Attribute-guided Cross-modal Interaction and Enhancement for Audio-Visual Matching

Jiaxiang Wang, Aihua Zheng*, Yan Yan, Ran He, Jin Tang

***Abstract*—Audio-visual matching is an essential task that measures the correlation between audio clips and visual images. However, current methods rely solely on the joint embedding of global features from audio clips and face image pairs to learn semantic correlations. This approach overlooks the importance of high-confidence correlations and discrepancies of local subtle features, which are crucial for cross-modal matching. To address this issue, we propose a novel Attribute-guided Cross-modal Interaction and Enhancement Network (ACIENet), which employs multiple attributes to explore the associations of different key local subtle features. The ACIENet contains two novel modules: the Attribute-guided Interaction (AGI) module and the Attribute-guided Enhancement (AGE) module. The AGI module employs global feature alignment similarity to guide cross-modal local feature interactions, which enhances cross-modal association features for the same identity and expands cross-modal distinctive features for different identities. Additionally, the interactive features and original features are fused to ensure intra-class discriminability and inter-class correspondence. The AGE module captures subtle attribute-related features by using an attribute-driven network, thereby enhancing discrimination at the attribute level. Specifically, it strengthens the combined attribute-related features of gender and nationality. To prevent interference between multiple attribute features, we design a multi-attribute learning network as a parallel framework. Experiments conducted on a public benchmark dataset demonstrate the efficacy of the ACIENet method in different scenarios. Code and models are available at https://github.com/w1018979952/ACIENet.**

***Index Terms*—Audio-visual cross-modal matching, attribute-guided cross-modal interaction, attribute-guided cross-modal enhancement.**

I. Introduction

Psychological studies have shown that people can match faces with their corresponding identities with a high degree of accuracy by hearing the voice of an unfamiliar person, and vice versa [1]–[4]. The human brain is capable of identifying the same identity by learning only the face or audio information in the multimodal brain regions, which generate correlations between the two modalities [5]. This has led to the emergence of a new research topic, known as audio-visual matching, which seeks to associate a face image with the voice information of the corresponding speaker. This technique is useful for various traditional machine learning tasks, including audio-visual speech separation [6], [7], face recognition [8], [9], speaker recognition [10]–[12], and audio-visual localization [13], [14].

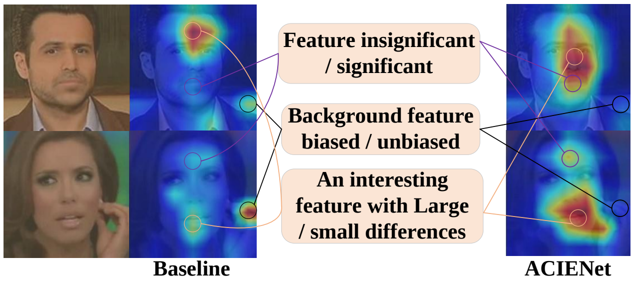

Feature insignificant / significant
Background feature biased / unbiased
An interesting feature with Large / small differences
Baseline
ACIENet

Fig. 1. Comparison of Baseline and ACIENet Methods for Audio-Visual Matching Using Cumulative Activation Maps (CAM). The current method for audio-visual matching involves concatenating global features for classification. However, our study has identified three main problems with this approach. Firstly, local cross-modal features do not appear to be significant. Secondly, background features are overemphasized. Thirdly, regions of interest are visually different. To address these issues, we propose the ACIENet method, which enhances local cross-modal features, suppresses background features, and expands the range of perception.

The major challenge in visual-audio matching is to precisely measure the similarity between the feature embeddings of the two modalities. Nagrani *et al.* [15] first launches the audio-visual cross-modality matching task by designing a binary classification network. Due to the heterogeneity of cross-modal features, Wang *et al.* [16] and Nawaz *et al.* [17] use a shared common space to map two modal features to mitigate the effect of modal differences. Furthermore, the distance loss is designed to effectively constrain the distribution of features to learn the joint global feature embedding. However, the cross-modal data has the problem of modality heterogeneity, which leads to an inconsistent distribution of modality features. Therefore, there emerge two types of methods, the common space feature mapping [17], [18] and the modality adversarial elimination [19]–[21]. By contrast, the latter approach can

This research is supported in part by the National Natural Science Foundation of China (62372003, U21B2045, and U20A20223), the University Synergy Innovation Program of Anhui Province (GXXT-2022-036), and the Natural Science Foundation of Anhui Province (2308085Y40). The corresponding author is Aihua Zheng.

A. Zheng is 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).

Y. Yan is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: yyan34@iit.edu).

R. He is with the School of Artificial Intelligence, 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, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China(e-mail: rhe@nlpr.ia.ac.cn).

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

better eliminate modal heterogeneity by generating adversarial networks (GAN) [22]. In spite of the great advances in audio-visual matching, there are two problems that have not been effectively addressed.

First, existing methods often utilize only global features and ignore the inter-correlation between local features [16], [17], [23]. However, the semantic information extracted by the respective modality only reflects the distribution of identity features under the same modality. In addition, modal heterogeneity is ubiquitous between cross-modal features. Therefore, it is necessary to bridge the correspondence between cross-modal features with the same identity information to reduce modal heterogeneity. Ning *et al.* [24] explores a disentangled latent variable method that separates cross-modal features into shared and private features. Shared features with the same identity undergo feature alignment, while joint private features perform intra-modal identity discrimination. An adaptive framework is proposed by Wen *et al.* [25] that considers not only cross-modal global feature alignment but also the diversity of learning difficulties between different objectives. However, the alignment methods with similarity measures often come from the complex aggregation of local similarities between audio-visuals, which is ignored by the cross-modal global feature alignment approach. Speaker voice and face differences usually occur in detail, resulting in suboptimal feature alignment in most existing schemes [26].

Herein, we propose an Attribute-guided Interaction (AGI) module to tackle the problem of potential invalid inter-modal interactions between audio clips and face image features. This module consists of an Inter-modal Interaction (IMI) structure, an Interactive Feature Combination (IFC), and an Identity Alignment Loss ($\mathcal{L}_{IA}$). To enhance the accuracy of audio-visual matching, we use a Compact Bilinear Pooling (CBP) [27] and an attribute classification network to obtain cross-modal identity similarity. This similarity guides the feature interactions between modalities and explores meaningful correlations between cross-modal features. Furthermore, we design the IFC scheme to leverage the discriminative ability of features by allowing cross-modal interaction features and the original dynamical feature fusion, ensuring intra-class discriminability and inter-class correspondence. To ensure consistency in cross-modal same-identity, we compute the Identity Alignment Loss ($\mathcal{L}_{IA}$). As different identities may have varying levels of difficulty in matching, we use the Simulated Annealing technique (SAT) [28] to weight the identity alignment loss ($\mathcal{L}_{IA}$), thus reducing the impact of hard-to-match samples on the network's robustness. The audio-visual matching model, as depicted in Fig. 1, utilizes an attribute-guided interaction and enhancement approach to direct the network's attention towards more discriminative feature regions, surpassing the baseline model's performance.

Second, the association between attribute information in audio clips and face images remains underexplored. The human brain is susceptible to correlations between multi-modal information, which are evident when recognizing gender and nationality by hearing an audio clip or seeing a face [1], [2], [5]. Fig. 2 (a) shows how short hair and beard features in a face image, or low tone and high loudness features in audio, can indicate the male gender. Similarly, in Fig. 2 (b), the high tone and low loudness features in audio, or long hair and thin eyebrows features in a face image, indicate female gender. To address this issue, Wen *et al.* [18] proposed a disjoint mapping network (DIMNet) for audio-visual matching, comprising a cross-modal embedding module and a multi-attribute collaborative supervised network training. However, a supervised network scenario with a multi-attribute serial approach presents difficulties in finding a locally optimal solution due to the mutual constraints between multiple attribute losses [29].

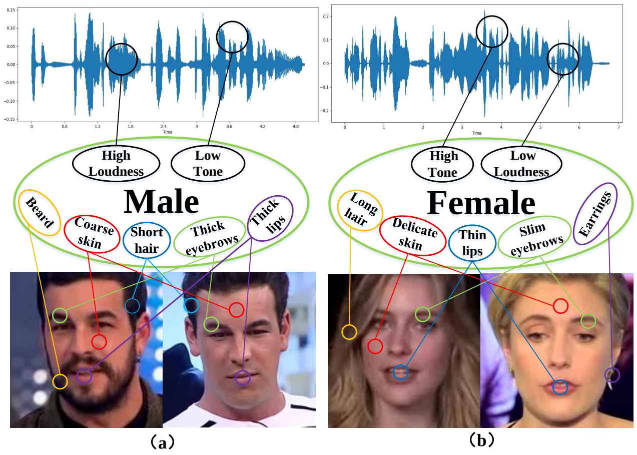


Fig. 2. Illustration of the attribute regions for audio and face images. The attribute annotations of two identity-identical face images are associated with the corresponding audio clip attribute annotations. In particular, the male image in (a) shows attribute regions of beard, rough skin, short hair, thick eyebrows, and thick lips, which correspond to the audio attributes of high loudness and low tone located in the upper left corner. Similarly, the female image in (b) displays attribute regions of long hair, delicate skin, thin eyebrows, thin lips, and earrings, which correspond to the audio attributes of high tone and low loudness located in the upper right corner.

Herein, we propose the Attribute-Guided Enhancement (AGE) module to take full advantage of attribute discrimination. The AGE module is an attribute-driven network designed to capture subtle attribute-related features, thereby improving attribute matching between audio-visual samples. Specifically, the AGE module consists of two parts. The first part is coding, which obtains relevant features of the predicted attributes. The second part is decoding, which generates attention weights for the attribute-related features. By weighting the extracted unimodal features based on attribute weight, we can highlight attribute-related local features. Unlike DIMNet [18], our proposed AGE module and AGI module operate in parallel, effectively mitigating interaction between multiple attribute losses.

It has been demonstrated by previous approaches [19]–[21] that modal heterogeneity can be eliminated using generative adversarial networks. We design the proposed ACIENet method to maintain this adversarial architecture. The main contributions of this work are summarized as follows:

- We propose an attribute-guided interaction (AGI) module to explore potential cross-modal local feature associations using cross-modal identity-aligned similarity-guided interaction correlation matrices. It can simultaneously reduce the discrepancy of the same identity and enhance the variability of different identities among cross-modal features.

- We propose an attribute-guided enhancement (AGE) module to capture subtle attribute-related features with attribute-driven networks. It can enhance the combined attribute-related features of gender and nationality raising hierarchical attribute discrimination. In addition, we design the parallel network with the identity attribute module to avoid mutual interference between multiple attribute features.
- Experiments show that the ACIENet method can effectively use multiple attributes for learning relationships between audio and visual. We perform audio-visual matching experiments on the Voxceleb [30] and VGGFace [31] datasets, which can achieve superior performance compared to state-of-the-art algorithms.

II. Related Work

A. Audio-visual Matching

Audio-visual matching is an important research topic in multimodal learning that is currently attracting a large number of researchers' interests. This topic, which originated from psychological research, was first proposed by Nagrani *et al.* [15] with the design of dual-stream deep neural network classification to achieve classification probabilities comparable to or even beyond the human baseline. Albanie *et al.* [32] proposed a joint course learning and contrast loss optimization embedding network to further mine the relationships between audio-visual data, which was extended to a broader range of application tasks. Wen *et al.* [18] employed more labels, such as identity, nationality, and gender, to co-supervise network training to learn shared representations instead of direct association of audio clips and face images. Wang *et al.* [16] used bidirectional ranking constraints, identity constraints, and centrality constraints to learn the association of face-voice discriminative features in small batches of data, which is a simple end-to-end joint embedding network.

The audio-visual matching task poses significant challenges due to modal heterogeneity and sample complexity across multi-modality. Despite achieving good performance, the problem remains to be addressed. To tackle these issues, Wen *et al.* [25] proposed a two-level modal alignment approach to learn hard but valuable identities, while filtering out identities that are difficult to learn. Additionally, Ning *et al.* [24] proposed a disentangled representation learning technique to decompose face and speech features into identity and modality-related features respectively, thereby reducing the feature differences of the same identity information by filtering out the modality features. Moreover, apart from feature alignment across modalities, exploring complementary cues between audio and visual modalities is also necessary. For this purpose, Saeed *et al.* [23] proposed a plug-and-play mechanism that decomposes and fuses features in a two-stream pipeline, thereby improving the discriminative joint feature embedding space for the face-voice association.

The presented method was inadequate in addressing the problem of heterogeneity across modalities, which can result in significant discrepancies in the extracted features for the same identity across modal samples. To overcome this limitation, Zheng *et al.* [20] proposed an adversarial measurement learning model for audio-visual matching that uses generative adversarial networks to learn modality-independent feature representations. Additionally, a similarity measure was employed to constrain the feature distribution and accelerate convergence. Similarly, Cheng *et al.* [19] proposed a similar approach, which used triple loss and modal center loss to eliminate modal heterogeneity and enhance the network's robustness. To further improve the correlation between audio and face features, Wang *et al.* [21] proposed a dual-enhanced siamese adversarial network to enhance the extracted audio and face features, respectively. Then a joint embedding representation was implemented using the siamese adversarial structure and structural metric learning. Lastly, Choi *et al.* [33] designed a CGAN-based generation framework to generate faces directly from speech. This method was used as an end-to-end network to achieve a seamless association between audio and face.

B. Cross-Modal Interaction and Enhancement

The cross-modal matching task has been extensively investigated and holds a pivotal role in facilitating an understanding of the relationships between cross-modal features. However, due to the presence of modal heterogeneity, Tu *et al.* [34] utilized the prior knowledge to guide an adversarial network capable of generating exceptionally realistic facial videos. To establish robust cross-modal correlations, Sun *et al.* [35] introduced a multi-subtitle attention mechanism designed to synthesize multi-word features, thereby generating highly semantically relevant facial images. Furthermore, the parsing-based method [36], [37] assisted the attention network in realizing fine-grained semantic correlations. Consequently, substantial research efforts have been directed toward enhancing cross-modal interaction and enhancement, building upon prior work [26], [38]–[40]. Attentional mechanisms have proven to be effective across a spectrum of tasks, including audio-visual event localization [41], [42], audio-visual expression recognition [43], [44], and image-text retrieval [45], [46].

The audio-visual matching task differs from other tasks in two main ways. First, instead of image and audio sequences, it employs paired face images and audio clips to learn shared feature representations. Cheng *et al.* [47] proposed a self-supervised framework that used joint attention mechanisms to focus on audio-visual synchronized sequence information with potential correlation, but it does not apply to audio-visual matching. Second, unlike image-text cross-modal alignment, an explicit cross-modal features alignment cannot be established between audio and visuals. This means that the attention approach used in previous studies may not be suitable for this task. Mercea *et al.* [48] proposed a cross-attention module to learn shared information between audio and visual representations, but without considering the impact of inter-modality associations and fine-grained attribute information for discrimination. To address this limitation, we design attribute-guided

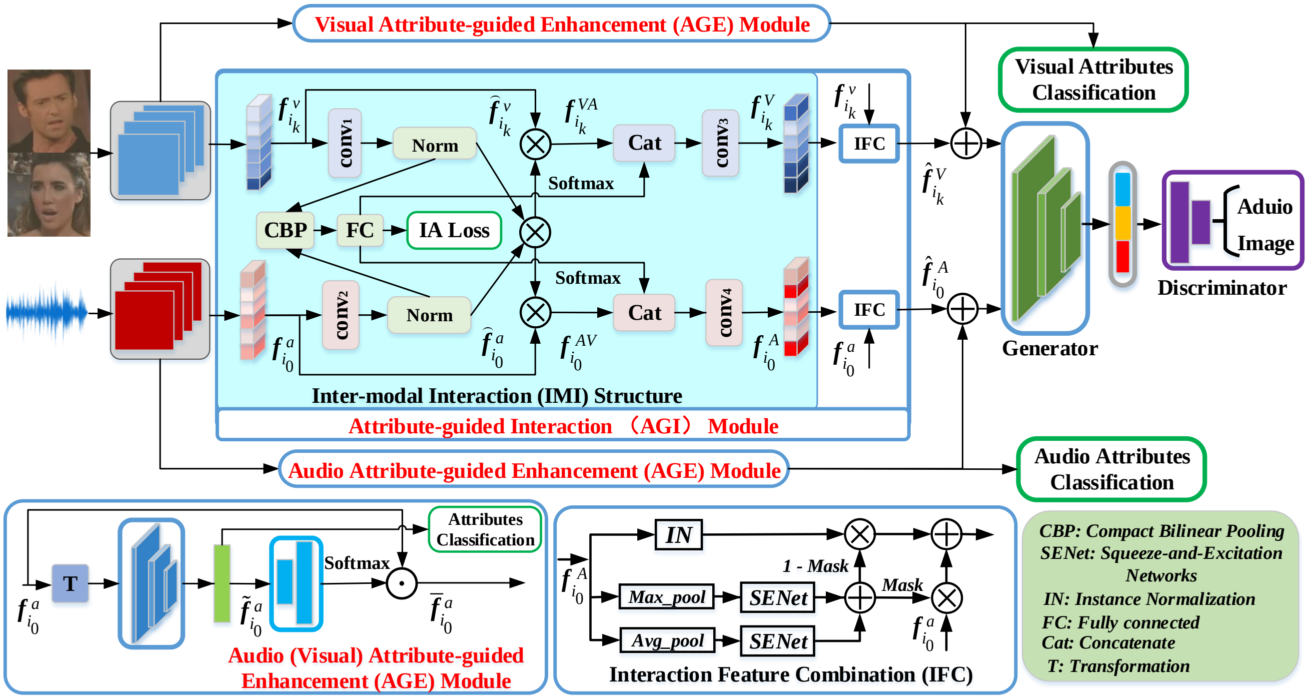

Visual Attribute-guided Enhancement (AGE) Module
Visual Attributes Classification
$f^v_{i_k}$
conv1
Norm
$\hat{f}^v_{i_k}$
$f^{VA}_{i_k}$
Cat
conv3
$f^V_{i_k}$
$f^v_{i_k}$
IFC
Softmax
CBP
FC
IA Loss
Softmax
conv2
Norm
Cat
conv4
$f^a_{i_0}$
$\hat{f}^a_{i_0}$
$f^{AV}_{i_0}$
$f^A_{i_0}$
IFC
$\hat{f}^V_{i_k}$
$\hat{f}^A_{i_0}$
$f^a_{i_0}$
Generator
Adiuo
Image
Discriminator
Inter-modal Interaction (IMI) Structure
Attribute-guided Interaction （AGI） Module
Audio Attribute-guided Enhancement (AGE) Module
Audio Attributes Classification
T
$f^a_{i_0}$
$\tilde{f}^a_{i_0}$
Softmax
Attributes Classification
$\bar{f}^a_{i_0}$
Audio (Visual) Attribute-guided Enhancement (AGE) Module
$f^A_{i_0}$
IN
1 - Mask
Max_pool
SENet
Mask
Avg_pool
SENet
$f^a_{i_0}$
Interaction Feature Combination (IFC)
CBP: Compact Bilinear Pooling
SENet: Squeeze-and-Excitation Networks
IN: Instance Normalization
FC: Fully connected
Cat: Concatenate
T: Transformation

Fig. 3. Overview of the overall architecture of ACIENet. It incorporates two novel components: the attribute-guided interaction (AGI) module and the attribute-guided enhancement (AGE) module. The AGI module comprises three components: the inter-modal interaction (IMI) structure, the interactive feature combination (IFC), and the identity alignment loss ($\mathcal{L}_{IA}$). The AGI module focuses on inter-modal local feature interactions between classes. AGE, on the other hand, leverages gender and nationality attributes to improve the subtle attribute-related features for hierarchical matching. The enhanced features, in conjunction with the interactive features, mitigate pattern heterogeneity by generating adversarial networks.

cross-modal interaction and enhancement networks to explore potential attribute feature correlations between the audio-visual modalities.

III. Method

The objective of this study is to investigate the interaction and enhancement of cross-modal attribute features to accurately perceive local features for reliable audio-visual matching. To achieve this goal, we employ the ResNet18 [49] and SE-ResNet-34 [49] as feature extraction architecture. An overview of our proposed method, the Attribute-guided Cross-modal Interaction and Enhancement Network (ACIENet), is presented in Fig. 3. The ACIENet comprises two modules, namely the attribute-guided interaction module and the attribute-guided enhancement module.

A. Audio-Visual Representations

To better comprehend audio-visual matching, we provide a detailed formulation of the V-F matching task in this paper. Its primary aim is to identify anchored audio clips and corresponding visual face images in a gallery with numerous candidates, as well as vice versa for the F-V matching task.

To achieve this, we utilize an anchor audio clip a_{i_0} and k visual face images $\{v_{i_1}, v_{i_2}, ..., v_{i_k}\}$ as a match in the gallery, where i denotes the i-th data tuple. Multiple audio clips and face images are paired and their features are extracted using ResNet18 [49] and SE-ResNet-34 [49], respectively. Research by Gu *et al.* [50] has revealed that cross-modal data may be better aligned by loading pre-training parameters exclusively for image modalities. For this reason, we load ResNet18 [49] with pre-trained parameters obtained by pre-training on ImageNet [51] data, while SE-ResNet-34 [49] does not import pre-trained model parameters. In this task, the extracted activation mappings for audio and face images are $\boldsymbol{f}^a \in \mathbb{R}^{C_1\times H_1\times W_1}$ and $\boldsymbol{f}^v \in \mathbb{R}^{C_2\times H_2\times W_2}$, respectively. Here, $C_1(C_2)$ represents the number of channels in the semantic features, and $H_1(H_2)$ and $W_1(W_2)$ represent the height and width of the semantic features. We pool the audio clip and face image features into a unified matrix dimension $\boldsymbol{f} \in \mathbb{R}^{C\times H\times W}$ to simplify subsequent computations. For a given data tuple, which consists of the audio clip $\boldsymbol{f}^a_{i_0}$ and the visual face images $\boldsymbol{f}^v_i = \{\boldsymbol{f}^v_{i_1}, \cdots, \boldsymbol{f}^v_{i_k}\}$.

B. Attribute-guided Interaction Module

In this paper, we present a novel method called the Attribute-guided Interaction (AGI) module, which distinguishes itself from prior attribute-based cross-modal methods, such as those proposed by β-VAE [24], Wen *et al.* [25], DIMNet [18], and AML [20]. Instead of relying on feature combinations of global features to enhance cross-modal interactions, the AGI module takes a different approach. It focuses on two main objectives: (1) learning local feature interactions to enhance the association between the same-identity attributes across modal features, and (2) alleviating the weak correlation between cross-modal data of different identities to improve the model's generalizability. To achieve these objectives, the AGI module comprises three components: the inter-modal interaction structure, the interaction feature combination, and the identity alignment loss. We will discuss each of these components in detail below.

1) Inter-modal Interaction Structure: The purpose of inter-modal interactions is to explore semantic relationships among inter-modal features that help establish inter-modal data relationships. The self-attention mechanism's input comprises a query vector (Q), a key (K), and a value (V) that are weighted by V to obtain crucial feature information related to the task. To learn inter-modal associations, we perform a cross-modal feature interaction using semantic features extracted from face images and audio clips to establish connections between the same categories. We initially consider features $\boldsymbol{f}^{a}_{i_0}$ as query vectors (Q) and $\boldsymbol{f}^{v}_{i_k}$ as keys (K) for channel dimension reduction to lower the computational load of feature interactions. Second, the distribution of sample features differs across modalities; Hence, we normalize cross-modal features to ensure that they have the same range of values, which hastens the model's convergence.

$$\widehat{\boldsymbol{f}^{v}_{i_k}} = Norm(conv_1(\boldsymbol{f}^{v}_{i_k})), \tag{1}$$

$$\widehat{\boldsymbol{f}^{a}_{i_0}} = Norm(conv_2(\boldsymbol{f}^{a}_{i_0})), \tag{2}$$

where $\widehat{\boldsymbol{f}^{a}_{i_0}}$ and $\widehat{\boldsymbol{f}^{v}_{i_k}}$ are the processed audio clip features and the kth face image features, respectively. To reduce computational effort, feature channel compression is performed on face image and audio clip features using 1×1 convolution operation denoted by $conv_1$ and $conv_2$, respectively. After normalization using $Norm$, the dot product between the query (Q) and the key (K) is computed to form the cross-modal interaction matrix. Applying a softmax operation on this matrix computes the attention values that help in emphasizing inter-class correlation of the same identity features and inter-class discrepancy of different identity features. The interaction between the kth face feature and audio features can be represented as:

$$\boldsymbol{f}^{VA}_{i_k} = softmax(\widehat{\boldsymbol{f}^{v}_{i_k}}\widehat{\boldsymbol{f}^{a}_{i_0}})\boldsymbol{f}^{v}_{i_k}, \tag{3}$$

$$\boldsymbol{f}^{AV}_{i_0} = softmax(\widehat{\boldsymbol{f}^{a}_{i_0}}\widehat{\boldsymbol{f}^{v}_{i_k}})\boldsymbol{f}^{a}_{i_0}, \tag{4}$$

where $\boldsymbol{f}^{AV}_{i_0}$ and $\boldsymbol{f}^{VA}_{i_k}$ are the audio clip features and the kth face image features after the cross-modal interaction. However, the pairwise relationship between audio clips and face images is not known, and exploring semantic relationships through direct feature interactions is impossible. To explicitly investigate the relationships between cross-modal features, identity labels are required to guide the feature interactions. Unfortunately, in testing situations, identity labeling is not available. To overcome this challenge, we estimate the cross-modal similarity using compact bilinear pooling (CBP) [27], which can be considered as an identity pseudo-label that guides the interaction. The computation of the cross-modal similarity is as follows:

$$S_{i_{0j}} = FC(CBP(\widehat{\boldsymbol{f}^{a}_{i_0}}, \widehat{\boldsymbol{f}^{v}_{i_j}})), \tag{5}$$

where FC stands for fully connected layer. $S_{i_{0j}}$ denotes the identity similarity between the audio clip $\boldsymbol{f}^{a}_{i_0}$ and the jth face image $\boldsymbol{f}^{v}_{i_j}$, which is [0, 1] for the same identity and [-1, 0] for different identities. Based on this similarity, we propose an identity similarity-guided cross-modal interaction method as follows:

$$\boldsymbol{f}^{V}_{i_k} = conv_3(cat[S_{i_{0k}}, \boldsymbol{f}^{VA}_{i_k}]), \tag{6}$$

$$\boldsymbol{f}^{A}_{i_0} = conv_4(cat[S_{i_{0k}}, \boldsymbol{f}^{AV}_{i_0}]), \tag{7}$$

where cat denotes the concatenated operation. Then $conv_3$ and $conv_4$ are decompressed by performing 1×1 convolution operations on the interactive face image and audio clip features.

2) Interaction Feature Combination: The cross-modal matching task aims to distinguish both intra-modality samples and eliminate inter-modality heterogeneity. While the initial features extracted from the data can differentiate intra-modal diversity, inter-modal interaction features help to correlate cross-modal samples. However, the two types of features differ from each other. Therefore, a method is needed to effectively fuse these two features. In this paper, we propose the Interaction Feature Combination (IFC) that can fuse cross-modal interaction features with the original features to explore valuable distinguishing category features.

To reduce the instance differences between inter-modal interaction features, we adopt instance normalization (IN) [52]. Then, we use SENet [53] to estimate mask values. We apply these masks to the corresponding audio and visual features before fusing them to obtain the final multimodal features.

$$\hat{\boldsymbol{f}}^{A}_{i_0} = m_0\boldsymbol{f}^{a}_{i_0} + (1 - m_0) \odot \ IN(\boldsymbol{f}^{A}_{i_0}), \tag{8}$$

$$\hat{\boldsymbol{f}}^{V}_{i_k} = m_k\boldsymbol{f}^{v}_{i_k} + (1 - m_k) \odot \ IN(\boldsymbol{f}^{V}_{i_k}), \tag{9}$$

where $\odot$ refers to the element-wise product. m_0 and m_k denote the sets of identity-related channels in the audio clip and the kth face image, respectively.

$$m_0 = SENet(avg(\boldsymbol{f}^{A}_{i_0}) + max(\boldsymbol{f}^{A}_{i_0})), \tag{10}$$

$$m_k = SENet(avg(\boldsymbol{f}^{V}_{i_k}) + max(\boldsymbol{f}^{V}_{i_k})), \tag{11}$$

where $avg(\cdot)$ and $max(\cdot)$ refer to global average pooling and global maximum average pooling, respectively. The purpose of these two pooling approaches is to ensure that the learned features are well-represented from multiple perspectives.

3) Identity Alignment Loss: The existing models have not adequately addressed the issue of identity similarity, which is critical for facilitating the interaction of features among modalities. To address this issue, we propose incorporating an identity alignment loss, which measures the disparity between the estimated identity similarity and the actual identity label. This measure can effectively guide Compact Bilinear Pooling Networks (CBPNet) [27] to more accurately estimate identity similarity.

Identity similarity can be derived using Eq. (5), which is determined by the values assigned to positive and negative identity labels. The identity alignment loss is computed by taking the difference between the activated identity similarity and the identity label as follows:

$$\boldsymbol{I}_{i_j} = (sigmoid(\tau\ S_{i_{0j}}) - l_{i_j} P_{i_j})^2, \tag{12}$$

where $l_{i_j} \in [1, k]$ is the matched identity label and P_{i_j} denotes the jth identity mask value. τ is the temperature control parameter and is set to 5. In addition, we introduce a modification to the CBP [27] to improve the accuracy of learning identity correlation in hard samples. The simulated annealing weights [28] are added to mitigate the negative impact of hard samples. The modified loss function is presented below:

$$\mathcal{L}_{IA} = \frac{1}{2k}(1 + \cos(\frac{epoch}{N}\pi)) \sum_{j=1}^{k} I_{i_j}, \tag{13}$$

where $epoch$ is the number of iterations and N represents the total number of iterations.

C. Attribute-guided Enhancement Module

The purpose of the attribute-guided enhancement module is to identify subtle features associated with attributes and enhance their relevance to improving audio-visual attribute hierarchy matching ability. Prior works have employed multiple attribute labels as supervised labels to train networks for feature representation learning [18]. However, using multiple attribute losses can result in challenges in achieving superior network performance due to their interactions with each other.

To address this limitation, we propose a novel attribute-guided enhancement module that focuses on learning subtle features related to gender and nationality. Our approach enhances the discrimination of these features by decompressing attribute features and directing attention to them. Inspired by the squeeze and excitation network [53], we design a simple coding-decoding network with two fully connected (FC) layers and a ReLU activation layer. The first step is to reduce the feature representation through the FC layers and apply the ReLU activation as follows:

$$\tilde{\boldsymbol{f}^{a}_{i_0}} = \delta(Fc(\mathbf{T}(\boldsymbol{f}^{a}_{i_0}))), \tag{14}$$

$$\tilde{\boldsymbol{f}^{v}_{i_k}} = \delta(Fc(\mathbf{T}(\boldsymbol{f}^{v}_{i_k}))), \tag{15}$$

where $\mathbf{T}$ is the feature transformation operation. The feature classification network uses $\tilde{\boldsymbol{f}^{a}_{i_0}}$ and $\tilde{\boldsymbol{f}^{v}_{i_k}}$, which are the vectorized feature, to classify attributes. The resulting attribute classification loss can be represented as follows:

$$\mathcal{L}_{Att} = -\frac{1}{kM} \sum_{i=1}^{M} (kY_{i_0} \log C_{att}(\tilde{\boldsymbol{f}^{a}_{i_0}}) + \sum_{j=0}^{k} Y_{i_j} \log C_{att}(\tilde{\boldsymbol{f}^{v}_{i_j}})), \tag{16}$$

where Y_{i_0} and Y_{i_j} are the audio attribute labels and the face image labels of the ith tuple, respectively. C_{att} denotes the attribute classification. The gender distribution is relatively balanced, while the nationality distribution has a severe long-tail distribution. To address this issue, we divide the nationality attributes into American, British, and other nationalities, thereby avoiding the long-tail problem. The attribute labels Y_{i_0} and Y_{i_j} consist of gender and nationality values in the range of $[0, 5]$. Here, M represents the number of training data tuples.

Subsequently, we decode $\tilde{\boldsymbol{f}^{a}_{i_0}}$ and $\tilde{\boldsymbol{f}^{v}_{i_k}}$ into attention values, which are then multiplied with the original attribute features as follows:

$$\overline{\boldsymbol{f}^{a}_{i_0}} = softmax(FC(\tilde{\boldsymbol{f}^{a}_{i_0}}))\boldsymbol{f}^{a}_{i_0}, \tag{17}$$

$$\overline{\boldsymbol{f}^{v}_{i_k}} = softmax(FC(\tilde{\boldsymbol{f}^{v}_{i_k}}))\boldsymbol{f}^{v}_{i_k}, \tag{18}$$

where, $\overline{\boldsymbol{f}^{a}_{i_0}}$ and $\overline{\boldsymbol{f}^{v}_{i_k}}$ are the enhanced common attribute features. We then add them to the corresponding inter-modal interaction identity attribute features $\hat{\boldsymbol{f}^{A}_{i_0}}$ and $\hat{\boldsymbol{f}^{V}_{i_k}}$, respectively. Finally, we eliminate the modal heterogeneity between the audio and visual features through the use of a generative adversarial network (GAN) [20].

D. Objective Function

In order to eliminate modal heterogeneity, we employ a method whereby we feed audio feature $\boldsymbol{f}^{a}_{i_0}$ and face images features $\{\boldsymbol{f}^{v}_{i_1}, \cdots, \boldsymbol{f}^{v}_{i_k}\}$ into the GAN, resulting in the generation of modality-independent features $\{\boldsymbol{h}_{i_0}, \cdots, \boldsymbol{h}_{i_k}\} \in \mathcal{H}$. The discriminator D is trained through a minimax two-player game, where it discriminates the $\boldsymbol{h}_{i_j}$ features, classifying them as belonging to either the visual or audio modality.

$$\mathcal{L}_{disc} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=0}^{k} N_{i_j} \log D(\boldsymbol{h}_{i_j}), \tag{19}$$

where N_{i_j} represents the modality label of the j-th sample in the i-th data tuple, and $D(\boldsymbol{h}_{i_j})$ denotes the modality probability of the output of D. The number of training data tuples is denoted by M.

Then, we utilize a fully connected neural network due to its nonlinear fitting capability in finding matching candidates. To achieve this, the feature residuals of the anchor sample features and each face image feature are computed and concatenated as the input to the matching classifier. The loss is computed using the common cross-entropy method [54].

$$\mathcal{L}_{Cls} = -\frac{1}{M} \sum_{i=1}^{M} (l_i \log C_m([\boldsymbol{h}_{i_0} - \boldsymbol{h}_{i_1}, \cdots, \boldsymbol{h}_{i_0} - \boldsymbol{h}_{i_k}]), \tag{20}$$

where C_m denotes the matching classification.

Inspired by Peng *et al.* [55] and Zheng *et al.* [20], we introduce a contrast loss to enhance network convergence. This loss serves to bring intra-class samples closer while simultaneously pushing inter-class samples farther apart. We display the following:

$$\mathcal{L}_{Contrast} = \frac{1}{2M} \sum_{i=1}^{M} max(D_i, 0), \tag{21}$$

$$D_i = log(\max_{j\in[1,k]} w_{l_i} e^{\theta - d_{i_0,i_j}} + \max_{q\in[1,k]} w_{l_i} e^{\theta - d_{i_1,i_q}} + d_{i_0,i_1}), \tag{22}$$

where w_{l_i} represents the matching label mask. If it is a match, w_{l_i} equals 0, otherwise, it equals 1. The Euclidean distance,

d_{i_0,i_1}, is utilized to compute the distance between the paired anchors, $\boldsymbol{h}_{i_0}$ and positive samples, $\boldsymbol{h}_{i_p}$ $(p \in [1,k])$. Similarly, d_{i_1,i_q} is used to calculate the Euclidean distance between same-modal anchors, $\boldsymbol{h}_{i_p}$, and negative samples, $\boldsymbol{h}_{i_q}$, whereas d_{i_1,i_j} measures the distance between cross-modal anchors, $\boldsymbol{h}_{i_0}$, and negative samples, $\boldsymbol{h}_{i_j}$. To better distinguish the candidate matching samples that are close to each other, the negative instances are activated by the maximum value. Moreover, a hyperparameter, θ, has been set to 1.2.

The total loss is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{disc} + \alpha\mathcal{L}_{IA} + \beta\mathcal{L}_{Contrast} + \gamma\mathcal{L}_{Cls} + \lambda\mathcal{L}_{Att}, \quad (23)$$

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

IV. Experiments

A. Implementation Details

Network architecture. We conducted all our experiments using an NVIDIA GeForce RTX 3090 graphics card. To ensure a fair comparison with advanced audio-visual cross-modal methods, we maintained the previous feature extraction structure, which was based on [25], for our feature extraction. For the image encoder, we used ResNet18 [49], pre-trained on ImageNet [51], as the feature extractor. Regarding the audio network, we used a three-layer multilayer perceptron as a starting point and then applied SE-ResNet-34 [49] without a pre-trained model as a feature extractor. The read-face images had dimensions of $224 \times 224 \times 3$, and the audio clips had a sequence length of 160000. Both audio and image features had dimensions of $512 \times 3 \times 3$ after each feature extractor. The input and output of the attribute-guided interaction and enhancement modules were consistent in their feature dimensions. To keep the output feature dimensions of both attribute guidance modules constant, we performed feature summation, which then vectorized the features into 4608-dimensional features. The identity features of the audio-visual pairs were fed into the adversarial network, which converted the features into 256 dimensions and then to 128 dimensions for modality-independent audio-visual features. A binary classification network was used as the discriminator to classify the probability of the corresponding modality. Finally, we used a fully connected network to obtain the matching results.

Training parameters. During training, we set the batch size to 50 and used Adaptive Moment Estimation (Adam) [56] with a momentum of 0.9 and weight decay of 0.0005 to fine-tune the network. The initial learning rates for each module were as follows: feature extractor (5×10^{-2}), attribute-guided interaction module (5×10^{-3}), attribute-guided enhancement module (5×10^{-3}), generator (5×10^{-3}), discriminator (5×10^{-3}), and matching classifier (5×10^{-2}). The delay was set to 0.1 at the 20th and 35th epochs. Depending on the matching case settings, the $k * 128$-dimensional features were divided into k output classes to represent the matching probabilities. For the validation trial, we treated it as a special matching task, which had only one candidate objective $(k = 1)$ to determine whether a match or not. In the matching task, there were k $(k >= 2)$ candidate matching samples in addition to the anchor sample, which was combined into a $k{*}128$-dimensional feature to represent the probability of matching between them. The classification network outputted k dimensions to calculate the probability of matching. We measured the cross-modal matching performance using accuracy (ACC) [25].

Dataset. We evaluated the performance of ACIENet, as proposed in this study, on the publicly available datasets Voxceleb [30] and VGGFace [31]. These datasets contain a total of 137,060 face images and 149,354 audio clips, respectively, and have 1,225 paired audio and visual data. To ensure a fair comparison, we followed the prevalent evaluation protocol [18], [25] presented in Table I for data sampling, analysis, and validation of our main experiments. Additionally, we employed another evaluation protocol [20], [32] to complement the experimental validity.

TABLE I
The data splitting to training, validation, and testing after sampling.

Item	Train	Validation	Test	Total
Identities	924	112	189	1225
Face Images	104724	12260	20076	137060
Audio Clips	113322	14182	21850	149354

B. Comparison Results

In this study, we evaluate the effectiveness of ACIENet by comparing it with seven state-of-the-art algorithms, namely SVHF-Net [15], DIMNet [18], Wang's [16], Wen's [25], AML [20], DCLR [57], and DSANet [21]. To demonstrate the efficacy of ACIENet, we perform audio-visual verification and matching tasks in both V-F (visual to audio) and F-V (audio to visual) scenarios. These tasks are illustrated in Table II. Our Baseline methodology involves advanced feature extraction, inspired by Wen *et al.* [25], which aims to obtain feature representations. Additionally, we employ generative adversarial networks (GANs) to mitigate modal heterogeneity, following the approach proposed by Wang *et al.* [21]. We also employ distance metrics to constrain the intra- and inter-modal feature distribution, as suggested by Zheng *et al.* [20]. It is worth noting that this approach achieves competitive performance when compared to existing state-of-the-art methods.

We have added the comparison with transformer-based methods in Table II. As we can see from them, transformer-based [58]–[60] models perform significantly worse than CNNs. The main reason may be that the transformer is susceptible to overfitting [61], [62], which may not be suitable for robust audio-visual matching in noisy medium-scale cross-modal data, especially in F-V scenarios. However, the proposed ACIENet method displays the potential to further enhance verification and matching task performance, building upon the foundation of the CNN baseline. Specifically, in V-F and F-V binary matching scenarios, the ACC accuracy of ACIENet is 3.89% and 4.41% higher than the existing state-of-the-art method, respectively. We also investigate the

TABLE II
THE QUALITATIVE RESULTS OF MATCHING TASKS. VERIFICATION INDICATES WHETHER $k = 1$ IS MATCHED OR NOT. BINARY DENOTES THE 1:2 MATCHING WHILE MULTI-WAY DENOTES THE $1 : k$ ($k = 10$) MATCHING. V-F REPRESENTATIVE FOR AUDIO AS AN ANCHOR TO MATCH GALLERY FACES. F-V REPRESENTS THE FACE AS AN ANCHOR TO MATCH THE AUDIO OF THE GALLERY. THE EXPERIMENTAL RESULTS IN THE TABLE ARE OBTAINED BY THE DATA SETTINGS PROPOSED BY WEN *et al.* [25].

Methods	Backbone	Venue	Binary (ACC)		Multi-way (ACC)		Verification (AUC)	
			V-F	F-V	V-F	F-V	V-F	F-V
SVHF [15]	CNN	CVPR2018	81.0	79.5	34.5	×	-	-
DIMNet [18]		ICLR2019	81.3	81.9	38.4	36.2	81.0	81.2
Wang's [16]		ACM2020	83.4	84.2	39.7	36.4	82.6	82.9
Wen's [25]		CVPR2021	87.2	86.5	48.2	44.8	87.2	87.0
AML [20]		TMM2021	90.2	86.3	46.2	43.7	86.4	86.2
DCLR [57]		ICDM2022	86.79	87.45	-	-	86.76	86.89
DSANet [21]		TMM2022	92.5	88.4	49.1	46.8	87.4	91.5
Transformer [58]	Transformer	NIPS2017	88.9	76.3	36.6	23.5	82.3	75.3
Mobilevit [59]		ICLR2022	94.1	88.2	48.2	33.6	87.7	89.5
Fastvit [60]		ICCV2023	94.2	77.1	37.3	20.1	82.7	80.5
Baseline	CNN	Ours	94.8	89.8	48.5	45.6	88.2	91.2
ACIENet	CNN	Ours	**96.0**	**92.3**	**49.5**	**47.1**	**90.1**	**91.9**

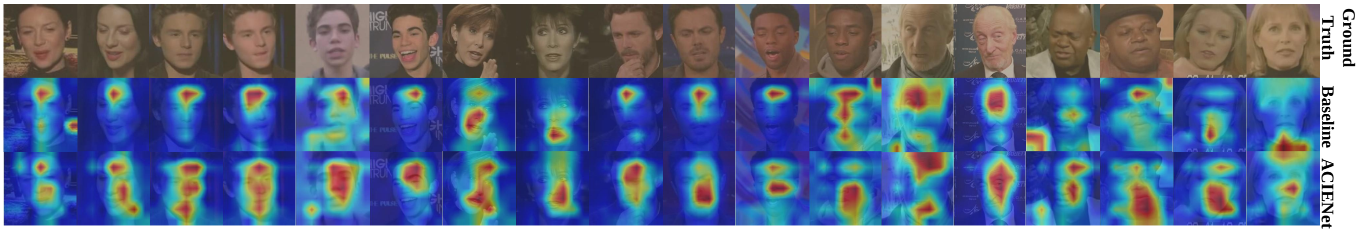


Fig. 4. Class activation maps (CAM) generated by the proposed ACIENet compared with the baseline. Visualization of features extracted from face images with different identity, gender, and nationality attributes is displayed.

performance of ACIENet in a more complex and challenging multi-way matching task, where $k = 10$, which involves multiple candidate objectives. As expected, the performance of ACIENet degrades in this task, but its accuracy in V-F and F-V scenarios is still 0.80% and 0.64% higher than the current state-of-the-art method, respectively.

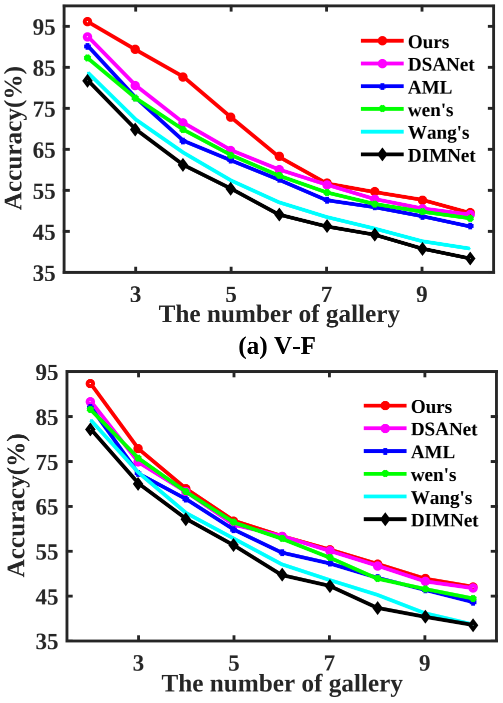
The performance of the audio-visual matching in the binary and multi-way cases is usually better with the V-F scenario. According to Wei *et al.* [66], audio signals are more susceptible to environmental noise than facial images, which display higher intra-class similarity. This results in lower performance in F-V scenarios. There are some state-of-the-art (SOTA) algorithms based on the data splitting scheme in Person Identification Networks (PINs) [32] for experiments, as reported by Nagrani *et al.* [32]. ACIENet is one of these algorithms, which also uses this data splitting scheme. As illustrated in Table III, ACIENet consistently outperforms the state-of-the-art methods in verification and matching tasks in all but the EFT's [65] verification results, thus validating the effectiveness of our proposed method. In the verification experiments, ACIENet, employing a single expert across two distinct unimodal modalities, attains a level of performance second only to EFT's [65] with multi-expert fusion. Due to the inconsistency of the data splitting, we only compare the results of state-of-the-art methods with the same data settings as the PINs [32] method in Table III for fairness. Note that we only compare V-F results since the state-of-the-art methods only provide V-F data-splitting for verification. Additionally, we provide a visualization of the features extracted by ACIENet in Fig. 4. The results indicate that ACIENet has the ability to focus on a wider range of valid features for a person with different identity, gender, and nationality attributes. For face images of different scenes of the same person, ACIENet focuses on a higher overlap part between the salient feature regions, indicating its capability of associating same-modality attribute information. Unless specifically stated, all subsequent experiments follow Wen 's [25] data-splitting scheme.

We conducted $1 : k$ multi-way cross-modal matching experiments to further validate the superiority of ACIENet. As the number of matching candidate objectives increases, the intra-class similarity also increases, leading to a gradual rise in cross-modal matching difficulty. Fig. 5 illustrates that the ACIENet method achieves competitive performance, but its matching performance gradually decreases with increasing k in both V-F and F-V scenarios. Notably, ACIENet outperforms other methods in the V-F scenario due to the significant intra-class discrepancy between face images and audio signals. When k is small, the ACIENet method achieves relatively high performance because there is a higher probability of attribute differences between matched candidate targets. However, with

TABLE III
COMPARISON RESULTS OF AUDIO-VISUAL MATCHING WITH THE STATE-OF-THE-ART METHOD IN THE BINARY ($k = 2$) AND MULTI-BINARY ($k = 10$) CASES. VERIFICATION INDICATES WHETHER $k = 1$ IS MATCHED OR NOT. WHERE " -" MEANS "NOT AVAILABLE". THE EXPERIMENTAL RESULTS IN THE TABLE ARE OBTAINED FOLLOWING THE DATA SETTINGS PROPOSED BY PINS [32].

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (AUC)	
		V-F	F-V	V-F	F-V	V-F	F-V
DIMNet [18]	ICLR2019	84.12	84.03	39.75	-	83.2	-
PINs [32]	ECCV2018	84.00	-	31.00	-	78.5	-
SSNet [17]	DIC2019	78.00	78.50	30.00	30.05	78.8	-
β-VAE [24]	TMM2021	84.15	84.22	41.30	40.02	84.64	-
AML [20]	TMM2021	92.72	93.3	43.45	39.35	80.6	-
CMPC [63]	IJCAI2022	82.2	81.7	-	-	84.6	-
FOP [23]	ICASSP2022	89.3	83.5	-	-	83.5	-
DSANet [21]	TMM2022	95.25	94.28	46.83	43.36	78.0	-
SBNet [64]	ICASSP2023	82.4	82.4	-	-	82.5	-
EFT [65]	ICME2023	89.6	89.6	-	-	**90.1**	-
ACIENet	Ours	**96.4**	**95.6**	**46.9**	**44.1**	84.8	-

an increasing number of matched candidate objectives, the method may weaken its attribute discrimination ability due to the presence of hard samples with the same attributes.



(a) V-F

(b) F-V

Fig. 5. The quantitative results of $1 : k$ matching task in V-F and F-V scenarios.

C. Ablation Study

Evaluation of Different Component Effectiveness. We conducted ablation experiments on the verification and binary matching tasks to assess the effectiveness of each component of ACIENet. The outcomes of these experiments are presented in Table IV. Specifically, the attribute-guided interaction (AGI) module is introduced to explore potential local feature associations using cross-modal identity-aligned similarity-guided interaction matrices. In comparison to existing aligned cross-modal matching methods, the AGI module aligns cross-modalities of the same identity and distinguishes the differences in cross-modal features of different identities simultaneously. Additionally, the attribute-guided enhancement (AGE) module is developed to complement the AGI module by guiding the network to learn attribute-related features, which enhances attribute discriminability between different identities. Table IV depicts the experimental performance of the AGI and AGE modules individually and jointly in the V-F and F-V scenarios. These tables' results demonstrate that each module is effective and performs best when combined with the other modules. Moreover, to further corroborate the efficacy of these components, we conduct experiments across various configurations, including 5-way and 10-way tasks. These experiments unveil a swift deterioration in matching performance as the number of matching candidates escalated. Nonetheless, the two modules consistently exhibit their effectiveness.

TABLE IV
THE ACIENET METHOD IS CONDUCTED IN V-F AND F-V SCENARIOS FOR VERIFICATION (WHEN $k = 1$), BINARY (WHEN $k = 2$), AND 10-WAY (WHEN $k = 10$) AUDIO-VISUAL MATCHING TASKS ON ABLATION STUDIES. '✓' MEANS THE CORRESPONDING COMPONENT IS INCLUDED.

Component		Binary		Verification		5-way		10-way	
AGI	AGE	V-F	F-V	V-F	F-V	V-F	F-V	V-F	F-V
		94.8	89.8	88.2	91.2	73.4	59.8	48.5	45.6
✓		95.4	91.6	88.9	91.6	74.2	60.5	49.0	46.1
	✓	95.8	91.9	89.9	91.7	74.5	61.2	49.2	46.8
✓	✓	**96.0**	**92.3**	**90.1**	**91.9**	**74.8**	**61.6**	**49.5**	**47.1**

Evaluation on Attribute-guided Interaction Module. The attribute-guided interaction (AGI) module comprises three components, namely the inter-model interaction (IMI) structure, the interaction feature combination (IFC), and the identity alignment loss ($\mathcal{L}_{IA}$). To evaluate the necessity of these components, we performed ablation tests on the baseline model. The results, presented in Table V (b), indicate that the AGI module has a significant impact on the model's performance on both matching and verification tasks. The root cause of the performance degradation is that the AGI module learns cross-modal interaction features for all candidate matching targets without regard to whether they can match each other, leading to the learning of useless interaction semantic relations. To address this issue, we used identity similarity to guide feature interactions between modalities explicitly, which enabled effective feature interactions. Additionally, the IFC facilitates the fusion of cross-modal interaction features with original features, thereby fully exploiting feature distinguishability. The results in Table V (c) demonstrate that the IFC can substantially improve the model's cross-modal matching performance. Moreover, we used the identity alignment loss ($\mathcal{L}_{IA}$) to constrain identity similarity, and simulated annealing weights to improve model generalization for hard-to-match samples. As shown in Table V (c) compared with (d), these measures further improved the model's performance. Overall, each of the proposed components contributes positively to the AGI module, and the best performance can be achieved by using all three components.

TABLE V
ABLATION EXPERIMENTS ON THE PROPOSED ATTRIBUTE-GUIDED INTERACTION (AGI) MODULE PERFORM AUDIO-VISUAL MATCHING TASKS IN THE VALIDATION AND BINARY (WHEN $k = 2$) CASES.

	AGI			Binary ($k = 2$)		Verification	
	IMI	IFC	$\mathcal{L}_{IA}$	V-F	F-V	V-F	F-V
a				94.8	89.8	88.2	91.2
b	✓			93.2	89.3	75.4	88.1
c	✓	✓		95.2	91.2	88.4	91.4
d	✓	✓	✓	**95.4**	**91.6**	**88.9**	**91.6**

Evaluation on Interaction Feature Combination. We analyzed to evaluate the impact of Interaction Feature Combining (IFC) operation on the network performance. To achieve this, we performed three different IFC operations on the ACIENet method. The three operations, marked as Residual IFC (a), Bidirectional mask IFC (b), and Adaptive IFC (c), are designed to combine original and interaction features using feature combinations. Fig. 6 shows the schematic of these operations. We observed that all three IFC operations have varying degrees of importance in the audio-visual cross-modal matching task. While Adaptive IFC (c) did not consistently achieve optimal performance in some experiments, it generally demonstrated competitive performance across different tasks and scenarios. The results are presented in Table VI. Therefore, we used the adaptive interaction feature combination as the feature fusion manner on the ACIENet method.

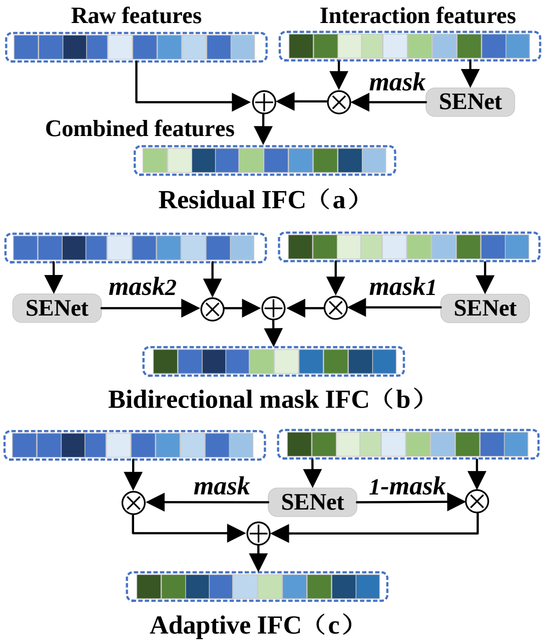


Fig. 6. Comparison of the three interaction feature combination operations.

TABLE VI
THE EXPERIMENTS ARE CONDUCTED TO COMPARE THE THREE INTERACTION FEATURE COMBINATION OPERATIONS IN BINARY (WHEN $k = 2$) AND 5-WAY (WHEN $k = 5$) FOR THE AUDIO-VISUAL MATCHING TASK IN THE V-F AND F-V SCENARIOS.

Demo	Methods	Binary		5-way	
		V-F	F-V	V-F	F-V
	Baseline	94.8	89.8	73.4	59.8
Fig. 6 (a)	R-IFC (a)	95.5	91.6	71.9	60.3
Fig. 6 (b)	BM-IFC (b)	95.7	91.8	72.1	61.2
Fig. 6 (c)	A-IFC (c)	95.4	91.6	74.2	60.5

Evaluation on Attribute-guided Enhancement Module. To evaluate the efficacy of the proposed Attribute-guided Enhancement (AGE) model, we compared it with a joint supervised approach that comprises three attributes suggested by Wen *et al.* [18]. The results of the comparison are presented in Table VII. Unlike the joint supervised approach, the AGE model is decoupled from the attribute features and operates in a parallel manner. This allows for superior performance by avoiding the mutual influence between multiple attribute losses. Notably, the ACIENet method, unlike Wen *et al.* [18] method, does not require the excessive adjustment of loss weights to achieve optimal results.

TABLE VII
COMPARING DIFFERENT WAYS OF UTILIZING ATTRIBUTE FEATURES.

Methods	Binary ($k = 2$)		Verification	
	V-F	F-V	V-F	F-V
Attribute-Serial [18]	93.7	88.6	90.0	91.2
Attribute-Parallel (Ours)	**96.0**	**92.3**	**90.1**	**91.9**

Evaluation on the Impact of Different Attributes on the AGE Module. In order to assess the impact of dif-

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.2024.3388949

ferent attribute-guided enhancement models on network performance, we utilized various attribute combinations on ACIENet. Our objective was to evaluate the performance of the enhanced features. Specifically, we incorporated two attributes: gender and nationality, and examined three attribute combinations: gender only, nationality only, and joint gender and nationality supervision. These attribute combinations corresponded to the three sets of experimental cases illustrated in Table VIII (b), (c), and (d), respectively. As depicted in Table VIII, the attribute-guided feature enhancements improved the discriminative ability of the features compared to the baseline network. The optimal performance is observed when the two attributes are co-supervised. This result highlights the potential benefits of the attribute-guided feature enhancement module in audio-visual cross-modal matching tasks.

TABLE VIII

COMPARISON OF THE PERFORMANCE OF DIFFERENT ATTRIBUTE-GUIDED FEATURE ENHANCEMENTS IN A BINARY AUDIO-VISUAL MATCHING TASK.

	Different Attribute-guided	V-F	F-V
a	Baseline	94.8	89.8
b	+ Gender	95.4	91.4
c	+ Nationality	95.3	91.1
d	+ Gender + Nationality	**95.8**	**91.9**

D. *Evaluation on Different Interaction Stages*

To evaluate the impact of the attribute-guided interaction module at different feature extraction stages, we have conducted experiments on early, middle, and late features. First, the attribute-guided interaction (AGI) model exhibits superior performance on late features compared to the baseline (Table IX (a)) as well as the early and middle feature interactions (Table IX (b) and (c)). Furthermore, the best performance can be achieved when integrating both AGE and AGI modules on late features, as shown in Table IX (h) compared to Tables IX (e), (f), and (g), which evidences the effectiveness of reducing interference between multiple attributes at the late stage. The overall result is that the late-stage features present a more comprehensive understanding of attribute semantics to the extent of learning meaningful cross-modal feature interactions.

TABLE IX

ANALYSIS OF AUDIO-VISUAL MATCHING EXPERIMENTS IN THE VALIDATION AND BINARY (WHEN $k = 2$) CASES FOR ATTRIBUTE-GUIDED INTERACTION (AGI) MODULES INSERTED INTO THE FEATURE EXTRACTION NETWORK'S EARLY, MIDDLE, AND LATER STAGES, RESPECTIVELY.

	AGE	AGI			Binary ($k = 2$)		Verification	
		Early	Middle	Late	V-F	F-V	V-F	F-V
a					94.8	89.8	88.2	91.2
b		✓			90.5	72.0	81.9	84.0
c			✓		92.6	85.2	88.2	77.5
d				✓	95.4	91.6	88.9	91.6
e	✓				95.8	91.9	89.9	91.7
f	✓	✓			93.6	75.2	85.5	87.0
g	✓		✓		94.5	86.6	89.6	88.9
h	✓			✓	**96.0**	**92.3**	**90.1**	**91.9**

E. *Hyper-parameters Analysis*

Fig. 7 presents the weights of hyperparameters for multiple losses in Eq. (23), which are determined by the control variables α, β, γ, and λ. These variables represent the weights of identity alignment loss, modality metric loss, matching classification loss, and attribute classification loss, respectively, in the cross-modal matching task. In the V-F and F-V scenarios, the first three hyperparameters result in slightly fluctuating performance for the cross-modal matching task, but they all outperform the existing state-of-the-art methods. However, for the same task, as the value of the parameter λ increases gradually, the performance of ACIENet initially increases before slowly declining. This gradual increase in λ results in a slowly decreasing accuracy of cross-modal matching, which indicates that the optimization with parallel attribute-guided interaction (AGI) and attribute-guided enhancement (AGE) modules is working effectively. Based on the experimental analysis for V-F and F-V scenarios, setting $\alpha = 3$, $\beta = 1$, $\gamma = 2$, and $\lambda = 1$, and $\alpha = 1$, $\beta = 1$, $\gamma = 1$, and $\lambda = 1$ respectively, ACIENet achieves excellent performance.

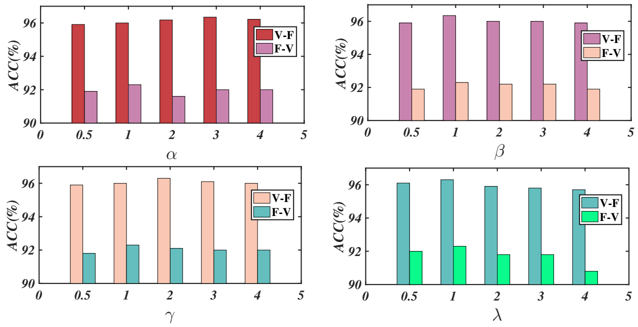


Fig. 7. The effects of hyperparameters of α, β, γ, and λ on binary matching task.

V. CONCLUSION

In this paper, we present ACIENet, an attribute-guided interaction and enhancement network that includes two modules: the attribute-guided interaction (AGI) module, which explores the semantic relationships between cross-modal features, and the attribute-guided enhancement (AGE) module, which enhances local attribute-related feature representations. The AGI module is further divided into three parts: inter-modal interaction (IMI) structure, interaction feature combination (IFC), and identity alignment loss ($\mathcal{L}_{IA}$). The combination of these three parts produces an efficient local feature interaction between modalities under global identity feature alignment. To enhance attribute-related subtle features, we propose an AGE module that focuses on local features corresponding to gender and nationality attributes. This module obtains attention weights by means of decoding structures and thus enhances these attributes. In our experiments, we demonstrate that

ACIENet outperforms several other state-of-the-art methods for cross-modal matching and validation on both Voxceleb and VGGFace datasets.

References

[1] V. Bruce and A. Young, "Understanding face recognition," *British journal of psychology*, vol. 77, no. 3, pp. 305–327, 1986.

[2] P. Belin, P. E. Bestelmeyer, M. Latinus, and R. Watson, "Understanding voice perception," *British Journal of Psychology*, vol. 102, no. 4, pp. 711–725, 2011.

[3] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention, Perception, & Psychophysics*, vol. 78, no. 3, pp. 868–879, 2016.

[4] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "Putting the face to the voice': Matching identity across modality," *Current Biology*, vol. 13, no. 19, pp. 1709–1714, 2003.

[5] A. W. Young, S. Frühholz, and S. R. Schweinberger, "Face and voice perception: Understanding commonalities and differences," *Trends in Cognitive Sciences*, vol. 24, no. 5, pp. 398–410, 2020.

[6] 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.

[7] 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.

[8] A. George, A. Mohammadi, and S. Marcel, "Prepended domain transformer: Heterogeneous face recognition without bells and whistles," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 133–146, 2022.

[9] Y. Fang, Z. Xiao, W. Zhang, Y. Huang, L. Wang, N. Boujemaa, and D. Geman, "Attribute prototype learning for interactive face retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2593–2607, 2021.

[10] 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.

[11] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2019.

[12] 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.

[13] 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*, 2021.

[14] 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.

[15] 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.

[16] 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.

[17] 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 *Digital Image Computing: Techniques and Applications*, pp. 1–7, IEEE, 2019.

[18] 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.

[19] 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.

[20] 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.

[21] 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*, pp. 1–12, 2022. doi: 10.1109/TMM.2022.3222936.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 27, 2014.

[23] 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.

[24] 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.

[25] 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.

[26] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10941–10950, 2020.

[27] 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.

[28] 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.

[29] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, 2022.

[30] 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.

[31] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, pp. 41.1–41.12, 2015.

[32] 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.

[33] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," *International Conference on Learning Representations*, 2020.

[34] X. Tu, Y. Zou, J. Zhao, W. Ai, J. Dong, Y. Yao, Z. Wang, G. Guo, Z. Li, W. Liu, *et al.*, "Image-to-video generation via 3d facial dynamics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1805–1819, 2021.

[35] J. Sun, Q. Li, W. Wang, J. Zhao, and Z. Sun, "Multi-caption text-to-face synthesis: Dataset and algorithm," in *Proceedings of the International Conference on Multimedia*, pp. 2290–2298, 2021.

[36] J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim, "Integrated face analytics networks through cross-dataset hybrid training," in *Proceedings of the International Conference on Multimedia*, pp. 1531–1539, 2017.

[37] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proceedings of the International Conference on Multimedia*, pp. 792–800, 2018.

[38] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 34, pp. 9694–9705, 2021.

[39] A. Zheng, P. Pan, H. Li, C. Li, B. Luo, C. Tan, and R. Jia, "Progressive attribute embedding for accurate cross-modality person re-id," in *Proceedings of the ACM International Conference on Multimedia*, pp. 4309–4317, 2022.

[40] K. Zhang, Z. Mao, Q. Wang, and Y. Zhang, "Negative-aware attention framework for image-text matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15661–15670, 2022.

[41] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 279–286, 2020.

[42] S. Liu, W. Quan, C. Wang, Y. Liu, B. Liu, and D.-M. Yan, "Dense modality interaction network for audio-visual event localization," *IEEE Transactions on Multimedia*, pp. 1–1, 2022. doi: 10.1109/TMM.2022.3150469.

[43] R. G. Praveen, W. C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. L. Koerich, S. Bacon, P. Cardinal, *et al.*, "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2486–2495, 2022.

[44] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2617–2629, 2021.

[45] Z. Ji, H. Wang, J. Han, and Y. Pang, "Sman: stacked multimodal attention network for cross-modal image-text retrieval," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 1086–1097, 2020.

[46] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3536–3545, 2020.

[47] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the ACM International Conference on Multimedia*, pp. 3884–3892, 2020.

[48] 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.

[49] 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.

[50] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *Proceedings of the International Conference on Learning Representations*, 2022.

[51] 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.

[52] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision*, pp. 464–479, 2018.

[53] 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.

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

[55] Y. Peng and J. Qi, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1, pp. 1–24, 2019.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv Preprint arXiv::1412.6980*, 2014.

[57] 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.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[59] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," 2022.

[60] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[61] B. Li, Y. Hu, X. Nie, C. Han, X. Jiang, T. Guo, and L. Liu, "Dropkey for vision transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22700–22709, 2023.

[62] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, *et al.*, "Robustart: Benchmarking robustness on architecture design and training techniques," *ArXiv Preprint arXiv:2109.05211*, 2021.

[63] 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.

[64] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[65] G. Chen, D. Zhang, T. Liu, and X. Du, "Eft: Expert fusion transformer for voice-face association learning," in *IEEE International Conference on Multimedia and Expo*, pp. 2603–2608, 2023.

[66] Y. Wei, D. Hu, Y. Tian, and X. Li, "Learning in audio-visual context: A review, analysis, and new perspective," *arXiv preprint arXiv:2208.09579*, 2022.