**90.1**	91.9
Baseline	Ours	CNN	94.8	89.8	48.5	45.6	88.2	91.2
BIANet	Ours	Transformer	80.7	93.0	18.9	25.6	89.3	91.6
BIANet	Ours	CNN	**97.3**	**96.9**	50.2	47.3	89.5	**92.8**

of 0.1 epoch. The validation experiment is treated as an independent matching task, where candidate samples are classified based on 256-dimensional features to determine whether they match. In this task, positive intervention audio–visual features are used as inputs to the classification network, and each audio clip feature is subtracted from the corresponding face image features to generate $K * 128$-dimensional features. The same operation is applied to the negative intervention features. Both positive and negative bidirectional features are input to the network to produce K match probabilities, the differences of which are used for the final prediction. The highest probability indicates the matched sample, and audio–visual matching performance is evaluated based on the accuracy (ACC).

(3) Dataset Description. The **VoxCeleb1** [16] dataset is derived from YouTube videos and contains a large collection of real-world speech samples from over a thousand celebrities. It is widely used for audio–visual tasks in complex acoustic environments to enhance recognition capabilities for security system authentication and forensic identification. This task is particularly challenging due to dual constraints from both external and internal factors. External interference includes background noise, music, laughter, echo effects, and distortions introduced by vocal tract or microphone characteristics. Internal variability arises from differences in speaker age, regional accent, emotional state, intonation, and speech style. Specifically, the combined dataset comprises 1225 distinct identities, with VoxCeleb1 contributing 137,060 facial images and VGGFace providing 149,354 audio clips. To enhance model generalization, previous research [8] divided the dataset alphabetically: 112 validation samples (prefixes "A" and "B"), 189 test samples (prefixes "C", "D", and "E"), and 924 training samples from the remaining prefixes.

VoxCeleb2 [17] is a large-scale audio–visual bimodal dataset comprising speech data from over one million YouTube clips featuring more than 6000 public figures. The dataset maintains a balanced gender distribution and demonstrates extensive diversity in ethnicity, accent, occupation, and age. Data are collected from a wide range of real-world scenarios, including red carpet events, outdoor sports venues, indoor studios, public speeches, professional film and television productions, and mobile device recordings. The audio data contain various forms of environmental interference, such as background noise, laughter, and overlapping sound sources, while the corresponding facial images exhibit natural variations in head pose, illumination, and motion blur. This dataset poses significant challenges for audio–visual in large-scale, unconstrained environments, providing new opportunities to advance research on robustness, scalability, and multimodal fusion in open-world settings. VoxCeleb2 supports multiple research tasks, including speaker recognition and visual speech synthesis. Regarding data partitioning, the training set comprises 5994 identity corresponding samples, while the test set consists of 118 independent identities.

5.2. Comparison to the SOTA

(1) Evaluation Results on VoxCeleb1 [16]. To evaluate the superiority of our method, we conduct comparative experiments against current mainstream audio–visual matching algorithms. Notably, most current mainstream models are based on Convolutional Neural Networks (CNNs). In contrast, Transformer [38]-based models often demonstrate suboptimal performance, as their attention mechanisms are prone to overfitting on disruptive information. As shown in Table 1, all comparative experiments are conducted using the data segmentation method proposed by Wen's method [8]. Specifically, DCLR [4] employs feature disentanglement to remove distracting features and improve audio–visual matching performance through debiasing, while Wen's method [8] achieves debiasing by filtering distracting features at the sample level. However, the performance of both methods is lower than that of attribute-supervised models [2,24], highlighting the limitations of existing feature- and sample-level debiasing approaches in reducing prediction bias.

To address prediction biases caused by interfering data, we propose the BIANet method, which demonstrates optimal or near-optimal performance across binary matching, multi-way matching, and verification tasks. Notably, BIANet achieves superior results in binary matching, illustrating that effective debiasing significantly enhances model performance. Although BIANet's performance in multi-way matching and verification tasks is marginally inferior to attribute-based methods [2, 24], the gap is negligible. Crucially, attribute-based methods require additional labeled data for supervision, while our approach achieves comparable or superior performance without relying on attribute labels, which highlights BIANet's strength in practical applications. Furthermore, as shown in Table 1, previous attribute-based methods [2, 24] exhibit significant discrepancies in matching performance between audio-to-face (V-F) and face-to-audio (F-V) scenarios, primarily due to interference in the data. In contrast, BIANet achieves more balanced performance across both scenarios, demonstrating its capability to effectively mitigate the influence of interfering features through model debiasing.

To comprehensively validate algorithmic effectiveness, we integrate the BIANet approach with existing feature enhancement techniques. Results in Table 2 demonstrate that this integration significantly boosts overall model performance. These studies indicate that DSANet and ACIEnet, as advanced feature enhancement methods, demonstrate robust performance in audio–visual matching. However, they inadequately address inherent model biases. In contrast, the core module of BIANet is designed to augment existing techniques by effectively suppressing interference, leading to optimized overall performance. Furthermore, when addressing complex interference patterns in real-world scenarios, the method avoids dependence on specific architectural designs. Instead, it establishes a universal interference processing framework, underscoring its broad applicability and practical potential.

Table 2
Comparison of integration results between the proposed method and feature enhancement techniques on the VoxCeleb1 [16] dataset. * indicates feature enhancement methods using the corresponding algorithm, while "Ours" refers to the inference architecture of BIANet.

Methods	Binary (ACC)		Multi-way (ACC)		Verification (ACC)	
	V-F	F-V	V-F	F-V	V-F	F-V
DSANet [7]	92.5	88.4	49.1	46.8	87.4	91.5
DSANet*+Ours	**97.5**	**97.1**	**50.5**	**47.5**	**88.5**	**93.0**
ACIENet [2]	96.0	92.3	49.5	47.1	90.1	91.9
ACIENet*+Ours	**97.6**	**97.3**	**50.4**	**47.8**	**90.5**	**93.2**

Table 3
Experimental comparisons of audio–visual matching task with SOTA methods for various scenarios performed on the VoxCeleb1 [16] dataset. These results are derived based on the PINs data configuration.

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (ACC)
		V-F	F-V	V-F	F-V	
DIMNet [19]	ICLR2019	84.12	84.03	39.75	–	83.2
PINs [39]	ECCV2018	84.00	–	31.00	–	78.5
SSNet [21]	DIC2019	78.00	78.50	30.00	30.05	78.8
β-VAE [3]	TMM2021	84.15	84.22	41.30	40.02	84.64
AML [5]	TMM2022	92.72	93.3	43.45	39.35	80.6
CMPC [40]	IJCAI2022	82.2	81.7	–	–	84.6
DSANet [7]	TMM2023	95.25	94.28	46.83	43.36	78.0
ACIENet [2]	TIFS2024	96.4	95.6	46.9	44.1	84.8
BIANet	Ours	**97.9**	**97.5**	**47.8**	**44.6**	**85.1**

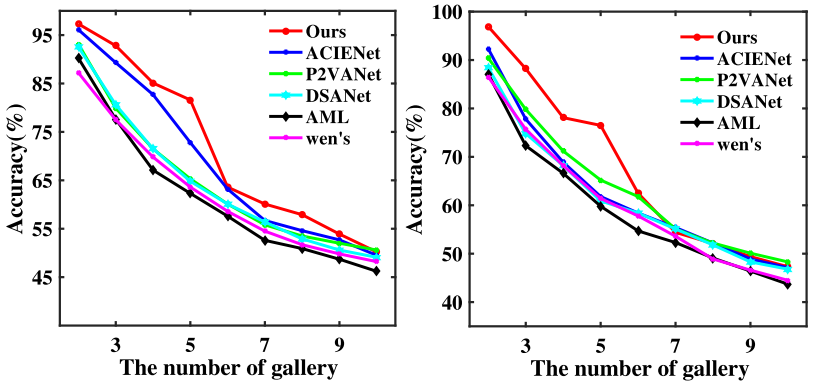

Fig. 4. The quantitative results of 2 : 10 matching task in V-F and F-V scenarios on VoxCeleb1 [16].

To further validate the effectiveness of the BIANet method, we are conducting a 2 : 10 multidimensional audio–visual matching experiment, as illustrated in Fig. 4. The results indicate that model performance declines rapidly as the number of matching candidates increases, highlighting the significant impact of candidate quantity on matching difficulty. Nevertheless, in the V-F scenario, our method exhibits notable superiority, particularly when the number of candidate samples is fewer than 6. Similarly, in the F-V scenario with fewer than 6 candidates, BIANet demonstrates a substantial advantage and remains competitive even as the number of candidates increases. These findings validate the BIANet method in performing debiased predictions to achieve reliable audio–visual matching. Additionally, we compare the BIANet method with other classical algorithms using PINs data segmentation, as shown in Table 3. Among these, β-VAE [3], an effective feature disentanglement method, performs significantly worse than attribute-guided enhancement-based approaches [2,24]. This outcome underscores the importance of both removing interference and exploring audio–visual matching association features. However, in practice, interference and association features often coexist, complicating the processing. As shown in Fig. 5, the baseline model tends to lock too many feature regions, leading to overfitting of background information. Although existing enhancement algorithms can mitigate interference to concentrate on foreground features, they still frequently introduce biased features. In contrast, BIANet can effectively focus on specific facial features, significantly reducing the fitting of interference features. Audio data similarly suffers from background information overfitting due to interference, which cannot be clearly visualized in audio. Future research must explore suitable approaches to represent this phenomenon. Experimental results show that BIANet effectively addresses this challenge and performs superiorly across various scenarios and tasks. The BIANet method enhances audio–visual matching performance by integrating feature disentanglement and counterfactual inference to learn feature associations and unbiased predictions.

The experimental analysis is conducted on specific interference datasets. We compare the performance of the proposed BIANet method with advanced feature enhancement algorithms DSANet and ACIEnet in the binary matching task. As shown in Fig. 6, existing models tend to overfit to intra-class data under interference conditions, which leads to a bias in feature attention. This feature bias causes the model to deviate from salient foreground features or allocate insufficient attention, ultimately resulting in incorrect matching predictions. In contrast, our method consistently focuses on salient foreground features under the same interference conditions and accurately retrieves correct matches, demonstrating robust matching capability on disturbed data.

(2) Evaluation Results on VoxCeleb2 [17]. VoxCeleb2 is a dataset enriched with diverse character information, making it an essential

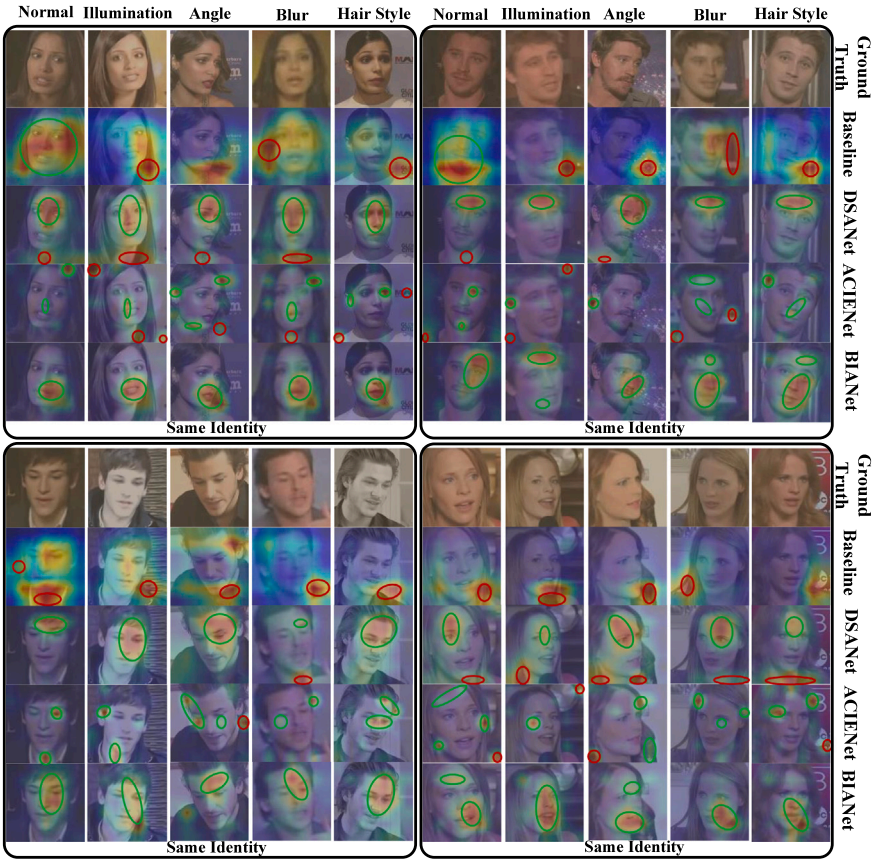

Fig. 5. Comparison of class activation maps (CAMs) generated by different methods on the VoxCeleb1 dataset. Red regions: overfitting-induced bias; green regions: model-focused foreground.

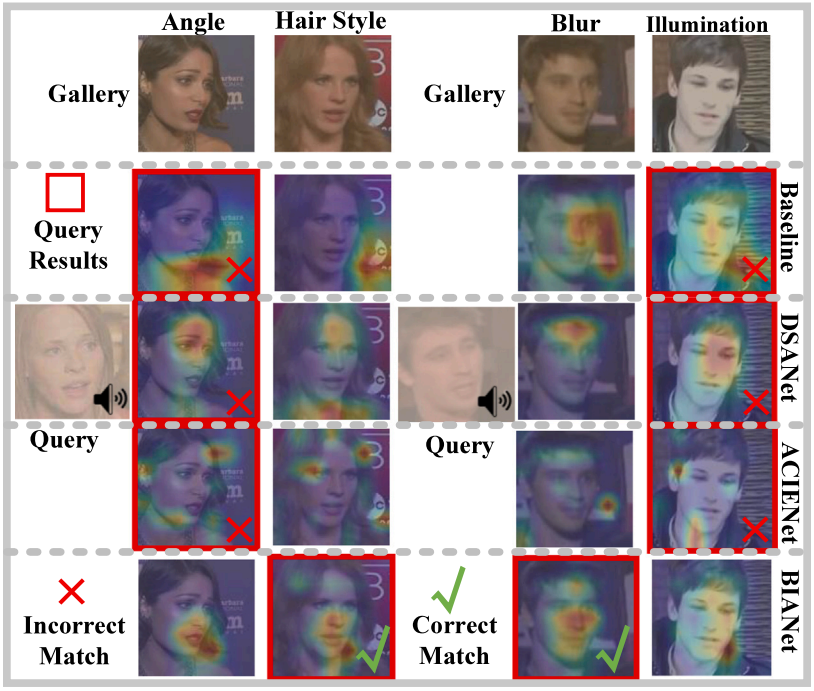

Fig. 6. In the $k = 2$ matching scenario, models are frequently prone to erroneous matches due to various forms of interference. The proposed method effectively corrects these matching results by mitigating such interference.

Table 4
The qualitative results for the task in different scenarios on VoxCeleb2 [17] dataset. '×' 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 [18]	CVPR2018	68.7	67.9	×	×	–	–
DIMNet [19]	ICLR2019	68.5	69.0	–	–	–	–
AML [5]	TMM2021	80.2	81.4	41.2	40.7	80.6	78.4
DSANet [7]	TMM2023	82.9	83.6	42.3	41.2	78.8	77.5
P^2VANet [24]	TCSVT2024	87.3	85.2	46.2	45.1	84.9	82.1
ACIENet [2]	TIFS2024	88.1	88.7	45.6	44.3	**86.3**	**91.5**
BIANet	Ours	**91.3**	**91.5**	**46.3**	**46.8**	85.8	91.1

Table 5
The experiments on component ablation of the BIANet method are performed in different audio–visual matching tasks.

Component			VoxCeleb1 [16]						VoxCeleb2 [17]					
			Binary		Multi-way		Verification		Binary		Multi-way		Verification	
	PIA	NIA	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	✓		96.8	96.6	49.8	46.8	89.3	92.6	91.1	91.1	45.8	45.9	84.2	90.9
c		✓	96.6	93.0	42.1	35.9	85.2	91.8	91.5	90.6	37.3	32.4	83.4	90.2
d	✓	✓	**97.3**	**96.9**	**50.2**	**47.3**	**89.5**	**92.8**	**92.0**	**91.5**	**46.3**	**46.8**	**85.8**	**91.1**

benchmark for assessing the generalization performance of algorithms. To evaluate the generalization ability of our model, we use the model trained on VoxCeleb1 [16] to process the test data in VoxCeleb2. As the latest algorithm tested for generalization capability on the VoxCeleb2 dataset, we conduct a comparative analysis between our method and P^2VANet [24]. As shown in Table 4, our method achieves SOTA performance in binary and multi-way matching tasks, while performing slightly below optimal in verification tasks. These results show that our method can perform debiased prediction, thus making the model resistant to interference to achieve superior generalizability.

5.3. Ablation study

(1) Evaluation of Different Component Effectiveness: Table 5 presents two intervention attention mechanisms designed to remove biases and demonstrates their effectiveness through experiments conducted on two datasets across multiple tasks and scenarios. The proposed BIANet integrates a Positive Intervention Attention (PIA) module and a Negative Intervention Attention (NIA) module. The PIA module reduces the impact of distractor features by aggregating identity features within the same modality. As evidenced in the comparison between Table 5(b) and (a), implementing the PIA module significantly enhances the performance of the baseline model, indicating that the correlation matrix disentanglement approach effectively reduces bias. However, interference features may sometimes be incorrectly associated with foreground features, leading the attention-based feature aggregation method to amplify the effects of these interference features. The NIA module leverages interference features for backward prediction and gradually introduces counterfactual inference to address this issue. From the comparison of Table 5(c) with (a), the NIA module performs well in binary matching tasks but introduces side effects in verification and multi-way matching tasks. In the verification task, the absence of direct comparisons between positive and negative sample information results in an inaccurate correlation matrix untangling, adversely affecting performance. In multi-way matching tasks, the difficulty of matching increases sharply as the number of candidate samples grows, making it challenging for the network to identify and de-bias interfering features within the samples.

Our approach addresses the problem of prediction bias by leveraging separation information from bidirectional interventions to facilitate counterfactual inference. As demonstrated in Tables 5(d) vs (a) and 5(d) vs (b), this method synergistically integrates both positive and negative perspectives to achieve effective counterfactual inference. In addition, we observe that baseline models significantly outperform the F-V scenario in both binary matching and multi-directional matching tasks under the V-F scenario. This performance gap may be attributed to data interference. Notably, the experimental results indicate that the BIANet model achieves a more significant performance improvement in the F-V scenario compared to the V-F scenario. This finding indicates that audio noise significantly impacts audio–visual matching, while the BIANet method effectively mitigates biases caused by such interference.

Table 6
The impact of the setting of the disentanglement threshold (ε) on the audio–visual matching task performance on the VoxCeleb1 [16] dataset.

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	96.9	96.3	**50.2**	46.0	89.2	92.6
0.1	96.8	96.5	50.1	46.3	**90.0**	91.4
0.2	97.1	96.4	49.6	45.9	88.5	92.7
Self-Learning	**97.3**	**96.9**	**50.2**	**47.3**	89.5	**92.8**

Table 7
The effect of curriculum learning strategy (CLS) on the BIANet method in various audio–visual matching tasks on the VoxCeleb1 [16] dataset.

BIANet	Binary		Multi-way		Verification	
(CLS)	V-F	F-V	V-F	F-V	V-F	F-V
w/o	94.9	96.8	45.6	44.4	89.4	84.6
w	**97.3**	**96.9**	**50.2**	**47.3**	**89.5**	**92.8**

(2) Evaluate the Impact of the Disentanglement Threshold ε Setting on Model Performance. To evaluate the impact of the disentanglement threshold parameter (ε), we compare the BIANet method by both the fixed hyperparameter and the self-learning approaches. As shown in Table 6, adjusting (ε) enhances the BIANet method performance in both scenarios, indicating the effectiveness of the disentanglement threshold in mitigating overfitting. Although the effect of adjusting the parameters for different tasks varies, the experimental results as a whole do not differ significantly. The self-learning approach is preferred in this paper to optimize the value of ε to achieve a more competitive performance.

(3) Evaluating the Impact of Curriculum Learning Strategy (CLS) on Audio–Visual Matching Performance. We perform a comparative analysis of BIANet methods with and without the CLS method. As shown in Table 7, directly implementing counterfactual feature

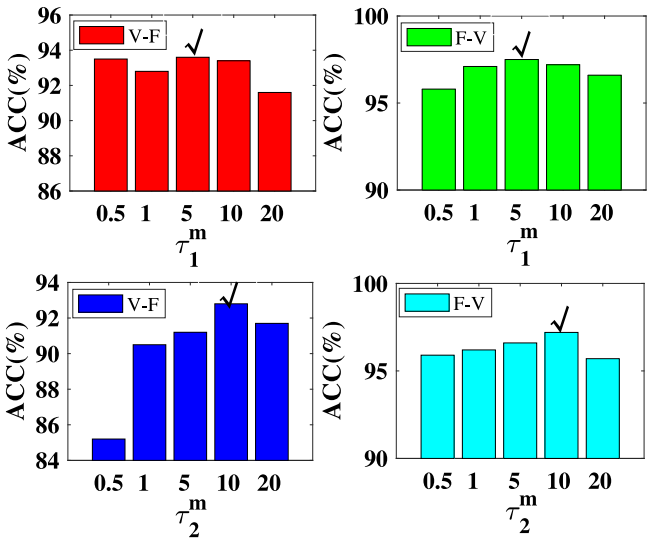

Fig. 7. The effects of two hyperparameters on the binary matching task using the VoxCeleb1 [16] dataset.

intervention without incorporating a CLS diminishes the performance of audio–visual information and worsens model performance. Conversely, integrating the CLS significantly bolsters the model's resilience to overfitting, improving the model's performance.

5.4. Hyper-parameters analysis

In our model, two key parameters influence the degree of correlation in the feature correlation matrix. The hyperparameter τ_1^m, as defined in Eq. (12), adjusts the autocorrelation of audio and face images. Fig. 7 illustrates that varying these parameters within the set $[0.5, 1, 5, 10, 20]$ results in only minor performance fluctuations, with the optimal parameter value being 5 for both V-F and F-V scenarios. This indicates that the autocorrelation matrix can enhance within-mode feature aggregation without requiring precise hyperparameter tuning. Conversely, the hyperparameters τ_2^m in Eq. (19) regulate the inter-correlation between audio and face images. We find that the performance of τ_2^m in V-F and F-V scenes gradually improves with increasing parameter values, and the higher performance can only be achieved at larger parameter values, although tremendous values should be avoided. This suggests that appropriate hyperparameter selection can emphasize relevant features and support the model in counterfactual Inference, enhancing audio–visual matching performance.

6. Conclusion and future works

In this work, we propose the BIANet framework as a universal solution for interference suppression. This approach integrates feature disentangling and counterfactual inference to address bias in audio–visual matching models caused by data interference. Specifically, we construct Positive Intervention Attention (PIA) and Negative Intervention Attention (NIA) modules by decoupling the correlation matrix, which together form a counterfactual inference architecture. The PIA module aggregates identity-related features based on strong correlations to enhance audio–visual feature recognition and mitigate interference, while the NIA module aggregates identity-irrelevant features through weak correlations. Since identity-related and unrelated features often become entangled during model training, and models tend to learn effective features before overfitting to interference features, a Curriculum Learning Strategy (CLS) is introduced. This strategy progressively strengthens counterfactual inference, ensuring that effective features are thoroughly learned during the early training phase. Experimental results demonstrate that BIANet achieves significant bias correction and superior overall performance.

Furthermore, this study presents a universal bias-mitigation framework that can be extended to other cross-modal perception tasks, thereby advancing the field. However, in real-world scenarios, the proportion of samples affected by interference in audio–visual datasets often follows a long-tail distribution, which limits the ability of general interference mitigation models to effectively address these effects. Future research will therefore focus on the statistical characterization of interference information to train targeted multi-perception mechanism groups. This technology aims to enhance interference-specific processing capabilities and enable the development of more robust models, ultimately improving overall system performance.

CRediT authorship contribution statement

Jiaxiang Wang: Writing – original draft. **Aihua Zheng:** Writing – review & editing. **Dequan Li:** Formal analysis. **Chenglong Li:** Methodology. **Wenjuan Cheng:** Investigation. **Ran He:** Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jiaxiang Wang reports financial support was provided by Anhui University of Science and Technology. Aihua zheng reports financial support was provided by Anhui University. Reports a relationship with that includes:. Has patent pending to. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work is supported by the Scientific Research Foundation for High-level Talents of Anhui University of Science and Technology (No. 2024yjrc95), the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202508), the Open Project of National Key Laboratory of Optoelectronic Information Acquisition and Protection Technology, Anhui University (No. OEIAPT202511), the National Natural Science Foundation of China under Grants (No. 62372003), the Natural Science Foundation of Anhui Province under Grant (No. 2308085Y40), the National Key Research and Development Program Project (No. 2023YFC3807501), and the Special Project of State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology (No. OEIAPT20250102).

Data availability

The authors do not have permission to share data.

References

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

[2] J. Wang, A. Zheng, Y. Yan, R. He, J. Tang, Attribute-guided cross-modal interaction and enhancement for audio-visual matching, IEEE Trans. Inf. Forensics Secur. 19 (2024) 4986–4998.

[3] H. Ning, X. Zheng, X. Lu, Y. Yuan, Disentangled representation learning for cross-modal biometric matching, IEEE Trans. Multimed. 24 (2021) 1763–1774.

[4] Z. Yu, X. Liu, Y.-M. Cheung, M. Zhu, X. Xu, N. Wang, 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, 2022, pp. 648–655.

[5] A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, B. Luo, Adversarial-metric learning for audio-visual cross-modal matching, IEEE Trans. Multimed. 24 (2021) 338–351.
[6] K. Cheng, X. Liu, Y.-m. Cheung, R. Wang, X. Xu, 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, 2020, pp. 448–455.
[7] J. Wang, C. Li, A. Zheng, J. Tang, B. Luo, Looking and hearing into details: Dual-enhanced siamese adversarial network for audio-visual matching, IEEE Trans. Multimed. 25 (2023) 7505–7516.
[8] P. Wen, Q. Xu, Y. Jiang, Z. Yang, Y. He, Q. Huang, Seeking the shape of sound: An adaptive framework for learning voice-face association, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16347–16356.
[9] D. Yang, M. Li, D. Xiao, Y. Liu, K. Yang, Z. Chen, Y. Wang, P. Zhai, K. Li, L. Zhang, Towards multimodal sentiment analysis debiasing via bias purification, in: Proceedings of the European Conference on Computer Vision, 2025, pp. 464–481.
[10] S. Zhao, Z. Jin, Q. Jiao, Q. Zhang, J. Han, Resolving semantic conflicts in RGB-t semantic segmentation, Pattern Recognit. 162 (2025) 111398.
[11] X. Gong, M. Liu, Q. Liu, Y. Guo, G. Wang, MDFCL: Multimodal data fusion-based graph contrastive learning framework for molecular property prediction, Pattern Recognit. 163 (2025) 111463.
[12] T. Sun, W. Wang, L. Jing, Y. Cui, X. Song, L. Nie, Counterfactual reasoning for out-of-distribution multimodal sentiment analysis, in: Proceedings of the ACM International Conference on Multimedia, 2022, pp. 15–23.
[13] X. Li, Y. Lu, B. Liu, Y. Liu, G. Yin, Q. Chu, J. Huang, F. Zhu, R. Zhao, N. Yu, Counterfactual intervention feature transfer for visible-infrared person re-identification, in: Proceedings of the European Conference on Computer Vision, 2022, pp. 381–398.
[14] D. Yang, K. Yang, H. Kuang, Z. Chen, Y. Wang, L. Zhang, Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 10663–10680.
[15] M. Ye, J. Shen, D. J. Crandall, L. Shao, J. Luo, Dynamic dual-attentive aggregation learning for visible-infrared person re-identification, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 229–247.
[16] A. Nagrani, J.S. Chung, A. Zisserman, Voxceleb: a large-scale speaker identification dataset, in: Proceedings of the International Speech Communication Association, 2017, pp. 2616–2620.
[17] J.S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, in: Proceedings of the International Speech Communication Association, 2018, pp. 1086–1090.
[18] A. Nagrani, S. Albanie, A. Zisserman, Seeing voices and hearing faces: Cross-modal biometric matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8427–8436.
[19] Y. Wen, M.A. Ismail, W. Liu, B. Raj, R. Singh, Disjoint mapping network for cross-modal matching of voices and faces, in: Proceedings of the International Conference on Learning Representations, 2019.
[20] R. Wang, X. Liu, Y.-m. Cheung, K. Cheng, N. Wang, 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, 2020, pp. 1881–1884.
[21] S. Nawaz, M.K. Janjua, I. Gallo, A. Mahmood, 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, 2019, pp. 1–7.
[22] R. He, J. Cao, L. Song, Z. Sun, T. Tan, Adversarial cross-spectral face completion for NIR-vis face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2019) 1025–1037.
[23] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, R. He, Deep audio-visual learning: A survey, Int. J. Autom. Comput. 18 (3) (2021) 351–376.
[24] A. Zheng, F. Yuan, H. Zhang, J. Wang, C. Tang, C. Li, Public-private attributes-based variational adversarial network for audio-visual cross-modal matching, IEEE Trans. Circuits Syst. Video Technol. 34 (9) (2024) 8698–8709.
[25] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, Beta-vae: Learning basic visual concepts with a constrained variational framework, in: Proceedings of the International Conference on Learning Representations, 2016.
[26] H. Kim, A. Mnih, Disentangling by factorising, in: Proceedings of the International Conference on Machine Learning, 2018, pp. 2649–2658.
[27] D. Peng, Z. Li, Unbiased VQA via modal information interaction and question transformation, Pattern Recognit. (2025) 111394.
[28] C. Huang, J. Chen, Q. Huang, S. Wang, Y. Tu, X. Huang, Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis, Inf. Fusion 114 (2025) 102725.
[29] L. Wang, Z. He, R. Dang, M. Shen, C. Liu, Q. Chen, Vision-and-language navigation via causal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13139–13150.
[30] Y. Wang, L. Meng, H. Ma, Y. Wang, H. Huang, X. Meng, Modeling event-level causal representation for video classification, in: Proceedings of the ACM International Conference on Multimedia, 2024, pp. 3936–3944.
[31] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, J.-R. Wen, Counterfactual vqa: A cause-effect look at language bias, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12700–12710.
[32] R. Yan, L. Xie, X. Shu, L. Zhang, J. Tang, Progressive instance-aware feature learning for compositional action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 10317–10330.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[34] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[35] P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, Cross-modality person re-identification with generative adversarial training, in: Proceedings of the International Joint Conference on Artificial Intelligence, vol. 1, (3) 2018, p. 6.
[36] A. Andonian, S. Chen, R. Hamid, Robust cross-modal representation learning with progressive self-distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16430–16441.
[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Advances in Neural Information Processing Systems, 30, 2017.
[39] A. Nagrani, S. Albanie, A. Zisserman, Learnable pins: Cross-modal embeddings for person identity, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 71–88.
[40] B. Zhu, K. Xu, C. Wang, Z. Qin, T. Sun, H. Wang, Y. Peng, Unsupervised voice-face representation learning by cross-modal prototype contrast, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2022, pp. 3787–3794.